УДК 621.391

МОДЕЛЬ ОЦІНКИ ЕФЕКТИВНОСТІ ФУНКЦІОНУВАННЯ ХМАРНОГО ЦЕНТРУ З ВИСОКИМ СТУПЕНЕМ ВІРТУАЛІЗАЦІЇ ТА В УМОВАХ ГРУПОВОГО НАДХОДЖЕННЯ ЗАПИТІВ

Н.В. Пелех, О.М. Шпур, М.М. Климаш
Національний університет «Львівська політехніка»

Abstract *– The current state of information technology is to develop and implement new approaches to the computational process. Evaluating the effectiveness of cloud centers is an important challenge for research, but it is complicated by the dynamic of cloud environments and a variety of user requests. This evaluation is vital in cases where virtualization is used to provide well-defined computing resources for users. The proposed model for evaluating the effectiveness of cloud centers in a high degree of virtualization to solve this problem has been proposed. Compared to existing, it considers the ability to service requests for group requests and the distributed time of service requests. The model is based on a two-stage approximation technique. The main non-Markov process is first modeled as an embedded semi-Markov process, then modeled as an approximated Markov process but only when receiving group request flows. The technique of constructing Markov links to build the model has been used. This model provides a full probability distribution of request waiting time, response time to execute requests, and the number of requests in the system. The results show that the performance of cloud centers is highly dependent on the coefficient of variation (CoV), request service time, and the size of the group flow (i.e., the number of requests in the group flow of requests). The larger the flow rate and/or the value of the coefficient of variation of the service time of requests, the longer the response time. But this helps reduce the use of resources by cloud providers. As a result, the work shows that in the conditions of large group flow of requests and/or large value of CoV, it is possible to increase the efficiency of cloud centers by grouping requests using the criterion of homogeneity.*

Анотація *– Сучасний стан розвитку інформаційних технологій спрямований на розроблення та впровадження нових підходів до організації обчислювального процесу. Оцінка ефективності хмарних центрів є важливим напрямком досліджень, проте вона ускладнюється динамічним характером хмарного середовища та різноманітністю запитів користувачів. Особливо важливою така оцінка є у випадку, де віртуалізація використовується для забезпечення чітко визначених обчислювальних ресурсів для користувачів. Для вирішення цієї проблеми пропонується модель для оцінки ефективності функціонування хмарних центрів в умовах високого ступеня віртуалізації, яка на відмінну від існуючих враховує можливість обслуговування запитів у разі групового надходження запитів і розподілений час їх обслуговування. Модель базується на двоступеневій техніці апроксимації, де основний немарківський процес спочатку моделюється як вкладений напівмарківський процес, який потім моделюється як апроксимований процес Маркова, але тільки в моменти надходження групових потоків запитів. Для побудови моделі використовується техніка побудови ланцюгів Маркова. Дана модель забезпечує повний імовірнісний розподіл часу очікування запитів, часу відгуку щодо виконання запитів і кількість запитів у системі. Результати показують, що продуктивність роботи хмарних центрів значно залежить від коефіцієнту варіації (CoV), часу обслуговування запитів і розміру групового потоку (тобто кількості запитів у груповому потоці запитів). Чим більшими є розмір потоку та/або значення коефіцієнта варіації часу обслуговування запитів, тим довша тривалість відгуку. Проте це сприяє зменшенню рівня використання ресурсів хмарними провайдерами. Дослідження показує, що в умовах надходження великих групових потоків запитів та/або великого значення CoV можна досягти підвищення ефективності функціонування хмарних центрів шляхом групування запитів за критерієм однорідності.*

Вступ

Сучасний стан розвитку інформаційних технологій спрямований на розроблення та впровадження нових підходів до організації обчислювального процесу. Оцінка ефективності функціонування хмарних центрів є важливим напрямком досліджень, проте вона ускладнюється динамічним характером хмарного середовища та різноманітністю

запитів користувачів. Із аналізу нещодавніх досліджень в області хмарних обчислень тільки незначна їх частина була присвячена оцінці ефективності [1-11], яка є важливою у випадку, коли віртуалізація використовується для забезпечення чітко визначених обчислювальних ресурсів для користувачів. Водночас важливе значення відіграє ступінь віртуалізації, тобто число віртуальних машин, що працюють на одній фізичній машині. Достатньо високий ступінь віртуалізації, і відповідно велика кількість запитів, які обслуговують віртуальні машини, може призвести до перевантаження системи загалом.

Зазвичай хмарний центр моделюється як класична відкрита мережа з одним входом, в якій розподіл часу відгуку набуває вигляду експоненти та приймає значення між часом прибуття та часом обслуговування запиту. Науковці [1] дослідили час відгуку з різними значеннями метрик, таких як накладні витрати на придбання та реалізацію віртуальних ресурсів тощо. Для зменшення часу відгуку вони розробили та впровадили портативну, масштабовану та просту в роботі платформу для генерування та відправки тестових навантажень до обчислювальних хмар. В роботі [2] хмарний центр був змодельований як система з чергою типу $M/M/m/m+r$ (де $m+r$ – розмір буферу m з максимальною кількістю запитів r, які одночасно знаходяться в системі: на обслуговуванні та в черзі), в якій вже є змодельований розподіл часу відгуку. Припускається, що розподіл часу між моментами надходження запитів і початком їх обслуговування експоненціальний. Час відгуку розбивається на час очікування, час обслуговування та період виконання за умови, що всі три періоди є незалежними (що є нереалістичним на думку авторів [2]). Ієрархічний підхід до моделювання системи для оцінки продуктивності функціонування в хмарному середовищі був запропонований у роботі [3]. Суть підходу полягала у застосуванні відомих моделей: класичної формули Ерланга для оцінки втрат та системи масового обслуговування типу M/M/m/K, як базис для моделювання системи – для оцінки пропускної здатності вихідного каналу та моделювання часу відгуку.

В роботах [4-9] науковці запропонували загальні аналітичні моделі, які базуються на підході до аналізу продуктивності хмарних сервісів. Запропонований у роботі [4] підхід зменшує складність аналізу продуктивності хмарних центрів шляхом поділу загальної моделі на простіші Марківські підмоделі та отримання загального результату з ітерації над окремими підмоделями. Проте, запропонований авторами підхід поділу на підмоделі та вимоги, які встановлюються під час проведення досліджень, є не цілком виправданими. Часи прибуття запитів у кожній із підмоделей підпорядковуються строго експоненціальному закону розподілу, а система загалом характеризується невисоким ступенем віртуалізації (в даному дослідженні кількість віртуальних машин в одній фізичній машині менша 10), що в реальних умовах надання хмарних сервісів є досить обмеженим.

Все вищевикладене визначає актуальність дослідження та його мету – розробка моделі оцінки ефективності функціонування хмарного центру з високим ступенем віртуалізації та в умовах групового надходження запитів від користувачів. У всіх подібних роботах час обслуговування підпорядковується експоненціальному закону розподілу. Крім того, ступінь віртуалізації не враховується належним чином або ігнорується вза-

галі [11]. Для вирішення цієї проблеми пропонується модель оцінки ефективності хмарних центрів в умовах високого ступеня віртуалізації, яка на відміну від існуючих враховує можливість обслуговування під час групового надходження запитів і розподілений час обслуговування запитів. Модель базується на двоступеневій техніці апроксимації, а для її побудови використовується техніка побудови ланцюгів Маркова. Такий підхід забезпечить точне обчислення важливих показників ефективності, таких як середня кількість завдань в системі, довжина черги, час відгуку та час очікування, ймовірність блокування та ймовірність негайного обслуговування.

I. Концепція побудови хмарного центру

Вважатимемо, що хмарний центр складається з M фізичних машин, кожна з яких може розмістити m віртуальних машин, як показано на рис. 1. Вхідні запити направляються через сервер розподілу навантаження до одного з M фізичних серверів. Користувачі можуть здійснити запит на одну або більше віртуальних машин одночасно, тобто від одного користувача можливе групове надходження запитів, що в умовах великої кількості користувачів є типовим. Приймемо, що групове надходження запитів буде підпорядковуватися Пуассонівському закону розподілу.

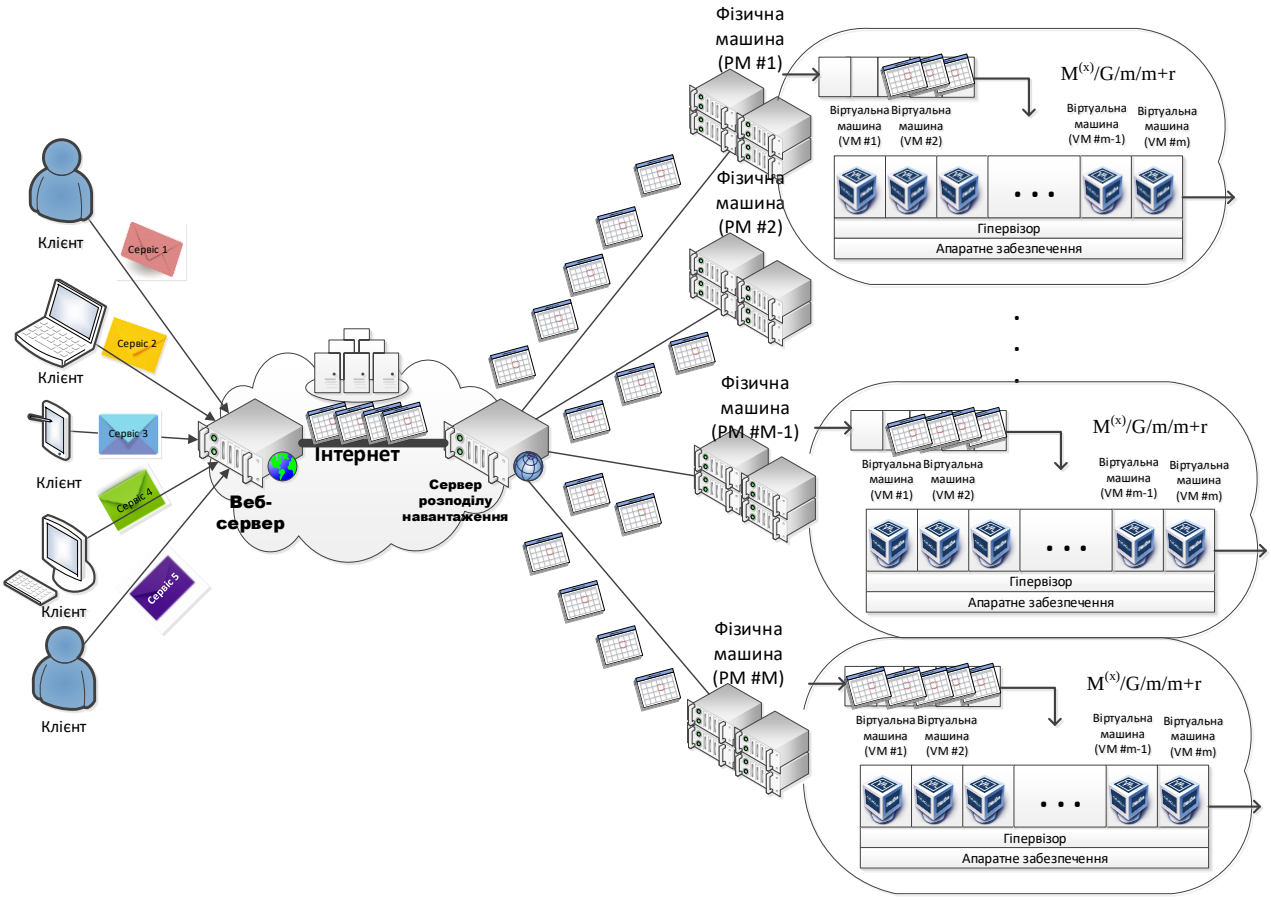

Рис. 1. Архітектура досліджуваного хмарного центру надання сервісів

Коли відбувається групове надходження запитів на сервер, сервер розподілу навантаження здійснює спробу обслужити його – тобто надати йому одну з M фізичних машин. Оскільки кожна фізична машина має скінченну вхідну чергу, такі вхідні групові запити обробляються наступним чином:

- якщо фізична машина з достатньою вільною ємністю знайдена, то групові запити розпаралелюються між доступними віртуальними машинами і виконуються негайно;
- якщо фізичну машину з достатнім розміром вхідної черги не знайдено, то запити в межах групового потоку запитів стають в чергу на виконання;
- інакше – потік запитів відкидається.

Такий спосіб відомий як загальний спосіб ефективного прийому/відхилення користувачів, які зазвичай не приймають часткового виконання своїх запитів на обслуговування [1].

На сьогоднішній день провайдери хмарних послуг не публікують інформацію щодо дослідження таких статистичних параметрів, як час виконання сервісу, тому можна припустити, що загальний розподіл часу обслуговування запитів переважно той, що враховує коефіцієнт варіації (він визначається відношенням середньоквадратичного відхилення та середнього значення величини). Коли загальне навантаження на одну фізичну машину збільшується, то зменшується продуктивність віртуальних машин, які працюють одночасно, і фактична тривалість надання сервісу зростатиме. Запропонована у статті модель враховує цю особливість.

Змоделюємо кожну фізичну машину в хмарному центрі, як систему масового обслуговування типу *M*/*G*/*m*/*m*+*r* (Пуассонівським потоком вхідних запитів з довільним розподілом часу обслуговування, *m* фізичних машин та розміром буферу *m* з максимальною кількістю запитів *r*, які одночасно знаходяться в системі: на обслуговуванні та в черзі, тобто з урахуванням запитів, які обслуговуються віртуальними машинами) Вважатимемо, що груповий потік запитів, які прибули на обслуговування, підпорядковується П уассонівському закону розподілу, а час прибуття A описується експоненціальним законом розподілу. Вважатимемо, що потоки запитів надходять з інтенсивністю λ_j (запитів за секунду). Кумулятивна функція розподілу ймовірностей часу прибуття групового потоку запитів протягом часового інтервалу x матиме вигляд $CDF(A) = P\{A < x\}$. Враховуючи закон розподілу часу прибуття групового потоку, ймовірність того, що протягом часового інтервалу x надійдуть групові потоки запитів з інтенсивністю λ_j визначатиметься як $PDF(x) = \lambda_j e^{-\lambda x}$. Тоді час очікування i-го запиту в черзі на обслуговування із використанням перетворення Лапласа-Стілтьєса визначатиметься як

$$A^j(S) = \int_0^\infty e^{-Sx} PDF(x)dx = \frac{\lambda_j}{\lambda_j + S}. \quad (1)$$

Нехай g_k – імовірність того, що груповий потік має розмір $k = 1,2,...MRS$, де MRS – максимальна кількість запитів в потоці, $\bar{g}$ – середнє значення розміру групового потоку, а $\prod_g(z)$ – імовірнісна твірна функція групового потоку, які визначаються відповідно:

$$\prod_g(z) = \sum_{k=1}^{MRS} g_k z^k, \ \bar{g} = \prod_g^{(1)}(1). \quad (2)$$

II. Аналітична модель оцінки ефективності хмарного центру з високим ступенем віртуалізації та в умовах групового надходження запитів

Припустимо, що час обслуговування запитів у межах групового запиту однаковий і рівномірно розподілений відповідно до загального розподілу. Проте, параметри загального потоку (наприклад, середнє значення, дисперсія тощо) залежать від ступеня віртуалізації фізичних машин. Вважатимемо, що зміна середнього часу обслуговування у зв'язку зі зміною робочого навантаження така ж, як у дослідженнях, проведених у статті [2]. На жаль, не можна зробити якийсь висновок щодо зміни моментів часу обслуговування, приведених у статті [3]. Наприклад, на рис. 2 наведені результати оцінки продуктивності засобами VMmark для сімейства фізичних машин з різною кількістю віртуальних машин. З даних на рис. 2 можна побудувати залежність середнього часу обслуговування від кількості розгорнутих на фізичних машинах віртуальних машин, яка нормалізована за часом обслуговування, отриманого за рахунок роботи фізичних машин як окремого елементу. Слід зазначити, що середній час обслуговування об'єднує як обробку запитів, так і тривалість обслуговування запиту в груповому потоці запитів.

Залежність, приведена на рис. 2, може бути апроксимованою і показувати нормалізоване значення середнього часу обслуговування $\overline{b_n}(y)$ як функцію, залежну від кількості елементів та ступеня віртуалізації:

$$\overline{b_n}(y) = \frac{\bar{b}(y)}{\bar{b}(1)}, \quad (3)$$

де y – кількість віртуальних машин на одній фізичній машині.

Кумулятивна функція розподілу часу обслуговування групового потоку запитів y віртуальними машинами матиме вигляд $B_y(x) = P\{B_y(x) < x\}$, а його функція щільності $b_y(x)$ визначатиметься як $PDF(x)$. Тоді час очікування j-го запиту в черзі на обслуговування y віртуальними машинами із використанням перетворення Лапласа-Стілтьєса визначатиметься таким чином:

$$B_y^j(S) = \int_0^\infty e^{-sx} b_y(x)dx\,. \tag{4}$$

У випадку, якщо кожна віртуальна машина зможе одночасно обслужити максимально Ω запитів, то функція розподілу часу обслуговування j-го запиту y віртуальними машинами визначатиметься:

$$B^j(s) = \sum_{y=1}^{\Omega} p_y B_y^j(s)\,, \tag{5}$$

де p_y – імовірність того, що на кожній фізичній машині розташовано y віртуальних машин.

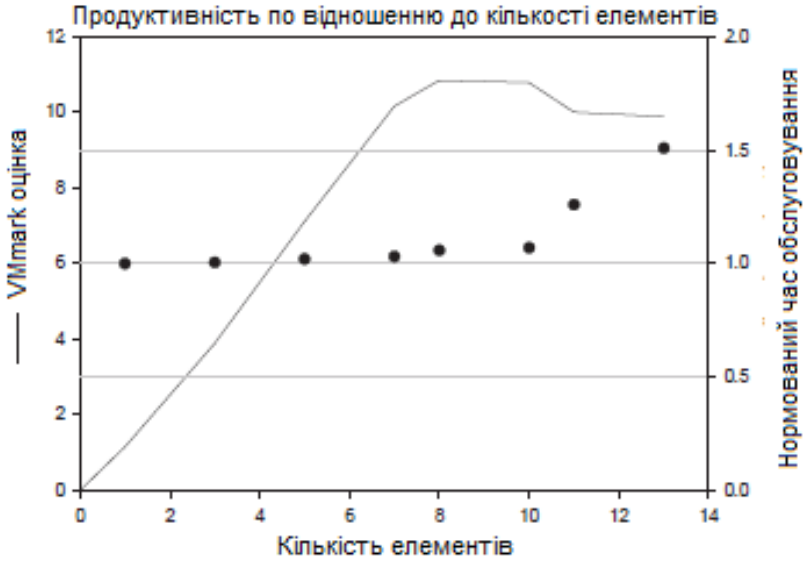

Рис. 2. Результати оцінки продуктивності фізичних машин із різною кількістю розгорнутих віртуальних машин засобами VMmark

Тоді загальний середній час обслуговування дорівнює

$$\bar{b} = \sum_{y=1}^{\Omega} p_y \bar{b}(y)\,. \tag{6}$$

Розглянемо час обслуговування як часовий інтервал, протягом якого розпочинається і закінчується обслуговування запиту, що надійшов на обслуговування. Позначимо через B_+ час обслуговування запитів у довільній точці перед початком та до закінчення інтервалу обслуговування, а B_- – інтервал часу від початку обслуговування до довільної точки в середині цього інтервалу. Вважатимемо, що вони мають однакову функцію розподілу часу обслуговування та визначатимуться як

$$B_+^j(s) = B_-^j(s) = \frac{1 - B^j(s)}{s\bar{b}}. \quad (7)$$

Це дозволить врахувати ступінь віртуалізації при визначенні функції розподілу часу перебування запитів у кожній «ділянці» при дослідженні переходу запитів зі стану в стан (надходження запиту – обслужений запит), який описуватиметься нижче.

III. Використання ланцюгів Маркова для побудови достовірної моделі оцінки ефективності функціонування хмарного центру

Система масового обслуговування типу M/G/m/m+r у разі надходження групового потоку запитів здійснює їх паралельно-групове обслуговування. За умови такого способу обслуговування важливим є процес переходу запитів із стану в стан (надходження запиту – обслужений запит). Для опису таких переходів і побудови достовірної моделі оцінки ефективності функціонування хмарного центру використаємо техніку побудови ланцюгів Маркова із деякою апроксимацією. Для початку змоделюємо немарківську систему з вкладеним напівмарківським процесом, який надалі матиме назву eSMP, що збігається з основною системою в моменти надходження групових потоків запитів і їх виходу з системи. Для визначення ймовірності переходу до вкладеного напівмарківського процесу необхідно підрахувати кількість обслужених запитів між двома запитами, які прибули один за одним, оскільки процес обслуговування є паралельним.

Наступним кроком після процесу підрахунку є введення вкладеного марківського процесу (позначимо як aEMP), який моделює напівмарківський процес дискретним способом (рис. 3). Стани нумеруються відповідно до кількості запитів, які наразі знаходяться в системі. Вважатимемо, що у aEMP розподілення вихідних запитів, котрі передаються між двома запитами, які прибули послідовно, відбувається за час надходження групового потоку.

Нехай маємо протікання трьох процесів: початковий, вкладений напівмарківський та один апроксимований вкладений процес. Пронумеруємо ділянки залежно від кількості запитів, що надходять у систему в даний момент часу (які включають в себе і ті, які обслуговуються, і ті, які очікують).

Нехай A_n та A_{n+1} вказують відповідно моменти n та $(n+1)$ надходження групового потоку запитів у систему, а q_n і q_{n+1} вказують на кількість запитів, які знайдені в системі перед тим, як прибули наступні, що схематично показано на рис. 4. Якщо k – кількість запитів у груповому потоці, а V_{n+1} вказує на кількість запитів, які обслужені системою між A_n та A_{n+1}, то $q_{n+1} = q_n - V_{n+1} + k$.

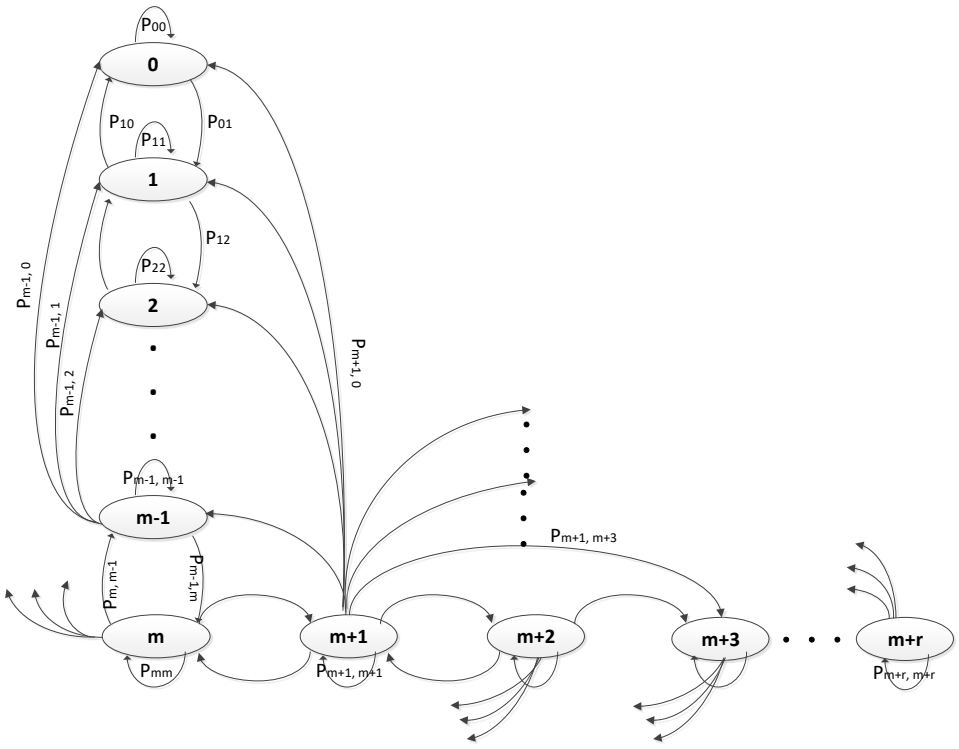

Рис. 3. Процес переходу запитів зі стану в стан з використанням *л*анцюгів Маркова

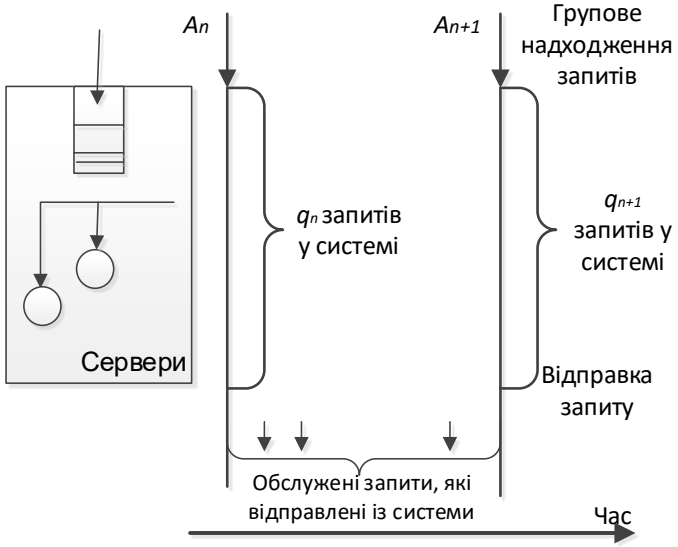

Рис. 4. Точки спостереження процесів залежно від кількості запитів, що надходять у систему в даний момент часу

Для знаходження вкладеного процесу переходу запитів зі стану в стан з використанням *л*анцюгів Маркова (позначимо його як аЕМС) необхідно розрахувати відповідні ймовірності переходів, які можуть визначатися як

$$P(i,j,k) \equiv P[q_{n+1} = j \mid q_n = i \quad \text{та} \quad X_g = k]. \tag{8}$$

Іншими словами, необхідно знайти ймовірність того, що $i+k-j$ клієнтів обслуговуються у проміжку часу між двома послідовними надходженнями групових потоків. Такий процес знаходження ймовірності вимагає точного знання поведінки системи між двома груповими потоками запитів, що надходять. Очевидно, що для $j > i+k$, $P(j,k,i)=0$. Звідси випливає, що існує принаймні не більше $i+k$ запитів, що потребують обслуговування між моментами надходження A_n та A_{n+1}. Для розрахунку інших імовірностей переходів у aEMC необхідно визначити функцію розподілу часу перебування запитів у кожній «ділянці» в eSMP. Для цього необхідно розглянути декілька випадків:

- випадок 1: час перебування в «ділянці» на початку відправки залишається часом обслуговування сервісу $B_+(x)$ з моменту останнього надходження запитів і є випадковою величиною в момент обслуговування поточного запиту;
- випадок 2: якщо починається надходження другого групового потоку запитів з того ж сервера, то час обслуговування запитів дорівнює $B(x)$;
- випадок 3: ні час відправки між надходженням двох групових потоків, ні останній час відправки перед наступним надходженням запитів не відповідатиме експоненціальному закону розподілу;
- випадок 4: якщо i – час відправки запитів з іншого серверу, то функція часу перебування ділянки CDF набуває значення $B_{i+}(x)$.

Розглянемо час відправки запиту D_{21} (рис. 5), який наступає після часу відправки D_{11}. Момент часу D_{11} можна розглядати як довільну точку в моменті часу обслуговування запитів на сервері 2, а отже, кумулятивна функція розподілу часу перебування в другому моменті відправки матиме вигляд $B_{2+}(x)$.

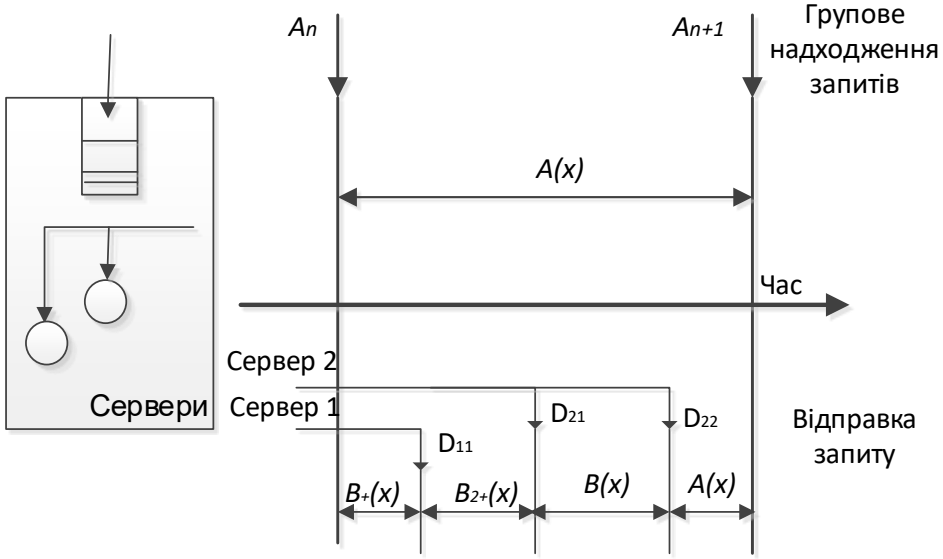

Рис. 5. Поведінка системи між двома точками спостереження

В результаті функція розподілу часу обслуговування разом із $B_{2+}(x)$ є такою ж, як і $B_{+}^{j}(s)$, аналогічно (7), але з додатковим кроком рекурсії. Зазвичай функція розподілу часу обслуговування з урахуванням часу знаходження запитів між різними серверами, може бути рекурсивно визначена та мати такий вигляд:

$$B_{i+}^{j}(s)=\frac{1-B_{(i-1)+}^{j}(s)}{s*\bar{b}_{(i-1)+}},\ i=1,2,3\ldots, \tag{9}$$

де $\bar{b}_{(i-1)+}=[\frac{d}{ds}B_{(i-1)+}^{j}(s)]_{s=0}$. Для підтримки узгодженості в позначеннях приймемо, що $\bar{b}_{0+}=\bar{b}, B_{0+}^{j}(s)=B^{j}(s)$ та $B_{1+}^{j}(s)=B_{+}^{j}(s)$.

IV. Імовірність відправки сервісу та рівняння балансу

Для оцінки продуктивності хмарних центрів з урахуванням переходів зі стану в стан, необхідно визначити передавальну ймовірнісну матрицю, елементами якої є кількість запитів, що відправляються із системи в проміжку часу між двома прибулими послідовними груповими потоками запитів. Перш за все, для цього необхідно обчислити ймовірність існування k запитів під час перебування в кожному елементі. Нехай $N(B_{+})$, $N(B)$ та $N(B_{i+})$, де $i=2,3,\ldots$, вказують на кількість запитів, які прибули в періоди часу $B_{+}(x)$, $B(x)$ і $B_{i+}(x)$ відповідно. Оскільки досліджувана система характеризується Пуассонівським потоком надходження запитів, можна визначити ці ймовірності:

$$\alpha_k\equiv P[N(B_{+})=k]=\int_0^{\infty}\frac{(\lambda x)^k}{k!}e^{-\lambda x}dB_{+}(x); \tag{10}$$

$$\beta_k\equiv P[N(B)=k]=\int_0^{\infty}\frac{(\lambda x)^k}{k!}e^{-\lambda x}dB(x); \tag{11}$$

$$\delta_{ik}\equiv P[N(B_{i+})=k]=\int_0^{\infty}\frac{(\lambda x)^k}{k!}e^{-\lambda x}dB_{i+}(x). \tag{12}$$

Проте, у подібних системах існує ймовірність того, що в певний момент часу черга запитів порожня і запити на обслуговування не надходять. Відштовхуючись від цього, можна визначити ймовірність переходу в стан eEMC наступним чином:

$$P_x\equiv P[N(B_{+})=0]=\int_0^{\infty}e^{-\lambda x}dB_{+}(x)=B_{+}^{j}(\lambda); \tag{13}$$

$$P_y\equiv P[N(B)=0]=\int_0^{\infty}e^{-\lambda x}dB_{+}(x)=B^{j}(\lambda); \tag{14}$$

$$P_{ix} \equiv P[N(B_{2+}) = 0] = \int_0^\infty e^{-\lambda x} dB_{i+}(x) = B_{i+}^j(\lambda); \tag{15}$$

$$P_{xy} = P_x P_y. \tag{16}$$

Імовірності переходу P можуть бути зображені у вигляді матриці переходу (17), де рядки і стовпці відповідають кількості запитів у системі безпосередньо перед прибуттям групового потоку (стовпці) та відповідно відразу ж після чергового прибуття групового потоку (рядки). Можна виділити чотири робочих області залежно від того, чи вхідна черга порожня, чи ні, до і після послідовних надходжень групових потоків запитів. Кожна область має яскраво виражену ймовірність переходу P-рівняння, яке залежить від поточного стану i, наступного стану j, кількості запитів в груповому потоці k і кількості вихідних запитів між надходженням двох групових потоків.

$$\left[\begin{array}{cccc|ccc}
(0,0) & (0,1) & \dots & (0,m) & (0,m+1) & \dots & (0,m+r) \\
\cdot & & \textit{Область_1} & & \cdot & \textit{Область_2} & \\
\cdot & & P_{LL}(i,j,k) & & \cdot & P_{LH}(i,j,k) & \\
\cdot & & & & \cdot & & \\
(m,0) & (m,1) & \dots & (m,m) & (m,m+1) & \dots & (m,m+r) \\
- & - & - & - & - & - & - \\
(m+1,0) & (m+1,1) & \dots & (m+1,m) & (m+1,m+1) & \dots & (m+1,m+r) \\
\cdot & & \textit{Область_4} & & \cdot & \textit{Область_3} & \\
\cdot & & P_{HL}(i,j,k) & & \cdot & P_{HH}(i,j,k) & \\
\cdot & & & & \cdot & & \\
(m+r,0) & (m+r,1) & \dots & (m+r,m) & (m+r,m+1) & \dots & (m+r,m+r)
\end{array}\right] \tag{17}$$

Область 1

В даній області визначається ймовірність переходу $P_{LL}(i,j,k)$, в якій вхідна черга порожня та залишається порожньою до наступного надходження запитів. Тому переходи починаються та закінчуються в області запитів з мітками m чи без них.

Область 2

В даній області визначається ймовірність переходу $P_{LH}(i,j,k)$, в якій черга порожня перед переходом, але потім зайнята, тому $i \le m, j > m$. Це означає, що надіслані групові потоки запитів стають у чергу, оскільки не можуть обслуговуватися відразу через недостатню кількість вільних фізичних машин.

Для розрахунку ймовірності переходу для областей 3 та 4 потрібно знайти ймовірність існування n непрацюючих серверів та m, які простоюють. Розглянемо надходження потоку запитів, який підпорядковується Пуассонівському закону розподілу. Кожен груповий потік запитів зберігається в кінцевій черзі. Зберігання групового надходження у черзі буде продовжено, поки черга або не буде повна, або останнє надходження запитів не буде вміщатися в ній.

Область 3

В даній області визначається ймовірність переходу $P_{HH}(i, j, k)$, в якій черга не порожня, тобто $i, j > m$. У цьому випадку переходи починаються і закінчуються в станах m, як на рис. 3

Область 4

У даній області визначається ймовірність переходу $P_{HL}(i, j, k)$, в якій черга не порожня під час першого надходження запиту, але стає порожньою в момент наступного надходження потоку запитів: $i > m, j \le m$.

Рівняння балансу матриці переходу матиме такий вигляд:

$$\pi_i = \sum_{j=\max[0,i-MBS]}^{m+r} \pi_j p_{ji}, 0 \le i \le m+r, \tag{18}$$

яке доповненюється рівнянням нормалізації $\sum_{i=0}^{m+r} \pi_i = 1$. Загальна кількість рівнянь становить $m+r+2$. Таким чином, необхідно видалити одне з рівнянь, щоб отримати єдине рівноважне рішення. Рівнянь (18) цілком досить, оскільки вони володіють мінімальним обсягом інформації про систему у порівнянні з усіма іншими.

V. Розподіл кількості запитів по фізичних машинах

Як тільки отримаємо ймовірності стійкого стану, встановимо значення твірної функції групового потоку для заданої кількості запитів у фізичних машинах у момент надходження групового потоку, що матиме вигляд

$$\Pi(z) = \sum_{k=0}^{m+r} \pi_z z^k . \tag{19}$$

Завдяки груповим надходженням не всі запити виконуються. Таким чином, функція $\prod(z)$ для розподілу кількості запитів фізичними машинами в довільний момент часу не співпадає з функцією $P(z)$ для розподілу кількості запитів у фізичних машинах у момент надходження групового потоку. Для отримання першої функції використовуємо техніку двокрокової апроксимації з вкладеним напівмарківським процесом та апроксимованим вкладеним марківським процесом, відповідно до рівнянь 8 та 9.

Функція $P(z)$ з кількістю запитів в фізичній машині задається у вигляді

$$P(z) = \sum_{i=0}^{m+r} p_i z^i . \tag{20}$$

Оскільки момент надходження запитів не залежить від стану буферу та розподілу кількості запитів у фізичній машині, то можна отримати ймовірність блокування

групових запитів по фізичних машинах з розміром буфера r. Вона визначатиметься таким чином

$$P_{block}(r) = \sum_{k=0}^{MRS-1} [\sum_{i=0}^{MRS} p_{m+r-i-k}(1-G(i))]^* P_i(k). \quad (21)$$

Розмір буфера r_φ зобов'язаний зберігати ймовірність блокування нижче певного значення ψ, а саме:

$$r_\varphi = \{r \geq 0 \mid P_{block}(r) \leq \psi, P_{block}(r-1) > \psi\}. \quad (22)$$

Отже, ймовірність блокування в хмарному центрі з M фізичними машинами розраховується як

$$P_{block.cloud} = (P_{block})^M. \quad (23)$$

VI. Дослідження адекватності моделі оцінки ефективності функціонування хмарного центру

Отримані рівняння балансу з аналітичної моделі були розв'язані за допомогою програмного засобу Maple 15 виробництва Maple Soft. Ця система комп'ютерної алгебри призначена для символьних обчислень, хоча має ряд засобів і для чисельного рішення диференціальних рівнянь, і знаходження інтегралів, володіє легко інтегрованими графічними засобами. Maple включає динамічну мову програмування в імперативному стилі (схожу на Паскаль), яка дозволяє визначати змінні будь якого формату. Також Maple оснащено програмними інтерфейсами, які дозволяють проводити інтеграцію з системами, написаними за допомого інших мов програмування (C, C #, Fortran, Java, MATLAB та Visual Basic).

Для перевірки аналітичних розв'язків побудовано дискретний симулятор хмарного центру за допомогою засобів Artifex. Було налаштовано хмарний центр на роботу з M = 1000 фізичних машин та двома ступенями віртуалізації: 100 та 200 віртуальних машин (Virtual Mashine, VM) на кожній фізичній машині. Розмір вхідної черги рівний r = 300 запитів. Припускаємо, що час обслуговування запитів змінюватиметься пропорційно експоненціальному розподілу. Важливим при такому розподілі є наявність коефіцієнту варіації, який встановлюється незалежно від середнього значення. Середнє значення залежить від ступеня віртуалізації з урахуванням формули (3).

Для більшої достовірності встановимо час обслуговування запиту рівним 120 мс та 150 мс відповідно для помірного та високого ступеня віртуалізації. Використовуємо також два значення коефіцієнту варіації CoV=0,5 та 1,4, які відповідно призводять до гіпо- та гіперекспоненціально розподілених часів обслуговування.

На рис. 7 представлені аналітичні результати для різних показників ефективності залежно від розміру групового потоку запитів. Червоною лінією позначено дослі-

дження, при якому CoV=0,5, а кількість VM=100; синьою пунктирною лінією позначено дослідження, при якому CoV=1,4, а кількість VM=100; зеленою пунктирною лінією позначено дослідження, при якому CoV=0,5, а кількість VM=200; фіолетовою пунктирною лінією позначено дослідження, при якому CoV=1,4, а кількість VM=200.

Як видно з графіків, розмір черги збільшується залежно від середнього розміру групового потоку, тоді як середня кількість запитів у системі зменшується (рис. 7 а).

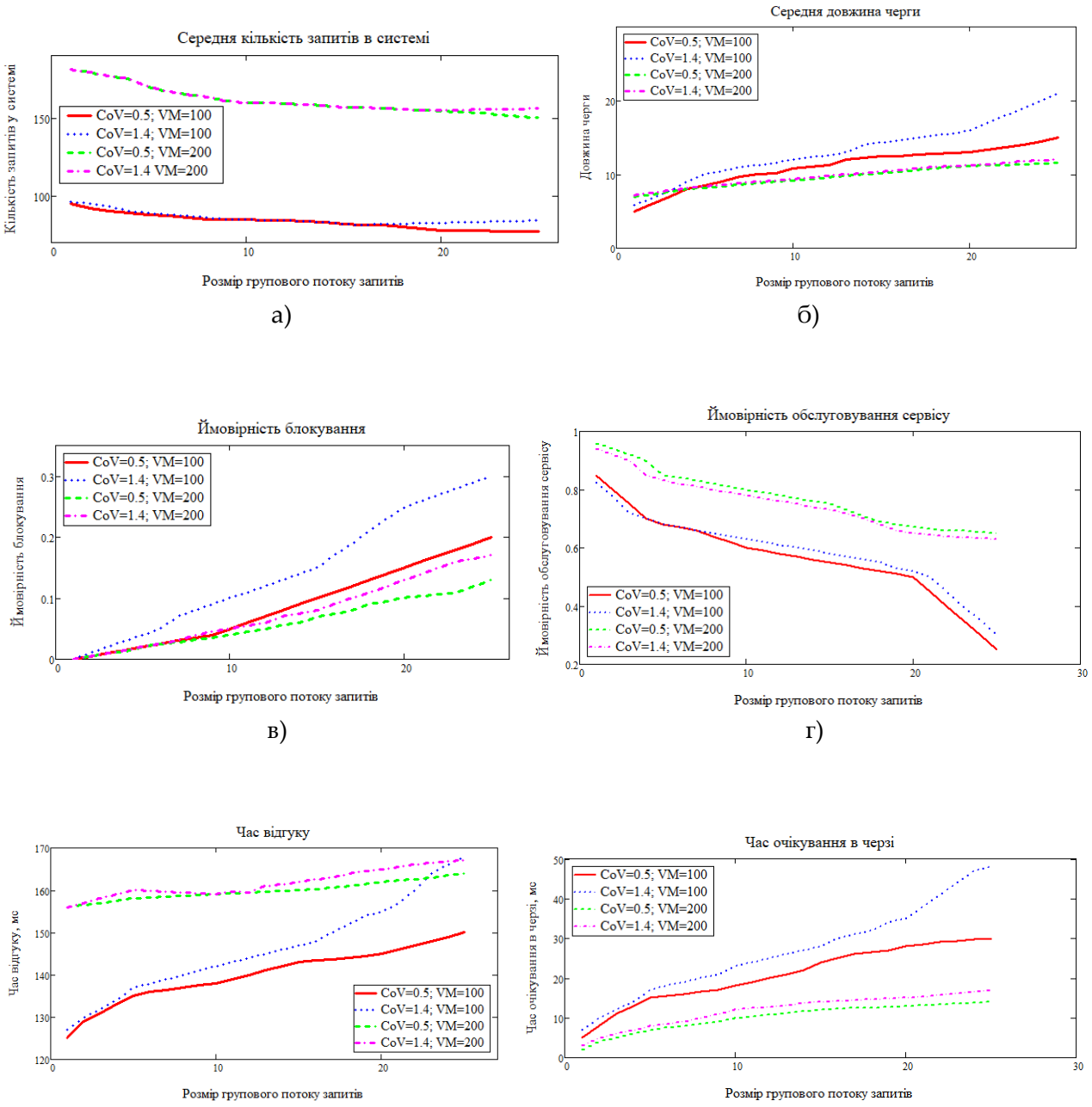

Рис. 7. Показники ефективності як функції середнього розміру потоку запитів

Як видно з отриманих результатів, імовірність блокування зростає, коли збільшується розмір групового потоку (рис. 7 в), а ймовірність обслуговування (яку можна на-

звати продуктивністю) спадає майже лінійно (рис. 7 г). Тим не менш, імовірність обслуговування є вищою за 0,5 у досліджуваному діапазоні середніх розмірів пакетів, яка означає, що більше 50% запитів від користувачів будуть обслуговуватися відразу ж після прибуття.

Як видно із залежностей, представлених на рис. 7 д, час відгуку, який є сумою часу очікування та часу обслуговування, та час очікування (рис. 7 е) повільно зростають зі збільшенням середнього розміру групового потоку. Це підтверджує обґрунтованість запропонованої процедури апроксимації з використанням функцій eSMP , aEMP та aEMC. З дослідження можна дійти висновку, що використання гіперекспоненціального розподілу часу обслуговування (при CoV=1,4) є менш ефективним, ніж використання гіпоекспоненціального (при CoV=0,5). Чим більший розмір групового потоку та/або значення коефіцієнта варіації часу обслуговування запиту, що не більший 1, тим довша тривалість відгуку. Проте, це сприяє зменшенню рівня використання ресурсів хмарними провайдерами. Обидва показники ефективності можуть бути поліпшені за рахунок гомогенізації, отриманої шляхом поділу вхідних групових потоків на основі кількості запитів та/або коефіцієнту варіації часу обслуговування запитів та обслуговування підпотоків окремими хмарними центрами. Таким чином, цей простий підхід пропонує значні переваги як для користувачів, так і для провайдерів хмари.

Висновки

Оцінка ефективності хмарних центрів є важливим завданням для дослідження, але воно ускладнюється динамічним характером хмарного середовища та різноманітністю запитів користувачів. Оцінка ефективності є особливо важливою у випадку, де віртуалізація використовується для забезпечення чітко визначених обчислювальних ресурсів для користувачів і навіть більше, коли ступінь віртуалізації, тобто число віртуальних машин, що працюють на одній фізичній машині, високий. У роботі запропоновано модель оцінки ефективності функціонування хмарного центру з високим ступенем віртуалізації та в умовах групового надходження запитів від користувачів, що дає змогу у разі погіршення продуктивності при великій завантаженості проводити додаткову корекцію параметрів функціонування системи.

Модель використовує теорію масового обслуговування та ймовірнісний аналіз для отримання ряду показників ефективності, в тому числі час відгуку, час очікування в черзі, довжину черги, ймовірність блокування, ймовірність обслуговування сервісу та ймовірність розподілу кількості запитів у системі. Вона базується на двоступеневій техніці апроксимації, де основний немарківський процес спочатку моделюється як вкладений напівмарківський процес, який потім моделюється як апроксимований процес Маркова, але тільки в моменти надходження групових потоків запитів. Для побудови моделі використовується техніка побудови ланцюгів Маркова. Дана модель забезпечує повний імовірнісний розподіл часу очікування запитів, часу відгуку щодо виконання запитів і кількості запитів у системі.

Отримані результати показують, що визначення запитів з однаковим часом обслуговування однорідного хмарного центру може призвести до більшого часу очікування і більш низької ймовірності обслуговування. Показано, що продуктивність роботи хмарних центрів значно залежить від коефіцієнту варіації (CoV), часу обслуговування запитів і розміру групового потоку (тобто кількості запитів у груповому потоці запитів). Чим більший розмір потоку та/або значення коефіцієнта варіації, часу обслуговування запитів, тим довша тривалість відгуку. Проте, це сприяє зменшенню рівня використання ресурсів хмарними провайдерами. Обидва показники ефективності можуть бути поліпшені за рахунок гомогенізації, отриманої шляхом поділу вхідних групових потоків на основі кількості запитів та/або коефіцієнту варіації часу обслуговування запитів та обслуговування підпотоків окремими хмарними центрами. Таким чином, цей простий підхід пропонує значні переваги, як для користувачів, так і для провайдерів хмари. Як результат, робота показує, що в умовах надходження великих групових потоків запитів та/або великого значення CoV можна досягти підвищення ефективності функціонування хмарних центрів шляхом групування запитів за критерієм однорідності.

Список літератури

1. *Ait-Salaht, F., Castel-Taleb, H.* (2015), "Stochastic bounding models for performance analysis of clouds", Proceedings of The 15th IEEE International Conference on Computer and Information Technology (CIT-2015), Liverpool, UK, 26-28 October 2015, P. 603-610. **DOI**: https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.86

2. *Yang, B., Tan, F., Dai, Y., Guo, S.* (2009), "Performance evaluation of cloud service considering fault recovery", Jaatun M.G., Zhao G., Rong C. (eds) Cloud Computing. CloudCom 2009. Lecture Notes in Computer Science, No. 5931, Springer, Berlin, Heidelberg, P. 571–576. **DOI**: https://doi.org/10.1007/978-3-642-10665-1_54

3. *Qian, H., Medhi, D., Trivedi, K. S.* (2011), "A hierarchical model to evaluate quality of experience of online services hosted by cloud computing", Proceedings of the 12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops, Dublin, Ireland, 23-27 May 2011, P. 105-112. **DOI**: https://doi.org/10.1109/INM.2011.5990680

4. *Li, K.* (2016), "Power and performance management for parallel computations in clouds and data centers", Journal of Computer and System Sciences, No. 82(2), P. 174-190. **DOI**: https://doi.org/10.1016/j.jcss.2015.07.001

5. *Khazaei, H., Misic, J., Misic, V. B.* (2013), "Performance of an iaas cloud with live migration of virtual machines", Proceedings of the 2013 IEEE Global Communications Conference (GLOBECOM), Aalanta, GA, USA, 9-13 December 2013, P. 2289-2293. **DOI**: https://doi.org/10.1109/GLOCOM.2013.6831415

6. *Xiong, K., Perros, H.* (2009), "Service performance and analysis in cloud computing", Proceedings of the 1st IEEE Congress on Services, SERVICES'09, Los Angeles, California, USA, 6-10 July 2009, P. 693-700. **DOI**: https://doi.org/10.1109/SERVICES-I.2009.121

7. *Guo, L., Yan, T., Zhao, S., Jiang, C.* (2014), "Dynamic performance optimization for cloud computing using M/M/m queueing system", Journal of Applied Mathematics, No. 2014, 756592, P. 1-8. **DOI**: https://doi.org/10.1155/2014/756592

8. *Bai, W.-H., Xi, J.-Q., Zhu, J.-X., Huang, S.-W.* (2015), "Performance analysis of heterogeneous data centers in cloud computing using a complex queuing model", Mathematical Problems in Engineering, No. 2015, 980945, P. 1-15. **DOI**: https://doi.org/10.1155/2015/980945

9. *El Kafhali, S., Salah, K.* (2017), "Stochastic Modelling and Analysis of Cloud Computing Data Center", Proceedings of the *20th Conference on Innovations in Clouds, Internet and Networks (ICIN)*, Paris, France, 7-9 March 2017, P .122-126. **DOI**: https://doi.org/10.1109/ICIN.2017.7899401

10. *Faychuk, V., Lavriv, O., Stryhalyuk, B., Shpur, O., Demydov, I., Bak, R.* (2019), "Performance of routing algorithm remote operation in cloud environment for IoT devices", International Journal of Electronics and Telecommunications, No. 65(3), P. 367–373. **DOI**: https://doi.org/10.24425/ijet.2019.129787

11. *Klymash, M., Shpur, O., Peleh, N., Lutsiuk, I.* (2018), "Clustering model of cloud centers for Big Data processing", Proceedings of the 2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), Lviv, Slavske, 20–24 February 2018, P. 268–281. **DOI**: https://doi.org/10.1109/TCSET.2018.8336200