УДК 519.872:621.394.74

Мехбалыева Э.В.

СХЕМА ОПРЕДЕЛЕНИЯ СКАЧКООБРАЗНЫХ ПРИОРИТЕТОВ В СИСТЕМАХ ОБСЛУЖИВАНИЯ С ГЕТЕРОГЕННЫМИ СЕРВЕРАМИ

Сумгаитский Государственный Университет, Азербайджан

esmira.mehbaliyeva@mail.ru

Введение

Системы обслуживания с гетерогенными серверами (Heterogeneous Servers, HS) очень часто встречаются при разработке математических моделей процессов обработки запросов в компьютерных и коммуникационных системах, так как в процессе расширения существующих систем приходится использовать сервера с различными показателями, т.е. в единой сети используются сервера с различными скоростями обработки запросов, различными надежностными показателями и т.д. Такие ситуации также встречаются в системах, где в процессе обработки запросов участвуют не машины (аппараты, компьютеры и т.д.), а люди.

В последние годы модели систем с HS интенсивно исследуются различными авторами. Подробный обзор работ в этом направление можно найти в работах [1-3]. Последние исследования показали, что мощным средством повышения эффективности этих систем является использования скачкообразных приоритетов (Jump Priorities, JP), которые впервые были введены в работах [4-6].

В известных до сих пор работах изучались модели систем обслуживания с HS при наличии идентичных заявок, т.е. не были изучены модели систем обслуживания с HS и разнотипными заявками. Однако очевидно, что для повышения экономической эффективности (относительно выбранного критерия качества) работы системы с HS следует выделить высокоприоритетные (h-заявки) и низкоприоритетные заявки (l-заявки), т.е. следует организовать обработки h-заявок в высокоскоростных серверах (f-серверах), а медленные серверы (s-сервера) назначаются для обслуживания l-заявок.

Первая публикация, которая была посвящена к изучению моделей систем с HS и разнотипными заявками при наличии JP, была работа [7]. В ней показана актуальность изучения систем с зависящими от состояния системы скачкообразными приоритетами, при этом зависимость JP от состояния может быть учтена различными способами. В указанной работе изучена модель, в которой JP зависят от разности текущего числа разнотипных заявок в системе, а именно, если разность числа разнотипных заявок в системе превышает определенное пороговое значение, то одна l-заявка с определенной вероятностью может переходит в очередь h-заявок, т.е. в указанной схеме определения JP имеется только один степень свободы. Такой переход позволяет улучшить показатели качества обслуживания l-заявок, но при этом ухудшается показатели качества обслуживания h-заявок. Потому в [7] были решены различные задачи оптимизации введенных приоритетов.

В данной работе предложена новая схема определения скачкообразных приоритетов: переход l-заявки в очередь h-заявок зависит от конкретного числа заявок каждого типа. Иными словами, здесь, в отличие от работы [7], имеются две степени свободы, т.е. за счет выбора двух параметров можно управлять показателями качества функционирования системы. Здесь разработана модель систем с гетерегенными серверами и указанными скачкообразными приоритетами, при этом рас-

сматривается модель с раздельными очередями разнотипных заявок. Предложен метод расчета ее характеристик.

Описание модели и постановка задачи

Изучаемая система содержит два гетерогенных сервера: быстрый (*f*-сервер) и медленный серверы (*s*-сервер). Эти серверы обслуживают потоки заявок двух типов – высокоприоритетных (*h*-заявки) и низкоприоритетных (*l*-заявки), при этом *h*-заявки обслуживаются в *f*-сервере, а *l*-заявки – в *s*-сервере. Оба потока заявок являются пуассоновскими с интенсивностями λ_h и λ_l для *h*-заявок и *l*-заявок соответственно. Время обслуживания заявок в обоих серверах являются случайными величинами (с.в.) с показательной ф.р., при этом среднее времена обслуживания в *f*-сервере и *s*-сервере равны μ_f^{-1} и μ_s^{-1} соответственно. Скорость обслуживания *f*-сервере больше, чем скорость обслуживания *s*-сервера, т.е. $\mu_f > \mu_s$.

Для ожидания заявок в очереди имеются раздельные буфера конечных размеров, при этом максимальное число *h*-заявок и *l*-заявок в системе равны K_h и K_l. Это означает, что размер буфера для *h*-заявок (*h*-буфер) равен $K_h - 1$, а соответствующий буфер для *l*-заявок (*l*-буфер) имеет размерность $K_l - 1$.

Здесь рандомизированные скачкообразные приоритеты определяются следующим образом. Прежде всего отметим, что введенные рандомизированные JP зависят от текущего состояния системы, при этом состояния системы в каждый момент времени задается двумерным вектором (h,l), где компоненты h и l указывают на число *h*-запросов и *l*-запросов в системе соответственно. Для определения указанных приоритетов вводятся два пороговых параметра $r_l, 1 \le r_l \le K_l$, и $r_h, 1 \le r_h \le K_h$.

- Поступающие *h*-заявки всегда присоединяются к *h*-буферу, если там имеется свободное место; иначе эти заявки теряются с вероятностью единица.

- Если в момент поступления *l*-заявки числа заявок такого типа в системе меньше параметра r_l, то не зависимо от числа *h*-заявок в системе эта *l*-заявка присоединяется к *l*-буферу.

- Если в момент поступления *l*-заявки числа заявок такого типа в системе не меньше параметра r_l, и числа *h*-заявок меньше порогового параметра r_h, то один из *l*-заявок (для определенности считается, что *l*-заявка, стоящей в начале очереди) согласно схеме Бернулли либо с вероятностью α присоединяется к *h*-буферу либо поступившая *l*-заявка с вероятностью $1-\alpha$ переходит в конец очереди *l*-буфера (если там имеется свободное место).

- Если в момент поступления *l*-заявки числа заявок такого типа в системе не меньше параметра r_l, и числа *h*-заявок не меньше порогового параметра r_h, то поступившая *l*-заявка с вероятностью единица присоединяется к *l*-буферу, если там имеется свободное место; иначе поступившая *l*-заявка теряется с вероятностью единица.

- Если в момент поступления *l*-заявки соответствующий буфер полностью заполнен и числа *h*-заявок меньше указаннного выше порогового параметра r_h, то *l*-заявка, стоящей в начале очереди либо с вероятностью α присоединяется к *h*-буферу либо поступившая *l*-заявка теряется с вероятностью $1-\alpha$.

Следовательно, в предложенной схеме рандомизированные JP определяются так:

$$J(h,l) = \begin{cases} \alpha, & \text{если } l > r_l, h < r_h, \\ 0 & \text{в других случаях.} \end{cases} \qquad (1)$$

Если в соотношение (1) принимать, что $\alpha = 1$, то получаются детерминированные JP, т.е. каждый раз, когда в момент поступления *l*-заявки выполняется условие $l > r_l, h < r_h$, то *l*-заявка присоединяется к *h*-буферу; в случае $\alpha = 0$ исходная система распадается на две сепарабельных систем обслуживания.

Задача состоит в нахождении совместного распределения числа разнотипных заявок в системе и разработке метода вычисления ее характеристик.

Расчет вероятностей состояний системы

Математической моделью системы является двумерная цепь Маркова (Two Dimensional Markov Chain, 2D MC) с состояниями вида (h,l), при этом ее пространство состояний (ПС) определяется так:

$$E = \{(h,l): h = \overline{0,K_h}, l = \overline{0,K_l}\}. \quad (2)$$

Находим элементы производящей матрицы данной 2D MC, которые определяют интенсивности переходов между ее состояниями. С этой целью в ПС (2) выделим следующее подпространство:

$$E_a = \{(h,l) \in E : h < r_h, l > r_l\}. \quad (3)$$

Интенсивность перехода из состояния $(h,l) \in E$ в состояние $(h',l') \in E$ обозначим через $q((h,l),(h',l'))$. Переходы между состояниями возможны лишь в моменты поступления разнотипных заявок и в моменты завершения их обслуживания. Исходя из этого заключаем, что указанные величины определяются из следующих соображений:

- Если в момент поступления *h*-заявки система находится в состоянии $(h,l) \in E$, где $h < K_h$, то система переходит в состояние $(h+1,l) \in E$ с интенсивностью λ_h.
- Если в момент поступления *l*-заявки система находится в состоянии $(h,l) \in E - E_a$, где $l < K_l$, то система переходит в состояние $(h,l+1) \in E$ с интенсивностью λ_l.
- Если в момент поступления *l*-заявки система находится в состоянии $(h,l) \in E_a$, где $l < K_l$, то система переходит в состояние $(h,l+1) \in E$ с интенсивностью $\lambda_l \alpha$.
- Если в момент поступления *l*-заявки система находится в состоянии $(h,l) \in E_a$, где $l < K_l$, то система переходит в состояние $(h+1,l) \in E$ с интенсивностью $\lambda_l (1-\alpha)$.
- Если в момент завершения обслуживания *h*-заявки система находится в состоянии $(h,l) \in E$, где $h > 0$, то система переходит в состояние $(h-1,l) \in E$ с интенсивностью μ_f.
- Если в момент завершения обслуживания *l*-заявки система находится в состоянии (h,l), где $l > 0$, то система переходит в состояние $(h,l-1)$ с интенсивностью μ_s.

Отсюда заключаем, что искомые величины определяются следующим образом:

случаи $(h,l) \in E - E_a$:

$$q((h,l),(h',l')) = \begin{cases} \lambda_h, & \text{если } h < K_h, h' = h+1, l' = l, \\ \lambda_l, & \text{если } l < K_l, h' = h, l' = l+1, \\ \mu_f, & \text{если } h > 0, h' = h-1, l' = l, \\ \mu_s, & \text{если } l > 0, h' = h, l' = l; \end{cases} \quad (4)$$

случаи $(h,l) \in E_a$:

$$q((h,l),(h',l')) = \begin{cases} \lambda_h + \alpha\lambda_l, & \text{если } h' = h+1, l' = l, \\ (1-\alpha)\lambda_l, & \text{если } h' = h, l' = l+1, \\ \mu_f, & \text{если } h > 0, h' = h-1, l' = l, \\ \mu_s, & \text{если } l > 0, h' = h, l' = l. \end{cases} \quad (5)$$

Из соотношений (4) и (5) заключаем, что построенная конечная 2D MC является неприводимой, т.е. при любых положительных значениях исходных параметров системы в ней существует стационарный режим.

Пусть $p(h,l),(h,l) \in E,$ обозначают вероятности состояний этой 2D MC. Эти вероятности находятся в результате решения соответствующей системы уравнений равновесия (СУР), которая составляется на основе соотношений (4-5). Явный вид этой СУР приводится ниже.

Случаи $(h,l), l \le r_l$:

$$\begin{aligned} &(\lambda_h I(h < K_h) + \lambda_l + \mu_h I(h > 0) + \mu_s I(l > 0)) p(h,l) = \\ &= \lambda_h p(h-1,l) I(h > 0) + \lambda_l p(h,l-1) I(l > 0) + \\ &+ \mu_f p(h+1,l) I(h < K_h) + \mu_s p(h,l+1). \end{aligned} \quad (6)$$

Случаи (h,l), $l > r_l$:

$$\begin{aligned}&\left(\lambda_h I(h<K_h)+\lambda_l I(l<K_l)+\mu_f I(h>0)+\mu_l I(l>0)\right)p(h,l)=\\&=(\lambda_h+\lambda_l\alpha)p(h-1,l)I(h<r_h)+\lambda_h p(h-1,l)I(h\ge r_h)+\\&+\lambda_l p(h,l-1)I(h\ge r_h)+\lambda_l(1-\alpha)p(h,l-1)I(h<r_h)+\\&+\mu_f p(h+1,l)I(h<K_h)+\mu_s p(h,l+1).\end{aligned}\quad(7)$$

В уравнениях (6-7) $I(A)$ обозначает индикаторную функцию события A. К этим уравнениям следует добавить условие нормировки:

$$\sum_{(h,l)} p(h,l)=1. \quad (8)$$

СУР (6-8) представляет собой систему линейных алгебраических уравнений размерности $K_h \cdot K_l$. Поскольку изучаемая 2D MC является неприводимой, то при любых положительных значениях исходных параметров СУР (6-8) всегда имеет единственное решение. Для решения СУР (6-8) могут быть использованы известные численные методы линейной алгебры, которые реализованы в доступных пакетах прикладных программ.

Здесь же отметим, что в случаях, когда указанная СУР (6-8) имеет очень высокую размерность (т.е. когда величина $K_h \cdot K_l$ имеет большое значение) или пространство состояний модели (1) имеет бесконечный размер, можно использовать эффективный приближенный метод, разработанный в работе [7].

Формулы для расчета характеристик системы

В качестве основных характеристик изучаемой системы принимаются следующие величины: вероятности потери разнотипных заявок; средняя интенсивность скачков l-заявок в h-буфер; среднее число разнотипных заявок в системе; среднее время ожидания в очереди разнотипных заявок.

Высокоприоритетные заявки теряются лишь тогда, когда в моменты их поступления в системе уже имеются K_h заявок данного типа, т.е. вероятность потери h-заявок (PB_h) определяется так:

$$PB_h=\sum_{l=0}^{K_l} p(K_h,l). \quad (9)$$

Для вычисления вероятности потери l-заявок следует рассматривать состояния типа (h,K_l), при этом необходимо различать два случая: 1) $0\le h<r_h$; 2) $r_h\le h\le K_h$. Если в момент поступления l-заявки имеет место случай 1), то она теряется с вероятностью $1-\alpha$, а в случаях 2) эта заявка теряется с вероятностью единица. Таким образом, вероятность потери l-заявок (PB_l) определяется следующим образом:

$$PB_l=(1-\alpha)\sum_{h=0}^{r_h-1} p(h,K_l)+\sum_{h=r_h}^{K_h} p(h,K_l). \quad (10)$$

Скачки l-заявок в h-буфер происходят в моменты поступления l-заявок с вероятностью α, если в эти моменты система находится в состояние $(h,l)\in E_a$. Иными словами, средняя интенсивность скачков l-заявок в h-буфер (RJ_{lh}) вычисляется как

$$RJ_{lh}=\lambda_L\alpha\sum_{l=r_l+1}^{K_l}\sum_{h=0}^{r_h-1} p(h,l). \quad (11)$$

Среднее число h-заявок (N_h) и l-заявок (N_l) в системе определяются как математические ожидания соответствующих с.в., т.е.

$$N_h=\sum_{h=1}^{K_h} h\sum_{l=0}^{K_l} p(h,l); \quad (12)$$

$$N_l=\sum_{l=1}^{K_l} l\sum_{h=0}^{K_h} p(h,l). \quad (13)$$

Среднее время ожидания в очереди h-заявок (W_h) и l-заявок (W_l) вычисляются из модифицированной формулы Литтла:

$$W_x=N_x/\lambda_x(1-PB_x),\ x\in\{h,\ l\}. \quad (14)$$

Таким образом, полученные формулы (9-14) позволяют вычислить характеристик изучаемой системы.

Выводы

Предложена марковская модель системы обслуживания с гетерогенными серверами и раздельными очередями разнотипных заявок при наличии скачкообразных приоритетов. Введена новая схема

определения рандомизированных скачкообразных приоритетов, которые зависят от текущего состояния системы. Показано, что математической моделью изучаемой системы является двумерная цепь Маркова. Предложен алгоритм построения производящей матрицы этой цепи, разработана система уравнений равновесия для стационарных вероятностей состояний и получены формулы для вычисления характеристик системы.

Литература

1. *Efrosinin D.* Controlled Queuing Systems with Heterogeneous Servers. – Saarbrucken: VDM Verlag, 2008. – 236 p.

2. *Dharmaraja S., Kumar R.* Transient Solution of a Markovian Queuing Models with Heterogeneous Servers and Catastrophes // OPSEARCH. – 2015. – Vol. 52, Iss. 4. – P. 810-8217.

3. *Xu J., Liu L., Zhu T.* Transient Analysis of Two- Heterogeneous Server Queue with Impatient Behavior and Multiple Vacations // J. of Systems Science and Information. – 2018. – Vol. 6, Iss. 1. – P. 69-84.

4. *Maertens T., Walraevens J. Bruneel H.* On Priority Queues with Priority Jumps // Performance Evaluation. – 2006. – Vol. 63, Iss. 12. – P. 1235–1252.

5. *Maertens T., Walraevens J., Bruneel H.* A Modified HOL Priority Scheduling Discipline: Performance Analysis // Europ. J. Operations Research. – 2007. – Vol. 180, Iss. 3. – P. 1168–1185.

6. *Maertens T., Walraevens J., Bruneel H.* Performance Comparison of Several Priority Schemes with Priority Jumps // Annals of Operations Research. – 2008. – Vol. 162. – P. 109-125.

7. *Melikov A.Z., Mekhbaliyeva E.V.* Analysis and optimization of system with heterogeneous servers and jump priorities // J. of Computer and Systems Sciences International. – 2019. – Vol. 58, Iss. 5. – P. 718-735.

Мехбалыева Э.В.

СХЕМА ОПРЕДЕЛЕНИЯ СКАЧКООБРАЗНЫХ ПРИОРИТЕТОВ В СИСТЕМАХ ОБСЛУЖИВАНИЯ С ГЕТЕРОГЕННЫМИ СЕРВЕРАМИ

В данной работе предложена математическая модель системы обслуживания с гетерогенными серверами, заявками различных типов и скачкообразными приоритетами. Оба типа заявок формируют пуассоновские потоки и они ожидают в сепаратных буферах конечных размеров. Заявки высокого приоритета обслуживаются в сервере с высокой скоростью, в то время как заявки низкого приортета обслуживаются в сервера с низкой скоростью. Скачкообразные приоритеты определяют правила перехода заявки низкого приоритета в очередь заявок высокого приоритета. Если в момент поступления заявки высокого (низкого) приоритета имеется хотя бы одно свободное место в соответствующем буфере, то она присоединяется в очередь; иначе она получает отказ. Времена занятия каналов имеют показательное распределения с различными средними. Показано, что математической моделью системы является двумерная цепь Маркова с конечным пространством состояний. Разработан алгоритм для построения производящей матрицы изучаемой цепи и показано, что она являетс неприводимой. Поэтому в ней существует стационарный режим. Приведен явный вид системы балансовых уравнений. Найдены формулы для вычисления характеристик системы. Главными характеристиками являются вероятности потери разнотипных заявок, средняя длина очереди заявок каждого типа и срднее время их ожидания в очереди. Разработанные формулы позволяют проводить численные эксперименты для изучения поведения характеристик системы относительно изменения ее параметров, а также решить проблемы их оптимизации относительно выбранного критерия качества функционирования системы.

Ключевые слова: *система обслуживания, гетерогенные серверы, скачкообразные приоритеты, заявки различного типа, метод расчета.*

Mekhbaliyeva E.V.

SCHEME FOR DETERMINING THE JUMP PRIORITIES IN QUEUING SYSTEMS WITH HETEREGENEOUS SERVERS

In this paper, we propose the mathematical model of a queuing system with heteregenous servers, calls of different types and jump priorities. Both type of calls are formed Poisson flows and they are waits in finite separate buffers at heterogeneous servers. Calls of high priority are served by fast server while calls of low priority are servered in slow server. Jump priorities are defined rules for transfer of low priority calls to buffer of high priority calls. If upon arrival of a high priority call (low priority call) there is one free position in appropriate buffer, then it occupies it; otherwise, the call is lost. The distribution functions of channel occupation time by heterogeneous calls are exponential with different average values. It is shown that the mathematical model of the system is a certain two-dimensional Markov chain with a finite set of states. An algorithm is proposed for constructing the generating matrix of this chain and it is proved that this chain is irreducible and therefore there exists stationary probability distribution of the states of this Markov chain. An explicit form of the balance equations is obtained. Explicit formulas have been developed for calculating the characteristics of the system under study. Main characteristics are call loss probabilities of each type flow, average length of the both queue of calls different types and their average waiting times in queue. The developed formulas allow us to conduct numerical experiments in order to study the characteristics of the system relative to changes in its parameters, as well as solve the problems of their optimization with respect to the selected quality criterion for the functioning of the system.

Keywords: *queuing system, heteregenous servers, jump prioritiy, calls of different types, calculation method.*