УДК 004.32

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

Аналітичні моделі багатопріоритетних хмарних дата-ценрів зі змішаною дисципліною надання послуг з урахуванням особливостей функціонування та можливих відмов

Одними із основних показників ефективності хмарних дата-центрів є показники, що базуються на оцінці часових характеристик. Порушення припустимих часових обмежень, наприклад, часу відгуку хмарних дата-центрів, негативно позначається на ефективності рішення цільових задач користувачів. Це має особливе значення для систем реального масштабу часу і, в першу чергу, для спеціальних інформаційних систем, що побудовані з використанням приватних хмар. Розроблено аналітичні моделі хмарних дата-центрів як системи масового обслуговування зі змішаною дисципліною надання ресурсів. Моделі враховують відмови й різні особливості функціонування та мають довільні закони розподілення для деяких вірогідних процесів. Для кожної моделі отримано аналітичні вирази часових характеристик. Такі моделі для хмарних дата-центрів використано вперше.

Ключові слова: *хмарний дата-центр, математична модель, дисципліна надання обчислювальних ресурсів, змішана дисципліна обслуговування, абсолютний і відносний пріоритети, часові характеристики, час відгуку, довільні закони розподілення.*

Вступ

Розробка математичних моделей хмарних дата-центрів, хмарних обчислень чи інформаційних систем, що створені з використанням хмар, є важливим напрямком для виявлення та покращення їхніх характеристик [1, 2, 4, 14–19]. Хмарні дата-центри є об’єктами з високим рівнем невизначеності процесу функціонування, головними чинниками якої є [2, 4, 5]:

— вірогідність потоку запитів на обчислювальні ресурси (ОР);

— наявність необхідних ОР і випадковість часу їхнього використання клієнтами;

— випадковість відмов інфраструктури хмарних дата-центрів і часу їхнього усунення;

— необхідність забезпечення певних часових характеристик для ряду клієнтів, наприклад, часу відгуку хмарних дата-центрів;

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

— необхідність введення адаптації у процес функціонування хмарних дата-центрів з метою забезпечення певних часових характеристик для ряду клієнтів і оптимального використання ОР.

Одними з основних показників ефективності хмарних дата-центрів є показники, що базуються на оцінці часових характеристик цих систем. Порушення припустимих часових обмежень, наприклад, часу відгуку хмарних дата-центрів, негативно позначається на ефективності рішення цільових задач користувачів, що має особливе значення для систем реального масштабу часу. В першу чергу це торкається спеціальних інформаційних систем, що побудовані з використанням приватних хмар.

Стохастичний характер головних чинників та необхідність кількісної оцінки масових процесів на основі теорії імовірності обумовлює використання теорії масового обслуговування. Тоді як механізми адаптації хмарних дата-центрів можливо і доцільно використовувати технологію динамічної адаптивної змішаної дисципліни надання ОР (обслуговування) користувачам хмарних дата-центрів [2].

Пропонуються аналітичні моделі для вирахування часових характеристик в умовах особливостей функціонування хмарних дата-центрів з використанням змішаної дисципліни обслуговування з абсолютно відносними пріоритетами та врахуванням відмов. Моделі базуються на роботах [7–10, 12, 13].

Опис моделі функціонування хмарної інфраструктури зі змішаною дисципліною обслуговування та адаптацією до відмов

Нехай на вхід хмарного дата-центру, в якому реалізована дисципліна обслуговування з відносно-абсолютним пріоритетом, надходять N пуасонівських потоків заявок інтенсивності $\lambda(m,n)$ $(m=\overline{1,M},\ \ n=\overline{1,N_m})$. Цим потокам поставлені у відповідність N пріоритетів [2].

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 А), 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 В).

Обслуговуючий прилад (хмарний дата-центр) 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 $n_i=1$ в тому випадку, якщо запити *i*-го потоку в чергу приймаються, і $n_i=0$, якщо запити одержують відмову.

Адаптацією до відмов буде те, що у період відновлення приладу вхідні заявки можуть або накопичуватись у черзі (дисципліна поповнення черги I), або одержувати відмову 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.

Різні особливості функціонування визначаються різними дисциплінами A, B, C, D, I і II. Сполучення дисциплін поновлення обслуговування та поповнення черги дозволяє розглядати самостійні моделі різних типів систем, що мають відповідне позначення: AC-I, AC-II, BD-I, BD-II.

Нехай дата-центр знаходиться у стаціонарному режимі, умовою якого для систем типу I є $R_M < K_r$, а для систем типу II — $R_M < 1$. Тут $R_M = \sum_{m=1}^{M}\sum_{n=1}^{N}\rho(m,n)$ — сумарне завантаження приладу заявками, де $(\rho(m,n) = \lambda(m,n)b(m,n)$ — 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 дата-центру для заявок кожного пріоритету.

Визначення часових характеристик моделі системи типу AC-I

Для визначення середнього часу перебування заявок у системі (часу відгуку системи) типу AC-I скористаємося відомим прямим методом [3].

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*):

$$v(j,k) = w(j,k) + b(j,k). \qquad (1)$$

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*):

$$w(j,k) = w_H(j,k) + u(j,k). \tag{2}$$

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

$$u(j,k) = u_3(j,k) + u_0(j,k). \tag{3}$$

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

$$\Theta(j,k) = b(j,k) + u(j,k). \tag{4}$$

Почнемо з обчислення $u(j, k)$, для чого застосуємо підхід, описаний у [2].

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 обслуговування.

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

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

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, що вимагають обслуговування раніш (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 в системі типу АС-I у період відновлення приладу знову надходять заявки, які продовжують накопичуватись у 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 раніш (*j*, *k*)-заявки тощо

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

В остаточному підсумку одержуємо:

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

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

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

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

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

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, з $\overline{j+1, M}$ груп;

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

4) дообслуговано заявки з $\overline{1, j}$ груп, що 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: $K_r\rho_0$ — ймовірність 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 $\overline{1, j-1}$: $\frac{K_{m-1}}{K_r - R_{m-1}}\rho(m,n)$ — ймовірність 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 $\overline{1, j}$, що перервані відмовленнями приладу:

$\frac{K_r\rho_0}{K_r - R_{m-1}}\rho(m,n)$ — ймовірність того, що в черзі маються (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:

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

$$\times \rho(m,n)\Delta(m,n) + \sum_{m=1}^{j-1} \sum_{n=1}^{N_m} w_H(m,n)\rho(m,n) + \sum_{n=1}^{k-1} w_H(j,n)\rho(j,n)\Bigg], \quad (13)$$

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 пріоритетами.

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

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

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

Ці формули дозволяють припустити загальне рішення у вигляді

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

де

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

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

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

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 вигляд формули для визначення $w_H(j,k)$ для змішаної дисципліни обслуговування:

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

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.

Визначення часових характеристик моделі системи типу АС-II

Для визначення середнього часу перебування (*j*, *k*)-заявки в системі типу АС-II і його складових необхідно скористатися тими ж підходами, що і при визначенні часових характеристик моделі системи типу АС-I. Відмінність полягає в тім, що в період відновлення обслуговуючого приладу заявки в систему не надходять.

Запишемо без висновку деякі складові середнього часу перебування (*j*, *k*)-заявки в системі типу АС-II.

Середній час очікування (*j, k*)-заявки в перерваному стані за рахунок дії перериваючих заявок із $\overline{1, j-1}$ груп і середній час відмовлень приладу знаходимо за формулами:

$$u_3(j,k) = b(j,k)\frac{R_{j-1}}{K_r(1 - R_{j-1})}, \quad (17)$$

$$u_0(j,k) = b(j,k)\rho_0. \quad (18)$$

Тоді сумарний середній час очікування (*j*, *k*)-заявки у перерваному стані має вигляд

$$u(j,k) = b(j,k)\frac{R_{j-1} + \rho_0}{1 - R_{j-1}}, \tag{19}$$

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

$$\Theta(j,k) = b(j,k)\frac{1}{K_r(1 - R_{j-1})}. \tag{20}$$

Визначимо середній час очікування (j, k)-заявки до початку обслуговування $w_H(j,k)$.

Нагадаємо, що в систему типу АС-II заявки надходять лише тоді, коли обслуговуючий прилад знаходиться в справному стані (вільний чи зайнятий обслуговуванням). У противному випадку заявки, що надходять, не приймаються. У зв'язку з цим інтенсивність прорідженого потоку будь-яких (m, n)-заявок буде дорівнювати $K_r\lambda(m,n)$, де K_r має сенс ймовірності справного стану приладу.

Нехай у систему надходить (j, k)-заявка. Перш ніж потрапити на прилад, ця заявка чекає виконання тих же подій, крім першого, що й у системі типу АС-I.

Для середніх тривалостей цих подій запишемо рівняння

$$\begin{aligned} w_H(j,k) &= \sigma(j,k) + \eta(j,k) + \eta_0(j,k) + \\ &+ K_r\sum_{m=1}^{j-1}\sum_{n=1}^{N_m} w_H(m,n)\rho(m,n) + K_r\sum_{n=1}^{k} w_H(j,n)\rho(j,n) + \\ &+ z_H(j,k)\frac{R_{j,k-1}}{K_r(1 - R_{j,k-1})} + z_H(j,k)\rho_0, \end{aligned} \tag{21}$$

в якому

$\sigma(j,k) = K_r\sum_{m=1}^{j}\sum_{n=1}^{N_m}\rho(m,n)\Delta(m,n)$ — середній час дообслуговування кожної (m, n)-заявки в присутності (j, k)-заявки;

$K_r\rho(m,n)$ — ймовірність обслуговування (m, n)-заявки;

$\eta(j,k) = \sum_{m=2}^{j}\sum_{n=1}^{N_m}\frac{R_{m-1}}{1 - R_{m-1}}\rho(m,n)\Delta(m,n)$ — середній час дообслуговування заявок, перерваних заявками з абсолютним пріоритетом;

$\frac{R_{m-1}}{1 - R_{m-1}}\rho(m,n)$ — ймовірність перебування в черзі (m, n)-заявки, перерваної заявками з груп $\overline{1, m-1}$. Ця ймовірність отримана з виразу (17) з урахуванням інтенсивності $K_r\lambda(m,n)$ потоку (m, n)-заявок;

$\eta_0(j,k) = K_r\rho_0\sum_{m=1}^{j}\sum_{n=1}^{N_m}\rho(m,n)\Delta(m,n)$ — середній час дообслуговування заявок, перерваних відмовленнями приладу: $K_r\rho_0\rho(m,n)$ — ймовірність того, що в

черзі є (*m*, *n*)-заявка, перервана відмовленням приладу. Ця ймовірність визначається з рівняння (18) з урахуванням $K_r\lambda(m,n)$;

$z_H(j,k)$ — сумарний середній час очікування (*j*, *k*)-заявки.

З рівняння (21) отримуємо наступне рекурентне співвідношення:

$$w_H(j,k)=\frac{1}{K_r(1-R_{j,k})}\left[\sum_{m=1}^{j}\sum_{n=1}^{N_m}\frac{1}{1-R_{m-1}}\rho(m,n)\Delta(m,n)+\right.$$
$$\left.+K_r\sum_{m=1}^{j-1}\sum_{n=1}^{N_m}w_H(m,n)\rho(m,n)+K_r\sum_{n=1}^{k-1}w_H(j,n)\rho(j,n)\right]. \quad (22)$$

У результаті аналізу, подібного проведеному для системи типу АС-I, зі співвідношення (22) можна одержати вираз для визначення $w_H(j,k)$ у системі з «чистим» відносним і абсолютним пріоритетами, а також у системі зі змішаним пріоритетом:

$$w_H(1,k)=\frac{\sum_{n=1}^{N_1}\rho(1,n)\Delta(1,n)}{K_r(1-R_{1,k-1})(1-R_{1,k})}, \quad (23)$$

$$w_H(j,1)=\frac{\sum_{m=1}^{j}\rho(m,1)\Delta(m,1)}{K_r(1-R_{j-1,1})(1-R_{j,1})}, \quad (24)$$

$$w_H(j,k)=\frac{\sum_{m=1}^{j}\sum_{n=1}^{N_m}\rho(m,n)\Delta(m,n)}{K_r(1-R_{j,k-1})(1-R_{j,k})}. \quad (25)$$

Шуканий середній час перебування (*j*, *k*)-заявки *v* (*j*, *k*) у системі типу АС-II визначається за виразами (20) і (25).

Визначення часових характеристик моделі системи типу BD-I

Система типу BD-I відрізняється від системи типу АС тим, що за час кожної перерви (*m*, *n*)-заявки обслуговуються всі заявки з потоків $\overline{(1,1),(m,n-1)}$. Нехай у систему типу BD-I надходить (*j*, *k*)-заявка. Для визначення *v* (*j*, *k*) скористаємося наступним прийомом. Припустимо, що заявки потоків $\overline{(j,1),(j,k-1)}$, так само як і заявки з груп $\overline{1,j-1}$, мають абсолютний пріоритет до заявок потоку (*j*, *k*). У цьому випадку маємо систему з абсолютним пріоритетом і (*j*, *k*)-потоками. Для такої модифікаційної системи відповідно до виразу (11) можна записати:

$$\Theta^*(j,k)=\frac{b(j,k)}{K_r-R_{j,k-1}}. \quad (26)$$

У досліджуваній системі зі змішаною дисципліною обслуговування

$$\Theta(j,k) = \Theta^*(j,k) - \Delta\Theta(j,k), \tag{27}$$

де величина $\Delta\Theta(j,k)$ обумовлена перерахованими нижче факторами. За час обслуговування (j, k)-заявки $\omega(j,k)$, що слідує за останньою її перервою, у системі може зібратися $\omega(j,k)\sum_{n=1}^{k-1}\lambda(j,n)$-заявок потоків $\overline{(j,1),(j,k-1)}$, які будуть обслуговуватися протягом часу $\omega(j,k)\sum_{n=1}^{k-1}\rho(j,n)$. Обслуговування цих заявок з урахуванням надходження заявок потоків $\overline{(1,1),(j,k-1)}$ і відмовлень приладу в модифікованій системі приведе до збільшення середнього повного часу обслуговування (j, k)-заявки на величину

$$\Delta\Theta(j,k) = \frac{\omega(j,k)\sum_{n=1}^{k-1}\rho(j,n)}{K_r - R_{j,k-1}}. \tag{28}$$

Оскільки в досліджуваній системі зі змішаним пріоритетом заявки потоків $\overline{(j,1),(j,k-1)}$ не переривають обслуговування (j, k)-заявки, то її повний час обслуговування $\Theta(j,k)$ буде менше на величину $\Delta\Theta(j,k)$, тобто має місце рівність (27), що з урахуванням (26) і (28) приймає вигляд:

$$\Theta(j,k) = \frac{b(j,k) - \omega(j,k)\sum_{n=1}^{k-1}\rho(j,n)}{K_r - R_{j,k-1}}. \tag{29}$$

Обчислимо $\omega(j,k)$. Нехай тривалість обслуговування (j, k)-заявки дорівнює t. З ймовірністю $\exp\{-(\Lambda_{j-1}+\lambda_0)t\}$ це обслуговування не буде перервано, і тоді тривалість останнього інтервалу дорівнює t. У випадку переривань з імовірністю $1-\exp\{-(\Lambda_{j-1}+\lambda_0)x\}$ тривалість останнього інтервалу менше деякого значення x, причому $x < t$. Звідси:

$$\begin{aligned}\omega(j,k) &= \int_0^\infty \{te^{-(\Lambda_{j-1}+\lambda_0)t} + \int_0^t x\,d[1-e^{-(\Lambda_{j-1}+\lambda_0)x}]\}\times \\ &\times dB_{j,k}(t) = \frac{1}{\Lambda_{j-1}+\lambda_0}\int_0^\infty [1-e^{-(\Lambda_{j-1}+\lambda_0)t}]dB_{j,k}(t).\end{aligned} \tag{30}$$

При показовому обслуговуванні $\omega(j,k) = b(j,k)/[+(\Lambda_{j-1}+\lambda_0)b(j,k)]$, а при детермінованому — $\omega(j,k) = [1-e^{-(\Lambda_{j-1}+\lambda_0)b(j,k)}]/(\Lambda_{j-1}+\lambda_0)$.

Залишилось обчислити $w_H(j,k)$. Представимо модифіковану систему з потоками $\overline{(1,1),(j,N_j)}$ і з абсолютним пріоритетом. Для такої системи відповідно до формули (16) маємо:

$$w_H^*(j,k) = \frac{K_r^3 \rho_0 \Delta_0 + \sum_{m=1}^{j-1} \sum_{n=1}^{N_m} \rho(m,n)\Delta(m,n)}{(K_r - R_{j,k-1})(K_r - R_{j,k})} + \frac{\sum_{n=1}^{k} \rho(j,n)\Delta(j,n)}{(K_r - R_{j,k-1})(K_r - R_{j,k})}. \quad (31)$$

У досліджуваній системі зі змішаним пріоритетом будемо мати:

$$w_H(j,k) = w_H^*(j,k) + \Delta w_H(j,k), \quad (32)$$

де $\Delta w_H(j,k)$ визначається з наступних розумінь. У момент надходження (j, k)-заявки може обслуговуватися будь-яка заявка з потоків $\overline{(j,k+1),(j,N_j)}$. Середній час обслуговування цієї заявки в присутності (j, k)-заявки $\sigma^*(j,k) = \sum_{n=k+1}^{N_j} \rho(j,n)v(j,n)$, де $v(j,n)$ — середня тривалість обслуговування (j, n)-заявки від моменту надходження (j, k)-заявки до першого переривання. За час $\sigma^*(j,k)$ у системі збереться $\sigma^*(j,k)\sum_{n=1}^{k-1}\lambda(j,n)$-заявок потоків $\overline{(j,1),(j,k-1)}$, повний час обслуговування яких з урахуванням заявок потоків $\overline{(1,1),(j,k-1)}$ і відмовлень приладу дорівнює $\sigma^*(j,k)\sum_{n=1}^{k-1}\rho(j,n)/(K_r - R_{j,k-1})$. Тоді повний час обслуговування заявок потоків $\overline{(j,k+1),(j,N_j)}$ буде дорівнювати: $(K_r - R_{j-1})\sigma^*(j,k)/(K_r - R_{j,k-1})$.

Оскільки заявки потоків $\overline{(j,1),(j,k-1)}$ не переривають обслуговування (j, n)-заявки, то середній час очікування (j, k)-заявки $w_H(j,k)$ збільшиться на $\Delta w_H(j,k)$:

$$\Delta w_H(j,k) = \frac{(K_r - R_{j-1}) \sum_{n=k-1}^{N_j} \rho(j,n)v(j,n)}{(K_r - R_{j,k-1})(K_r - R_{j,k})}. \quad (33)$$

Величину $w_H(j,k)$ знаходимо з виразу (32) з урахуванням формул (31) і (33).

Визначимо $v(j,n)$. Нехай у момент надходження (j, k)-заявки час дообслуговування (j, n)-заявки, що знаходиться на приладі, дорівнює t. Вона буде або дообслугована до кінця, або перервана заявкою, що надходить, з абсолютним пріоритетом чи відмовленням приладу. Щільність імовірності часу t дорівнює $[1 - B_{j,n}(t)]/b(j,n)$ [9]. Аналогічно формулі (30) маємо:

$$v(j,n) = \int_0^\infty \frac{1 - e^{-(\Lambda_{j-1}+\lambda_0)t}}{\Lambda_{j-1} + \lambda_0} \frac{1 - B_{j,n}(t)}{b(j,n)} = \frac{b(j,n) - \omega(j,n)}{(\Lambda_{j-1} + \lambda_0)b(j,n)}, \quad (34)$$

де $\omega(j,n)$ визначається за формулою (30).

Величину $v(j,k)$ знаходимо, використовуючи вирази (29), (31) і (33).

Визначення часових характеристик моделі системи типу BD-II

Аналогічно системі BD-I можуть бути визначені тимчасові характеристики моделі системи типу BD-II. Запишемо без обчислення виразу для визначення $\Theta(j,k)$ та $w_H(j,k)$:

$$\Theta(j,k) = \frac{b(j,) - \omega(j,k)\sum_{n=1}^{k-1} \rho(j,n)}{K_r(1 - R_{j,k-1})}, \tag{35}$$

$$w_H(j,k) = \frac{\sigma_{j,k} + (1 - R_{j-1}) \sum_{n=k+1}^{N_j} \rho(j,n)v(j,n)}{K_r(1 - R_{j,,k-1})(1 - R_{j,k})}, \tag{36}$$

де

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

Величина $v(j,k)$ обчислюється шляхом підсумовування рівностей (35) і (36).

Висновок

Розроблені аналітичні моделі хмарних дата-центрів як системи масового обслуговування зі змішаною (абсолютно-відносними пріоритетами) дисципліною надання обчислювальних ресурсів користувачам, які враховують різні особливості функціонування та мають довільні закони розподілення для деяких вірогідних процесів. Моделі дозволяють враховувати часові характеристики (наприклад, час відгуку системи) дата-центрів та інформаційних систем, які створюються із використанням хмарних дата-центрів. Такі моделі для хмарних дата-центрів використані вперше.

Ці моделі враховують такі фізичні властивості хмарних дата-центрів як миттєву еластичність (динамічне виділення та звільнення ресурсів для швидкого масштабування відповідно до потреб) та вимірювальний сервіс (керування і оптимізація ресурсів за допомогою засобів вимірювання).

1. Matov A.Y. Mathematical models of cloud computing with absolute — relative priorities of providing computer resources to users in conditions of functioning features and failures. Материалы XVIII международной научно-практической конференции «Информационные технологии и безопасность». Киев ИПРИ НАН Украины. 2018. Вып. 18. С. 319–329. ISBN 978-966 2344-69-1.

2. Матов О.Я. Оптимізація надання послуг обчислювальними ресурсами адаптивної хмарної інфраструктури. *Реєстрація, зберігання і оброб. даних*. 2018. Т. 20. № 3. С. 83–90.

3. 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.

4. Матов О.Я., Храмова І.О. Проблеми користування і математичне моделювання хмарних обчислень для інтегрованої інформаційно-аналітичної системи державного управління. *Реєстрація, зберігання і оброб. даних*. 2010. Т. 12. № 2. С. 113-127.

5. Матов О.Я., Храмова І.О. Сучасні технології інтеграції інформаційних ресурсів. *Реєстрація, зберігання і оброб. даних.* 2009. Т. 11. № 1. С. 33–42.

6. Matov A.Y. Optimum operational training in systems containing components subject to failure. *Engineering Cybernetic.* 1975. **13**(5). C. 87–90.

7. Matov A.Y. Two modes of continuous completion of a queue when the instrument is restored in a servicing system with a relative priority. *Avtomat. i Telemekh*. 1974. P. 66–70.

8. Matov A.Y. Two priority system with an unreliable device and period of servicing. *Engineering Cybernetics*. 1973. **10**(5). P. 849–852.

9. Matov A.Y. Two continuous queue disciplines for service-resumption period in a nonpreemptive-priority queuing system. *Automation and remote control*. 1974. **35**(4). P. 575–578.

10. Матов А.Я., Тищенко Н.Ф. Математические модели вычислительных систем с приоритетным отказом в обслуживании. Изв. АН СССР. *Техническая кибернетика*. 1980. № 3. С. 190–194.

11. Матов А.Я., Шпилев В.Н. Использование комбинированных приоритетов для повышения эффективности вычислительных процессов в АСУ. *Механизация и автоматизация управления*. 1983. № 4. С. 58–60.

12. Matov A.Y., Zhluktenko V.I., Chernous K.A., Tishchenko N.F. Two continuous queuing disciplines in mixed priority systems. *Cybernetics and Systems Analysis*. 1978. **14**(3). C. 421–426.

13. Matov A.Y., Tishenko N.F., Zhluktenko V.I.. Numerical-method for investigation of priority systems with finite queues and non-failproof maintenance devices. *Avtomatika i vychislitelnaya tekhnika*. 1978. C. 48–53.

14. Мокров Е.В., Самуйлов К.Е. Модель системы облачных вычислений в виде системы массового обслуживания с несколькими очередями и с групповым поступлением заявок. URL: https://cyberleninka.ru/article/n/model-sistemy-oblachnyh-vychisleniy-v-vide-sistemy-massovogo-obsluzhivaniya-s-neskolkimi-ocheredyami-i-s-gruppovym-postupleniem-zayavok

15. Горбунова А.В., Зарядов И.С., Матюшенко С.И., Самуйлов К.Е., Шоргин С.Я. Аппроксимация времени отклика системы облачных вичислений. *Информатика и ее применения*. 2015. Т. 9. № 3. С. 32–38. doi: 10.14357/19922264150304.

16. Беззатеев С.В., Елина Т.Н., Мыльников В.А. Моделирование процессов подбора параметров облачных систем для обеспечения их устойчивости с учетом надежности и безопасности. *Научно-технический вестник информационных технологий, механики и оптики*. 2018. Т. 18. № 4. С. 654–662. doi: 10.17586/2226-1494-2018-18-4-654-662.

17. Гудкова И.А., Масловская Н.Д. Вероятностная модель для анализа задержки доступа к инфраструктуре облачных вычислений с системой мониторинга // T-Comm: Телекоммуникации и транспорт. 2014. № 6. С. 13-15.

18. Tsai J.M., Hung S.W. A novel model of technology diffusion:system dynamics perspective for cloud computing. *Journal of Engineering and Technology Management*. 2014. Vol. 33. P. 4762. doi: 10.1016/j.jengtecman.2014.02.003.

19. 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.

20. 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). doi: 10.14357/1992264150410.

21. Gudkova I.A., Maslovskaya N.D. Probability model for analysing impact of delays due to monitoring on mean service time in cloud computing. *T-Comm: Telecommunications and Transport*. 2014. No 6. P. 13–15 (in Russian).

Надійшла до редакції 20.02.2019