«Системні технології» 6 (161) 2025 «System technologies»
DOI 10.34185/1562-9945-5-161-2025-15
УДК 004.042, 004.02, 004.67

Д. Божуха, О. Байбуз

МОДЕЛЬ СПІЛЬНОГО ДИНАМІЧНОГО РОЗВАНТАЖЕННЯ ХМАРНОЇ АРХІТЕКТУРИ З БАЛАНСУВАННЯМ РІВНІВ

Анотація. Останніми роками для вирішення задачі оброблення великого обсягу запитів з низькою затримкою науковцями запропоновано велика кількість підходів для отримання характеристик хмарної архітектури, що можуть використовувати принципи побудови системи.

У роботі розглядається можливість створення гібридної системи, яка у своєму складі має типову систему граничних, туманних та хмарних обчислень для дослідження питання оптимізації навантаження за рахунок перерозподілу частини завдань між пристроями на рівнях. Наведено опис моделі, яка поєднує принципи проєктування граничних, туманних та хмарних обчислень. Для перевірки опису побудованої моделі використана модель з обмеженою кількістю пристроїв у своєму складі. Проведено аналіз отриманих даних про зв'язок між стратегією та функцією траєкторії навантаження хмарної системи.

Ключові слова: математична модель, система хмарних обчислень, послідовність процесів, балансування навантаження.

Постановка проблеми

Активне використання хмарних систем різних типів та архітектури в прикладних задачах та отримання великої кількості даних на реальних експериментах відкриває нові напрями дослідження мінімізації та прогнозування навантаження, вибору стратегії швидкого розвантаження системи, що формує достатньо велику кількість інформаційних технологій для моніторингу, управління та адаптації ресурсів хмарного сервісу у режимі реального часу.

Важливою задачею стає вибір архітектури хмарної системи та стратегії її розвантаження на етапі проєктування з використанням наявних ресурсів та їх характеристик, що вимагає аналізу структури хмарного сервісу та вибору існуючих інформаційних технологій.

Аналіз останніх досліджень і публікацій

Останні дослідження, які пов'язані з хмарними системами, пропонують парадигми обчислень для вирішення великого обсягу завдань різного типу з мінімальною затримкою. Запропоновані парадигми можуть використовувати ресурси, пристрої, вузли та кластери хмарного центру. Але основним завданням створення таких парадигм є вирішення проблеми розвантаження системи хмарних обчислень при використанні ком-

бінації існуючих та нових підходів. Наприклад, при дослідженні хмарної архітектури, яка запропонована у вигляді ієрархічної системи граничного туману [1], запропоновано звернути увагу на зміщення між рівнями системи відповідно до толерантності користувача до затримки. У багатьох роботах науковців приділена увага задачі використання методів машинного навчання для вирішення проблем розвантаження обчислень та управління мобільністю хмарної системи на різних рівнях її архітектури. Актуальним напрямом розвитку хмарних обчислень є вирішення проблеми прогнозування навантаження (проактивна оптимізація) з вбудовуванням в систему інтелектуальних агентів для моніторингу, управління та адаптації ресурсів хмарного сервісу режимі реального часу, які взаємодіють між собою та центральними оркестраторами для автономного масштабування та самовідновлення системи [2]. Для дослідження більш складної структури системи хмарних обчислень використана ідея автора роботи [3] щодо запропонованої стратегії динамічного спільного розвантаження хмарних граничних пристроїв з урахуванням балансування навантаження.

Мета дослідження

Метою дослідження є розгляд стратегії динамічного спільного розвантаження хмарних пристроїв з урахуванням балансування навантаження на граничних серверах та балансування підключення на туманних вузлах для багаторівневої структури системи хмарних обчислень. Проведення експериментів та аналіз отриманих результатів.

Викладення основного матеріалу дослідження

Автором роботи [3] представлено архітектурне рішення хмарної системи з рівнями кінцевих пристроїв (DL), граничних серверів (EL) та хмарного центру (CL). В запропонованому дослідженні додано рівень туману (FL) для наближення моделі системи хмарних обчислень до реальної. Типовий робочий процес системи хмарних обчислень формується з декількох етапів: користувачі створюють завдання, кінцеві пристрої рівня DL розподіляють отримані завдання на граничні сервери рівня EL через основну мережу під керуванням схеми розвантаження рівня DL, граничні сервери розміщують отримані завдання у черзі, на кожному граничному сервері частина завдань оброблюється локально, а інша частина може переноситися на інші граничні сервери для балансування навантаження на рівні EL; паралельно завдання можуть перенаправлятися за рахунок роботи схеми керування щодо вибору маршрутизаторів підключення рівня FL; частина завдань рівня FL може бути також перенаправлена за рахунок роботи хмарного центру для віддаленої допомоги під керуванням схеми розвантаження хмари, після оброблення завдання надходять на кінцеві пристрої до користувачів з граничних серверів, туманних вузлів або з хмарного центру.

Математичне формулювання задачі.

Розглянемо гібридну систему, яка включає в себе типову систему граничних, туманних та хмарних обчислень для дослідження питання оптимізації навантаження за рахунок перенесення частини завдань з граничних серверів до маршрутизатора підключення, а з туманних вузлів до хмарного центру. Частина цієї системи при включенні першого та останнього рівнів з архітектури (граничний та хмарний) представляє собою

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

CEE, що забезпечує послуги з низькою затримкою для кінцевих користувачів через граничні сервери [3].

Розглянемо систему хмарних обчислень, у складі якої є N кінцевих користувачів, ES граничних серверів та FR маршрутизаторів навантаження. Нехай множина кінцевих користувачів $U = \{u_1, u_2, \dots, u_N\}$, множина граничних серверів $S = \{s_1, s_2, \dots, s_{ES}\}$, множина туманних вузлів (маршрутизаторів підключення) $R = \{r_1, r_2, \dots, r_{FR}\}$. В будь-який момент часу $t \in [0, T]$ швидкість виконання завдань кінцевими пристроями u_i рівня DC позначимо як $v_{1,i}^{task}(t)$, швидкість виконання завдань граничними серверами s_j рівня EC позначимо як $v_{2,j}^{task}(t)$ та швидкість виконання завдань туманними вузлами r_k рівня FC позначимо як $v_{3,k}^{task}(t)$. Швидкість хмарного центру позначимо як $v_{4,1}^{task}(t)$. Частка завдань, які є вивантаженими у момент часу $t \in [0, T]$ з кінцевого пристрою u_i на граничний сервер s_j, з граничного серверу s_j на туманний вузол r_k та з туманного вузлу до хмарного центру позначимо відповідно $x_{ij}^{US}(t)$, $x_{jk}^{SR}(t)$ та $x_{k1}^{RC}(t)$. Функція $x(t) = \{x_{ik}(t)\}_{N \times FR}$ для $t \in [0, T]$ є стратегією розвантаження рівнів DC, EC та FC. Навантаження (довжина черги завдань) граничних серверів s_j та туманних вузлів r_k позначено як $q_j^{EC}(t)$ та $q_k^{FC}(t)$ відповідно. Навантаження хмарного центру позначимо функцією $q^{bal}(t)$. Прослідкувати за траєкторією навантаження можна за допомогою функції

$$Balance(t) = \{q_1^{EC}(t), \dots, q_{ES}^{EC}(t), q_1^{FC}(t), \dots, q_{FR}^{FC}(t), q^{bal}(t)\}$$

Після надходження завдань до черг пристроїв виконується часткове їх локальне оброблення або перенесення на інші пристрої за допомогою схем розвантаження/підключення з подальшим вивантаженням на наступні рівні. Визначимо швидкості перенесення власних завдань з s_n на інший s_m для контролю балансування навантаження рівня EC через функцію $f_{level}(n, m, q_n^{level}(t), t)$.

Стратегію розвантаження хмари визначає множина

$$y(t) = \{y_k(t)\}_{k=1}^{FR} = \left\{v_{3,k}^{task}(t) x_{k1}^{RC}(t)\right\}_{k=1}^{FR}, \qquad t \in [0; T]$$

Математичний вираз стратегій розвантаження системи

Розглянемо набір функцій $x(t)$ та $y(t)$, які є стратегіями розвантаження рівнів системи (у визначеному прикладі, це рівні DC, EC та FC) та хмарного рівня для $t \in [0; T]$. Множина стратегій розвантаження системи може бути представлена як:

$$\Phi = \{(x,y): x \in \mathbb{R}^{N\times ES} \times \mathbb{R}^{ES\times FR}, y\epsilon\mathbb{R}^{FR} |$$

$$x_{ij}(t) \in [0;1], x_{jk}(t) \in [0;1], \sum_{j=1}^{ES} x_{ij}^{US}(t) = 1, \sum_{k=1}^{FR} x_{jk}^{SR}(t) = 1,$$

$$y_k(t) \in [0; y_{max}], i = 1,\dots,N, j = 1,\dots,ES, k = 1,\dots,FR, t \in [0;T]\}$$

Розглянемо можливість збільшення/зменшення навантаження на рівень DC, яке пов'язано з змінами кількості запитів від користувача $\Delta R_i(t) = R_i(t+\Delta t) - R_i(t)$ протягом інтервалу часу $[t; t+\Delta t]$.

Дослідження зв'язку між стратегією та функції траєкторією навантаження

Надалі вважаємо, що кожний об'єкт побудованої схеми системи має чергу завдань нескінченої довжини (без пріоритетів).

Навантаження рівня СС в момент часу $t + \Delta t$ дорівнює:

$$q^{bal}(t+\Delta t) =$$

$$= q^{bal}(t) + \Delta t \sum_{k=1}^{FR} \delta\left(q_k^{FC}(t)\right) v_{3,k}^{task}(t) x_{k1}^{RC}(t) - \Delta t\, \delta\left(q^{bal}(t)\right) v_{4,1}^{task}(t)$$

Навантаження рівнів FC та EC в момент часу $t + \Delta t$ залежить від швидкості виконання/опрацювання завдань пристроями рівнів, частки вивантажених завдань з пристроїв рівнів, швидкостей перенесення завдань між пристроями на рівні (балансування), наявності черги та її довжини.

Траєкторія навантаження функції $Balance(t)$ залежить від множини стратегій розвантаження $\Phi(x,y)$ динамічної системи:

$$q^{bal}(t) = Q_0, t \in [0,T]$$

$$(q_k^{FC}(t+\Delta t) - q_k^{FC}(t)) / \Delta t, k = 1,2,\dots,FR, t \in [0,T]$$

$$(q_j^{EC}(t+\Delta t) - q_j^{EC}(t)) / \Delta t, j = 1,2,\dots,ES, t \in [0,T]$$

Висновки

У результаті дослідження отримано модель еволюції навантаження багаторівневої системи для аналізу впливу обраних стратегій для вирішення задачі розвантаження рівнів системи. Запропоновано модель багаторівневої системи, яка поєднує принципи проєктування граничних, туманних та хмарних обчислень, в якій при перевантаженості пристроїв на рівнях системи частина завдань може бути частково перерозподілена.

Надалі можна розглянути гібридну систему, яка включає в себе типову систему граничних, туманних та хмарних обчислень для дослідження питання оптимізації навантаження за рахунок перенесення частини завдань з граничних серверів та з туманних вузлів одразу до хмарного центру або з граничних серверів, туманних вузлів, хмарного центру до центру контролю навантаження.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

ЛІТЕРАТУРА

1. Diamanti M, Charatsaris P, Tsiropoulou EE, Papavassiliou S. Incentive mechanism and resource allocation for edge-fog networks driven by multi-dimensional contract and game theories. IEEE Open Journal of the Communications Society. 2022;3:435–452.
2. Bodra D and Khairnar S (2025) Machine learning-based cloud resource allocation algorithms: a comprehensive comparative review. Front. Comput. Sci. 7:1678976. - DOI: 10.3389/fcomp.2025.1678976
3. Fan Y (2024) Load balance -aware dynamic cloud-edge-end collaborative offloading strategy // PLOS ONE 19(1): e0296897. - URL: https://doi.org/10.1371/journal.pone.0296897

REFERENCES

1. Diamanti M, Charatsaris P, Tsiropoulou EE, Papavassiliou S. Incentive mechanism and resource allocation for edge-fog networks driven by multi-dimensional contract and game theories. IEEE Open Journal of the Communications Society. 2022;3:435–452.
2. Bodra D and Khairnar S (2025) Machine learning-based cloud resource allocation algorithms: a comprehensive comparative review. Front. Comput. Sci. 7:1678976. - DOI: 10.3389/fcomp.2025.1678976
3. Fan Y (2024) Load balance -aware dynamic cloud-edge-end collaborative offloading strategy // PLOS ONE 19(1): e0296897. URL:https://doi.org/10.1371/journal.pone.0296897

Received 25.09.2025.
Accepted 30.09.2025.

A model of joint dynamic offloading of cloud architecture with load balancing of layers

Analysis of recent studies and publications. Recent research related to cloud systems offers computing paradigms for solving a large number of tasks of various types with minimal latency. The proposed paradigms can use resources, devices, nodes and clusters of the cloud center. But the main task of creating such paradigms is to solve the problem of offloading the cloud computing system using a combination of existing and new approaches. For example, when studying the cloud architecture, which is proposed in the form of a hierarchical boundary fog system [1], it is proposed to pay attention to the shift between the levels of the system according to the user's tolerance for latency. Many works of scientists have paid attention to the problem of using machine learning methods to solve the problems of offloading calculations and managing the mobility of the cloud system at different levels of its architecture. The current direction of cloud computing development is to solve the problem of load forecasting (proactive optimization) with the integration of intelligent agents into the system for monitoring, managing and adapting cloud service resources in real time, which interact with each other and central orchestrators for autonomous scaling and self-healing of the system [2].

To study a more complex structure of the cloud computing system, the author's idea of the work [3] was used regarding the proposed strategy of dynamic joint unloading of cloud edge devices taking into account load balancing.

Purpose of research. The purpose of the study is to consider a strategy for dynamic joint offloading of cloud devices, taking into account load balancing on edge servers and connection balancing on fog nodes for a multi-tier structure of a cloud computing system. Conducting experiments and analyzing the results.

Presentation of the main research material. The author of the work [3] presented an architectural solution of a cloud system with the levels of end devices (DL), edge servers (EL) and cloud center (CL). In the proposed study, a fog level (FL) was added to approximate the model of the cloud computing system to the real one. A typical workflow of a cloud computing system is formed from several stages: users create tasks, end devices of the DL level distribute the received tasks to edge servers of the EL level through the main network under the control of the DL level offloading scheme, edge servers place the received tasks in a queue, on each edge server part of the tasks is processed locally, and the other part can be transferred to other edge servers for load balancing at the EL level; in parallel, tasks can be redirected due to the operation of the control scheme for selecting connection routers of the FL level; Part of the FL level tasks can also be redirected through the work of the cloud center for remote assistance under the control of the cloud offload scheme, after processing the tasks are delivered to the end devices to users from edge servers, fog nodes or from the cloud center.

Conclusions. As a result of the study, a model of the evolution of the load of a multi-tier system was obtained to analyze the impact of the selected strategies for solving the problem of unloading the system levels. A model of a multi-tier system was proposed that combines the principles of edge, fog and cloud computing design, in which, when devices at the system levels are overloaded, part of the tasks can be partially redistributed.

In the future, a hybrid system can be considered that includes a typical edge, fog and cloud computing system to study the issue of load optimization by transferring part of the tasks from edge servers and fog nodes directly to the cloud center or from edge servers, fog nodes, cloud center to the load control center.

Keywords: mathematical model, cloud computing system, process sequence, load balancing.

Байбуз Олег Григорович - доктор технічних наук, професор, завідувач кафедри інженерії програмного забезпечення та інформаційних технологій, Дніпровський національний університет імені Олеся Гончара, bozhukha_d@365.dnu.edu.ua, ORCID: https://orcid.org/0000-0001-7489-6952

Божуха Даніїл Ігорович – аспірант, Дніпровський національний університет імені Олеся Гончара, bozhukha_d@365.dnu.edu.ua, ORCID: https://orcid.org/0009-0007-6869-4415

Baibuz Oleh - Doctor of Technical Sciences, Professor, Head of the Department of Software Engineering and Information Technologies, Oles Honchar Dnipro National University, bozhukha_d@365.dnu.edu.ua, ORCID: https://orcid.org/0000-0001-7489-6952

Bozhukha Daniil – postgraduate student, Oles Honchar Dnipro National University, bozhukha_d@365.dnu.edu.ua, ORCID: https://orcid.org/0009-0007-6869-4415

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)