**КОМП'ЮТЕРНІ НАУКИ**

**УДК 004.738.5** https://doi.org/10.31713/vt3202317

**Красько Б. В., аспірант, Грицюк П. М., д.е.н., професор** (Національний університет водного господарства та природокористування, м. Рівне, b.v.krasko@nuwm.edu.ua, p.m.hrytsiuk@nuwm.edu.ua)

# ЕФЕКТИВНИЙ МОНІТОРИНГ ТА УПРАВЛІННЯ РЕСУРСАМИ В AMAZON EC2 ПРИ МАСШТАБУВАННІ

**Amazon Elastic Compute Cloud (Amazon EC2) – це сервіс, який надає гнучку та масштабовану обчислювальну потужність в хмарі Amazon Web Services (AWS). За допомогою Amazon EC2 користувачі можуть запускати віртуальні сервери (екземпляри) з різними конфігураціями процесора, пам'яті, сховища та мережевої ємності, а також налаштовувати безпеку та мережеве з'єднання. Amazon EC2 також дозволяє користувачам моніторити та управляти ресурсами своїх екземплярів за допомогою різних інструментів та сервісів AWS.**

**Моніторинг та управління ресурсами в Amazon EC2 є важливою частиною підтримки надійності, доступності та продуктивності екземплярів та AWS-рішень. Моніторинг дозволяє користувачам виявляти та вирішувати проблеми з продуктивністю, безпекою, доступністю або помилками. Управління дозволяє користувачам оптимізувати використання ресурсів, зменшити витрати, покращити продуктивність та адаптуватися до зміни навантаження.**

**Управління ресурсами в Amazon EC2 полягає в обранні оптимального типу інстансу для певної задачі. Типи інстансів складаються з різних комбінацій процесора, пам'яті, сховища та мережевої пропускної здатності. Можливо вибрати тип інстансу, який найкраще підходить для певного додатку, а також змінювати його за потребою.**

**Масштабування – це процес збільшення або зменшення кількості інстансів Amazon EC2 залежно від потреб певного додатку. Використовуючи автоматичне масштабування для динамічного регулювання ресурсів інфраструктури за допомогою правил та метрик. Автоматичне масштабування допомагає покращити продуктивність, доступність та ефективність додатка.**

***Ключові слова*: Amazon EC2; Amazon CloudWatch хмарні платформи; автомасштабування; моніторинг; управління; балансувальники навантаження.**

**Постановка проблеми.** Amazon EC2 – це сервіс, який надає гнучку та масштабовану обчислювальну потужність в хмарі AWS. За допомогою Amazon EC2 користувачі можуть запускати віртуальні сервери (екземпляри) з різними конфігураціями процесора, пам'яті, сховища та мережевої ємності, а також налаштовувати безпеку та мережеве з'єднання. Amazon EC2 також дозволяє користувачам моніторити та управляти ресурсами своїх екземплярів за допомогою різних інструментів та сервісів AWS.

Однак, моніторинг та управління ресурсами в Amazon EC2 не є простим завданням. Інженери стикаються з такими проблемами, як:

- визначення оптимального типу екземпляра для своїх застосунків за параметрами процесора, пам'яті, сховища та мережевої ємності;
- виявлення та вирішення проблем з продуктивністю, безпекою, доступністю або помилками на своїх екземплярах;
- оптимізація використання ресурсів, зменшення витрат, покращення продуктивності та адаптація до зміни навантаження;
- створення та управління постійними сховищами даних для своїх екземплярів, які можуть бути динамічно змінені за розміром, типом та шифруванням;
- збереження та отримання будь-яких обсягів даних у будь-який час із будь-якого місця;
- створення та управління стеками ресурсів AWS за допомогою шаблонів.

З наведеного вище стає зрозумілим, наскільки важливо для раціонального проєктування хмарної інфракструктури розробити гнучку систему моніторнигу для забезпечення стабільної роботи додатку.

**Аналіз останніх публікацій за темою дослідження.** На момент написання статті є невелика кількість публікацій за темою дослідження. В основному публікації є англомовними, оскільки це новий напрям в IT-технологіях, що активно розвивається. У роботах [1–7] наведено підходи як використовувати допоміжні інструменти для ефективного масштабування та наводять загальні приклади як інтегрувати інструменти з існуючою інфраструктурою.

У цьому дослідженні будуть частково використані результати робіт, цитованих вище.

**Мета роботи** дослідження та розробка ефективного підходу до моніторингу та управління ресурсами в хмарному середовищі Amazon Elastic Compute Cloud (EC2) з урахуванням масштабування системи.

**Виклад основного змісту дослідження**

Amazon EC2 є популярним вибором для обчислення у хмарі, оскільки він дозволяє користувачам запускати та видаляти віртуальні сервери (екземпляри) за потребою, платячи лише за використані ресурси. EC2 також надає різноманітні варіанти масштабування, як-от ручне масштабування, планове масштабування, динамічне масштабування та передбачуване масштабування.

Однак масштабування екземплярів EC2 – це нетривіальна задача. Воно передбачає балансування кількох факторів, таких як продуктивність, вартість, доступність, надійність, безпека та відповідність. Крім того, масштабування потребує постійного моніторингу та управління ресурсами EC2, щоб забезпечити їх ефективну та результативну роботу.

Amazon CloudWatch збирає та відстежує метрики, журнали та події з ресурсів AWS. За допомогою CloudWatch можливо контролювати різні аспекти екземплярів EC2, включаючи використання процесора, мережевий трафік, операції з диском та інше. Шляхом налаштування сповіщень та створення спеціальних панелей інструментів, CloudWatch дозволяє прогнозовано виявляти проблеми продуктивності та оптимізувати використання ресурсів.

Прогнозоване масштабування є функцією EC2 Auto Scaling, яка використовує машинне навчання для передбачення майбутнього попиту для автоматичного налаштування кількості екземплярів відповідно до цього. Прогнозоване масштабування аналізує історичні дані та метрики у реальному часі, щоб створити план масштабування, який передбачає оптимальну кількість екземплярів для кожного періоду часу. Прогнозоване масштабування може допомогти користувачам досягти кращої продуктивності, знизити витрати та підвищити доступність, уникнувши надлишкового або недостатнього виділення ресурсів для екземплярів.

Прогнозоване масштабування працює, створюючи план масштабування на основі таких вхідних даних:

- політика відстеження цілі, яка визначає бажане значення метрики (наприклад, використання процесора чи кількість запитів) та цільове значення (наприклад, 50% або 1000 запитів на хвилину);

- прогнозований горизонт, який визначає, на який термін прогнозоване масштабування має дивитися при передбаченні попиту (наприклад 24 години або 7 днів);
- гранулярність прогнозу, яка визначає, як часто прогнозоване масштабування має оновлювати прогноз (наприклад 5 хвилин або 1 година);
- рівень впевненості, який визначає, наскільки впевненим має бути прогнозоване масштабування при здійсненні передбачень (наприклад, 80% або 95%).

На основі цих даних прогнозоване масштабування генерує прогнозоване значення метрики для кожного періоду часу у прогнозованому горизонті. Потім порівнюється прогнозоване значення метрики з цільовим значенням та розраховує оптимальну кількість екземплярів для кожного періоду часу. В результаті, створюється розклад дій з масштабування, який налаштовує кількість екземплярів згідно з розрахованими значеннями.

Прогнозоване масштабування можна активувати через консоль управління AWS, командний рядок AWS Command Line Interface (CLI) або SDK AWS. Також є функціонал для перегляду графіків попереднього масштабування, які відображають прогнозовані значення метрик, цільові значення, фактичні значення метрик, поточну та заплановану потужність, а також діапазони впевненості. Графіки прогнозованого масштабування можуть допомогти зрозуміти, як працює прогнозоване масштабування та оцінити його продуктивність.

Ключовими показниками моніторингу продуктивності EC2 є наступні: CPU utilization (%), DiskQueueDepth, FreeableMemory, FreeStorageSpace, NetworkReceiveThoughput та ReadLatency.

Відстеження використання процесора (CPU) у відсотках допомагає ідентифікувати екземпляри із високим навантаженням на CPU. Високе використання CPU може призвести до зменшення продуктивності та збільшення часу відповіді. Оптимізуючи розподіл ресурсів на основі використання процесора, допоможе забезпечити безперебійну роботу та економічно ефективне використання ресурсів. На рис. 1 наведено приклад метрики CPU utilization (%).

Глибина дискової черги представляє кількість незавершених дискових операцій рис. 2. Відстеження цього показника допомагає виявити вузькі місця дискового введення-виведення та оптимізувати

конфігурації сховища для EC2 екземплярів. На рис. 2 наведено приклад метрики DiskQueueDepth.

Відстеження обсягу вільної пам’яті допомагає забезпечити ефективне використання пам’яті в екземплярах EC2 рис. 3. Відстеження цього показника дозволяє оптимізувати розподіл пам’яті та запобігти проблемам з продуктивністю, викликаним обмеженнями пам’яті. На рис. 3 наведено приклад метрики FreeableMemory.

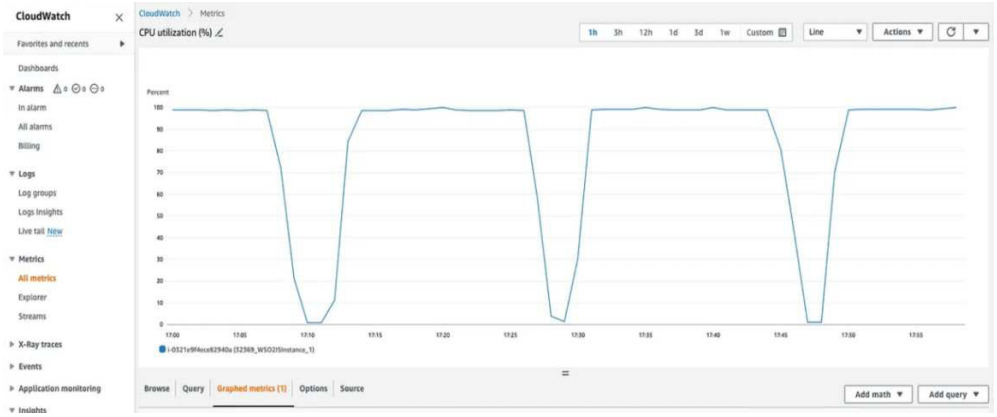


Рис. 1. Метрика CPU utilization (%)

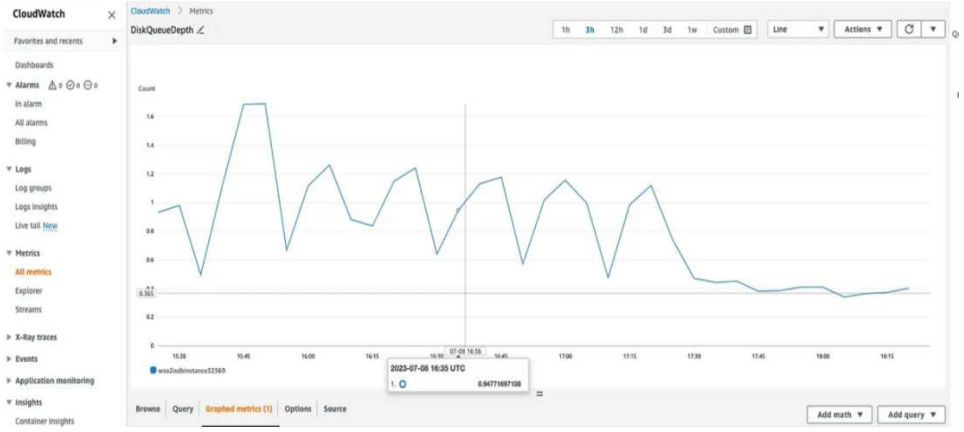


Рис. 2. Метрика DiskQueueDepth

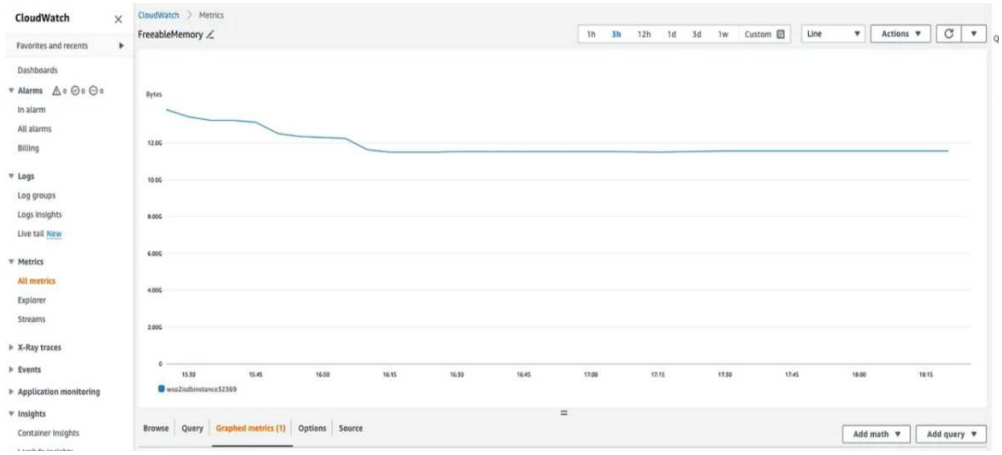


Рис. 3. Метрика FreeableMemory

Відстеження обсягу вільного місця для зберігання в екземплярах EC2 допомагає запобігти проблемам продуктивності, пов’язаним зі сховищем. Забезпечивши достатній вільний простір для зберігання даних, можливо уникнути потенційної втрати даних і оптимізувати використання сховища. На рис. 4 наведено приклад метрики FreeStorageSpace.

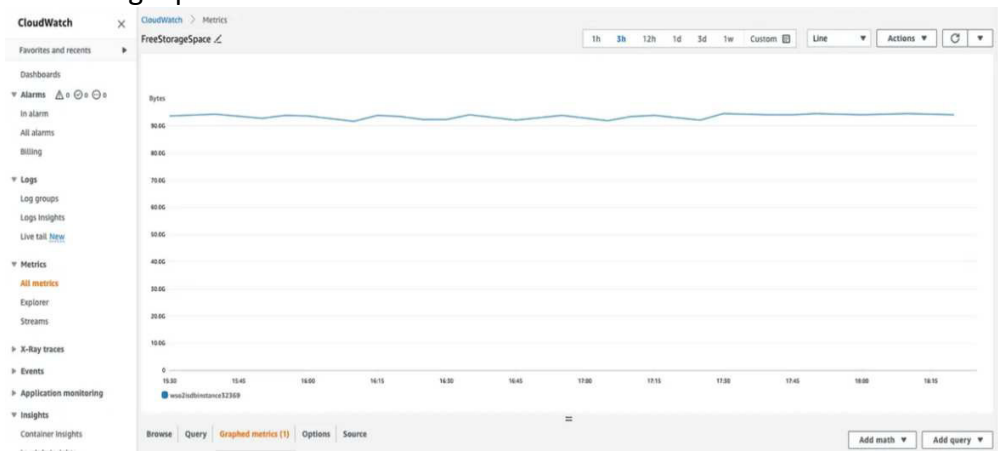


Рис. 4. Метрика FreeStorageSpace

Відстеження пропускної здатності мережі прийому та передачі допомагає виявити потенційні вузькі місця мережі. Аналізуючи ці показники, можливо оптимізувати мережеві конфігурації для забезпечення ефективної передачі даних для екземплярів EC2. На рис. 5 наведено приклад метрики NetworkReceiveThoughput.

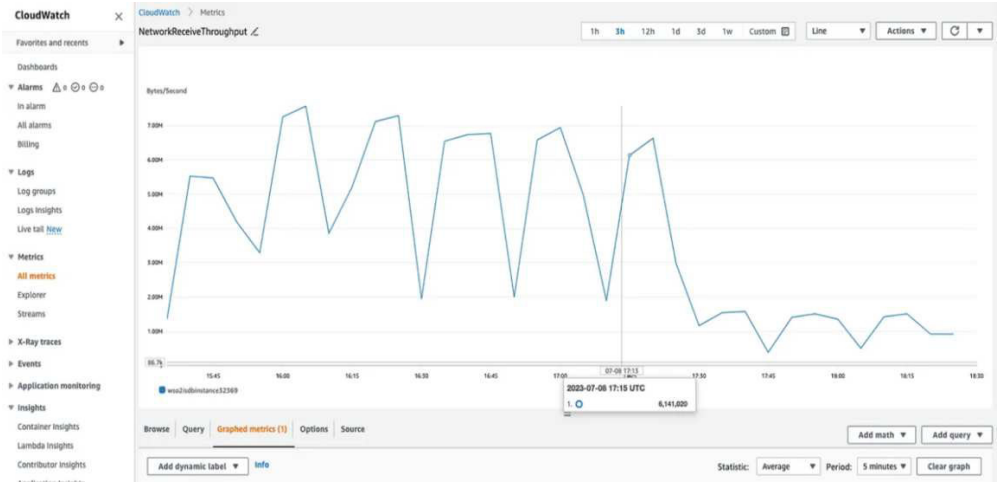


Рис. 5. Метрика NetworkReceiveThoughput

Моніторинг IOPS і затримки читання дає змогу зрозуміти продуктивність систем зберігання, підключених до екземлярів EC2. Відстежуючи ці показники, можливо виявити будь-які вузькі місця в отриманні даних і оптимізувати конфігурації зберігання або шаблони доступу до даних для підвищення продуктивності. На рис. 6 наведений приклад метрики ReadLatency.

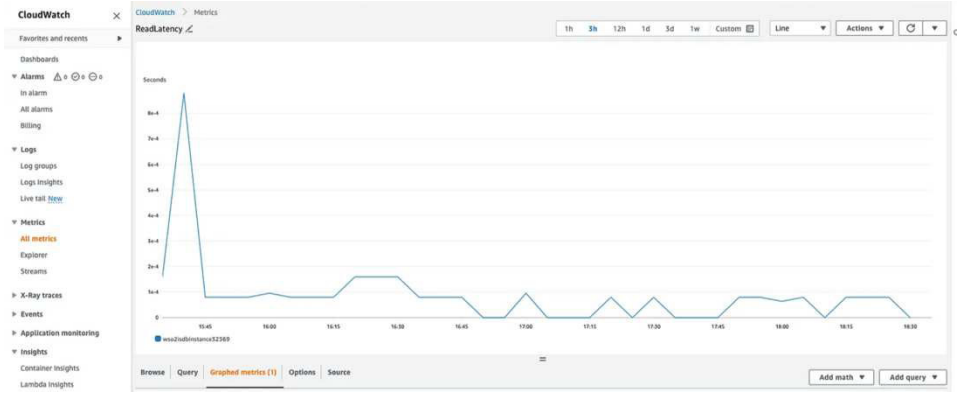


Рис. 6. Метрика ReadLatency

Загальний результат впровадження системи моніторингу та управління ресурсами в Amazon EC2 полягає в тому, що отримуються інструменти та знання для ефективного контролю, оптимізації та автоматизації хмарного середовища.

**Висновки**

Оптимізація продуктивності EC2 життєво важлива для компаній, які

використовують сервіси AWS. Ефективно відстежуючи та аналізуючи ключові показники продуктивності за допомогою CloudWatch, включаючи CPU utilization (%), DiskQueueDepth, FreeableMemory, FreeStorageSpace, NetworkReceiveThoughput та ReadLatency – можливо виявити вузькі місця, оптимізувати розподіл ресурсів і підвищити продуктивність додатків. Використовуючи потужність системи моніторингу AWS і CloudWatch, можливо завчасно вирішувати проблеми з продуктивністю, оптимізувати використання ресурсів і забезпечувати безперебійну роботу сервісу. Завдяки добре оптимізованим екземплярам EC2 можливо максимально використати переваги хмарних обчислень, зменшити витрати та забезпечити високу продуктивність сервісу. Постійний моніторинг, аналіз і оптимізація є важливими для підтримки оптимальної продуктивності в динамічному середовищі AWS.

**1.** Скляр В. О. Дослідження методів і технологій автоматизації моніторингу інфраструктур для хмарних інформаційних систем : пояснювальна записка до кваліфікаційної роботи здобувача вищої освіти на другому (магістерському) рівні, спеціальність 122 Комп’ютерні науки / М-во освіти і науки України. Харків. нац. ун-т радіоелектроніки. Харків, 2022. 96 с. **2.** Богатиренко А. М. Моделі архітектури високонавантажених web-застосунків на платформі хмарних сервісів Amazon : пояснювальна записка до атестаційної роботи здобувача вищої освіти на другому (магістерському) рівні, спеціальність 123 – Комп’ютерна інженерія / М-во освіти і науки України. Харків. нац. ун-т радіоелектроніки. Харків, 2019. 71 с. **3.** Azeez M. A. Autoscaling webservices on Amazon EC2 [Master's theses, University of Moratuwa]. Institutional Repository University of Moratuwa. 2010. URL: http://dl.lib.mrt.ac.lk/handle/123/2018 (дата звернення: 12.06.2023). **4.** Arvindhan, M and Anand, Abhineet, Scheming an Proficient Auto Scaling Technique for Minimizing Response Time in Load Balancing on Amazon AWS Cloud (March 15, 2019). *International Conference on Advances in Engineering Science Management & Technology (ICAESMT) – 2019.* Uttaranchal University, Dehradun, India, Available at SSRN. URL: https://ssrn.com/abstract=3390801 or http://dx.doi.org/10.2139/ssrn.3390801 (дата звернення: 10.09.2023). **5.** Yi Li, Fangming Liu, Qiong Chen, Yibing Sheng, Miao Zhao, Jianping Wang. MarVeLScaler: A Multi-View Learning-Based Auto-Scaling System for MapReduce. *IEEE Transactions on Cloud Computing*. 2022. Vol. 10, no. 1. Pp. 506–520. **6.** Hanieh Alipour, Yan Liu. Online machine learning for cloud resource provisioning of microservice backend systems. *IEEE International Conference on Big Data (Big Data)*. 2017. Pp. 2433–2441. **7.** Vahid Mirzaebrahim Mostofi, Evan Krul, Diwakar Krishnamurthy, Martin Arlitt. Trace-Driven Scaling of Microservice Applications. *IEEE Access*. 2023. Vol. 11. Pp. 29360–29379.

**REFERENCES:**

**1.** Skliar V. O. Doslidzhennia metodiv i tekhnolohii avtomatyzatsii monitorynhu infrastruktur dlia khmarnykh informatsiinykh system : poiasniuvalna zapyska do kvalifikatsiinoi roboty zdobuvacha vyshchoi osvity na druhomu (mahisterskomu) rivni, spetsialnist 122 Kompiuterni nauky / M-vo osvity i nauky Ukrainy. Kharkiv. nats. un-t radioelektroniky. Kharkiv, 2022. 96 s. **2.** Bohatyrenko A. M. Modeli arkhitektury vysokonavantazhenykh web-zastosunkiv na platformi khmarnykh servisiv Amazon : poiasniuvalna zapyska do atestatsiinoi roboty zdobuvacha vyshchoi osvity na druhomu (mahisterskomu) rivni, spetsialnist 123 – Kompiuterna inzheneriia / M-vo osvity i nauky Ukrainy. Kharkiv. nats. un-t radioelektroniky. Kharkiv, 2019. 71 s**. 3.** Azeez M. A. Autoscaling webservices on Amazon EC2 [Master's theses, University of Moratuwa]. Institutional Repository University of Moratuwa. 2010. URL: http://dl.lib.mrt.ac.lk/handle/123/2018 (data zvernennia: 12.06.2023). **4.** Arvindhan, M and Anand, Abhineet, Scheming an Proficient Auto Scaling Technique for Minimizing Response Time in Load Balancing on Amazon AWS Cloud (March 15, 2019). *International Conference on Advances in Engineering Science Management & Technology (ICAESMT) – 2019.* Uttaranchal University, Dehradun, India, Available at SSRN. URL: https://ssrn.com/abstract=3390801 or http://dx.doi.org/10.2139/ssrn.3390801 (data zvernennia: 12.06.2023). **5.** Yi Li, Fangming Liu, Qiong Chen, Yibing Sheng, Miao Zhao, Jianping Wang. MarVeLScaler: A Multi-View Learning-Based Auto-Scaling System for MapReduce. *IEEE Transactions on Cloud Computing*. 2022. Vol. 10, no. 1. Pp. 506–520. **6.** Hanieh Alipour, Yan Liu. Online machine learning for cloud resource provisioning of microservice backend systems. *IEEE International Conference on Big Data (Big Data)*. 2017. Pp. 2433–2441. **7.** Vahid Mirzaebrahim Mostofi, Evan Krul, Diwakar Krishnamurthy, Martin Arlitt. Trace-Driven Scaling of Microservice Applications. *IEEE Access*. 2023. Vol. 11. Pp. 29360–29379.

---

**Krasko B. V., Post-graduate Student, Hrytsiuk P. M., Doctor of Economics, Professor** (National University of Water and Environmental Engineering, Rivne, b.v.krasko@nuwm.edu.ua, p.m.hrytsiuk@nuwm.edu.ua)

## EFFECTIVE MONITORING AND MANAGEMENT OF RESOURCES IN AMAZON EC2 WHEN SCALING

**Amazon Elastic Compute Cloud (Amazon EC2) is a service that provides flexible and scalable computing power in the Amazon Web Services (AWS) cloud. With Amazon EC2, users can run virtual servers (instances) with different configurations of CPU, memory, storage, and network capacity, as well as configure security and network connectivity. Amazon EC2 also allows users to monitor and manage the resources of their instances using various AWS tools and services.**

**Monitoring and managing resources in Amazon EC2 is an important part of maintaining the reliability, availability, and performance of AWS instances and solutions. Monitoring allows users to identify and resolve performance, security, availability, or error issues. Management enables users to optimize resource utilization, reduce costs, improve performance, and adapt to changing workloads.**

**Resource management in Amazon EC2 is about choosing the optimal instance type for a given task. Instance types consist of different combinations of CPU, memory, storage, and network bandwidth. You can choose the instance type that best suits your application and change it as needed.**

**Scaling is the process of increasing or decreasing the number of Amazon EC2 instances depending on the needs of a particular application. Using auto-scaling to dynamically adjust infrastructure resources using rules and metrics. Auto-scaling helps improve application performance, availability, and efficiency.**

***Keywords*: Amazon EC2; Amazon CloudWatch cloud platforms; autoscaling; monitoring; management; load balancers.**

**_____**