УДК 519.872

ИССЛЕДОВАНИЕ ХАРАКТЕРИСТИК ПРОИЗВОДИТЕЛЬНОСТИ СИСТЕМЫ ПОТОКОВЫХ ВЫЧИСЛЕНИЙ

А.М. Соколов

Институт проблем управления им. В.А. Трапезникова РАН

Россия, 117997, Москва, Профсоюзная ул., 65

E-mail: aleksandr.sokolov@phystech.edu

Ключевые слова: потоковые вычисления, система массового обслуживания, цепь Маркова, пуассоновский процесс.

Аннотация: В статье представлено исследование характеристик производительности системы потоковых вычислений с использованием аппарата теории массового обслуживания. Исследуемая система массового обслуживания состоит из буфера емкости N и M обслуживающих приборов. Время обслуживания имеет экспоненциальное распределение. Заявки, поступающие в систему в пуассоновском потоке, состоят из случайного числа задач в диапазоне от 1 до K. Каждая отдельная задача обслуживается на отдельном приборе в порядке очереди. Порядок обслуживания определяется по принципу FIFO. В работе представлено описание системы, цепь Маркова, описывающая систему, и формулы для расчета характеристик производительности.

1. Введение

При исследовании большого класса задач для получения достаточно точных результатов требуется много однотипных вычислений. Например, в задачах прогнозирования, визуализации данных или задачах машинного обучения, где для получения результатов требуется сгенерировать синтетический набор данных. Время получения численных результатов в задачах данного типа варьируется от нескольких минут до нескольких недель в зависимости от сложности вычислений. Для ускорения получения результатов существуют системы для потоковых вычислений. Суть таких систем заключается в параллельном выполнении сразу нескольких задач. Каждая задача занимает определенный сервер (обслуживающий прибор) и выполняется на нем. Если в системе предусмотрено наличие M серверов, то, соответственно, есть возможность одновременно выполнять M задач. Примерами таких систем являются [1, 2]. Авторы данной статьи использовали систему потоковых вычислений [3, 4] для генерации синтетического набора данных для исследования сложной системы массового обслуживания (СМО) [5] с приоритетным обслуживанием. Используя подобные инструменты, можно сократить время получения численных результатов в десятки раз.

Аналитическая модель системы потоковых вычислений по классу является СМО

с групповым поступлением. Аналитические модели с групповым поступлением были исследованы и ранее. Например, в работе [6] исследована системе $M[b]/M/1$, где поступают группы с фиксированным размером, в системе предусмотрен только один обслуживающий прибор. В работе [7] исследована СМО вида $M[b]/M/N$, где так же поступают групповые заявки фиксированного размера, но предусмотрено N обслуживающих приборов.

В данной статье приведено описание математической модели системы потоковых вычислений в предположении, что входящий поток заявок является пуассоновским с интенсивностью λ, время обслуживания имеет экспоненциальное распределение с показателем μ, а число задач в заявке случайно и распределено по некоторому закону. Вероятность того, что в заявке содержится ровно $k > 0$ задач равна b_k, $\sum_{k=1}^{K} b_k$, где K – максимальное число задач в заявке.

2. Аналитическая модель

В системе, изображенной на рис.1, на вход поступают групповые заявки с интенсивностью λ, состоящие из нескольких задач. Число задач в заявке обозначим через $k = 1, .., K$, где K – максимальное число задач в заявке. Обозначим b_k – вероятность того, что в заявке k задач для обслуживания, $\sum_{k=1}^{K} b_k = 1$. Система состоит из буфера конечной емкости N, где находятся заявки в ожидании обслуживания и M обслуживающих приборов. Стоит отметить, что одновременно на обслуживании могут находиться задачи разных заявок, например, в случае, когда в заявке осталось обработать задач меньше M (число приборов). Время обслуживания задачи на обслуживающем приборе имеет экспоненциальное распределение с параметром μ. Также для построения цепи Маркова будем рассматривать случай, когда $K > M$, то есть максимальное число задач в заявке больше числа обслуживающих приборов.

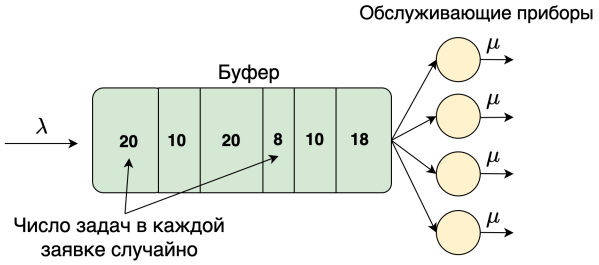

Рис. 1. СМО с групповым поступлением задач. СМО содержит буфер конечной емкости N, M обслуживающих приборов, число задач в каждой приходящей заявке случайно и распределено по некоторому распределению B

2.1. Цепь Маркова

Для построения цепи Маркова будем говорить, что заявка находится на обслуживании или частично выполнена, если хотя бы одна из задач данной заявки находится на обслуживании, при этом есть задачи данной заявки, которые ожидают

обслуживания. Будем говорить, что заявка находится в буфере или ожидает обслуживания, если все ее задачи ожидают обслуживания.

Каждое состояние цепи Маркова данной СМО описывается тройкой чисел n, r, m:

- n – число заявок в буфере;
- r – общее число задач частично выполненных заявок, ожидающих выполнения;
- m – число занятых приборов обслуживания.

Изменение состояний описывается цепью Маркова ξ_t с непрерывным временем

$$\xi_t = \{(n, r, m), t \geqslant 0\} \tag{1}$$

и пространством состояний

$$\Omega = \{(n, r, m), m \leqslant M, n = 0, r = 0 \cup (n, r, m), m = M, 0 \leqslant n \leqslant N, 0 \leqslant r \leqslant K\} \tag{2}$$

Обозначим $p(n, r, m)$ вероятность нахождения в состоянии (n, r, m) в стационарном режиме. Составив уравнение баланса для каждого состояния, и, записав условие нормировки $\sum p(n, r, m) = 1$, можно найти стационарное распределение вероятностей для данной СМО.

2.2. Характеристики производительности системы

Формулы для вычисления характеристик производительности системы:

- среднее число заявок в буфере:

$$\bar{N_b} = \sum_{n=1}^{N} \sum_{r=1}^{K} np(n, r, M) \tag{3}$$

- среднее число занятых приборов:

$$\bar{M} = \sum_{m=1}^{M} mp(0, 0, m) + M(1 - \sum_{m=0}^{M} p(0, 0, m)) \tag{4}$$

первая сумма суммирование состояний, где $m \leqslant M$ и $r = 0$ (в системе отсутствуют заявки, ожидающие обслуживания), вторая сумма – соответствует всем состояниям, в которых заняты все обслуживающие приборы;

- среднее число задач в буфере:

$$\bar{T_b} = \sum_{r=1}^{K} kp(0, r, M) + \sum_{n=1}^{N} \sum_{r=1}^{K} p(n, r, M)(n\bar{B} + r), \tag{5}$$

где первое слагаемое учет состояний, в которых нет заявок в буфере. Во втором слагаемом учитываются задачи нераспакованных заявок, где $\hat{B}$ – среднее число задач в заявке, соответственно, $n\hat{B}$ – среднее число задач в n заявках, также здесь учитываются задачи, принадлежащие частично выполненным заявкам;

- среднее число задач в системе:

$$\bar{T} = \bar{T}_b + \bar{M} \tag{6}$$

складывается из среднего числа задач, находящихся в буфере и числа занятых приборов;

3. Численные результаты

Решение системы уравнений баланса, получение стационарного состояния и вычисление характеристик производительности системы реализовано на языке Python. Код доступен по ссылке. В рамках численного эксперимента исследовалось влияние интенсивности входного потока заявок, интенсивности обслуживания и числа обслуживающих приборов на основные характеристики производительности системы.

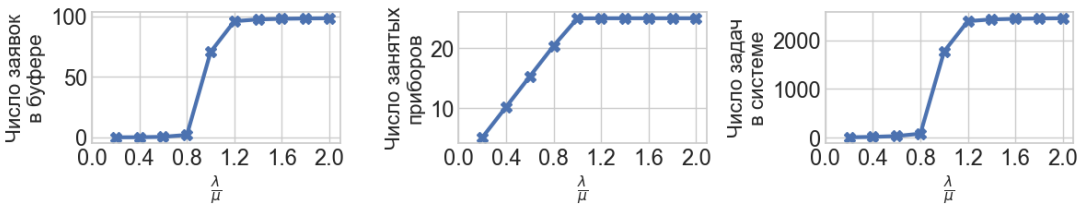

Рис. 2. Характеристики производительности СМО в зависимости от $\frac{\lambda}{\mu}$

На рис.2 показаны графики зависимости характеристик производительности системы от отношения $\frac{\lambda}{\mu}$. Эксперимент проводился при следующих значениях параметров: $\mu = 0.5$, $M = 25$, $b_k = 0.02$, $K = 50$, $N = 100$, а значение λ изменялось в пределах $(0.1, 0.2, .., 1)$. Из графиков видно, что при соотношении $\frac{\lambda}{\mu} = 1.2$ размер системы достигает своего максимума, буфер системы полностью заполнен.

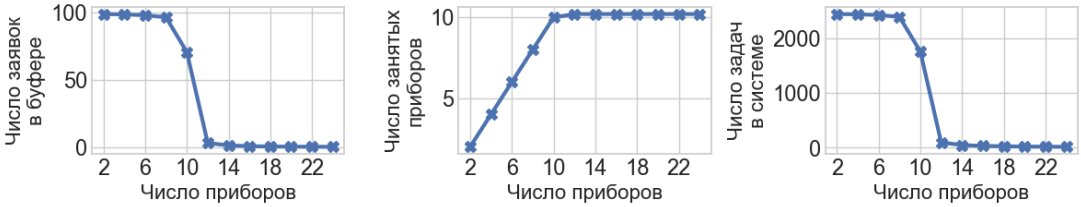

Рис. 3. Характеристики производительности СМО в зависимости от числа обслуживающих приборов

На рис.3 показана зависимость характеристик производительности системы от числа обслуживающих приборов. В данном эксперименте параметры системы были

равны: $\mu = 0.5$, $\lambda = 0.2$, $b_k = 0.02$, $K = 50$, $N = 100$, а значение M менялось в пределах $(2, 4, 6, .., 24)$. При $m = 10$ система выходит из режима насыщения, буфер не полностью заполнен.

4. Заключение

В работе приведено описание математической модели системы для потоковых вычислений. В систему поступают заявки, состоящие из нескольких задач. В статье рассмотрен случай, при котором поток входящих заявок пуассоновский, время обработки задачи распределено по экспоненциальному закону, число задач имеет произвольное дискретное распределение. Приведены формулы для вычисления характеристик производительности системы.

Исследование выполнено за счет гранта Российского научного фонда № 22-49-02023, https://rscf.ru/project/22-49-02023/.

Список литературы

1. Arora R., Redondo C., Joshua G. Scalable software infrastructure for integrating supercomputing with volunteer computing and cloud computing // Communications in Computer and Information Science. 2019. P. 105–119.
2. Sukhoroslov O., Putilina E. Cloud Services for Automation of Scientific and Engineering Computations Science // Business. Soc. 2018. Vol. 1, No. 2. P. 6–9.
3. Соколов А.М., Ларионов А.А., Вишневский В.М., Мухтаров А.А. Архитектура Распределенной Системы Для Потоковых Вычислений С Контейнеризацией И Приоритизацией Задач // Информационные Технологии И Вычислительные Системы. 2023. № 4. С. 5-18.
4. Sokolov A., Larionov A., Mukhtarov A., Fedotov I. Architecture of a Distributed Parallel Computing System Using Docker Cluster // Proc. 2022 Int. Conf. Information, Control. Commun. Technol. ICCT 2022. 2022.
5. Vishnevsky V., Klimenok V., Sokolov A., Larionov A. Performance evaluation of the priority multi-server system mmap/ph/m/n using machine learning methods // Mathematics. 2021. Vol. 9, No. 24.
6. Ghimire S., Ghimire R. P., Thapa G. B. Mathematical Models of Mb/M/1 Bulk Arrival Queueing System // J. Inst. Eng. 2014. Vol. 10, No. 1. P. 184–191.
7. Kumar J., Shinde V. Performance Evaluation Bulk Arrival and Bulk Service with Multi Server using Queue Model // International Journal of Research in Advent Technology. 2018. Vol.6, No.11.