УДК 004.9 : 621.39

Г.А. Кучук

Харківський університет Повітряних Сил імені Івана Кожедуба, Харків

МЕТОД ОПТИМАЛЬНОГО РОЗПОДІЛУ РЕСУРСІВ CLOUD-СИСТЕМИ

Розглянутий двохетапний підхід до оптимального розподілу ресурсів CLOUD-системи. На першому етапі аналізується процес розподілу множини транзакцій, декомпозований по двох рівнях: "користувацькі та комутаційні вузли системи" та "гібридне хмарне сховище даних". На другому етапі оптимізації розподіляються ресурси, котрі надаються CLOUD-системою при багатофазній обробці транзакцій безпосередньо на серверах системи. Сформульовані відповідні оптимізаційні задачі математичного програмування, наведені шляхи їх розв'язання.

***Ключові слова**: сховище даних, гібридність, хмарні технології, CLOUD-система.*

Вступ

В останні роки величезної популярності в галузі інформаційних технологій набула «хмарна технологія» (англ. *cloud technologies*) [1, 2]. Сучасний IT-ринок пропонує велику кількість програмних продуктів, що підтримують дану технологію (у подальшому – CLOUD-системи) та забезпечують функціонування як хмарного сховища даних (англ. *cloud storage*) – модель онлайн-сховища, в якому дані зберігаються на численних розподілених в мережі серверах, що надаються в користування клієнтам, так і хмарних обчислень (англ. *cloud computing*)**,** котрі забезпечують розподілену обробку даних, при якій комп'ютерні ресурси та потужності користувач отримує як інтернет-сервіс [2].

Серед ряду проблем, що виникають при впровадженні CLOUD-систем, найбільш актуальними є такі, котрі пов'язані з розподілом ресурсів системи, що викликано як суттєвим зростанням користувачів, так і розширенням послуг, котрі надаються сучасними CLOUD-системами [3, 4]. Зокрема, стоїть завдання оптимального розподілу ресурсів у CLOUD-системах, котрі підтримують гібридні хмарні сховища даних (ГХСД), що і є **метою** даної статті. Пропонується окремо розглянути процес оптимального розподілу множини транзакцій до ГХСД та процес розподілу ресурсів, котрі надаються CLOUD-

системою при багатофазній обробці транзакцій безпосередньо на серверах системи.

1. Розподіл транзакцій до гібридного хмарного сховища даних

Задачу розподілу транзакцій будемо розглядати з врахуванням двох ієрархічних рівнів хмарного середовища (ХС), визначеного CLOUD-системою: "користувацькі та комутаційні вузли системи" (ККС) та "гібридне хмарне сховище даних" (ГХСД).

На ієрархічному рівні ККС необхідно для кожного k-го вузла ХС ($k = \overline{1,K}$) знайти мінімум витрат мережевого ресурсу CLOUD-системи [5]:

$$\varphi_k^{(0)} = \min_{x_k \in D_k(\bar{y}_k)} \varphi(\bar{x}_k) \qquad (1)$$

при таких обмеженнях

$$D_k(\bar{y}_k) = \left\{ \begin{array}{l} \bar{x}_k \left| \sigma_k^* \cdot \bar{x}_k = \bar{y}_k; \right. \\ R_k^{(\ell)}(\bar{x}_k) \le M_{0,k}^{(\ell)}, \\ \ell = \overline{1,L};\ x_k \ge 0, \end{array} \right\}, \qquad (2)$$

де $\varphi(\bar{x}_k)$ – функціонал k-го вузла ККС;

$\bar{x}_k$ – вектор стану k-го вузла ККС;

$\bar{y}_k$ – вектор узагальненого стану для k-го вузла ККС, що розглядається в дискретні моменти часу (на інтервалі [0, T] розглядається t_n відліків];

$D_k(\bar{y}_k)$ – множина допустимих станів;

σ_k^* – транспонована матриця інцидентності k-го вузла ККС;

$R_k^{(\ell)}(\bar{x}_k)$ – функція витрат ℓ-ої складової мережевого ресурсу CLOUD-системи;

$M_{0,k}^{(\ell)}$ – допустимий об'єм витрат ℓ-ої складової мережевого ресурсу CLOUD-системи для k-го вузла ККС;

L – кількість обмежених мережевих ресурсів.

На рівні ГХСД необхідно знайти глобальний мінімум витрат мережевого ресурсу CLOUD-системи [6]:

$$\varphi^{(0)} = \min_{M_{0,k} \in D(M), \bar{y}_k \in G(\bar{y})} \sum_{k=1}^{K} \varphi_k^{(0)}\left(M_{0,k}; \bar{y}_k\right), \qquad (3)$$

де

$$D(M) = \left\{ M_{0,k} \left| \sum_{k=1}^{K} M_{0,k}^{(\ell)} \le M_0^{(\ell)};\ M_0^{(\ell)} \ge 0;\ \ell = \overline{1,L} \right. \right\}; \qquad (4)$$

$$G(y) = \left\{ \bar{y}_k \left| \psi^* \cdot \bar{y}_k \ge \bar{B};\ \bar{y}_k \ge 0 \right. \right\}, \qquad (5)$$

$\varphi_k^{(0)}$ – функція, що апроксимує мінімальні значення затрат для k-го вузла ККС при зміні векторів відтворюваних ресурсів $M_{0,k}$ та невідтворюваних ресурсів, визначених вектором $\bar{y}_k$, що розглядаються у дискретні проміжки часу;

$G(\bar{y}_k)$ – множина допустимих значень векторів узагальненого стану $\bar{y}_k$;

$M_0^{(\ell)}$ – загальна кількість ℓ-ої складової мережевого ресурсу, виділеного CLOUD-системою;

$\bar{B}$ – вектор потреби ресурсів;

ψ^* – транспонована матриця інцидентності ГХСД.

Багатократно розв'язуючи задачу (1) – (2) на рівні ККС при фіксованих значеннях вектору $\bar{y}_k$ та виділених ресурсів $M_0^{(\ell)}$ в допустимих діапазонах їх змін [7], знаходимо множину значень

$$\varphi_k^{(0)} = \left(M_{0,k}; \bar{y}_k\right). \qquad (6)$$

Апроксимувавши цю множину безперервною функцією, одержимо динамічну характеристику k-го ККС. Після цього можна перейти до другого етапу – безпосередньому аналізу транзакцій гібридного хмарного сховища даних.

Припустимо, що загальний неоднорідний трафік транзакцій ХС можна розбити на R видів однорідних потоків. Трафік k-го КлС тоді можна розбити на m_k ($m_k \le R$) однорідних потоків у часовому періоді [0, T], що характеризуються множиною

$$Q^{(k,t_n)} = \left\{ Q_{i_1}^{(k,t_n)}, \ ..., \ Q_{i_{m_k}}^{(k,t_n)} \right\}, \qquad (7)$$

де індекси визначають належність до відповідного типу потоку, причому для вхідного трафіка значення $Q_{i_j}^{(k)} \ge 0$, для вихідного – $Q_{i_j}^{(k)} < 0$. Тоді траєкторію однорідного потоку можна задати таким кортежем [8]:

$$< u, k_b, k_e, r >, \qquad (8)$$

де u – номер ділянки; k_b – початковий вузол ККС траєкторії; k_e – кінцевий вузол ККС траєкторії; r – тип потоку.

Сумарний потік на ділянці u за часовий період [0, T] можна записати, як

$$Q_u^{(t_n)} = \sum_{k_b=1}^{K} \sum_{k_e=1}^{K} \sum_{r=1}^{R} Z_{u,k_b,k_e,r}^{(t_n)}, \qquad (9)$$

де складання проводиться по усім траєкторіям, що проходять крізь розглядаєму ділянку, а $Z_{u,k_b,k_e,r}^{(t_n)}$ – потік відповідної траєкторії, причому

$$Z_{u,k_b,k_e,r}^{(t_n)} = \xi_{u,k_b,k_e,r}^{(t_n)} + \zeta_{u,k_b,k_e,r}^{(t_n)}, \qquad (10)$$

де перша із складових – це частина потоку, що знаходиться в межах пропускної здатності ділянки, а друга – частина потоку, що перевищує пропускну здатність ділянки.

Тоді

$$\sum_{k_b=1}^{K}\sum_{k_e=1}^{K}\sum_{r=1}^{R}\xi_{u,k_b,k_e,r}^{(t_n)} \le \le P_u^{(0)} + \sum_{i=1}^{t_n-1}\sum_{k_b=1}^{K}\sum_{k_e=1}^{K}\sum_{r=1}^{R}\zeta_{u,k_b,k_e,r}^{(i)}, \quad (11)$$

де $P_u^{(0)}$ – пропускна здатність ділянки u.

Якщо $v_{i_j}^{(k,t_n)}$ – об‘єм потоку i_j в k-му вузлі ККС за часовий період [0, T], то

$$0 \le v_{i_j}^{(k,t_n)} \le -\min\left(0, Q_{i_j}^{(k,t_n)}\right) \quad (12)$$

і базисні рівняння мають такий вигляд [9]:

$$\sum_{u\in U_1} Z_{u,k_b,k_e,r}^{(t_n)} - \sum_{u\in U_2} Z_{u,k_b,k_e,r}^{(t_n)} + v_{i_j}^{(k,t_n)} = = P_{i_j}^{(k,t_n)}, \quad (13)$$

де $$P_{i_j}^{(k,t_n)} = \max\left(0, Q_{i_j}^{(k,t_n)}\right), \quad (14)$$

U_1, U_2 – множини всіх траєкторій типу i_j, що входять і виходять із k-го вузла ККС за часовий період [0, T] відповідно.

Тоді цільова функція задачі оптимізації розподілу транзакцій гібридного хмарного сховища даних може бути подана у лінійному вигляді таким чином:

$$\sum_{t_n}\left(\sum_{u\in U} W_u^{(1)}\sum_{k_b=1}^{K}\sum_{k_e=1}^{K}\sum_{r=1}^{R}\zeta_{u,k_b,k_e,r}^{(t_n)} + + W_u^{(2)}\left(\sum_{k_b=1}^{K}\sum_{k_e=1}^{K}\sum_{r=1}^{R}\xi_{u,k_b,k_e,r}^{(t_n)} + \varsigma_{u,k_b,k_e,r}^{(t_n)}\right) + + W_u^{(3)}\sum_{k=1}^{K}\left|\overline{x}_k\right|\cdot v_{i_j}^{(k,t_n)}\right) \to \min, \quad (15)$$

при відповідних обмеженнях на ресурси та затрати $\left(W_u^{(1)},\ W_u^{(2)},\ W_u^{(3)}\right)$.

Задачу оптимізації із цільовою функцією (15) та обмеженнями (11) – (14) можна розв’язати відомими методами дискретної оптимізації [11], наприклад, методом гілок та меж. Після її вирішення можна починати другий етап оптимізації, працюючи безпосередньо з ресурсами обслуговування ГХСД.

2. Розподіл ресурсів гібридного хмарного сховища даних

Виходячи з (7) – (9), розглянемо множину транзакцій, що надходять до гібридного хмарного сховища даних за часовий період [0, T], як лінійний замкнутий стохастичний ланцюг [10]. Це дозволяє сформулювати завдання оптимального розподілу ресурсів ГХСД в термінах теорії масового обслуговування таким чином [10]. Припустимо, що при обслуговуванні запитів в ГМХС транзакція проходить N фаз обслуговування, причому кожна j-та фаза має в розпорядженні n_j каналів масового обслуговування. Відома інтенсивність μ_j обслуговування одним каналом j-ої фази, також ГМХС за часовий період [0, T] обслуговує m транзакцій.

Для експлуатації ГМХС за часовий період [0, T] передбачаються витрати ресурсу CLOUD-системи у розмірі W, втрати на обслуговування одного каналу для фази j-го типу – ω_j. Необхідно визначити кількість каналів для кожної фази обслуговування, при якому середній час перебування транзакції в CLOUD-системі (t_{cp}) буде мінімальним.

Розглянемо роботу CLOUD-системи для обробки заритів до ГХСД у стаціонарному режимі, причому всі транзакції, що надходять, повинні бути обслуговані, а ймовірність переходу з j-ої фази до (j + 1)-ої дорівнює $P_{i,i+1}$. Якщо $t_j^{(cp)}$ – середня тривалість перебування транзакції в j-й фазі, то

$$t_{cp} = \sum_{j=1}^{N} P_{j-1,j}\cdot t_j^{(cp)}. \quad (16).$$

Для виконання вимоги обов'язкового обслуговування всіх транзакцій, що поступають до CLOUD-системи, повинна дотримуватися умова ненасичення, тобто сумарна інтенсивність обслуговування всіх каналів j-ої фази повинна бути більше, ніж інтенсивність вхідного потоку транзакцій цієї фази (кількість каналів n_j повинне бути не менше фіксованого мінімуму, тобто $n_j \ge n_j^{(\phi)}$). При виконанні даної умови для розподілу залишається такий обсяг ресурсу CLOUD-системи:

$$W^{(ост)} = W - \sum_{j=1}^{N}\omega_j\cdot n_j^{(\phi)}, \quad (17)$$

причому повинне виконуватись обмеження

$$W^{(ост)} \ge 0. \quad (18)$$

Для вирішення задачі мінімізаціїї цільової функції (16) з виконанням обмежень (17) – (18) методом динамічного програмування [11] необхідно провести поетапну оптимізацію розглянутої системи масового обслуговування (СМО), причому етапом оптимізації в даному випадку буде оптимізація фаз системи, починаючи з останньої.

Введемо такі позначення: x_j – ресурс, що виділений j-й фазі понад обов'язкового $\omega_j\cdot n_j^{(\phi)}$, причому в θ-му варіанті побудови системи $\left(\theta = \overline{1,\Theta}\right)$ змінна x_j набуває фіксованого значення $x_j^{(\theta)}$; $y_j^{(\theta)}$ – частина $x_j^{(\theta)}$, що виділяється на внутрішні потреби

системи обслуговування ГХСД фази j; $f_{j,N}\left(x_j^{(\theta)}\right)$ – мінімальна середня тривалість перебування транзакції у фазах від j-ої до N-ої включно при оптимальному розподілі засобів між цими фазами у варіанті з номером θ.

При оптимізації кожної j-ої фази розглядатимемо гіпотези про ресурс, виділений ланцюгу фаз від j-ої до N-ої, причому будь-яке припущення повинне відповідати умові $0 \le y_j \le x_j^{(\theta)}$ і для кожного значення θ повинне вирішуватися рівняння

$$f_{j,N}\left(x_j^{(\theta)}\right) = \min_{0\le y_j\le x_j^{(\theta)}}\left(P_{j-1,j}\cdot t_j^{(cp)} + f_{j+1,N}\left(x_j^{(\theta)} - y_j\right)\right), \theta \in \overline{1,L}, \quad (19)$$

тобто мінімальний середній час проходження транзакції в ланцюзі фаз (від j до N) дорівнює мінімуму з суми часу перебування транзакції в j-й фазі (з урахуванням y_j) і умовно мінімального часу перебування заявки у фазах від (j + 1)-ої до N-ої фази.

Для кожного етапу оптимізації функціональне рівняння (19) вирішується Θ разів, причому у результаті отримуємо вектор з Θ умовних мінімальних середніх часів і вектор найбільш вигідних значень y_j при заданих $x_j^{(\theta)}$. На першому кроці процесу оптимізації $0 \le y_N \le \le W^{(ост)}$, кількість каналів N-ої фази $n_N = n_N^{(ф)} + \left[y_N / \omega_N\right]$, а функціональне рівняння (19) приймає наступний вигляд:

$$f_{N,N}\left(x_N^{(\theta)}\right) = P_{N-1,N}\cdot t_N^{(cp)}\left(x_N^{(\theta)}\right), \quad \theta \in \overline{1,\Theta} \quad (20)$$

На другому кроці оптимізації (19) з обліком (20) приводиться до такого вигляду:

$$f_{N-1,N}\left(x_{N-1}^{(l)}\right) = \min_{0\le y_{N-1}\le x_{N-1}^{(l)}}\left(\begin{array}{c} P_{N-2,N-1}\cdot t_{N-1}^{(cp)}\left(x_{N-1}^{(\theta)}\right) + \\ + f_{NN}\left(x_{N-1}^{(\theta)} - y_{N-1}\right)\end{array}\right). \quad (21)$$

Далі, на кожному k-му кроці значення умовно мінімальної середньої тривалості перебування транзакції у фазах від від j-ої до N-ої дорівнює $f_{N-k+1,N}\left(x_{N-k+1}^{(\theta)}\right)$ та знаходиться, виходячи з результатів вирішення функціонального рівняння для значення $f_{N-k+2,N}\left(x_{N-k+2}^{(\theta)}\right)$.

Описаний ітераційний процес дозволяє на N-му кроці оптимізації знайти безумовно оптимальний ресурс $y_1^{(0)}$ для побудови цієї фази з такого функціонального рівняння:

$$f_{1,N}\left(W^{(ост)}\right) = \min_{0\le y_1\le L}\left(P_{0,1}\cdot t_1^{(cp)}\left(y_1\right) + f_{2,N}\left(L - y_1\right)\right), \quad (22)$$

причому величина $f_{1,N}\left(W^{(ост)}\right)$ є середнім часом t_{cp}.

Розгортаючи процес у зворотному напрямі від першої до останньої фази можна розрахувати безумовно оптимальні розміри ресурсу, що виділяються на побудову кожної фази, використовуючи такий ітераційний процес:

$$f_{j,N}\left(W^{(ост)} - \sum_{i=1}^{j-1} y_i^{(0)}\right) = \min_{0\le y_j\le W-\sum_{i=1}^{j-1} y_i^{(0)}}\left(\begin{array}{c} P_{i-1,i}\cdot t_j^{(cp)}\left(y_j\right) + \\ f_{(j+1),N}\left(L - \sum_{i=1}^{j-1} y_j^{(0)} - x_j\right)\end{array}\right). \quad (23)$$

При цьому необхідна кількість каналів в кожній фазі розраховується як

$$n_j = n_j^{(ф)} + \left[y_j^{(0)} \Big/ \omega_j\right]. \quad (24)$$

Висновки

Таким чином, у статті запропонований двохетапний метод оптимального розподілу ресурсів CLOUD-системи. На першому етапі аналізується процес розподілу множини транзакцій. Проведена декомпозиція даного процесу по двох рівнях: "користувацькі та комутаційні вузли системи" та "гібридне хмарне сховище даних". На другому етапі запропонованого методу розподіляються ресурси, котрі надаються CLOUD-системою при багатофазній обробці транзакцій безпосередньо на серверах системи. Сформульовані відповідні оптимізаційні задачі математичного програмування, наведені шляхи їх розв'язання.

Перспектива подальших досліджень у даному напрямі пов'язана з розширенням області застосування запропонованого методу для всіх типів хмарних сховищ даних.

Список літератури

1. Широкова Е.А. Облачные технологии / Е.А. Широкова // Современные тенденции техн. наук: мат. межд. науч. конф.; Уфа, 2011 г. – Уфа: Лето, 2011. – С. 30 – 33.

2. Риз Д. Облачные вычисления [Текст] / Джордж Риз. – СПб.: 2011. – 288 с.

3. Google Cloud Platform [Электронный ресурс]. – Режим доступа: http://cloud.google.com. – 12.04.2013.

4. Бородакий Ю.В. Эволюция информационных систем (современное состояние и перспективы) / Ю.В. Бородакий. – М. : Горячая Линия – Телеком, 2011. – 368 с.

5. Кучук Г.А. Распределение каналов по трактам узла коммутации при адаптивной маршрутизации / Г.А. Кучук // Вестник НТУ «ХПИ». Тем. вып. «Автоматика и приборостроение». – Х.: НТУ «ХПИ», 2003. – № 26. – С. 167 – 172.

6. Кучук Г.А. Розрахунок навантаження мультисервісної мережі / Г.А. Кучук, Я.Ю. Стасєва, О.О. Болюбаш // Системи озброєння і військова техніка. – 2006. – № 4 (8). – С. 130 – 134.

7. Кучук Г.А. Управление ресурсами инфотелекоммуникаций / Г.А. Кучук, Р.П. Гахов, А.А. Пашнев. – М.: Физматлит, 2006. – 220 с.

8. Кучук Г.А. Моделювання трафіка мультисервісної розподіленої телекомунікаційної мережі / Г.А. Кучук, І.Г. Кіріллов, А.А. Пашнєв // Системи обробки інформації. – Х.: ХУ ПС, 2006. – Вип. 9 (58). – С. 50 – 59.

9. Поповский В.В. Математические основы управления и адаптации в телекоммуникационных системах / В.В. Поповский, В.Ф. Олейник. – Х.: ООО "Компания СМИТ", 2011. – 362 с.

10. Саймак А. Обработка транзакций / А. Саймак // СУБД. – 1997. – № 2. – С. 70 - 82.

11. Сергиенко И.В. Модели и методы решения на ЭВМ комбинаторных задач оптимизации / И.В. Сергиенко, М.Ф. Каспшицкая. – К.: Наук. думка, 1981. – 287 с.

Надійшла до редколегії 29.05.2013

Рецензент: д-р техн. наук проф. Ю.В. Стасєв, Харківський університет Повітряних Сил ім. І. Кожедуба, Харків.

МЕТОД ОПТИМАЛЬНОГО РАСПРЕДЕЛЕНИЯ РЕСУРСОВ CLOUD-СИСТЕМЫ

Г.А. Кучук

Рассмотрен двухэтапный подход к оптимальному распределению ресурсов CLOUD-системы. На первом этапе анализируется процесс распределения множества транзакций, проведена его декомпозиция по двум уровням: "узлы пользователей и коммутационных систем" и "гибридное облачное хранилище данных". На втором этапе оптимизации распределяются ресурсы, которые предоставляются CLOUD-системой при многофазной обработке транзакций непосредственно на серверах системы. Сформулированы соответствующие оптимизационные задачи математического программирования, приведены пути их решения.

***Ключевые слова**: хранилище данных, гибридность, облачные технологии, CLOUD-система.*

METHOD OF OPTIMUM ALLOCATION OF CLOUD-SYSTEM RESOURCES

G.A. Kuchuk

A twostage going is considered near optimum allocation of resources of the CLOUD-system. On the first stage the process of distributing of great number of transactions is analysed, his decouplig is conducted on two levels: "knots of users and interconnect systems" and "hybrid cloudy depository of information". On the second stage of optimization resources which are given the CLOUD-system at polyphase treatment of transactions directly on the servers of the system are distributed. The proper optimization tasks of the mathematical programming are formulated, the ways of their decision are resulted.

***Keywords**: depository of information, hybrid, cloudy technologies, CLOUD-system.*