ВІДГУК

офіційного опонента – кандидата педагогічних наук, доцента,
завідувача кафедри комп'ютерних систем, мереж та кібербезпеки
факультету інформаційних технологій
Національного університету біоресурсів і природокористування України
КАСАТКІНА Дмитра Юрійовича
на дисертаційне дослідження **КИРИЧЕНКА Олександра Олексійовича**
на тему: «Оптимізація безсерверних обчислень у хмарних середовищах»,
подану на здобуття наукового ступеня доктора філософії
за спеціальністю 121 – Інженерія програмного забезпечення
галузі знань 12 – Інформаційні технології

Актуальність дисертаційного дослідження.

Сучасний етап розвитку інформаційних технологій характеризується міграцією обчислювальних робочих навантажень у хмарні середовища, серед яких безсерверні обчислення (serverless computing) посідають окреме місце. Парадигма Function-as-a-Service дозволяє розробникам зосередитися на бізнес-логіці застосунків, делегуючи постачальнику хмарних послуг завдання забезпечення масштабованості, відмовостійкості та керування інфраструктурою. Водночас широке впровадження безсерверних архітектур виявило низку фундаментальних обмежень, серед яких основними є проблема холодного старту, реактивне автомасштабування та невідповідність класичних аналітичних моделей фактичній стохастичній природі вхідних потоків подій.

Загальноприйняті підходи до автоматичного масштабування ресурсів у безсерверних системах базуються переважно на евристичних порогових правилах або на застосуванні методів машинного навчання, зокрема рекурентних нейронних мереж. Ці підходи демонструють певні переваги, але вимагають значних обсягів історичних даних для навчання,

потребують періодичного перенавчання при зміні характеру навантаження та не забезпечують аналітичних гарантій якості обслуговування у вигляді явних формул, придатних для перевірки відповідності вимогам Service Level Agreement.

За цих умов розробка математичних моделей безсерверних обчислювальних систем як неоднорідних систем масового обслуговування зі змішаними режимами роботи, які враховують гетерогенність джерел подій, нестаціонарність вхідних потоків та змінну інтенсивність обслуговування, є актуальною науково-прикладною задачею. Її розв'язання має створити теоретичне підґрунтя для побудови проактивних алгоритмів автомасштабування, що поєднують аналітичні гарантії продуктивності з можливостями динамічної адаптації до зміни навантаження у реальному хмарному середовищі.

Викладене вище зумовлює актуальність теми дисертаційної роботи Кириченка О.О., яка присвячена розв'язанню науково-прикладної задачі оптимізації безсерверних обчислень у хмарних середовищах на основі побудови математичних моделей теорії систем масового обслуговування неоднорідної структури та розробки відповідної інформаційної технології з проактивним автомасштабуванням обчислювальних ресурсів.

Зв'язок роботи з науковими програмами, планами, темами.

Дисертаційне дослідження виконано відповідно до планів науково-дослідних робіт Чернівецького національного університету імені Юрія Федьковича, а саме:

– кафедри програмного забезпечення комп'ютерних систем за держбюджетною тематикою «Дослідження, моделювання та розробка програмного забезпечення складних динамічних систем» (державний реєстраційний номер 0121U109232;

– кафедри математичних проблем управління і кібернетики за держбюджетною тематикою «Інформаційні технології в аспекті сучасних

задач прийняття рішень» (державний реєстраційний номер 0121U109159).

Наукова новизна одержаних результатів.

До основних наукових результатів дисертаційної роботи слід віднести:

– вперше доведено граничні еволюції для процесу довжини черги у схемі усереднення та схемі дифузійної апроксимації для побудованої неоднорідної системи масового обслуговування $Hk(t)/M/\infty$ зі змішаними режимами роботи з використанням апарату напівмарковських випадкових еволюцій, що, на відміну від відомих результатів для стаціонарних СМО, дозволяє враховувати нестаціонарність та неоднорідність вхідного процесу та отримати нормальне наближення для розподілу довжини черги;

– набула подальшого розвитоку математична модель безсерверних обчислень (на прикладі AWS Lambda), яку представлено як неоднорідну систему масового обслуговування зі змішаними режимами функціонування. На відміну від традиційного використання MMPP-моделей, запропонований підхід базується на трактуванні вхідного потоку як суміші незалежних неоднорідних пуассонівських процесів від різних джерел подій. Це дозволило з вищою точністю описати дисперсію навантаження та забезпечити достовірнішу оцінку ризиків перевищення лімітів одночасного виконання (concurrency limits);

– удосконалено підхід до оцінювання частки відхилених завдань у безсерверних системах з обмеженою чергою. Завдяки застосуванню нормального наближення вдалося узагальнити класичні результати для $M/M/\infty$ СМО з марковською модуляцією на випадок неоднорідних вхідних потоків зі змішаними режимами роботи в неперервному часі. Запропоноване рішення дозволяє підтверджувати відповідність системи критеріям SLA за допомогою прямих аналітичних розрахунків;

– 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 і дозволяє отримати явні формули для оптимальних параметрів provisioned concurrency без потреби у великих обсягах історичних даних;

– розроблено архітектуру фреймворку для розподіленої обробки даних у середовищі AWS, що базується на принципах безсерверних обчислень. Основною відмінністю від традиційних рішень із реактивним масштабуванням є впровадження модуля прогнозного керування ресурсами на основі аналітичної моделі. Використання систем черг повідомлень дозволило досягти слабкої зв'язності (decoupling) компонентів, забезпечуючи високу стійкість та гнучкість системи при нерівномірному навантаженні;

– Реалізовано інформаційну технологію для розподіленої обробки даних з використанням безсерверних обчислень та проактивним автоматичним масштабуванням обчислювальних ресурсів. Прогнозне автомасштабування на основі аналітичної моделі забезпечило прискорення обробки даних на 25,8%, зростання пропускної здатності на 21,3% та зменшення холодних стартів до 3% порівняно з класичним реактивним масштабуванням (1 025 132 записи).

Аналіз основного змісту дисертації.

Науковий рівень викладення дисертації відповідає вимогам МОН України. Назва дисертації адекватно і в повній мірі відображає її зміст. Дисертаційна робота складається з переліку умовних позначень, вступу, чотирьох розділів із підрозділами, висновків, списку використаних джерел та додатків.

У вступі обґрунтовано актуальність теми дисертації, визначено мету, об'єкт і предмет дослідження, сформульовано основні завдання, відображено наукову новизну та практичне значення одержаних результатів.

У першому розділі проведено теоретичне дослідження парадигми безсерверних обчислень. Розглянуто ключові концепції моделей Function-as-a-Service та Backend-as-a-Service, проаналізовано архітектурні особливості подієво-орієнтованих систем та механізми автоматичного масштабування на провідних хмарних платформах AWS Lambda, Google Cloud Functions та Azure Functions. Систематизовано основні переваги та обмеження безсерверних обчислень з акцентом на проблему холодного старту. Розглянуто теоретичні моделі для аналізу ефективності та сформовано перелік відкритих проблем, що потребують подальших досліджень.

У другому розділі проведено аналіз архітектурних рішень для безсерверних обчислень та розроблено архітектуру фреймворку для розподіленої обробки даних. Обґрунтовано вибір системи черг повідомлень як базового компонента архітектури для асинхронної обробки даних, що забезпечує буферизацію та згладжування пікових навантажень. Проведено порівняльний аналіз шаблонів комунікації між компонентами, визначено критерії вибору архітектурних рішень, проаналізовано обмеження реактивного масштабування.

У третьому розділі побудовано математичну модель безсерверної обчислювальної системи як неоднорідної системи масового обслуговування $Hk(t)/M/\infty$ зі змішаними режимами роботи. Досліджено властивості моделі суміші потоків завдань, доведено необхідну та достатню умову обмеженості черги, отримано граничні теореми для процесу довжини черги у схемах усереднення та дифузійної апроксимації з використанням апарату напівмарковських випадкових еволюцій. Розроблено метод оцінки параметрів суміші вхідного процесу на основі EM-алгоритму з використанням метрик AWS CloudWatch. Запропоновано алгоритм параметричної оцінки та оптимізації конфігурації безсерверної системи на основі аналітичних оцінок.

У четвертому розділі здійснено практичну верифікацію теоретичних результатів та реалізовано програмне забезпечення інформаційної технології оптимізації безсерверних обчислень на хмарній платформі AWS. У трьох серіях експериментів досліджено: прогнозне автомасштабування на основі нейронної мережі DeepAR (зменшення кількості холодних стартів на 27%, необроблених запитів на 14%); прогнозне автомасштабування на основі аналітичної моделі (прискорення обробки на 25,8%, зменшення холодних стартів до 3%); імітаційне моделювання методом Монте-Карло (4 конфігурації, 100 прогонів), що підтвердило перевагу моделі суміші потоків над MMPP-наближенням та оптимального протоколу вибору сервера над випадковим.

У висновках подано отримані основні наукові та практичні результати дослідження.

Ступінь обґрунтованості наукових положень, висновків і рекомендацій, їх достовірність.

Сформульовані у дисертації наукові положення, висновки та рекомендації є аргументованими і підкріплені практичною реалізацією. Наукова обґрунтованість положень і висновків забезпечена детальним аналізом сучасних літературних джерел, чітким формулюванням завдань дослідження та коректним застосуванням математичного апарату теорії масового обслуговування, теорії випадкових процесів, теорії напівмарковських еволюцій, стохастичного аналізу та методів математичної статистики.

Достовірність одержаних результатів підтверджена строгим математичним доведенням ключових теоретичних положень, зокрема Леми 3.2.1 щодо умов обмеженості черги та Теореми 3.2.3 про граничні еволюції процесу довжини черги, узгодженістю результатів імітаційного моделювання методом Монте-Карло на 100 прогонах для 4 конфігурацій з аналітичними оцінками, апробацією на реальних наборах даних

безсерверних систем (1082178 подій за 90 хвилин у першому експерименті, 1025132 записи у другому), а також практичним впровадженням розроблених моделей та алгоритмів. Достовірність результатів додатково підтверджується їх апробацією на семи міжнародних та всеукраїнських наукових конференціях.

Практичне значення результатів роботи.

Практичне значення одержаних результатів полягає у реалізації інформаційної технології оптимізації безсерверних обчислень та програмного забезпечення, що забезпечує прогнозування навантаження та проактивне масштабування ресурсів на хмарній платформі AWS. Експериментальна апробація на реальних даних продемонструвала, що аналітична модель на основі 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.

Розроблений фреймворк використовує AWS SQS як чергу повідомлень, AWS Lambda для подієво-орієнтованої обробки, DynamoDB для збереження стану та AppSync для моніторингу в реальному часі. Запропоновані моделі та алгоритми надають готовий аналітичний інструментарій для оптимізації конфігурації безсерверних систем без потреби у великих обсягах історичних даних, що особливо актуально для нових систем або систем зі змінним характером навантаження.

Теоретичні та практичні результати дисертаційного дослідження впроваджено у роботу компаній Finker Finance B.V. та ФОП Вербицької С.І., а також в освітньому процесі відділу комп'ютерних технологій Навчально-наукового інституту фізико-технічних та комп'ютерних наук Чернівецького національного університету імені Юрія Федьковича при викладанні дисциплін «Безсерверні обчислення у хмарних середовищах» та «Проектування інформаційних систем».

Оформлення дисертації, дотримання вимог академічної доброчесності та повнота викладу наукових положень і результатів в опублікованих працях

Дисертаційна робота має логічну структуру. Загальний обсяг роботи становить 245 сторінок, з них основного тексту – 142 сторінки, література – 26 сторінок (194 позиції), додатки – 47 сторінок.

У дисертації не виявлено текстових запозичень і використання наукових результатів інших науковців без посилань на відповідні джерела. Робота відповідає принципам академічної доброчесності.

За результатами досліджень опубліковано 11 наукових праць, з них 4 статті у рецензованих виданнях (2 – у журналах, що індексуються у наукометричній базі Scopus, 2 – в українських фахових виданнях), у матеріалах міжнародних наукових конференцій – 7 робіт (одна з яких індексована у Scopus).

Мова та стиль дисертаційної роботи.

Текст дисертаційної роботи викладено у логічній послідовності з використанням сучасної наукової термінології. Дисертація містить достатню кількість ілюстративного матеріалу – схем, рисунків, графіків та таблиць, що сприяє кращому розумінню результатів дослідження. Мова викладу, стиль та оформлення роботи в цілому відповідають установленим вимогам до наукових праць.

Зауваження та дискусійні положення щодо змісту дисертаційного дослідження

Незважаючи на загалом високу наукову якість дисертаційної роботи, низку її положень доцільно піддати дискусійному обговоренню.

1. У першому розділі дисертаційного дослідження проаналізовано переважно англомовні джерела щодо безсерверних обчислень, проте недостатньо повно охоплено наукові праці українських дослідників у галузі хмарних обчислень та систем масового обслуговування з

неоднорідним вхідним потоком, які могли б додатково обґрунтувати теоретичні засади запропонованих моделей.

2. У роботі не охоплено клас моделей систем масового обслуговування з повторними викликами (retrial queues). У безсерверних архітектурах механізм повторних спроб після throttling-помилок через DLQ концептуально близький до retrial-механізмів, що могло б бути розглянуто як додатковий теоретичний інструмент аналізу.

3. Запропонований фреймворк (підрозділ 2.4) реалізовано виключно на базі сервісів AWS (Lambda, SQS, DynamoDB, AppSync, CloudWatch). У роботі недостатньо повно проаналізовано особливості перенесення фреймворку на інші провідні хмарні платформи: зокрема, Google Cloud Functions використовує Pub/Sub з власною моделлю гарантії доставки, а Azure Functions – Event Grid з відмінною семантикою тригерів та моделлю біллінгу. Це обмежує універсальність декларованого фреймворку та ускладнює його порівняння з альтернативними рішеннями.

4. У моделі (формула 4.5) та таблиці 4.1 враховано вартісні характеристики виконання, проте у експериментальній частині відсутня кількісна оцінка економічної ефективності запропонованого проактивного автомасштабування у безпосередньому порівнянні зі стратегією provisioned concurrency. Зокрема, не наведено розрахунку економії на типовому місячному робочому навантаженні, що було б важливим аргументом для практичного впровадження фреймворку.

5. В алгоритмі налаштування параметрів СМО (підрозділ 4.3.2) використано тест Колмогорова–Смірнова на ковзних інтервалах $\Delta = 60$ с з прив'язкою до інтервалу моніторингу AWS CloudWatch. Однак у роботі недостатньо обґрунтовано саме такий розмір вікна: чи аналізувалася чутливість алгоритму при $\Delta = 30, 90, 300$ с (при детальному CloudWatch-моніторингу), та як вибір розміру вікна впливає на швидкість виявлення зміни режиму навантаження порівняно з ризиком хибного спрацювання

при сезонних коливаннях.

Зазначені зауваження мають дискусійний характер, не є принциповими, не впливають на загальний зміст дисертаційної роботи та не знижують її наукову новизну і практичну цінність.

Висновки щодо дисертації в цілому.

На основі викладеного вище вважаю, що дисертаційна робота Кириченка Олександра Олексійовича на тему «Оптимізація безсерверних обчислень у хмарних середовищах», яка подана на здобуття наукового ступеня доктора філософії, за її актуальністю, науковим рівнем, практичною цінністю, новизною розв'язання поставлених завдань, змістом та оформленням повністю відповідає вимогам пунктів 6, 7, 8, 9 «Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», затвердженого постановою Кабінету Міністрів України від 12 січня 2022 р. № 44 (зі змінами, внесеними згідно з Постановою Кабінету Міністрів України № 341 від 21.03.2022 р., № 502 від 19.05.2023 р , № 507 від 03.05.2024 р.), а її автор, Кириченко Олександр Олексійович, заслуговує на присудження йому наукового ступеня доктора філософії за спеціальністю 121 – «Інженерія програмного забезпечення» галузі знань 12 – «Інформаційні технології».

Офіційний опонент –
кандидат педагогічних наук, доцент,
завідувач кафедри комп'ютерних систем,
мереж та кібербезпеки
факультету інформаційних технологій
Національного університету біоресурсів
і природокористування України Дмитро КАСАТКІН

Підпис Касаткіна Д.Ю. засвідчую
секретар НУБіП України
Олеф /Барановської О.Д/

27.05.2026 р.