Вісник ЛДУБЖД
Bulletin of Lviv State University of Life Safety
https://journal.ldubgd.edu.ua/index.php/Visnuk
ISSN 2078-4643 (print), ISSN 2708-1389 (online)

UDC 004.75:004.042
DOI: 10.32447/20784643.33.2026.15

Д. Д. Смик, Н. Є. Бурак
Львівський державний університет безпеки життєдіяльності, м. Львів Україна
ORCID: https://orcid.org/0009-0002-9925-2349 – Д. Д. Смик
https://orcid.org/0000-0002-3880-4077 – Н. Є. Бурак
nazar.burak@ukr.net

МЕТОД АДАПТИВНОГО ПРОГНОЗНОГО МАСШТАБУВАННЯ СЕРВІСІВ У РОЗПОДІЛЕНИХ СИСТЕМАХ

У статті досліджено метод адаптивного прогнозного масштабування сервісів у розподілених системах як комплексний підхід до керування обчислювальними ресурсами в умовах динамічного навантаження. Обґрунтовано обмеженість традиційного реактивного масштабування, яке орієнтується виключно на поточний стан системи та не враховує майбутніх змін інтенсивності запитів. Встановлено, що адаптивне прогнозне масштабування забезпечує перехід від реактивної моделі реагування до випереджального керування, у межах якого рішення щодо зміни ресурсів приймається на основі аналізу історичних даних, прогнозу майбутнього навантаження та оцінки фактичних результатів функціонування системи.

У роботі визначено сутність методу як поєднання прогнозної, аналітичної та керувальної складових, що функціонують у межах безперервного циклу. Обґрунтовано п'ятиетапну структуру реалізації методу, яка включає моніторинг стану сервісів, аналітичну обробку даних і прогнозування навантаження, прийняття рішення про масштабування, практичну реалізацію зміни ресурсів та адаптаційне коригування правил керування. Показано, що ефективність такого підходу залежить від якості вхідних даних, релевантності обраних метрик, особливостей архітектури розподіленої системи та узгодженості між прогнозною моделлю і механізмом виконання рішень.

Проаналізовано сучасні підходи до реалізації адаптивного прогнозного масштабування, зокрема на основі часових рядів, методів машинного навчання, гібридних моделей, метрико-орієнтованих та архітектурно-узгоджених рішень. Визначено їх сильні сторони та обмеження. Доведено, що адаптивне прогнозне масштабування забезпечує більш гнучке керування ресурсами, підвищення стабільності сервісів та раціональніше використання інфраструктури порівняно з традиційними реактивними підходами, однак потребує складнішої організації процесів моніторингу та коригування параметрів моделі.

Ключові слова: адаптивне масштабування; прогнозне масштабування; розподілені системи; керування ресурсами; хмарні обчислення.

D. D. Smyk, N. Ye. Burak
Lviv State University of Life Safety, Lviv, Ukraine

METHOD OF ADAPTIVE PREDICTIVE SCALING OF SERVICES IN DISTRIBUTED SYSTEMS

The paper investigates the method of adaptive predictive scaling of services in distributed systems as a comprehensive approach to managing computing resources under dynamic workloads. The limitations of traditional reactive scaling, which relies solely on the current state of the system and does not account for future workload changes, are substantiated. It is established that adaptive predictive scaling enables the transition from reactive response to proactive resource management, where decisions on resource adjustment are based on historical data analysis, workload forecasting, and evaluation of actual system performance.

The essence of the method is defined as a combination of predictive, analytical, and control components operating within a continuous management cycle. A five-stage implementation structure is substantiated, including system monitoring, data analysis and load forecasting, decision-making on scaling, practical execution of resource changes, and adaptive correction of management rules. It is demonstrated that the effectiveness of this approach depends on data quality, relevance of selected metrics, architectural characteristics of distributed systems, and coordination between forecasting models and execution mechanisms.

Modern approaches to implementing adaptive predictive scaling are analysed, including time-series forecasting, machine learning methods, hybrid models, metric-oriented, and architecture-aware solutions. Their strengths and limitations are identified. The study confirms that adaptive predictive scaling provides more flexible resource management, improved service stability, and more rational infrastructure utilisation compared to traditional reactive approaches, while requiring a more sophisticated monitoring and model adjustment framework.
Keywords: adaptive scaling; predictive scaling; distributed systems; resource management; cloud computing.

Вступ. Стрімкий розвиток хмарних технологій, контейнеризації та мікросервісних архітектур зумовив суттєве ускладнення механізмів керування обчислювальними ресурсами в розподілених системах. Умови функціонування сучасних сервісів характеризуються високою динамічністю навантаження, нерівномірністю запитів та необхідністю забезпечення стабільної якості обслуговування за одночасного контролю інфраструктурних витрат. У сучасних дослідженнях також підкреслюється, що функціонування технічних систем дедалі більше залежить від складності зовнішнього середовища, невизначеності та впливу кризових чинників [26]. У таких умовах особливої актуальності набувають підходи, що дозволяють не лише реагувати на зміну стану системи, а й передбачати її майбутню динаміку. Саме тому, дослідження методів адаптивного прогнозного масштабування сервісів є важливим напрямом розвитку сучасних систем керування ресурсами в розподілених середовищах.

Постановка проблеми. Сучасні розподілені системи працюють в умовах постійної зміни навантаження, тому питання ефективного масштабування сервісів є одним із ключових для забезпечення їх стабільності та продуктивності. Традиційні підходи до масштабування переважно мають реактивний характер, тобто система збільшує або зменшує ресурси вже після зміни навантаження. На практиці це може спричиняти затримки в роботі сервісів, перевантаження окремих компонентів і нераціональне використання обчислювальних ресурсів. Саме тому, зростає інтерес до адаптивного прогнозного масштабування, яке дає змогу враховувати майбутній стан навантаження під час прийняття рішень.

Аналіз останніх досліджень і публікацій. У сучасних дослідженнях значна увага приділяється поєднанню масштабування з методами машинного навчання, прогнозування часових рядів і адаптивного керування ресурсами. Зокрема, Тхай О., Шаповал Н. [3] показують доцільність використання моделей штучного інтелекту для вдосконалення масштабування в середовищі Kubernetes. Пінтьє І., Ковач Я., Ловаш Р. [11] доводять, що результативність такого підходу залежить не лише від моделі прогнозування, а й від правильного вибору метрик для конкретного застосунку. Сопов О., Цитовцева А. [1] розглядають інтелектуальне масштабування як складову розвитку сучасних хмарних і мікросервісних систем, а Маєвський Я. Ю. [2] пов'язує підвищення ефективності масштабування з автоматизацією керування контейнеризованими застосунками в Kubernetes.

У закордонних дослідженнях проблема автоматичного масштабування розглядається ширше – як складова самокерованих і самоадаптивних хмарних систем. Зокрема, Чен Т., Бахсун Р., Яо С. [12] запропонували ґрунтовну класифікацію autoscaling-систем, у якій масштабування трактується як комплексний механізм керування ресурсами, що охоплює моніторинг, аналіз, прийняття рішень і адаптацію до змін середовища. Такий підхід формує теоретичну основу для розуміння адаптивного прогнозного масштабування як цілісного методу керування ресурсами.

Окремий напрям досліджень пов'язаний із використанням навчальних і самонавчальних моделей для прийняття рішень щодо масштабування. Так, Гарі Й., Монге Д. А., Пачіні Е., Матеос К., Гарсія Гаріно К. [13] показують, що autoscaling на основі reinforcement learning (навчання з підкріпленням) дає змогу враховувати складну динаміку хмарного середовища та вдосконалювати правила керування ресурсами залежно від отриманих результатів. Водночас Догані Дж., Намвар Р., Хунджуш Ф. [14], узагальнюючи сучасні auto-scaling (автоматичного масштабування) техніки для контейнерних платформ, підкреслюють необхідність урахування не лише поточного навантаження, а й архітектурних особливостей сервісів, взаємозв'язків між компонентами та інфраструктурних обмежень.

Сучасні оглядові праці також засвідчують зміщення autoscaling від реактивних до прогнозних і гібридних моделей. Зокрема, Альхарті С., Альшамсі А., Альсеярі А., Альварафі А. [15] виділяють reactive (реактивні), proactive (випереджальні), machine learning (методи машинного навчання) та time-series (методи на основі часових рядів) підходи як основні напрями розвитку автоматичного масштабування. Подібну логіку простежують і Чонг Б., Чонг Й.-С. [16], які в дослідженні з cloud-native computing (хмарно-орієнтованих обчислень) показують, що масштабування має забезпечувати не лише продуктивність сервісів, а й еластичність, стабільність та економічну доцільність.

Практичну цінність прогнозного підходу підтверджують і нові прикладні дослідження

Kubernetes-середовищ. Так, Гуруге М., Пріядаршана Я. [17] запропонували модель proactive autoscaling (випереджального автоматичного масштабування) на основі поєднання Facebook Prophet (інструменту прогнозування часових рядів) і LSTM (Long Short-Term Memory, різновиду рекурентної нейронної мережі), яка дає змогу точніше передбачати майбутню інтенсивність HTTP-запитів (запитів за протоколом передавання гіпертексту) і завчасно визначати необхідну кількість pod-реплік (екземплярів сервісу в Kubernetes).

Вагомий внесок у розвиток досліджень автоматичного масштабування зробили також Луонг Д.-Х., Тхіє Х.-Т., Ауттагартс А., Гамрі-Дудан Я. [19], Алі-Ельдін А., Ілюшкін А., Гіт Б., Хербст Н. Р., Пападопулос А., Іосуп А. [20], Агарвал П., Лакшмі Дж. [21], Чої Б., Парк Дж., Лі К., Хан Д. [22], Шеганаку Г., Шульте С., Вайбель П., Вебер І. [23], Денаро Г., Ель Мусса Н., Гейдаров Р., Ломіо Ф., Пецце М., Цю К. [24], Сачідананда В., Сівараман А. [25], у працях яких висвітлено спільно важливі аспекти сучасного autoscaling: прогнозне й проактивне керування ресурсами, оцінювання ефективності алгоритмів масштабування, урахування вартості ресурсів, архітектурних особливостей мікросервісів, а також ризиків і збоїв у розподілених системах. Узагальнення цих підходів дає змогу глибше розкрити адаптивне прогнозне масштабування як комплексний метод, що поєднує технічну результативність, гнучкість керування та практичну придатність у хмарних середовищах. Отже, аналіз сучасних публікацій свідчить, що адаптивне прогнозне масштабування розглядається як багатокомпонентний підхід, у межах якого поєднуються моніторинг, аналіз даних, прогнозування навантаження та інтелектуальне прийняття рішень.

Виділення невирішених частин загальної проблеми. Попри значну кількість праць, присвячених масштабуванню сервісів, прогнозуванню навантаження та використанню інтелектуальних моделей у хмарних середовищах, у науковій літературі недостатньо цілісно узагальнено метод адаптивного прогнозного масштабування сервісів саме як комплексний підхід до керування ресурсами в розподілених системах. Більшість досліджень висвітлює окремі аспекти проблеми – алгоритми прогнозування, вибір метрик, застосування машинного навчання, особливості Kubernetes або мікросервісних архітектур. Водночас потребують уточнення сутність цього методу, логіка його реалізації, послідовність основних етапів, сучасні підходи до застосування, а також його переваги, обмеження й практичні умови використання.

Метою статті є визначення сутності методу адаптивного прогнозного масштабування сервісів, аналіз сучасних підходів до його реалізації та обґрунтування практичних особливостей його використання в розподілених системах.

Методи дослідження. У дослідженні використано методи аналізу та узагальнення наукових джерел, порівняльний аналіз підходів до масштабування, структурно-функціональний аналіз етапів реалізації методу та елементи системного підходу до оцінювання його застосування в розподілених системах.

Результати. Поява методу адаптивного прогнозного масштабування сервісів пов’язана з ускладненням сучасних розподілених систем, у яких навантаження змінюється нерівномірно, а окремі сервіси працюють у тісній взаємодії між собою. У таких умовах традиційне реактивне масштабування вже не завжди забезпечує належну швидкість реагування, оскільки збільшення ресурсів відбувається після того, як навантаження вже зросло, що особливо відчутно в контейнерних і хмарних середовищах, де затримка між виявленням проблеми та фактичним запуском нових екземплярів сервісу може призвести до зростання часу відповіді, погіршення якості обслуговування і перевантаження окремих компонентів системи. Саме тому виникла потреба в такому підході, який дозволяє не лише реагувати на зміну стану системи, а й передбачати майбутнє навантаження та завчасно коригувати обсяг ресурсів – зміщення від реакції до випередження і стало головною передумовою формування адаптивного прогнозного масштабування як окремого напряму в дослідженнях керування ресурсами розподілених систем.

У праці Тхай О. та Шаповал Н. [3] метод розглядається у зв’язку з оптимізацією масштабування в середовищі Kubernetes за допомогою засобів штучного інтелекту. Автори виходять із того, що звичайні механізми масштабування орієнтуються переважно на поточні показники навантаження, тому мають обмеження в умовах швидких і різких змін попиту. На цій основі вони пропонують використовувати прогнозування навантаження на процесор як основу для більш своєчасного прийняття рішень щодо зміни кількості ресурсів. У межах цієї роботи сутність методу фактично зводиться до поєднання двох процесів: прогнозування майбутнього стану системи та автоматичного коригування її ресурсної конфігурації. Цінність цієї праці полягає в тому, що вона показує: прогнозне масштабування є не абстрактною ідеєю, а практичним способом підвищення точності рішень, зменшення затримок і кращого використання ресурсів у реальному програмному середовищі.

Пінтьє І., Ковач Я., Ловаш Р. [11] розкривають сутність цього підходу дещо ширше. Для них прогнозне масштабування – це не просто застосування певної моделі передбачення навантаження, а система керування, ефективність якої визначається якістю вхідних даних і правильним вибором показників, що характеризують стан застосунку. Автори наголошують, що для різних типів сервісів однакові метрики не дають однакового результату, тому метод повинен бути чутливим до особливостей конкретної системи. Такий підхід істотно доповнює попереднє трактування, оскільки дозволяє зрозуміти: адаптивність методу полягає не лише в автоматичній зміні ресурсів, а й у здатності спиратися на релевантні для конкретного середовища сигнали. Отже, у цій праці сутність методу розкривається через поєднання прогнозування, добору значущих метрик і керувального рішення, яке змінюється залежно від типу навантаження та поведінки застосунку.

Маєвський Я. Ю. [2] інтелектуальне масштабування розглядає як складову ширшого розвитку хмарних вебсервісів і мікросервісної архітектури. Досліджено не лише окремі алгоритми, а й загальну логіку переходу від звичайного масштабування до такого, що враховує попередні дані, архітектурні особливості системи та засоби машинного навчання, що дає підстави трактувати досліджуваний метод як комплексний механізм керування ресурсами, який об’єднує спостереження за системою, аналітичну обробку даних і прийняття рішення про зміну конфігурації сервісів. Важливо, що в цій праці масштабування не подається як ізольована технічна дія, а розглядається як частина загальної організації роботи сучасних розподілених сервісів. Саме тому вона є важливою для теоретичного осмислення сутності методу: вона показує, що йдеться не про окремий алгоритм, а про цілісний підхід до керування ресурсами в мінливому середовищі.

Опрацювання наведених праць дає можливість виділити кілька спільних положень. По-перше, в усіх дослідженнях вихідною точкою є визнання обмеженості реактивного масштабування в умовах динамічного навантаження; по-друге, сутність нового підходу пов’язується з використанням прогнозу майбутнього стану системи як підстави для прийняття рішення про зміну обсягу ресурсів; по-третє, дослідники підкреслюють, що ефективність такого підходу залежить не лише від самого прогнозу, а й від того, наскільки правильно система інтерпретує поточний стан сервісів, добирає показники для аналізу і коригує власні дії після отриманого результату. Отже, метод адаптивного прогнозного масштабування не можна зводити лише до математичного прогнозування навантаження. Його доцільно розглядати як поєднання прогнозної, аналітичної та керувальної складових, які спільно забезпечують випереджальне й гнучке масштабування сервісів.

З урахуванням опрацьованих наукових праць метод адаптивного прогнозного масштабування сервісів доцільно визначити як підхід до керування ресурсами розподіленої системи, за якого рішення про збільшення або зменшення обчислювальних ресурсів приймається на основі прогнозу майбутнього навантаження, аналізу поточного стану сервісів і подальшого коригування правил масштабування відповідно до фактичних результатів функціонування системи. У такому визначенні поєднано три ключові ознаки методу: випереджальний характер, тобто орієнтацію на майбутній стан; адаптивність, тобто здатність змінювати логіку керування залежно від нових даних; практичну спрямованість, тобто націленість на підтримання стабільності сервісів і раціональне використання ресурсів.

Реалізація методу адаптивного прогнозного масштабування сервісів у розподілених системах ґрунтується на послідовному поєднанні кількох взаємопов’язаних етапів . На відміну від звичайного реактивного масштабування, де рішення приймається лише після зміни навантаження, у цьому методі керування будується як безперервний цикл спостереження, аналізу, прогнозування, прийняття рішення та коригування подальших дій. Багатокомпонентна будова забезпечує його здатність не просто реагувати на стан системи, а завчасно враховувати можливі зміни в роботі сервісів. У сучасних дослідженнях підкреслюється, що ефективність цього методу визначається не окремим алгоритмом, а узгодженою роботою всіх його складових [3].

Першим етапом є моніторинг стану системи та збір даних – фіксуються показники, що характеризують навантаження на сервіси та використання обчислювальних ресурсів. До них можуть належати завантаження процесора, обсяг використаної пам’яті, кількість вхідних запитів, час відповіді сервісу, інтенсивність мережевого обміну, довжина черг запитів, кількість активних користувачів тощо. Значення цього етапу полягає в тому, що саме зібрані дані формують основу для подальшого аналізу і прогнозування. Якщо система отримує неповні, запізнілі або нерелевантні дані, то навіть якісна модель прогнозування не забезпечить правильного керувального рішення. На цю обставину звертають увагу Пінтьє І., Ковач Я., Ловаш Р. [2], підкреслюючи, що результативність масштабування значною мірою залежить від

правильного вибору метрик для конкретного типу застосунку.

Другим етапом є аналітична обробка даних і прогнозування майбутнього навантаження. На відміну від традиційних підходів, тут оцінюється не лише поточний стан системи, а й майбутня динаміка її роботи. Подібна логіка відповідає сучасним підходам до моделювання складних технічних процесів, у межах яких аналітичне оцінювання майбутнього стану системи розглядається як основа для прийняття обґрунтованих рішень [27]. На відміну від традиційних підходів, тут оцінюється не лише поточний стан системи, а й майбутня динаміка її роботи. Для цього використовуються історичні дані, закономірності зміни навантаження, періодичні коливання та короткочасні відхилення. У сучасних роботах для цього застосовуються як статистичні методи, так і моделі машинного навчання, нейронні мережі, гібридні схеми прогнозування [19]. Логіка цього етапу полягає в тому, що система має не просто визначити, що навантаження вже зросло, а спрогнозувати, чи зростатиме воно далі, наскільки швидко і який обсяг ресурсів буде необхідний для підтримання стабільної роботи сервісу. Саме прогнозний компонент перетворює масштабування з реактивного на випереджальне.

Третім етапом є прийняття рішення про масштабування. На основі прогнозу система визначає, чи потрібно збільшити, зменшити або залишити незмінним обсяг доступних ресурсів. На практиці це може означати збільшення кількості екземплярів сервісу, зміну обсягу процесорного часу чи пам’яті, перерозподіл навантаження між вузлами або використання комбінованих рішень. Важливо, що рішення не повинно бути механічним наслідком одного показника. Воно має враховувати не лише прогноз, а й поточний стан системи, її архітектурні особливості, чутливість до перевантаження, а також економічну доцільність зміни ресурсів.

Четвертим етапом є реалізація масштабування в інфраструктурі розподіленої системи– практична фаза, у межах якої ухвалене рішення втілюється в роботі конкретного середовища. У контейнерних системах це може бути запуск нових контейнерів або збільшення кількості реплік сервісу, у хмарних середовищах – додавання або виведення обчислювальних вузлів, а в мікросервісних архітектурах – зміна конфігурації окремих функціональних компонентів. Особливість цього етапу полягає в тому, що саме тут перевіряється реальна ефективність попередніх аналітичних дій. Якщо реалізація масштабування відбувається із затримкою або не відповідає архітектурі сервісу, прогнозна перевага методу може бути частково втрачена. Отже, практична результативність методу залежить не лише від точності прогнозу, а й від технічної здатності системи швидко виконати керувальне рішення.

П’ятим етапом є адаптація і коригування правил масштабування – забезпечує адаптивний характер методу. Після реалізації рішення система оцінює, наскільки воно виявилося правильним: чи дало змогу уникнути перевантаження, чи не спричинило надлишкового резервування ресурсів, чи відповідав фактичний стан системи прогнозованому. На основі такого зіставлення уточнюються параметри моделі, вагомість окремих метрик, часовий горизонт прогнозу та правила переходу до наступного рішення. У результаті метод не залишається статичним, а змінюється разом із системою, в якій він застосовується. Саме це дає підстави вважати його не просто прогнозним, а адаптивним. Як показують сучасні праці, без такого зворотного зв’язку прогнозне масштабування перетворюється лише на разове передбачення, тоді як повноцінний метод має функціонувати як циклічний механізм самокорекції [8].

Отже, структура методу адаптивного прогнозного масштабування охоплює п’ять основних компонентів: моніторинг, аналітичну обробку і прогнозування, прийняття рішення, практичну реалізацію масштабування та адаптаційне коригування – рис.1. Логіка його реалізації полягає в послідовному переході від збору інформації про стан системи до формування прогнозу, від прогнозу – до керувального рішення, а від результату цього рішення – до уточнення наступних дій. Саме така структура робить метод придатним для роботи в умовах змінного навантаження, де недостатньо просто реагувати на події, а необхідно передбачати їх і пристосовувати механізм керування до поведінки конкретної розподіленої системи.

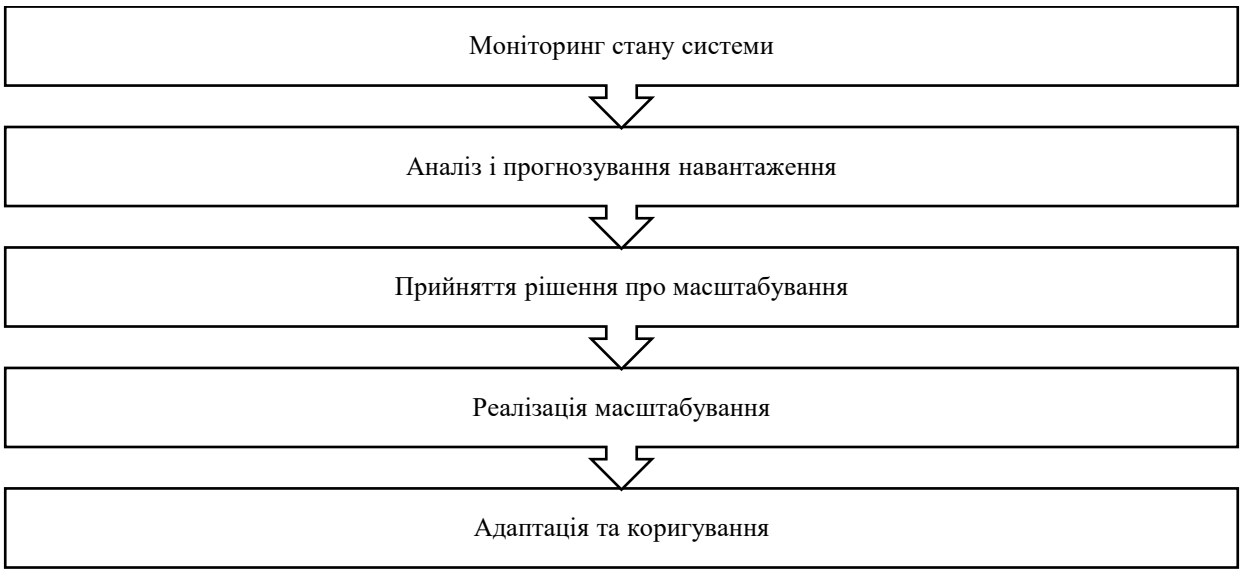

Рисунок 1 – Послідовність етапів методу адаптивного прогнозного масштабування сервісів у розподілених системах

Як показано на рис. 1, метод адаптивного прогнозного масштабування реалізується як послідовний і замкнений цикл керування ресурсами. Спочатку система безперервно збирає дані про власний стан, зокрема про рівень навантаження на сервіси, використання процесора, пам'яті, мережеву активність і час відповіді. Далі ці дані аналізуються, що дає змогу не лише оцінити поточну ситуацію, а й спрогнозувати її подальшу зміну. На основі такого прогнозу формується рішення щодо масштабування, тобто щодо збільшення, зменшення або збереження наявного обсягу ресурсів.

У практичному вимірі означає, що система не чекає моменту критичного перевантаження, а намагається підготувати необхідні ресурси завчасно. Наприклад, якщо прогноз показує зростання кількості запитів у найближчий проміжок часу, система може заздалегідь збільшити кількість екземплярів сервісу або перерозподілити ресурси між вузлами. Якщо ж очікується зниження навантаження, система, навпаки, скорочує обсяг задіяних ресурсів, що дозволяє уникнути їх надлишкового використання.

Важливою особливістю є те, що після виконання масштабування система знову оцінює фактичний результат своїх дій. Якщо прогноз виявився неточним або умови функціонування змінилися, правила керування коригуються. Саме тому цей метод є не лише прогнозним, а й адаптивним, оскільки він постійно уточнює свою роботу відповідно до реального стану розподіленої системи. У цілому наведена схема відображає практичну логіку методу як інструмента, що забезпечує баланс між стабільністю роботи сервісів і раціональним використанням обчислювальних ресурсів.

У сучасних наукових дослідженнях реалізація методу адаптивного прогнозного масштабування сервісів розглядається як багатоваріантний процес, у якому поєднуються засоби прогнозування навантаження, моделі прийняття рішень та механізми практичного керування ресурсами. Загальна тенденція розвитку цього напряму полягає у поступовому переході від простих реактивних схем до складніших інтелектуальних підходів, здатних враховувати часову динаміку навантаження, характер роботи конкретного сервісу та обмеження інфраструктури. Тобто в центрі уваги дослідників перебуває вже не саме масштабування як технічна дія, а логіка його випереджального та адаптивного здійснення. Саме тому сучасні підходи до реалізації цього методу доцільно розглядати не як ізольовані технічні рішення, а як різні способи організації єдиного процесу керування ресурсами в умовах змінного навантаження.

Одним із найбільш поширених підходів є реалізація методу на основі прогнозування часових рядів. У цьому випадку рішення про масштабування приймається після аналізу історичних даних про зміну навантаження на сервіс, що дозволяє виявити періодичні коливання, короткочасні піки та загальну тенденцію розвитку навантаження. Такий підхід є важливим насамперед тому, що він дає змогу перейти від фіксації вже наявної проблеми до оцінки її ймовірного виникнення в найближчому майбутньому. Саме таку логіку реалізації методу розглядають Ковальов А. В., Алєксєєв М. О. [8], використання часових рядів у середовищі Kubernetes дає можливість більш своєчасно готувати систему до зміни інтенсивності запитів.

Наукове значення цього підходу полягає в тому, що він формує базову прогнозну основу методу й демонструє, що навіть без складних інтелектуальних моделей випереджальне масштабування вже може бути ефективнішим за звичайне реактивне керування.

Подальший розвиток цього напряму пов'язаний із використанням методів машинного навчання, які дозволяють враховувати складніші залежності між станом сервісу та майбутнім навантаженням. На відміну від часових рядів у їх традиційному вигляді, такі моделі здатні виявляти приховані закономірності, які не завжди очевидні при звичайному статистичному аналізі. У праці Тхай О. та Шаповал Н. [1] показано, що застосування моделей штучного інтелекту в середовищі Kubernetes підвищує точність рішень щодо масштабування, оскільки система починає орієнтуватися не лише на поточний стан, а й на прогнозовану зміну навантаження. Особливістю цього підходу є те, що він розглядає масштабування як результат інтелектуальної інтерпретації даних, а не як механічну реакцію на перевищення певного порога. Отже, у межах цього напряму реалізації метод адаптивного прогнозного масштабування набуває ознак самонавчального механізму керування, який поступово пристосовується до поведінки конкретної системи.

Водночас сучасні дослідження показують, що сама по собі модель машинного навчання ще не гарантує високої ефективності масштабування. Не менш важливим є правильний добір показників, за якими система оцінює власний стан і формує основу для прогнозу. Саме на цьому наголошують Пінтьє І., Ковач Я., Ловаш Р. [2], які доводять, що результативність масштабування залежить від того, наскільки обрані метрики відповідають особливостям конкретного застосунку. Сучасний підхід до реалізації методу має бути не лише прогнозним, а й контекстно чутливим. Фактично йдеться про перехід від універсальних схем до спеціалізованих моделей, де одна й та сама логіка масштабування може давати різний результат залежно від архітектури сервісу, типу навантаження та характеру взаємодії між компонентами системи. У науковому плані цей підхід поглиблює розуміння адаптивності методу, оскільки показує, що адаптація стосується не тільки оновлення прогнозу, а й вибору релевантних сигналів для керування.

Окремий напрям сучасних досліджень пов'язаний із використанням гібридних моделей реалізації методу, коли в одному рішенні поєднуються кілька способів прогнозування. Такі моделі створюються з метою підвищення точності прогнозу в умовах нестабільного або складно передбачуваного навантаження. У праці Сімакін С. К., Божуха Л. М. [6] запропоновано гібридний підхід, у якому поєднуються різні інструменти прогнозування для підвищення якості рішень щодо масштабування контейнеризованих сервісів. Наукова цінність цього напряму полягає в тому, що він демонструє: жодна окрема модель не є універсальною для всіх типів навантаження, а отже, підвищення ефективності методу часто потребує комбінування кількох аналітичних інструментів. Гібридизація дозволяє зробити масштабування стійкішим до коливань навантаження і водночас зберегти достатню гнучкість керування ресурсами.

Ще один важливий підхід стосується реалізації методу в мікросервісних і контейнерних системах, де об'єктом масштабування є не окремий ізольований сервіс, а ціла структура взаємопов'язаних компонентів. У таких умовах масштабування одного сервісу не завжди розв'язує проблему продуктивності, оскільки вузьке місце може виникнути в іншому компоненті системи. Саме тому Помелуйко Д. А., Чалий М. Ф. [7] розглядають ефективне масштабування як завдання, що потребує врахування складності контейнеризованих мікросервісів і координації рішень між кількома сервісними елементами. Сучасна реалізація методу дедалі більше відходить від спрощеного розуміння масштабування як збільшення кількості екземплярів сервісу та переходить до системного керування сервісною архітектурою в цілому. Отже, практична реалізація методу в новітніх умовах вимагає врахування не лише прогнозу навантаження, а й внутрішньої будови розподіленої системи.

Не менш важливим є підхід, у якому масштабування розглядається з позицій поєднання технічної ефективності й економічної доцільності. У сучасних хмарних середовищах збільшення обчислювальних ресурсів безпосередньо пов'язане зі зростанням витрат, тому надмірне резервування інфраструктури може бути не менш небажаним, ніж нестача потужностей. Саме тому Бугров А. А. [9] пропонує розглядати реалізацію методу як процес оптимального розподілу ресурсів, де метою є не максимальне нарощування потужностей, а досягнення збалансованого стану між продуктивністю системи та вартістю її підтримки.

Отже, сучасні підходи до реалізації методу адаптивного прогнозного масштабування сервісів можна поділити на кілька основних груп: прогнозування за часовими рядами, використання методів машинного навчання, гібридні моделі, реалізація в мікросервісних середовищах і підходи, орієнтовані на баланс між продуктивністю та витратами. Їх об'єднує спільна мета – забезпечення випереджального й гнучкого

керування ресурсами розподіленої системи, однак вони відрізняються глибиною аналізу, характером використаних даних і практичними умовами застосування.

Таблиця 1

Сучасні підходи до реалізації методу адаптивного прогнозного масштабування сервісів

Підхід	Зміст підходу	Сильні сторони	Обмеження
На основі часових рядів	Прогнозування майбутнього навантаження за історичними даними про роботу сервісу	Відносна простота реалізації, зрозуміла логіка прогнозу	Менш ефективний за нестабільних і різко змінних навантажень
На основі машинного навчання	Використання моделей, що навчаються на даних системи та виявляють складні залежності	Вища точність рішень, здатність враховувати приховані закономірності	Високі вимоги до обсягу та якості даних
Метрико-орієнтований	Добір релевантних показників для конкретного типу застосунку	Краще узгодження рішення зі станом реального сервісу	Потребує попереднього глибокого аналізу системи
Гібридний	Поєднання кількох моделей прогнозування в одному рішенні	Вища стійкість прогнозу та гнучкість використання	Складність інтеграції та налаштування
Мікросервісно-архітектурний	Урахування взаємозв'язків між кількома сервісами в межах однієї системи	Краще відображає реальну будову розподіленого середовища	Складніший у практичній реалізації
Ресурсно-економічний	Поєднання технічної ефективності масштабування з контролем витрат	Дає змогу уникати надлишкового резервування ресурсів	Може вимагати компромісу між швидкодією і вартістю

Джерело: узагальнено авторами на основі [1;2;3;5]

Дані табл. 1 свідчать, що сучасні підходи до реалізації методу адаптивного прогнозного масштабування не є взаємовиключними, а радше відображають різні акценти в межах одного дослідницького напряму. Якщо підходи, засновані на часових рядах, забезпечують базову прогнозну логіку й відносну простоту впровадження, то моделі машинного навчання дозволяють істотно поглибити аналіз навантаження та підвищити точність рішень. Водночас метрико-орієнтований підхід показує, що якість масштабування залежить не тільки від самої моделі, а й від того, наскільки правильно обрано показники, за якими оцінюється стан системи.

Поряд із цим гібридні та мікросервісно-архітектурні підходи демонструють подальше ускладнення сучасних механізмів реалізації методу. Їх поява зумовлена тим, що реальні розподілені системи дедалі рідше є простими за структурою, а отже, потребують врахування як різнорідності навантаження, так і внутрішніх залежностей між сервісами. Ресурсно-економічний підхід, своєю чергою, підкреслює, що сучасне масштабування вже не можна оцінювати лише з позиції технічної продуктивності, оскільки в хмарних середовищах будь-яке рішення про додаткові ресурси має ще й економічний вимір. Отже, сучасна реалізація адаптивного прогнозного масштабування розвивається в напрямі багатокритеріального керування, де одночасно враховуються точність прогнозу, архітектура системи, характер навантаження та вартість використання ресурсів.

У прикладних дослідженнях технічних систем також підкреслюється, що своєчасне визначення критичних параметрів є важливою умовою ефективного керування та запобігання небажаним наслідкам [28]. На відміну від реактивного масштабування, цей підхід орієнтований на випереджальне керування, тому його ефективність залежить від узгодженості між моніторингом, прогнозною моделлю та механізмом виконання рішень – рис.1. Ключовою умовою практичного використання методу є якість вхідних даних. Для коректного прогнозування недостатньо спиратися лише на загальні системні показники, оскільки для різних типів сервісів визначальними можуть бути різні метрики: інтенсивність запитів, час відповіді, навантаження на базу даних або довжина черг. Тому результативність методу значною мірою залежить від правильного добору показників, які найбільш точно відображають стан конкретного застосунку [10]. Крім того, у мікросервісних середовищах масштабування має враховувати не лише окремий сервіс, а й його взаємозв'язки з іншими компонентами системи, оскільки локальне збільшення ресурсів не завжди усуває проблему на рівні всієї архітектури [6].

Не менш важливими є часовий горизонт прогнозування та економічна доцільність масштабування. Надто короткий горизонт не забезпечує випереджального характеру керування, тоді як надто тривалий підвищує ймовірність похибки прогнозу. Водночас у хмарних середовищах кожне рішення про розширення

ресурсів супроводжується додатковими витратами, тому метод має забезпечувати баланс між продуктивністю сервісів і вартістю інфраструктури [7]. Отже, практична ефективність адаптивного прогнозного масштабування визначається поєднанням якісного моніторингу, релевантних метрик, урахування архітектури системи та економічно обґрунтованого керування ресурсами.

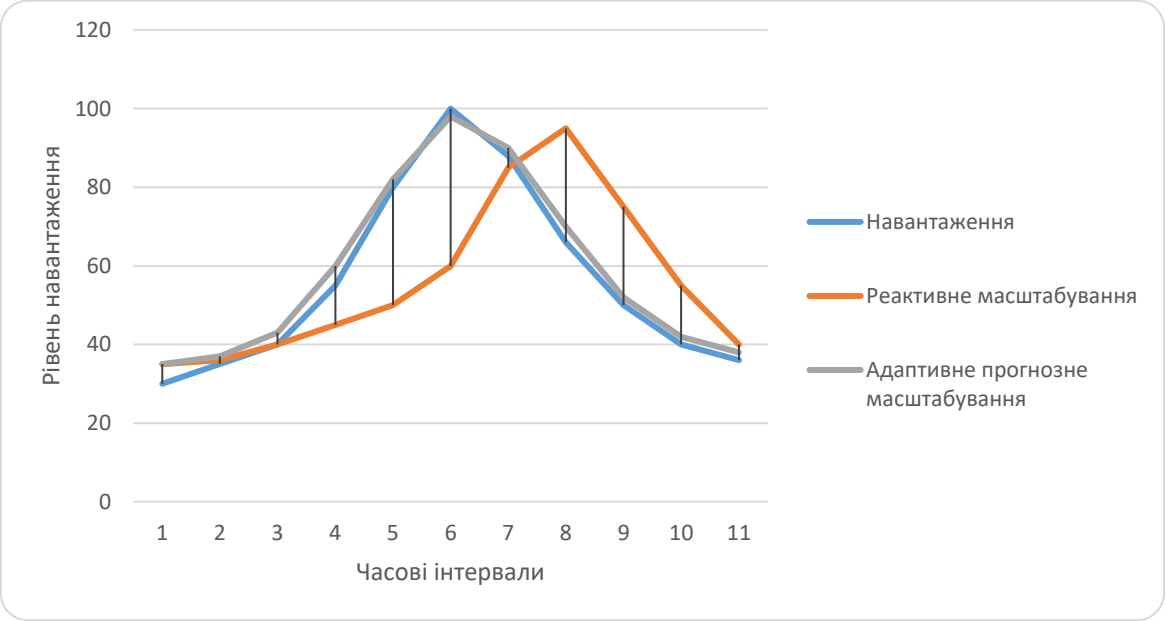

Рисунок 2 – Ілюстративне порівняння реактивного та адаптивного прогнозного масштабування сервісів

З метою більш наочного відображення сильних і слабких сторін методу адаптивного прогнозного масштабування доцільно подати їх у вигляді таблиці 2.

Таблиця 2

Переваги та недоліки методу адаптивного прогнозного масштабування сервісів у розподілених системах

Переваги	**Недоліки**
Забезпечує випереджальне керування ресурсами на основі прогнозу майбутнього навантаження	Потребує якісних і релевантних даних моніторингу
Зменшує ймовірність перевантаження сервісів у пікові періоди	Ефективність залежить від точності прогнозної моделі
Сприяє підвищенню стабільності роботи сервісів і зниженню часу відповіді	Складніший у реалізації порівняно з традиційними підходами
Дає змогу раціональніше використовувати обчислювальні ресурси	Вимагає постійного коригування параметрів моделі
Може враховувати особливості архітектури та динаміку навантаження конкретної системи	У мікросервісних середовищах потребує врахування міжсервісних залежностей
Поєднує технічну ефективність із можливістю оптимізації інфраструктурних витрат	За помилкового прогнозу можливе надлишкове або недостатнє масштабування

Наведені в табл. 2 переваги та недоліки свідчать, що метод адаптивного прогнозного масштабування має суттєвий потенціал для використання в сучасних розподілених системах, насамперед завдяки випереджальному характеру керування ресурсами. Його основна цінність полягає в можливості завчасно реагувати на зміну навантаження, підвищувати стабільність сервісів і забезпечувати більш раціональне використання обчислювальних потужностей. Водночас ефективність цього методу не є автоматично гарантованою, оскільки вона безпосередньо залежить від якості даних, точності прогнозу та здатності системи коректно реалізувати рішення про масштабування. Отже, адаптивне прогнозне масштабування доцільно розглядати як перспективний, але методично й технічно складний підхід, ефективність якого визначається рівнем узгодженості всіх елементів системи керування ресурсами.

Сутність методу адаптивного прогнозного масштабування сервісів у розподілених системах полягає у випереджальному та адаптивному керуванні обчислювальними ресурсами, за якого рішення про їх зміну приймається на основі прогнозу майбутнього навантаження, аналізу поточного стану сервісів і подальшого коригування параметрів масштабування відповідно до фактичних результатів функціонування системи. У результаті дослідження встановлено, що ключовою ознакою цього методу є поєднання прогнозної та адаптаційної складових, що забезпечує перехід від реактивної моделі реагування до випереджального керування ресурсами. Доведено, що метод не зводиться до окремого

алгоритму прогнозування, а функціонує як цілісний циклічний механізм, який охоплює моніторинг, аналітичну обробку даних, прийняття рішення, реалізацію масштабування та зворотний зв'язок. Визначено, що саме узгодженість цих етапів забезпечує стабільність сервісів у змінному середовищі навантаження.

Практична значущість методу адаптивного прогнозного масштабування сервісів підтверджується його впровадженням у сучасних хмарних платформах. Зокрема, у сервісі Google Compute Engine Managed Instance Groups реалізовано механізм predictive autoscaling, який дає змогу автоматично змінювати кількість віртуальних машин на основі прогнозу майбутнього навантаження. Такий механізм орієнтований насамперед на системи, у яких додавання нових екземплярів потребує певного часу, а навантаження має повторювані добові або тижневі коливання. Аналогічний принцип використовується і в Amazon EC2 Auto Scaling, де predictive scaling аналізує історичні дані про трафік, виявляє закономірності зміни навантаження та забезпечує проактивне збільшення обчислювальних потужностей до моменту очікуваного піку. Отже, метод адаптивного прогнозного масштабування сервісів уже має практичне застосування в комерційних хмарних середовищах, що свідчить про його прикладну цінність для цифрових сервісів із динамічним характером споживання ресурсів.

Таблиця 3

Приклади практичної реалізації адаптивного прогнозного масштабування сервісів у хмарних платформах

Платформа / сервіс	Реалізований механізм	Які дані враховуються	Особливість практичного застосування	Значення для адаптивного прогнозного масштабування
Google Compute Engine Managed Instance Groups	Predictive autoscaling	Історичні дані про навантаження, зокрема показники CPU	Дає змогу завчасно додавати або вилучати VM-екземпляри в керованій групі, якщо застосунок має тривалий час ініціалізації	Забезпечує підготовку ресурсів до очікуваного піку навантаження, а не реакцію після його настання
Amazon EC2 Auto Scaling	Predictive scaling	Історичні дані про навантаження з виявленням добових і тижневих патернів трафіку	Дає змогу прогнозувати потребу в потужностях і проактивно масштабувати Auto Scaling group	Зменшує ризик дефіциту ресурсів у період прогнозованого зростання навантаження

Отже, адаптивне прогнозне масштабування сервісів уже реалізується в провідних хмарних екосистемах у вигляді механізмів прогнозного автоскейлінгу. Спільною ознакою наведених рішень є орієнтація не лише на поточний стан системи, а й на прогноз майбутньої динаміки навантаження, побудований на основі історичних метрик та дає змогу розглядати адаптивне прогнозне масштабування не лише як теоретичну модель керування ресурсами, а як апробований інструмент забезпечення стійкості та ефективності функціонування цифрових сервісів у середовищах із нерівномірним попитом.

Таким чином, адаптивне прогнозне масштабування сервісів доцільно розглядати як один із найбільш перспективних напрямів розвитку сучасних механізмів керування ресурсами в розподілених системах. Його значення визначається не лише здатністю забезпечувати своєчасну зміну обсягу ресурсів, а насамперед переходом до випереджального та гнучкого керування, орієнтованого на прогноз майбутнього стану системи. У такому розумінні адаптивне прогнозне масштабування виходить за межі окремого алгоритму чи локального інструмента й постає як цілісний механізм узгодження моніторингу, аналізу, прогнозування, реалізації рішень і подальшого коригування правил керування. Саме така логіка робить його особливо актуальним для хмарних, контейнеризованих і мікросервісних середовищ, де стабільність сервісів дедалі більше залежить від здатності системи не лише реагувати на зміни навантаження, а й завчасно пристосовуватися до них.

Висновки. Ефективність адаптивного прогнозного масштабування визначається сукупністю конкретних чинників: якістю вхідних даних, релевантністю обраних метрик, типом архітектури розподіленої системи, правильністю визначення горизонту прогнозування та економічною доцільністю рішень щодо розподілу ресурсів. Встановлено, що використання моделей прогнозування, зокрема на основі часових рядів, машинного навчання та гібридних підходів, підвищує точність масштабування лише за умови їх інтеграції з адаптаційним механізмом самокорекції. Обґрунтовано, що основними перевагами методу є зменшення ризику

перевантаження сервісів, підвищення стабільності їх роботи та більш раціональне використання обчислювальних ресурсів, тоді як до його обмежень належать складність реалізації та залежність від точності прогнозу. Отримані результати підтверджують доцільність використання адаптивного прогнозного масштабування в хмарних, контейнеризованих і мікросервісних середовищах як перспективного напряму розвитку систем керування ресурсами.

Список літератури:

1. Сопов О., Цитовцева А. Особливості масштабування контейнерного навантаження на базі системи Kubernetes. *Технічні науки та технології.* 2021. № 1(23). С. 103–108.
2. Маєвський Я. Ю. Підвищення ефективності автоматизації масштабування мікросервісів у системі керування контейнеризованими застосунками Kubernetes. *Вісник Хмельницького національного університету. Технічні науки.* 2022. № 5. С. 260–264. DOI: https://doi.org/10.31891/2307-5732-2024-333-2-12
3. Тхай О., Шаповал Н. Оптимізація горизонтального масштабування контейнеризованих застосунків у середовищі Kubernetes засобами штучного інтелекту. *Системні дослідження та інформаційні технології.* 2025. № 1. С. 45–58.
4. Федоришин Б., Краско О. Міграція сервісів в кластері Kubernetes на основі архітектурного шаблону “sidecar”. *Information and Telecommunication Technologies, Radio Electronics.* 2024. Вип. 4, № 2. DOI:https://doi.org/10.23939/ictee2024.02.082
5. Пазинін А. С. Методи автоматичного масштабування в хмарних середовищах. *Кібербезпека: освіта, наука, техніка,* 2024. № 2(26). DOI: https://doi.org/10.28925/2663-4023.2024.26.715 (дата звернення: 07.03.2026).
6. Сімакін С. К., Божуха Л. М. Прогнозування навантаження на сервер з використанням ШІ для оптимізації веб-сервісів. *Актуальні проблеми автоматизації та інформаційних технологій.* 2024. № 28. С. 234–243.
7. Помелуйко Д. А., Чалий М. Ф. Методи масштабування Kubernetes-кластеру в хмарному середовищі. *Системи управління, навігації та зв’язку.* 2025. № 2. DOI: https://doi.org/10.26906/SUNZ.2025.2.175
8. Ковальов А. В., Алєксєєв М. О. Підхід до автомасштабування кластерів Kubernetes на основі персональних обчислювальних ресурсів користувачів. Збірник матеріалів Міжнародної науково-технічної конференції *«Перспективи телекомунікацій».* 2025. С. 282–286.
9. Бугров А. А. Моделі і методи вдосконалення високонавантажених розподілених систем : дис. … доктора філософії : 126 «Інформаційні системи та технології». Київ : Київський національний університет будівництва і архітектури, 2025.
10. Шушура О. М., Ігнатов Д. А. Моделювання прийняття рішень з управління ресурсами мікросервісів та вибору віртуальних машин у середовищі Kubernetes. *Вісник Херсонського національного технічного університету.* 2025. Т. 2, № 3(94). С. 526–532. DOI: https://doi.org/10.35546/kntu2078-4481.2025.3.2.67
11. Pintye I., Kovács J., Lovas R. Impact of metric selection on the effectiveness of machine-learning autoscalers for cloud applications. *Journal of Grid Computing,* 2024. Vol. 22, № 4. P. 1–20.
12. Chen T., Bahsoon R., Yao X. A survey and taxonomy of self-aware and self-adaptive cloud autoscaling systems. *ACM Computing Surveys.* 2018. Vol. 51, № 3. Article 61. DOI: 10.1145/3190507
13. Garí Y., Monge D. A., Pacini E., Mateos C., García Garino C. Reinforcement learning-based application autoscaling in the cloud: A survey. *Engineering Applications of Artificial Intelligence.* 2021. Vol. 102. DOI: 10.1016/j.engappai.2021.104288
14. Dogani J., Namvar R., Khunjush F. Auto-scaling techniques in container-based cloud and edge/fog computing: Taxonomy and survey. *Computer Communications.* 2023. Vol. 209. P. 120–150. DOI: 10.1016/j.comcom.2023.06.010.
15. Alharthi S., Alshamsi A., Alseiari A., Alwarafy A. Auto-scaling techniques in cloud computing: Issues and research directions. *Sensors.* 2024. Vol. 24, № 17. Article 5551. DOI: 10.3390/s24175551
16. Jeong B., Jeong Y.-S. Autoscaling techniques in cloud-native computing: A comprehensive survey. *Computer Science Review.* 2025. Vol. 58. Article 100791. DOI: https://doi.org/10.1016/j.cosrev.2025.100791
17. Guruge P. B., Priyadarshana Y. H. P. P. Time series forecasting-based Kubernetes autoscaling using Facebook Prophet and Long Short-Term Memory. *Frontiers in Computer Science.* 2025. Vol. 7. DOI:10.3389/fcomp.2025.1509165
18. Smyk D., Burak N. Методи та засоби обробки даних в сучасних автоматизованих системах. *Вісник Львівського державного університету безпеки життєдіяльності.* 2025. № 31. С. 41–49. DOI: 10.32447/20784643.31.2025.05
19. Luong D.-H., Thieu H.-T., Outtagarts A., Ghamri-Doudane Y. Predictive Autoscaling Orchestration for Cloud-native Telecom Microservices. 2018 IEEE 5G World Forum (5GWF). 2018. P. 153–158. DOI:10.1109/ACCESS.2020.3025032
20. Ali-Eldin A., Ilyushkin A., Ghit B., Herbst N. R., Papadopoulos A., Iosup A. Which Cloud Auto-Scaler Should I Use for my Application?

Benchmarking Auto-Scaling Algorithms. *Proceedings of the 2016 ACM/SPEC International Conference on Performance Engineering*. 2016. P. 131–132.

21. Agarwal P., Lakshmi J. Cost Aware Resource Sizing and Scaling of Microservices. Proceedings of the 2019 International Conference on Cloud Computing and Internet of Things. 2019. P. 66–74. DOI: 10.1145/3361821.3361823.

22. Choi B., Park J., Lee C., Han D. pHPA: A Proactive Autoscaling Framework for Microservice Chain // Proceedings of the 5th Asia-Pacific Workshop on Networking. 2021. P. 65–71. DOI:10.1145/3469393.3469401

23. Sheganaku G., Schulte S., Waibel P., Weber I. Cost-efficient auto-scaling of container-based elastic processes. *Future Generation Computer Systems*. 2023. Vol. 138. P. 296–312. DOI: https://doi.org/10.1016/j.future.2022.09.001

24. Denaro G., El Moussa N., Heydarov R., Lomio F., Pezzè M., Qiu K. Predicting Failures of Autoscaling Distributed Applications. *Proceedings of the ACM on Software Engineering*. 2024. Vol. 1, Issue FSE. P. 1960–1981.

25. Sachidananda V., Sivaraman A. Erlang: Application-Aware Autoscaling for Cloud Microservices. *Proceedings of the Nineteenth European Conference on Computer Systems.* 2024. P. 888–923. DOI: 10.1145/3627703.3650084.

26. Bun R., Marland G., Oda T., See L., Puliafito E., Nahorski Z., Jonas M., Kovalyshyn V., Ialongo I., Yashchun O., Romanchuk Z. Tracking unaccounted greenhouse gas emissions due to the war in Ukraine since 2022. *Science of the Total Environment*. 2024. Vol. 914. DOI: https://doi.org/10.1016/j.scitotenv.2024.169879

27. Tatsiy R. M., Pazen O. Yu., Vovk S. Ya., Ropyak L. Ya., Pryhorovska T. O. Numerical study on heat transfer in multilayered structures of main geometric forms made of different materials. *Journal of the Serbian Society for Computational Mechanics*. 2019. Vol. 13, № 2. P. 36–55. DOI: https://doi.org/10.24874/jsscm.2019.13.02.04

28. Hulida E., Pasnak I., Koval O., Tryhuba A. Determination of the critical time of fire in the building and ensure successful evacuation of people. *Periodica Polytechnica Civil Engineering.* 2019. Vol. 63, № 1. P. 308–316. DOI: https://doi.org/10.3311/PPci.12760

References:

1. Sopov, O., & Tsytovtseva, A. (2021). Osoblyvosti masshtabuvannia konteinernoho navantazhennia na bazi systemy Kubernetes [Peculiarities of scaling container workload based on the Kubernetes system]. Tekhnichni nauky ta tekhnolohii, 1(23), 103–108 [in Ukrainian].

2. Maievskyi, Ya. Yu. (2022). Pidvyshchennia efektyvnosti avtomatyzatsii masshtabuvannia mikroservisiv u systemi keruvannia konteineryzovanymy zastosunkamy Kubernetes [Improving the efficiency of automation of microservice scaling in the management system of containerized Kubernetes applications]. Visnyk Khmelnytskoho natsionalnoho universytetu. Tekhnichni nauky, 5, 260–264. https://doi.org/10.31891/2307-5732-2024-333-2-12 [in Ukrainian].

3. Tkhai, O., & Shapoval, N. (2025). Optymizatsiia horyzontalnoho masshtabuvannia konteineryzovanykh zastosunkiv u seredovyshchi Kubernetes zasobamy shtuchnoho intelektu [Optimization of horizontal scaling of containerized applications in the Kubernetes environment using artificial intelligence tools]. Systemni doslidzhennia ta informatsiini tekhnolohii, 1, 45–58 [in Ukrainian].

4. Fedoryshyn, B., & Krasko, O. (2024). Mihratsiia servisiv v klasteri Kubernetes na osnovi arkhitekturnoho shablonu “sidecar” [Migration of services in a Kubernetes cluster based on the “sidecar” architectural pattern]. Information and Telecommunication Technologies, Radio Electronics, 4(2). https://doi.org/10.23939/ictee2024.02.082 [in Ukrainian].

5. Pazynin, A. S. (2024). Metody avtomatychnoho masshtabuvannia v khmarnykh seredovyshchakh [Methods of automatic scaling in cloud environments]. Kiberbezpeka: osvita, nauka, tekhnika, 2(26). https://doi.org/10.28925/2663-4023.2024.26.715 [in Ukrainian].

6. Simakin, S. K., & Bozhukha, L. M. (2024). Prohnozuvannia navantazhennia na server z vykorystanniam ShI dlia optymizatsii veb-servisiv [Server load forecasting using AI for web service optimization]. Aktualni problemy avtomatyzatsii ta informatsiinykh tekhnolohii, 28, 234–243 [in Ukrainian].

7. Pomeluiko, D. A., & Chalyi, M. F. (2025). Metody masshtabuvannia Kubernetes-klasteru v khmarnomu seredovyshchi [Methods of scaling a Kubernetes cluster in a cloud environment]. Systemy upravlinnia, navihatsii ta zviazku. https://doi.org/10.26906/SUNZ.2025.2.175 [in Ukrainian].

8. Kovalov, A. V., & Alieksieiev, M. O. (2025). Pidkhid do avtomasshtabuvannia klasteriv Kubernetes na osnovi personalnykh obchysliuvalnykh resursiv korystuvachiv [An approach to autoscaling Kubernetes clusters based on users’ personal computing resources]. Zbirnyk materialiv Mizhnarodnoi naukovo-tekhnichnoi konferentsii “Perspektyvy telekomunikatsii”, 282–286 [in Ukrainian].

9. Buhrov, A. A. (2025). Modeli i metody vdoskonalennia vysokonavantazhenykh rozpodilenykh system [Models and methods for improving high-load distributed systems]. Doctor’s thesis. Kyiv: Kyivskyi natsionalnyi universytet budivnytstva i arkhitektury [in Ukrainian].

10. Shushura, O. M., & Ihnatov, D. A. (2025). Modeliuvannia pryiniattia rishen z upravlinnia resursamy mikroservisiv ta vyboru virtualnykh mashyn u seredovyshchi Kubernetes [Modeling decision-making on microservice resource management and virtual machine selection in the Kubernetes environment]. Visnyk Khersonskoho natsionalnoho tekhnichnoho universytetu, 2, 3(94), 526–532. https://doi.org/10.35546/kntu2078-4481.2025.3.2.67 [in Ukrainian].

11. Pintye, I., Kovács, J., & Lovas, R. (2024). Impact of metric selection on the effectiveness of machine-learning autoscalers for cloud applications. Journal of Grid Computing, 22(4), 1–20

12. Chen, T., Bahsoon, R., & Yao, X. (2018). A survey and taxonomy of self-aware and self-adaptive cloud autoscaling systems. ACM Computing Surveys, 51(3), Article 61. https://doi.org/10.1145/3190507

13. Garí, Y., Monge, D. A., Pacini, E., Mateos, C., & García Garino, C. (2021). Reinforcement learning-based application autoscaling in the cloud: A survey. Engineering Applications of Artificial Intelligence, 102, Article 104288. https://doi.org/10.1016/j.engappai.2021.104288

14. Dogani, J., Namvar, R., & Khunjush, F. (2023). Auto-scaling techniques in container-based cloud and edge/fog computing: Taxonomy and survey. Computer Communications, 209, 120–150. https://doi.org/10.1016/j.comcom.2023.06.010

15. Alharthi, S., Alshamsi, A., Alseiari, A., & Alwarafy, A. (2024). Auto-scaling techniques in cloud computing: Issues and research directions. Sensors, 24(17), Article 5551. https://doi.org/10.3390/s24175551

16. Jeong, B., & Jeong, Y.-S. (2025). Autoscaling techniques in cloud-native computing: A comprehensive survey. Computer Science Review, 58, Article 100791. https://doi.org/10.1016/j.cosrev.2025.100791

17. Guruge, P. B., & Priyadarshana, Y. H. P. P. (2025). Time series forecasting-based Kubernetes autoscaling using Facebook Prophet and Long Short-Term Memory. Frontiers in Computer Science, 7. https://doi.org/10.3389/fcomp.2025.1509165

18. Smyk, D., & Burak, N. (2025). Metody ta zasoby obrobky danykh v suchasnykh avtomatyzovanykh systemakh [Methods and means of data processing in modern automated systems]. Visnyk Lvivskoho derzhavnoho universytetu bezpeky zhyttiediialnosti, 31, 41–49. doi.org/10.32447/20784643.31.2025.05 [in Ukrainian].

19. Luong, D.-H., Thieu, H.-T., Outtagarts, A., & Ghamri-Doudane, Y. (2018). Predictive autoscaling orchestration for cloud-native telecom microservices. In 2018 IEEE 5G World Forum (5GWF) (pp. 153–158). https://doi.org/10.1109/ACCESS.2020.3025032

20. Ali-Eldin, A., Ilyushkin, A., Ghit, B., Herbst, N. R., Papadopoulos, A., & Iosup, A. (2016). Which cloud auto-scaler should I use for my application? Benchmarking auto-scaling algorithms. In Proceedings of the 2016 ACM/SPEC International Conference on Performance Engineering (pp. 131–132)

21. Agarwal, P., & Lakshmi, J. (2019). Cost aware resource sizing and scaling of microservices. In Proceedings of the 2019 International Conference on Cloud Computing and Internet of Things (pp. 66–74). https://doi.org/10.1145/3361821.3361823

22. Choi, B., Park, J., Lee, C., & Han, D. (2021). pHPA: A proactive autoscaling framework for microservice chain. In Proceedings of the 5th Asia-Pacific Workshop on Networking (pp. 65–71). https://doi.org/10.1145/3469393.3469401

23. Sheganaku, G., Schulte, S., Waibel, P., & Weber, I. (2023). Cost-efficient auto-scaling of container-based elastic processes. Future Generation Computer Systems, 138, 296–312. https://doi.org/10.1016/j.future.2022.09.001

24. Denaro, G., El Moussa, N., Heydarov, R., Lomio, F., Pezzè, M., & Qiu, K. (2024). Predicting failures of autoscaling distributed applications. Proceedings of the ACM on Software Engineering, 1(FSE), 1960–1981

25. Sachidananda, V., & Sivaraman, A. (2024). Erlang: Application-aware autoscaling for cloud microservices. In Proceedings of the Nineteenth European Conference on Computer Systems (pp. 888–923). https://doi.org/10.1145/3627703.3650084

26. Bun, R., Marland, G., Oda, T., See, L., Puliafito, E., Nahorski, Z., Jonas, M., Kovalyshyn, V., Ialongo, I., Yashchun, O., & Romanchuk, Z. (2024). Tracking unaccounted greenhouse gas emissions due to the war in Ukraine since 2022. Science of the Total Environment, 914. https://doi.org/10.1016/j.scitotenv.2024.169879

27. Tatsiy, R. M., Pazen, O. Yu., Vovk, S. Ya., Ropyak, L. Ya., & Pryhorovska, T. O. (2019). Numerical study on heat transfer in multilayered structures of main geometric forms made of different materials. Journal of the Serbian Society for Computational Mechanics, 13(2), 36–55. https://doi.org/10.24874/jsscm.2019.13.02.04

28. Hulida, E., Pasnak, I., Koval, O., & Tryhuba, A. (2019). Determination of the critical time of fire in the building and ensure successful evacuation of people. Periodica Polytechnica Civil Engineering, 63(1), 308–316. https://doi.org/10.3311/PPci.12760

Оглядова стаття.
Надійшла до редакції 20.03.2026.
Прийнята до друку 29.04.2026.
Опублікована 25.05.2026.