МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ
MATHEMATICAL MODELING

УДК 519.872
https://doi.org/10.37661/1816-0301-2022-19-2-56-67

Оригинальная статья
Original Paper

Система массового обслуживания с разделением процессора, повторными вызовами и нетерпеливостью запросов

В. И. Клименок

Белорусский государственный университет,
пр. Независимости, 4, Минск, 220030, Беларусь
✉*E-mail: klimenok@bsu.by*

Аннотация

Цели. Рассматривается задача построения и исследования математической модели стохастической системы с разделением процессора, повторными вызовами и нетерпеливостью запросов. Данная система формализована в виде системы массового обслуживания, построен процесс функционирования системы, найдено условие существования стационарного распределения и предложены алгоритмы вычисления стационарного распределения и стационарных характеристик производительности системы.

Методы. Используются методы теории вероятностей, теории массового обслуживания и теории матриц.

Результаты. Функционирование системы описано в терминах многомерной цепи Маркова. Показано, что эта цепь имеет стационарное распределение, совпадающее с эргодическим, при любых приемлемых значениях параметров, описывающих входной поток, время обслуживания, процесс повторных вызовов и процесс ухода запросов из системы вследствие нетерпеливости.

Заключение. Исследован стационарный режим функционирования системы массового обслуживания с повторными вызовами, разделением процессора и двумя типами запросов, поступающих в систему в соответствии с маркированным марковским потоком. Пропускная способность канала делится между запросами двух типов в некоторой пропорции, а число запросов каждого из типов, одновременно находящихся на приборе, ограничено. Запросы одного из типов, заставшие все отведенные для них каналы занятыми, с некоторой вероятностью уходят из системы необслуженными и с дополнительной вероятностью идут на орбиту бесконечного объема, откуда делают попытки попасть на обслуживание через случайные промежутки времени. Запросы второго типа, заставшие все отведенные для них каналы занятыми, теряются. Запросы, находящиеся на орбите, проявляют нетерпеливость: каждый из них может покинуть орбиту навсегда по истечении экспоненциально распределенного времени при условии, что он не попадет на обслуживание за это время. Времена обслуживания запросов распределены по фазовому закону с разными параметрами. Функционирование системы описано в терминах многомерной цепи Маркова. Доказано, что при любых значениях параметров системы эта цепь имеет стационарное распределение. Предложены алгоритмы вычисления стационарного распределения и ряда характеристик производительности системы. Результаты исследования могут быть использованы для моделирования работы соты фиксированной емкости в беспроводной сотовой сети связи и других реальных систем, функционирующих в режиме разделения процессора.

Ключевые слова: система массового обслуживания, неоднородный входной поток, повторные вызовы, ограниченное разделение процессора, стационарное распределение, характеристики производительности

Для цитирования. Клименок, В. И. Система массового обслуживания с разделением процессора, повторными вызовами и нетерпеливостью запросов / В. И. Клименок // Информатика. – 2022. − Т. 19, № 2. – С. 56–67. https://doi.org/10.37661/1816-0301-2022-19-2-56-67

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию | Received 01.03.2022
Подписана в печать | Accepted 18.04.2022
Опубликована | Published 29.06.2022

A retrial queueing system with processor sharing and impatient customers

Valentina I. Klimenok

Belorussian State University,
av. Nezavisimosti, 4, Minsk, 220030, Belarus
✉*E-mail: klimenok@bsu.by*

Abstract
Objectives. The problem of constructing and investigating a mathematical model of a stochastic system with processor sharing, repeated calls, and customer impatience is considered. This system is formalized in the form of a queueing system. The operation of the queue is described in terms of multi-dimensional Markov chain. A condition for the existence of a stationary distribution is found, and algorithms for calculating the stationary distribution and stationary performance characteristics of the system are proposed.
Methods. Methods of probability theory, queueing theory and matrix theory are used.
Results. The steady state operation of a queueing system with repeated calls, processor sharing and two types of customers arriving in a marked Markovian arrival process is studied. The channel bandwidth is divided between two types of customers in a certain proportion, and the number of customers of each type simultaneously located on the server is limited. Customers of one of the types that have made all the channels assigned to them busy leave the system unserved with some probability and, with an additional probability, go to the orbit of infinite size, from where they make attempts to get service at random time intervals. Customers of the second type, which caused all the channels assigned to them to be busy, are lost. Customers in orbit show impatience: each of them can leave orbit forever if the time of its stay in orbit exceeds some random time distributed according to an exponential law. Service times of customers of different types are distributed according to the phase law with different parameters. The operation of the system is described in terms of a multi-dimensional Markov chain. It is proved that for any values of the system parameters this chain has a stationary distribution. Algorithms for calculating the stationary distribution and a number of performance measures of the system are proposed. The results of the study can be used to simulate the operation of a fixed capacity cell in a wireless cellular communication network and other real systems operating in the processor sharing mode.

Keywords: queueing system, heterogeneous input, repeated calls, limited processor sharing, stationary distribution, performance measures

For citation. Klimenok V. I. *A retrial queueing system with processor sharing and impatient customers.* Informatika [*Informatics*], 2022, vol. 19, no. 2, pp. 56‒67 (In Russ.). https://doi.org/10.37661/1816-0301-2022-19-2-56-67

Conflict of interest. The author declare of no conflict of interest.

Введение. Одним из важных разделов математического моделирования является теория массового обслуживания, представляющая собой теоретические основы эффективного конструирования и эксплуатации систем массового обслуживания. Система массового обслуживания характеризуется своими структурой и дисциплиной обслуживания. С развитием телекоммуникационных и компьютерных сетей все большее внимание уделяется дисциплине разделения процессора, при которой все или некоторые пользователи могут получать обслуживание одновременно. В частности, разделение процессора используется при планировании задач в многопрограммных компьютерных системах и в сетях мобильной сотовой связи, при кешировании популярного мультимедийного контента.

Существует множество научных работ, в которых исследователи в области телекоммуникаций и теории массового обслуживания изучают системы и сети с дисциплиной разделения процессора (см., например, статьи [1–5] и ссылки в них). В большинстве работ предполагается, что входной поток в систему является стационарным пуассоновским, а времена обслуживания запросов распределены по экспоненциальному закону. Такие предположения снижают ценность результатов для приложений, так как в них не учитывается реальный характер трафика в современных сетях. В частности, не учитывается коррелированность потоков информации и то, что времена обслуживания не имеют свойства отсутствия последействия, характерного для экспоненциального распределения. Упомянутые ограничения сняты в работах [6, 7], где рассматриваются марковский поток и фазовое распределение времени обслуживания. Вместе с тем в этих работах не принимается во внимание тот факт, что реальные потоки в телекоммуникационных сетях, как правило, являются не только коррелированными, но и неоднородными. В таких потоках запросы разных типов могут делить пропускную способность канала в разных пропорциях и иметь разные распределения времени обслуживания. Неоднородность потока информации учтена в статье [8], где рассмотрена система массового обслуживания с маркированным марковским потоком *MMAP* (общепризнанная в мировой литературе аббревиатура от англ. Marked Markovian Arrival Process), ограниченным разделением процессора и повторными вызовами. Явление повторных вызовов – неотъемлемая черта многих современных телекоммуникационных и компьютерных сетей. Повторные вызовы характеризуются тем, что пользователь, получив отказ в обслуживании, не уходит из системы, а становится в режим ожидания (поступает на так называемую орбиту) и через случайные моменты времени повторно пытается получить обслуживание. Системы массового обслуживания с повторными вызовами менее изучены, чем системы с ожиданием и потерями, что объясняется сложностями при их исследовании, вызванными пространственной неоднородностью случайных процессов, описывающих функционирование таких систем.

В настоящей статье рассматривается система массового обслуживания с повторными вызовами, *MMAP*-потоком запросов двух типов и ограниченным разделением процессора. Главной отличительной чертой данной системы по сравнению с системой [8] является нетерпеливость запросов, находящихся на орбите. Каждый из таких запросов может покинуть систему навсегда в течение экспоненциально распределенного времени при условии, что он не попадет на обслуживание за это время. Рассматриваемая система может служить для математического моделирования соты мобильной сети связи с двумя типами пользователей: новыми (new), которые находятся в зоне видимости соты постоянно, и хэндовер (handover), которые пересекают территорию соты транзитом. Процесс функционирования системы описан многомерной цепью Маркова, найдено ее условие эргодичности и вычислено стационарное распределение вероятностей состояний. На основе стационарного распределения получены формулы для ряда характеристик производительности системы.

Описание системы. Представляется система массового обслуживания с разделением процессора между двумя типами запросов, которые поступают в систему в *MMAP*-потоке. В общем случае *MMAP* является коррелированным потоком $K(0 < K < \infty)$ типов запросов. Запросы поступают под управлением неприводимой цепи Маркова с непрерывным временем ν_t, $t \geq 0$, которая принимает значения в множестве $\{0, 1, 2, \dots, W\}$ и называется управляющим процессом *MMAP*. Процесс ν_t пребывает в состоянии ν в течение экспоненциально распределенного времени с параметром λ_ν, $\nu = \overline{0,W}$, после чего с вероятностью $p_k(\nu,\nu')$ переходит в состояние ν' и генерируется запрос k-го типа, $k \in \{1, 2, \dots, K\}$, или с вероятностью $p_0(\nu,\nu')$ цепь переходит в состояние ν' без генерации запроса, причем $p_0(\nu,\nu) = 0$. Для указанных вероятностей выполняются естественные ограничения $\sum_{k=1}^{K} \sum_{\nu'=0}^{W} p_k(\nu,\nu') = 1$,

$\nu,\nu' = \overline{0,W}$. Всю информацию о $MMAP$ удобно хранить в виде набора матриц D_k, $k = \overline{1,K}$, порядка $(W+1) \times (W+1)$, элементы которых определяются следующим образом:

$$(D_k)_{\nu,\nu'} = \lambda_\nu p_k(\nu,\nu'),\ \nu,\nu' = \overline{0,W}, k = \overline{1,K};$$

$$(D_0)_{\nu,\nu'} = \begin{cases} \lambda_\nu p_0(\nu,\nu'),\ \nu \neq \nu',\nu,\nu' = \overline{0,W}; \\ -\lambda_\nu,\ \nu = \nu' = \overline{0,W}. \end{cases}$$

Из формул видно, что элементами матриц D_k, $k = \overline{1,K}$, являются интенсивности переходов процесса ν_t, сопровождающиеся генерацией запроса k-го типа. Аналогичный смысл имеют недиагональные элементы матрицы D_0, а диагональные элементы этой матрицы есть взятые с противоположным знаком интенсивности выхода процесса ν_t из соответствующих состояний. Матрицы D_k, $k = \overline{1,K}$, можно задавать их матричной производящей функцией $D(z) = \sum_{k=0}^{K} D_k z^k$, $|z| < 1$. Значение этой функции в точке $z = 1$ является инфинитезимальным генератором управляющего процесса ν_t, $t \geq 0$. Стационарное распределение данного процесса, представленное в виде вектора-строки $\boldsymbol{\theta}$, определяется как решение системы линейных алгебраических уравнений: $\boldsymbol{\theta} D(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$. Здесь и далее $\mathbf{0}$ – вектор-строка, состоящая из нулей, а $\mathbf{e}$ – вектор-столбец, состоящий из единиц. Интенсивность λ_k поступления запросов k-го типа в $MMAP$-потоке задается формулой $\lambda_k = \boldsymbol{\theta} D_k \mathbf{e}$. Более подробное описание $MMAP$, включающее формулы для дисперсии длин интервалов между моментами поступления запросов k-го типа и коэффициентов корреляции длин двух соседних интервалов между моментами поступления запросов каждого из типов, можно найти, например, в публикациях [9, 10].

В настоящей работе, как уже было сказано ранее, сделано предположение, что в систему поступают в $MMAP$-потоке запросы двух типов, т. е. $K = 2$. На обслуживающем приборе одновременно могут обслуживаться до N запросов первого типа и до R запросов второго типа. Если на приборе обслуживается только один запрос k-го типа, $k = 1, 2$, то время его обслуживания имеет фазовое распределение PH (общепризнанная аббревиатура от англ. phase type distribution), заданное неприводимым представлением $(\boldsymbol{\beta}_k,S_k)$ и управляющим процессом $m_t^{(k)}$, $t \geq 0$, с пространством состояний $\{1,\ldots,M_k, M_k + 1\}$, где состояние $M_k + 1$ является поглощающим. Первоначальное состояние (фаза) обслуживания выбирается в соответствии со стохастическим вектором-строкой $\boldsymbol{\beta}_k$. Интенсивности переходов в поглощающее состояние определяются вектором-столбцом $\mathbf{S}_0^{(k)} = -S_k\mathbf{e}$. Более детальное описание PH-распределения можно найти в работах [10, 11].

Запросы каждого из типов делят отведенную им пропускную способность обслуживающего прибора поровну. Если на приборе одновременно обслуживается $n^{(k)}$ запросов k-го типа, то время обслуживания любого из этих запросов имеет PH-распределение, заданное представлением $(\boldsymbol{\beta}^{(k)}, \frac{1}{n^{(k)}} S^{(k)})$.

Если поступающий запрос первого типа видит, что в системе уже имеется N запросов его типа, то с вероятностью q поступающий запрос уходит из системы навсегда (теряется) и с дополнительной вероятностью $\bar{q} = 1 - q$ уходит на орбиту бесконечного объема, откуда делает попытки попасть на обслуживание через случайные моменты времени, распределенные по экспоненциальному закону с параметром γ. В аналогичной ситуации, если поступающий запрос второго типа видит, что в системе уже имеется R запросов его типа, то он теряется. Запросы, находящиеся на орбите, проявляют нетерпеливость: через экспоненциально распределенное с параметром α время каждый из таких запросов покидает систему навсегда.

Цепь Маркова, описывающая процесс функционирования системы. Функционирование системы описывается регулярной неприводимой цепью Маркова с непрерывным временем

$$\xi_t = \{i_t, n_t, \eta_t^{(1)}, \eta_t^{(2)}, \ldots, \eta_t^{(M_1)}, r_t, \tau_t^{(1)}, \tau_t^{(2)}, \ldots, \tau_t^{(M_2)}, \nu_t\},$$

где i_t – число запросов первого типа на орбите, $i \geq 0$;

n_t – число запросов первого типа на приборе, $n_t = \overline{0,N}$;

$\eta_t^{(m^{(1)})}$– число запросов первого типа, которые обслуживаются на фазе $m^{(1)}$, $\eta_t^{(m^{(1)})} = \overline{0,n_t}$, $m^{(1)} = \overline{1,M_1}$;

r_t – число запросов второго типа на приборе, $r_t = \overline{0,R}$;

$\tau_t^{(m^{(2)})}$– число запросов второго типа, которые обслуживаются на фазе $m^{(2)}$, $\tau_t^{(m^{(2)})} = \overline{0,r_t}$, $m^{(2)} = \overline{1,M_2}$;

ν_t – состояние управляющего процесса $MMAP$, $\nu_t = \overline{0,W}$, в момент времени t.

Введем некоторые обозначения:

$R = \sum_{r=0}^{R} C_{r+M_2-1}^{M_2-1}$, $\bar{W} = W + 1$;

$\otimes$ ($\oplus$) – символ кронекерова произведения (суммы) матриц (см., например, [12]);

$diag\{a_1, a_2, ..., a_n\}$ – блочная диагональная матрица, у которой диагональные блоки равны элементам, перечисленным в скобках, а остальные блоки нулевые;

$diag^{+}\{a_1, a_2, ..., a_n\}$ – квадратная блочная матрица, у которой наддиагональные блоки равны элементам, перечисленным в скобках, а остальные блоки нулевые;

$diag^{-}\{a_1, a_2, ..., a_n\}$ – квадратная блочная матрица, у которой поддиагональные блоки равны элементам, перечисленным в скобках, а остальные блоки нулевые;

$\mathbf{u}_t^{(1)} = \{\eta_t^{(1)}, \eta_t^{(2)}, \ldots, \eta_t^{(M_1)}\}$, $\mathbf{u}_t^{(2)} = \{\tau_t^{(1)}, \tau_t^{(2)}, \ldots, \tau_t^{(M_2)}\}$;

C_n^m – биномиальный коэффициент.

Упорядочим состояния рассматриваемой цепи Маркова $\xi_t, t \geq 0$, следующим образом. Перенумеруем компоненты i_t, n_t, r_t в прямом лексикографическом порядке и при фиксированных значениях этих компонент перенумеруем состояния процессов $\mathbf{u}_t^{(1)}$ и $\mathbf{u}_t^{(2)}$ в обратном лексикографическом порядке. Упорядочение в обратном лексикографическом порядке требуется для дальнейшего описания интенсивностей переходов процессов $\mathbf{u}_t^{(1)}$ и $\mathbf{u}_t^{(2)}$ с использованием введенных в статьях [13, 14] матриц $P_i(\cdot)$, $A_i(\cdot,\cdot)$, $L_i(\cdot,\cdot)$.

Введем в рассмотрение матрицы $\tilde{S}_l = \begin{pmatrix} 0 & O \\ \mathbf{S}_0^{(l)} & S_l \end{pmatrix}$, $l = 1, 2$.

Дадим краткое объяснение вероятностных значений матриц $P_i(\cdot)$, $A_i(\cdot,\cdot)$, $L_i(\cdot,\cdot)$:

$L_k(n, \tilde{S}_l)$ – матрица порядка $C_{n-k+M_l-1}^{M_l-1} \times C_{n-k+M_l-2}^{M_l-1}$, которая содержит интенсивности переходов процесса $\mathbf{u}_t^{(l)}$, приводящих к завершению обслуживания одного из $n-k$ запросов l-го типа (k есть число свободных каналов для запросов l-го типа, n – общее число каналов, отведенных для запросов l-го типа);

$P_n(\boldsymbol{\beta}_l)$ – матрица порядка $C_{n+M_l-1}^{M_l-1} \times C_{n+M_l}^{M_l-1}$, которая содержит вероятности переходов процесса $\mathbf{u}_t^{(l)}$, приводящих к увеличению числа запросов l-го типа, находящихся на приборе, с n до $n+1$;

$A_n(k, S_l)$ – матрица порядка $C_{n+M_l-1}^{M_l-1} \times C_{n+M_l}^{M_l-1}$, которая содержит интенсивности переходов процесса $\mathbf{u}_t^{(l)}$ в его пространстве состояний без увеличения или уменьшения числа запросов l-го типа, находящихся на обслуживании (n есть число запросов l-го типа, находящихся на обслуживании; k – общее число каналов, отведенных для запросов l-го типа).

В дальнейшем полагаем $L_0(0) = A_0(\cdot) = P_{-1}(\cdot) = 0$.

Алгоритм вычисления матриц $P_i(\cdot)$, $A_i(\cdot,\cdot)$, $L_i(\cdot,\cdot)$ следует из результатов В. Рамасвами и Д. Лукантони, опубликованных в статьях [13, 14]. Шаги этого алгоритма описаны в работе [15] и приложении настоящей статьи.

Введем обозначение $Q_{n,n'}$ для интенсивностей переходов цепи из состояний, соответствующих значению n первой компоненты, в состояния, соответствующие значению n' этой компоненты, $n,n' = \overline{0,N}$. Тогда инфинитезимальный генератор цепи определяется следующей леммой.

Лемма. Инфинитезимальный генератор цепи Маркова ξ_t, $t \geq 0$, имеет блочную структуру

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & O & \cdots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & O & \cdots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & O & \cdots \\ O & O & Q_{3,2} & Q_{3,3} & Q_{3,4} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

где
$$Q_{i,i-1} = i\gamma\, diag^{+}\{P_n(\boldsymbol{\beta}_1), n = \overline{0,N-1}\} \otimes I_R \otimes I_{\overline{W}} + i\alpha I, i \geq 1;$$

$$Q_{i,i+1} = diag\{O_{C_{n+M_1-1}^{M_1-1}}, n = \overline{0,N-1}, I_{C_{N+M_1-1}^{M_1-1}}\} \otimes I_R \otimes \bar{q} D_1, i \geq 0;$$

$$(Q_{i,i})_{n,n-1} = \frac{1}{n} L_{N-n}(N,\tilde{S}_1) \otimes I_R \otimes I_{\overline{W}}, n = \overline{1,N}, i \geq 0;$$

$$(Q_{i,i})_{n,n+1} = P_n(\boldsymbol{\beta}_1) \otimes I_R \otimes D_1, n = \overline{0,N-1}, i \geq 0;$$

$$(Q_{i,i})_{n,n} = \Psi_n - i(\gamma+\alpha) I_{C_{n+M_1-1}^{M_1-1} R\overline{W}} + \Delta_{i,n}, n = \overline{0,N-1}, i \geq 0;$$

$$(Q_{i,i})_{N,N} = \Psi_N + qD_1 - i\alpha I_{C_{n+M_1-1}^{M_1-1} R\overline{W}} + \Delta_N, i \geq 0;$$

$$\Psi_n = I_{C_{n+M_1-1}^{M_1-1}} \otimes diag^{-}\{\tfrac{1}{r} L_{R-r}(R,\tilde{S}_2), r = \overline{1,R}\} \otimes I_{\overline{W}} +$$

$$+\frac{1}{n} A_n(N,S_1) \oplus diag\{0, \frac{1}{r} A_r(R,S_2);$$

$$r = \overline{1,R}\} \oplus D_0 + I_{C_{n+M_1-1}^{M_1-1}} \otimes diag\{O_{\sum_{r=0}^{R-1} C_{r+M_2-1}^{M_2-1}}, I_{C_{R+M_2-1}^{M_2-1}}\} \otimes D_2 +$$

$$+I_{C_{n+M_1-1}^{M_1-1}} \otimes diag^{+}\{P_r(\boldsymbol{\beta}_2), r = \overline{0,R-1}\} \otimes D_2.$$

Здесь $\Delta_{i,n}$, $i \geq 0$, $n = \overline{0,N-1}$, и Δ_N – диагональные матрицы, которые обеспечивают выполнение равенства $Q\mathbf{e} = \mathbf{0}$.

Доказательство леммы проводится путем анализа поведения цепи Маркова ξ_t, $t \geq 0$, на бесконечно малом интервале времени. Опишем кратко смысл ненулевых блоков генератора.

Блок $Q_{i,i-1}$, $i \geq 1$, состоит из интенсивностей переходов рассматриваемой цепи Маркова, сопровождающихся «удачной» повторной попыткой с орбиты запроса первого типа либо потерей запроса с орбиты из-за нетерпеливости. В любом из этих случаев число запросов на орбите уменьшается с i до $i-1$.

Блок $Q_{i,i+1}$, $i \geq 0$, состоит из интенсивностей переходов, сопровождающихся поступлением первичного запроса первого типа, который находит все каналы, отведенные для данного типа запросов, занятыми и уходит на орбиту. При этом число запросов на орбите увеличивается с i до $i+1$.

Блок $(Q_{i,i})_{n,n-1}$, $n = \overline{1,N}$, $i \geq 0$, состоит из интенсивностей переходов, сопровождающихся окончанием обслуживания одного из n запросов первого типа, находящихся на приборе. При этом число запросов первого типа на приборе уменьшается с n до $n-1$.

Блок $(Q_{i,i})_{n,n+1}, n = \overline{0,N-1}, i \geq 0$, состоит из интенсивностей переходов, сопровождающихся поступлением первичного запроса первого типа, заставшего свободные каналы, отведенные для запросов данного типа. При этом число запросов первого типа на приборе увеличивается с n до $n+1$.

Недиагональные элементы блока $(Q_{i,i})_{n,n}, n = \overline{0,N}, i \geq 0$, есть интенсивности переходов, не вызывающих ни изменения числа запросов на орбите, ни изменения числа запросов первого типа на приборе. Соответствующие переходы могут быть вызваны либо окончанием обслуживания одного из запросов второго типа (матрица $I_{C_{n+M_1-1}^{M_1-1}} \otimes diag^{-}\{\frac{1}{r}L_{R-r}(R,\tilde{S}_2), r = \overline{1,R}\} \otimes I_{\overline{W}}$), либо перераспределением числа приборов, обслуживающих запросы на разных фазах, и холостыми переходами управляющего процесса *MMAP* (матрица $\frac{1}{n}A_n(N,S_1) \oplus \oplus\, diag\{0, \frac{1}{r}A_r(R,S_2), r = \overline{1,R}\} \oplus D_0$), либо поступлением запроса второго типа, заставшего все места занятыми (матрица $I_{C_{n+M_1-1}^{M_1-1}} \otimes diag\{O_{\sum_{r=0}^{R-1} C_{r+M_2-1}^{M_2-1}}, I_{C_{R+M_2-1}^{M_2-1}}\} \otimes D_2$, либо поступлением запроса второго типа, который застает свободные места и поступает на обслуживание (матрица $I_{C_{n+M_1-1}^{M_1-1}} \otimes diag^{+}\{P_r(\boldsymbol{\beta}), r = \overline{0,R-1}\} \otimes D_2$). Диагональные элементы рассматриваемого блока есть взятые с противоположным знаком интенсивности выхода цепи из состояний, соответствующих i запросам на орбите и n запросам первого типа на приборе.

Следствие. *Цепь Маркова* $\xi_t, t \geq 0$, *принадлежит классу многомерных асимптотически квазитеплицевых цепей Маркова (АКТЦМ)* [16] *с непрерывным временем.*

Доказательство. Пусть T_i – диагональная матрица, диагональными элементами которой являются модули диагональных элементов матрицы $Q_{i,i}, i \geq 0$. Согласно [16], где приведено определение АКТЦМ, следствием будет доказано, что существуют пределы

$$Y_0 = \lim_{i\to\infty} T_i^{-1} Q_{i,i-1},\ Y_1 = \lim_{i\to\infty} T_i^{-1} Q_{i,i} + I,\ Y_2 = \lim_{i\to\infty} T_i^{-1} Q_{i,i+1} \quad (1)$$

и матрица $Y_0 + Y_1 + Y_2$ является стохастической.

Простые вычисления приводят к следующим выражениям для матриц Y_k:

$$Y_0 = diag^{+}\{P_n(\boldsymbol{\beta}_1), n = \overline{0,N-1}\} \otimes I_{R\overline{W}} + diag\{I_{C_{n+M_1-1}^{M_1-1}R\overline{W}}, n = \overline{0,N-1}, I_{C_{N+M_1-1}^{M_1-1}R\overline{W}}\},$$

$$Y_1 = 0,\ Y_2 = 0.$$

Таким образом, пределы (1) существуют. Сумма $Y_0 + Y_1 + Y_2$ есть стохастическая матрица, поскольку она равна Y_0 и выполняется равенство $Y_0\mathbf{e} = \mathbf{e}$. Из этого следует, что цепь Маркова ξ_t принадлежит классу АКТЦМ.

Стационарное распределение. Интуитивно понятно, что стационарное (совпадающее с эргодическим) распределение рассматриваемой цепи Маркова существует при любых допустимых значениях параметров вследствие того, что запросы могут уходить с орбиты из-за нетерпеливости. Строгое доказательство этого факта следует из результатов [16] для АКТЦМ и приведено в следующей теореме.

Теорема. *Цепь Маркова* $\xi_t, t \geq 0$, *является эргодической при любых значениях параметров.*

Доказательство. Как следует из работы [16], АКТЦМ $\xi_t, t \geq 0$, является эргодической, если выполняется неравенство

$$\mathbf{x}Y_2 < \mathbf{x}Y_0, \quad (2)$$

где вектор $\mathbf{x}$ – единственное решение системы линейных алгебраических уравнений

$$\mathbf{x}(Y_0 + Y_1 + Y_2) = \mathbf{x}, \quad (3)$$

$$\mathbf{x}\mathbf{e} = 1. \quad (4)$$

Поскольку выполняется равенство $Y_0 + Y_1 + Y_2 = Y_0$ и матрица Y_0 является неприводимой стохастической, решение системы (3) определяется с точностью до константы, которая единственным образом находится из уравнения (4). В итоге получаем стохастический вектор $\boldsymbol{x}$. Подставляя этот вектор в неравенство (2) и учитывая, что $Y_2 = 0$, убеждаемся, что данное неравенство выполняется всегда. Из последнего факта следует, что рассматриваемая цепь Маркова эргодична при любых значениях параметров.

Далее будем обозначать через $\mathbf{p}_i$ вектор-строку стационарных вероятностей, соответствующий значению i первой компоненты цепи ξ_t, $i \geq 0$. Для вычисления векторов $\mathbf{p}_i$, $i \geq 0$, используется численно устойчивый алгоритм, который был разработан в работе [16] для многомерных квазитеплицевых цепей Маркова общего вида и адаптирован на случай рассматриваемой системы.

Алгоритм [16] *состоит из следующих шагов:*

1. Находим матрицы G_i, $i \geq 0$, из уравнения обратной рекурсии:

$$G_i = \left(-Q_{i+1,i+1} - Q_{i+1,i+2}G_{i+1}\right)^{-1} Q_{i+1,i}.$$

При реализации шага 1 используем факт существования предела $\lim\limits_{i\to\infty} G_i = Y_0$, чтобы найти начальное условие для уравнения обратной рекурсии. Для этого выбираем некоторое число i_0, полагаем $G_{i_0+1} = Y_0$, вычисляем по уравнению G_{i_0} и проверяем условие $\| G_{i_0} - Y_0 \| < \epsilon$, где ϵ – заранее заданное малое число. Если условие выполняется, то полагаем все матрицы G_i для $i \geq i_0$ равными Y_0. Остальные матрицы G_i находим из уравнения обратной рекурсии. Если для этого i_0 условие не выполняется, то выбираем новое (большее) значение i_0.

2. Вычисляем матрицы $\bar{Q}_{i,i}, \bar{Q}_{i,i+1}$ по формулам

$$\bar{Q}_{i,i} = Q_{i,i} + Q_{i,i+1}G_i, \ i \geq 0, \quad \bar{Q}_{i,i+1} = Q_{i,i+1}, \ i \geq 0.$$

3. Находим матрицы F_i из рекуррентных соотношений

$$F_0 = I, F_i = F_{i-1}\bar{Q}_{i-1,i}\left(-\bar{Q}_{i,i}\right)^{-1}, i \geq 1.$$

4. Вычисляем вектор $\mathbf{p}_0$ как единственное решение системы линейных алгебраических уравнений

$$\mathbf{p}_0(-\bar{Q}_{0,0}) = \mathbf{0}, \ \mathbf{p}_0 \sum_{i=0}^{\infty} F_i\boldsymbol{e} = 1.$$

5. Вычисляем векторы $\mathbf{p}_i$ по формулам $\mathbf{p}_i = \mathbf{p}_0 F_i$, $i \geq 0$.

Характеристики производительности. Вычислив стационарное распределение $\mathbf{p}_i$, $i \geq 0$, можно найти ряд вероятностных характеристик производительности системы. Приведем важнейшие из них:

• Распределение числа запросов первого типа, находящихся на орбите, $p_i = \mathbf{p}_i\mathbf{e}$, $i \geq 0$.

• Среднее число запросов первого типа, находящихся на орбите, $Z_{orbit} = \sum_{i=1}^{\infty} ip_i$.

• Совместное распределение числа запросов первого типа, находящихся на приборе, и состояний *MMAP*-потока:

$$\boldsymbol{\pi}_n^* = \sum_{i=0}^{\infty} \mathbf{p}_i \left[\begin{pmatrix} \mathbf{0}^T_{\sum_{m=0}^{n-1} C_{m+M_1-1}^{M_1-1} R} \\ \mathbf{e}_{C_{n+M_1-1}^{M_1-1} R} \\ \mathbf{0}^T_{\sum_{m=n+1}^{N} C_{m+M_1-1}^{M_1-1} R} \end{pmatrix} \otimes I_{\overline{W}} \right], n = \overline{0,N}.$$

• Распределение числа запросов первого типа, находящихся на приборе, $\pi_n = \boldsymbol{\pi}_n^* \mathbf{e}_{\overline{W}}$, $n = \overline{0,N}$.

• Среднее число запросов первого типа, находящихся на приборе, $\bar{N} = \sum_{n=1}^{N} n\pi_n$.

• Вероятность того, что запрос первого типа пойдет на обслуживание, не посещая орбиту,

$$P_{imm} = 1 - \frac{1}{\lambda_1}\boldsymbol{\pi}_N^* \bar{q} D_1 \mathbf{e}.$$

Пояснение. Здесь вычитаемое представляет собой отношение интенсивности запросов первого типа, заставших в системе N запросов и ушедших на орбиту, к интенсивности поступления всех запросов первого типа. Другими словами, вычитаемое есть вероятность того, что любой поступивший запрос первого типа застанет в системе N запросов первого типа и уйдет на орбиту. Дополнительная вероятность дает искомую вероятность P_{imm}.

• Вероятность потери запроса первого типа из-за недостатка свободных каналов

$$P_{loss,1} = \frac{1}{\lambda_1}\boldsymbol{\pi}_N^* q D_1 \mathbf{e}.$$

• Вероятность потери запроса первого типа из-за нетерпеливости

$$P_{loss,1}^{imp} = \frac{\alpha}{\lambda_1} Z_{orbit}.$$

• Совместное распределение числа запросов второго типа, находящихся на приборе, и состояний $MMAP$-потока

$$\mathrm{q}_r^* = \sum_{i=0}^{\infty} \mathbf{p}_i \sum_{n=0}^{N} \left[\begin{pmatrix} \mathbf{0}^T_{\sum_{m=0}^{n-1} C_{m+M_1-1}^{M_1-1}} \\ \mathbf{e}_{C_{n+M_1-1}^{M_1-1}} \\ \mathbf{0}^T_{\sum_{m=n+1}^{N} C_{m+M_1-1}^{M_1-1}} \end{pmatrix} \otimes \begin{pmatrix} \mathbf{0}^T_{\sum_{m=0}^{r-1} C_{m+M_2-1}^{M_2-1}} \\ \mathbf{e}_{C_{r+M_2-1}^{M_2-1}} \\ \mathbf{0}^T_{\sum_{m=r+1}^{R} C_{m+M_2-1}^{M_2-1}} \end{pmatrix} \otimes I_{\bar{W}} \right], r = \overline{0,R}.$$

• Распределение числа запросов второго типа на приборе $q_r = \mathrm{q}_r^* \mathbf{e}$, $r = \overline{0,R}$.

• Среднее число запросов второго типа на приборе $\bar{R} = \sum_{r=1}^{R} r q_r$.

• Вероятность потери запроса второго типа $P_{loss,2} = \frac{1}{\lambda_2}\mathrm{q}_R^* D_2 \mathbf{e}$.

Заключение. В статье исследована система массового обслуживания с повторными вызовами, разделением процессора и двумя типами запросов, поступающих в систему в соответствии с $MMAP$. Времена обслуживания запросов разных типов имеют PH-распределения с разными параметрами. Запросы, находящиеся на орбите, проявляют нетерпеливость, вследствие чего могут уходить из системы необслуженными. Функционирование системы описано в терминах многомерной цепи Маркова. Доказано, что при любых допустимых значениях параметров системы эта цепь имеет стационарное распределение. Предложен алгоритм вычисления стационарного распределения и формулы для ряда характеристик производительности системы. Результаты исследования могут быть использованы для моделирования работы соты фиксированной емкости в беспроводной сотовой сети связи и других реальных систем, функционирующих в режиме разделения процессора.

References

1. Ghosh A., Banik A. D. An algorithmic analysis of the *BMAP*/*MSP*/1 generalized processor-sharing queue. *Computers and Operations Research*, 2017, vol. 79, pp. 1–11.

2. Telek M., van Houdt B. Response time distribution of a class of limited processor sharing queues. *Performance Evaluation Review*, 2018, vol. 45, no. 3, pp. 143–155. https://doi.org/10.1145/3199524.3199548

3. Yashkov S., Yashkova A. Processor sharing: a survey of the mathematical theory. *Automation and Remote Control*, 2007, vol. 68, pp. 662–731.

4. Zhen Q., Knessl C. On sojourn times in the finite capacity *M*/*M*/1 queue with processor sharing. *Operations Research Letters*, 2009, vol. 37, pp. 447–450.

5. Masuyama H., Takine T. Sojourn time distribution in a *MAP*/*M*/1 processor-sharing queue. *Operations Research Letters*, 2003, vol. 31, pp. 406–412.

6. Dudin S., Dudin A., Dudina O., Samouylov K. Analysis of a retrial queue with limited processor sharing operating in the random environment. *Lecture Notes in Computer Science*, 2017, vol. 10372, pp. 38–49.

7. Dudin A., Dudin S., Dudina O., Samouylov K. Analysis of queuing model with limited processor sharing discipline and customers impatience. *Operations Research Perspectives*, 2018, vol. 5, pp. 245–255.

8. Klimenok V., Dudin A. A retrial queueing system with processor sharing. *Communications in Computer and Information Science*, 2021, vol. 1391, pp. 46–60.

9. He Q. M. Queues with marked customers. *Advances in Applied Probability*, 1996, vol. 28, pp. 567–587.

10. Dudin A. N., Klimenok V. I., Vishnevsky V. M. *The Theory of Queuing Systems with Correlated Flows*. Springer, 2020, 410 p.

11. Neuts M. F. *Matrix-Geometric Solutions in Stochastic Models*. Baltimore, the Johns Hopkins University Press, 1981, 352 p.

12. Graham A. *Kronecker Products and Matrix Calculus with Applications*. Cichester, Ellis Horwood, 1981, 130 p.

13. Ramaswami V. Independent Markov processes in parallel. *Communications in Statistics. Stochastic Models*, 1985, vol. 1, pp. 419–432.

14. Ramaswami V., Lucantoni D. M. Algorithms for the multi-server queue with phase-type service. *Communications in Statistics. Stochastic Models*, 1985, vol. 1, pp. 393–417.

15. Dudina O., Kim C. S., Dudin S. Retrial queueing system with Markovian arrival flow and phase type service time distribution. *Computers and Industrial Engineering*, 2013, vol. 66, pp. 360–373.

16. Klimenok, V. I., Dudin A. N. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems*, 2006, vol. 54, pp. 245–259.

Приложение [15]

Алгоритм вычисления матриц $A_i(N,S)$, $i \in \{0,\dots,N\}$.

1. Вычисляем матрицы $\tau^{(k)}(S)$, $k \in \{0,\dots,M-1\}$, которые получаются удалением k первых строк и k первых столбцов из матрицы S.

2. Вычисляем матрицы $T_j = \tau^{(M-2-j)}(S)$, $j \in \{1,\dots,M-2\}$.

3. Вычисляем матрицы $L_i^{(w)}(T_j)$, используя рекуррентные формулы

$$L_i^{(0)}(T_j) = (N-i)t^j_{r_j,1},\ i \in \{0,\dots,N-1\}, j \in \{1,\dots,M-2\},$$

$$L_i^{(w)}(T_j) = \begin{pmatrix} (N-i)t^j_{r_j-w,1}I & O & \cdots & O \\ L_{N-1}^{(w-1)}(T_j) & (N-i-1)t^j_{r_j-w,1}I & \cdots & O \\ O & L_{N-2}^{(w-1)}(T_j) & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & t^j_{r_j-w,1}I \\ O & O & \cdots & L_i^{(w-1)}(T_j) \end{pmatrix},$$

$$w \in \{1,\dots,r_j-2\},\ i \in \{0,\dots,N-1\}, j \in \{1,\dots,M-2\},$$

где $t^j_{k,l}$ – (k,l)-й элемент матрицы T_j, r_j – число строк матрицы T_j.

4. Вычисляем матрицы $U_i^{(w)}(T_j)$, используя рекуррентные формулы

$$U_i^{(0)}(T_j) = t_{1,r_j}^j,\ i \in \{1,\ldots,N\},\ j \in \{1,\ldots,M-2\},$$

$$U_i^{(w)}(T_j) = \begin{pmatrix} t_{1,r_j-w}^j I & U_N^{(w-1)}(T_j) & O & \cdots & O & O \\ O & t_{1,r_j-w}^j I & U_{N-1}^{(w-1)}(T_j) & \cdots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \cdots & t_{1,r_j-w}^j I & U_i^{(w-1)}(T_j) \end{pmatrix},$$

$$w \in \{1,\ldots,r_j-2\},\ i \in \{1,\ldots,N\},\ j \in \{1,\ldots,M-2\}.$$

5. Вычисляем матрицы $L_i(N,T_j) = L_i^{(r_j-2)}(T_j),\ i \in \{0,\ldots,N-1\}$, и $U_i(N,T_j) = iU_i^{(r_j-2)}(T_j)$, $i \in \{1,\ldots,N\},\ j \in \{1,\ldots,M-2\}$.

6. Вычисляем матрицы $A_i^{(w)}$, используя рекуррентные формулы

$$A_i^{(0)} = \begin{pmatrix} O & iS_{M-1,M} & O & \cdots & O & O \\ S_{M,M-1} & 0 & (i-1)S_{M-1,M} & \cdots & O & O \\ O & 2S_{M,M-1} & O & \cdots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \cdots & O & S_{M-1,M} \\ O & O & O & \cdots & iS_{M,M-1} & O \end{pmatrix},\ i \in \{1,\ldots,N\},$$

$$A_i^{(j)} = \begin{pmatrix} O & \dfrac{iU_N(N,T_j)}{N} & O & \cdots & O & O \\ L_{N-1}(N,T_j) & A_1^{(j-1)} & \dfrac{(i-1)U_{N-1}(N,T_j)}{N-1} & \cdots & O & O \\ O & L_{N-2}(N,T_j) & A_2^{(j-1)} & \cdots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \cdots & A_{i-1}^{(j-1)} & \dfrac{U_{N-i+1}(N,T_j)}{N-i+1} \\ O & O & O & \cdots & L_{N-i}(N,T_j) & A_i^{(j-1)} \end{pmatrix},$$

$$i \in \{1,\ldots,N\},\ j \in \{1,\ldots,M-2\}.$$

7. Вычисляем матрицы $A_i(N,S)$ следующим образом: $A_0(N,S) = O_{1\times 1}, A_i(N,S) = A_i^{(M-2)}$, $i \in \{1,\ldots,N\}$.

Замечание 1. Если $M \leq 2$, то шаги 1–5 пропускаем и вычисляем только матрицы $A_i^{(0)}$, $i \in \{1,\ldots,N\}$, на шаге 6.

Алгоритм вычисления матриц $L_i(N,\tilde{S}),\ i \in \{0,\ldots,N-1\}$.

1. Вычисляем матрицы $L_i(\tilde{S})$, используя рекуррентные формулы

$$L_i^{(0)}(\tilde{S}) = (N-i)t_{r,1},\ i \in \{0,\ldots,N-1\},$$

$$L_i^{(w)}(\tilde{S}) = \begin{pmatrix} (N-i)t_{r-w,1}I & O & \cdots & O \\ L_{N-1}^{(w-1)}(\tilde{S}) & (N-i-1)t_{r-w,1}I & \cdots & O \\ O & L_{N-2}^{(w-1)}(\tilde{S}) & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & t_{r-w,1}I \\ O & O & \cdots & L_i^{(w-1)}(\tilde{S}) \end{pmatrix},$$

$$w \in \{1,\ldots,r-2\},\ i \in \{0,\ldots,N-1\},$$

где $t_{k,l}$ – (k,l)-й элемент матрицы $\tilde{S}$, r – число строк матрицы $\tilde{S}$.

2. Вычисляем матрицы $L_i(N,\tilde{S}) = L_i^{(r-2)}(\tilde{S})$, $i \in \{0,\ldots,N-1\}$.

3. Вычисляем матрицы $L_i(N,\tilde{S})$ следующим образом: $L_i(N,\tilde{S}) = L_i^{(M-1)}(\tilde{S})$, $i \in \{0,\ldots,N-1\}$, $L_N(N,\tilde{S}) = O_{1\times 1}$.

Замечание 2. Здесь наиболее трудоемкий шаг 1 повторяет шаг 3 при вычислении матриц $A_i(N,S)$, только матрицы T_j заменяются на матрицу $\tilde{S}$ и не зависят от j.

Алгоритм вычисления матриц $P_i(\boldsymbol{\beta})$, $i \in \{1,\ldots,N-1\}$.

1. Вычисляем матрицы $P_i^{(j)}$ размерности $(i+1)\times(i+2)$, используя рекуррентные формулы

$$P_i^{(0)} = \begin{pmatrix} \beta_{M-1} & \beta_M & 0 & \cdots & 0 & 0 \\ 0 & \beta_{M-1} & \beta_M & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \beta_{M-1} & \beta_M \end{pmatrix}, i \in \{1,\ldots,N-1\},$$

$$P_i^{(j)} = \begin{pmatrix} \beta_{M-j-1} & z^{(j)} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & \beta_{M-j-1}I & P_1^{(j-1)} & O & \cdots & O & O \\ \mathbf{0}^T & O & \beta_{M-j-1}I & P_2^{(j-1)} & \cdots & O & O \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}^T & O & O & O & \cdots & \beta_{M-j-1}I & P_i^{(j-1)} \end{pmatrix},$$

$$j \in \{1,\ldots,M-2\}, i \in \{1,\ldots,N-1\},$$

где $z^{(j)} = (\beta_{M-j},\beta_{M-j+1},\ldots,\beta_M)$, $j \in \{1,\ldots,M-2\}$.

2. Вычисляем матрицы $P_i(\boldsymbol{\beta})$ следующим образом: $P_0(\boldsymbol{\beta}) = \boldsymbol{\beta}$, $P_i(\boldsymbol{\beta}) = P_i^{(M-2)}$, $i \in \{1,\ldots,N-1\}$.

Информация об авторе

Клименок Валентина Ивановна, доктор физико-математических наук, профессор, главный научный сотрудник научно-исследовательской лаборатории прикладного вероятностного анализа, Белорусский государственный университет.
E-mail: klimenok@bsu.by

Information about the author

Valentina I. Klimenok, D. Sc., Prof., Chief Scientific Researcher of Laboratory of Applied Probability, Belorussian State University.
E-mail: klimenok@bsu.by