МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ "ЛЬВІВСЬКА ПОЛІТЕХНІКА"

ПЛЕСКАНКА МАР'ЯНА ВІКТОРІВНА

УДК 621.391

ПІДВИЩЕННЯ ЯКОСТІ ОБСЛУГОВУВАННЯ У МЕРЕЖІ ДОСТАВКИ КОНТЕНТУ

05.12.02 – телекомунікаційні системи та мережі

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Львів - 2025

Дисертацією є рукопис.

Робота виконана у Національному університеті «Львівська політехніка» Міністерства освіти і науки України

Науковий керівник - доктор технічних наук, професор
Кирик Мар'ян Іванович,
Національний університет «Львівська політехніка»,
професор кафедри інформаційно-комунікаційних технологій.

Офіційні опоненти - доктор технічних наук, професор
Толюпа Сергій Васильович,
Київський національний університет
імені Тараса Шевченка,
професор кафедри кібербезпеки та захисту інформації;

доктор технічних наук, професор
Жураковський Богдан Юрійович,
Національний технічний університет України
«Київський політехнічний інститут
імені Ігоря Сікорського»
професор кафедри інформаційних систем та технологій.

Захист відбудеться "16" травня 2025 р. о 11^{00} год. на засіданні спеціалізованої вченої ради Д 35.052.10 у Національному університеті "Львівська політехніка" (79013, Львів, вул. С. Бандери, 12, ауд. 226 головного навчального корпусу).

З дисертацією можна ознайомитись у науковій бібліотеці Національного університету "Львівська політехніка" (79013, м. Львів, вул. Професорська, 1).

Автореферат розісланий "11" квітня 2025 р.

Вчений секретар спеціалізованої
вченої ради, д.т.н., доцент М.І. Бешлей

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Сучасний рівень розвитку інформаційних технологій характеризується інтенсивним впровадженням нових послуг та платформ. Досить поширеними в наш час стали Cloud-мережі та технологія CDN (Content Delivery Network). Все це призводить до підвищення вимог до каналів зв'язку та обслуговуючих пристроїв, які виконують обробку трафіку, щоб забезпечити необхідну якість послуг QoS (Quality of Service), оскільки відповідно до вимог часу, сучасні мережі передачі даних повинні будуватися як високонадійні системи, здатні забезпечити необхідні показники якості обслуговування.

Умови, які склались в сучасному світі, а це для прикладу COVID, війна та багато інших чинників диктують свої правила розвитку технологій та способів надання послуг. Для забезпечення різних вимог параметрів QoS мультисервісних послуг в системах передавання даних необхідно впроваджувати алгоритми та методи управління трафіком, які повинні враховувати особливості різних видів послуг, а також забезпечувати ефективне використання ресурсів мережі та вузлів обслуговування.

Окрему увагу слід приділити і новим архітектурним рішенням, які використовуються в наш час для розробки сучасних сервісів та додатків. Особливий акцент робиться на підходи, які дозволяють швидко наростити ресурси в моменти зростання навантаження, та так само швидко все згорнути при його відсутності. Не менш важливими є також методи, які дозволяють переносити обробку даних якнайближче до користувача. Якщо раніше такі підходи застосовувалися переважно для статичного контенту, такого як відео та аудіо трафік чи різного роду зображення, то зараз цього вже недостатньо. У сучасних CDN значну частку трафіку становить динамічний контент, вміст якого може змінюватися залежно від часу, місця, дій користувача та інших факторів.

Варто зазначити, що досить багато досліджень було зосереджено саме на покращення якості доставки як статичних так і динамічних даних в CDN мережах. Серед провідних наукових досліджень в даному напрямку, можна виділити авторів: Nakanishi K., Suzuki F., Dong Y., Gupta P., Goya M., Mijumbi R., Shitole A., Lugones D., Л. Глоба, О. Лемешко, С. Толюпа, В. Безрук, Б. Жураковський. У більшості розглянутих робіт автори досліджують методи балансування навантаження та ефективності кешування, проте недостатньо уваги приділено критеріям, за якими можна розподіляти типи запитів між граничними серверами, а також аналітичним даним, що прогнозують динаміку поведінки користувачів. Щодо обробки динамічних даних, не було враховано можливості перенесення самого процесу обробки та формування даних якомога ближче до локації користувача з метою ефективного використання обчислювальної інфраструктури.

Таким чином, зростання обсягів та різноманітності статичних і динамічних даних у мережах доставки контенту обумовлює необхідність розв’язання науково-практичного завдання підвищення якості обслуговування користувачів в умовах обмеженої обчислювальної інфраструктури шляхом розроблення методів і алгоритмів обробки запитів, що потребують значних обчислювальних ресурсів, кешування даних, реалізації балансування навантаження та адаптивного розгортання мікросервісів у точках присутності CDN мереж.

Зв’язок роботи з науковими програмами, планами, темами. Тематика дисертаційного дослідження виконувалась у відповідності до наукового напряму кафедри інформаційно-комунікаційних технологій (кафедри телекомунікацій) Національного університету «Львівська політехніка» - «Інфокомунікаційні системи та мережі», в межах низки держбюджетних науково-дослідних робіт: «Методи побудови та моделі інформаційно-телекомунікаційної інфраструктури на основі SDN-технологій для систем електронного урядування» (ДБ/SDN), (№ держреєстрації 0115U000444, (2015-2016 рр.), «Методи побудови гетерогенних інформаційно-комунікаційних систем для розгортання програмно-конфігурованих мереж 5G подвійного використання» (№ держреєстрації 0117U004449, (2017–2018 рр.)).

Мета і завдання досліджень. Метою дисертаційної роботи є підвищення якості обслуговування у мережах доставки контенту шляхом розробки методів та алгоритмів ефективного кешування, балансування трафіку та адаптивного розгортання мікросервісів у точках присутності CDN мереж.

Для досягнення поставленої мети необхідно розв’язати такі завдання:

1. Провести аналіз існуючих підходів та архітектурних рішень, які застосовуються під час розробки та впровадження нових сервісів та послуг в CDN мережах.
2. Удосконалити метод обробки запитів, що потребують значних обчислювальних ресурсів у точках присутності CDN мережі.
3. Запропонувати інтегральний ключ кешування для ефективного кешування статичних даних в мережах доставки контенту.
4. Удосконалити метод балансування навантаження у мережах доставки контенту для забезпечення ефективного використання кеш-пам'яті та обчислювальних ресурсів граничних серверів та мережевих пристроїв.
5. Розробити метод адаптивного розгортання мікросервісів для обробки динамічних даних у режимі реального часу в точках присутності CDN-мереж.
6. Провести експериментальне дослідження для оцінювання ефективності запропонованих рішень на основі розробленого прототипу мережі доставки контенту.

Об’єкт дослідження – процеси обробки даних в точках присутності CDN мережі.

Предмет дослідження – методи та алгоритми обробки статичних та динамічних даних в мережах доставки контенту.

Методи дослідження. Дослідження виконано на основі використання методів об’єктно-орієнтованого проєктування, теорії систем масового обслуговування, паралельних та розподілених обчислень, математичного та комп’ютерного моделювання, а також на основі результатів експериментів.

Наукова новизна отриманих результатів.

1. Набув подальшого розвитку метод обробки запитів, що потребують значних обчислювальних ресурсів, який, на відміну від раніше відомих, для розподіленої обробки запитів враховує дані аналітики щодо популярності певних ресурсів (веб-сторінок, мультимедійного контенту) серед списку найбільш запитуваних, що дало змогу забезпечити ефективне використання кешованих

даних, а також зменшити навантаження на кореневий сервер та час відповіді на запити кінцевого користувача.

2. Удосконалено метод балансування трафіку у точках присутності CDN мережі, який, на відміну від відомих, враховує значення інтегрального ключа кешування сформованого на основі розробленого методу обробки запитів, локацію клієнта, наявність контенту на граничному сервері та стан функціонування доступних серверів, що дало змогу підвищити якість обслуговування в мережах доставки контенту.

3. Вперше запропоновано метод адаптивного розгортання мікросервісів для обробки динамічних даних у режимі реального часу в точках присутності CDN-мереж, який, на відміну від відомих, частково дублює бізнес-логіку сервісу з кореневого сервера, що надається кінцевим користувачам, здійснює в режимі реального часу аналіз параметрів, які визначають якість послуги, та адаптивно розподіляє запити на основі оцінювання рівня завантаженості граничних серверів для забезпечення необхідної якості обслуговування.

Практичне значення одержаних результатів полягає у можливості їх безпосереднього застосування в існуючих мережах доставки контенту та провайдерів хмарних сервісів:

1. Удосконалений метод обробки запитів статичних даних у мережах доставки контенту дав змогу покращити якість обслуговування користувачів шляхом зменшення часу відповіді для кінцевого користувача до 9%, та на 10% завантаженість кореневого сервера.

2. Комплексне використання інтегрального ключа кешування, методу обробки запитів та балансування трафіку у точках присутності CDN мережі дало змогу підвищити коефіцієнт ефективності кешування на 30 %, а також зменшити час відповіді на запити кінцевого користувача на 40%.

3. Розроблений прототип мережі доставки контенту для передавання статичних та динамічних даних дав змогу підтвердити на практиці ефективність розробленого методу адаптивного розгортання мікросервісів для обробки динамічних даних у режимі реального часу в точках присутності CDN-мереж, а саме забезпечити необхідну якість обслуговування в умовах обмежених ресурсів кореневого сервера. Результати експериментального дослідження, проведеного для оцінки якості надання певного типу сервісу кінцевому користувачеві за умов високого навантаження системи, показали, що час відповіді на запити користувача становить 25 мс, що на 7% швидше порівняно з часом відповіді від кореневого сервера для клієнтів у тій самій локації.

Основні результати дисертаційної роботи використано і впроваджено в телекомунікаційних корпоративних мережах ТОВ “Телекомунікаційна компанія”, ТОВ ВТФ “Контех”, ТОВ “МаксіТех”, що підтверджено актами впровадження, а також у навчальному процесі кафедри інформаційно-комунікаційних технологій Національного університету «Львівська політехніка».

Особистий внесок здобувача. Усі результати наукових, теоретичних і практичних досліджень, викладені в дисертації, автор одержав особисто. У працях, опублікованих у співавторстві, дисертантові належать: у роботах [1,3,9] – представлено новий механізм розгортання інфраструктури, створення

мікросервісів для майбутніх хмарних мереж, процес міграції монолітної програми, [2,10-12] – запропоновано метод обробки запитів, що вимагають значних обчислювальних ресурсів, що враховує дані аналітики та інтегральний ключ кешування для забезпечення високої ефективності кешування даних та використання ресурсів мережі доставки; розроблено метод адаптивного створення мікросервісу у граничній локації CDN мережі, який призначений для забезпечення необхідної якості обслуговування, [4,13,14] – запропоновано використання нового методу оптимізованого кешування даних як складової площини балансування навантаження, [5,15-17] – представлені основні методи та технології розповсюдження та доставки даних, виходячи з цільових функцій, [6] – розглянуто методи спектральної мобільності для когнітивного радіо, що надають змогу когнітивним користувачам перемикатися в частотні канали, [7] – запропоновано ряд моделей та методів для конвергенції гетерогенних мереж мобільного зв'язку п'ятого покоління з використанням технології D2D, [8,18] – представлено технічні та архітектурні підходи щодо підвищення енергоефективності телекомунікаційних мереж, [19] – представлено генетичний алгоритм для визначення маршруту передачі даних у мережах із змінною структурою.

Апробація результатів дисертації. Основні результати наукових досліджень доповідалися та обговорювались на всеукраїнських та міжнародних науково-технічних конференціях: Міжнародних науково-технічних конференціях «Сучасні проблеми радіоелектроніки, телекомунікацій, комп'ютерної інженерії» (м. Львів-Славське 2016, 2020, 2022 рр.); Problems of infocommunications science and technology, PIC S and T 2018: Proceedings of 5th International scientific-practical conference, Kharkiv, Ukraine, 9–12 October 2018.; Proceedings of the Second International Conference on Advanced Information and Communication Technologies (AICT'2017), Lviv, Ukraine.:, Міжнародних науково-технічних конференціях «Досвід розробки та застосування приладо-технологічних САПР в мікроелектроніці» (Поляна-Свалява, 2017, 2019 рр.); 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET, Львів 2024).

Публікації. За результатами досліджень, які викладені у дисертаційній роботі, опубліковано 19 наукових праць, серед них 1 стаття у закордонному виданні, що входить до наукометричних баз даних Scopus [1], 7 статей у наукових фахових виданнях згідно з переліком МОН України [2-8] та 11 публікацій у збірниках праць міжнародних і всеукраїнських конференцій [9-19].

Структура та обсяг роботи. Робота складається з переліку умовних скорочень, вступу 4 розділів, висновків, списку використаних джерел і 3 додатків. Загальний обсяг роботи складає 199 сторінок друкарського тексту, із них 7 сторінок вступу, 160 сторінок основного тексту, 82 рисунки, 3 таблиці, список використаних джерел із 114 найменувань, 3 додатки на 13 сторінках.

ОСНОВНИЙ ЗМІСТ ДИСЕРТАЦІЙНОЇ РОБОТИ

У вступі обґрунтовано актуальність теми дисертаційної роботи, констатовано зв'язок роботи з науковими програмами, темами, сформульовано мету і завдання дослідження, наукову новизну та практичне значення отриманих результатів.

Наведено дані про впровадження результатів роботи, її апробацію, публікації та особистий внесок здобувача.

У першому розділі **«Аналіз методів міграції сервісів від монолітної до мікросервісної архітектури»** проведено дослідження принципів та методів переходу від монолітної архітектури до мікросервісної. Визначено основні фактори які визначають основу функціонування мікросервісного застосунку.

Для розгортання нового сервісу запропоновано декілька підходів, кожен з яких має свої особливості. Використовуючи процес автоматичного розгортання, ми зводимо до мінімуму людський фактор та робимо сервіс більш стабільним та надійним. Розглянуто сучасні методи щодо створення та управління інфраструктурою не шляхом фізичного налаштування обладнання чи використання інтерактивних інструментів налаштування, а за допомогою файлів конфігурації та підходів «Інфраструктура як код» (IaC, Infrastructure as Code). Такі методи створення та керування виробничими середовищами дають можливість створювати, змінювати та адмініструвати свою інфраструктуру безпечним, узгодженим і повторюваним способом, визначаючи конфігурації ресурсів, та застосовуючи при цьому методи контролю версій.

Проведено аналіз наукових досліджень, присвячених динамічному балансуванню навантаження та кешуванню даних, з метою узагальнення їхніх переваг, обмежень та впливу на ефективніс

ть мереж доставки контенту й якість наданих послуг. Попри те, що розглянуті підходи демонструють значний потенціал у підвищенні продуктивності та масштабованості, залишається потреба в розробці інтелектуальних методів балансування навантаження, кешування статичних даних, а також перенесення процесу обробки динамічних даних на граничні локації CDN мереж, які здатні відповідати динамічним вимогам сучасних хмарних обчислювальних середовищ.

У другому розділі «Методи і алгоритми балансування та розподіленої обробки запитів в CDN мережі» досліджено основні архітектурні принципи роботи CDN мережі, методи балансування навантаження які використовуються в сучасних мережа, технології, які забезпечують надійність та контроль за якістю доставки даних, а також забезпечують можливість ефективно використовувати ресурси мережі та обслуговуючих пристроїв. У багатьох випадках, коли користувач запитує один статичний файл, запит перенаправляється до граничного сервера, який у випадку наявності файлу в кеші відразу надсилає його клієнту. Однак, часто бувають ситуації, у яких користувач запитує багато даних в межах одного запиту. Найпростішим прикладом такого запиту може бути сторінка інтернет-магазину із списком предметів в певній категорії, або перегляд програми в записі, використовуючи сервіс інтерактивного цифрового телебачення. У варіанті інтернет-магазину, в запиті від клієнта буде міститись велика кількість даних(предметів), на які потрібно отримати інформацію про їхній опис, характеристики, ціни, наявність на складі, рейтинг, відгуки і інша інформація. Якщо такий запит перенаправити на один граничний сервер, то ймовірність того, що всі ці дані будуть міститися в кеші даного сервера дуже мала. Очевидним також є факт, що чим більша кількість предметів буде міститись в запиті користувача, тим меншою буде ймовірність того, що всі вони будуть присутні в

кеш пам'яті граничного сервера. А це означає, що збільшиться кількість запитів до кореневого сервера, зросте час відповіді для кінцевого користувача, а ефективність використання кешованих даних відповідно зменшується.

З метою покращення якості сервісу, ефективності кешування та обробки запитів такого типу, запропоновано схему із розділення одного запиту, що потребує значних обчислювальних ресурсів, на велику кількість простих. Для прикладу, якщо користувач запитує інформацію про 10 предметів, Product1, Product2, … Product10, то балансувальник навантаження розділить його на 10 простих запитів по одному предмету. Балансувальник навантаження, при направленні запиту на граничний сервер, буде використовувати алгоритм роботи, який також враховує завантаженість граничного сервера та наявність контенту в його пам'яті. В такому випадку, кількість запитів до граничних серверів зросте, але тривалість їх обробки знизиться, оскільки буде затрачатись менше ресурсів на їх обробку, а також покращиться ефективність використання кешованих даних.

Запропонований метод розподілення запитів можна розвинути ще більше, зробивши інтеграцію із даними аналітики. Зазвичай, на будь якій веб-сторінці збираються метрики про те, що переглядає користувач, які саме категорії, фільтри та інші параметри вибору. Всі ці метрики в подальшому надсилаються для аналітики та прогнозування поведінки користувача, наприклад Google аналітика. Маючи такі дані, можна сформувати інтегральні ключі кешування, базуючись на найбільш популярних сторінках інтернет-магазину, які можуть включати в себе назву категорії, фільтри, кількість предметів на сторінці та навіть рекомендації для користувача. Варто зазначити, що інтегральний ключ кешування використовується і для інших типів мережевого трафіку, для забезпечення ефективного використання закешованих даних. Прикладом формування інтегрованого ключа може бути сервіс реального часу - інтерактивне цифрове телебачення, рисунок 1.

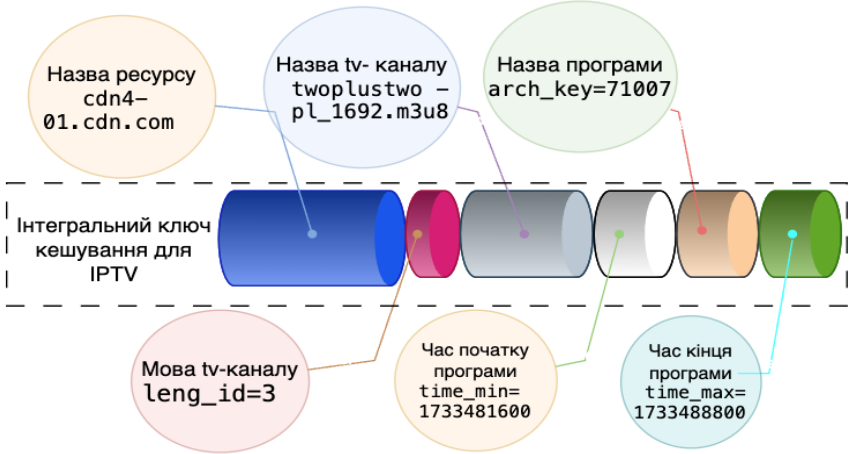

Рис.1. Приклад інтегрованого ключа кешування для сервісу IPTV

Провайдери мультимедійних послуг мають аналітичні дані про перегляд тих чи інших каналів, програм передач на певному каналі, та часових інтервалів, які користуються найбільшою популярністю.

Якщо зробити інтеграцію всіх цих параметрів, то ключ кешування матиме наступний вигляд:

cdn401.cdntvc.com/twoplustwopl_1692.m3u8?ticket=559225136&leng_id=3&arch_key=71007&time_min=1733481600&time_max=1733488800

Інтегральний ключ кешування, можна представити наступним математичним виразом:

$$K = R_n + f(C_n, P_n, L_i, T_s, T_e) \quad (1)$$

$$f(C_n, P_n, L_i, T_s, T_e) = C_n + P_n + L_i + T_s + T_e \quad (2)$$

де R_n – назва ресурсу, до якого звертається користувач, C_n – назва каналу, який переглядає користувач, P_n – назва програми передач, L_i –ідентифікатор мови трансляції програми, T_s – часова мітка початку трансляції програми, T_e – часова мітка завершення трансляції програми.

Значення $f(C_n, P_n, L_i, T_s, T_e)$, може змінюватись та залежить від кількості складових, які враховуються при визначенні інтегрального ключа кешування. Блок схему алгоритму роботи балансувальника із використанням інтегрального ключа кешування та контролем завантаженості серверів представлено на рисунку 2.

Рис.2. Блок схема алгоритму роботи балансувальника навантаження в граничній локації, який забезпечує максимальну ефективність кешування та контроль завантаженості кешуючих серверів

В роботі проведено дослідження, як впливає розподілена обробка даних у поєднанні із інтегральним ключем кешування та даними аналітики на ефективність кешування та час відповіді для кінцевого користувача. Для дослідження було використано сервіс нереального часу, а саме ефективність кешування даних, представлених на веб-сторінках. Результати ефективності кешування представлені на рисунку 3.

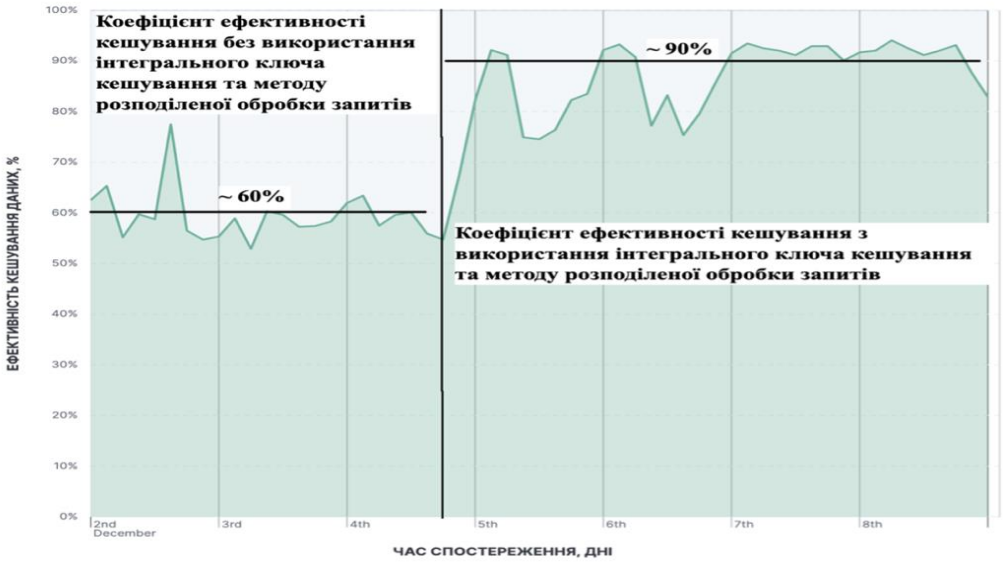

Рис.3. Ефективність кешування даних із використанням інтегрального ключа кешування та методу розподіленої обробки запитів

Час дослідження поділено на два часові інтервали. У першому інтервалі, 2–5 грудня (December), метод розподіленої обробки запитів та інтегральний ключ кешування не використовувався. Балансування навантаження в граничній локації здійснювалось на основі TCP-сесії користувача. Значення ефективності кешування в такому випадку становить 50–60%. На другому часовому інтервалі, 5–8 грудня, було використано аналітичні дані за попередні дні та застосовано метод розподіленої обробки ресурсозатратних запитів і інтегральний ключ кешування на рівні балансувальника навантаження. Як бачимо, ефективність кешування зросла до 90%.

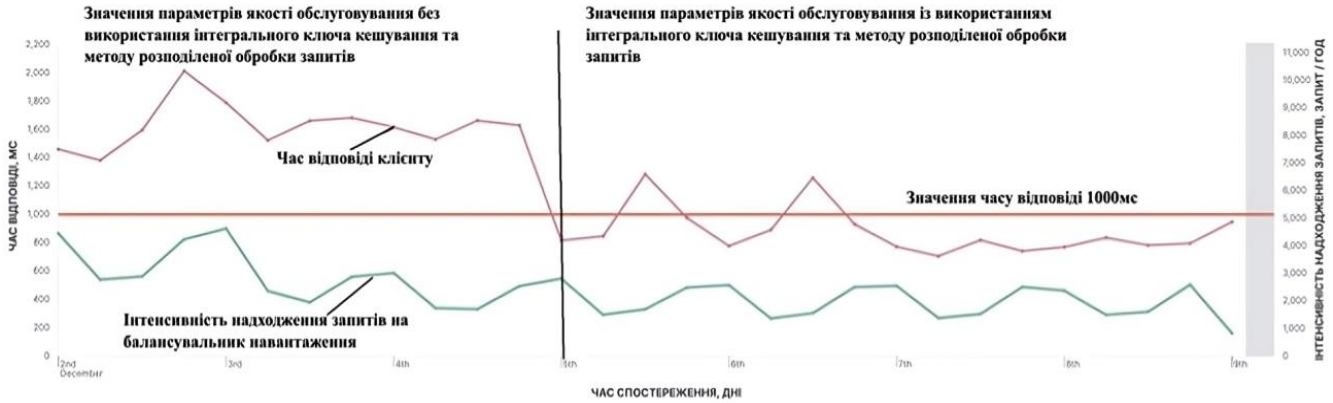

Рис.4. Час відповіді для кінцевого користувача із використанням інтегрального ключа кешування та методу розподіленої обробки запитів

На рисунку 4 показано, як змінювалися час завантаження веб-сторінки та кількість запитів до балансувальника навантаження до і після використання інтегрального ключа кешування та методу розподіленої обробки запитів. Значення часу завантаження сторінки зменшилось із 1600мс до 800мс.

Також варто зауважити, що час завантаження сторінки досить суттєво впливає на рейтинг веб ресурсу в пошукових системах. Якщо на веб ресурсі буде присутня велика кількість сторінок, час завантаження яких буде значним, тоді пошукова

система суттєво понизить рейтинг такого ресурсу. Саме тому, забезпечення мінімального часу завантаження сторінки веб ресурсів є важливим не тільки для кінцевого користувача, а безумовно і для збереження хорошого рейтингу ресурсу в пошукових системах, які використовуються в глобальних мережах.

У третьому розділі **«Модель оцінки ефективності роботи систем обробки мережевого трафіку у точках присутності CDN-мереж»** представлено комплексну модель масового обслуговування з трьома об'єднаними системами масового обслуговування для характеристики процесу обробки запитів у хмарних сервісах із динамічним виділенням ресурсів.

Три системи масового обслуговування складають комплексну модель, яка характеризує неоднорідність ЦОД та сервісні процеси, які задіяні в розподілених хмарних обчисленнях. Схема роботи моделі представлена на рисунку 5.

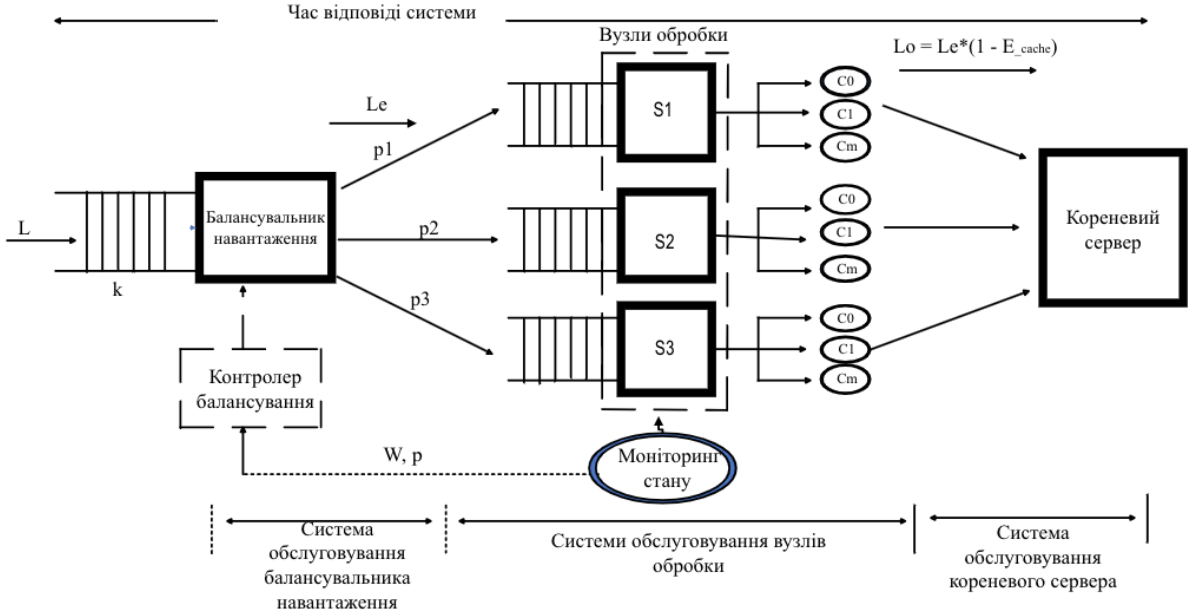

Рис.5. Комплексна модель обслуговування запитів в точках присутності CDN мережі

Перша – це рівень входу в точку присутності сервісу в хмарі, друга – виконання обробки даних на рівні вузла обробки запитів, який є найближче до користувача та має для цього наявні ресурси, третя – система обслуговування на рівні кореневого сервера.

Для ефективного розподілу навантаження, балансувальник повинен отримувати значення показників ефективності та завантаженості з усіх вузлів обробки, а саме кількість завдань в черзі, Lq, середній час очікування завдання в черзі, Wq, і коефіцієнт завантаженості вузла обробки, ρ.

В якості математичної моделі першої підсистеми, а саме балансувальника навантаження, запропоновано систему масового обслуговування класу $M/M/1/k$ – система масового обслуговування з обмеженою довжиною черги k, та 1 обслуговуючим пристроєм. Оскільки кожне вузол обробки, а також кореневий сервер, можуть паралельно обробляти кілька завдань, то розглядаються як система масового обслуговування $M/M/$n. Вхідний потік заявок описується розподілом Пуассона із інтенсивністю надходження λ (часові інтервали між поступленням завдань на обслуговування є незалежними). Оскільки надходження запитів не залежать від стану черги та балансувальник навантаження є системою масового

обслуговування з обмеженою довжиною черги, то завжди існує ймовірність блокування системи розподілу вхідних запитів, яка визначається як pb і буде більше або дорівнює 0 ($pb \geq 0$) і $\lambda \geq \lambda e$, де λ - інтенсивність поступлення запитів на вхід системи балансування, λe - сумарна інтенсивність запитів на виході системи балансування. Основні параметри системи обслуговування визначатимуться наступним чином:

$$\lambda_e = \lambda(1 - \rho_b) \tag{3}$$

$$\lambda_i = \lambda_e p_i \quad 1 \leq i \leq n \quad \Sigma\rho_i = 1 \tag{4}$$

$$\mu_i = \frac{1}{\left(\frac{\underline{I}}{f_i}\right)} = \frac{f_i}{\underline{I}} \tag{5}$$

$$\rho_i = \frac{\lambda_i}{(C_i\mu_i)} = \frac{(\lambda_i \underline{I})}{(C_i f_i)} \tag{6}$$

де f_i – кількість операції, які може опрацювати одне ядро, C_i – кількість ядер на обслуговуючому пристрої, I – сумарна кількість операцій яку може опрацювати обслуговуючий пристрій.

Для розрахунку параметрів балансувальника навантаження вибрано система $M/M/1/k$. Якщо коефіцієнт завантаженості системи рівний:

$$\rho_s = \frac{\lambda}{\mu_s} \tag{7}$$

та $\rho_s < 1$, тоді ймовірність того, що в системі перебуває i задач буде визначатись наступним співвідношенням:

$$\pi_i^{(s)} = \frac{1-\rho_s}{1-\rho_s^{k+1}}\rho_s^i = 1,2,\dots,k \tag{8}$$

У випадку, якщо i > k , нові задачі не можуть надходити в систему і в такому випадку ймовірність блокування буде рівною:

$$P_b = \pi_k^{(s)} = \frac{1-\rho_s}{1-\rho_s^{k+1}}\rho_s^k \tag{9}$$

Враховуючи співвідношення 3 та 9, ефективна інтенсивність обробки на сервері балансування буде визначатись як:

$$\lambda_e = \lambda(1 - p_b) = \lambda\frac{1-\rho_s^k}{1-\rho_s^{k+1}} \tag{10}$$

Наступним параметром є кількість запитів, які знаходяться в балансувальнику навантаження. Ця кількість включає в себе запити які знаходяться в самому балансувальному пристрої та відправляються на обслуговуючі вузли, а також запити які знаходяться в черзі на обслуговування.

Сумарна кількість таких запитів рахується за формулою:

$$\underline{C}_{total}^{(s)} = \underline{C}_{\omega}^{(s)} + \underline{C}_{sch}^{(s)} = \frac{\rho_s}{1-\rho_s} - \frac{(k+1)\rho_s^{k+1}}{1-\rho_s^{k+1}} \tag{11}$$

Враховуючи формулу Літтла та вирази 10 та 11 , час відповіді можна отримати за наступним виразом:

$$\underline{T}^{(s)} = \frac{\underline{C}_{total}^{(s)}}{\lambda_e} = \frac{1-(k+1)\rho_s^k+k\rho_s^{k+1}}{\mu_s-(1-\rho_s)(1-\rho_s^k)} = \frac{1}{\mu_s-\lambda} - \frac{k\lambda^k}{\mu_s^{k+1}-\lambda^k} \tag{12}$$

Ймовірність того, що запити, які приходять від балансувальника навантаження на обслуговуючі пристрої будуть оброблені відразу:

$$q_i = 1 - \sum_{m=C_i}^{\infty}\pi_{i,m}^{(e)} = 1 - \sum_{m=C_i}^{\infty}\frac{C_i^{C_i}}{C_i!}\rho_i^m\pi_{i,0}^{(e)} = \frac{C_i-\rho_{i0}-C_i\pi_{i,C_i}^{(e)}}{C_i-\rho_{i0}} \tag{13}$$

Кількість задач, які знаходяться обслуговуючому пристрої

$$\underline{C}_{total\_i}^{(e)} = \underline{C}_{\omega\_i}^{(e)} + \underline{C}_{exc\_i}^{(e)} = \pi_{i,0}^{(e)} \cdot \frac{\rho_{i0}^{C_{i+1}}}{(C_i-1)!(C_i-\rho_{i0})^2} + \rho_{i0} \quad (14)$$

Враховуючи формулу Літтла та вирази 14 , час відповіді обслуговуючого пристрою можна отримати за наступним виразом:

$$\underline{T}_i^{(e)} = \frac{\underline{C}_{total\_i}^{(e)}}{\lambda_i} = \frac{I}{f_i}\left(1 + \pi_{i,0}^{(e)} \cdot \frac{\rho_{i0}^{C_i}}{(\mathrm{C}_i-1)!(C_i-\rho_{i0})^2}\right) \quad (15)$$

Що стосується кореневого сервера, то інтенсивність надходження запитів, яка буде перенаправлятись на кореневий сервер буде визначатись коефіцієнтом ефективності кешування.

$$\mathrm{E}_{\mathrm{cache}} = \frac{\mathrm{N}_{\mathrm{edge}}}{\mathrm{N}_{\mathrm{total}}}, \quad (16)$$

$\mathrm{E}_{\mathrm{cache}}$ - ефективність кешування, $\mathrm{N}_{\mathrm{edge}}$ - кількість запитів, відповіді на які були сформовані на граничному сервері (без звернення до кореневого сервера), $\mathrm{N}_{\mathrm{total}}$ – загальна кількість отриманих запитів.

Відповідно, використавши вирази 10 та 16, визначимо:

$$\lambda_o = \lambda_e(1-\mathrm{E}_{\mathrm{cache}}) = \lambda(1-p_b)(1-\mathrm{E}_{\mathrm{cache}}) = \lambda\frac{1-\rho_s^k}{1-\rho_s^{k+1}}(1-\mathrm{E}_{\mathrm{cache}}) \quad (17)$$

Враховуючи формулу Літтла та вирази 14 та 17, час відповіді від кореневого сервера можна отримати за наступним виразом:

$$\overline{T}^{(o)} = \frac{\overline{C}_{total\_i}^{(e)}}{\lambda_o} = \frac{\pi_{i,0}^{(e)} \cdot \frac{\rho_{i0}^{C_{i+1}}}{(C_i-1)!(C_i-\rho_{i0})^2} + \rho_{i0}}{\lambda\frac{1-\rho_s^k}{1-\rho_s^{k+1}}(1-\mathrm{E}_{\mathrm{cache}})} \quad (18)$$

Сумарне значення часу затримки для клієнта буде визначатись значенням часу затримки на рівні балансувальника, вузла обслуговування та кореневого сервера.

$$T_{sum} = \overline{T}^{(s)} + \overline{T}_i^{(e)} + \overline{T}^{(o)} \quad (19)$$

В результаті моделювання отримано графічні залежності, які показують, що системи обслуговування даних працюють досить добре при середніх навантаженнях вузлів обслуговування (Рис.6.). При пікових навантаженнях, ефективність роботи системи різко знижується.

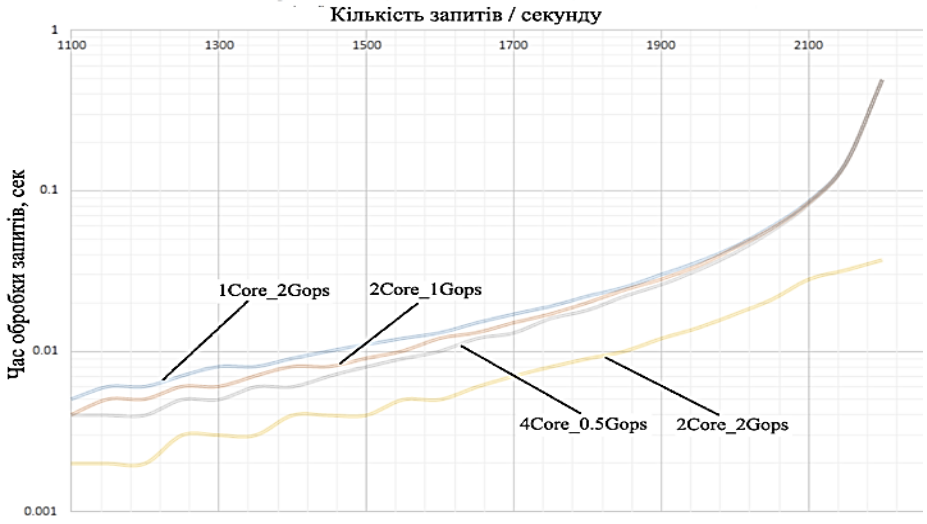

Рис.6. Залежність часу відповіді від інтенсивності поступлення заявок

Розрахунок часу затримки проводився для 4-х варіантів роботи підсистеми вузлів обслуговування:

- 1 вузол обробки даних, продуктивністю 2000 запитів/секунду;
- 2 вузли обробки даних, продуктивністю 1000 запитів/секунду кожен;
- 4 вузли обробки даних, продуктивністю 500 запитів/секунду кожен;
- 2 вузли обробки даних, продуктивністю 2000 запитів/секунду кожен.

Для перших трьох варіантів організації підсистеми роботи вузлів обслуговування, бачимо що час відповіді різко зростає в момент коли завантаженість вузлів обробки сягає > 95% та стає практично однаковим. Однак при завантаженості вузлів в межах до 75%, можна побачити що система із більшою кількості вузлів обробки показує кращі результати в плані часу відповіді.

Не менш важливим параметром, який характеризує якісні показники роботи системи в цілому є час затримки обслуговування на кореневому сервері. Враховуючи особливість запропонованої системи, цей час буде залежати від коефіцієнта ефективності кешування даних на граничних серверах. Чим меншим буде значення коефіцієнта ефективності кешування даних, тим більше запитів буде перенаправлено до кореневого сервера, відповідно тим більшим буде значення часу відповіді системи вцілому. В процесі математичного моделювання було вибрано три значення коефіцієнта ефективності кешування, а саме: 0 – кешування відсутнє, всі запити перенаправляються на кореневий сервер; 0.6 (60%) – ефективність кешування, яку забезпечують існуючі методи балансування та обробки даних в граничних локаціях мереж доставки; 0.9 (90%) – ефективність кешування, яку забезпечує запропонований в роботі метод розподіленої обробки ресурсозатратних запитів та інтегральний ключ кешування на рівні балансувальника навантаження.

Результати залежності часу відповіді від інтенсивності надходження запитів наведено на рис. 7. Як показано на графіку, за відсутності кешування даних (коефіцієнт ефективності кешування E_cache = 0), ас відповіді починає суттєво зростати зі збільшенням навантаження.

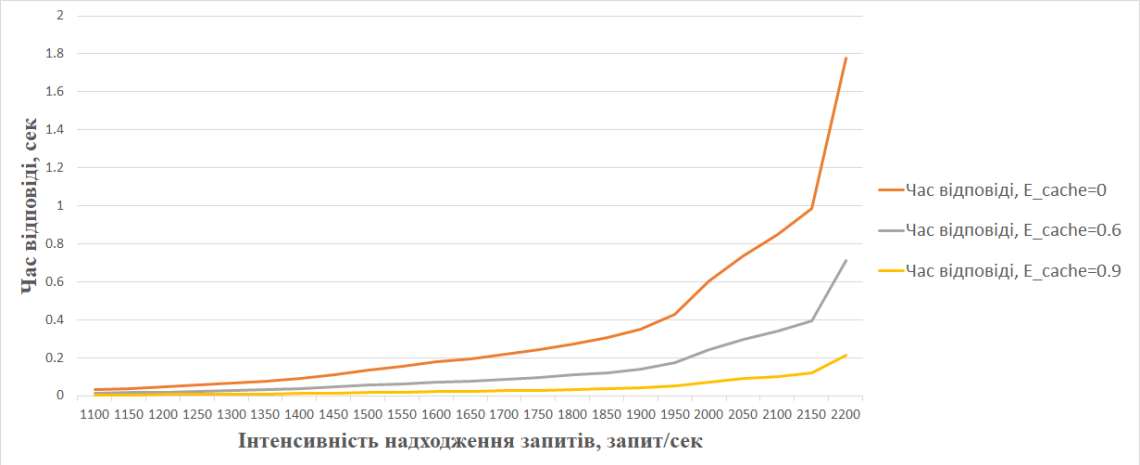

Рис.7. Залежність часу відповіді від інтенсивності поступлення заявок при різних значеннях коефіцієнта ефективності кешування

Це пов’язано з тим, що всі запити повинні бути оброблені всіма системами: балансувальником навантаження, вузлом обробки в граничній локації CDN-мережі та кореневим сервером. При використанні кешування даних, час відповіді суттєво

зменшується із збільшенням коефіцієнта ефективності кешування. При значенні E_cache = 0,9 навіть при суттєвому зростанні навантаження час відповіді залишається в межах 180–200 мс.

Також встановлено, що система, яка складається із більшої кількості менш продуктивних вузлів обробки даних, буде працювати краще, ніж система, яка складається із одного вузла більшої продуктивності, за умови, що особливість трафіку не вимагає складних обчислень та великих затрат процесорного часу. Щодо мікросервісної архітектури, то завжди ефективніше використовувати більшу кількість маленьких за продуктивністю вузлів обслуговування та в будь-який час мати змогу здійснити швидке масштабування.

У четвертому розділі **«Реалізація методу адаптивного створення мікросервісу в точці присутності CDN-мережі»** представлено результати роботи та використання технології розподілених обчислень, які використовуються для швидкої обробки даних у найближчій до користувача локації. Наведено результати роботи методу адаптивного створення мікросервісу в точці присутності мережі CDN, який призначений для забезпечення задовільних параметрів якості обслуговування в мережах передачі даних. Даний метод ґрунтується на використанні алгоритму, який здійснює збір метрик і аналіз параметрів якості обслуговування. Він приймає рішення про початок обробки даних на граничному сервері на основі певних заздалегідь заданих значень. Блок схема роботи розробленого алгоритму представлено на рисунку 8. Використання системи, що працює на базі даного алгоритму, дає змогу завжди мати актуальну інформацію про стан кореневого сервера, актуальні значення параметрів якості обслуговування та гранично допустимі межі, при яких ефективно застосовувати розподілені обчислення в CDN мережі. На основі зібраних статистичних даних, можна прогнозувати періоди підвищеного навантаження та застосовувати метод граничних обчислень для підтримання якості обслуговування в заданих межах.

Рис.8. Блок схема роботи алгоритму адаптивного створення мікросервісу для обробки даних в CDN мережі

На створеному в роботі прототипі мережі доставки контенту, проведено імітаційне моделювання роботи методу адаптивного створення мікросервісу в точках присутності CDN мережі, враховуючи задані граничні значення часу затримки при певній кількості запитів від клієнтів. У результаті моделювання було створено графічну інтерпретацію отриманих результатів. Кількість запитів, яка використовувалась під час експерименту становить 1500 запитів/секунду.

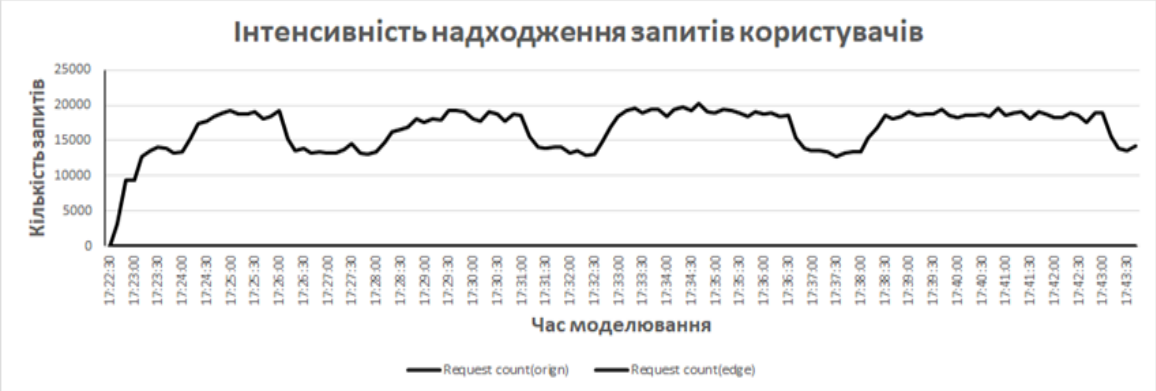

Рис.9. Кількість запитів на проміжку часу моделювання

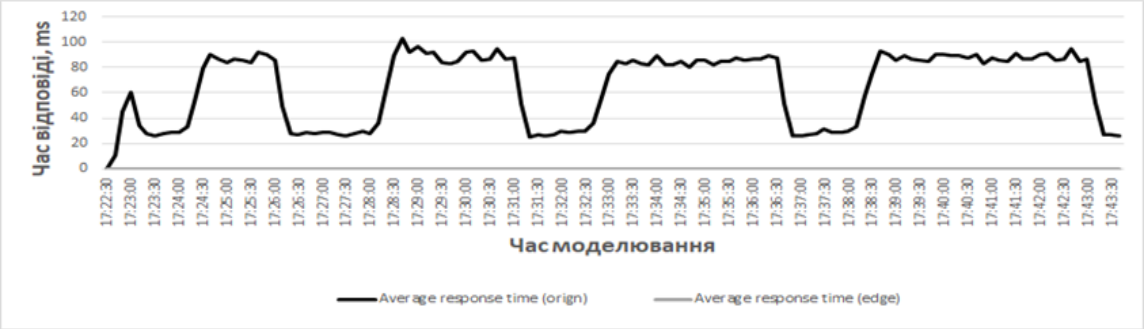

Рис.10. Час затримки на проміжку часу моделювання

Під час моделювання, було проведено чотири стрибки навантаження до 2000 запитів/секунду. Під час першого сценарію, весь трафік оброблявся кореневим сервером, розподілені граничні обчислення не використовувались. Як можна побачити із рисунків 9-10, при звичайному навантаженні, 1500 запитів/секунду, час відповіді був в межах 30-40 мс. Перший стрибок навантаження тривав 2 хв, при цьому весь трафік балансувальник навантаження надсилав тільки на кореневий сервер. Час відповіді при такому навантаженні був в межах 90-100 мс. Наступні стрибки були дещо довші, тривали 3, 4 та 5 хв. Навантаження було в межах 2000 запитів/секунду, а час відповіді зберігався в межах 90-100мс. Підсумовуючи результати першого експерименту, можна стверджувати, що при зростанні навантаження на третину, час відповіді зростав більше як у два рази без використання розподілених граничних обчислень.

Наступний сценарій моделювання був дуже наближеним до попереднього. Вхідні параметри задавались такими ж, основною відмінністю було те, що балансувальник навантаження відразу перенаправив 50% запитів на граничний сервер розподілених обчислень. Результати моделювання представлено на рис 11-12.

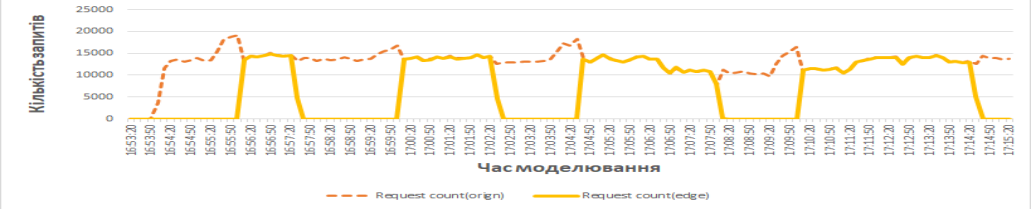

Рис.11. Кількість запитів на кореневий та граничний сервер

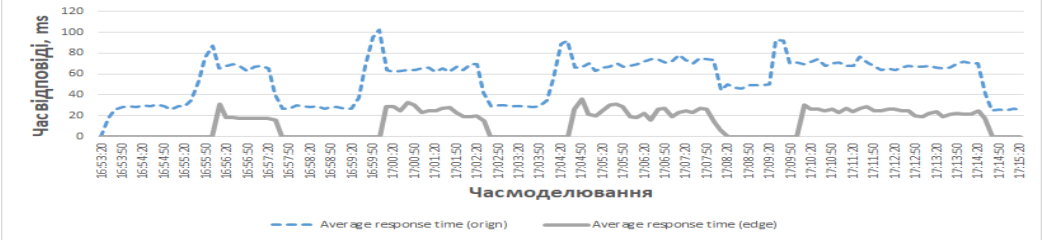

Рис.12. Час відповіді від кореневого/граничного серверів

Аналізуючи результати моделювання, можна побачити, що як тільки час відповіді від кореневого сервера досягав встановленого порогового значення, а саме 80 мс, балансувальник навантаження половину всіх запитів перенаправив на граничний сервер розподілених обчислень. На протязі хвилини, час відповіді від кореневого сервера вернувся до початкових значень, а час відповіді від граничного сервера розподілених обчислень був у межах 20-30 мс. Якщо порівняти результати із попереднім сценарієм моделювання, то перевагою є те, що навантаження на кореневий сервер зменшується практично відразу, час відповіді також вертається в межі 60-70 мс.

Із результатів експериментальних досліджень можна зазначити, що використання методу адаптивного створення мікросервісу в граничних локаціях CDN мережі, дозволяє значно зменшити навантаження на кореневий сервер, скоротити час затримки для кінцевого користувача при отриманні контенту, а також знизити ймовірність втрати даних під час їх передачі. Усі ці переваги сприяють покращенню якості обслуговування в мережах передачі даних.

ОСНОВНІ РЕЗУЛЬТАТИ ТА ВИСНОВКИ

В дисертаційній роботі розв’язано науково-практичне завдання підвищення якості обслуговування користувачів в умовах обмеженої обчислювальної інфраструктури, шляхом розроблення методів і алгоритмів обробки запитів, що потребують значних обчислювальних ресурсів, кешування даних, реалізації балансування навантаження та адаптивного розгортання мікросервісів у точках присутності CDN мереж.

Основні наукові та практичні результати полягають у наступному:

1. Проведено аналіз існуючих підходів та архітектурних рішень, які використовуються під час проєктування та впровадження сучасних сервісів і послуг у мережах доставки контенту. Встановлено, що для ефективного використання ресурсів CDN-мереж, а також забезпечення можливості застосування розподілених граничних обчислень у найближчих до користувача локаціях, доцільно застосовувати мікросервісну архітектуру. Такий підхід дозволяє гнучко управляти компонентами аплікацій і використовувати методи їх

швидкого розгортання, що підвищує масштабованість, стійкість до збоїв і можливість адаптації системи до змін у моменти зростання навантаження.

2. Удосконалено метод обробки запитів, що потребують значних обчислювальних ресурсів у точках присутності CDN мережі, який враховує дані аналітики щодо популярності певних типів контенту, таких як веб-сторінки, мультимедійний контент, серед списку найбільш запитуваних, що дало змогу забезпечити ефективне використання кешованих даних, зменшити навантаження на кореневий сервер та час відповіді на запити кінцевого користувача. Застосування даного методу дозволило скоротити час відповіді для кінцевого користувача на 9% та на 10% завантаженість кореневого сервера, що сприяло покращенню загальної продуктивності системи та підвищенню якості взаємодії з користувачем.

3. Запропоновано інтегральний ключ кешування, що являє собою унікальний ідентифікатор, який формується шляхом поєднання кількох параметрів або компонентів, чим дає можливість деталізувати критерії вибору даних, які зберігатимуться як одне ціле та забезпечує ефективне кешування статичних даних в мережах доставки контенту. Інтегральний ключ кешування може бути адаптований для роботи з різними типами мережевого трафіку, а його складові можуть змінюватися відповідно до специфіки застосування. Правильно сформований ключ кешування, який також враховує аналітичні дані щодо популярності контенту в локаціях CDN мережі, забезпечує підвищення коефіцієнта ефективності кешування, зменшення навантаження на кореневі сервери та зниження затримок при доступі до даних, що є важливим аспектом підвищення якості обслуговування у мережі доставки контенту.

4. Удосконалено метод балансування навантаження у мережах доставки контенту, який враховує значення інтегрального ключа кешування сформованого на основі розробленого методу обробки запитів, локацію клієнта, наявність контенту на граничному сервері та стан функціонування доступних серверів, для забезпечення ефективного використання кеш-пам'яті та ресурсів обчислювальної інфраструктури. Застосування такого типу балансування у граничній локації CDN мережі, забезпечує ефективність кешування даних до 90% та зменшення часу відповіді із 1600 мс до 800 мс для кінцевого користувача. Забезпечення таких показників є важливим не тільки для кінцевого користувача, а безумовно і для збереження хорошого рейтингу ресурсу в пошукових системах, які використовуються в глобальних мережах.

5. Запропоновано використання комплексної математичної моделі з трьома об’єднаними системами масового обслуговування для характеристики процесу обробки запитів у хмарних сервісах із динамічним виділенням ресурсів. Встановлено, що система, яка складається із більшої кількості менш продуктивних вузлів обробки даних, буде працювати краще, ніж система, яка складається із одного вузла більшої продуктивності за умови що особливість трафіку не вимагає складних обчислень та великих затрат процесорного часу.

6. Розроблено метод адаптивного розгортання мікросервісів для обробки динамічних даних у режимі реального часу в точках присутності CDN-мереж, який частково дублює бізнес-логіку сервісу з кореневого сервера, здійснює в режимі

реального часу аналіз параметрів, які визначають якість послуги, та адаптивно розподіляє запити на основі оцінювання рівня завантаженості граничних серверів для забезпечення необхідної якості обслуговування. Результати застосування методу демонструють зменшення часу затримки для кінцевого користувача, навіть за умов двократного зростання кількості запитів.

7. Наведено результати практичного використання методу адаптивного розгортання мікросервісу та дослідження його роботи на прикладі хмарного сервісу Google Cloud Run, що забезпечує хорошу продуктивність програм та аплікацій. Із результатів досліджень можна зазначити, що використання методу адаптивного розгортання мікросервісу в граничних локаціях CDN мережі, дозволяє значно зменшити навантаження на кореневий сервер, скоротити час затримки для кінцевого користувача при отриманні контенту, а також знизити ймовірність втрати даних під час їх передачі, що забезпечує покращення якості обслуговування.

ОСНОВНІ РОБОТИ, ОПУБЛІКОВАНІ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Статті у наукових фахових виданнях України та наукових періодичних виданнях інших держав, що входять до міжнародних наукометричних баз даних

1. M. Kyryk, N. Pleskanka, M. Pitsyk, V. Kyryk, “Infrastructure as code and microservices for intent-based cloud networking,” *Lecture Notes in Electrical Engineering: Future intent-based networking. On the QoS robust and energy efficient heterogeneous software defined networks*, vol. 831, pp. 51–68, 2022.
2. М.В. Плесканка, “Покращення параметрів якості обслуговування QoS в CDN мережі за рахунок використання модуля Edge Compute,” *Інфокомунікаційні технології та електронна інженерія,* Випуск.3, № 2, с. 64-73, 2023.
3. Н.М. Плесканка, М.В. Плесканка, Т.С. Слободзян,Б.М. Марко, “Аналіз ефективності використання мікросервісів при розробці web додатків,” *Комп'ютерні системи проектування. Теорія і практика.* Випуск 6, № 2, с. 146-157, 2024.
4. М.І. Кирик, Н.М. Плесканка, М.В. Плесканка, “Аналіз роботи методу оптимізованого кешування даних в мережі доставки,” *Проблеми телекомунікацій.* № 1 (22), с. 43–55, 2018.
5. М.І. Кирик, Н.М. Плесканка, М.В. Плесканка, “Дослідження ефективності використання мережі CDN,” *Вісник Національного університету “Львівська політехніка”. Радіоелектроніка та телекомунікації,* № 885, с. 31–40, 2017.
6. М.І. Кирик, В.Б. Янишин, М.В. Плесканка, “Оцінка ефективності методів спектральної мобільності у когнітивних радіомережах,”. *Вісник Національного університету “Львівська політехніка”. Радіоелектроніка та телекомунікації,* №849, с. 194 – 202, 2016.
7. Т.А. Максимюк,О.М. Яремко, М.В. Піцик, “Моделі конвергенції гетерогенних мереж мобільного зв’язку 5-го покоління на основі технології D2D,”. *Телекомунікаційні та інформаційні технології, Київ, ДУТ,* № 3, с. 91-102, 2016.
8. М.В. Кайдан, В.С. Андрущак, М.В. Піцик, В.З. Пашкевич, “Аналіз енергетичного балансу оптичної транспортної мережі з урахуванням

технологічних і архітектурних підходів," *Вісник Національного університету "Львівська політехніка". Радіоелектроніка та телекомунікації,* №818, с. 120-129, 2015.

Публікації у матеріалах конференцій, що входять до міжнародних наукометричних баз даних:

9. M. Kyryk, O. Tymchenko, N. Pleskanka, and M. Pleskanka, "Methods and process of service migration from monolithic architecture to microservices," in 2022 IEEE 16th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), 2022.
10. M. Kyryk, N. Pleskanka, M. Pleskanka, "Adaptive Edge Compute module in CDN networks," Advanced Information and Communication Technologies: Proceedings of the 5th IEEE International Conference, Lviv, Ukraine, 2023
11. M. Kyryk, N. Pleskanka, M. Pleskanka, "Dynamic Data Processing on Edge Locations of CDN Network," in 2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), 2024.
12. M. Kyryk, N. Pleskanka, M. Pleskanka, and P. Nykonchuk, "Load balancing method in edge computing," in 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), 2020.
13. M. Kyryk, N. Pleskanka, and M. Pleskanka, "The analysis of the optimal data distribution method at the content delivery network," in 2019 IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), 2019.
14. M. Kyryk, N. Pleskanka, and M. Pleskanka, "Analysis of the technologies and methodologies of data transmission in distributed information systems," in 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), 2018.
15. M. Kyryk, M. Pleskanka, and N. Pleskanka, "The efficiency and productivity of the CDNs," in 2017 2nd International Conference on Advanced Information and Communication Technologies (AICT), 2017.
16. M. Kyryk, N. Pleskanka, and M. Pitsyk, "QoS mechanism in content delivery network," in 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), 2016.
17. M. Kyryk, N. Pleskanka, and M. Pleskanka, "Content delivery network usage monitoring," in 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), 2017.
18. M. Kaidan, V. Andrushchak, and M. Pitsyk, "Calculation model of energy efficiency in optical transport networks," in 2015 Second International Scientific-Practical Conference Problems of Infocommunications Science and Technology (PIC S&T), 2015.
19. Y. Pyrih, M. Kaidan, I. Tchaikovskyi, and M. Pleskanka, "Research of genetic algorithms for increasing the efficiency of data routing," in 2019 3rd International Conference on Advanced Information and Communications Technologies (AICT), 2019.

АНОТАЦІЯ

Плесканка М.В. “Підвищення якості обслуговування у мережі доставки контенту”. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за фахом 05.12.02 – телекомунікаційні системи та мережі. – Національний університет “Львівська політехніка”, Львів, 2025.

Дослідження спрямоване на розробку методів і алгоритмів обробки запитів, що потребують значних обчислювальних ресурсів, кешування даних, реалізації балансування навантаження та адаптивного розгортання мікросервісів у точках присутності CDN мереж, для підвищення якості обслуговування користувачів в умовах обмеженої обчислювальної інфраструктури.

В роботі запропоновано інтегральний ключ кешування, що формується шляхом поєднання кількох параметрів, та може використовуватись для різних типів мережевого трафіку. Представлено метод обробки запитів, що потребують значних обчислювальних ресурсів, який забезпечує можливість розподіленої обробки даних в точках присутності CDN мережі. Удосконалено метод балансування навантаження в граничних локаціях CDN мережі, який враховує значення інтегрального ключа кешування, наявність контенту на граничному сервері, а також його доступність.

Запропоновано комплексну математичну модель з трьома об’єднаними системами масового обслуговування для характеристики процесу обробки запитів у хмарних сервісах із динамічним виділенням ресурсів. На основі математичного моделювання проведено оцінку параметрів якості обслуговування.

Вперше розроблено метод адаптивного розгортання мікросервісу в точці присутності CDN мережі, який забезпечує дублювання бізнес-логіки сервісу з кореневого сервера на граничні сервери, здійснює аналіз параметрів, що визначають якість послуги в режимі реального часу та адаптивно розподіляє запити на основі оцінювання рівня завантаженості граничних серверів.

Практичне значення отриманих результатів полягає у тому, що удосконалено метод обробки запитів статичних даних у мережах доставки контенту, який дав змогу покращити якість обслуговування користувачів шляхом зменшення часу відповіді для кінцевого користувача до 9%, та на 10% завантаженість кореневого сервера. Комплексне використання інтегрального ключа кешування, методу обробки запитів та балансування трафіку у точках присутності CDN мережі, дало змогу підвищити коефіцієнт ефективності кешування на 30 %, а також зменшити час відповіді на запити кінцевого користувача на 40%. Розроблений метод адаптивного розгортання мікросервісів для обробки динамічних даних у режимі реального часу в точках присутності CDN-мереж, що дає змогу забезпечити необхідну якість обслуговування в умовах обмежених ресурсів кореневого сервера.

Ключові слова: CDN мережа, балансування навантаження, мікросервісна архітектура, хмарні обчислення, граничний сервер, кореневий сервер, маршрутизація запитів, хмарні сервіси, інфраструктура як код.

SUMMARY

Pleskanka M. V. "Improving the quality of service in the content delivery network". – Manuscript.

Dissertation on the competition of scientific degree of engineering's science' candidate on specialty 05.12.13 - Radio Engineering and telecommunications devices – Lviv Polytechnic National University, Lviv, 2025.

The research is aimed at developing methods and algorithms for processing requests, that require significant computing resources, data caching, implementing load balancing and adaptive microservices deployment at CDN network points of presence, to improve the quality of user service in conditions of limited computing infrastructure.

The paper considers the main principles and approaches related to the microservice architecture. A comprehensive analysis of the literature on dynamic load balancing and data caching, focusing on the current state of the strategies used, their advantages, disadvantages, implementation challenges, and their impact on the efficiency of CDN networks.

An integral caching key is proposed, which is formed by connecting several parameters, and can be used for various types of network traffic. A method of processing requests that requires significant computing resources is proposed, which provides the possibility of distributed data processing at CDN network. A load balancing method in edge locations of the CDN network is improved, which takes into account the value of the integral caching key, the content availability on the edge server, and functioning state of available servers.

A complex mathematical model of the queuing system, comprising three integrated subsystems, is presented to characterize the process of servicing requests in cloud services with dynamic resource allocation. Based on mathematical modeling, the quality of service parameters were assessed.

For the first time, a method for adaptive microservice deployment at CDN network point of presence was developed. This method provides duplication of the business logic service from the origin server to edge servers, analyzes the parameters that trigger the quality of service in real time, and adaptively distributes requests based on the edge servers load.

The practical significance of the results is that the method for processing static data requests in content delivery networks has been improved, which allows improve the quality of user service by reducing the response time for the end user by up to 9%, and the load of the origin server by 10%. The integrated use of the integral caching key, the method of processing and balancing traffic at CDN network points of presence allowed increase the caching efficiency by 30%, as well as reduce response time for end user requests by 40%. A method of adaptive microservices deployment for processing dynamic data in real-time mode at CDN networks points of presence has been developed, which makes it possible to actually provide the required quality of service in conditions of limited origin server resources.

Keywords: CDN network, load balancing, microservices architecture, cloud computing, edge server, origin server, request routing, cloud services, Infrastructure as Code.

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

GCP	Google Cloud Platform	хмарна платформа Google
API	Application Programming Interface	інтерфейс програмування додатків
CDN	Content Delivery Network	мережа доставки контенту
QoS	Quality of Service	якість обслуговування
LB	Load Balancer	балансувальник навантаження
EC	Edge Compute	обробка даних на граничному сервері

Підписано до друку 09.04.2025 р.
Формат 60×90 1/16. Папір офсетний.
Друк на різографі. Умовн. друк. арк. 1,5. Обл.-видав. арк. 0,89.
Тираж 100 прим. Зам. 250184.

Поліграфічний центр
Видавництва Національного університету “Львівська політехніка”
вул. Ф.Колесси, 4, 79013, Львів
Реєстраційне свідоцтво серії ДК № 4459 від 27.12.2012 р.