Теория вероятностей и математическая статистика

Probability theory and mathematical statistics

УДК 519.872

СИСТЕМА МАССОВОГО ОБСЛУЖИВАНИЯ С ГРУППОВЫМ МАРКОВСКИМ ПОТОКОМ И МЕНЯЮЩИМИСЯ ПРИОРИТЕТАМИ

В. И. КЛИМЕНОК[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Рассматривается однолинейная система массового обслуживания с конечным буфером и групповым марковским потоком. Запросы, принятые в буфер, могут иметь низший или высший приоритет. Сразу после поступления каждому из запросов назначается низший приоритет и для него устанавливается таймер, который задается случайной величиной, распределенной по фазовому закону и имеющей два поглощающих состояния. После попадания таймера в одно из поглощающих состояний запрос может уйти из системы навсегда (потеряться) или изменить свой приоритет на высший. При попадании таймера в другое поглощающее состояние запрос с некоторой вероятностью теряется и с дополнительной вероятностью таймер устанавливается заново. Если запрос поступает в полностью заполненную систему, он теряется. Такого типа системы могут служить математическими моделями многих реальных систем оказания медицинской помощи, контакт-центров, систем хранения скоропортящихся продуктов и т. д. Функционирование системы описывается в терминах многомерной цепи Маркова, вычисляется стационарное распределение и ряд важных характеристик производительности системы. Отличие данной работы от имеющихся литературных источников заключается в формулировке модели, в более общем и реалистичном

Образец цитирования:
Клименок ВИ. Система массового обслуживания с групповым марковским потоком и меняющимися приоритетами. *Журнал Белорусского государственного университета. Математика. Информатика.* 2022;2:47–56.
https://doi.org/10.33581/2520-6508-2022-2-47-56

For citation:
Klimenok VI. A queueing system with a batch Markovian arrival process and varying priorities. *Journal of the Belarusian State University. Mathematics and Informatics.* 2022;2:47–56. Russian.
https://doi.org/10.33581/2520-6508-2022-2-47-56

Автор:
Валентина Ивановна Клименок – доктор физико-математических наук, профессор; главный научный сотрудник научно-исследовательской лаборатории прикладного вероятностного анализа факультета прикладной математики и информатики.

Author:
Valentina I. Klimenok, doctor of science (physics and mathematics), full professor; chief researcher at the laboratory of applied probabilistic analysis, faculty of applied mathematics and computer science.
klimenok@bsu.by
https://orcid.org/0000-0002-3903-6444

характере распределений, описывающих происходящие в системе процессы, а также в исчерпывающих результатах, включающих формулы и алгоритмы для вычисления стационарного распределения и характеристик производительности системы.

Ключевые слова: система массового обслуживания; конечный буфер; групповой марковский поток; меняющиеся приоритеты; стационарное распределение; характеристики производительности.

A QUEUEING SYSTEM WITH A BATCH MARKOVIAN ARRIVAL PROCESS AND VARYING PRIORITIES

V. I. KLIMENOK[a]

[a]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*

We consider herein a single-server queueing system with a finite buffer and a batch Markovian arrival process. Customers staying in the buffer may have a lower or higher priority. Immediately after arrival each of the customer is assigned the lowest priority and a timer is set for it, which is defined as a random variable distributed according to the phase law and having two absorbing states. After the timer enters one of the absorbing states, the customer may leave the system forever (get lost) or change its priority to the highest. When the timer enters another absorbing state, the customer is lost with some probability and the timer is set again with an additional probability. If a customer enters a completely full system, it is lost. Systems of this type can serve as mathematical models of many real-life medical care systems, contact centers, perishable food storage systems, etc. The operation of the system is described in terms of a multidimensional Markov chain, the stationary distribution and a number of performance characteristics of the system are calculated.

Keywords: queueing system; finite buffer; batch Markovian arrival process; changing priorities; stationary distribution; performance characteristics.

Введение

Важная ветвь теории массового обслуживания – исследование приоритетных систем с потоком запросов, приоритеты которых могут меняться в процессе ожидания обслуживания. Существует обширная классификация приоритетов, согласно которой выделяются классы относительных, абсолютных, статических, динамических и ряда других приоритетов. Сравнивая статические и динамические приоритеты, можно заметить, что динамические приоритеты, смена которых зависит от длины очереди, эффективнее статических, однако имеют более узкую область применения, поскольку иногда длины очередей не полностью наблюдаемы, а управление приоритетами обходится дорого. Поэтому статические приоритеты по-прежнему популярны во многих реальных системах. Основным недостатком классических статических приоритетов являются их негибкость и возможная несправедливость по отношению к низкоприоритетным клиентам. Для преодоления этой несправедливости могут быть предложены различные улучшения статических приоритетов, например ограничение слишком быстрого доступа приоритетных запросов или обязательное обслуживание неприоритетного запроса после обслуживания фиксированного числа приоритетных запросов. Еще одно популярное улучшение состоит в возможности повысить приоритет запроса во время его нахождения в очереди. Есть работы (см., например, [1–5]), где запрос за время пребывания в системе накапливает приоритет от начального значения, зависящего от приоритета запроса, в соответствии с некоторой линейной или нелинейной функцией. Другая группа работ предполагает, что повышение приоритета происходит не детерминированно, а через случайное время. Поскольку в настоящей статье делается аналогичное предположение, чтобы прояснить новизну представленных в ней модели и результатов, кратко опишем соответствующие результаты, имеющиеся в литературе.

В публикациях [6; 7] рассматриваются несколько приоритетных классов запросов, поступающих в маркированном марковском потоке. Времена обслуживания имеют распределения фазового типа, времена до повышения приоритета распределены по закону Кокса. В результате получено условие эргодичности цепи Маркова, описывающей функционирование системы. Более полно система с маркированным марковским потоком и двумя приоритетами исследована в статье [8]. Для этой системы вычислены стационарное распределение и характеристики производительности, включая распределение времени обслуживания приоритетных запросов. В остальных известных автору работах [9–13] предполагаются

стационарный пуассоновский поток и экспоненциальное распределение времени обслуживания, а также времени до повышения приоритета. Результатом этих исследований в зависимости от статьи являются условия эргодичности, численная оптимизация и анализ асимптотического поведения соответствующих систем.

В данной работе рассматривается однолинейная система с групповым марковским потоком (*batch Markovian arrival process*, BMAP). Времена обслуживания, как и времена до повышения приоритета, распределены по фазовому закону (*phase type distribution*, PH). Процесс обслуживания имеет одно поглощающее состояние, переход в которое означает окончание обслуживания, а время до повышения приоритета имеет два поглощающих состояния, переход в которые влечет за собой изменение приоритета запроса на высший либо уход необслуженного запроса из системы. В статье построена и исследована многомерная цепь Маркова, которая описывает функционирование системы, и получено ее стационарное распределение. Таким образом, преимущество настоящей работы заключается в более общем и реалистичном характере распределений, описывающих происходящие в системе процессы, а также в исчерпывающих результатах, включающих формулы и алгоритмы для вычисления стационарного распределения и характеристик производительности системы. Стоит отметить, что выбор более реалистичных, чем экспоненциальное, распределений, описывающих процессы поступления и обслуживания запросов и смену приоритетов, очень важен для потенциальных реальных приложений. Одним из наиболее популярных приложений такого рода моделей является здравоохранение (упоминается практически во всех цитируемых выше публикациях). Например, рассматриваемая система массового обслуживания подходит для описания работы отделения неотложной помощи в больнице, имеющей операционную и бригаду необходимых врачей и медсестер. В эту больницу доставляют пациентов, пострадавших в результате несчастного случая. В процессе сортировки и предварительной обработки пациент может быть выписан в удовлетворительном состоянии, либо оставлен для дальнейшей обработки, либо отправлен на срочное лечение (в этом случае он становится высокоприоритетным пациентом). Понятно, что групповое поступление больных характерно для ситуаций, связанных с дорожными или другими инцидентами (см. [14]), а распределение фазового типа намного лучше, чем экспоненциальное распределение, подходит для описания времени, пока пациент не покинет больницу (например, уйдет удовлетворенный предварительной обработкой, или переведется в другую больницу, или умрет), либо пока будет продолжена его предварительная обработка, либо пока он станет приоритетным пациентом.

Рассмотренная в данной статье модель также может быть использована для описания работы контакт-центра. Как отмечается в литературе, телефонные звонки имеют высокий приоритет, в то время как запросы, отправленные по электронной почте или через мессенджер, имеют низкий приоритет, но клиент, который воспользовался мессенджером для получения информации, может сделать телефонный звонок, если он слишком долго ждет ответа. Заметим, что в работе [15] поясняется, что экспоненциальные допущения могут быть совершенно неадекватными, и для моделирования контакт-центра следует использовать более общие распределения или потоки, как это сделано в настоящей статье.

Описание системы

Рассматривается система массового обслуживания с конечным буфером размера N и BMAP-потоком запросов. В этом потоке группы запросов случайного размера поступают под управлением неприводимой цепи Маркова с непрерывным временем ν_t, $t \geq 0$, которая принимает значения в множестве $\{0, 1, 2, \ldots, W\}$ и называется управляющим процессом BMAP. 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' с генерацией группы запросов размера k, $k \geq 1$, или с вероятностью $p_0(\nu, \nu')$ цепь переходит в состояние ν' без генерации запросов, причем $p_0(\nu, \nu) = 0$. Для указанных вероятностей выполняются естественные ограничения $\sum\limits_{k=1}^{\infty} \sum\limits_{\nu'=0}^{W} p_k(\nu, \nu') = 1$, ν, $\nu' = \overline{0, W}$. Всю информацию о BMAP удобно хранить в виде набора матриц D_k, $k \geq 0$, порядка $(W+1) \times (W+1)$, элементы которых определяются как

$$(D_k)_{\nu, \nu'} = \lambda_\nu p_k(\nu, \nu'),\ \nu, \nu' = \overline{0, W},\ k \geq 1,\ (D_0)_{\nu, \nu'} = \begin{cases} \lambda_\nu p_0(\nu, \nu'),\ \nu \neq \nu',\ \nu, \nu' = \overline{0, W}, \\ -\lambda_\nu,\ \nu = \nu' = \overline{0, W}. \end{cases}$$

Из формул видно, что элементами матриц D_k, $k \geq 1$, являются интенсивности переходов процесса ν_t, сопровождающихся генерацией группы запросов размера k. Аналогичный смысл имеют недиагональные

элементы матрицы D_0, а диагональные элементы этой матрицы есть взятые с противоположным знаком интенсивности выхода процесса ν_t из своих состояний. Матрицы D_k, $k \geq 0$, можно задавать их матричной производящей функцией $D(z) = \sum_{k=0}^{\infty} D_k z^k$, $|z| < 1$. Значение этой функции в точке $z = 1$ является инфинитезимальным генератором управляющего процесса ν_t, $t \geq 0$. Стационарное распределение данного процесса, представленное в виде вектор-строки $\boldsymbol{\theta}$, определяется как решение системы линейных алгебраических уравнений $\boldsymbol{\theta} D(1) = \mathbf{0}$, $\boldsymbol{\theta e} = 1$. Здесь и далее $\mathbf{0}$ – вектор-строка, состоящая из нулей, $\boldsymbol{e}$ – вектор-столбец, состоящий из единиц. Средняя интенсивность поступления запросов в BMAP-потоке задается формулой $\lambda = \boldsymbol{\theta} \sum_{k=1}^{\infty} k D_k \boldsymbol{e}$. Более подробное описание BMAP, включающее формулы для дисперсии длин интервалов между моментами поступления групп запросов и коэффициентов корреляции длин двух соседних интервалов между моментами поступления групп запросов, можно найти, например, в работах [16; 17].

Времена обслуживания запросов имеют PH-распределение с неприводимым представлением $(\boldsymbol{\beta}, S)$ и управляющим процессом (цепью Маркова) m_t, $t \geq 0$, принимающим значения в множестве $\{1, \ldots, M, M+1\}$, где состояние $M+1$ является поглощающим. Первоначальное состояние процесса m_t определяется в множестве несущественных состояний $\{1, \ldots, M\}$ в соответствии со стохастической вектор-строкой $\boldsymbol{\beta}$. Интенсивности переходов в множестве несущественных состояний задаются $(M \times M)$-матрицей S, интенсивности переходов в поглощающее состояние определяются вектор-столбцом $\boldsymbol{S}_0 = -S\boldsymbol{e}$. Более подробную информацию о PH-распределении и его свойствах можно найти, например, в работах [17; 18].

Предполагается, что новый запрос, поступивший в систему, обладает низшим приоритетом. Для каждого такого запроса устанавливается таймер, который задается PH-распределением с управляющим процессом (цепью Маркова) r_t, $t \geq 0$, имеющим множество несущественных состояний $\{1, 2, \ldots, R\}$ и два поглощающих состояния – состояние $*$ и состояние $**$. Интенсивности переходов управляющего процесса в множестве несущественных состояний задаются $(R \times R)$-матрицей Г. Интенсивности переходов в поглощающие состояния определяются вектор-столбцом $\boldsymbol{\Gamma}_0 = -\Gamma \boldsymbol{e}$. Интенсивности переходов в поглощающее состояние $*$ и поглощающее состояние $**$ задаются вектор-столбцами $\boldsymbol{\Gamma}_0^{(*)} = æ\boldsymbol{\Gamma}_0$ и $\boldsymbol{\Gamma}_0^{(**)} = (1 - æ)\boldsymbol{\Gamma}_0$, где $0 < æ < 1$. Когда таймер достигает состояния $*$, запрос с вероятностью p уходит из системы необслуженным и с дополнительной вероятностью $\bar{p} = 1 - p$ остается в очереди неприоритетных запросов. В последнем случае таймер для него устанавливается заново. Когда таймер достигает состояния $**$, запрос с вероятностью q уходит из системы необслуженным и с дополнительной вероятностью $\bar{q} = 1 - q$ приобретает высший приоритет. Во втором случае таймер сворачивается и запрос становится впереди всех неприоритетных запросов и в конце очереди приоритетных запросов.

Цепь Маркова, описывающая процесс функционирования системы

Положим в момент времени t:

- i_t – число заявок в очереди, $i_t = \overline{0, N}$;
- j_t – число неприоритетных заявок в очереди, $j_t = \overline{0, i_t}$;
- $n_t = 0$, если прибор свободен, и $n_t = 1$, если прибор обслуживает запрос;
- m_t – состояние управляющего процесса PH-обслуживания на приборе, $m_t = \overline{1, M}$;
- $r_t^{(n)}$ – состояние управляющего процесса PH-таймера для n-й неприоритетной заявки, стоящей в очереди, $r^{(n)} = \overline{1, R}$, $n = \overline{1, j}$;
- ν_t – состояние управляющего процесса BMAP-потока, $\nu_t = \overline{0, W}$.

Процесс функционирования системы описывается неприводимой цепью Маркова ξ_t, $t \geq 0$, с пространством состояний

$$\left\{(0, n, \nu),\ i = 0,\ n = 0, 1;\ \nu = \overline{0, W}\right\} \cup$$

$$\cup \left\{\left(i, j, \nu, m, r^{(1)}, \ldots, r^{(j)}\right),\ i = \overline{0, N},\ j = \overline{0, i},\ \nu = \overline{0, W},\ m = \overline{1, M},\ r^{(n)} = \overline{1, R},\ n = \overline{1, j}\right\}.$$

Упорядочим состояния цепи при каждом фиксированном значении компоненты i_t в лексикографическом порядке и образуем матрицы $Q_{i,\,l}$ интенсивностей переходов из множества состояний, соответ-

ствующих значению $i_t = i$, в состояния, соответствующие значению $i_t = l$. Тогда инфинитезимальный генератор Q рассматриваемой цепи формируется как $Q = \left(Q_{i,l}\right)_{i,l=\overline{0,N}}$.

Введем обозначения:

• $\bar{W} = W + 1$;

• O – матрица, состоящая из нулей, I – тождественная матрица (при необходимости порядок матрицы определяется нижним индексом);

• $\otimes$ ($\oplus$) – символ кронекерова произведения (кронекеровой суммы) матриц (см., например, [19]);

• $\mathrm{diag}\{a_1, a_2, \ldots, a_n\}$ – диагональная блочная матрица, у которой диагональные блоки равны элементам, перечисленным в скобках, а остальные блоки являются нулевыми;

• $\mathrm{diag}^{-}\{a_1, a_2, \ldots, a_n\}$ – квадратная блочная матрица, у которой поддиагональные блоки равны элементам, перечисленным в скобках, а остальные блоки являются нулевыми.

Справедливо следующее утверждение.

Лемма 1. *Инфинитезимальный генератор цепи Маркова ξ_t, $t \geq 0$, имеет блочную структуру*

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

где ненулевые блоки описываются следующим образом:

$$Q_{0,0} = \begin{pmatrix} D_0 & D_1 \otimes \boldsymbol{\beta} \\ I_{\bar{W}} \otimes \boldsymbol{S}_0 & D_0 \oplus S \end{pmatrix},$$

$$Q_{0,k} = \begin{pmatrix} O_{\bar{W}M\frac{1-R^k}{1-R}} & D_{k+1} \otimes \boldsymbol{\beta} \otimes \boldsymbol{\gamma}^{\otimes k} \\ O & D_k \otimes \boldsymbol{\gamma}^{\otimes k} \end{pmatrix}, \quad k = \overline{1, N-1},$$

$$Q_{0,N} = \begin{pmatrix} O_{\bar{W}M\frac{1-R^N}{1-R}} & \sum\limits_{k=N+1}^{\infty} D_k \otimes \boldsymbol{\beta} \otimes \boldsymbol{\gamma}^{\otimes N} \\ O & \sum\limits_{k=N+1}^{\infty} D_k \otimes \boldsymbol{\gamma}^{\otimes k} \end{pmatrix},$$

$$Q_{1,0} = \begin{pmatrix} I_{\bar{W}} \otimes \boldsymbol{S}_0\boldsymbol{\beta} & O_{\bar{W}M} \\ I_{\bar{W}} \otimes \boldsymbol{S}_0\boldsymbol{\beta} \otimes \boldsymbol{e}_R + I_{\bar{W}M} \otimes \left(p\boldsymbol{\Gamma}_0^{(*)} + q\boldsymbol{\Gamma}_0^{(**)}\right) & O \end{pmatrix},$$

$$Q_{i,i-1} = \begin{pmatrix} I_{\bar{W}} \otimes \boldsymbol{S}_0\boldsymbol{\beta} & O & O & \ldots & O & O \\ O & I_{\bar{W}} \otimes \boldsymbol{S}_0\boldsymbol{\beta} \otimes I_R & O & \ldots & O & O \\ O & O & I_{\bar{W}} \otimes \boldsymbol{S}_0\boldsymbol{\beta} \otimes I_{R^2} & \ldots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \ldots & O & I_{\bar{W}} \otimes \boldsymbol{S}_0\boldsymbol{\beta} \otimes I_{R^{i-1}} \\ O & O & O & \ldots & O & I_{\bar{W}} \otimes \boldsymbol{S}_0\boldsymbol{\beta} \otimes \boldsymbol{e}_R \otimes I_{R^{i-1}} \end{pmatrix} +$$

$$+ \mathrm{diag}^{-}\left\{I_{\bar{W}M} \otimes \left[p\left(\boldsymbol{\Gamma}_0^{(*)}\right)^{\oplus j} + q\left(\boldsymbol{\Gamma}_0^{(**)}\right)^{\oplus j}\right], \ j = \overline{1, i}\right\}, \quad i = \overline{2, N},$$

$$Q_{i,\,i} = \begin{pmatrix} \tilde{D} \oplus S & O & \dots & O & O \\ I_{\bar{W}M} \otimes \bar{q}\left(\boldsymbol{\Gamma}_0^{(**)}\right)^{\oplus 1} & \left(\tilde{D} \oplus S\right) \otimes I_R & \dots & O & O \\ O & I_{\bar{W}M} \otimes \bar{q}\left(\boldsymbol{\Gamma}_0^{(**)}\right)^{\oplus 2} & \dots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \dots & \left(\tilde{D} \oplus S\right) \otimes I_{R^{i-1}} & O \\ O & O & \dots & I_{\bar{W}M} \otimes \bar{q}\left(\boldsymbol{\Gamma}_0^{(**)}\right)^{\oplus i} & \left(\tilde{D} \oplus S\right) \otimes I_{R^i} \end{pmatrix} +$$

$$+ \operatorname{diag}\left\{ I_{\bar{W}M} \otimes \left[\Gamma^{\oplus j} + \bar{p}\left(\boldsymbol{\Gamma}_0^{(*)}\boldsymbol{\gamma}\right)^{\oplus j} \right],\ j = \overline{0,\,i} \right\},\ i = \overline{1,\,N},$$

здесь $\tilde{D} = D_0$, *если* $i = \overline{1,\,N-1}$, *и* $\tilde{D} = D(1)$, *если* $i = N$,

$$Q_{i,\,i+k} = \left(O_{\bar{W}M\frac{1-R^{i+1}}{1-R} \times \bar{W}M\frac{1-R^{k}}{1-R}} \left| \operatorname{diag}\left\{ D^{(k,\,j)},\ j = \overline{0,\,i} \right\} \right. \right),\ i = \overline{1,\,N-1},\ k = \overline{1,\,N-i},$$

здесь $D^{(k,\,j)} = D_k \otimes I_M \otimes I_{R^j} \otimes \boldsymbol{\gamma}^{\oplus k}$, *если* $k = \overline{1,\,N-i-1}$, *и* $D^{(N-i,\,j)} = \sum_{l=N-i}^{\infty} D_l \otimes I_M \otimes I_{R^j} \otimes \boldsymbol{\gamma}^{\oplus N-i}$, *если* $k = N-i$.

Доказательство леммы проводится путем анализа поведения цепи Маркова ξ_t, $t \geq 0$, на бесконечно малом интервале времени. Опишем кратко вероятностный смысл ненулевых блоков генератора. Блок $Q_{i,\,i-1}$, $i \geq 1$, состоит из интенсивностей переходов рассматриваемой цепи Маркова, в результате которых число запросов в буфере уменьшается с i до $i-1$. Если в буфере находится j неприоритетных запросов, то такие переходы могут быть вызваны либо окончанием текущего обслуживания и занятием прибора приоритетным запросом (соответствующие интенсивности задаются матрицей $I_{\bar{W}} \otimes \boldsymbol{S}_0\boldsymbol{\beta} \otimes I_{R^j}$), либо окончанием обслуживания и занятием прибора неприоритетным запросом, если в буфере нет приоритетных запросов (матрица $I_{\bar{W}} \otimes \boldsymbol{S}_0\boldsymbol{\beta} \otimes \boldsymbol{e}_R \otimes I_{R^{i-1}}$), либо уходом из системы одного из неприоритетных запросов вследствие попадания установленного для него таймера в одно из поглощающих состояний (матрица $I_{\bar{W}M} \otimes \left[p\left(\boldsymbol{\Gamma}_0^{(*)}\right)^{\oplus j} + q\left(\boldsymbol{\Gamma}_0^{(**)}\right)^{\oplus j} \right]$).

Блок $Q_{i,\,i+k}$, $i \geq 0$, $k \geq 1$, состоит из интенсивностей переходов, сопровождающихся поступлением группы запросов. Если размер группы не превышает числа $N-i$ свободных мест в буфере, то вся группа принимается в буфер (соответствующие интенсивности задаются матрицами $D_k \otimes I_M \otimes I_{R^j} \otimes \boldsymbol{\gamma}^{\oplus k}$). В противном случае в буфер принимается только $N-i$ запросов группы, а остальные запросы теряются (матрица $D^{(N-i,\,j)} = \sum_{l=N-i}^{\infty} D_l \otimes I_M \otimes I_{R^j} \otimes \boldsymbol{\gamma}^{\oplus N-i}$).

Блок $Q_{i,\,i}$ состоит из интенсивностей переходов, не влекущих за собой изменения числа i запросов в буфере. Поддиагональные блоки $I_{\bar{W}M} \otimes \bar{q}\left(\boldsymbol{\Gamma}_0^{(**)}\right)^{\oplus j}$ описывают интенсивности переходов таймера какого-то из j неприоритетных запросов в поглощающее состояние $**$, в результате которых этот запрос становится приоритетным. Недиагональные элементы диагональных блоков матрицы $Q_{i,\,i}$ описывают интенсивности переходов управляющих процессов BMAP, таймеров и времени обслуживания, не влекущих изменения числа запросов в системе (матрицы $\left(\tilde{D} \oplus S\right) \otimes I_{R^j} + I_{\bar{W}M} \otimes \Gamma^{\oplus j}$), либо переход таймера какого-то из j неприоритетных запросов в буфере в поглощающее состояние $*$ (матрицы $\bar{p}\left(\boldsymbol{\Gamma}_0^{(*)}\boldsymbol{\gamma}\right)^{\oplus j}$), после которого запрос не меняет приоритет, а таймер на нем устанавливается заново. Диагональные элементы диагональных блоков матрицы $Q_{i,\,i}$ есть взятые с противоположным знаком интенсивности выхода рассматриваемой цепи Маркова ξ_t, $t \geq 0$, из состояний, соответствующих i запросам в буфере.

Стационарное распределение

Поскольку исследуемая цепь Маркова ξ_t, $t \ge 0$, является неприводимой непериодической цепью с конечным пространством состояний, то она имеет эргодическое распределение, совпадающее с единственным стационарным распределением. Пусть $\boldsymbol{p}_i$ – вектор-строка стационарных вероятностей состояний, имеющих значение i первой компоненты, $i = \overline{0, N}$. Элементы вектора $\boldsymbol{p}_0$ дают стационарные вероятности того, что буфер пуст, прибор простаивает или занят, управляющий процесс BMAP находится в любом из $W + 1$ состояний и, если прибор занят, время обслуживания находится в одной из M фаз. Заметим, что j-й подвектор вектора $\boldsymbol{p}_i$, $i = \overline{1, N}$, дает стационарные вероятности того, что в буфере есть i запросов, из них j неприоритетных, управляющий процесс BMAP находится в любом из $W + 1$ состояний, время обслуживания – в одной из M фаз, а процесс таймеров, установленных для неприоритетных запросов, стоящих в буфере, – в любом из R^j состояний.

Сформируем из этих векторов вектор $\boldsymbol{p} = (\boldsymbol{p}_0, \boldsymbol{p}_i, \ldots, \boldsymbol{p}_N)$ стационарных вероятностей рассматриваемой цепи. Хорошо известно, что этот вектор является единственным решением следующей системы линейных алгебраических уравнений:

$$\boldsymbol{p}Q = \boldsymbol{0}, \boldsymbol{p}\boldsymbol{e} = 1. \tag{1}$$

В случае малой размерности система (1) может быть решена на компьютере стандартными методами. Однако при более или менее больших значениях N, R порядок этой системы становится настолько большим, что решить ее напрямую (например, методом обратной матрицы) невозможно. В таком случае используется алгоритм, который был разработан в статье [20]. Алгоритм устойчив, учитывает верхнюю хессенбергову структуру генератора Q и работает с блоками генератора $Q_{i,\,l}$. Размеры этих блоков определяются значениями $N_0 = \bar{W}(1 + M)$ при $i = 0$ и $N_i = \bar{W}M \sum_{j=0}^{i} R^i$ при $i = \overline{1, N}$, а размер всей системы (1) равен $\sum_{i=0}^{N} N_i$. Для удобства читателя приведем принципиальные шаги алгоритма.

Шаг 1. Находим матрицы G_{N-1}, G_{N-2}, …, G_0 из уравнения обратной рекурсии

$$G_i = \left(-Q_{i+1,\,i+1} - Q_{i+1,\,i+2}G_{i+1}\right)^{-1} Q_{i+1,\,i},\ i = N-2,\ N-1, \ldots, 0,$$

где полагаем, что $G_{N-1} = \left(-Q_{N,\,N}\right)^{-1} Q_{N,\,N-1}$.

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

$$\bar{Q}_{N,\,N} = Q_{N,\,N},\ \bar{Q}_{i,\,i} = Q_{i,\,i} + Q_{i,\,i+1}G_i,\ i = \overline{0,\ N-1},\ \bar{Q}_{i,\,i+1} = Q_{i,\,i+1},\ i = \overline{0,\ N-1}.$$

Шаг 3. Находим матрицы F_i из рекуррентных соотношений

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

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

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

Шаг 5. Вычисляем векторы $\boldsymbol{p}_i$ по формуле $\boldsymbol{p}_i = \boldsymbol{p}_0 F_i$, $i = \overline{1,\ N}$.

Стационарные характеристики производительности

Вычислив векторы стационарных вероятностей $\boldsymbol{p}_i$, $i = \overline{0,\ N}$, можем вычислить ряд представляющих интерес характеристик производительности системы. Ниже приведены выражения для наиболее важных характеристик вместе с краткими пояснениями к нетривиальным формулам.

1. Вероятность того, что система свободна,

$$p_0 = \boldsymbol{p}_0 \begin{pmatrix} \boldsymbol{e}_{\bar{W}} \\ \boldsymbol{0}^T_{\bar{W}M} \end{pmatrix}.$$

2. Вероятность того, что в системе один запрос (обслуживается на приборе),

$$p_1 = \boldsymbol{p}_0 \begin{pmatrix} \boldsymbol{0}^T_{\bar{W}} \\ \boldsymbol{e}_{\bar{W}M} \end{pmatrix}.$$

3. Вероятность того, что в буфере находится i запросов,

$$p_i = \sum_{j=0}^{i} \boldsymbol{p}_i \boldsymbol{e},\ i = \overline{0,\ N}.$$

4. Вероятность того, что в буфере находится i запросов, из них j неприоритетных,

$$p_{i,\,j} = \begin{pmatrix} \mathbf{0}^T_{\sum_{n=0}^{j-1} \bar{W} M \frac{R^{n+1}-1}{R-1}} \\ \boldsymbol{e}_{\bar{W} M \frac{R^{j+1}-1}{R-1}} \\ \mathbf{0}^T_{\sum_{n=j+1}^{i} \bar{W} M \frac{R^{n+1}-1}{R-1}} \end{pmatrix},\ j = \overline{0,\ i},\ \ i = \overline{1,\ N}.$$

5. Среднее число запросов в буфере

$$L = \sum_{i=1}^{N} i \boldsymbol{p}_i \boldsymbol{e}.$$

6. Среднее число неприоритетных запросов в буфере

$$L^{(\text{non-prior})} = \sum_{i=1}^{N} \sum_{j=1}^{i} j p_{i,\,j}.$$

7. Среднее число приоритетных запросов в буфере

$$L^{(\text{prior})} = L - L^{(\text{non-prior})}.$$

8. Вероятность того, что произвольный запрос будет потерян либо из-за недостатка мест в буфере, либо из-за попадания таймера в поглощающее состояние (далее будем говорить «вследствие нетерпеливости»),

$$P_{\text{loss}} = 1 - \frac{1}{\lambda} \left[\boldsymbol{p}_0 \begin{pmatrix} O_{\bar{W} \times M} \\ \boldsymbol{e}_{\bar{W}} \otimes I_M \end{pmatrix} + \sum_{i=1}^{N} \boldsymbol{p}_i H_i \right] \boldsymbol{S}_0, \tag{2}$$

где

$$H_i = \begin{pmatrix} \boldsymbol{e}_{\bar{W}} \otimes I_M \\ \boldsymbol{e}_{\bar{W}} \otimes I_M \otimes \boldsymbol{e}_R \\ \boldsymbol{e}_{\bar{W}} \otimes I_M \otimes \boldsymbol{e}_{R^2} \\ \vdots \\ \boldsymbol{e}_{\bar{W}} \otimes I_M \otimes \boldsymbol{e}_{R^i} \end{pmatrix}.$$

Краткое пояснение к формуле (2) состоит в следующем. Выражение $\left[\boldsymbol{p}_0 \begin{pmatrix} O_{\bar{W} \times M} \\ \boldsymbol{e}_{\bar{W}} \otimes I_M \end{pmatrix} + \sum_{i=1}^{N} \boldsymbol{p}_i H_i \right] \boldsymbol{S}_0$ есть интенсивность выходящего потока, а λ – интенсивность входящего потока обслуженных запросов. Тогда отношение этих интенсивностей дает вероятность того, что произвольный запрос будет принят в буфер и обслужен, а дополнительная вероятность дает искомую вероятность P_{loss}.

9. Вероятность того, что произвольный запрос будет потерян из-за недостатка мест в буфере,

$$P_{\text{loss}}^{(\text{buff})} = 1 - \lambda^{-1} \sum_{i=0}^{N-1} \boldsymbol{p}_i \hat{H}_i \sum_{k=0}^{N-1} (k + i - N) D_k \boldsymbol{e}, \tag{3}$$

где

$$\hat{H}_i = \begin{pmatrix} I_{\bar{W}} \otimes \boldsymbol{e}_M \\ I_{\bar{W}} \otimes \boldsymbol{e}_{MR} \\ I_{\bar{W}} \otimes \boldsymbol{e}_{MR^2} \\ \vdots \\ \bar{W} \otimes \boldsymbol{e}_{MR^i} \end{pmatrix}.$$

Приведем краткий вывод формулы (3). Согласно формуле полной вероятности $P_{\text{loss}}^{(\text{buff})}$ вычисляется как

$$P_{\text{loss}}^{(\text{buff})} = 1 - \sum_{i=0}^{N-1} \sum_{k=1}^{\infty} P_k P_i^{(k)} R^{(i,k)}, \tag{4}$$

где P_k – вероятность того, что произвольный запрос поступает в группе размера k; $P_i^{(k)}$ – вероятность того, что в момент поступления группы размера k в буфере находится i запросов; $R^{(i,k)}$ – вероятность того, что произвольный запрос будет потерян из-за недостатка мест в буфере при условии, что он поступит в составе группы размера k, которая застанет i запросов в буфере.

Нетрудно видеть, что

$$P_i^{(k)} = \frac{\boldsymbol{p}_i \hat{H}_i D_k \boldsymbol{e}}{\boldsymbol{\theta} D_k \boldsymbol{e}}, \quad i = \overline{0,\ N-1},\ k \geq 1, \tag{5}$$

$$P_k = \frac{k \boldsymbol{\theta} D_k \boldsymbol{e}}{\boldsymbol{\theta} \sum_{l=1}^{\infty} l D_l \boldsymbol{e}} = k \frac{\boldsymbol{\theta} D_k \boldsymbol{e}}{\lambda},\ k \geq 1, \tag{6}$$

$$R^{(i,k)} \begin{cases} 1,\ k \leq N-1, \\ \dfrac{N-i}{k},\ k > N-i,\ i = \overline{0,\ N-1}. \end{cases} \tag{7}$$

Подставляя формулы (5)–(7) в выражение (4), после некоторых алгебраических преобразований, учитывающих, что $\sum_{k=N-i+1}^{\infty} D_k \boldsymbol{e} = -\sum_{k=0}^{N-i} D_k \boldsymbol{e}$, получаем формулу (3).

10. Вероятность того, что произвольный неприоритетный запрос будет потерян вследствие нетерпеливости,

$$P_{\text{loss}}^{(\text{imp})} = P_{\text{loss}} - P_{\text{loss}}^{(\text{buff})}.$$

11. Вероятность того, что произвольный неприоритетный запрос, принятый в буфер, будет потерян вследствие нетерпеливости,

$$\overline{P}_{\text{loss}}^{(\text{imp})} = \frac{P_{\text{loss}}^{(\text{imp})}}{1 - P_{\text{loss}}^{(\text{buff})}}.$$

Заключение

В статье исследовано стационарное поведение однолинейной системы массового обслуживания с BMAP-потоком, конечным буфером и меняющимся приоритетом. Запрос с низким приоритетом может стать запросом с высоким приоритетом после случайного времени нахождения в буфере, имеющего PH-распределение. Данная система массового обслуживания может быть полезна для моделирования работы отделения неотложной помощи в больнице, контакт-центра, а также для моделирования инвентаризации скоропортящихся продуктов и т. п. Анализ системы осуществляется при более реалистичных, чем в большинстве существующих литературных источников, предположениях о характере входного потока, распределениях времени обслуживания и времени до смены приоритета. Приведены алгоритм для вычисления стационарного распределения и формулы для характеристик производительности системы, в том числе таких важных для приложений, как вероятности потерь запросов из-за отсутствия свободных мест в буфере и вследствие нетерпеливости.

Библиографические ссылки / References

1. Bilodeau B, Stanford DA. Average waiting times in the two-class *M*/*G*/1 delayed accumulating priority queue. arXiv:2001.06054v1 [Preprint]. 2020 [cited 2022 March 10]: [19 p.]. Available from: https://arxiv.org/abs/2001.06054v1.

2. Fajardo VA, Drekic S. Waiting time distributions in the preemptive accumulating priority queue. *Methodology and Computing in Applied Probability*. 2017;19:255–284. DOI: 10.1007/s11009-015-9476-1.

3. Mojalal M, Stanford DA, Caron RJ. The lower-class waiting time distribution in the delayed accumulating priority queue. *INFOR: Information Systems and Operational Research*. 2020;58(1):60–86. DOI: 10.1080/03155986.2019.1624473.

4. Sharma KC, Sharma GC. A delay dependent queue without pre-emption with general linearly increasing priority function. *Journal of the Operational Research Society*. 1994;45(8):948–953. DOI: 10.1057/jors.1994.147.

5. Stanford DA, Taylor P, Ziedins I. Waiting time distributions in the accumulating priority queue. *Queueing Systems*. 2014;77: 297–330. DOI: 10.1007/s11134-013-9382-6.

6. Qi-Ming He, Jingui Xie, Xiaobo Zhao. Stability conditions of a preemptive repeat priority *MMAP[N]*/*PH[N]*/*S* queue with customer transfers (short version). In: *Proceedings of the 13th International conference advanced stochastic models and data analysis (ASMDA-2009); 2009 June 30 – July 3; Vilnius, Lithuania*. p. 463–467.

7. Qi-Ming He, Jingui Xie, Xiaobo Zhao. Priority queue with customer upgrades. *Naval Research Logistics*. 2012;59(5):362–375. DOI: 10.1002/nav.21494.

8. Klimenok V, Dudin A, Dudina O, Kochetkova I. Queuing system with two types of customers and dynamic change of a priority. *Mathematics*. 2020;8(5):824. DOI: 10.3390/math8050824.

9. Jingui Xie, Qi-Ming He, Xiaobo Zhao. On the stationary distribution of queue lengths in a multi-class priority queueing system with customer transfers. *Queueing Systems*. 2009;62:255–277. DOI: 10.1007/s11134-009-9130-0.

10. Jingui Xie, Ping Cao, Boray Huang, Marcus Eng Hock Ong. Determining the conditions for reverse triage in emergency medical services using queuing theory. *International Journal of Production Research*. 2016;54(11):3347–3364.

11. Jingui Xie, Taozeng Zhu, An-Kuo Chao, Shuaian Wang. Performance analysis of service systems with priority upgrades. *Annals of Operations Research*. 2017;253:683–705. DOI: 10.1007/s10479-016-2370-6.

12. Ping Cao, Jingui Xie. Optimal control of a multiclass queueing system when customers can change types. *Queueing Systems*. 2016;82:285–313. DOI: 10.1007/s11134-015-9466-6.

13. Jingui Xie, Qi-Ming He, Xiaobo Zhao. Stability of a priority queueing system with customer transfers. *Operations Research Letters*. 2008;36:705–709. DOI: 10.1016/j.orl.2008.06.007.

14. Cildoz M, Ibarra A, Mallor F. Accumulating priority queues versus pure priority queues for managing patients in emergency departments. *Operations Research for Health Care*. 2019;23:100224. DOI: 10.1016/j.orhc.2019.100224.

15. Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, et al. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association*. 2005;100:36–50. DOI: 10.2307/27590517.

16. Lucantoni DM. New results on the single server queue with a batch Markovian arrival process. *Communications in Statistics. Stochastic Models.* 1991;7(1):1–46. DOI: 10.1080/15326349108807174.

17. Dudin AN, Klimenok VI, Vishnevsky VM. *The theory of queuing systems with correlated flows.* Berlin: Springer Nature; 2019. 410 p. DOI: 10.1007/978-3-030-32072-0.

18. Neuts MF. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Baltimore: The Johns Hopkins University Press; 1981. 348 p.

19. Graham A. *Kronecker products and matrix calculus with applications*. Cichester: Ellis Horwood; 1981. 130 p.

20. Klimenok VI, Kim CS, Orlovsky DS, Dudin AN. Lack of invariant property of Erlang loss model in case of *MAP* input. *Queueing Systems*. 2005;49:187–213. DOI: 10.1007/s11134-005-6481-z.

Получена 18.04.2022 / исправлена 05.05.2022 / принята 22.06.2022.
Received 18.04.2022 / revised 05.05.2022 / accepted 22.06.2022.