УДК 519.216.3:004.9

Т. О. Савчук
А. В. Козачук

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ МАСШТАБУВАННЯ ХМАРНОГО ЗАСТОСУНКУ З НЕПЕРІОДИЧНИМИ ПІКАМИ НАВАНТАЖЕННЯ

Вінницький національний технічний університет, Україна

Анотація

Запропоновано математичну модель, що описує роботу хмарного застосунку як багатоканальну систему масового обслуговування з необмеженою чергою. На основі побудованої математичної моделі розраховано середній час виконання мережевого запиту, значення якого необхідне для прийняття рішень щодо масштабування хмарного застосунку.

Ключові слова: хмарні обчислення; прогнозування часу виконання мережевого запиту; СМО.

Abstract

Model that describes cloud application state as a multichannel queuing system with infinitive queue was build. Based on this model we calculate average network request execution time, that can be used when making a decision about cloud application scaling strategy.

Keywords: cloud computing; forecasting of a network request execution time; queuing theory.

Використання хмарних обчислень дозволяє значно полегшити підтримку інфраструктури, збільшити швидкість розгортування застосунку, а також адаптуватися під змінний режим навантаження з періодичними піками. Нерівномірність інтенсивності використання хмарних застосунків робить актуальною задачу оптимізації обчислювальних ресурсів, що виділяються для підтримання роботи хмарного застосунку. Ця задача розв'язується завдяки масштабуванню хмарного застосунку – адаптації кількості виділених обчислювальних ресурсів до завантаженості хмарного застосунку.

Для збільшення ефективності масштабування доцільно здійснювати прогнозування стану хмарного застосунку, зокрема такого параметру його як середній час виконання мережевих запитів.

Представимо хмарний застосунок зі станом інфраструктури $\langle S, N, P \rangle$ як багатоканальну систему масового обслуговування (СМО). Кількість каналів приймемо рівною кількості процесорних ядер *C*, яка залежить від кількості віртуальних машин та їх типу: $C = C(S, N)$. Заявками у такій системі масового обслуговування є мережеві запити до хмарного застосунку. Усі канали обслуговування є однорідними з інтенсивністю потоку обслуговування μ та математичним сподіванням часу обслуговування $\overline{T}_{об}, \overline{T}_{об} = 1/\mu$.

При заповненості всіх каналів очікування запити потрапляють у чергу. При перебуванні у СМО протягом тривалого часу, запит відхиляється, формуючи потік відхилених запитів з інтенсивністю *v*. Чергу будемо вважати необмеженою. Таким чином за дисципліною обслуговування введена СМО належить до СМО змішаного типу. Так як кількість джерел запитів-користувачів хмарного застосунку може бути великою та від одного користувача може надходити кілька паралельних запитів, СМО є відкритою [1].

Багатоканальна система масового обслуговування з необмеженою чергою містить нескінченну кількість станів $S_0, S_1, ..., S_C, S_{C+1}, ...$, де перші *C* станів відповідають хмарному застосунку у якого є незадіяні процесорні ядра та черга запитів – пуста. Граф станів СМО зображений на рисунку 1.

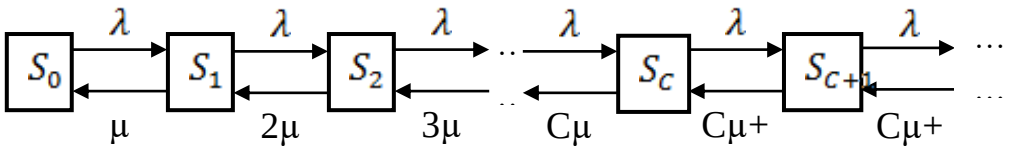

Рисунок 1 – Граф станів СМО хмарного застосунку

Потік запитів послідовно переводить систему з довільного стана у стан, що знаходиться правіше на рисунку з постійною інтенсивністю λ. Інтенсивність потоку обслуговування зростає зі збільшенням кількості задіяних каналів до $C\mu$, коли задіяні всі канали. Також на систему діє потік відхилених запитів, який переводить систему зі стану S_i у стан S_{i-1}. Інтенсивність потоку відхилених запитів зростає при збільшенні черги на v для кожного наступного стану СМО. Відповідно до граничної теореми про складання потоків [2], сумарний потік обслуговування становить $i\mu$ при $i \le C$ та $C\mu + (i-C)\upsilon$ при $i > C$, де i – номер стану СМО. Така СМО може бути описана за допомогою системи диференційних рівнянь (1), розв'язок яких дозволяє визначити середню кількість мережевих запитів у черзі і, відповідно, середній час виконання мережевого запиту.

Об'єднаємо отримані співвідношення в систему диференційних рівнянь:

$$\begin{cases} \dfrac{dp_0}{dt} = \mu p_1 - \lambda p_0 \\ \dfrac{dp_1}{dt} = 2\mu p_2 - (\lambda + \mu)p_1 + \lambda p_0, \\ \dots \\ \dfrac{dp_i}{dt} = (i+1)\mu p_{i+1} - (\lambda + i\mu)p_i + \lambda p_{i-1}, i < C \\ \dots \\ \dfrac{dp_i}{dt} = (C\mu + (i-C+1)\nu)p_{i+1} - (\lambda + C\mu + (i-C)\nu)p_i + \lambda p_{i-1}, i \ge C \\ \dots \end{cases} \quad (1)$$

Тоді, середній час перебування мережевого запиту у СМО хмарного застосунку визначається як сума середнього часу виконання мережевого запиту та часу перебування в черзі. При цьому, останній, в свою чергу, дорівнює часу обробки одного запиту помноженому на кількість запитів в черзі:

$$\overline{R} = \overline{T}_{об} + k\overline{T}_{об} = \frac{1}{\mu} + \bar{k}\frac{1}{\mu} = \frac{\bar{k}+1}{\mu}, \quad (2)$$

де $\bar{k}$ – середня кількість запитів у черзі, $\overline{R}$ – середній час виконання мережевого запиту.

Такии чином, було розраховано середній час виконання мережевого запиту за рахунок застосування теорії систем масового обслуговування до моделі хмарного застосунку, значення якого необхідно для роботи інформаційної технолонії масштабування хмарного застосунку з неперіодичними піками навантаження.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Саакян Г.Р. Теория массового обслуживания[Текст] / Г. Р. Саакян // Шахты, 2006.
2. Гнеденко Б. В. Введение в теорию массового обслуживания/ Б. В. Гнеденко, И. Н. Коваленко// М. "Наука", 1966 - 434с.

Савчук Тамара Олександрівна — к.т.н,,доцент, систем, професор кафедри комп'ютерних наук ВНТУ, Вінницький національний технічний університет, м. Вінниця

Козачук Андрій Валерійович — асистент кафедри комп'ютерних наук ВНТУ, Вінницький національний технічний університет, м. Вінниця, e-mail: kozachuk35@rambler.ru

Tamara O. Savchuk — Cand. Sc. (Eng), Assistant Professor, Professor of the Computer Sciences Chair, Vinnytsia National Technical University, Vinnytsia

Andriy V. Kozachuk — assistent of the Computer Sciences Chair, Vinnytsia National Technical University, Vinnytsia, e-mail: kozachuk35@rambler.ru