

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

MATHEMATICAL MODELING

Научная статья

УДК 001.891.573

doi: 10.17223/19988605/60/3

**Анализ системы обслуживания с коррелированными входными потоками
и изменяемыми приоритетами****Александр Николаевич Дудин¹, Сергей Александрович Дудин², Ольга Сергеевна Дудина³**^{1, 2, 3} *Белорусский государственный университет, Минск, Республика Беларусь*¹ *dudin@bsu.by*² *dudins@bsu.by*³ *dudina@bsu.by*

Аннотация. Приоритетные системы массового обслуживания могут использоваться для решения задач проектирования и оптимизации различных реальных систем, в частности телекоммуникационных систем, систем оказания скорой медицинской помощи. Различные схемы предоставления приоритетов в последние годы дополнились динамическими приоритетами. Такие схемы предполагают изменение приоритетов запросов во время ожидания начала обслуживания. В данной работе рассматривается приоритетная система массового обслуживания, на вход которой поступает марковский входной поток разнотипных запросов. Приоритеты запросов изменяются динамически. Предусмотрена зависимость поведения запросов, находящихся в буфере, а также интенсивности обслуживания от типа запроса.

Ключевые слова: приоритетная система массового обслуживания; динамические приоритеты; оптимизация; коррелированные потоки

Для цитирования: Дудин А.Н., Дудин С.А., Дудина О.С. Анализ системы обслуживания с коррелированными входными потоками и изменяемыми приоритетами // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2022. № 60. С. 21–31. doi: 10.17223/19988605/60/3

Original article

doi: 10.17223/19988605/60/3

Analysis of a queuing system with correlated arrival flows and variable priorities**Alexander N. Dudin¹, Sergei A. Dudin¹, Olga S. Dudina¹**^{1, 2, 3} *Belarusian State University, Minsk, Belarus*¹ *dudin@bsu.by*² *dudins@bsu.by*³ *dudina@bsu.by*

Abstract. Priority queuing systems can be used to solve the problems of designing and optimizing various real systems, in particular, telecommunications systems, emergency medical systems. Various priority schemes have been supplemented in recent years with dynamic priorities. Such schemes involve changing the priorities of requests while waiting for service to begin. In this paper, we consider a priority queuing system with a Markovian arrival process of

heterogeneous requests. Request priorities change dynamically. The dependence of the behavior of requests in the buffer, as well as the intensity of service on the type of request, is provided.

Keywords: priority queuing system; dynamic priorities; optimization; correlated arrival process

For citation: Dudin, A.N., Dudin, S.A., Dudina, O.S. (2022) Analysis of a queuing system with correlated arrival flows and variable priori. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 60. pp. 21–31. doi: 10.17223/19988605/60/3

Теория массового обслуживания представляет собой хороший математический аппарат для решения проблем оптимального распределения некоторого ограниченного ресурса между конкурирующими пользователями. В большинстве существующих работ предполагается, что пользователи являются однородными, в то время как во многих реальных системах имеются потоки пользователей, разнородных по требованиям к времени обслуживания и их важности для системы. Например, в моделях функционирования приемных покоев учреждений здравоохранения пациенты делятся на требующих неотложной помощи, скорой помощи и менее срочной помощи [1]. В сетях когнитивного радио имеются потоки лицензированных пользователей и нелицензированных (когнитивных) пользователей [2, 3]. Последние подразделяются на пользователей, требующих и не требующих обслуживания в режиме реального времени, вновь поступивших и вытесненных с обслуживания и т.д. В моделях функционирования интеллектуальных транспортных систем потоки информации, критичной для безопасности движения, заведомо более важны, чем потоки информационно-развлекательные. Работа таких систем может быть описана в терминах приоритетных систем массового обслуживания.

Наряду с рассмотрением различных разновидностей статических относительных и абсолютных приоритетов в последние годы появились исследования систем с динамически изменяющимися приоритетами, в которых во время ожидания запросов пользователей в очереди некоторым детерминированным или рандомизированным образом может происходить изменение изначально назначенного пользователю приоритета. Системам с динамическими изменениями приоритетов запросов посвящены, например, работы [4–11]. Стоит отметить, что в большинстве работ делается практически нереалистичное в современных системах, включая сети телекоммуникаций, предположение о том, что потоки запросов являются стационарными пуассоновскими, в то время как во многих реальных системах потоки не являются стационарными, а имеют высокие коэффициенты корреляции и вариации длин интервалов между моментами поступления.

В данной статье исследуется однолинейная система массового обслуживания с коррелированным входным потоком разнотипных запросов с приоритетами. При выборе следующего запроса на обслуживание приоритеты являются относительными. Прерывание текущего обслуживания в случае прихода более приоритетного запроса не допускается. Система имеет конечный буфер, общий для запросов всех типов. При поступлении новых запросов приоритет является рандомизированно абсолютным. В случае заполнения буфера поступающий запрос удаляет из него запрос более низкого приоритета, если таковой имеется, с фиксированной вероятностью, зависящей от типа поступающего запроса. Поступление запросов описывается ММАР-поток (маркированный марковский входной поток) [12], позволяющим учитывать характерные черты многих реальных потоков, в частности потоков информации в современных телекоммуникационных сетях. Важнейшей особенностью рассмотренной системы является возможность повышения приоритета произвольного запроса во время его пребывания в буфере. Учтена возможность ухода запросов из системы из-за нетерпеливости.

В работе [13] была исследована система массового обслуживания с динамически изменяемыми приоритетами и коррелированным входным потоком. Однако в [13] сделано предположение, что запросы всех типов имеют одинаковые требования к обслуживанию. Данное предположение существенно упрощает исследование системы, но в то же время значительно сужает область применения модели. В реальных системах запросы разных типов могут иметь совершенно разный объем данных или работы, что приводит к разному времени обслуживания. В данной статье предусмотрена зависимость интенсивности обслуживания от типа запроса.

1. Математическая модель

Рассматривается однолинейная система массового обслуживания с конечным буфером, структура которой представлена на рис. 1.

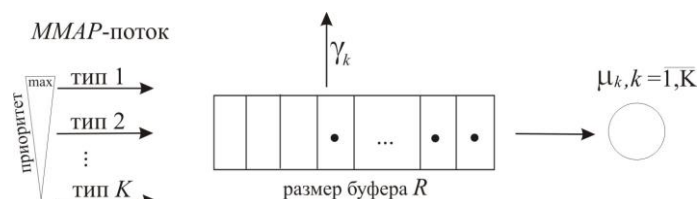


Рис. 1. Структура исследуемой системы
Fig. 1. Structure of the system under study

Входной поток запросов K типов задается ММАР-поток. Этот процесс является обобщением более известного марковского входного потока [14–18] на случай разнотипных запросов. Использование такой модели входного потока вместо широко популярной модели суперпозиции стационарных пуассоновских потоков, являющейся очень частным случаем ММАР-потока, позволяет учитывать флуктуацию мгновенной интенсивности потока, корреляцию длин последовательных интервалов между моментами поступления и их вариацию.

Запросы $k, k = \overline{1, K}$, типов поступают в ММАР-поток, который определяется неприводимой цепью Маркова $\gamma_t, t \geq 0$, с непрерывным временем и конечным пространством состояний $\{1, 2, \dots, W\}$, а также набором квадратных матриц $D_k, k = \overline{0, K}$. Обозначим через λ среднюю интенсивность поступления запросов, а через λ_k среднюю интенсивность поступления запросов типа $k, k = \overline{1, K}$. Формулы для нахождения данных величин, а также таких характеристик ММАР-потока, как коэффициенты корреляции и вариации, приведены в [18].

Если в момент поступления запроса любого типа в систему прибор свободен, то этот запрос сразу начинает обслуживание. Для хранения запросов различных типов, поступающих, когда прибор занят, в системе имеется буфер емкостью R . Если прибор занят, но буфер не заполнен, запрос любого типа помещается в буфер. Конкретного ограничения на число запросов каждого типа, находящихся в буфере одновременно, нет, но общее количество запросов, находящихся в буфере, не должно превышать его емкость R . Очевидно, что в системе одновременно может находиться не более $R + 1$ запросов (один запрос на обслуживании и R запросов в буфере).

Запросы разных типов имеют разные приоритеты. Считаем, что приоритет типов запросов уменьшается с увеличением номера типа. То есть запросы первого типа имеют наивысший приоритет среди всех типов запросов, а запросы типа K имеют самый низкий приоритет. Приоритет определяет порядок приема запроса в систему, когда буфер заполнен, и порядок выбора запроса из буфера на обслуживание. При выборе запроса из буфера на обслуживание приоритет предполагается *относительным*. Если в момент завершения обслуживания в буфере присутствуют запросы первого типа, то один из этих запросов начинает обслуживание. Запрос типа $k, k \geq 2$, может начать обслуживание, только если в буфере отсутствуют запросы типов $1, 2, \dots, k-1$. Обслуживание любого запроса не может быть прервано в случае поступления запроса, имеющего более высокий приоритет.

Предполагается также, что при принятии в систему запросы типа $k, k = \overline{1, K}$, имеют рандомизированно *абсолютный* приоритет над запросами типа $l, l = \overline{k+1, K}$. То есть если в момент прихода запроса типа k прибор занят, число запросов в буфере равно R и нет запросов типа $l, l = \overline{k+1, K}$, то поступивший запрос покидает систему (теряется). Если же в буфере есть запросы типа $l, l = \overline{k+1, K}$, то поступивший запрос с вероятностью q_k принимается в буфер, а один из запросов из буфера с самым низким приоритетом теряется. С дополнительной вероятностью $1 - q_k$ поступивший запрос

покидает систему, несмотря на наличие в буфере запросов с более низким приоритетом. Такая рандомизация имеет целью уменьшить потери низкоприоритетных запросов. Считаем, что во время пребывания в буфере запросы типа $k, k = \overline{2, K}$, могут увеличить свой приоритет. Полагаем, что после экспоненциально распределенного с параметром α_k времени запрос типа k становится запросом типа l с вероятностью $p_{k,l}, l = \overline{1, k-1}, \sum_{l=1}^{k-1} p_{k,l} = 1, k = \overline{2, K}$.

Запросы, находящиеся в буфере, могут проявлять нетерпеливость и покидать систему без обслуживания независимо от других запросов, если время ожидания слишком велико. Запрос типа k покидает систему без обслуживания по истечении экспоненциально распределенного с параметром γ_k времени ожидания, $\gamma_k \geq 0$. Предполагается, что если запрос меняет приоритет, то отсчет его времени «терпеливости» начинается с самого начала с параметром, соответствующим новому приоритету.

Время обслуживания запроса типа $k, k = \overline{1, K}$, прибором имеет экспоненциальное распределение с параметром $\mu_k, 0 < \mu_k < \infty$. Если в момент окончания обслуживания в буфере есть запросы, то обслуживание начинает первый запрос, имеющий наивысший приоритет. В противном случае прибор простаивает до следующего момента поступления запроса.

2. Процесс изменения состояний системы

Поведение рассматриваемой системы можно описать регулярной неприводимой цепью Маркова с непрерывным временем $\xi_t = \{i_t, m_t, v_t, n_t^{(1)}, \dots, n_t^{(K)}\}, t \geq 0$, где $i_t, i_t = \overline{0, R+1}$ – число запросов в системе; $m_t, m_t = \overline{1, K}$ – тип обслуживаемого запроса; $v_t, v_t = \overline{1, W}$ – состояние управляющего процесса ММАР-потока; $n_t^{(k)}$ – число запросов типа k в буфере, $n_t^{(k)} = \overline{0, \max\{0, i_t - 1\}}$, $\sum_{l=1}^K n_t^{(l)} = \max\{0, i_t - 1\}, k = \overline{1, K}$, в момент времени t .

Цепь Маркова $\xi_t, t \geq 0$, является неприводимой и имеет конечное пространство состояний. Следовательно, стационарные вероятности состояний системы существуют при любых значениях параметров системы:

$$\pi(i, m, v, n^{(1)}, \dots, n^{(K)}) = \lim_{t \rightarrow \infty} P\{i_t = i, m_t = m, v_t = v, n_t^{(1)} = n^{(1)}, \dots, n_t^{(K)} = n^{(K)}\}, i = \overline{0, R+1},$$

$$m = \overline{1, K}, v = \overline{1, W}, n^{(k)} = \overline{0, \max\{0, i - 1\}}, \sum_{l=1}^K n^{(l)} = \max\{0, i - 1\}, k = \overline{1, K}.$$

Из этих вероятностей, перенумерованных в обратном лексикографическом порядке компонент $n^{(1)}, \dots, n^{(K)}$, сформируем вектор-строки $\pi(i, m, v), v = \overline{1, W}$, и векторы π_i следующим образом:

$$\pi(i, m) = (\pi(i, m, 1), \pi(i, m, 2), \dots, \pi(i, m, W)), m = \overline{1, K}, \pi_i = (\pi(i, 1), \pi(i, 2), \dots, \pi(i, K)), i = \overline{2, R+1},$$

$$\pi_0 = (\pi(0, 0, 1, 0), \pi(0, 0, 2, 0), \dots, \pi(0, 0, W, 0)),$$

$$\pi(1, m) = (\pi(1, m, 1, 0), \pi(1, m, 2, 0), \dots, \pi(1, m, W, 0)), \pi_1 = (\pi(1, 1), \pi(1, 2), \dots, \pi(1, K)).$$

Векторы π_i удовлетворяют следующей системе линейных алгебраических уравнений:

$$(\pi_0, \pi_1, \dots, \pi_{R+1})Q = 0, (\pi_0, \pi_1, \dots, \pi_{R+1})e = 1,$$

где Q – инфинитезимальный генератор цепи Маркова $\xi_t, t \geq 0$, e – вектор-столбец, состоящий из единиц, 0 – вектор-строка, состоящая из нулей. Для решения данной системы может быть применен, например, алгоритм, описанный в [17].

Теорема 1. Инфинитезимальный генератор $Q = (Q_{i,j})_{i,j=\overline{0, R+1}}$ имеет блочно-трехдиагональную структуру, где ненулевые блоки $Q_{i,j}, |i-j| \leq 1$, определяются следующим образом:

$$\begin{aligned}
 Q_{0,0} &= D_0, \quad Q_{1,1} = I_K \otimes D_0 - C \otimes I_W, \quad Q_{i,i} = I_K \otimes D_0 \otimes I_{T_{i-1}} - C \otimes I_{WT_{i-1}} + I_{KW} \otimes (Y_{i-1} + \hat{I}_{i-1}), i = \overline{2, R}, \\
 Q_{R+1, R+1} &= I_K \otimes D_0 \otimes I_{T_R} - C \otimes I_{WT_R} + I_{KW} \otimes (Y_R + \hat{I}_R) + I_K \otimes \sum_{k=1}^K \left(D_k \otimes \left[(1 - q_k) I_{T_R} + q_k \hat{E}_k \right] \right), \\
 Q_{0,1} &= \sum_{k=1}^K (\mathbf{h}_k \otimes D_k), \quad Q_{i,i+1} = I_K \otimes \sum_{k=1}^K (D_k \otimes A_{i-1}(\mathbf{h}_k)), i = \overline{1, R}, \\
 Q_{1,0} &= C \mathbf{e} \otimes I_W, \quad Q_{i,i-1} = C \otimes I_W \otimes E_{i-1}^- + I_{KW} \otimes L_{i-1}(\gamma), i = \overline{2, R+1}.
 \end{aligned}$$

Здесь I_i – единичная матрица размера i ; $\otimes$ – символ Кронекера произведения матриц [18]; $L_i(\gamma), i = \overline{1, R}$, – матрица, элементы которой определяют интенсивности переходов, когда запрос покидает буфер из-за нетерпеливости; $Y_i = Y_i(H), i = \overline{1, R}$, – матрица, элементы которой определяют интенсивности переходов, происходящих, когда запрос повышает свой приоритет. Матрица H размера $K \times K$ определяет интенсивность возрастания приоритетов и имеет все нулевые элементы, кроме элементов $(H)_{i,j} = p_{i,j} \alpha_i, i = \overline{2, K}, j = \overline{1, i-1}$; $A_i(\mathbf{h}_k), i = \overline{0, R-1}$, – матрица, элементы которой определяют вероятности перехода в момент поступления в систему нового запроса типа k , когда в буфере находится $i, 0 \leq i < R$, запросов (в буфере есть место для поступившего запроса); вектор $\mathbf{h}_k = (\underbrace{0, \dots, 0}_{k-1}, \underbrace{1, 0, \dots, 0}_{K-k})$ задает вероятность того, что поступивший в систему запрос имеет тип $k, k = \overline{1, K}$. $\hat{E}_k, k = \overline{1, K}$, – квадратная матрица размера T_R , элементы которой определяют вероятности перехода в момент, когда запрос типа k поступает в систему, в буфере есть R запросов и поступивший запрос пытается вытеснить из буфера запрос с более низким приоритетом; $E_i^-, i = \overline{1, R}$, – матрица, элементы которой определяют вероятности перехода в момент выбора запроса с максимальным (из присутствующих в системе в данный момент) приоритетом для обслуживания; $\hat{I}_i = -\text{diag}\{Y_i \mathbf{e} + L_i \mathbf{e}\}, i = \overline{1, R}$; $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_K)$; $T_i = \frac{(i+K-1)!}{i!(K-1)!}, i = \overline{1, R}$; $C = \text{diag}\{\mu_1, \mu_2, \dots, \mu_K\}$. Формулы для вычисления матриц $Y_i(H), L_i(\gamma), E_i^-, i = \overline{1, R}, A_i(\mathbf{h}_k), i = \overline{0, R-1}, \hat{E}_k, k = \overline{1, K}$, приведены в [19].

Доказательство основано на анализе всех возможных переходов цепи Маркова за бесконечно малый интервал времени. Поступление и завершение обслуживания двух и более запросов за бесконечно малый интервал времени невозможны, поэтому блоки $Q_{i,j}, |i-j| > 1$, являются нулевыми матрицами.

Диагональные элементы матриц $Q_{i,i}, i = \overline{0, R+1}$, отрицательны, а модули этих элементов определяют суммарную интенсивность выхода цепи из соответствующего состояния. Если система пуста ($i=0$), то цепь Маркова $\xi_t, t \geq 0$, может выйти из своего состояния только тогда, когда: (а) управляющий процесс ММАР-потока совершает выход из текущего состояния; интенсивности таких выходов определяются модулями диагональных элементов матрицы D_0 . Если $i=1$, т.е. запрос находится на обслуживании, а буфер пуст, то цепь Маркова $\xi_t, t \geq 0$, может выйти из своего состояния только в случае (а), описанном выше; (б) в случае окончания обслуживания. Интенсивности таких выходов определяются модулем диагональных элементов матрицы $I_K \otimes D_0$ в случае (а) и матрицы $C \otimes I_W$ в случае (б). Если $i = \overline{2, R}$, прибор занят и в буфере есть хотя бы одно свободное место, то цепь Маркова $\xi_t, t \geq 0$, может выйти из своего состояния по тем же причинам, что и в предыдущем случае (интенсивности таких переходов определяются модулями диагональных элементов матриц $I_K \otimes D_0 \otimes I_{T_{i-1}}$ для случая (а) и матриц $C \otimes I_{WT_{i-1}}$ для случая (б)). Также возможно, что (в) какой-то запрос из буфера может изменить свой приоритет или покинуть буфер из-за нетерпеливости, интенсивности таких переходов задаются модулями диагональных элементов матриц $I_{KW} \otimes \hat{I}_{i-1}$. Если

$i = R + 1$, то прибор занят и буфер заполнен полностью. Значит, цепь Маркова ξ_t , $t \geq 0$, может покинуть свое состояние по причинам (а), (б) и (в) (интенсивности таких переходов определяются модулями диагональных элементов матриц $I_K \otimes D_0 \otimes I_{T_R}$, $C \otimes I_{WT_R}$ и $I_{KW} \otimes \hat{I}_R$ соответственно). Также возможны ситуации (г), когда приходит запрос типа k и уходит из системы, несмотря на возможное наличие в буфере запросов с более низким приоритетом, или (д) – этот запрос пытается попасть в буфер и вытеснить запрос с самым низким приоритетом. Интенсивности таких выходов определяются диагональными элементами матрицы $\sum_{k=1}^K (1 - q_k)(I_K \otimes D_k \otimes I_{T_R})$ в случае (г) и матрицы $\sum_{k=1}^K q_k (I_K \otimes D_k \otimes \hat{E}_k)$ в случае (д).

Недиагональные элементы матриц $Q_{i,j}$, $i = \overline{0, R+1}$, неотрицательны и определяют интенсивности переходов цепи Маркова ξ_t , $t \geq 0$, не приводящие к изменению числа запросов i в системе. Такими переходами являются: переходы управляющего процесса ММАР-потока без генерации запроса (интенсивности определяются недиагональными элементами матрицы D_0 , если $i = 0$, матрицы $I_K \otimes D_0$ при $i = 1$, матрицы $I_K \otimes D_0 \otimes I_{T_{i-1}}$ при $i = \overline{2, R+1}$); переходы из-за смены приоритета некоторым запросом в буфере (интенсивности определяются элементами матрицы $I_{KW} \otimes Y_{i-1}$, $i = \overline{2, R+1}$); переходы, когда запрос при поступлении застаёт заполненный буфер и покидает систему или вытесняет запрос с наименьшим приоритетом из буфера (интенсивности определяются недиагональными элементами матрицы $\sum_{k=1}^K (I_K \otimes D_k) \otimes [(1 - q_k)I_{T_R} + q_k \hat{E}_k]$, $i = R + 1$). Таким образом, после некоторых алгебраических преобразований с учетом приведенных выше объяснений мы получаем вид диагональных блоков $Q_{i,j}$, $i = \overline{0, R+1}$, генератора Q .

Элементы матриц $Q_{i,i+1}$, $i = \overline{0, R}$, неотрицательны и определяют интенсивности переходов цепи Маркова ξ_t , $t \geq 0$, приводящих к увеличению количества запросов в системе на единицу. Такие переходы происходят, когда в систему поступает запрос любого типа. Интенсивности этих переходов определяются матрицей $\sum_{k=1}^K (\mathbf{h}_k \otimes D_k)$, если система пуста, и матрицей $\sum_{k=1}^K I_K \otimes D_k \otimes A_{i-1}(\mathbf{h}_k)$, если в системе находится i , $i = \overline{1, R}$, запросов.

Элементы матриц $Q_{i,i-1}$, $i = \overline{1, R+1}$, неотрицательны и определяют интенсивности переходов цепи Маркова ξ_t , $t \geq 0$, приводящих к уменьшению количества запросов в системе на единицу. Такое уменьшение может произойти из-за (а) успешного обслуживания запроса или (б) ухода запроса из буфера из-за нетерпеливости. В случае (а) интенсивности переходов задаются матрицей $C \otimes I_W$, если буфер пуст, и матрицей $C \otimes I_W \otimes E_{i-1}^-$, иначе. В случае (б) интенсивности переходов определены только для случая $i = \overline{2, R+1}$, и задаются элементами матриц $I_{KW} \otimes L_{i-1}(\gamma)$. Теорема доказана.

3. Характеристики производительности

Среднее число запросов в системе вычисляется по формуле $L = \sum_{i=1}^{R+1} i \pi_i \mathbf{e}$. Среднее число запросов в буфере вычисляется по формуле $N_{buf} = \sum_{i=2}^{R+1} (i-1) \pi_i \mathbf{e}$. Среднее число запросов типа k в буфере вычисляется как $N_{buf}^{(k)} = \sum_{i=2}^{R+1} \pi_i (I_{KW} \otimes L_{i-1}(\mathbf{h}_k)) \mathbf{e}$, $k = \overline{1, K}$. Средняя интенсивность выходного потока успешно обслуженных запросов определяется как $\lambda_{serv} = \sum_{i=1}^{R+1} \sum_{m=1}^K \mu_m \pi(i, m) \mathbf{e}$. Средняя интенсивность выходного потока запросов, которые покинули буфер из-за нетерпеливости, вычисляется по формуле

$\lambda_{leave} = \sum_{i=2}^{R+1} \pi_i (I_{KW} \otimes L_{i-1}(\gamma)) \mathbf{e}$. Вероятность потери произвольного запроса находится как $P_{loss} = 1 - \frac{\lambda_{serv}}{\lambda}$. Вероятность потери произвольного запроса из-за нетерпеливости находится как $P_{leave-loss} = \frac{\lambda_{leave}}{\lambda}$. Интенсивность выходного потока запросов типа k , которые покинули буфер из-за нетерпеливости, определяется как $\lambda_{leave}^{(k)} = \sum_{i=2}^{R+1} \pi_i (I_{KW} \otimes L_{i-1}(\gamma_k)) \mathbf{e} = \gamma_k N_{buf}^{(k)}$, $k = \overline{1, K}$, где γ_k – вектор-строка размера K со всеми нулевыми элементами, кроме k -го, который равен γ_k . Средняя интенсивность переходов запросов типа l , $l = \overline{k+1, K}$, в тип k , $k = \overline{1, K-1}$, вычисляется как $\lambda_{inc}^{(k)} = \sum_{l=k+1}^K \alpha_l N_{buf}^{(l)} p_{l,k}$. Вероятность потери произвольного запроса типа k из-за нетерпеливости вычисляется по формуле $P_{leave-loss}^{(k)} = \frac{\lambda_{leave}^{(k)}}{\lambda_k + \lambda_{inc}^{(k)}}$, $k = \overline{1, K}$. Здесь предполагается, что $\lambda_{inc}^{(K)} = 0$. Вероятность потери произвольного запроса типа k по прибытии без попытки вытеснить запрос с более низким приоритетом равна $P_{loss-no-push}^{(k)} = (1 - q_k) \lambda_k^{-1} \pi_{R+1} (I_K \otimes D_k \otimes I_{T_R}) \mathbf{e}$, $k = \overline{1, K}$. Вероятность потери произвольного запроса типа k по прибытии, несмотря на неудачную попытку вытеснить запрос с более низким приоритетом, равна $P_{loss-push}^{(k)} = q_k \lambda_k^{-1} \pi_{R+1} (I_K \otimes D_k \otimes \tilde{E}_k) \mathbf{e}$, $k = \overline{1, K}$, где матрица $\tilde{E}_k$ имеет все нулевые элементы, кроме диагональных элементов, которые равны диагональным элементам матрицы $\hat{E}_k$. Вероятность потери произвольного запроса по прибытии равна $P_{arr-loss} = \lambda^{-1} \sum_{k=1}^K \lambda_k (P_{loss-no-push}^{(k)} + P_{loss-push}^{(k)})$. Вероятность потери произвольного запроса типа k по прибытии равна $P_{arr-loss}^{(k)} = P_{loss-no-push}^{(k)} + P_{loss-push}^{(k)}$, $k = \overline{1, K}$. Вероятность того, что запрос произвольного типа k по прибытии встретит полный буфер и вытеснит запрос с более низким приоритетом, равна $P_{push-out}^{(k)} = q_k \lambda_k^{-1} \pi_{R+1} (I_K \otimes D_k \otimes (\hat{E}_k - \tilde{E}_k)) \mathbf{e}$, $k = \overline{1, K}$. Вероятность того, что произвольный запрос по прибытии встретит полный буфер и вытеснит запрос с более низким приоритетом, равна $P_{push-out} = \lambda^{-1} \sum_{k=1}^K q_k \pi_{R+1} (I_K \otimes D_k \otimes (\hat{E}_k - \tilde{E}_k)) \mathbf{e}$.

4. Численный эксперимент

Исследуем влияние коэффициента корреляции во входном потоке на основные характеристики производительности исследуемой системы. Рассмотрим четыре ММАР-потока с тремя типами запросов $K=3$, которые определяются матрицами D_k , $k = \overline{0, 3}$, имеют одинаковую среднюю интенсивность поступления запросов $\lambda = 20$ и одинаковые интенсивности поступления запросов каждого типа $\lambda_1 = 4$, $\lambda_2 = 6$, $\lambda_3 = 10$, но различные коэффициенты корреляции.

Первый ММАР имеет $W=1$, коэффициент корреляции $c_{cor} = 0$, коэффициент вариации $c_{var} = 1$ и определяется матрицами $D_0 = -20$, $D_1 = 4$, $D_2 = 6$, $D_3 = 10$.

Второй ММАР имеет коэффициент корреляции $c_{cor} = 0,1$ и характеризуется матрицами

$$D_0 = \begin{pmatrix} -53,3383 & 2,3544 & 2,4692 \\ 2,7708 & -9,78638 & 1,6916 \\ 2,7292 & 1,6576 & -6,54159 \end{pmatrix}, D_1 = \begin{pmatrix} 9,2375 & 0,29047 & 0,17496 \\ 0,307359 & 0,69271 & 0,06471 \\ 0,22816 & 0,03399 & 0,1688 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 13,8563 & 0,435719 & 0,26244 \\ 0,461039 & 1,03908 & 0,09707 \\ 0,342239 & 0,05099 & 0,2532 \end{pmatrix}, D_3 = \begin{pmatrix} 23,0938 & 0,72619 & 0,43739 \\ 0,768399 & 1,7318 & 0,1618 \\ 0,57039 & 0,08499 & 0,42199 \end{pmatrix}.$$

Третий ММАР имеет коэффициент корреляции $c_{cor} = 0,2$ и характеризуется матрицами

$$D_0 = \begin{pmatrix} -62,9307 & 2,4248 & 2,3696 \\ 2,0712 & -9,15879 & 1,8716 \\ 2,3884 & 2,2612 & -7,83879 \end{pmatrix}, D_1 = \begin{pmatrix} 11,3201 & 0,24168 & 0,06543 \\ 0,08567 & 0,825599 & 0,13192 \\ 0,06863 & 0,15832 & 0,41087 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 16,9802 & 0,36251 & 0,09815 \\ 0,12852 & 1,2384 & 0,19788 \\ 0,10296 & 0,23748 & 0,61631 \end{pmatrix}, D_3 = \begin{pmatrix} 28,3004 & 0,60419 & 0,1636 \\ 0,2142 & 2,064 & 0,32979 \\ 0,1716 & 0,39579 & 1,0272 \end{pmatrix}.$$

Четвертый ММАР имеет коэффициент корреляции $c_{cor} = 0,3$ и характеризуется матрицами

$$D_0 = \begin{pmatrix} -102,15 & 1,57306 & 1,44467 \\ 0,58074 & -8,928 & 0,79992 \\ 1,18389 & 1,54946 & -7,0097 \end{pmatrix}, D_1 = \begin{pmatrix} 19,3919 & 0,37348 & 0,06103 \\ 0,02727 & 1,33332 & 0,14886 \\ 0,00719 & 0,20438 & 0,64370 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 29,0879 & 0,56023 & 0,09157 \\ 0,040916 & 1,99998 & 0,2233 \\ 0,010799 & 0,306572 & 0,96553 \end{pmatrix}, D_3 = \begin{pmatrix} 48,4798 & 0,93376 & 0,15256 \\ 0,06819 & 3,3333 & 0,37216 \\ 0,01784 & 0,51094 & 1,60925 \end{pmatrix}.$$

Три последних входных потока имеют коэффициент вариации $c_{var} = 4$.

Считаем, что $\mu_1 = 100, \mu_2 = 70, \mu_3 = 60, \gamma_k = 0,1, k = \overline{1,3}, \alpha_2 = \alpha_3 = 0,2, q_1 = q_2 = 0,5, p_{2,1} = 1, p_{3,1} = p_{3,2} = 0,5$.

Будем изменять размерность буфера R на отрезке $[1, 50]$. На рис. 2–4 приведены зависимости основных характеристик производительности системы для четырех ММАР-потоков с различными коэффициентами корреляции, приведенных выше.

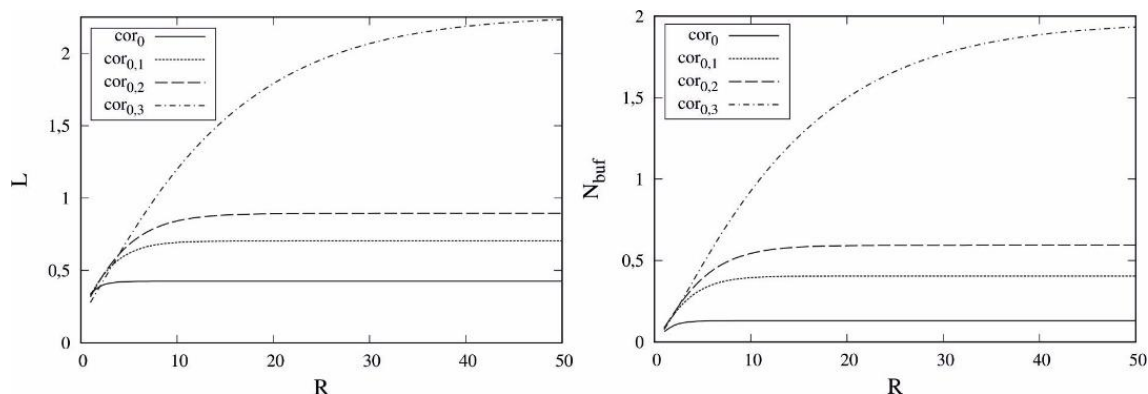


Рис. 2. Среднее число запросов L в системе и N_{buf} в буфере при различных значениях размерности буфера R
Fig. 2. The average number of requests L in the system and N_{buf} in the buffer for different values of the buffer size R

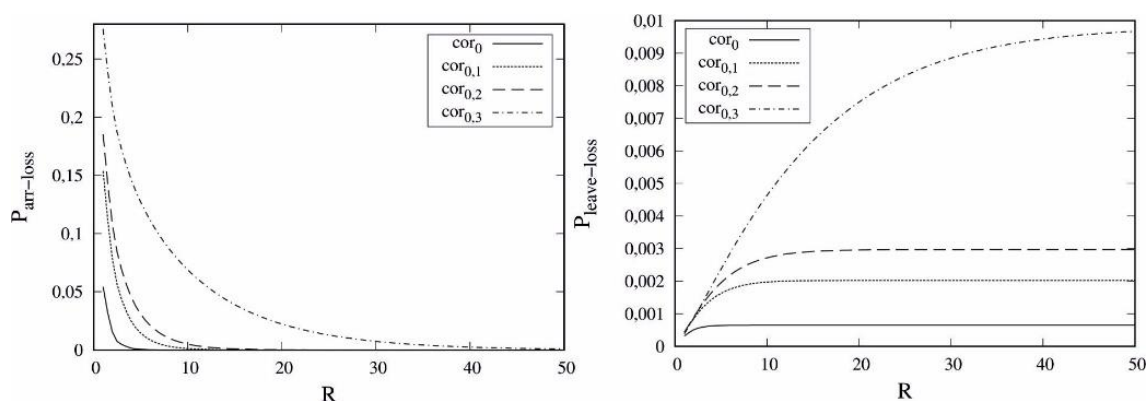


Рис. 3. Вероятности потери $P_{arr-loss}$ и $P_{leave-loss}$ при различных значениях размерности буфера R
Fig. 3. The loss probabilities $P_{arr-loss}$ and $P_{leave-loss}$ for different values of the buffer size R

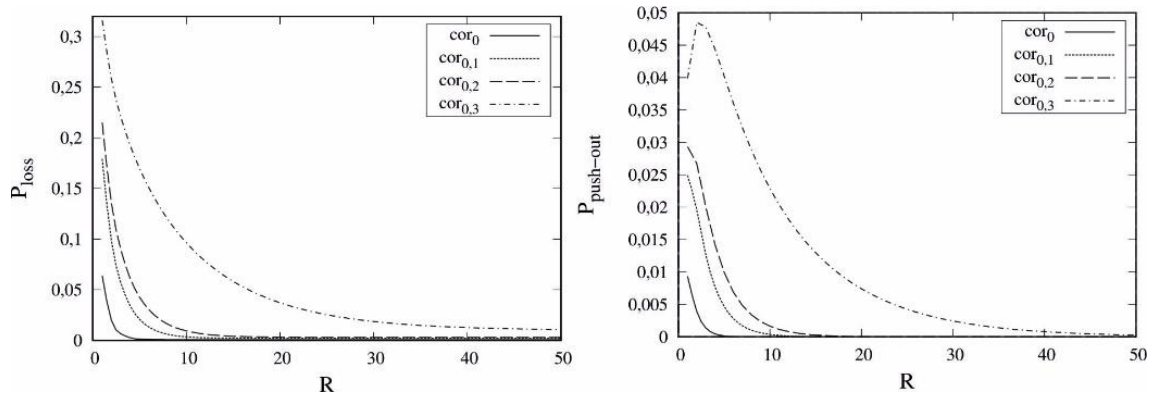


Рис. 4. Вероятности потери P_{loss} и $P_{push-out}$ при различных значениях размерности буфера R

Fig. 4. The loss probabilities P_{loss} and $P_{push-out}$ for different values of the buffer size R

На основании приведенных графиков можно сделать вывод, что увеличение емкости буфера приводит к увеличению среднего числа запросов как в самом буфере N_{buf} , так и системе в целом L . Это происходит из-за того, что при увеличении емкости буфера очевидно уменьшается число запросов, потерянных на входе в систему из-за заполненности буфера, что подтверждается графиком для вероятности $P_{arr-loss}$. Достаточно предсказуемо, что чем больше запросов в буфере, тем большее их число уходит из системы из-за нетерпеливости, так и не дождавшись обслуживания, поэтому вероятность $P_{leave-loss}$ увеличивается с ростом емкости буфера R . Поскольку интенсивность λ_{leave} выходного потока запросов, которые покинули буфер из-за нетерпеливости, можно выразить через вероятность $P_{leave-loss}$, то поведение этих характеристик совпадает. Очевидно, что с ростом R вероятность $P_{push-out}$ должна уменьшаться, так как уменьшается вероятность прихода запроса в полную систему. Однако, как видно из рис. 4, поведение вероятности $P_{push-out}$ для случая $c_{cor} = 0,3$ не является монотонным. Это объясняется тем, что при малой емкости буфера при ее увеличении растет число запросов с низким приоритетом в буфере, принимаемых в систему, поэтому возрастает вероятность того, что произвольный запрос на входе встретит полный буфер и вытеснит запрос с более низким приоритетом. При малых R , например $R = 1$, велика вероятность, что все места в буфере занимают запросы с наивысшим приоритетом, которые не могут быть выбиты.

Можно также наблюдать, что при фиксированном значении емкости буфера R рост коэффициента корреляции во входном потоке негативно влияет на процесс функционирования системы и приводит к ухудшению качества обслуживания в ней. Отметим, что для трех ММАР-потоков с коэффициентами корреляции 0, 0,1 и 0,2 значения характеристик L , N_{buf} , P_{loss} и $P_{leave-loss}$ стабилизируются, а значения вероятностей потерь $P_{arr-loss}$ и $P_{push-out}$ приближаются к нулю при значениях R , больших чем 8, 14 и 21 соответственно. При этом дальнейшее увеличение размера буфера не приводит к улучшению качества обслуживания, поскольку для приведенных значений R емкость буфера достаточна, чтобы запросы практически не терялись на входе ($P_{arr-loss}$ и $P_{push-out}$ принимают значения, близкие к нулю), а потеря запросов из системы возможна только из-за их нетерпеливости (P_{loss} практически совпадает с $P_{leave-loss}$). Дальнейшее улучшение качества обслуживания возможно теперь только за счет увеличения интенсивности обслуживания. Отметим также, что для ММАР-потока с коэффициентом корреляции 0,3 требуется гораздо большая емкость буфера R , чтобы минимизировать вероятности потерь $P_{arr-loss}$ и $P_{push-out}$. Таким образом, проведенные численные эксперименты доказывают важность учета коэффициента корреляции во входном потоке при оценке качества функционирования системы.

Заключение

Исследована приоритетная система массового обслуживания с маркированным марковским входным потоком запросов. Запросы различаются интенсивностью обслуживания и их важностью для системы. Приоритет запроса меняется динамически. Предусмотрен уход запросов из буфера из-за нетерпеливости. Получено компактное описание генератора цепи Маркова, задающей динамику ис-

следуемой системы, и применены численно устойчивые методы для решения системы уравнений равновесия для векторов стационарных вероятностей состояний цепи. Получены аналитические зависимости характеристик производительности от этих векторов. Численно показана зависимость данных характеристик от емкости буфера. Проиллюстрирована важность учета коэффициента корреляции во входном потоке при оценке производительности системы.

Список источников

1. Elalouf A., Wachtel G. Queueing Problems in Emergency Departments: A Review of Practical Approaches and Research Methodologies // *Operations Research Forum*. 2022. V. 3, № 1. P. 1–46.
2. Maharaj B.T., Awoyemi B.S. *Developments in Cognitive Radio Networks: Future Directions for Beyond 5G*. Springer Nature, 2021.
3. Goel S., Kulshrestha R. Queueing based spectrum management in cognitive radio networks with retrial and heterogeneous service classes // *Journal of Ambient Intelligence and Humanized Computing*. 2022. V. 13. P. 2429–2437.
4. Cao P., Xie J. Optimal control of a multiclass queueing system when customers can change types // *Queueing Systems*. 2016. V. 82, № 3. P. 285–313.
5. Xie J. et al. Determining the conditions for reverse triage in emergency medical services using queueing theory // *International Journal of Production Research*. 2016. V. 54, № 11. P. 3347–3364.
6. Fajardo V.A., Drekić S. Waiting time distributions in the preemptive accumulating priority queue // *Methodology and Computing in Applied Probability*. 2017. V. 19, № 1. P. 255–284.
7. Mojalal M., Stanford D.A., Caron R.J. The lower-class waiting time distribution in the delayed accumulating priority queue // *INFOR: Information Systems and Operational Research*. 2020. V. 58, № 1. P. 60–86.
8. Stanford D.A., Taylor P., Ziedins I. Waiting time distributions in the accumulating priority queue // *Queueing Systems*. 2014. V. 77, № 3. P. 297–330.
9. Xie J. et al. Performance analysis of service systems with priority upgrades // *Annals of Operations Research*. 2017. V. 253, № 1. P. 683–705.
10. 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. V. 23. Art. 100224.
11. Dudin A. et al. Analysis of single-server multi-class queue with unreliable service, batch correlated arrivals, customers impatience, and dynamical change of priorities // *Mathematics*. 2021. V. 9, № 11. Art. 1257.
12. He Q.M. Queues with marked customers // *Advances in Applied Probability*. 1996. V. 28, № 2. P. 567–587.
13. Lee S. et al. A Priority Queue with Many Customer Types, Correlated Arrivals and Changing Priorities // *Mathematics*. 2020. V. 8, № 8. Art. 1292.
14. Chakravarthy S.R. et al. The batch Markovian arrival process: a review and future work // *Advances in Probability Theory and Stochastic Processes*. 2001. V. 1. P. 21–49.
15. Lucantoni D.M. New results on the single server queue with a batch Markovian arrival process // *Communications in Statistics. Stochastic Models*. 1991. V. 7, № 1. P. 1–46.
16. Dudin A., Klimenok V.I., Vishnevsky V.M. *The theory of queueing systems with correlated flows*. Cham : Springer, 2020.
17. Dudin S.A., Dudina O.S. Call center operation model as a MAP/PH/N/R–N system with impatient customers // *Problems of Information Transmission*. 2011. V. 47, № 4. P. 364–377.
18. Graham A. *Kronecker products and matrix differentiation with applications*. Chichester : Ellis Horwood, 1981.

References

1. Elalouf, A. & Wachtel G. (2022) Queueing Problems in Emergency Departments: A Review of Practical Approaches and Research Methodologies. *Operations Research Forum*. Springer International Publishing. 3(1). pp. 1–46. DOI: 10.1007/s43069-021-00114-8
2. Maharaj, B.T., Awoyemi, B.S. (2021) *Developments in Cognitive Radio Networks: Future Directions for Beyond 5G*. Springer Nature.
3. Goel, S. & Kulshrestha, R. (2022) Queueing based spectrum management in cognitive radio networks with retrial and heterogeneous service classes. *Journal of Ambient Intelligence and Humanized Computing*. 2429–2437. DOI: 10.1007/s12652-021-03442-z
4. Cao, P. & Xie, J. (2016) Optimal control of a multiclass queueing system when customers can change types. *Queueing Systems*. 82(3). pp. 285–313. DOI: 10.1007/s11134-015-9466-6
5. Xie, J. et al. (2016) Determining the conditions for reverse triage in emergency medical services using queueing theory. *International Journal of Production Research*. 54(11). pp. 3347–3364. DOI: 10.1080/00207543.2015.1109718
6. Fajardo, V.A. & Drekić, S. (2017) Waiting time distributions in the preemptive accumulating priority queue. *Methodology and Computing in Applied Probability*. 19(1). pp. 255–284. DOI: 10.1007/s11009-015-9476-1
7. Mojalal, M., Stanford, D.A. & Caron, R.J. (2020) The lower-class waiting time distribution in the delayed accumulating priority queue. *INFOR: Information Systems and Operational Research*. 58(1). pp. 60–86. DOI: 10.1080/03155986.2019.1624473

8. Stanford, D.A., Taylor, P. & Ziedins, I. (2014) Waiting time distributions in the accumulating priority queue. *Queueing Systems*. 77(3). pp. 297–330. DOI: 10.1007/s11134-013-9382-6
9. Xie, J. et al. (2017) Performance analysis of service systems with priority upgrades. *Annals of Operations Research*. 253(1). pp. 683–705. DOI: 10.1007/s10479-016-2370-6
10. Cildoz, M., Ibarra, A. & Mallor, F. (2019) Accumulating priority queues versus pure priority queues for managing patients in emergency departments. *Operations Research for Health Care*. 23. 100224. DOI: 10.1016/j.orhc.2019.100224
11. Dudin, A. et al. (2021) Analysis of single-server multi-class queue with unreliable service, batch correlated arrivals, customers impatience, and dynamical change of priorities. *Mathematics*. 9(11). pp. 1257. DOI: 10.3390/math9111257
12. He, Q.M. (1996) Queues with marked customers. *Advances in Applied Probability*. 28(2). pp. 567–587.
13. Lee, S. et al. (2020) A Priority Queue with Many Customer Types, Correlated Arrivals and Changing Priorities. *Mathematics*. 8(8). pp. 1292. DOI: 10.3390/math8081292
14. Chakravathy, S.R. et al. (2001) The batch Markovian arrival process: a review and future work. *Advances in Probability Theory and Stochastic Processes*. 1. pp. 21–49.
15. Lucantoni, D.M. (1991) New results on the single server queue with a batch Markovian arrival process. *Communications in Statistics. Stochastic Models*. 7(1). pp. 1–46. DOI: 10.1080/15326349108807174
16. Dudin, A., Klimenok, V.I. & Vishnevsky, V.M. (2020) *The theory of queuing systems with correlated flows*. Cham: Springer. pp. 63–392.
17. Dudin, S.A. & Dudina, O.S. (2011) Call center operation model as a MAP/PH/N/R– N system with impatient customers. *Problems of Information Transmission*. 47(4). pp. 364–377. DOI: 10.1134/S0032946011040053
18. Graham, A. (1981) *Kronecker products and matrix differentiation with applications*. Chichester: Ellis Horwood.

Информация об авторах:

Дудин Александр Николаевич – профессор, доктор физико-математических наук, заведующий НИЛ прикладного вероятностного анализа Белорусского государственного университета (Минск, Республика Беларусь); директор Центра прикладного вероятностного анализа Института прикладной математики и телекоммуникаций Российского университета дружбы народов (Москва, Россия). E-mail: dudin@bsu.by

Дудин Сергей Александрович – кандидат физико-математических наук, ведущий научный сотрудник НИЛ прикладного вероятностного анализа Белорусского государственного университета (Минск, Республика Беларусь). E-mail: dudins@bsu.by

Дудина Ольга Сергеевна – кандидат физико-математических наук, ведущий научный сотрудник НИЛ прикладного вероятностного анализа Белорусского государственного университета (Минск, Республика Беларусь). E-mail: dudina@bsu.by

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:

Dudin Alexander N. (Doctor of Physical and Mathematical Sciences, Professor, Belarusian State University, Minsk, Belarus). E-mail: dudin@bsu.by

Dudin Sergei A. (Candidate of Physical and Mathematical Sciences, Belarusian State University, Minsk, Belarus). E-mail: dudins@bsu.by

Dudina Olga S. (Candidate of Physical and Mathematical Sciences, Belarusian State University, Minsk, Belarus). E-mail: dudina@bsu.by

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Поступила в редакцию 21.04.2022; принята к публикации 30.08.2022

Received 21.04.2022; accepted for publication 30.08.2022