DOI: https://doi.org/10.15276/ict.02.2025.36
УДК: 004.3+658.5

Математична модель для оптимізації хмарної IT інфраструктури

Лисевич Євген Ігорович[1)]
Аспірант каф./Механіки, автоматизації та інформаційних технологій
ORCID: https://orcid.org/0009-0004-6161-9453; lisevich.e.i@gmail.com
Волков Віктор Едуардович[1)]
Д-р техніч. наук, професор каф. Механіки, автоматизації та інформаційних технологій
ORCID: https://orcid.org/0000-0002-3990-8126; viktor@te.net.ua. Scopus Author ID: 57220703810
[1)] Одеський національний університет імені І. І.Мечникова, Всеволода Змієнка, 2. Одеса, 65082, Україна

АНОТАЦІЯ

Це дослідження присвячене формалізації математичної моделі для оптимізації хмарної IT-інфраструктури в умовах стрімкого зростання попиту на обчислювальні ресурси та підвищення вимог до продуктивності й надійності сервісів. Хмарні обчислення є ключовою парадигмою сучасної IT-інфраструктури організацій, адже вони забезпечують масштабованість, гнучкість, скорочення витрат та швидке розгортання сервісів. Разом із тим, зростання складності інфраструктури породжує нові виклики: необхідність балансування між вартістю ресурсів і якістю обслуговування кінцевих користувачів.

У роботі проаналізовано основні метрики, що визначають якість функціонування хмарних систем: затримку відповіді, частку помилок, пропускну здатність та доступність. Підкреслено важливість підходу «кількості дев'яток» при оцінці доступності сервісів. Важливу роль відіграють Service Level Objectives (SLO) та Service Level Agreements (SLA), які формалізують вимоги до системи та встановлюють наслідки їх невиконання. Розібрано, як саме SLO та SLA можна інтегрувати у математичні моделі оптимізації, щоб забезпечити контрольовану якість послуг при збереженні економічної ефективності.

Запропоновано використання класичних методів лінійного програмування для задач розподілу обчислювальних ресурсів. У моделі змінні відображають кількість та типи ресурсів (CPU, пам'ять, пропускна здатність), а обмеження відповідають вимогам SLO, SLA та фінансовим бюджетам. Такий підхід дозволяє знайти оптимальні конфігурації інфраструктури, що забезпечують максимальне співвідношення «продуктивність/вартість». Додатково розглядається задача масштабування, де оптимізація зводиться до підбору комбінацій віртуальних машин провайдерів хмарних сервісів різних типів у межах заданих обмежень.

Також увагу приділено сучасним дослідженням у сфері автоматичного масштабування та використання методів цілочислового лінійного програмування для задач консолідації віртуальних машин. Це дозволяє розширити класичну математичну модель і зробити його придатним для вирішення реальних проблем провайдерів хмарних послуг.

Результати дослідження демонструють, що формалізація задачі оптимізації хмарної інфраструктури через систему метрик, SLO та SLA створює основу для побудови ефективних алгоритмів автоматизованого управління ресурсами. Запропонований підхід може бути використаний як у наукових дослідженнях, так і у практичній діяльності компаній, які прагнуть досягти балансу між витратами та надійністю у високонавантажених хмарних середовищах.

Ключові слова: хмарна інфраструктура; оптимізація; математичний модель; лінійне програмування; SLO; SLA; масштабування; продуктивність; вартість

Вступ. Хмарні обчислення сьогодні є однією з провідних парадигм розвитку інформаційних технологій. Під хмарними обчисленнями слід розуміти набор сервісів, об'єднаних в одну структуру, яка дозволяє повсюдний, орієнтований на користувача допуск до спільного пулу настроюваних обчислювальних активів, які можуть видаватись по запиту через хмару (інтернет) без прямого активного управління користувачем. Основні переваги хмарних обчислень обмежуються не лише скороченням часу та витрат, але й гнучкістю та масштабованістю. Ідея хмарних обчислень спочатку була пов'язана з концепціями розподілених паралельних обчислень, обчислень з оплатою за безпосереднє користування та автономних обчислень. Хмарні обчислення мають різні моделі, засновані на розгортанні та наданні послуг. На основі доступу до хмари існує чотири моделі: публічна хмара, приватна хмара, гібридна хмара та хмара підприємства (організації), заснована на наданні послуг; моделі можна класифікувати як SaaS (програмне забезпечення як послуга), PaaS (платформа як послуга) та IaaS (інфраструктура як послуга) [1]. За оцінками провідних дослідницьких агентств, обсяг ринку хмарних послуг перевищує 600 млрд доларів США в 2023 році [2], і ця величина демонструє стійку тенденцію до зростання. З одного боку, хмарні рішення надають організаціям значні переваги: масштабованість, гнучкість, швидке розгортання сервісів та

економію витрат. З іншого боку, виникають нові виклики, зокрема забезпечення стабільної роботи високонавантажених систем, які мають відповідати вимогам користувачів та бізнес-цілям при обмежених ресурсах.

Метою цієї роботи є побудова концептуальних засад математичної моделі для оптимізації хмарної IT-інфраструктури шляхом формалізації ключових метрик та визначення цільових показників їх функціонування.

Мета. Формування концептуальних засад побудови математичної моделі для оптимізації хмарної IT-інфраструктури, яка дозволить знайти оптимальні конфігурації інфраструктури, що забезпечують максимальне співвідношення «продуктивність/вартість».

Теоретичні основи вибору метрик. Фундаментальним кроком у створенні математичної моделі є визначення системи метрик, які відображають як якість сервісу, так і витрати на його підтримку. Значний внесок у цю сферу зроблено в роботі [3], що стала орієнтиром для побудови сучасних моделей керування надійністю.

Основними метриками виступають:

- **Затримка відповіді (request latency):** час, необхідний серверу для обробки клієнтського запиту;
- **Частка помилок (error rate):** відношення кількості помилкових запитів до їх загального числа;
- **Пропускна здатність (throughput):** кількість успішно оброблених запитів за одиницю часу;
- **Доступність (availability):** частка часу, коли сервіс є працездатним.

У практиці застосовується так званий підхід «кількості дев'яток». Наприклад, доступність 99% трактується як «дві дев'ятки», а 99,999% – як «п'ять дев'яток». Для Google Compute Engine типовим цільовим показником є 99,95% («три з половиною дев'ятки»). Очевидним є той факт, що максимальна доступність прикладного сервісу не може перевищувати доступності базового сервісу, на якому він розгорнутий.

SLO та SLA як основа оптимізаційної моделі. Зібрані метрики самі по собі не надають вичерпної інформації щодо адекватності якості сервісу. Для оцінки цього використовуються такі поняття, як Service Level Objectives (SLO) – цільові показники для метрик сервісу, та Service Level Agreements (SLA) – угоди про рівень обслуговування.

Service Level Objectives (SLO)

SLO визначають цільові значення або допустимі діапазони для вибраних метрик. Вони слугують формальним інструментом оцінки відповідності роботи системи очікуванням користувачів. Прикладом може бути вимога забезпечити середню затримку відповіді на пошуковий запит не більше 100 мс.

Формально: **метрика ≤ ціль** або **нижня межа ≤ метрика ≤ верхня межа**

Service Level Agreements (SLA)

SLA є контрактами (явними або неявними) з кінцевими користувачами, що закріплюють наслідки недотримання SLO. У більшості випадків SLA передбачає фінансову відповідальність провайдера, проте можливі й нефінансові санкції (наприклад, репутаційні втрати чи зниження рівня довіри).

Відмінність між SLO та SLA можна визначити через запитання: *що відбувається у випадку недотримання мети?* Якщо наслідки відсутні – це лише SLO; якщо ж передбачені санкції – це SLA [3].

В умовах, де поширені завдання з інтенсивною роботою з даними, функції хмарних обчислень, такі як еластичність, розподіл ресурсів та їх отримання «на замовлення», відіграють важливу роль. Масштабованість, ключовий компонент хмарних обчислень, дає можливість організаціям змінювати використання своїх ресурсів відповідно до своїх потреб, пропонуючи не тільки економічні переваги, але й важливі покращення продуктивності хмарних додатків. Автомасштабування, яке передбачає динамічне збільшення або

зменшення ресурсів на основі конкретних потреб та стратегій, вирішує цю проблему, забезпечуючи гнучкість [4].

Варто зазначити, що тема використання тих чи інших видів лінійного програмування набуває все більшої популярності, наприклад, алгоритми цілочислових задач лінійного програмування використовується для розрахунків розподілення віртуальних машин на серверах у провайдерів хмарних сервісів [5].

Методологія і рішення

Враховуючи перелічені параметри, вирішення задачі оптимізації лягає в класичну задачу лінійного програмування[6] і може бути описане як

$$\sum_{i=0}^{n} p_i V_i \rightarrow max, \quad (1)$$

де p_i – вагові коефіцієнти (наприклад, ціна за одиницю тарифікації);

V_i – змінні, що описують ресурси (мінімальні гранулярні показники, які тарифікуються. Наприклад, 1vCPU за годину, 1GB трафіку чи місця на віртуального диска, тощо).

Обмеження на змінні

$$a_i \leq V_i \leq b_i, (i = 1, \ldots, n), \quad (2)$$

де a_i, b_i – мінімальні та максимальні допустимі значення (наприклад, SLA/SLO межі продуктивності або доступності).

Балансові обмеження

$$\underline{\boldsymbol{V_j}} \leq \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{V_i} \leq \overline{\boldsymbol{V_j}}, (j = 1, \ldots, m), \quad (3)$$

де $\underline{V_i}, \overline{V_j}$ – нижні та верхні межі агрегованих ресурсів або метрик (наприклад, загальна кількість запитів, бюджет витрат, пропускна здатність системи).

Відповідно, зараз ми маємо все для того, щоб вирішити класичну задачу лінійного програмування, яку можна використати для моделювання оптимального розподілу хмарних ресурсів.

1. Оптимізація вартості і продуктивності.
   - p_i може відображати економічну ефективність (наприклад, співвідношення «продуктивність/вартість» для конкретного типу інстансу AWS або GCP).
   - Оптимізація $\sum p_i V_i$ відповідає знаходженню конфігурації ресурсів, яка максимізує віддачу від інфраструктури.
2. Врахування SLO.
   - Обмеження $a_i \leq V_i \leq b_i$ можна пов’язати з гарантіями SLO: наприклад, мінімальна кількість CPU для обробки 10k RPS, або верхня межа затримки відповіді <100 мс.
   - Це дозволяє формально закласти в модель вимоги, описані вище.
3. Врахування SLA.
   - Балансові обмеження $\underline{\boldsymbol{V_j}} \leq \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{V_i} \leq \overline{\boldsymbol{V_j}}$ можуть виражати:
     - загальний бюджет на інфраструктуру;
     - загальний рівень доступності («кількість дев’яток»);
     - сумарну пропускну здатність, яку потрібно забезпечити.
4. Задача масштабування.
   - Якщо розглядати V_i як кількість інстансів різних типів (наприклад, t3.micro, m5.large тощо), тоді ця модель зводиться до вибору оптимальної комбінації машин під обмеження вартості й вимог до SLO/SLA.

Практичний приклад застосування моделі

Для ілюстрації можливостей розробленої математичної моделі розглянемо спрощений сценарій оптимізації хмарних ресурсів для веб-додатку, розгорнутого в середовищі AWS. Припустимо, що компанія має забезпечити обробку не менше 10 000 запитів на секунду при середньому часі відгуку не більше ніж 100 мс та гарантованій доступності на рівні 99,9 %. Додатковим обмеженням є бюджет у розмірі 500 доларів США на місяць.

У рамках задачі лінійного програмування змінними виступають кількості інстансів різних типів (наприклад, t3.micro, m5.large, c5.xlarge). Вартість оренди кожного типу інстансу враховується як ваговий коефіцієнт, а продуктивність визначається через пропускну здатність та латентність. Обмеження задаються через цільові SLO та SLA: мінімальний рівень продуктивності, максимальна затримка відповіді, граничний бюджет.

Розв'язавши таку задачу, можна отримати оптимальну комбінацію інстансів, яка забезпечить необхідний рівень продуктивності при дотриманні фінансових обмежень. При цьому модель дозволяє порівняти декілька альтернативних рішень і вибрати найбільш економічно вигідне. Такий підхід є універсальним і може застосовуватися не лише до EC2, а й до інших сервісів, наприклад, до вибору оптимального обсягу сховищ S3 чи параметрів бази даних RDS.

Переваги і обмеження моделі

Запропонована модель оптимізації має низку переваг. По-перше, вона забезпечує формальну постановку задачі, що дозволяє застосовувати відомі методи лінійного та цілочислового програмування. По-друге, такий підхід є прозорим та інтерпретованим: кожна змінна та обмеження мають чітке практичне значення. По-третє, використання моделей цього класу добре інтегрується з існуючими інструментами автоматизації та може бути реалізоване за допомогою відкритих бібліотек оптимізації.

Разом із тим, існують певні обмеження. Лінійні моделі погано враховують динамічний характер навантажень, які можуть змінюватися протягом доби або тижня. Для реальних масштабних систем складність задачі зростає експоненційно, що може ускладнити знаходження оптимального розв'язку. Крім того, для якісного застосування потрібні точні дані про вартість і продуктивність ресурсів, які не завжди доступні або можуть варіюватися залежно від регіону чи політики постачальника.

Таким чином, розроблена модель є потужним інструментом для аналізу й початкової оптимізації хмарної інфраструктури, проте для комплексного застосування в промислових середовищах вона потребує доповнення евристичними та машинно-навчальними підходами.

Висновки. У ході дослідження було сформовано концептуальні засади побудови математичної моделі для оптимізації хмарної IT-інфраструктури. Показано, що ключовим етапом у цьому процесі є визначення релевантних метрик – затримки відповіді, пропускної здатності, доступності та частки помилок – які в сукупності характеризують якість функціонування системи. На основі цих метрик формалізуються Service Level Objectives (SLO) та Service Level Agreements (SLA), що забезпечують узгодженість між технічними показниками інфраструктури та бізнес-вимогами користувачів.

Запропонований підхід доводить, що задачі розподілу ресурсів та масштабування хмарних систем доцільно розглядати у термінах класичного лінійного програмування. Така постановка дозволяє одночасно враховувати обмеження за бюджетом, вимогами до продуктивності та гарантіями доступності. Розроблена модель створює підґрунтя для побудови автоматизованих рішень, що можуть забезпечити динамічне балансування між вартістю інфраструктури та якістю обслуговування.

Перспективним напрямом подальших досліджень є інтеграція методів цілочислового програмування, евристичних алгоритмів та підходів машинного навчання, зокрема reinforcement learning, для удосконалення процесів автоматичного масштабування та прогнозування навантажень. Це дозволить розширити математичну модель та адаптувати його до реальних сценаріїв експлуатації хмарних середовищ.

Отже, отримані результати мають як теоретичне значення – формалізація задач оптимізації у хмарній інфраструктурі, так і практичну цінність – можливість використання розроблених моделей у корпоративних середовищах для зниження витрат і підвищення надійності сервісів.

СПИСОК ЛІТЕРАТУРИ

1. Malik A. W., Bhatti, D. S., Park T.-J., Ishtiaq H. U., Ryou J.-C., Kim K.-I. “Cloud Digital Forensics: Beyond Tools, Techniques, and Challenges”. *Sensors*. 2024; 24: 433. DOI: https://doi.org/10.3390/s24020433.

2. Zheng X., Li L. “Trading Cloud Computing Stocks Using SMA”. *Information*. 2024; 15: 506. DOI: https://doi.org/10.3390/info15080506 .

3. Beyer B., Jones C., Petoff J., Murphy N. R. “Site Reliability Engineering: How Google Runs Production Systems”. DOI: https://dl.acm.org/doi/book/10.5555/3006357 .

4. Alharthi S., et al. “Auto-Scaling Techniques in Cloud Computing: Issues and Research Directions”. *Sensors* 2024; 24 (17): 5551. DOI: https://doi.org/10.3390/s24175551 .

5. Luo J.-Y., Chen L., Chen W.-K., Yuan J.-H., Dai Y.-H. “A cut-and-solve algorithm for virtual machine consolidation problem”. *Future Gener. Comput. Syst*. 2024; 154: 359–372. DOI: https://doi.org/10.1016/j.future.2024.01.010 .

6. Taha H. A. “Operations Research: An Introduction”. 2003. ISBN-10: 0130323748.

DOI: https://doi.org/10.15276/ict.02.2025.36
UDC: 004.3+658.5

Mathematical model for optimizing cloud IT infrastructure

Yevhen I. Lysevych[1)]
PhD Student of the Department of Mechanics, Automation and Information Technologies
ORCID: https://orcid.org/0009-0004-6161-9453; lisevich.e.i@gmail.com
Viktor E. Volkov[1)]
Doctor of Engineering Sciences, Professor of the Department of Mechanics, Automation and Information Technologies
ORCID: https://orcid.org/0000-0002-3990-8126; viktor@te.net.ua. Scopus Author ID: 57220703810
[1)] Odesa I. I. Mechnikov National University, 2, Vsevoloda Zmiienka Str. Odesa, 65082, Ukraine

ABSTRACT

This research is dedicated to the formalization of a mathematical model for optimizing cloud IT infrastructure under conditions of rapidly growing demand for computing resources and increasing requirements for service performance and reliability. Cloud computing is a key paradigm of modern organizational IT infrastructure, as it provides scalability, flexibility, cost reduction, and rapid service deployment. At the same time, the growing complexity of infrastructure generates new challenges: the need to balance between resource costs and the quality of end-user service.

The study analyzes the main metrics that determine the quality of cloud systems: response, latency, error rate, throughput, and availability. Particular attention is given to the “number of nines” approach in evaluating service availability. Service Level Objectives (SLOs) and Service Level Agreements (SLAs) play an important role, as they formalize system requirements and define the consequences of non-compliance. The paper explains how SLOs and SLAs can be integrated into mathematical optimization models to ensure controlled service quality while maintaining economic efficiency.

The use of classical linear programming methods is proposed for solving resource allocation tasks. In the model, variables represent the quantity and types of resources (CPU, memory, bandwidth), while constraints correspond to SLO, SLA, and budget requirements. This approach makes it possible to find optimal infrastructure configurations that maximize the “performance-to-cost” ratio. Additionally, the scaling problem is considered, where optimization reduces to selecting combinations of cloud service providers’ virtual machines of different types within defined constraints.

Attention is also given to modern research in the field of automatic scaling and the use of integer linear programming methods for virtual machine consolidation tasks. This extends the classical mathematical model and makes it applicable to solving real-world problems faced by cloud service providers.

The results of the study demonstrate that formalizing the cloud infrastructure optimization problem through a system of metrics, SLOs, and SLAs creates a foundation for building efficient algorithms for automated resource management. The proposed approach can be applied both in scientific research and in the practical activities of companies striving to achieve a balance between cost and reliability in high-load cloud environments.

Keywords: Cloud infrastructure; optimization; mathematical model; linear programming; SLO; SLA; scaling; performance; cost