DOI: 10.35681/1560-9189.2020.22.4.225913

УДК 004.32

О. Я. Матов
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Методи і аналітичні умови адаптації надання ресурсів користувачам хмарних обчислень

Розглянуто питання подальшого розвитку засад створення адаптивних інфраструктур хмарних обчислень, здатних динамічно адаптуватися до вимог користувачів і діючих особливостей і змін умов функціонування. Розроблено методи і аналітичні умови адаптації надання ресурсів користувачам хмарних обчислень. Ці умови надають можливість розробляти технологію (механізми і алгоритми) використання адаптивної дисципліни (порядку) надання обчислювальних ресурсів користувачам. У свою чергу, це дозволяє забезпечувати часові вимоги різних користувачів на отримання своєчасних результатів обчислень чи найбільш ефективно використовувати наявні ресурси хмарних обчислень. Це є актуальним для систем реального масштабу часу і, в першу чергу, для спеціальних інформаційних систем, що побудовані з використанням приватних хмар, і може бути критичним при обмежених обчислювальних ресурсах хмарних обчислень.

Ключові слова: *хмарні обчислення, математична модель, дисципліна надання обчислювальних ресурсів, змішана дисципліна обслуговування, абсолютний і відносний пріоритети, часові характеристики, час відгуку, ефективність хмарних обчислень.*

Вступ

Створення адаптивних інфраструктур**,** які здатні адаптуватися до змін умов функціонування та підтримувати системи в оптимальному, а іноді просто в працездатному стані, є важливим напрямком розвитку сучасних глобальних інформаційно-аналітичних систем із застосуванням технологій хмарних обчислень (ХО). Для такої адаптації запропоновано динамічну адаптивну змішану дисципліну надання обчислювальних ресурсів користувачам ХО [1, 4, 5].

Аналітичні (формульні) умови адаптації розробляються на основі відповідних показників ефективності та математичних моделей хмарних обчислень. Стохастичний характер головних чинників і необхідність кількісної оцінки масових

процесів на основі теорії імовірності обумовлює використання аналітичної моделі хмарних обчислень як багатопотокової і багатопріоритетної системи масового обслуговування з чергами зі змішаною дисципліною обслуговування. Модель враховує вірогідні відмови й різні особливості та має довільні закони розподілу для деяких вірогідних процесів. Модель дозволяє вираховувати часову характеристику — час відгуку системи в умовах особливостей функціонування та відмов хмарних обчислень. Стаття є продовженням циклу робіт про адаптивне управління надання ресурсів користувачам хмарних обчислень. Деякі результати попередніх статей використано у цій статті.

Розглянемо дві практичні задачі динамічної адаптації змішаної дисципліни надання ресурсів з відносно-абсолютними пріоритетами. Одними із основних показників ефективності ХО є ті, що базуються на оцінці часових характеристик цих систем і які необхідно підтримувати на заданому рівні. Такі показники можуть задаватися договором між постачальником і користувачем ХО та здобувають особливе значення для систем, що функціонують у реальному масштабі часу і, в першу чергу, для спеціальних інформаційних систем на базі приватних хмар. Унаслідок випадкового характеру обчислювального процесу виникають додаткові затримки в обробці інформації, порушуються припустимі обмеження на час її перебування в ХО, що негативно позначається на ефективності рішення цільових задач користувачів.

Для забезпечення необхідної ефективності ХО в таких ситуаціях необхідно підтримувати часові характеристики системи на заданому рівні. В умовах дефіциту обчислювальних ресурсів це можливо тільки за рахунок підвищення ефективності обчислювального процесу, зокрема, за рахунок адаптації дисципліни обслуговування.

Поряд із цим виникає задача найбільш ефективного використання наявних обчислювальних ресурсів у кожен момент часу функціонування керуючих ХО. Цю задачу також можна вирішити шляхом адаптації дисципліни обслуговування.

Показники ефективності надання ресурсів користувачам хмарних обчислень

Метою роботи є розробка методів і аналітичних умов адаптації надання обчислювальних ресурсів користувачам ХО задля забезпечення часових характеристик інформаційно-аналітичних систем та оптимізації використання ресурсів ХО.

За показник ефективності ХО беремо середню сумарну вартість (штрафу) часу відгуку ХО (часу затримки в ХО, часу очікування в чергах та часу надання ресурсів, тобто перебування в ХО як у системі масового обслуговування (СМО)) на заявки (вимоги) користувачів. Для цього використаємо відомий функціонал [5]

$$C^{(S)} = \sum_{i=1}^{n} \alpha_i \lambda_i v_i^{(s)},$$

з чого маємо

$$C^{(\varphi)} = \sum_{m=1}^{M} \sum_{n=1}^{N} \alpha(m,n)\lambda(m,n)v^{(\varphi)}(m,n), \qquad (1)$$

де α_i — вартість (штраф) за одиницю часу відгуку ХО (затримки, перебування в ХО) заявок *i*-го потоку;

λ_i — інтенсивність *i*-го потоку заявок;

$v_i^{(S)}$ — середній час відгуку ХО заявок *i*-го потоку;

n — кількість типів заявок;

s — параметр, що характеризує спосіб організації обчислювального процесу;

$v^{(\varphi)}(m,n)(m=\overline{1,M},n=\overline{1,N_m})$ — середній час відгуку ХО заявок (*m*, *n*)-го потоку;

α (*m*, *n*) — вартість одиниці часу відгуку ХО (затримки в ХО) заявок (*m*, *n*)-го потоку;

λ (*m*, *n*) — інтенсивність (*m*, *n*)-потоку.

Цей показник ефективності базується на припущенні, що результати використання ресурсів користувачем знецінюються пропорційно часу їхньої затримки в ХО, тобто перебування в ХО як у СМО. Тоді цілями адаптації змішаної дисципліни обслуговування будуть або задоволення вимог вчасного перебування (*m*, *n*) заявок у системі, що задаються припустимими значеннями цього часу $v_Д(m,n)$, або мінімізація функціоналу (1). Ці цілі досягаються шляхом пошуку відповідних оптимальних розбивок φ^0 на відносні та абсолютні пріоритети, тобто задачі адаптації змішаної дисципліни обслуговування з відносно-абсолютним пріоритетом являють собою оптимізаційні задачі, загальну постановку яких розглянуто вище.

Оскільки сформульовані вище цілі адаптації змішаної дисципліни обслуговування можуть бути досягнуті при декількох різних розбивках потоків заявок на групи абсолютного пріоритету, то виникає необхідність уведення додаткового обмеження на вибір розбивки φ.

Наявність у ХО абсолютного пріоритету потребує деяких технологічних втрат ресурсів, які пропорційні числу груп (рівнів) абсолютного пріоритету. У зв'язку із цим оптимальною необхідно вважати таку розбивку, яка забезпечує досягнення цілей адаптації при мінімальній кількості груп абсолютного пріоритету *M*.

Тоді розглянуті задачі адаптації змішаної дисципліни обслуговування можуть бути формально поставлені в такий спосіб:

$$\begin{gathered} v^{(\varphi)}(m,n) \le v_Д(m,n) \Rightarrow \varphi^0, \\ \varphi \in \Phi, \\ M = \min, \end{gathered} \tag{2}$$

$$\begin{gathered} C^{(\varphi)} \to \min \Rightarrow \varphi^0, \\ \varphi \in \Phi, \\ M = \min. \end{gathered} \tag{3}$$

Рішення задач пошуку оптимальної розбивки (2) і (3) за допомогою відомих аналітичних методів оптимізації не представляється можливим. Єдиний шлях рішення цих задач — евристичний підхід, що не має формального обґрунтування, а

спирається лише на специфіку задач (математичних моделей) і зв'язані з ними розуміння.

Із виразів (1)–(3) випливає, що досягнення цілей адаптації змішаної дисципліни обслуговування сполучено з потребою оцінки значення середнього часу відгуку ХО (перебування в ХО) заявок (m, n)-типу $\nu(m, n)$ на ресурси. Тому виникає необхідність синтезу математичної моделі ХО зі змішаною дисципліною надання обчислювальних ресурсів (обслуговування).

Клас моделі хмарної інфраструктури

Розробка математичних моделей хмарних обчислень чи інформаційних систем, що створені з використанням хмар, є важливим напрямком для виявлення та покращення їхніх характеристик [2, 3, 5–9]. Хмарні обчислення є об'єктами з високим рівнем невизначеності процесу функціонування. Тут зовнішню невизначеність потоку запитів на обчислювальні ресурси (ОР) (середовища) доповнює внутрішня невизначеність ХО (об'єкта), що пов'язана з наявністю чи відсутністю необхідних ОР, випадкових відмов системи ХО, а також необхідність забезпечення певних часових характеристик для багатьох клієнтів. Саме це визначає необхідність введення адаптації у процес функціонування ХО.

Крім того, введення адаптації у процес функціонування ХО пов'язано з необхідністю підтримки системи в оптимальному, а іноді й просто працездатному стані незалежно від численних факторів зовнішнього та внутрішнього характеру, що виводять ХО з необхідного цільового стану.

Хмарні обчислення є об'єктами з високим рівнем невизначеності процесу функціонування, головними чинниками якої є [1, 4]:

— вірогідність потоку запитів на обчислювальні ресурси (ОР);

— наявність необхідних ОР і випадковість часу їхнього використання клієнтами;

— випадковість відмов інфраструктури ХО та часу їхнього усунення;

— необхідність забезпечення певних часових характеристик для ряду клієнтів, наприклад, часу відгуку ХО;

— необхідність оптимального використання ОР залежно від вартості часу затримки замовлених клієнтами результатів обчислень та умов функціонування;

— необхідність введення адаптації у процес функціонування ХО з метою забезпечення певних часових характеристик для ряду клієнтів та оптимального використання ОР.

Стохастичний характер головних чинників і необхідність кількісної оцінки масових процесів на основі теорії імовірності обумовлює використання теорії масового обслуговування. Тоді як механізми адаптації ХО можливо та доцільно використовувати технологію динамічної адаптивної змішаної дисципліни надання ОР (обслуговування) користувачам ХО [1, 4].

Пропонуються аналітичні моделі для вирахування часових характеристик в умовах особливостей функціонування ХО з використанням змішаної дисципліни обслуговування з абсолютними і відносними пріоритетами та врахуванням відмов. Моделі базуються на роботах [2, 3, 6].

Математичний опис багатопотокової і багатопріоритетної моделі функціонування хмарної інфраструктури з чергами зі змішаною дисципліною обслуговування та адаптацією до відмов

Нехай на вхід системи ХО, в якій реалізована дисципліна обслуговування з відносно-абсолютним пріоритетом, надходять N пуасоновських потоків заявок інтенсивності $\lambda(m,n)$ $m=\overline{1,M}$, $n=\overline{1,N_m}$. Цим потокам поставлені у відповідність N пріоритетів.

Тривалість обслуговування заявок пріоритету (m, n) є випадковою величиною з функцією розподілу $B_{m,n}(t)$, першим $b(m, n)$ і другим $b^{(2)}(m,n)$ початковими моментами.

Заявка пріоритету (m, n), обслуговування якої перерване заявками з груп з номерами $\overline{1,m-1}$, повертається в чергу. Поновлення її обслуговування можливо або після обслуговування всіх заявок, що були перервані (дисципліна поновлення обслуговування A), або після обслуговування всіх заявок, які переривають, і всіх заявок потоків, що накопичилися, m-ї групи з номерами $\overline{(m,1),(m,n-1)}$ (дисципліна поновлення обслуговування B).

Обслуговуючий прилад (ХО) виходить з ладу за пуасонівським законом з параметром λ_0. Період відновлення приладу є випадковою величиною, що має довільний закон розподілу $B_0(t)$ з першим b_0 і другим b_0^2 початковими моментами.

У період відновлення обслуговуючого приладу запити одних потоків у чергу приймаються, а інші — не приймаються. Ця умова задається матрицею-рядком коефіцієнтів $n_i, i=\overline{1,N}$, причому $n_i=1$ у тому випадку, якщо запити i-го потоку в чергу приймаються, і $n_i=0$, якщо запити одержують відмовлення.

Адаптацією до відмов буде те, що в період відновлення приладу вхідні заявки можуть або накопичуватися в черзі (дисципліна поповнення черги I), або одержувати відмовлення та залишати систему (дисципліна поповнення черги II).

Вихід з ладу обслуговуючого приладу може відбутися як під час його вільного стану, так і під час обслуговування заявки. В останньому випадку поновлення обслуговування здійснюється або з перерваної заявки, якщо немає заявок, що переривають її обслуговування (дисципліна поновлення обслуговування C), або із заявок старшого відносного пріоритету відповідної групи, якщо такі є (дисципліна поновлення обслуговування D).

При повторному надходженні на обслуговуючий прилад перервана заявка дообслуговується з місця її переривання. У межах одного пріоритету заявки обслуговуються в порядку надходження.

Сполучення дисциплін поновлення обслуговування та поповнення черги дозволяє розглядати самостійні моделі різних типів систем, що мають відповідне позначення. Різні особливості функціонування складаються з різних сполучень дисциплін A, B, C, D, I і II.

Нехай ХО знаходиться в стаціонарному режимі, умовою якого для систем типу I є $R_M < K_r$, а для систем типу II — $R_M < 1$. Тут $R_M = \sum_{m=1}^{M}\sum_{n=1}^{N}\rho(m,n)$ — су-

марне завантаження приладу заявками ($\rho(m,n) = \lambda(m,n)b(m,n)$ — завантаження приладу (m, n)-заявками), а $K_r = 1/(1+\rho_0)$ — коефіцієнт готовності системи ($\rho_0 = \lambda_0 b_0$ — завантаження приладу відмовленнями).

Необхідно визначити середній час перебування в системі заявок кожного (m, n)-пріоритету $v(m,n)$, тобто час відгуку системи ХО.

Визначення часових характеристик моделі системи типу АС-I

Для визначення середнього часу перебування заявок у системі (часу відгуку системи) типу АС-I скористаємося відомим прямим методом [5].

Нехай у систему надходить деяка заявка (j, k)-пріоритету ($j \in M$, $k \in N_j$). Середній час перебування цієї заявки в системі $v(j, k)$ складається із середнього часу очікування в черзі $w(j, k)$ і середнього часу обслуговування b (j, k):

$$v(j,k) = w(j,k) + b(j,k).$$

Середній час очікування в черзі $w(j, k)$ складається із середнього часу очікування до початку обслуговування $w_H(j,k)$ і середнього часу очікування в перерваному стані $u(j, k)$:

$$w(j,k) = w_H(j,k) + u(j,k).$$

Останній доданок у цій формулі обумовлено перериваннями обслуговування заявки (j, k)-пріоритету заявками з груп $\overline{1, j-1}$ і відмовленнями, тобто:

$$u(j,k) = u_3(j,k) + u_0(j,k).$$

Середній час від початку обслуговування (j, k)-заявки до завершення є середній повний час обслуговування:

$$\Theta(j,k) = b(j,k) + u(j,k). \qquad (4)$$

Почнемо з обчислення u (j, k), для чого застосуємо підхід, описаний у [5].

За час обслуговування (j, k)-заявки в середньому відбудеться $b(j,k)\Lambda_{j-1}$ переривань, де $\Lambda_{j-1} = \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\lambda(m,n)$ — інтенсивність сумарного потоку заявок, що переривають.

У результаті цих переривань (j, k)-заявка повертається в чергу і очікує закінчення обслуговування заявок, які переривають, що буде тривати в $b(j,k)R_{j-1}$ середньому одиниць часу, де

$$R_{j-1} = \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\lambda(m,n)b(m,n). \qquad (5)$$

За цей час ще надійдуть заявки з груп $\overline{1, j-1}$, обслуговування яких приведе до збільшення часу очікування (j, k)-заявки на величину $b(j,k)R_{j-1}^2$. Крім того, обслуговування цих заявок буде супроводжуватися додатковим накопиченням

заявок тих же пріоритетів, які вимагають обслуговування раніше (j, k)-заявки. Цей процес нескінченний, причому добавки до часу очікування (j, k)-заявки утворюють спадну геометричну прогресію зі знаменником $R_{j-1} < 1$. Сума членів такої геометричної прогресії являє собою середній час усіх перерв обслуговування (j, k)-заявки:

$$T^{(1)} = b(j,k)\frac{R_{j-1}}{1-R_{j-1}}. \tag{6}$$

За час $T^{(1)}$ у середньому відбудеться $T^{(1)}\lambda_0$ відмовлень приладу, у результаті чого він буде відновлюватися протягом $T^{(1)}\lambda_0 b_0 = T^{(1)}\rho_0$ одиниць часу. Оскільки у системі типу АС-І у період відновлення приладу знову надходять заявки, які продовжують накопичуватися в черзі, то після відновлення приладу середній час очікування (j, k)-заявки в перерваному стані збільшиться на

$$T^{(2)} = T^{(1)}\rho_0\frac{R_{j-1}}{1-R_{j-1}} = b(j,k)\rho_0\frac{R_{j-1}^2}{(1-R_{j-1})^2}. \tag{7}$$

За цей час можуть відбутися відмовлення приладу, відновлення якого буде супроводжуватися нагромадженням нових заявок, що обслуговуються раніше (j, k)-заявки і т.д.

Сумарний час усіх перерв обслуговування заявки (j, k)-пріоритету заявками груп $\overline{1, j-1}$ з урахуванням відмовлень приладу $u_3(j,k) = T^{(1)} + T^{(2)} + \ldots + T^{(\infty)}$. Цей вираз являє собою суму двох нескінченно спадних геометричних прогресій. Обчисливши суму членів кожної з них і склавши результати, отримаємо:

$$u_3(j,k) = b(j,k)\frac{R_{j-1}}{K_r - R_{j-1}}. \tag{8}$$

Аналогічним чином визначається середній час очікування (j, k)-заявки в перерваному стані за рахунок відмовлень приладу $u_0(j,k)$. Відрізняється лише початок міркувань. За час обслуговування (j, k)-заявки в середньому відбудеться $b(j,k)\lambda_0$ відмовлень приладу, що приведе до його відновлення протягом $b(j,k)\rho_0$ одиниць часу. З урахуванням можливості накопичення в період відновлення приладу та першочергового обслуговування заявок з абсолютним пріоритетом із групи $\overline{1, j-1}$ середній час очікування (j, k)-заявки збільшиться на

$$b(j,k)\rho_0\frac{R_{j-1}}{1-R_{j-1}}.$$

За цей час знову можуть відбутися відмовлення приладу, що додатково збільшують час очікування (j, k)-заявки на величину

$$b(j,k)\rho_0^2\frac{R_{j-1}}{1-R_{j-1}} \text{ і т.д.}$$

В остаточному підсумку отримаємо:

$$u_0(j,k) = b(j,k)\frac{K_r\rho_0}{K_r - R_{j-1}}. \tag{9}$$

Тоді сумарний середній час очікування (j, k)-заявки в перерваному стані буде

$$u(j,k) = b(j,k)\frac{R_{j-1} + K_r\rho_0}{K_r - R_{j-1}}, \tag{10}$$

а повний середній час обслуговування (j, k)-заявки відповідно

$$\Theta(j,k) = b(j,k)\frac{1}{K_r - R_{j-1}}. \tag{11}$$

Тепер обчислимо $w_H(j,k)$. Перед тим як (j, k)-заявка надійшла в систему перший раз, повинно бути виконане наступне:

1) відновлено прилад;

2) обслуговано заявку з $\overline{1,j}$ груп чи заміщено заявку, що обслуговується, із $\overline{j+1,M}$ груп;

3) дообслуговано заявки з $\overline{2,j}$ груп, перервані заявками з $\overline{1,j-1}$ груп;

4) дообслуговано заявки з $\overline{1,j}$ груп, перервані відмовленнями приладу;

5) обслуговано наявні заявки потоків з номерами $\overline{(1,1),(j,k)}$;

6) обслуговано заявки потоків з номерами $\overline{(1,1),(j,k-1)}$, що надійшли за час очікування (j, k)-заявки з урахуванням відмовлень приладу.

Для середніх тривалостей зазначених подій запишемо рівняння:

$$\begin{aligned} w_H(j,k) &= \sigma_0 + \sigma(j,k) + \eta(j,k) + \eta_0(j,k) + \\ &+ \sum_{m=1}^{j-1}\sum_{n=1}^{N_m} w_H(m,n)\rho(m,n) + \sum_{n=1}^{k} w_H(j,n)\rho(j,n) + \\ &+ [\sigma_0 + z_H(j,k)]\frac{R_{j,k-1}}{K_r - R_{j,k-1}} + z_H(j,k)\frac{K_r\rho_0}{K_r - R_{j,k-1}}. \end{aligned} \tag{12}$$

Тут $\sigma_0 = K_r\rho_0\Delta_0$ — середній час довідновлення приладу в присутності (j, k)-заявки: $K_r\rho_0$ — імовірність відновлення приладу, $\Delta_0 = b_0^{(2)}/2b_0$;

$\sigma(j,k) = \sum_{m=1}^{j}\sum_{n+1}^{N_m}\rho(m,n)\Delta(m,n)$ — середній час дообслуговування заявки приладом у присутності (j, k)-заявки: $\Delta(m,n) = b^{(2)}(m,n)/2b(m,n)$;

$\eta(j,k) = \sum_{m=2}^{j}\sum_{n=1}^{N_m}\frac{R_{m-1}}{K_r - R_{m-1}}\rho(m,n)\Delta(m,n)$ — середній час дообслуговвування заявок із груп $\overline{2,j}$, перерваних заявками з груп $\overline{1,j-1}$, де $\frac{R_{m-1}}{K_r - R_{m-1}}\rho(m,n)$ —

імовірність перебування в черзі (m, n)-заявки, перерваної заявками з груп $\overline{1,m-1}$. Ця ймовірність визначається за формулою (8) з урахуванням інтенсивності $\lambda(m,n)$ потоку (m, n)-заявок;

$\eta_0(j,k)=\sum_{m=1}^{j}\sum_{n=1}^{N_m}\frac{K_r\rho_0}{K_r-R_{m-1}}\rho(m,n)\Delta(m,n)$ — середній час дообслуговування заявок із груп $\overline{1,j}$, перерваних відмовленнями приладу, де $\frac{K_r\rho_0}{K_r-R_{m-1}}\rho(m,n)$ — імовірність того, що в черзі мається (m, n)-заявки, перервані відмовленням приладу. Ця ймовірність визначається на основі (9) з обліком $\lambda(m,n)$;

$z_H(j,k)$ — середній час очікування (j, k)-заявки, рівний сумі розглянутих складових без обліку σ_0;

$$R_{j,k-1}=\sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\rho(m,n)+\sum_{n=1}^{k-1}\rho(j,n).$$

Помітимо, що в кожній черзі може бути не більше однієї заявки, перерваної заявками з абсолютним пріоритетом чи відмовленням.

Після нескладних перетворень з рівняння (12) отримуємо наступне рекурентне співвідношення:

$$w_H(j,k)=\frac{1}{K_r-R_{j,k}}\Bigg[K_r^2\rho_0\Delta_0+\sum_{m=1}^{j}\sum_{n=1}^{N_m}\frac{1}{K_r-R_{m-1}}\times \\ \times\rho(m,n)\Delta(m,n)+\sum_{m=1}^{j-1}\sum_{n=1}^{N_m}w_H(m,n)\rho(m,n)+\sum_{n=1}^{k-1}w_H(j,n)\rho(j,n)\Bigg], \tag{13}$$

де $R_{j,k}=\sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\rho(m,n)+\sum_{n=1}^{k}\rho(j,n)$.

Щоб одержати формулу для визначення $w_H(j,k)$ в явному вигляді, проаналізуємо співвідношення (13) для «чистих» дисциплін обслуговування з відносним і абсолютним пріоритетом.

Для дисципліни обслуговування з відносним пріоритетом ($M=1,\ N_1=N$) отримаємо:

— для першого потоку

$$w_H(1,1)=\frac{K_r^3\rho_0\Delta_0+\sum_{n=1}^{N_1}\rho(1,n)\Delta(1,n)}{K_r[K_r-\rho(1,1)]};$$

— для другого потоку

$$w_H(1,2)=\frac{K_r^3\rho_0\Delta_0+\sum_{n=1}^{N_1}\rho(1,n)\Delta(1,n)}{[K_r-\rho(1,1)]\times[K_r-\rho(1,1)-\rho(1,2)]}.$$

Ці формули дозволяють припустити загальне рішення у вигляді

$$w_H(1,k) = \frac{K_r^3 \rho_0 \Delta_0 + \sum_{n=1}^{N_1} \rho(1,n)\Delta(1,n)}{(K_r - R_{1,k-1})(K_r - R_{1,k})}, \qquad (14)$$

де $R_{1,k-1} = \sum_{n=1}^{k-1} \rho(1,n),\ R_{1,k} = \sum_{n=1}^{k} \rho(1,n).$

Для дисципліни обслуговування з абсолютним пріоритетом ($M = N,\ N_m = 1$ для усіх $m = \overline{1,M}$) з виразу (13) отримуємо:

— для потоку першої групи

$$w_H(1,1) = \frac{K_r^3 \rho_0 \Delta_0 + \rho(1,1)\Delta(1,1)}{K_r[K_r - \rho(1,1)]};$$

— для потоку другої групи

$$w_H(2,1) = \frac{K_r^3 \rho_0 \Delta_0 + \rho(1,1)\Delta(1,1) + \rho(2,1)\Delta(2,1)}{[K_r - \rho(1,1)][K_r - \rho(1,1) - \rho(2,1)]}.$$

Тоді на основі цих рівностей одержимо загальний вираз:

$$w_H(j,1) = \frac{K_r^3 \rho_0 \Delta_0 + \sum_{m=1}^{j} \rho(m,1)\Delta(m,1)}{(K_r - R_{j-1,1})(K_r - R_{j,1})}, \qquad (15)$$

де $R_{j-1,1} = \sum_{m=1}^{j-} \rho(m,1),\ R_{j,1} = \sum_{m=1}^{j} \rho(m,1)$.

Аналізуючи вирази (14) і (15), неважко припустити загальний вид формули для визначення $w_H(j,k)$ для змішаної дисципліни обслуговування:

$$w_H(j,k) = \frac{K_r^3 \rho_0 \Delta_0 + \sum_{m=1}^{j} \sum_{n=1}^{N_m} \rho(m,n)\Delta(m,n)}{(K_r - R_{j,k-1})(K_r - R_{j,k})}. \qquad (16)$$

Підставивши формулу (16) у (13) і зробивши нескладні перетворення, можна переконатись у справедливості такого припущення.

За виразами (11) і (16) обчислюємо шуканий середній час перебування (j, k)-заявки $v(j, k)$ у системі типу АС-I.

Аналогічно, як і для систем типу АС-I, можливо вивести формули для визначення часових характеристик для інших систем типу АС-II, BD-I, BD-II. Ці моделі враховують такі фізичні властивості ХО як миттєву еластичність (динамічне виділення та звільнення ресурсів для швидкого масштабування відповідно до потреб) і вимірювальний сервіс (керування й оптимізація ресурсів за допомогою засобів вимірювання).

Методи та аналітичні умови адаптації надання ресурсів користувачам хмарних обчислень

Адаптація змішаної дисципліни обслуговування з моделлю ХО полягає в пошуку оптимальної розбивки потоків заявок за групами (рівнями) абсолютного пріоритету (φ^0), тобто такої сукупності чисел $\{N_m\} m=\overline{1,M}$, при якій тимчасові характеристики моделі ХО забезпечували би відповідно до задачі (6) рівність

$$\varphi^0 = opt\{N_1, N_2, \ldots, N_M / \nu^{(\varphi)}(m,n) \le \nu_{Д}(m,n), \varphi \in \Phi, M = \min\}, \qquad (17)$$

а відповідно до задачі (7) рівність

$$\varphi^0 = opt\{N_1, N_2, \ldots, N_M / C^{(\varphi)} = \min, \varphi \in \Phi, M = \min\}. \qquad (18)$$

Оскільки кількість потоків заявок N є скінченною, то задача пошуку оптимальної розбивки φ^0 може бути вирішена шляхом повного перебору усіх можливих розбивок і вибору з них такої, котра задовольняє рівностям (17) і (18). Однак цей шлях для ХО реального масштабу часу неприйнятний, тому що кількість усіх можливих розбивок $\Phi = 2^{N-1}$ при великих N — велика, і реалізація методу повного перебору вимагає значних часових витрат. Тому виникає необхідність розробки таких методів адаптації, що дозволяють одержувати оптимальну розбивку в результаті розгляду обмеженої кількості варіантів групування.

Для пошуку розбивки φ^0, що забезпечує рівність (17), пропонується метод, суть якого полягає в почерговому задоволенні вимог до часу перебування заявок у системі, починаючи з першого потоку, шляхом послідовного формування спочатку першої, потім другої і т.д. груп абсолютного пріоритету. При цьому процес адаптації починається з розбивки, що відповідає дисципліні обслуговування з «чистим» відносним пріоритетом (M = 1, N_1 = N). У зв'язку з цим перша з розбивок, за якої виконується ціль адаптації, характеризується мінімально можливою кількістю груп абсолютного пріоритету M, тобто є оптимальною.

Для пошуку розбивки φ^0, що задовольняє рівності (18), пропонується метод адаптації, суть якого полягає в цілеспрямованому формуванні груп абсолютного пріоритету, починаючи з останньої, на основі аналізу знаку прирощення величини середньої сумарної вартості перебування заявок у системі $\Delta C^{(\varphi)}$. При формуванні чергової групи заявки потоків сформованих груп виключаються з розгляду, оскільки вони не впливають на середній час перебування в системі заявок потоків попередніх груп абсолютного пріоритету. Процес адаптації у цьому випадку починається з розбивки, що відповідає дисципліні обслуговування з «чистим» абсолютним пріоритетом (M = N, N_m = 1), що також забезпечує мінімальне число груп M при виконанні мети адаптації.

Визначимо $\Delta C^{(\varphi)}$. Припустимо, що попередня *q-розбивка* має вигляд $N_1 = N_2 = \ldots = N_{l+1} = 1, \quad N_{l+2} = N_j = P - l - 1,$ де P — кількість потоків заявок, розглянутих на етапі формування чергової групи з номером j при нумерації груп з останньої $P = N - \sum_{m=1}^{j-1} N_m$. Наступна φ-розбивка відрізняється від *q-розбивки* тим,

що потоки заявок двох останніх груп об'єднані в одну, тобто $N_1 = N_2 = \ldots = N_l = 1, \quad N_{l+1} = N_j = P - l$.

$\Delta C^{(q,\varphi)}$ при переході від *q-розбивки* до φ-розбивки визначимо в такий спосіб:

$$\Delta C^{(q,\varphi)} = C^{(\varphi)} - C^{(q)} = \sum_{i=1}^{N} \alpha_i \lambda_i \Delta \nu_i^{(q,\varphi)}, \tag{19}$$

де $\Delta \nu_i^{(q,\varphi)} = \nu_i^{(\varphi)} - \nu_i^{(q)}, \quad i = \overline{1,N}$.

З формули (19) випливає, що φ-розбивка вважається кращою порівняно з *q-розбивкою*, якщо $\Delta C^{(q,\varphi)} \le 0$. У випадку $\Delta C^{(q,\varphi)} > 0$ кращою є *q-розбивка*. Вираз $\Delta C^{(q,\varphi)} = 0$ означає, що φ-розбивка за критерієм середньої сумарної вартості перебування заявок у системі не гірше *q-розбивки*, але забезпечує меншу кількість груп абсолютного пріоритету *M*.

Обчислимо $\Delta C^{(q,\varphi)}$ на прикладі системи типу АС-I, для якої на підставі виразів (11) і (16) можна записати:

$$\nu_i^{(\varphi)} = \begin{cases} \dfrac{b_i}{K_r - R_{i-1}} + \dfrac{K_r^3 \rho_0 \Delta_0 + \sum\limits_{r=1}^{i} \rho_r \Delta_r}{(K_r - R_{i-1})(K_r - R_i)}, & i = \overline{1,l}; \\ \dfrac{b_i}{K_r - R_l} + \dfrac{K_r^3 \rho_0 \Delta_0 + \sum\limits_{r=1}^{P} \rho_r \Delta_r}{(K_r - R_{i-1})(K_r - R_i)}, & i = \overline{l+1,P}. \end{cases} \tag{20}$$

При переході від *q-розбивки* до φ-розбивки збільшення середнього часу перебування заявок у системі $\Delta \nu_i^{(q,\varphi)}$:

$$\Delta \nu_i^{(q,\varphi)} = \begin{cases} 0, & i = \overline{1,l}; \\ \dfrac{\sum\limits_{r=l+2}^{P} \rho_r \Delta_r}{(K_r - R_l)(K_r - R_{l+1})}, & i = l+1; \\ -\dfrac{b_i \rho_{l+1}}{(K_r - R_l)(K_r - R_{l+1})}, & i = \overline{l+2,P}. \end{cases} \tag{21}$$

Тоді збільшення $\Delta C^{(q,\varphi)}$ запишемо у вигляді

$$\Delta C^{(q,\varphi)} = \alpha_{l+1} \lambda_{l+1} \Delta \nu_{l+1}^{(q,\varphi)} + \sum_{i=l+2}^{P} \alpha_i \lambda_i \Delta \nu_i^{(q,\varphi)} = \\ = \frac{\rho_{l+1}}{2(K_r - R_l)(K_r - R_{l+1})} \sum_{i=l+2}^{P} \lambda_i b_i^{(2)} \left(\frac{\alpha_{l+1}}{b_{l+1}} - \frac{2}{1+\psi_i^2} \frac{\alpha_i}{b_i} \right), \tag{22}$$

де $\psi_i = \sqrt{D[t_i]} / b_i$ — коефіцієнт варіації часу обслуговування заявок *i*-го потоку ($D[t_i]$ — дисперсія часу обслуговування). При показовому обслуговуванні заявок *i*-го потоку $\psi_i = 1$, а при детермінованому обслуговуванні — $\psi_i = 0$.

Аналіз виразу (22) показує, що доцільність переходу від *q-розбивки* до φ-розбивки визначається знаком вхідної в нього суми:

$$\Delta C_l = \sum_{i=l+2}^{P} \lambda_i b_i^{(2)} \left(\frac{\alpha_{l+1}}{b_{l+1}} - \frac{2}{1+\psi_i^2} \frac{\alpha_i}{b_i} \right). \qquad (23)$$

З цієї рівності випливає:

1) якщо $\Delta C_l \le 0$ для усіх $l = \overline{0, N-2}$, то оптимальною є дисципліна обслуговування з «чистим» відносним пріоритетом;

2) при $\Delta C_l > 0$ для всіх *l* оптимальною є дисципліна обслуговування з «чистим» абсолютним пріоритетом;

3) у випадку $\Delta C_l \le 0$ або $\Delta C_l > 0$ не для всіх *l* оптимальною є змішана дисципліна обслуговування.

Таким чином, для визначення доцільності переходу від *q-розбивки* до φ-розбивки досить за формулою (23) обчислити ΔC_l і проаналізувати результат. Зміна розбивки доцільна, якщо $\Delta C_l \le 0$.

Висновки

1. Виконано подальший розвиток засад створення адаптивних інфраструктур хмарних обчислень, здатних динамічно адаптуватися до вимог користувачів та діючих особливостей і змін умов функціонування. Цей науковий напрямок залишається актуальним і вимагає подальших досліджень.

2. Хмарні центри обробки даних є об'єктами з високим рівнем випадковості процесу функціонування, головними чинниками якої є: ймовірність потоку запитів на обчислювальні ресурси; наявність необхідних ресурсів і випадковість часу їхнього використання споживачами; випадковість відмов інфраструктури та часу їхнього усунення.

Унаслідок випадкового характеру обчислювального процесу виникають додаткові затримки в обробці інформації, порушуються припустимі обмеження на час її перебування в системі (на час відгуку системи), що негативно позначається на ефективності рішення цільових задач користувачів. Це є актуальним для систем реального масштабу часу і, в першу чергу, для спеціальних інформаційних систем, що побудовані із використанням приватних хмар, і може бути критичним при обмежених обчислювальних ресурсах ХО.

3. Позбавитися чи зменшити вплив вірогідних явищ на функціонування можливо введенням адаптації у процес функціонування інфраструктури ХО. Крім того, введення адаптації пов'язано з необхідністю підтримки ХО в оптимальному (ефективному використанню ресурсів), а іноді й просто працездатному стані незалежно від численних факторів, що виводять інфраструктуру ЦОД із необхідного цільового стану. Метою адаптації може бути максимізація доходу від обслугову-

вання користувачів, ліквідація перевантаження системи та підтримки її в стаціонарному режимі функціонування.

4. Задача адаптації може бути вирішена використанням адаптивної дисципліни (порядку) надання обчислювальних ресурсів користувачам. Непередбачені та неконтрольовані зміни в середовищі та системі неминуче змінюють оптимальне настроювання дисципліни, якщо така була в системі реалізована. Тому систематичні підстроювання (адаптація) дисципліни неминучі при бажанні підтримувати систему в оптимальному режимі незалежно від змін, що відбуваються в середовищі та системі.

Задля адаптації використано динамічну адаптивну змішану дисципліну з абсолютно-відносними пріоритетами надання обчислювальних ресурсів користувачам хмарних обчислень, один із варіантів створення технології якої розглянуто автором у [1, 4]. При цьому адаптація дисципліни складається з оптимальної зміни кількості та положення границь, що розділяють потоки заявок користувачів на ресурси на групи абсолютного пріоритету, всередині яких діє відносний пріоритет, тобто в зміні кількості груп і кількості потоків у групах.

5. Розробку аналітичних умов адаптації надання ресурсів користувачам хмарних обчислень виконано на основі аналітичної моделі. Аналітичні умови дозволяють розробляти механізми і алгоритми адаптації ХО, які враховують такі фізичні властивості ХО як миттєву еластичність (динамічну міграцію, виділення та звільнення ресурсів для швидкого масштабування відповідно до потреб) і сервіси вимірювання (керування і оптимізація ресурсів за допомогою засобів вимірювання). Момент включення алгоритму адаптації визначається системою управління ХО при порушенні припустимих обмежень на час відгуку, зміні контрольованих параметрів системи (наприклад, її сумарного завантаження) чи показника ефективності системи понад граничні значення.

6. Стохастичний характер головних чинників і необхідність кількісної оцінки масових процесів на основі теорії імовірності обумовлює використання аналітичної моделі хмарних обчислень як багатопотокової та багатопріоритетної системи масового обслуговування зі змішаною дисципліною обслуговування. Модель враховує вірогідні відмови і різні особливості та має довільні закони розподілу для деяких вірогідних процесів. Модель дозволяє вираховувати часову характеристику — час відгуку системи в умовах особливостей функціонування та відмов ХО. Тоді як механізм адаптації ХО можливо і доцільно використовувати технологію динамічної адаптивної змішаної дисципліни надання ресурсів користувачам ХО.

1. Aleksandr Matov. Adaptation of cloud computing as optimization of the process of rendering services to users in the conditions of limited computing resources. Selected Papers of the XIX International Scientific and Practical Conference «Information Technologies and Security» (ITS 2019). CEUR Workshop Proceedings (ceur-ws.org). Vol-2577. P. 210–221. ISSN 1613-0073.

2. Матов О.Я. Аналітичні моделі багатопріорітетних хмарних дата-ценрів зі змішаною дисципліною надання послуг з урахуванням особливостей функціонування і можливих відмов. *Реєстрація, зберігання і оброб. даних.* 2019. Т. 21. № 1. С. 32–45. doi: 10.35681/1560-9189.2019.1.1.179445

3. Aleksandr Matov. Mathematical models of cloud computing with absolute-relative priorities of providing of computer resources to users in conditions of functioning features and failures. CEUR

Workshop Proceedings (ceur-ws.org). Vol-2318 urn:nbn:de:0074-2318-4. Selected Papers of the XVIII International Scientific and Practical Conference on Information Technologies and Security (ITS 2018). P. 150–159.

4. Матов О.Я. Оптимізація надання послуг обчислювальними ресурсами адаптивної хмарної інфраструктури. *Реєстрація, зберігання і оброб. даних*. 2018. Т. 20. № 3. С. 83–90. doi: 10.35681/1560-9189.2018.20.3.158262.

5. Матов А.Я., Шпилев В.Н., Комов А.Д. и др. Организация вычислительных процессов в АСУ / под ред. А.Я. Матова. Киев, 1989. 200 с.

6. Mokrov E.V., Samuilov K.E. Cloud computing system model in the form of a queuing system with multiple queues and with a group of requests. https://cyberleninka.ru/article/n/model-sistemy-oblachnyh-vychisleniy-v-vide-sistemy-massovogo-obsluzhivaniya-s-neskolkimi-ocheredyami-i-s-gruppovym-postupleniem-zayavok. Rus.

7. Bezzateev S.V., Elina T.N., Mylnikov V.A. Modeling the processes of selecting parameters of cloud systems to ensure their stability, taking into account reliability and security. Scientific and technical bulletin of information technologies, mechanics and optics. 2018. Vol. 18. No. 4. P. 654–662. doi: 10.17586-2226-1494-2018-18-4-654-662 (in Russian).

8. Grusho A.A., Zabezhailo M.I., Zatsarinny A.A. Information flow monitoring and control in the cloud computing environment. *Informatics and Applications*, 2015. Vol. 9. No 4. P. 91–97 (in Russian).

9. Singh P., Dutta M., Aggarwal N. A review of task scheduling based on meta-heuristics approach in cloud computing. *Knowledge and Information Systems*. 2017. Vol. 52. No. 1. doi: 10.1007/s10115-017-1044-2.

Надійшла до редакції 10.12.2020