УДК 004.7+621.391

ОРГАНІЗАЦІЯ УПРАВЛІННЯ В БАГАТОЕТАПНИХ СИСТЕМАХ МАСОВОГО ОБСЛУГОВУВАННЯ

Скулиш М.А., Суліма С.В.
Інститут телекомунікаційних систем НТУУ "КПІ", Київ, Україна
E-mail: mb_s@ukr.net, lilthirteen@gmail.com

Management of multiple stage queuing systems

The problems of multiple stage multiple resource multiple service systems are discussed. A dynamic provisioning technique for multi-stage systems that employs a flexible analytical model to determine how much resources to allocate to each stage of the system is presented.

Розглянуто проблеми систем з кількома етапами, кількома типами ресурсів, кількома типами сервісів. Представлено метод динамічного управління ресурсами для багатоетапних систем, який використовує гнучку аналітичну модель для визначення кількості ресурсів, яку необхідно виділити для кожного етапу системи.

Широкий спектр сучасних систем обслуговування передбачає поетапну обробку запитів. Так само система розрахунків мобільного оператора передбачає виконання ряду операцій. Сьогодні все більшої популярності набуває використання віртуальних серверів, які легко налаштовуються під зміну навантаження. Існують різні підходи до управління числом виділених ресурсів для кожної підсистеми.

Динамічне надання ресурсів вивчалося в контексті програм одного рівня. Розширення таких механізмів на багаторівневі сценарії є нетривіальною задачею. Класичні підходи можуть просто змістити вузьке місце на інший рівень [1]. Крім цього, як правило, розглядаються системи з єдиним типом ресурсів. При цьому задачі розподілу ресурсів в системах багатостадійної обробки розглядаються переважно для Інтернет-систем, тоді як такі задачі виникають в різних областях, в тому числі і в процесі роботи серверів оператора мобільного зв'язку. Крім цього, в більшості робіт стадії (або рівні) розглядаються незалежно або з незалежними пулами ресурсів, наприклад в [2]. Проте в певних випадках необхідно розглядати загальний пул ресурсів, і виникає додаткова задача – задача оптимального розподілу загальних ресурсів між різними етапами. Через труднощі в математичному моделюванні складних характеристик трафіку (наприклад, багатоетапної обробки заявок), більшість досліджень в літературі була виконана з використанням імітаційного моделювання. Тим не менш, аналітична модель системи буде привабливою, оскільки вона зможе оцінити характеристики системи в широкому діапазоні умов, і бути обчисленою порівняно легко.

Таким чином, актуальною є задача побудови аналітичної моделі системи з кількома типами послуг, кількома типами використовуваних ресурсів та кількома етапами обслуговування запитів, особливо у сфері мобільних мереж

зв'язку. Jordan ввів метод моделювання систем з багатьма типами сервісів та ресурсів у [3]. У даній роботі модифіковано та розширено запропонований метод на випадок задачі багатоетапного надання ресурсів та застосовано його для управління датацентром оператора мобільного зв'язку.

Нехай кожному сервісу потрібна певна кількість ресурсів мережі. Нехай K – це кількість вузлів системи. Змоделюємо систему як K-мірний ланцюг Маркова, стан якого визначається наступним чином:

$Z=(Z^1,..., Z^K)$, де Z^j позначає стан вузла j.

Кожен вузол з ланцюга функціональних блоків (рис. 1), який обробляє заявку, моделюється наступним чином.

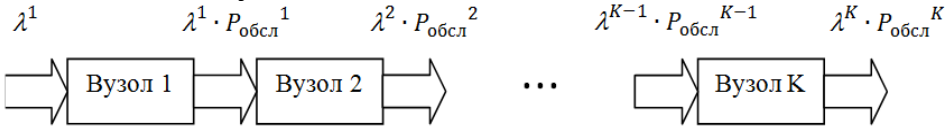

Рис. 1 Модель системи

Розглянемо систему з n типами сервісів, де кожному сервісу потрібна множина з m типів ресурсів. Заявки надходять як незалежні Пуассонівські процеси і займають необхідну кількість кожного ресурсу на однакову тривалість часу, що має експоненційний розподіл. Змоделюємо систему як ланцюг Маркова і використовуємо наступні позначення. Для вузла j, $j=1,2,...K$:

A^j – матриця розміром m x n, де у стовпці i визначено кількість кожного з m типів ресурсів, що потребує сервіс типу i;

b^j – вектор довжиною m, що визначає кількість ресурсів в системі;

$\lambda^j – (\lambda_1{}^j, ..., \lambda_n{}^j)$, інтенсивності надходження заявок на обслуговування;

$\mu^j – (\mu_1{}^j, ..., \mu_n{}^j)$, інтенсивності обслуговування;

$\rho^j – (\rho_1{}^j, ..., \rho_n{}^j)$, навантаження, визначається як $\rho_i^j = \lambda_i^j/\mu_i^j$;

$x^j – (x_1{}^j, ..., x_n{}^j)$, стан системи, де x_i^j – кількість заявок типу i, що оброблюються системою;

$Z^j – \{x^j \mid A^j x^j \le b^j\}$, тобто, x^j заявок можуть одночасно оброблюватись доступними ресурсами;

$F_i^j – \{x^j \mid x^j \in Z^j \text{ but } (x_1{}^j, ..., x_i{}^j+1, x_n{}^j) \notin Z^j\}$, підмножина станів множини Z^j, таких що надходження заявки типу i спричинить її відкидання;

$Eij – \{x^j \mid x^j \in Z^j \text{ but } (x_1{}^j, ..., x_i{}^j–1, x_n{}^j) \notin Z^j\}$, порожня множина для сервісу i;

$\pi(x)$ – стаціонарні ймовірності.

Припущення стосовно процесів надходження та обслуговування дають ланцюг Маркова з простором станів Z^j з інтенсивностями переходів:

$$r_{xy}{}^j = \begin{cases} \lambda_i{}^j, if\ x \notin F_i{}^j\ and\ y = (x_1{}^j, \dots, x_i{}^j + 1, \dots, x_n{}^j) \\ x_i \mu_i, if\ x \notin E_i{}^j\ and\ y = (x_1{}^j, \dots, x_i{}^j - 1, \dots, x_n{}^j) \\ 0, else \end{cases} \quad (1)$$

Ланцюг Маркова зворотний і має добре відому мультиплікативну форму стаціонарного розподілу:

$$\pi^j(x) = \pi^j(0) \prod_{i=1}^{n} \frac{(\rho_i{}^j)^{x_i{}^j}}{x_i{}^j!}, \pi^j(0) = \frac{1}{\sum_{x^j \in Z^j} \prod_{i=1}^{n} \frac{(\rho_i{}^j)^{x_i{}^j}}{x_i{}^j!}} \quad (2)$$

Використовуючи отриманий стаціонарний розподіл, ймовірність успішного обслуговування P_i^j на вузлі j для сервісу i може бути обчислена як:

$$P_i^j = 1 - \sum_l \pi_l^j(x), where\ l = \{x^j | x^j \in F_i^j\} \quad (3)$$

А інтенсивність надходження заявок на обслуговування для вузла j, $j=2,...K$ можна виразити як: $\lambda^j=(\lambda_1^{j-1}P_1^{j-1}, ..., \lambda_n^{j-1}P_n^{j-1})$.

Оскільки чисельний розрахунок добутків стане нерозв'язним для великої кількості класів трафіку, можна застосовувати простіші рекурсивні або приблизні методи розрахунку ймовірностей станів моделі.

Моделювання запропонованого вище методу було проведено в системі Mathcad з використанням чисельної процедури з [4] для знаходження стаціонарного розподілу, яке показало, що можливо отримати виграш у тисячі разів при використанні динамічної процедури перерозподілу ресурсів між етапами при зміні інтенсивності вхідного навантаження. Виграш запропонованого вище методу у порівнянні з методом, за яким ресурси розподіляються між етапами рівномірно, проілюстровано на прикладі рис. 2.

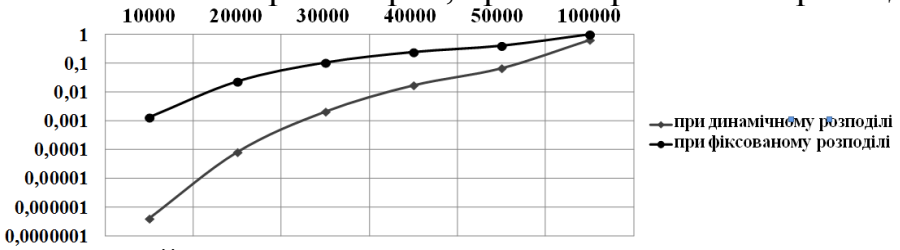

Рис. 2 Ймовірність відмови в обслуговуванні при зміні інтенсивності надходження запитів в логарифмічному масштабі

Слід відмітити, що визначена вище модель є дуже загальною і може застосовуватись для аналізу різних систем. Запропонований інструмент може застосовуватись для вирішення задачі оптимального розподілу ресурсів між вузлами.

Висновки. У роботі встановлено, що динамічне надання ресурсів у багатоетапних телекомунікаційних системах ставить нові задачі, не вирішені у попередніх дослідженнях систем надання ресурсів. Запропоновано розширений динамічний метод управління ресурсами для багатоетапних систем, який застосовує гнучку аналітичну модель системи масового обслуговування для визначення кількості ресурсів, яку необхідно надати кожному етапу системи.

Література

1. Urgaonkar B. Agile dynamic provisioning of multi-tier Internet applications / B. Urgaonkar, P. Shenoy, A. Chandra, P. Goyal, T. Wood // ACM Transactions on Autonomous and Adaptive Systems (TAAS). – 2008. – Vol. 3, No. 1. – pp.1-39.
2. Han R. Enabling cost-aware and adaptive elasticity of multi-tier cloud applications / R. Han, M. M. Ghanem, L. Guo, Y. Guo, M. Osmond // Future Generation Computer Systems. – 2014. – Vol. 32. – pp.82-98.
3. Jordan S. Control of Multiple Service, Multiple Resource Communication Networks / S. Jordan, P. Varaiya // IEEE INFOCOM'91 : proceedings. – 1991. – pp.648-657.
4. Меликов А. З. Телетрафик: модели, методы, оптимизация / А. З. Меликов, Л. А. Пономаренко, В. В. Паладюк. – К.: ИПК "Политехника", 2007. – 256 с.