Теоретичні основи розробки систем озброєння

УДК 621.3

О.В. Барабаш[1], М.І. Науменко[2], Ю.В. Стасєв[3]

[1]*Національний університет оборони України, Київ*
[2]*Міністерство Оборони України, Київ*
[3]*Харківський університет Повітряних Сил ім. І. Кожедуба, Харків*

ОПТИМІЗАЦІЯ РОЗПОДІЛУ ПОТОКУ ЗАПИТІВ МІЖ СЕРВЕРАМИ ПРИ ЦЕНТРАЛІЗОВАНІЙ ОБРОБЦІ ДАНИХ

Розглянута багатосерверна система з централізованою обробкою даних, в якій за рахунок високої інтенсивності потоків запитів на виконання додатків утворюються черги до вузлів обробки запитів. Запропонований метод розподілу ресурсів серверів обробки запитів, що дозволяє мінімізувати величину витрат, пов'язаних з простоями запитів в черзі на обробку і простоями серверів.

Ключові слова: *сервер, запит, обробка даних, інтенсивність, черга.*

Вступ

У сучасних мультисервісних мережах (МСС), що включають багатосерверні вузли з централізованою обробкою запитів підмереж вузла, досить часто виникають черги запитів додатків до ряду серверів при недозавантаженості останніх [1, 2]. Тому в даному випадку необхідне використання управління потоками внутрішніх запитів вузла для більш рівномірного розподілу його ресурсів, особливо при перевантаженнях серверних буферів. Дане питання розглядалося в [3 – 5], проте в пропонованих підходах не враховувалися коефіцієнти витрат при знаходженні додатку в черзі. Тому **метою даної статті** є розробка методу оптимізації розподілу потоку запитів додатків між серверами локального вузла мультисервісної мережі при централізованій обробці даних усередині вузла по критерію зменшення витрат.

Основна частина

Задача розподілу ресурсів багатосерверного вузла обробки інформації зустрічається на практиці в тих випадках, коли в кожній підмережі перевага віддається централізованій обробці і зберіганню даних. Проте при цьому різко збільшується число користувачів централізованих засобів обробки та інтенсивність потоків запитів на виконання додатків. Для виконання вимог до якості вирішення задач (наприклад, часу рішення) необхідно використовувати багатосерверні вузли зберігання даних і обробки запитів. Для таких вузлів характерна поява задач управління розподілом потоку запитів між серверами. Цю задачу і розглядатимемо нижче.

Розглядатимемо вузол, що складається з N серверів, кожен з яких може обслуговувати всі додатки, відповідні задачам, що вирішуються в мережі, та які складають множину $Q = \{q_1, ..., q_J\}$, де J – кількість додатків.

На вхід вузла надходять пуассоновські потоки запитів на запуск додатків, інтенсивності потоків відповідають інтенсивностям запуску задач, що включають ці додатки.

Позначимо як λ_j інтенсивність потоку запитів на виконання додатку q_j. Інтенсивності λ_j складають вектор $\Lambda = (\lambda_1, \lambda_2, ..., \lambda_D)$. Величина інтенсивності λ_j може бути обчислена як

$$\lambda_j = \sum_{k=1}^{K} \lambda_K \cdot a_{Kj}, \qquad (1)$$

де $a_{Kj} = \begin{cases} 1, \text{якщо додаток } q_j \text{ використовується} \\ \text{при вирішенні задачи } Z_k; \\ 0, \text{впротивномувипадку;} \end{cases}$

$Z = (Z_1, ..., Z_K)$ – множина вирішуваних задач за умови що додаток встановлений тільки один раз.

Позначимо ймовірність напряму запиту на запуск додатку q_j на сервер з номером n як P_{jK}. Для ймовірностей повинні виконуватися наступні умови:

1. Для кожного $j = 1, 2, ..., J$ $\sum_{n=1}^{N} p_{jn} = 1$, тобто потік запитів на запуск додатку кожного типу повинен розподілятися між серверами повністю.

2. Для кожного $n = 1, 2, ..., N$ – $\sum_{j=1}^{N} p_{jn} \geq 0$, тобто на кожен сервер повинні поступати запити на виконання додатків.

Імовірності p_{jn} складають матрицю $P = \|p_{jn}\|, (j = 1,2...,J;\ n = 1,2,...,N)$.

Тривалість роботи додатку q_j на сервері n вважаємо випадковою величиною з функцією розподілу $F_{jn}(t)$. Ця величина має кінцеві перший і другий моменти:

$$0 < m_{jn} = \int_0^\infty t dF_{jn}(t) < \infty; \qquad (2)$$

$$0 < d_{jn} = \int_0^\infty t^2 dF_{jn}(t) < \infty. \qquad (3)$$

Вважатимемо також, що всі сервери працюють незалежно один від одного.

При цьому як модель досліджуваної системи серверів можна розглядати сукупність з K однолінійних систем масового обслуговування (СМО) типу $M/G/1/\infty$ [5, 6]. Отже, розглядатимемо модель роботи кожного сервера як СМО типу $M/G/1/\infty$, на вхід якої поступають пуассоновські потоки запитів на запуск додатків.

Сервер відповідає обслуговуючому пристрою в СМО. Номер СМО співпадає з номером сервера.

Інтенсивність потоку запитів на запуск додатку P_j, що поступає на вхід СМО номер n (сервер n), обчислюється за формулою:

$$\lambda_{jn} = \lambda_j p_{jn}, \quad (j = 1,2,...,J;\ n = 1,2,...,N). \qquad (4)$$

Даний потік є пуассоновським, оскільки виходить з потоку запитів на запуск додатку p_j із застосуванням процедури просіювання [7].

Спочатку досліджуємо роботу однієї СМО. Вважаємо, що всі запити на кожному сервері утворюють одну чергу і обслуговуються в порядку надходження в чергу.

Сумарний потік запитів на сервер номер n має інтенсивність:

$$\Lambda_n = \sum_{j=1}^{J} \lambda_{jn} = \sum_{j=1}^{J} p_{jn} \lambda_j. \qquad (5)$$

Імовірність того, що запит, узятий з черги до сервера номер n, буде запитом на запуск додатку номер p_j, обчислюється за формулою:

$$p_{jn}^{(n)} = \frac{\lambda_{jn}}{\Lambda_n}. \qquad (6)$$

Використовуючи перетворення Лапласа-Стилтьєса функції розподілу тривалості обробки довільного запиту на сервері n, отримаємо:

$$\beta_n(s) = \int_0^\infty e^{-st} dF_n(t); \qquad (6)$$

$$\beta_{nj}(s) = \int_0^\infty e^{-st} dF_{nj}(t), \qquad (7)$$

де F_n – функції розподілу тривалості обробки запитів при зайнятості сервера n.

Використовуючи (6), (7) і відомі результати для СМО типу $M/G/1/\infty$, середній час очікування в черзі для будь-якого запиту на сервері номер n можна обчислити за формулою [8]:

$$T_n = \frac{\Lambda_n d_n}{2(1 - \Lambda_n m_n)}, \qquad (8)$$

де $0 < m_n = \int_0^\infty t dF_n(t) < \infty$ і $0 < d_n = \int_0^\infty t^2 dF_n(t) < \infty$.

При розподілі запитів повинні виконуватися умови, що запобігають перевантаженню серверів, [8]:

$$p_n = \Lambda_n m_n < 1, \ (n = 1,2,...,N). \qquad (9)$$

Імовірність простою сервера n обчислюється за формулою:

$$P_{0n} = 1 - \Lambda_n m_n. \qquad (10)$$

Таким чином, отримані формули для обчислення характеристик роботи одного сервера. Проте, всі сервери розділяють потоки запитів між собою, тому необхідно досліджувати їх спільну роботу по обслуговуванню запитів.

Визначимо якість роботи багатосерверного вузла, і, відповідно, якість управління розподілом ресурсів вузла, наступним функціоналом:

$$Z(N, P, \Lambda) = \sum_{n=1}^{N} a_n T_n + \sum_{n=1}^{N} s_n P_{0n}.$$

Даний функціонал дозволяє обчислити величину витрат, пов'язаних з простоями запитів в черзі на обробку і простоями серверів. Коефіцієнти a_n, s_n – це штрафи за одиницю часу очікування запиту в черзі до сервера номер n і одиницю часу простою сервера номер n.

В цьому випадку завдання оптимального управління розподілом ресурсів багатосерверного вузла ставиться таким чином.

Початкові дані:

- число задач, що вирішуються на мережі, – K; число додатків, що виконуються при вирішенні завдань, – J; число серверів – N;
- матриця інтенсивностей потоків запитів на запуск завдань $\Lambda = \|\lambda_{nk}\|, (n = 1,2,...,N;\ k = 1,2,...,K)$;
- матриця імовірностей розподілу потоків запитів по серверах $P = \|p_{jn}\|, (j = 1,2,...,J;\ n = 1,2,...,N)$;
- матриця апріорно заданих імовірностей розподілу потоків запитів по серверах –

$P^* = \left\| p^*_{jn} \right\|, (j = 1,2,...,J; n = 1,2,...,N)$, де $p^*_{jn} = 1$, якщо потік запитів на запуск додатку q_j обов'язково направляється на сервер номер n, і $p^*_{jn} = 0$, якщо потік запитів на запуск додатку номер J може розподілятися між декількома серверами, так, що в матриці $p = \left\| p_{jn} \right\|$, $p_{jn} = 1$, якщо $q^*_{jn} = 1$ для всіх $j = 1,2,...,J; n = 1,2,...,N$;

- множина вагових (вартісних) коефіцієнтів витрат, пов'язаних з очікуванням запитів в черзі протягом одиниці часу, – $A = \{a_1, a_2, ..., a_N\}$;

- множина допустимих значень інтенсивностей потоків запитів, що поступають на сервери, – $(\bar{\Lambda}_1, \bar{\Lambda}_2, ..., \bar{\Lambda}_N)$.

Цільова функція задачі:

$$Z(N, p^*, \Lambda) =$$

$$= \min_Q \left\{ F(N, p, \lambda) = \sum_{n=1}^{N} a_n T_n(P, \Lambda) + \sum_{n=1}^{N} s_n P_{0n}(p, \Lambda) \right\}.$$

Обмеження завдачі:

1) $\sum_{n=1}^{N} P_{jn} = 1$, для всіх $j = 1,2,...,J$;

2) $p_n = \Lambda_n m_n < 1$, для всіх n;

3) $\Lambda_i \le \bar{\Lambda}_i$, для всіх $i = 1,2,...,N$;

4) $p_{jn} = 1$, якщо, для всіх $j = 1,2,...,J$; $n = 1,2,...,N$.

У даній постановці задача дозволяє визначити оптимальну за рівнем витрат матрицю $P^* = \left\| p^*_{jn} \right\|, (j = 1,2,...,J; n = 1,2,...,N)$.

Дана задача є задачею математичного програмування і для її вирішення можна застосовувати відомі методи [9, 10].

Висновки

Розглянутий підхід до рішення задачі розподілу потоку запитів між серверами при централізованій обробці даних. Запропонований метод, що дозволяє оптимально управляти смугою пропускання магістрального каналу мультисервісної мережі, що дозволяє мінімізувати витрати з метою підвищення доходу від експлуатації каналу. **Напрям подальших досліджень** пов'язаний з розробкою алгоритму та програмної реалізації запропонованого методу.

Список літератури

1. *Телекоммуникационные системы и сети. Т. 3. Мультисервисные сети / В.В. Величко, Е.А. Субботин, В.П. Шувалов, А.Ф. Ярославцев. – М.: Горячая линия – Телеком, 2005. – 592 с.*
2. *Гургенидзе А.Т. Мультисервисные сети и услуги широкополосного доступа / А.Т. Гургенидзе, В.И. Кореш. – М.: Наука и техника, 2003. – 400 с.*
3. *Основи інформаційних систем / За ред. В.Ф. Ситника. – К.: КНЕУ, 2001. – 420 с.*
4. *Ромашкова О.Н. Обработка пакетной нагрузки в информационных сетях / О.Н. Ромашкова. – М.: МИИТ, 2001. – 244 с.*
5. *Олифер В.Г. Компьютерные сети: принципы, технологии, протоколы. 3-е изд. / В.Г. Олифер, Н.А. Олифер. – СПб.: Питер, 2008. – 958 с.*
6. *Танненбаум Э. Распределенные системы. Принципы и парадигмы / Э. Танненбаум, М. Ван Стен. – СПб.: Питер, 2003. – 877 с.*
7. *Кокс Д. Теория восстановления / Д. Кокс, В. Смит. – М.: Сов. радио, 1967. – 300 с.*
8. *Феллер В. Введение в теорию вероятностей и ее применения. В 2-х т.; пер с англ. / В. Феллер. – М.: Мир, 1987. – Т.1. – 528 с. – Т.2. – 738 с.*
9. *Моисеев Н.Н. Математические задачи системного анализа / Н.Н. Моисеев. – М.: Наука, 1981. – 488 с.*
10. *Кофман А. Методы и модели исследования операций / А. Кофман, А. Анри-Лабордер. – М.: Мир, 1977. – 432 с.*

Надійшла до редколегії 26.05.2011

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

ОПТИМИЗАЦИЯ РАСПРЕДЕЛЕНИЯ ПОТОКА ЗАПРОСОВ МЕЖДУ СЕРВЕРАМИ ПРИ ЦЕНТРАЛИЗОВАННОЙ ОБРАБОТКЕ ДАННЫХ

О.В. Барабаш, Н.И. Науменко, Ю.В. Стасев

Рассмотрена многосерверная система с централизованной обработкой данных, в которой за счет высокой интенсивности потоков запросов на выполнение приложений образуются очереди к узлам обработки запросов. Предложен метод распределения ресурсов серверов обработки запросов, позволяющий минимизировать величину затрат, связанных с простоями запросов в очереди на обработку и простоями серверов.

Ключевые слова: *сервер, запрос, обработка данных, интенсивность, очередь.*

OPTIMIZATION OF DISTRIBUTING OF STREAM OF QUERIES BETWEEN SERVERS AT THE CENTRALIZED PROCESSING OF DATA

O.V. Barabash, N.I. Naumenko, Yu.V. Stasev

The multiserver system is considered with the centralized processing of data, in which due to high intensity of streams of requests for implementation of appendixes turns appear to the knots of processing of inquiries. The method of allocation of resources of servers of processing of inquiries is offered, allowing to minimize the size of expenses, related to the outages of queries in and process queue and outages of servers.

Keywords: *server, query, processing of data, intensity, turn.*