АНОТАЦІЯ

Кириченко О.О. **Оптимізація безсерверних обчислень у хмарних середовищах.** – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 121 – «Інженерія програмного забезпечення» – Чернівецький національний університет імені Юрія Федьковича, Чернівці, 2026.

Дисертаційна робота присвячена оптимізації безсерверних обчислень у хмарних середовищах в умовах нерівномірного навантаження та неоднорідності вхідних потоків завдань. Сучасні підходи до автоматичного масштабування обчислювальних ресурсів мають переважно реактивний характер, що призводить до проблем холодного старту, неефективного використання ресурсів та збільшення затримок у системах черг. Застосування методів машинного навчання та нейронних мереж для прогнозування навантаження потребує значних обсягів історичних наборів даних і не забезпечує аналітичних гарантій якості обслуговування. У дисертаційній роботі розв'язується актуальна науково-прикладна задача розробки інформаційної технології оптимізації безсерверних обчислень у хмарних середовищах шляхом побудови математичних моделей на основі теорії систем масового обслуговування неоднорідної структури та створення відповідного програмного забезпечення.

Об’єктом дослідження є процеси розподіленої обробки даних у хмарних середовищах з використанням безсерверних технологій та систем керування чергами повідомлень.

Предметом дослідження є методи та моделі оптимізації безсерверних обчислень у хмарних середовищах на основі теорії систем масового обслуговування неоднорідної структури.

Метою дослідження є розробка підходів оптимізації безсерверних обчислень у хмарних середовищах, що забезпечують підвищення продуктивності, зниження витрат і мінімізацію холодних стартів шляхом побудови інформаційної технології на основі теорії масового обслуговування

для прогнозування навантаження та проактивного масштабування обчислювальних ресурсів.

У дисертаційній роботі **значно удосконалено** математичну модель безсерверної обчислювальної системи на прикладі хмарної платформи AWS Lambda як неоднорідної системи масового обслуговування зі змішаними режимами роботи, у якій вхідний процес визначається сумішшю незалежних неоднорідних пуассонівських процесів з різних джерел подій. **Встановлено необхідну та достатню умову** обмеженості черги для неоднорідної системи масового обслуговування. **Вперше доведено** граничні еволюції для процесу довжини черги у схемі усереднення та схемі дифузійної апроксимації з використанням апарату напівмарковських випадкових еволюцій, що базується на теорії Марковських ланцюгів та напівмарковських процесів. Схема усереднення визначає детерміновану траєкторію середнього навантаження, а схема дифузійної апроксимації описує випадкові відхилення від цієї траєкторії через стохастичне диференціальне рівняння Ito, що є основою для прогнозування поведінки інтелектуальної системи автоматичного масштабування.

У дисертаційній роботі також **удосконалено алгоритм** параметричної оцінки та оптимізації конфігурації безсерверної системи, який, на відміну від існуючих підходів на основі машинного навчання та нейронних мереж, використовує аналітичні оцінки теорії масового обслуговування, дозволяючи отримати явні формули для оптимальних параметрів та уникнути необхідності навчання моделей штучного інтелекту на великих наборах даних.

Розроблено архітектуру фреймворку для розподіленої обробки даних з використанням безсерверних технологій на хмарній платформі AWS. Особливістю даної архітектури є використання системи черг повідомлень, реалізація слабкої зв'язності між окремими елементами інтелектуальної системи, незалежне їх масштабування в разі потреби, швидке відновлення після збоїв та оптимізація використання обчислювальних ресурсів. **Розроблено та реалізовано інформаційну технологію** для розподіленої

обробки даних з використанням безсерверних обчислень та проактивним автоматичним масштабуванням обчислювальних ресурсів, що є основою створеного програмного забезпечення.

Практичне значення отриманих результатів полягає у реалізації інформаційної технології оптимізації безсерверних обчислень та програмного забезпечення, що забезпечує прогнозування навантаження та проактивне масштабування ресурсів на хмарній платформі. Експериментальна апробація на реальних наборах даних продемонструвала, що аналітична модель на основі Марковських ланцюгів та напівмарковських процесів забезпечила прискорення обробки даних на 25,8%, зростання пропускної здатності на 21,3% та зменшення холодних стартів до 3% порівняно з класичним реактивним масштабуванням. Порівняння підходів показало, що аналітична модель перевищує алгоритм на основі машинного навчання за основними показниками продуктивності та забезпечує гарантії якості обслуговування без потреби у великих обсягах історичних наборів даних.

Результати інтелектуального аналізу даних та прогнозування, отримані за допомогою розробленої інформаційної технології, можуть бути використані для оптимізації безсерверних обчислень у хмарних середовищах. Розроблене програмне забезпечення є відкритим фреймворком, що може бути адаптоване для різних хмарних платформ та використане для побудови інтелектуальних систем автоматичного масштабування ресурсів на основі запропонованих моделей та алгоритмів.

Дисертація складається зі вступу, чотирьох розділів, висновків, переліку використаних джерел та чотирьох додатків.

У **вступі** обґрунтовано актуальність теми дослідження, сформульовано мету, предмет, об’єкт, завдання та методи дослідження, вказано наукову новизну, подано та проаналізовано зв’язок роботи з науковими темами. Зазначено особистий внесок здобувача, а також наведено відомості про апробацію та публікації основних результатів дисертації. Описано структуру та обсяг дисертаційної роботи.

Перший розділ дисертації містить теоретичні основи та огляд літератури у галузі безсерверних обчислень на хмарних платформах. Розглянуто ключові концепції парадигми безсерверних обчислень, зокрема моделі Function-as-a-Service та Backend-as-a-Service, архітектурні особливості подійно-орієнтованих систем, механізми автоматичного масштабування обчислювальних ресурсів та інтеграцію з мікросервісними архітектурами. Проаналізовано основні переваги та обмеження безсерверних обчислень, серед яких найбільш критичними визначено проблему холодного старту, залежність від постачальника хмарних послуг та складність моніторингу розподілених систем. Систематизовано метрики оцінювання ефективності безсерверних інформаційних систем та визначено відкриті проблеми, що потребують подальших досліджень та оптимізації.

Другий розділ присвячений аналізу архітектурних рішень для безсерверних обчислень та розробці архітектури фреймворку для розподіленої обробки даних. Проаналізовано роль подієво-орієнтованих архітектур у побудові інтелектуальних систем на хмарних платформах та проведено порівняльний аналіз ключових шаблонів комунікації. Обґрунтовано вибір системи черг повідомлень як базового компонента архітектури для задач асинхронної розподіленої обробки даних, що забезпечує буферизацію та згладжування пікових навантажень. Проведено детальний порівняльний аналіз підходів до реалізації комунікації в реальному часі. Визначено обмеження реактивного масштабування, зокрема холодні старти та непередбачувані затримки, що підтверджує актуальність розробки проактивних алгоритмів масштабування на основі прогнозування навантаження.

У **третьому розділі** побудовано математичну модель безсерверної інформаційної системи як неоднорідної системи масового обслуговування зі змішаними режимами роботи та розроблено методи оцінювання її ключових параметрів. Досліджено властивості моделі суміші потоків завдань, що формується як поєднання незалежних неоднорідних пуассонівських процесів, та встановлено її відмінності від класичного Markov-Modulated Poisson

Process. Доведено необхідну та достатню умову обмеженості системи черг та граничні еволюції процесу довжини черги у схемі усереднення та схемі дифузійної апроксимації з використанням апарату напівмарковських випадкових еволюцій, що базується на теорії Марковських ланцюгів та напівмарковських процесів. Розроблено метод оцінки параметрів суміші вхідного процесу на основі EM-алгоритму з використанням метрик хмарної платформи AWS. Запропоновано алгоритм параметричної оцінки та оптимізації конфігурації безсерверної системи, який на відміну від підходів на основі машинного навчання та нейронних мереж використовує аналітичні оцінки, що дозволяє отримати явний вигляд для прогнозування оптимальних параметрів без потреби у великих наборах даних для навчання моделей штучного інтелекту.

У **четвертому розділі** здійснено практичну верифікацію теоретичних результатів дисертаційного дослідження та реалізовано програмне забезпечення інформаційної технології оптимізації безсерверних обчислень на хмарній платформі AWS. У першому експерименті досліджено прогнозне автомасштабування на основі нейронних мереж DeepAR, яке на реальних наборах даних зменшило кількість холодних стартів на 27% та кількість необроблених запитів на 14%, при цьому встановлено обмеження підходу, зокрема потребу у великих обсягах історичних даних для навчання моделі. У другому експерименті досліджено прогнозне автомасштабування на основі аналітичної моделі з напівмарковськими процесами, що забезпечило прискорення обробки на 25,8% та зменшення холодних стартів до 3%. У третьому експерименті проведено імітаційне моделювання інтелектуальної системи методом Монте-Карло, що підтвердило переваги моделі суміші потоків. Результати інтелектуального аналізу даних показали, що аналітична модель перевищує алгоритм на основі машинного навчання за основними показниками продуктивності та забезпечує гарантії якості обслуговування та підтверджує ефективність розробленої інформаційної технології та програмного забезпечення для оптимізації безсерверних обчислень.

У **висновках** підсумовано основні результати дисертаційного дослідження.

У **додатках** подано наукові публікації, в яких відображено основні наукові результати роботи, відомості про апробацію результатів дисертації – акти та довідки про впровадження результатів роботи, лістинг частини коду програмного забезпечення.

Запропонована інформаційна технологія для прогнозного автомасштабування, що дозволяє заздалегідь розгортати обчислювальні ресурси на хмарній платформі використовується у роботі компанії Finker Finance B.V. та ФОП Вербицької С.І. А результати теоретичних та практичних досліджень використовуються у навчальному процесі кафедри математичних проблем управління і кібернетики та кафедри програмного забезпечення комп'ютерних систем Чернівецького національного університету імені Юрія Федьковича.

Ключові слова: інформаційна технологія, хмарна платформа, оптимізація, модель, алгоритм, машинне навчання, нейронні мережі, штучний інтелект, система черг, інтелектуальна система, інтелектуальний аналіз даних, Марковські та напівмарковські ланцюги, прогнозування, набір даних, програмне забезпечення.

ABSTRACT

Kyrychenko O. **Optimization of serverless computing in cloud environments.** – Qualification research work published in the manuscript.

Dissertation for the degree of Doctor of Philosophy, speciality 121 – "Software Engineering" – Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 2026.

This dissertation addresses the optimization of serverless computing in cloud environments under conditions of uneven workloads and heterogeneous incoming task streams. Existing approaches to automatic scaling of computing resources are predominantly reactive, leading to cold-start problems, inefficient resource

utilization, and increased latency in queueing systems. Although machine learning methods and neural networks can be used for workload prediction, they require large volumes of historical data and do not provide analytical guarantees of quality of service. The dissertation addresses the relevant scientific and applied problem of developing information technology to optimize serverless computing in cloud environments by constructing mathematical models based on the theory of heterogeneous queueing systems and developing the corresponding software.

The object of the study is the processes of distributed data processing in cloud environments using serverless technologies and message queue management systems.

The subject of the study is the methods and models for optimizing serverless computing in cloud environments using the theory of heterogeneous queueing systems.

The study aims to develop approaches to optimizing serverless computing in cloud environments that improve performance, reduce costs, and minimize cold starts by creating an information technology based on queueing theory for workload prediction and proactive scaling of computing resources.

The dissertation has **significantly further developed and improved** the mathematical model of a serverless computing system, using the AWS Lambda cloud platform as an example, by representing it as a heterogeneous queueing system with mixed operating modes, in which the input process is defined as a mixture of independent, non-homogeneous Poisson processes generated by different event sources. **A necessary and sufficient condition for stability condition** for the queue in such a heterogeneous queueing system **is established**. **For the first time**, limiting results for the queue-length process **are obtained** under both the averaging and diffusion approximation schemes, using the framework of semi-Markov random evolutions based on the theory of Markov chains and semi-Markov processes. The averaging scheme defines the deterministic trajectory of the mean workload, whereas the diffusion approximation scheme describes random deviations from this trajectory through an

Itô stochastic differential equation. This provides the foundation for predicting the behavior of an intelligent autoscaling system.

The dissertation has **also further developed and improved** the algorithm for parametric estimation and configuration optimization of a serverless system. Unlike existing approaches based on machine learning and neural networks, the proposed algorithm uses analytical estimates derived from queueing theory, enabling explicit formulas for optimal parameters and eliminating the need to train artificial intelligence models on large datasets.

An architectural framework for distributed data processing using serverless technologies on the AWS cloud platform **has been developed**. A distinctive feature of this architecture is its use of a message queue system, which enables loose coupling among the intelligent system's components, independent scaling when necessary, rapid recovery after failures, and more efficient use of computing resources. **An information technology** for distributed data processing based on serverless computing and proactive autoscaling of computing resources **has been developed and implemented**, forming the basis of the software created in this research.

The practical significance of the results lies in the implementation of an information technology for serverless computing optimization and software that provides workload prediction and proactive resource scaling on a cloud platform. Experimental validation on real-world datasets showed that the analytical model based on Markov chains and semi-Markov processes achieved a 25.8% increase in data processing speed, a 21.3% increase in throughput, and a reduction in cold starts to 3% compared with conventional reactive scaling. A comparative analysis also showed that the analytical model outperforms the machine-learning-based approach in the main performance indicators while providing quality-of-service guarantees without requiring large historical datasets.

The results of data mining and prediction obtained using the developed information technology can be used to optimize serverless computing in cloud environments. The developed software is an open framework that can be adapted

to different cloud platforms and used to build intelligent resource autoscaling systems based on the proposed models and algorithms.

The dissertation consists of an introduction, four chapters, a conclusion, a list of references, and four appendices.

The introduction substantiates the relevance of the research topic, formulates the aim, object, subject, objectives, and research methods, highlights the scientific novelty, and examines the connection of the work with existing scientific themes. It also specifies the author's personal contribution and provides information on the approval and publication of the main dissertation results. The structure and scope of the dissertation are described as well.

The first chapter presents the theoretical foundations and literature review in the field of serverless computing on cloud platforms. It examines the key concepts of the serverless computing paradigm, including the Function-as-a-Service and Backend-as-a-Service models, the architectural features of event-driven systems, mechanisms of automatic resource scaling, and integration with microservice architectures. The main advantages and limitations of serverless computing are analyzed, with particular emphasis on the cold-start problem, vendor lock-in, and the complexity of monitoring distributed systems. Metrics for evaluating the efficiency of serverless information systems are systematized, and open problems requiring further research and optimization are identified.

The second chapter is devoted to the analysis of architectural solutions for serverless computing and the development of a framework architecture for distributed data processing. It examines the role of event-driven architectures in building intelligent systems on cloud platforms and presents a comparative analysis of the main communication patterns. The choice of a message queue system as the core architectural component for asynchronous distributed data processing is justified by its ability to buffer requests and smooth peak workloads. A detailed comparative analysis of approaches to real-time communication is also provided. The limitations of reactive scaling, particularly cold starts and

unpredictable delays, are identified, which confirms the importance of developing proactive scaling algorithms based on workload prediction.

The third chapter develops a mathematical model of a serverless information system as a heterogeneous queueing system with mixed operating modes and proposes methods for estimating its key parameters. The properties of the task-stream mixture model, formed as a combination of independent non-homogeneous Poisson processes, are investigated, and its differences from the classical Markov-Modulated Poisson Process are established. A necessary and sufficient condition for the boundedness of the queueing system is established, and limiting results for the queue-length process under the averaging and diffusion approximation schemes are derived within the framework of semi-Markov random evolutions, based on the theory of Markov chains and semi-Markov processes. A method for estimating the parameters of the input-process mixture based on the EM algorithm and AWS platform metrics is developed. In addition, an algorithm for parametric estimation and optimization of the serverless system configuration is proposed. Unlike machine-learning- and neural-network-based approaches, it relies on analytical estimates, enabling explicit prediction of optimal parameters without requiring large training datasets.

The fourth chapter presents the practical verification of the theoretical results and the implementation of the proposed software for optimizing serverless computing on the AWS cloud platform. In the first experiment, predictive autoscaling based on DeepAR neural networks was studied; on real-world datasets, it reduced the number of cold starts by 27% and the number of unprocessed requests by 14%, while also revealing limitations of the approach, especially the need for large volumes of historical data for model training. In the second experiment, predictive autoscaling based on the analytical model with semi-Markov processes was investigated, resulting in a 25.8% increase in processing speed and a reduction in cold starts to 3%. In the third experiment, a Monte Carlo simulation of the intelligent system was carried out, confirming the advantages of the task-stream mixture model. The results of the data mining showed that the

analytical model outperforms the machine-learning-based approach in the main performance indicators, provides quality-of-service guarantees, and confirms the effectiveness of the developed information technology and software for serverless computing optimization.

The conclusions summarize the dissertation's main results.

The appendices contain scientific publications reflecting the main scientific results of the work, information on the approbation of the dissertation results, including acts and certificates of implementation, and a partial listing of the software code.

The proposed information technology for predictive autoscaling, which enables provisioning computing resources on a cloud platform in advance, is used by Finker Finance B.V. and Individual Entrepreneur S. I. Verbytska. The theoretical and practical results of the study are also used in the educational process of the Departments of Mathematical Problems of Control and Cybernetics and Software of Computer Systems at Yuriy Fedkovych Chernivtsi National University.

Keywords*:* information technology, cloud platform, optimization, model, algorithm, machine learning, neural networks, artificial intelligence, queuing system, intelligent system, data mining, Markov and semi-Markov chains, forecasting, dataset, software.

СПИСОК ПУБЛІКАЦІЙ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні наукові результати дисертації :

Наукові праці у виданнях, включених до переліку наукових фахових видань України та проіндексованих у наукометричній базі даних Scopus:

1. **Kyrychenko O.,** Ostapov S., Kyrychenko O. Design of a framework for serverless distributed data processing using queues. *Eastern-European Journal of Enterprise Technologies*. 2025. Vol.4, №9. P. 19–25. (**Scopus**) URL:

https://doi.org/10.15587/1729-4061.2025.335723 (*Особистий внесок: Кириченко О. О. – теоретична та практична розробка положень, розробка фреймворку, проведення дослідження, написання; Остапов С. Е. – постановка задачі, визначення загальної схеми дослідження, обговорення результатів; Кириченко О. Л. – аналіз літературних джерел, візуалізація, обговорення результатів*)

Наукові праці у періодичних наукових виданнях, проіндексованих у наукометричній базі даних Scopus:

2. **Kyrychenko O. O.**, Ostapov S. E., Kyrychenko O. L. Optimization of SQS Configurations for Efficient Batch Data Processing. *WSEAS Transactions on Systems.* 2025. Vol. 24. P. 36–43. (**Scopus**) URL: https://doi.org/10.37394/23202.2025.24.4 (*Особистий внесок: Кириченко О. О. – проведення дослідження, розробка хмарного прототипу, огляд літератури, написання, редагування рукопису; Остапов С. Е. – постановка задачі, концептуалізація; Кириченко О. Л. – валідація результатів дослідження, аналіз літературних джерел, написання, редагування*)

Наукові праці у виданнях, включених до переліку наукових фахових видань України:

3. Кириченко О., **Кириченко О**. Кешування даних у додатках з використанням безсерверної архітектури. *Information Technology: Computer Science, Software Engineering and Cyber Security*. 2024. №2. С. 42–49. URL: https://doi.org/10.32782/IT/2024-2-6 (*Особистий внесок: Кириченко О. О. – проведення дослідження, написання, редагування рукопису; Кириченко О. Л. – постановка задачі, керівництво дослідженням, рецензування та редагування*).

4. Кириченко О. Л., **Кириченко О. О.** Використання AWS APPSYNC для комунікації вебдодатків у реальному часі. *Вчені записки Таврійського національного університету імені В.І. Вернадського.* Серія : Технічні науки.

2024. 35(74), №4. С. 105–110. URL: https://doi.org/10.32782/2663-5941/2024.4/16 (*Особистий внесок: Кириченко О. О. – огляд літературних джерел, проведення дослідження, написання, редагування рукопису; Кириченко О. Л. – постановка задачі, визначення загальної схеми досліджень, обговорення результату*).

Наукові праці, які засвідчують апробацію матеріалів дисертації:

5. Антонюк Д. В., **Кириченко О. О.** Пакетна обробка даних на безсерверній архітектурі. *Проблеми інформатики та комп'ютерної техніки (ПІКТ–2023)* : праці XII праці Міжнар. наук.-практ. конф., м. Чернівці, 10–12 листопада 2023 р. Чернівці: Черн. нац. ун-т, 2023. С. 106–109. (*Особистий внесок: Кириченко О. О. – постановка задачі, методологія дослідження, проведення дослідження; Антонюк Д. В. – тестування, обговорення результатів, написання*).

6. Кириченко О. Л., **Кириченко О. О.** Покращення швидкодії безсерверних додатків за допомогою кешування. *Problems of Science and Technology: the Search for Innovative Solutions :* Proceedings of the XXIII International scientific and practical conference, Munich, Germany, 15–17 May, 2024. International Scientific Unity, 2024. P. 65–67. (*Особистий внесок: Кириченко О. О. – огляд літературних джерел, проведення досліджень, написання; Кириченко О. Л. – постановка задачі, редагування, рецензування)*.

7. **Kyrychenko O.**, Kyrychenko O. Real-time communication tools for web applications in a cloud environment. *The 13 th International Conference on Electronics, Communications and Computing's (IC ECCO)* : Materials of the Intern. Conf., Chisinau, Moldova, 17–18 October, 2024. P. 126–127. (*Особистий внесок: Кириченко О. О. – огляд літературних джерел, написання, рецензування та редагування; Кириченко О. Л. – постановка задачі, рецензування*).

8. **Кириченко О.О.,** Кириченко О. Л., Остапов С.Е. Аналіз продуктивності AWS SQS в умовах високих навантажень. *Проблеми*

інформатики та комп'ютерної техніки (ПІКТ–2024) : праці XIII Міжнар. наук.-практ. конф., м. Чернівці, 01–03 листопада 2024 р. Чернівці : Черн. нац. ун-т, 2024. С. 42–44. (*Особистий внесок: Кириченко О. О. – проведення досліджень, аналіз та опис результатів; Кириченко О. Л. – рецензування, обговорення результатів; Остапов С. Е. – постановка задачі, обговорення результатів*).

9. **Кириченко О**., Остапов С., Кириченко О. Безсерверний фреймворк із прогнозним масштабуванням на основі DeepAR для розподілених обчислень. *Trends and Prospects for the Development of Science and Education* : Proceedings of the 2nd International Scientific Conference (Oxford, United Kingdom, 10 July 2025). Lulu Press, Inc., 2025. P. 159–162. (*Особистий внесок: Кириченко О. О. – методологія дослідження, розробка програмного забезпечення, дослідження, написання; Остапов С. Е. – постановка задачі, загальне керівництво; Кириченко О. Л. – тестування, обговорення результатів*).

10. **Кириченко О.**, Малик І., Кириченко О. Оцінки довжини черги в неоднорідних системах масового обслуговування зі змінними режимами роботи. *Scientific Progress: Theories, Applications and Global Impact* : Proceedings of the 3rd International Scientific and Practical Conference (Braga, Portugal, March 2-4, 2026). European Open Science Space, 2026. P. 258–260. (*Особистий внесок: Кириченко О. О – аналіз та опис результату, написання; Малик І. В. – концептуалізація дослідження, аналіз результатів; Кириченко О. Л. – обговорення результатів, рецензування*).

Наукові праці, які додатково відображають наукові результати дисертації:

11. **Kyrychenko O.**, Ostapov S., Kyrychenko O. L. Predictive autoscaling in AWS Serverless by means of machine learning and SQS metrics. *CEUR Workshop Proceedings* : 6th International Workshop on Intelligent Information

Technologies & Systems of Information Security (IntelITSIS 2025, April 4, 2025). 2025. Vol.-3963. P. 77–87. (**Scopus)** URL: https://ceur-ws.org/Vol-3963/ (*Особистий внесок: Кириченко О. О. – проведення дослідження, тестування, написання; Остапов С. Е. – постановка задачі, концептуалізація; Кириченко О. Л. – обговорення результатів, рецензування*).