



2026, № 1

Журнал обчислювальної та
прикладної математики
Заснований в 1965 році



Journal of Numerical
& Applied Mathematics

ISSN (Print): 2706-9680, ISSN (Online): 2706-9699

ЄДРПОУ видавця: 02070944

ROR: 02aaqv166

E-mail:
jnameditorialteam@gmail.com

Головний редактор:
Ляшко Сергій Іванович

Сіренко А., Вінце К.

ОПТИМІЗАЦІЯ ОБЧИСЛЮВАЛЬНИХ РЕСУРСІВ У БАГАТОХМАРНИХ АРХІТЕКТУРАХ З ВИКОРИСТАННЯМ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ ОРКЕСТРАЦІЇ

Адреса редакції:
Кафедра обчислювальної математики,
факультет комп'ютерних наук та кібернетики,
Київський національний університет імені Тараса Шевченка,
Україна, 01601, Київ, вул. Володимирська, 64/13.

©Київський національний університет імені Тараса Шевченка, 2026

Свідцтво про державну реєстрацію КВ 4246 від 26.05.2000
Включено в Перелік наукових фахових видань у відповідності до Наказу МОН
України № 409 від 17.03.2020

DOI: 10.17721/2706-9699.2026.1.05

УДК 004.7:004.8:519.85

MSC 68M14, 68T05, 90B50

OPTIMIZATION OF COMPUTATIONAL RESOURCES IN MULTI-CLOUD ARCHITECTURES USING INTELLIGENT ORCHESTRATION SYSTEMS

A. SIRENKO¹, K. VINTSE²

¹Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, E-mail: sirenko_andrii@knu.ua, ORCID: 0000-0003-1171-7894

²Faculty of Computer Science and Cybernetics, Taras Shevchenko Kiev National University of Kyiv, Kyiv, Ukraine, E-mail: krisvintse6@gmail.com

ОПТИМІЗАЦІЯ ОБЧИСЛЮВАЛЬНИХ РЕСУРСІВ У БАГАТОХМАРНИХ АРХІТЕКТУРАХ З ВИКОРИСТАННЯМ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ ОРКЕСТРАЦІЇ

А. СІРЕНКО¹, К. ВІНЦЕ²

¹Факультет комп'ютерних наук та кібернетики, Київський національний університет імені Тараса Шевченка, Київ, Україна, E-mail: sirenko_andrii@knu.ua, ORCID:0000-0003-1171-7894

²Факультет комп'ютерних наук та кібернетики, Київський національний університет імені Тараса Шевченка, Київ, Україна, E-mail: krisvintse6@gmail.com

ABSTRACT. At this work modern approaches are being explored to optimization of computational resources in multi-cloud architectures using intelligent orchestration systems. The main challenges which associated with managing heterogeneous cloud environments are considered and also the benefits of applying AI/ML-methods to automate workload placement and scaling. Was suggested a formal mathematical model of the optimization problem taking into account resource costs, network delays, SLA requirements and migration costs. It was shown that using of predictive models and RL-agents allows to increase the efficiency of resource allocation, reduce financial costs and ensure resilience to load changes. This work contains of analysis modern research, practical recommendations, and outlines directions for further scientific developments in the field of multi-cloud infrastructures.

KEYWORDS: multi-cloud architecture, resource optimization, intelligent orchestration systems, cloud computing.

Corresponding author: A. Sirenko (sirenko_andrii@knu.ua).

© A. Sirenko, K. Vintse, 2026. This is an open-access article distributed under the terms of **Creative Commons Attribution Licence (CC BY)**.

АНОТАЦІЯ. У роботі досліджується сучасні підходи до оптимізації обчислювальних ресурсів у багатохмарних (multi-cloud) архітектурах із використанням інтелектуальних систем оркестрації. Розглянуто основні виклики, пов'язані з управлінням різномірними хмарними середовищами, а також переваги застосування AI/ML-методів для автоматизації розміщення та масштабування робочих навантажень. Запропоновано формальну математичну модель задачі оптимізації з урахуванням вартості ресурсів, мережних затримок, вимог SLA та витрат на міграцію. Показано, що використання прогнозних моделей та RL-агентів дає змогу підвищити ефективність розподілу ресурсів, зменшити фінансові витрати та забезпечити стійкість до змін навантаження. Робота містить аналіз сучасних досліджень, практичні рекомендації та окреслює напрями подальших наукових розробок у сфері багатохмарних інфраструктур.

КЛЮЧОВІ СЛОВА: багатохмарна архітектура, оптимізація ресурсів, інтелектуальні системи оркестрації, хмарні обчислення.

1. INTRODUCTION

The modern digital space requires from organizations unprecedented flexibility, scalability, and financial gain. These challenges have led to widespread use of multi-cloud structures, where workloads are distributed across multiple public cloud providers (AWS, Microsoft Azure, Google Cloud Platform, and others). Strategic multicloudiness allows to reduce the risk of being tied to a single manufacturer, using the best services from each provider, and, most importantly, optimize costs through constant cost monitoring. However, managing a heterogeneous environment that contains of different application programming interfaces (API), different tariffs and controls, is extremely difficult. Traditional management systems (such as standard Kubernetes or simple scripts) where apply fixed guidelines which are unable to respond flexibly to market dynamics (changing Spot Instance prices) and load changes (unpredictable traffic spikes). Inefficient allocation of resources and suboptimal capacity increases lead to significant financial losses, which, according to some researchers, can reach 30% of the cloud budget. To overcome these difficulties, it is necessary switchover to methods of Intelligent Control Systems (ICS), which use artificial intelligence (AI) and machine learning (ML). These systems are able to predict, learn, and automatically make multi-objective decisions about workload placement and migration, balancing cost, performance, and regulatory compliance.

2. LITERATURE REVIEW

The issue of multi-cloud environments in modern research occupies one of the leading positions in the field of cloud computing, which is due to the growing complexity of digital infrastructures, increasing requirements for reliability, performance, and economic feasibility of using computing resources.

The fundamental definitions and conceptual principles of cloud computing were formulated in the works of the US National Institute of Standards and

Technology (NIST) where proposed a unified classification of service models and methods of deploying cloud infrastructure. These provisions have become the theoretical basis for the further development of multi-cloud computing approaches [1].

In Gholami's research, multi-cloud architecture is viewed as an evolutionary stage in the development of cloud technologies, aimed at overcoming the limitations inherent in using a single service provider. The authors identify the key motivations for implementing Multi-Cloud with a Best-of-Breed approach is need to comply with regional and regulatory requirements for data placement and the reduction of risks associated with dependence on a single provider [2]. Similar conclusions are confirmed in works devoted to the analysis of network challenges and prospects for the development of multi-cloud environments, where attention is focused on the growth of inter-cloud delays and the complexity of distributed resource management processes [3]. An important place in scientific literature is occupied by the problem of multi-criteria resource normalization in Multi-Cloud environments. In particular, the work of Wang, Li shows that simultaneous optimization of cost, performance and reliability indicators has a conflicting nature and requires used of formalized models of support making decision [4]. Further development concept of the AIOps has led to increased interest in intelligent cloud infrastructure management systems capable of automating the processes of monitoring, analyzing, and orchestrating resources. In the work of Shukla and Purbey summarizes modern approaches to the application of artificial intelligence methods in the management of cloud environments and substantiates the role of AIOps as a key technology for the transition from reactive to proactive management models [6]. In this context, particularly relevant become reinforcement learning methods. Research results of Alipour and Mohsenzadeh, also Lu, demonstrate the effectiveness of using reinforcement learning for dynamic resource allocation and scaling tasks in cloud and multi-cloud environments. The application of RL-approaches allows agents to adapt to changes in resource costs, load levels and network characteristics without need for pre-labeled training samples, which is especially important in resource utilization scenarios [7, 8]. A separate direction of scientific research is related to predictive scaling and deployment of services in multi-cloud environments. In particular, Zhang proposes a cost-aware approach to service placement with guaranteed SLA realization, which combines load forecasting methods with cost optimization in a multi-cloud infrastructure [9]. The obtained results confirm the feasibility of integrating demand and cost forecasting models into intelligent resource orchestration systems. Summarizing the results of the analysis of modern scientific sources, we can make conclusion that the effective functioning of multi-cloud environments requires using of complex intellectual approaches to resource normalization.

The combination methods of multi-criteria optimization, reinforcement learning and predictive analytics are considered by the scientific community as a promising direction for increasing the performance, reliability and cost-effectiveness of cloud infrastructures.

3. PURPOSE

The purpose of the research is to develop and substantiate approaches that will allow for more efficient use of computing resources in multi-cloud architectures. For this we proposed using intelligent orchestration systems based on methods of artificial intelligence and machine learning.

The main focus is on creating a mathematically formalized and practically applicable model of multi-criteria management. Such model should help to optimize space and scale workloads, balancing cost, performance, reliability and compliance requirements.

Ultimately, this will allow to minimize operating costs and increase the efficiency of multi-cloud infrastructure in conditions of constantly changing load and pricing models of different cloud providers.

4. METHODOLOGY

The formation of a multi-cloud infrastructure is caused by several key factors: using best-of-breed approach for using specialized services from different providers; necessity to comply with regional and regulatory requirements for data placement; and also the desire to minimize the risks of vendor dependence and ensure the competitiveness of pricing models. But optimization in multi-cloud environments is a multi-criteria task that includes economic, performance, reliability, and regulatory aspects.

Reducing the cost of resources often leads to worsening of latency indicators or stability, while productivity increases are accompanied by rising costs.

This necessitates using intellectual methods for searching compromise solutions. And here comes on help intelligent orchestration system, which functions as a superstructure over heterogeneous cloud resources and consists from data collection levels, resource abstraction, an analytical and decision-making core based on ML/AI and an execution level.

Such architecture provides an unified representation of resources and centralized decision-making.

In such systems, reinforcement learning is considered as a key method of dynamic resource management.

The agent interacts with the multi-cloud environment, receiving rewards or penalties which depend on the effectiveness of the decisions made.

This allows to adapt orchestration strategy to changes in load and pricing patterns in real time.

Unlike reactive approaches, predictive scaling is based on analysis of load time series and resource costs.

Using LSTM or Prophet models allows for early allocation of resources in the most economically viable clouds, reducing delays and financial losses.

But the problem of optimal workload placement in a multi-cloud environment is formalized as a multi-objective discrete optimization problem. Let given the set of workloads :

$$W = \{W_1, W_2, \dots, W_N\}$$

and a multitude of cloud providers:

$$C = \{C_1, C_2, \dots, C_M\}.$$

Each load W_i is described by the parameters

$$(cpu_i, mem_i, io_i, sla_i, reg_i),$$

where $reg_i \subseteq R$ are the permissible regions of placement.

Each provider C_j is characterized by

$$(p_j(t), cap_j^{cpu}, cap_j^{mem}, region_j).$$

Let us introduce a binary assignment variable:

$$x_{ij}(t) = \begin{cases} 1, & \text{if } W_i \text{ is placed in } C_j \text{ at time } t, \\ 0, & \text{otherwise.} \end{cases}$$

Single placement of each load:

$$\sum_{j=1}^M x_{ij} = 1, \quad \forall i \in W.$$

Computing resource limitations:

$$\begin{aligned} \sum_{i=1}^N cpu_i x_{ij}(t) &\leq cap_j^{cpu}, \\ \sum_{i=1}^N mem_i x_{ij}(t) &\leq cap_j^{mem}. \end{aligned}$$

Regional restrictions:

$$x_{ij}(t) = 0, \quad \text{if } region(W_i) \notin region(C_j).$$

The overall objective function at time t has the form

$$\min Z(t) = \sum_{i=1}^N \sum_{j=1}^M x_{ij}(t) (\alpha cost_{ij}(t) + \beta latency_{ij}(t) + \gamma mig_{ij}(t)),$$

where

$$\alpha, \beta, \gamma \in [0, 1], \quad \alpha + \beta + \gamma = 1.$$

Here $cost_{ij}(t)$ is the maintenance cost of W_i in C_j , $latency_{ij}(t)$ is the expected delay, $mig_{ij}(t)$ denotes migration costs (egress, downtime).

Delay for load:

$$latency_i(t) = \sum_{j=1}^M latency_{ij}(t) x_{ij}(t).$$

To guarantee the SLA, a penalty function is introduced:

$$penalty_{ij}(t) = \max(0, latency_{ij}(t) - SLA_i).$$

Migration cost:

$$Mig_i(t) = \begin{cases} cost_{ab}^{mig}(W_i), & \text{if } W_i : C_a \rightarrow C_b, \\ 0, & \text{otherwise.} \end{cases}$$

System state:

$$S_t = \{load(t), prices(t), latency(t)\}.$$

Agent action:

$$A_t \in \{migrate(W_i, C_j), scale(W_i), \emptyset\}.$$

Reward function:

$$R_t = -Z(t) - \lambda \sum penalty_{ij}(t),$$

where λ is the penalty coefficient.

Therefore, the task is reduced to a Markov decision process.

The optimal strategy is defined as

$$\pi^* = \arg \max_{\pi} \mathbb{E} \left[\sum_{t=0}^{\infty} \delta^t R_t \right], \quad \delta \in (0, 1).$$

5. DISCUSSION

Approach which proposed facilitates the transition from manual financial and operational management to an autonomous AIOps model.

The expected effect is to significantly reduce the cost of using cloud resources and increase the stability of services.

Promising areas of further research is an integration of edge and fog-computing, development of explanatory decision-making models and application of quantum optimization methods.

6. CONCLUSIONS

Intelligent orchestration systems play a key role in the development of multi-cloud infrastructures.

The combination of AIOps, predictive analytics and multi-criteria optimization creates the prerequisites for proactive, efficient and cost-effective resource management, which is critical for today's digital ecosystems.

The authors declare that there is no conflict of interest concerning the publication of this article.

REFERENCES

1. Mell P., Grance T. The NIST Definition of Cloud Computing. National Institute of Standards and Technology. Special Publication. 2011. 800–145.
2. Gholami R., Nazari M., Khani F. Multi-cloud resource management and orchestration: A comprehensive review. *Journal of Network and Computer Applications*. 2020. 151. 102506.
3. Bresnahan J. et al. Multi-cloud networking: Challenges and future directions. *IEEE Communications Magazine*. 2022, 60(2). P. 126–132.

4. Wang W., Li S., Li R. Multi-criteria optimization for cloud service selection and deployment. *Future Internet*. 2020. 12(11). 185.
5. Buyya R., Srirama S. N., Manvi S. S. Cloud resource provisioning and allocation: An economic approach. In: *Principles of Cloud Computing*. Springer, Cham, 2019.
6. Shukla S. P., Purbey S. A survey on AIOps and its application in cloud computing resource management. *Archives of Computational Methods in Engineering*. 2021. 28. P. 301–320.
7. Alipour Y., Mohsenzadeh Y. Resource allocation in multi-cloud environments using deep reinforcement learning. *Future Generation Computer Systems*. 2021. 124. P. 218–232.
8. Lu T. et al. Reinforcement learning-based dynamic resource provisioning in cloud data centers. *IEEE Transactions on Cloud Computing*. 2018. 6(3). P. 856–869.
9. Zhang C., et al. Cost-aware and SLA-guaranteed multi-cloud service placement using deep learning. *Journal of Parallel and Distributed Computing*. 2023. 172. 104712.

Received: 23.02.2026 / Revised: 20.03.2026 /
Accepted: 10.04.2026 / Published: 24.04.2026