УДК 519.872

АНАЛИТИЧЕСКОЕ МОДЕЛИРОВАНИЕ СИСТЕМ С ЭЛЕКТРОННОЙ ОЧЕРЕДЬЮ

О. С. ДУДИНА[1)]

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

Аннотация. Рассматривается система массового обслуживания типа *MAP*/*GPH*/*N*/*K* как модель системы с электронной очередью. Предполагается, что прибывающие пользователи после получения талона на обслуживание (номера в очереди) могут покидать систему с вероятностью, зависящей от количества пользователей перед ними, если сочтут очередь слишком длинной. Кроме того, пользователи могут покидать систему во время ожидания из-за нетерпеливости. Система не знает о присутствии (отсутствии) вызываемого пользователя и тратит некоторое время на обслуживание, даже если соответствующий пользователь уже покинул систему. Вычисляется стационарное распределение рассматриваемой системы. Приводятся формулы для нахождения основных характеристик производительности системы, а также численный эксперимент, показывающий возможность использования результатов работы в оптимизационных целях.

Ключевые слова: электронная очередь; коррелированный входной поток; нетерпеливые запросы; обобщенное распределение фазового типа.

ANALYTICAL MODELLING OF SYSTEMS WITH A TICKET QUEUE

O. S. DUDINA[a]

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

Abstract. A queuing system of *MAP*/*GPH*/*N*/*K* type as a model of a ticket queue is herein considered. It is assumed that arriving users, after receiving a service ticket (place in the queue), can leave the system with a probability based on the number of users in front of them if they find the queue too long. In addition, users may leave the system during waiting due to impatience. The system does not know about the presence (absence) of the called users for service and spends some time servicing them, even if the corresponding user has already left the system. The stationary distribution of the system under consideration is calculated. Formulas for finding the main characteristics of the system performance are given. The presented numerical experiment shows the possibility of using the results for optimisation purposes.

Keywords: ticket queue; correlated arrival process; impatience customers; generalised phase-type distribution.

Образец цитирования:
Дудина ОС. Аналитическое моделирование систем с электронной очередью. *Журнал Белорусского государственного университета. Математика. Информатика.* 2024;2:40–53.
EDN: NSXUQQ

For citation:
Dudina OS. Analytical modelling of systems with a ticket queue. *Journal of the Belarusian State University. Mathematics and Informatics.* 2024;2:40–53. Russian.
EDN: NSXUQQ

Автор:
Ольга Сергеевна Дудина – кандидат физико-математических наук; ведущий научный сотрудник научно-исследовательской лаборатории прикладного вероятностного анализа кафедры теории вероятностей и математической статистики факультета прикладной математики и информатики.

Author:
Olga S. Dudina, PhD (physics and mathematics); leading researcher at the laboratory of applied probabilistic analysis, department of probability theory and mathematical statistics, faculty of applied mathematics and computer science.
dudina@bsu.by
https://orcid.org/0000-0002-6788-8783

Введение

Теория массового обслуживания широко используется для оценки производительности и оптимизации различных промышленных, логистических, телекоммуникационных систем и сетей связи. Классические модели массового обслуживания предполагают, что каждый входящий пользователь принимается в систему при наличии хотя бы одного свободного места в очереди и обслуживается в определенном порядке. Существуют также системы массового обслуживания (СМО) с так называемой видимой очередью, в которых приходящий пользователь наблюдает длину очереди и принимает решение присоединиться к очереди или уйти, даже если буфер не полон (см., например, [1]).

В классических СМО предполагается, что принятые пользователи всегда ждут своей очереди и обязательно будут обслужены. Однако в реальных системах пользователи часто могут проявлять нетерпеливость и покидать систему после некоторого времени ожидания, если их обслуживание не началось. Литература, посвященная исследованию таких систем, достаточно обширна (см., например, [2–4]).

Еще один вид практически важных и интересных СМО – это так называемые СМО с талонами (*ticket systems*), которые могут использоваться для моделирования электронной очереди. В данных СМО каждый пришедший пользователь получает номерной билет (талон, жетон и т. п.) и наблюдает за номером обслуживаемого пользователя, который транслируется на табло. Когда обслуживание пользователя с отображаемым номером завершается, система (оператор) вызывает пользователя со следующим номером, т. е. обслуживание осуществляется по принципу «первым пришел – первым ушел». В СМО с талонами пользователь может видеть только свой номер талона и номер обслуживаемого пользователя. На основании разницы между ними пользователь решает уйти или дождаться обслуживания. В отличие от обычных СМО в СМО с талонами отказывающийся от ожидания пользователь покидает систему физически, но система (оператор) не имеет никакой информации об этом. Далее поступающих пользователей будем называть активными пользователями, а пользователей, покинувших систему, не дождавшись обслуживания, – неактивными пользователями. Во время ожидания активный пользователь также может проявить нетерпеливость и уйти из системы. Таким образом, он становится неактивным пользователем, но его талон остается в очереди.

СМО с талонами рассматриваются в литературе из-за их высокой практической значимости и множества преимуществ (см., например, [5]). Электронные очереди получили широкое распространение в финансовых и государственных учреждениях, организациях здравоохранения и розничных магазинах. Более подробную информацию и конкретные примеры можно найти в статьях [5; 6]. В них представлены хорошие обзоры соответствующих исследований и проанализирована СМО с талонами типа $M/M/1$.

В реальных системах уход пользователей из-за отказа ждать и (или) нетерпеливости негативно влияет на доход, получаемый системой. Кроме того, система может понести репутационные потери из-за недовольства пользователей, покинувших ее, не получив обслуживание. По этой причине системные администраторы должны постараться свести к минимуму негативные последствия, связанные с таким поведением пользователей. Наличие видимой очереди, нетерпеливости и необходимости обслуживать запросы ушедших пользователей создает больше проблем для управления системой, чем в классических СМО, и требует более тщательного анализа.

В статье [5] предполагается, что пришедший пользователь отказывается ждать, если разница между его номером и номером обслуживаемого пользователя превышает заранее определенный порог. Уход пользователей из-за нетерпеливости не допускается. Авторы исследуют цепь Маркова (ЦМ), описывающую процесс функционирования СМО с талонами, и разрабатывают эффективные инструменты для приближенной оценки производительности системы.

Похожая модель описана и проанализирована в работе [7], но авторы дополнительно позволяют пользователям уходить из очереди. В статье содержатся аппроксимации показателей производительности системы, основанные на интенсивном трафике, и их сравнение с аналогичными показателями соответствующей классической СМО.

В работе [8] анализ, представленный в статье [7], расширен для использования в управленческих целях. Предполагается, что существует два уровня работы прибора, отличающихся интенсивностью обслуживания и вероятностью отказа в зависимости от текущего уровня.

В публикации [9] рассмотрена СМО с талонами типа $MAP/M/1/K$. В отличие от ранее исследованных моделей в данной системе учитывается время обслуживания неактивных пользователей, предполагается, что любой пользователь может покинуть систему с вероятностью, произвольно зависящей от разницы между собственным номером пользователя и отображаемым на табло номером, а также рассматривается более общая модель входного потока (MAP-поток) вместо стационарного пуассоновского потока. В настоящей статье модель из работы [9] существенно обобщена следующим образом. Предполагается, что

система является многолинейной, а времена обслуживания активных и неактивных пользователей имеют распределение фазового типа, тогда как в работе [9] рассматривались однолинейная система и экспоненциальное распределение времен обслуживания. Данные обобщения существенно усложняют анализ системы и повышают ее адекватность реальным системам.

Математическая модель

Рассмотрим многолинейную СМО, состоящую из N приборов и конечного буфера емкостью K, структура которой представлена на рис. 1.

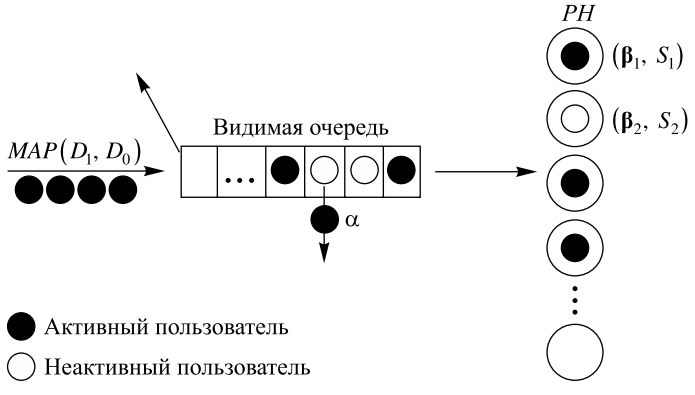

Рис. 1. Структура системы
Fig. 1. The structure of the queue

В систему поступает *MAP*-поток пользователей. Этот входной поток задается управляющим процессом ν_t, $t \geq 0$, который представляет собой неприводимую ЦМ с непрерывным временем и конечным пространством состояний $\{1, 2, \ldots, W\}$, и матрицами D_0, D_1. Средняя интенсивность поступления пользователей обозначается через λ и рассчитывается как $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$, где $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_W)$ – инвариантный вектор вероятностей ЦМ ν_t, $t \geq 0$. Он определяется как единственное решение системы $\boldsymbol{\theta}(D_0 + D_1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Здесь и далее $\mathbf{e}$ – вектор-столбец соответствующего размера, состоящий из единиц, а $\mathbf{0}$ – вектор-строка соответствующего размера, состоящая из нулей. Более подробное описание *MAP*-потока и формулы для определения его характеристик (например, коэффициентов корреляции и вариации) можно найти в работах [10–13].

Если пришедший пользователь обнаруживает, что один из приборов простаивает, он занимает его и начинает обслуживание. Если в момент поступления буфер заполнен, пользователь покидает систему навсегда. В противном случае он берет талон и присоединяется к системе. После этого пользователь может наблюдать длину очереди. Предположим, что пользователь решает, что длина очереди для него слишком велика, и становится неактивным пользователем с вероятностью q_k, $0 \leq q_k \leq 1$, где k – количество пользователей (активных и неактивных) в буфере в момент прибытия. С дополнительной вероятностью $1 - q_k$ пришедший пользователь остается в буфере как активный. Активных пользователей будем называть запросами первого типа, а неактивных пользователей – запросами второго типа.

Неактивный пользователь занимает место в буфере, но не требует полного обслуживания, и система не получает прибыли от обслуживания такого пользователя. Обратим внимание, что система не может распознать, неактивен ли пользователь, до начала его обслуживания. Считаем, что время обслуживания активного пользователя имеет фазовое распределение (*PH*) с неприводимым представлением $(\boldsymbol{\beta}_1, S_1)$, а время обслуживания неактивного пользователя имеет фазовое распределение с неприводимым представлением $(\boldsymbol{\beta}_2, S_2)$. Фазовое распределение времени обслуживания запроса типа l, $l = 1, 2$, означает следующее. Пусть есть ЦМ с непрерывным временем $\eta_t^{(l)}$, $t \geq 0$, имеющая пространство состояний $1, 2, \ldots, M_l, M_l + 1$, $l = 1, 2$. Состояния $1, 2, \ldots, M_l$ называются несущественными, а состояние $M_l + 1$ считается поглощающим. Начальное состояние данной ЦМ выбирается из множества несущественных состояний в соответствии с вероятностным вектором $\boldsymbol{\beta}_l$. Интенсивности выходов из текущих состояний и переходов между несущественными состояниями задаются субгенератором S_l. Интенсивности переходов в поглощающее состояние задаются элементами вектора $\boldsymbol{S}_0^{(l)} = -S_l \mathbf{e}$. Время обслуживания интерпретируется как время, за которое данная ЦМ достигнет поглощающего состояния. Подробная информация о распределении фазового типа и его свойствах представлена в работе [12].

Активные пользователи, находящиеся в буфере, могут проявлять нетерпеливость. То есть каждый активный пользователь независимо от других пользователей и собственного места в очереди может стать неактивным через экспоненциально распределенное время с параметром α, $\alpha > 0$.

Исследуем описанную систему.

Процесс изменения состояний системы и его стационарное распределение

Для облегчения исследования системы вместо отдельных распределений времени обслуживания активных и неактивных пользователей предлагаем ввести обобщенное распределение фазового типа (*GPH*) с неприводимым представлением $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, S)$, подробная информация о котором представлена в работе [14]. Данный прием позволяет существенно упростить процесс изменения состояний системы и облегчить его исследование.

Обобщенное время обслуживания можно интерпретировать как время до тех пор, пока управляющий марковский процесс m_t, $t \geq 0$, с конечным пространством состояний $\{1, \ldots, M, M+1\}$ достигнет единственного поглощающего состояния $M+1$. Здесь $M = M_1 + M_2$, где M_1 и M_2 – размер пространства состояний фазового распределения времени обслуживания активных и неактивных пользователей соответственно. Исходное состояние процесса m_t, $t \geq 0$, выбирается среди состояний $\{1, \ldots, M\}$ в зависимости от типа обслуживаемого пользователя. Если для обслуживания выбран активный пользователь, то начальное состояние этого процесса выбирается в соответствии с вероятностной вектор-строкой $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_1, \mathbf{0}_{M_2})$, а если для обслуживания выбран неактивный пользователь, то согласно вероятностной вектор-строке $\boldsymbol{\beta}_2 = (\mathbf{0}_{M_1}, \boldsymbol{\beta}_2)$. Интенсивности переходов процесса η_t, $t \geq 0$, внутри множества $\{1, \ldots, M\}$ определяются субгенератором $S = \begin{pmatrix} S_1 & O \\ O & S_2 \end{pmatrix}$, а интенсивности переходов в поглощающее состояние (что приводит к завершению обслуживания) задаются элементами вектор-столбца $\boldsymbol{S}_0 = -S\mathbf{e}$.

Пусть i_t, $i_t = \overline{0, N+K}$, – количество пользователей в системе (в буфере и на приборах); k_t, $k_t = \overline{0, \max\{0, i_t - N\}}$, – количество активных пользователей в буфере; ν_t, $\nu_t = \overline{1, W}$, – состояние управляющего процесса *MAP*, а $m_t^{(l)}$, $l = \overline{1, M}$, $m_t^{(l)} = \overline{0, \min\{i_t, N\}}$, $\sum_{l=1}^{M} m_t^{(l)} = \min\{i_t, N\}$ – количество приборов, обеспечивающих обслуживание на l-й фазе обобщенного процесса обслуживания, в момент времени t, $t \geq 0$.

Многомерный случайный процесс $\xi_t = \{i_t, k_t, \nu_t, m_t^{(1)}, \ldots, m_t^{(M)}\}$, $t \geq 0$, является немарковским, поскольку распределение времени обслуживания следующего пользователя зависит от того, активен или неактивен этот пользователь. Следовательно, чтобы получить марковский случайный процесс, введенные компоненты необходимо дополнить компонентами, определяющими статус каждого пользователя в буфере. Самый простой способ – явно указать статус каждого пользователя. Таким образом, можно исключить компоненту k_t. При $i_t > N$ каждый из $i_t - N$ пользователей в буфере помечается номером 1, если он активен, и номером 0 в противном случае. Пространство состояний этого набора компонент для каждого $i_t > N$ равно $2^{i_t - N}$. Соответственно, размер уравнений системы равновесия будет большим для сколь-либо большего значения K, и решение этой системы станет невозможным.

Другой возможный способ получить марковский процесс – дополнить процесс ξ_t, $t \geq 0$, указанием мест в буфере, занимаемых активными пользователями, и количеством неактивных пользователей в буфере, остающихся в очереди после любого активного пользователя. Такой способ обсуждается в статьях [5; 6]. Однако, как упоминается в работе [5], ее авторам не удалось справиться с вычислениями при $K > 9$. Напомним, что модель, рассмотренная в статье [5], существенно проще для анализа, чем исследуемая модель, поскольку предполагает нулевое время обслуживания неактивных пользователей и параметр W, равный единице, тогда как в данной работе допускаются произвольное конечное W и распределение времен обслуживания фазового типа. По этой причине делаем следующее упрощающее предположение. Если количество пользователей в буфере на момент завершения обслуживания пользователя равно i, $i = \overline{1, K}$, а количество активных пользователей равно k, $k = \overline{1, k}$, то с вероятностью $\frac{k}{i}$ следующим будет обслуживаться активный пользователь, а с дополнительной вероятностью будет обслуживаться неактивный пользователь. При этом предположении процесс $\xi_t = \{i_t, k_t, \nu_t, m_t^{(1)}, \ldots, m_t^{(M)}\}$, $t \geq 0$, становится неприводимой ЦМ с непрерывным временем.

Пусть состояния ЦМ ξ_t, $t \geq 0$, пронумерованы в лексикографическом порядке компонент.

Теорема 1. *Генератор Q ЦМ* ξ_t, $t \geq 0$, *имеет следующую блочно-трехдиагональную структуру:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & \dots & O & O & O \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & \dots & O & O & O \\ O & Q_{2,1} & Q_{2,2} & \dots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ O & O & O & \dots & Q_{K+N-1,K+N-2} & Q_{K+N-1,K+N-1} & Q_{K+N-1,K+N} \\ O & O & O & \dots & O & Q_{K+N,K+N-1} & Q_{K+N,K+N} \end{pmatrix},$$

где

$$Q_{0,0} = D_0,$$

$$Q_{i,i} = D_0 \otimes I_{T_i} + I_W \otimes \left(A_i + \Delta_i\right),\ 1 \leq i \leq N,$$

$$Q_{i,i} = I_{i-N+1} \otimes D_0 \otimes I_{T_N} + I_{(i-N+1)W} \otimes \left(A_N + \Delta_N\right) -$$

$$- \alpha C_{i-N+1} \otimes I_{WT_N} + \alpha C_{i-N+1} E^-_{i-N+1} \otimes I_{WT_N},\ i = \overline{N+1,\ N+K-1},$$

$$Q_{K+N,\,K+N} = I_{K+1} \otimes \left(D_0 + D_1\right) \otimes I_{T_N} + I_{(K+1)W} \otimes \left(A_N + \Delta_N\right) -$$

$$- \alpha C_{K+1} \otimes I_{WT_N} + \alpha C_{K+1} E^-_{K+1} \otimes I_{WT_N},$$

$$Q_{i,i+1} = D_1 \otimes P_i(\beta_1),\ i = \overline{0,\ N-1},$$

$$Q_{i,i+1} = \left(1 - q_{i-N}\right) E^+_{i-N+1} \otimes D_1 \otimes I_{T_N} + q_{i-N} \tilde{E}_{i-N+1} \otimes D_1 \otimes I_{T_N},\ i = \overline{N,\ N+K-1},$$

$$Q_{i,i-1} = I_W \otimes L_i,\ i = \overline{1,\ N},$$

$$Q_{i,i-1} = H_{i-N+1} \tilde{E}^-_{i-N+1} \otimes I_W \otimes L_N P_{N-1}(\beta_1) +$$

$$+ \left(I_{i-N+1} - H_{i-N+1}\right) \hat{E}_{i-N+1} \otimes I_W \otimes L_N P_{N-1}(\beta_2),\ i = \overline{N+1,\ N+K}.$$

Здесь $\otimes$ *– символ кронекерова произведения матриц;* $C_k = \operatorname{diag}\{0, 1, 2, \dots, k-2, k-1\}$, $k = \overline{2,\ K+1}$; $\operatorname{diag}\{\dots\}$ *– диагональная матрица с диагональными элементами, перечисленными в скобках;* E^-_k, $k = \overline{2,\ K+1}$, *– квадратная матрица размера k со всеми нулевыми элементами, кроме элементов* $\left(E^-_k\right)_{l,l-1}$, $l = \overline{1,\ k-1}$, *которые равны единице;* E^+_k, $k = \overline{1,\ K}$, *– матрица размера* $k \times (k+1)$ *со всеми нулевыми элементами, кроме элементов* $\left(E^+_k\right)_{l,l+1}$, $l = \overline{0,\ k-1}$, *которые равны единице;* $\tilde{E}_k$, $k = \overline{1,\ K}$, *– матрица размера* $k \times (k+1)$ *со всеми нулевыми элементами, кроме элементов* $\left(\tilde{E}_k\right)_{l,l}$, $l = \overline{0,\ k-1}$, *которые равны единице;* $\tilde{E}^-_k$, $k = \overline{2,\ K+1}$, *– матрица размера* $k \times (k-1)$ *со всеми нулевыми элементами, кроме элементов* $\left(\tilde{E}^-_k\right)_{l,l-1}$, $l = \overline{1,\ k-1}$, *которые равны единице;* $\hat{E}_k$, $k = \overline{2,\ K+1}$, *– матрица размера* $k \times (k-1)$ *со всеми нулевыми элементами, кроме элементов* $\left(\hat{E}_k\right)_{l,l}$, $l = \overline{0,\ k-1}$, *которые равны единице;* $H_k = \operatorname{diag}\left\{0, \frac{1}{k-1}, \frac{2}{k-1}, \dots, \frac{k-2}{k-1}, 1\right\}$, $k = \overline{2,\ K+1}$; $T_i = \frac{(i+M-1)!}{i!(M-1)!}$, $i = \overline{1,\ N}$, *– мощность пространства состояний процесса* $\left\{m_t^{(1)}, \dots, m_t^{(M)}\right\}$, $t \geq 0$, *при одновременном обслуживании i пользователей.*

Замечание 1. Матрицы A_i, $i = \overline{1,\ N}$, определяют интенсивности переходов компонент $\left\{m_t^{(1)}, \dots, m_t^{(M)}\right\}$, которые не влекут за собой окончание обслуживания при условии, что в данный момент заняты i при-

боров. Матрицы L_i, $i = \overline{1, N}$, определяют интенсивности переходов компонент $\left\{m_t^{(1)}, \ldots, m_t^{(M)}\right\}$, которые приводят к окончанию обслуживания в одном из занятых приборов при условии, что в данный момент заняты i приборов. Матрицы $P_i(\beta_l)$, $i = \overline{0, N-1}$, определяют вероятности переходов компонент $\left\{m_t^{(1)}, \ldots, m_t^{(M)}\right\}$ при условии, что в данный момент заняты i приборов и запрос типа l, $l = 1, 2$, начинает обслуживание. Диагональные элементы диагональной матрицы Δ_i, $i = \overline{1, N}$, определяют с точностью до знака интенсивности выхода процесса $\left\{m_t^{(1)}, \ldots, m_t^{(M)}\right\}$, $t \geq 0$, из соответствующего состояния. Алгоритмы вычисления матриц A_i, L_i, Δ_i, $i = \overline{1, N}$, $P_i(\beta_l)$, $i = \overline{0, N-1}$, $l = 1, 2$, представлены в работе [15].

Доказательство. Теорема доказывается путем анализа интенсивностей всех возможных переходов ЦМ ξ_t, $t \geq 0$, за бесконечно малый интервал времени. Блочно-трехдиагональная форма генератора Q легко объясняется тем, что пользователи поступают в систему и уходят из нее по одному.

Если система пуста (буфер пуст, и все приборы простаивают), поведение ЦМ ξ_t, $t \geq 0$, определяется только процессом ν_t, $t \geq 0$. Интенсивности его переходов, которые не приводят к изменению числа запросов в системе, задаются недиагональными элементами матрицы D_0, а интенсивности выхода из соответствующих состояний определяются с точностью до знака диагональными элементами этой матрицы, следовательно, $Q_{0,0} = D_0$.

Далее поясним вид блоков $Q_{i,i}$, $i = \overline{1, N+K}$. Так как это диагональный блок генератора, то все его диагональные элементы отрицательны, а модули этих элементов определяют интенсивности выхода ЦМ ξ_t, $t \geq 0$, из соответствующих состояний. Выход ЦМ ξ_t, $t \geq 0$, из текущего состояния возможен в следующих случаях.

1. Управляющий процесс ν_t, $t \geq 0$, поступления пользователей выходит из текущего состояния. Соответствующие интенсивности переходов определяются с точностью до знака диагональными элементами матриц $D_0 \otimes I_{T_i}$, если $i = \overline{1, N}$, и диагональными элементами матриц $I_{i-N+1} \otimes D_0 \otimes I_{T_N}$, если $i = \overline{N+1, N+K}$.

2. Процесс $\left\{m_t^{(1)}, \ldots, m_t^{(M)}\right\}$, $t \geq 0$, в одном из занятых приборов совершает выход из своего текущего состояния. Соответствующие интенсивности переходов определяются с точностью до знака диагональными элементами матриц $I_W \otimes \Delta_i$, если $i = \overline{1, N}$, и диагональными элементами матриц $I_{(i-N+1)W} \otimes \Delta_N$, если $i = \overline{N+1, N+K}$.

3. Активный пользователь, находящийся в буфере, покидает систему из-за нетерпеливости (становится неактивным). Соответствующие интенсивности задаются матрицами $\alpha C_{i-N+1} \otimes I_{WT_N}$, $i = \overline{N+1, N+K}$.

Недиагональные элементы матриц $Q_{i,i}$, $i = \overline{1, N+K+1}$, определяют интенсивности переходов ЦМ ξ_t, $t \geq 0$, без изменения значения i первой компоненты. Эти переходы определяются следующими элементами:

- недиагональными элементами матриц $D_0 \otimes I_{T_i}$, $i = \overline{1, N}$, и $I_{i-N+1} \otimes D_0 \otimes I_{T_N}$, $i = \overline{N+1, N+K}$, когда управляющий процесс ν_t, $t \geq 0$, совершает переход без генерации пользователя;
- элементами матриц $I_W \otimes A_i$, $i = \overline{1, N}$, и $I_{(i-N+1)W} \otimes A_N$, $i = \overline{N+1, N+K}$, когда процесс обслуживания в одном из занятых приборов совершает переход, не приводящий к окончанию обслуживания;
- элементами матриц $\alpha C_{i-N+1} E_{i-N+1}^{-} \otimes I_{WT_N}$, $i = \overline{N+1, N+K}$, когда активный пользователь становится неактивным из-за нетерпеливости;
- элементами матрицы $I_{K+1} \otimes D_1 \otimes I_{T_N}$, если пользователь поступает в систему при заполненности буфера (когда $i = N+K$) и покидает систему, не получив талон.

В результате имеем блоки $Q_{i,i}$, $i = \overline{0, N+K}$, представленные выше.

Форма блоков $Q_{i,i+1}$, $i = \overline{0, N+K-1}$, объясняется следующим образом. Данные блоки содержат интенсивности переходов ЦМ ξ_t, $t \geq 0$, которые приводят к увеличению количества пользователей в системе (в буфере и на приборах) на единицу. Если $i = \overline{0, N-1}$ (есть хотя бы один свободный прибор), эти переходы происходят, когда пользователь приходит в систему и начинает обслуживание. Соответствующие интенсивности определяются элементами матриц $D_1 \otimes P_i(\beta_1)$. Если $i = \overline{N, N+K-1}$ (все приборы заняты, но есть хотя бы одно свободное место в буфере), то увеличение количества пользователей в системе на единицу происходит, когда новый пользователь присоединяется к системе и становится в буфер для ожидания обслуживания. Интенсивности наступления этого события определяются элементами матриц

$\left(1-q_{i-N}\right)E^{+}_{i-N+1}\otimes D_1\otimes I_{T_N}$, если поступивший пользователь присоединяется к буферу как активный пользователь, и элементами матриц $q_{i-N}\tilde{E}_{i-N+1}\otimes D_1\otimes I_{T_N}$, если пришедший пользователь становится неактивным.

Теперь рассмотрим блоки $Q_{i,\,i-1}$, $i=\overline{1,\,N+K}$. Эти блоки содержат интенсивности переходов ЦМ ξ_t, $t\geq 0$, из состояния со значением i первой компоненты в состояние со значением $i-1$ данной компоненты. Такие переходы возможны только в случае завершения обслуживания. Если в момент завершения обслуживания буфер пуст, интенсивности наступления этого события определяются элементами матриц $I_W\otimes L_i$. В противном случае соответствующие интенсивности задаются элементами матриц $H_{i-N+1}\tilde{E}^{-}_{i-N+1}\otimes I_W\otimes L_N P_{N-1}(\beta_1)$, если активный пользователь начинает обслуживание, и элементами матриц $\left(I_{i-N+1}-H_{i-N+1}\right)\hat{E}_{i-N+1}\otimes I_W\otimes L_N P_{N-1}(\beta_2)$, если неактивный пользователь идет на обслуживание. Учитывая все эти пояснения, получаем формулы для блоков $Q_{i,\,i-1}$, $i=\overline{1,\,N+K}$, представленные выше. Теорема доказана.

Очевидно, что стационарные вероятности состояний системы $\pi\left(i,\,k,\,\nu,\,m^{(1)},\,\ldots,\,m^{(M)}\right)$, $i=\overline{0,\,K+N}$, $k=\overline{0,\,\max\{0,\,i-N\}}$, $\nu=\overline{1,\,W}$, $m^{(l)}=\overline{0,\,\min\{i,\,N\}}$, $\sum\limits_{l=1}^{M}m^{(l)}=\min\{i,\,N\}$, $l=\overline{1,\,M}$, существуют для всех возможных значений параметров системы. Сформируем вектор-строки $\boldsymbol{\pi}_i=\left(\boldsymbol{\pi}(i,\,k),\,k=\overline{0,\,\max\{0,\,i-N\}}\right)$ из этих вероятностей, пронумерованных в лексикографическом порядке компонент k, ν, $m^{(1)}$, …, $m^{(M)}$. Хорошо известно, что данные векторы удовлетворяют следующей системе линейных алгебраических уравнений:

$$\left(\boldsymbol{\pi}_0,\,\boldsymbol{\pi}_1,\,\ldots,\,\boldsymbol{\pi}_{K+N}\right)Q=\mathbf{0},\ \left(\boldsymbol{\pi}_0,\,\boldsymbol{\pi}_1,\,\ldots,\,\boldsymbol{\pi}_{K+N}\right)\mathbf{e}=1,$$

где Q – инфинитезимальный генератор ЦМ ξ_t, $t\geq 0$. Для решения этой системы может быть использован эффективный и численно устойчивый алгоритм, разработанный в статье [2].

Характеристики производительности

Среднее количество пользователей в системе рассчитывается по формуле

$$L_{\text{system}}=\sum_{i=1}^{K+N}i\boldsymbol{\pi}_i\mathbf{e}.$$

Среднее количество занятых приборов вычисляется следующим образом:

$$N_{\text{server}}=\sum_{i=1}^{K+N}\min\{i,\,N\}\boldsymbol{\pi}_i\mathbf{e}.$$

Среднее количество пользователей в буфере определяется как

$$N_{\text{buffer}}=\sum_{i=N+1}^{K+N}(i-N)\boldsymbol{\pi}_i\mathbf{e}=L_{\text{system}}-N_{\text{server}}.$$

Среднее количество активных пользователей в буфере вычисляется по формуле

$$N_{\text{buffer-active}}=\sum_{i=N+1}^{K+N}\sum_{k=1}^{i-N}k\boldsymbol{\pi}(i,\,k)\mathbf{e}.$$

Среднее количество неактивных пользователей в буфере рассчитывается следующим образом:

$$N_{\text{buffer-inactive}}=\sum_{i=N+1}^{K+N}\sum_{k=0}^{i-N-1}(i-N-k)\boldsymbol{\pi}(i,\,k)\mathbf{e}=N_{\text{buffer}}-N_{\text{buffer-active}}.$$

Средняя интенсивность обслуживания активных пользователей определяется как

$$\lambda_{\text{out-active}}=\sum_{i=1}^{K+N}\boldsymbol{\pi}_i\left(I_{(\max\{0,\,i-N\}+1)W}\otimes L^{(1)}_{\min\{i,\,N\}}\right)\mathbf{e}.$$

Средняя интенсивность обслуживания неактивных пользователей находится по формуле

$$\lambda_{\text{out-inactive}}=\sum_{i=1}^{K+N}\boldsymbol{\pi}_i\left(I_{(\max\{0,\,i-N\}+1)W}\otimes L^{(2)}_{\min\{i,\,N\}}\right)\mathbf{e}.$$

Замечание 2. Матрицы $L_i^{(l)} = L_i^{(l)}\left(\tilde{\boldsymbol{S}}_0^{(l)}\right)$, $i = \overline{1, N}$, $l = 1, 2$, вычисляются по тому же алгоритму, что и матрицы L_i. Здесь $\tilde{\boldsymbol{S}}_0^{(1)}$ – вектор-столбец размера M вида $\left(\left(\boldsymbol{S}_0^{(1)}\right)^T, \mathbf{0}_{M_2}\right)^T$, а $\tilde{\boldsymbol{S}}_0^{(2)}$ – вектор-столбец размера M вида $\left(\mathbf{0}_{M_1}, \left(\boldsymbol{S}_0^{(2)}\right)^T\right)^T$.

Вероятность того, что в произвольный момент система простаивает, определяется следующим образом:

$$P_{\text{idle}} = \boldsymbol{\pi}_0 \mathbf{e}.$$

Вероятность того, что пришедший пользователь начнет обслуживание по прибытии, вычисляется как

$$P_{\text{imm}} = \frac{1}{\lambda}\sum_{i=0}^{N-1} \boldsymbol{\pi}_i \left(D_1 \otimes P_i(\boldsymbol{\beta}_1)\right)\mathbf{e}.$$

Вероятность потери произвольного активного пользователя из буфера из-за нетерпеливости рассчитывается по формуле

$$P_{\text{imp-loss}} = \frac{\alpha}{\lambda}\sum_{i=N+1}^{K+N}\sum_{k=1}^{i-N} k\boldsymbol{\pi}(i, k)\mathbf{e} = \frac{\alpha N_{\text{buffer-active}}}{\lambda}.$$

Вероятность потери произвольного пользователя на входе из-за отказа присоединяться к длинной очереди вычисляется следующим образом:

$$P_{\text{balk-loss}} = \frac{1}{\lambda}\sum_{i=N}^{K+N-1} q_{i-N}\boldsymbol{\pi}_i\left(I_{i-N+1} \otimes D_1 \otimes I_{T_N}\right)\mathbf{e}.$$

Вероятность потери произвольного пользователя на входе из-за переполненности буфера определяется по формуле

$$P_{\text{full-buffer-loss}} = \frac{1}{\lambda}\boldsymbol{\pi}_{K+N}\left(I_{K+1} \otimes D_1 \otimes I_{T_N}\right)\mathbf{e}.$$

Общая вероятность потери произвольного пользователя рассчитывается как

$$P_{\text{loss}} = 1 - \frac{\lambda_{\text{out-active}}}{\lambda} = P_{\text{imp-loss}} + P_{\text{balk-loss}} + P_{\text{full-buffer-loss}}.$$

Последнее выражение можно использовать для проверки точности при отладке программы, а также вычисления стационарных вероятностей системы и показателей ее производительности.

Численный эксперимент

В численном эксперименте исследуем зависимость основных характеристик системы от числа приборов (N) и емкости буфера (K), а также коснемся вопроса выбора оптимальных параметров системы. Предположим, что входной *MAP*-поток пользователей в систему задается матрицами

$$D_0 = \begin{pmatrix} -1,6 & 0 \\ 0 & -0,4 \end{pmatrix},\ D_1 = \begin{pmatrix} 1,5 & 0,1 \\ 0,02 & 0,38 \end{pmatrix}.$$

Средняя интенсивность поступления пользователей составляет 0,6, коэффициенты корреляции и вариации последовательных времен между поступлениями равны 0,170 673 и 1,625 соответственно.

Пусть время обслуживания активного пользователя имеет распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}_1, S_1)$, где $\boldsymbol{\beta}_1 = (0,7,\ 0,3)$, а $S_1 = \begin{pmatrix} -0,5 & 0 \\ 0,1 & -0,2 \end{pmatrix}$. Среднее время обслуживания активного пользователя равно 3,2.

Пусть время обслуживания неактивного пользователя имеет распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}_2, S_2)$, где $\boldsymbol{\beta}_2 = (1, 0)$ и $S_2 = \begin{pmatrix} -4 & 4 \\ 0 & -4 \end{pmatrix}$. Среднее время обслуживания неактивного пользователя равно 0,5.

Предположим, что интенсивность ухода пользователей из-за нетерпеливости составляет $\alpha = 0,05$.

В данном численном эксперименте будем изменять емкость буфера в интервале $[1, 30]$ и число приборов в интервале $[1, 5]$ с шагом 1.

Вероятности ухода пользователей из-за нежелания присоединяться к длинной очереди определим как $q_k = \dfrac{k+1}{31}$, $k = \overline{0, 29}$.

На рис. 2–6 представлены зависимости значений L_{system}, N_{server}, N_{buffer}, $N_{\text{buffer-active}}$ и $N_{\text{buffer-inactive}}$ от параметров N и K.

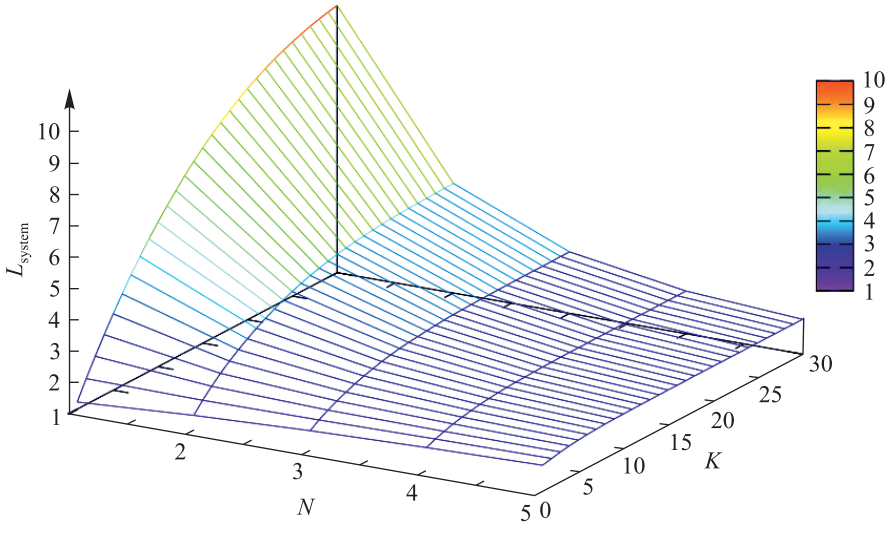

Рис. 2. Зависимость L_{system} от N и K
Fig. 2. Dependence L_{system} on N and K

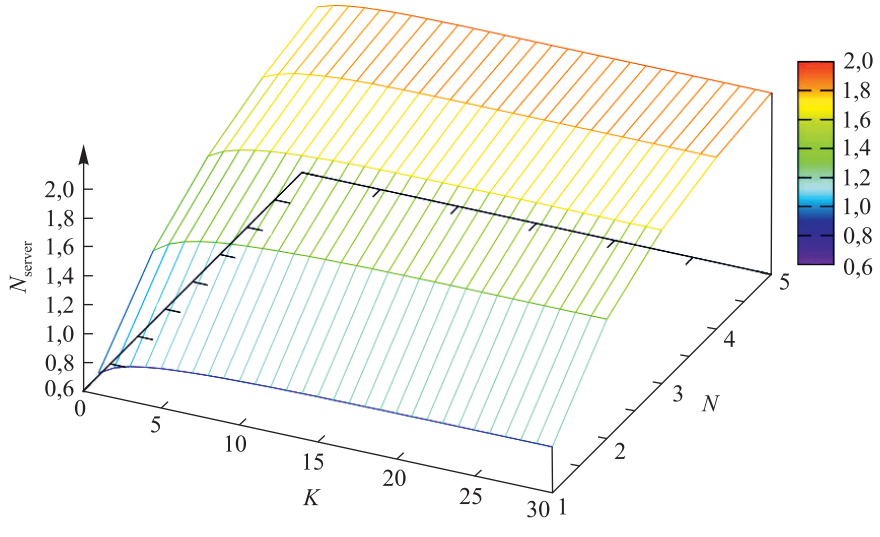

Рис. 3. Зависимость N_{server} от N и K
Fig. 3. Dependence N_{server} on N and K

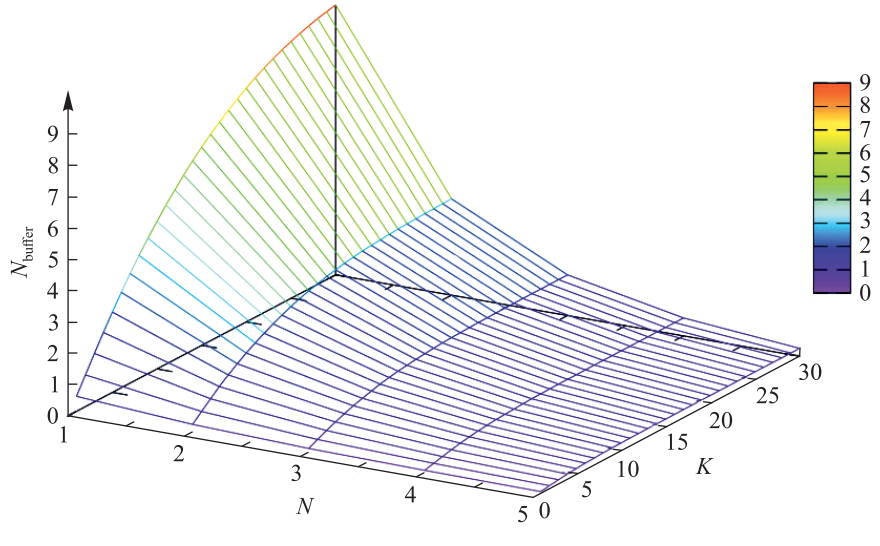

Рис. 4. Зависимость N_{buffer} от N и K
Fig. 4. Dependence N_{buffer} on N and K

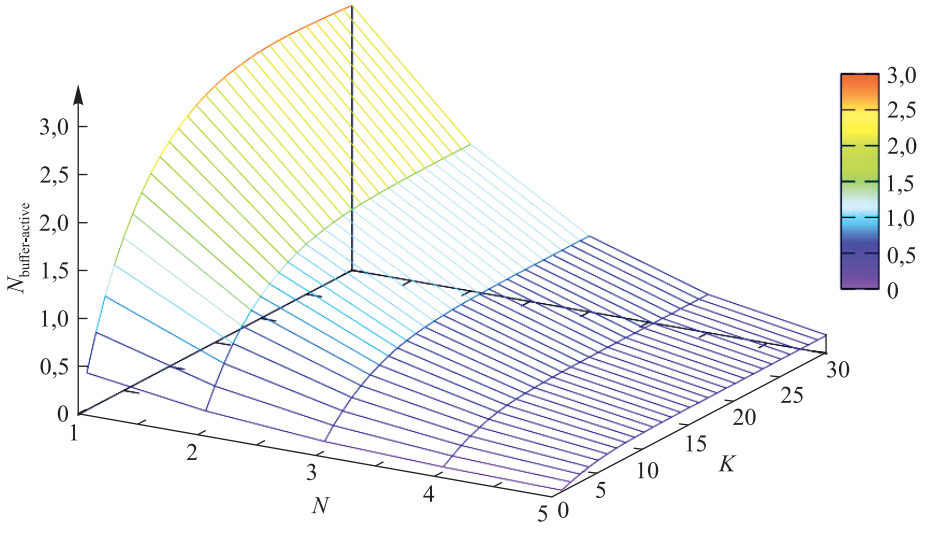

Рис. 5. Зависимость $N_{buffer\text{-}active}$ от N и K
Fig. 5. Dependence $N_{buffer\text{-}active}$ on N and K

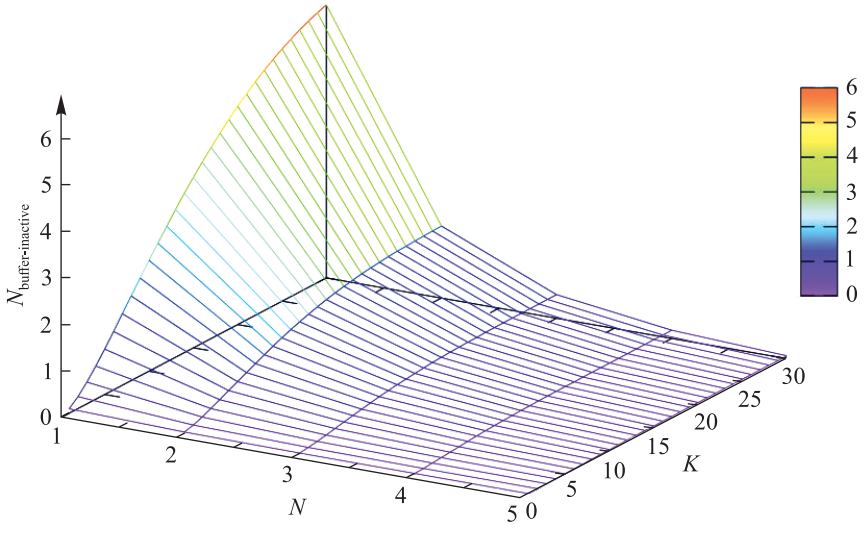

Рис. 6. Зависимость $N_{buffer\text{-}inactive}$ от N и K
Fig. 6. Dependence $N_{buffer\text{-}inactive}$ on N and K

Как видно из рис. 2, значение L_{system} растет с увеличением параметра K, поскольку в результате роста емкости буфера увеличивается емкость системы и уменьшается вероятность того, что пользователь покинет систему из-за занятости буфера. Следовательно, в буфере находится большее количество пользователей, что приводит к увеличению значения L_{system}. В то же время при фиксированной емкости буфера значение L_{system} может как увеличиваться, так и уменьшаться с ростом N. С одной стороны, при увеличении параметра N растет емкость системы и снижается вероятность потери пользователей на входе, следовательно, в системе находится больше запросов. С другой стороны, при увеличении числа приборов пользователи меньше ждут в очереди, что приводит к сокращению их количества в системе.

Как следует из рис. 3, значение N_{server} растет как с увеличением параметра N, так и с увеличением параметра K. В то же время значения N_{buffer}, $N_{buffer\text{-}active}$ и $N_{buffer\text{-}inactive}$ уменьшаются при росте N, но увеличиваются при росте K, что можно объяснить, используя приведенные выше рассуждения. Из рис. 4–6 становится очевидно, что при малых значениях N и больших значениях K существенная часть запросов, находящихся в очереди, являются неактивными, а соответственно, при таких параметрах система работает плохо.

На рис. 7 показана зависимость вероятности P_{imm} от параметров N и K. Очевидно, что данная вероятность увеличивается с ростом числа приборов и уменьшается с ростом емкости буфера.

На рис. 8–11 представлены зависимости вероятностей $P_{balk\text{-}loss}$, $P_{imp\text{-}loss}$, $P_{full\text{-}buffer\text{-}loss}$ и P_{loss} от параметров N и K.

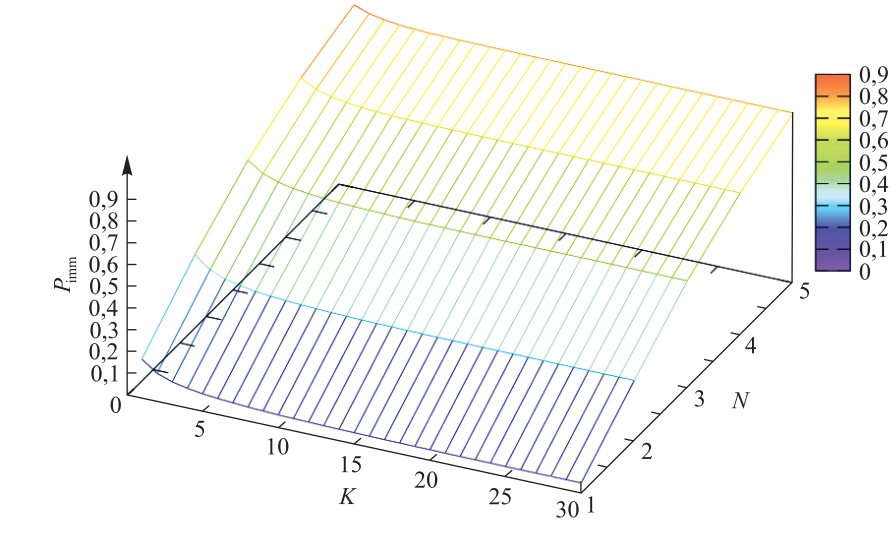

Рис. 7. Зависимость P_{imm} от N и K
Fig. 7. Dependence P_{imm} on N and K

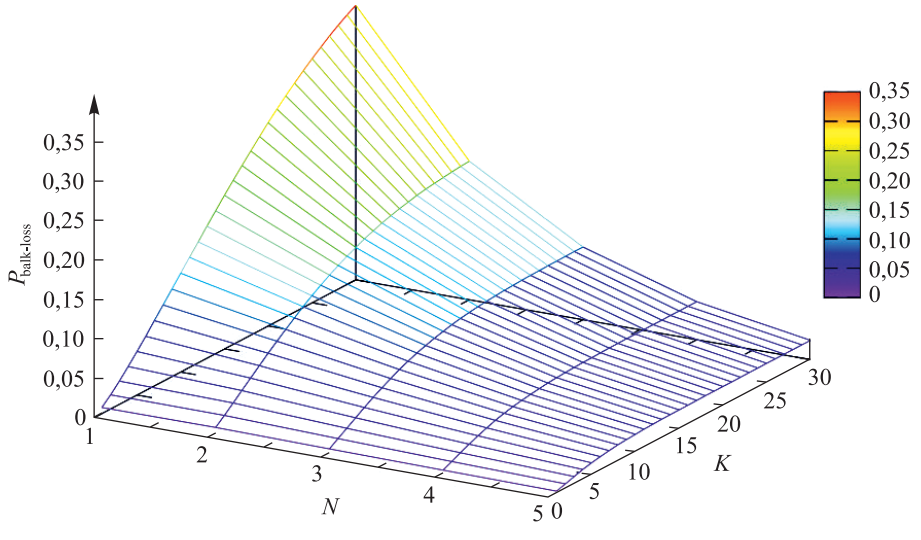

Рис. 8. Зависимость $P_{balk\text{-}loss}$ от N и K
Fig. 8. Dependence $P_{balk\text{-}loss}$ on N and K

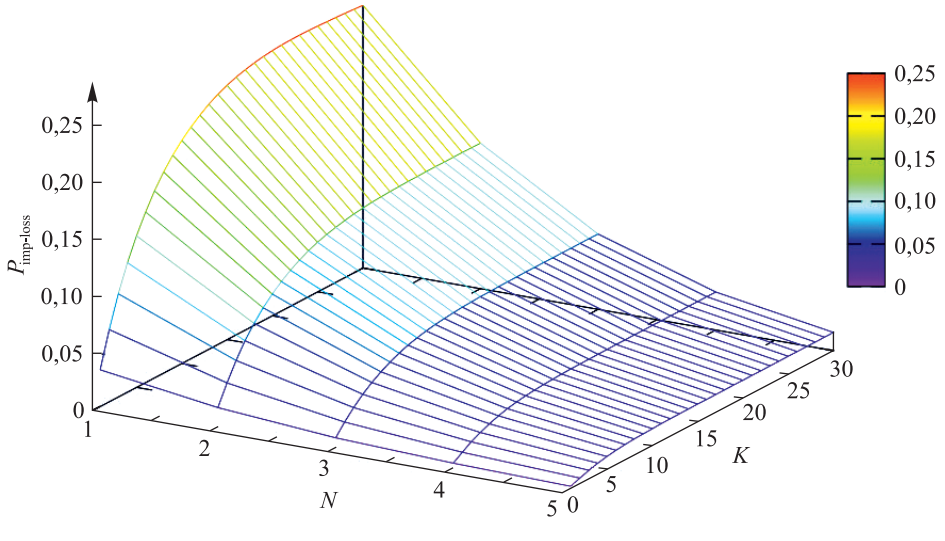

Рис. 9. Зависимость $P_{imp\text{-}loss}$ от N и K
Fig. 9. Dependence $P_{imp\text{-}loss}$ on N and K

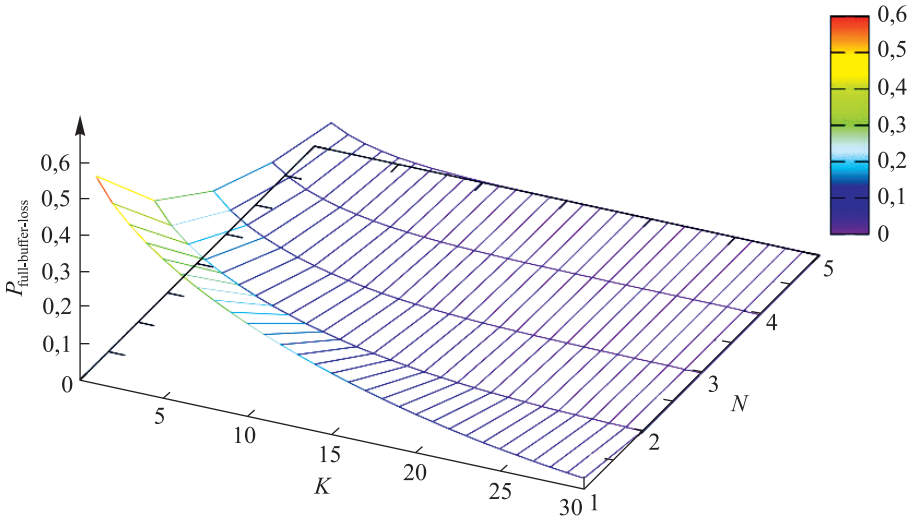

Рис. 10. Зависимость $P_{\text{full-buffer-loss}}$ от N и K
Fig. 10. Dependence $P_{\text{full-buffer-loss}}$ on N and K

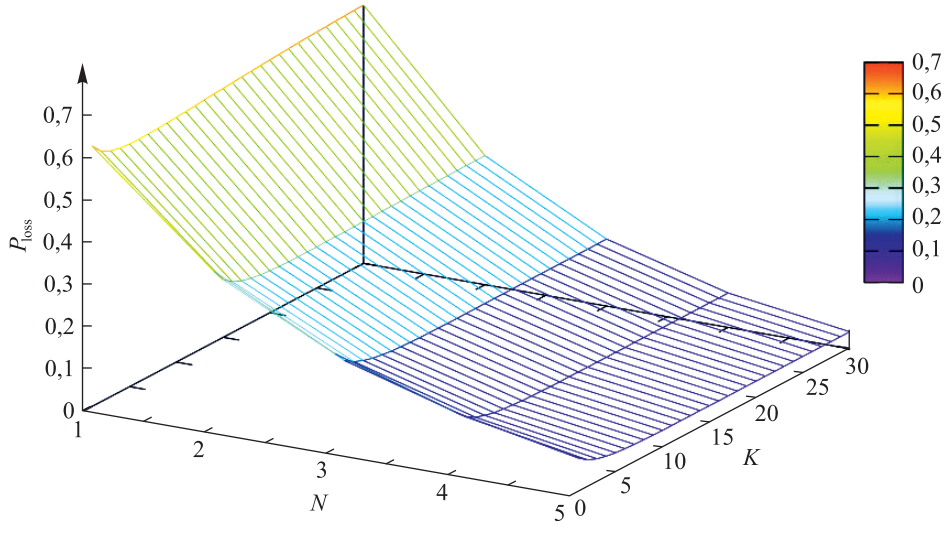

Рис. 11. Зависимость P_{loss} от N и K
Fig. 11. Dependence P_{loss} on N and K

Как видно из рис. 8 и 9, вероятности $P_{\text{balk-loss}}$ и $P_{\text{imp-loss}}$ уменьшаются с ростом N и увеличиваются с ростом K. В то же время, как следует из рис. 10, вероятность $P_{\text{full-buffer-loss}}$ уменьшается при увеличении значений N и K. Вероятность P_{loss} является суммой трех перечисленных вероятностей. В отличие от классических СМО, где с ростом емкости буфера общая вероятность потери произвольного пользователя уменьшается, в данном случае при увеличении значения K вероятность P_{loss} ведет себя немонотонно. Например, при фиксированном значении $N = 1$ ее величина составляет $P_{\text{loss}} = 0{,}616\,31$ при $K = 1$, $P_{\text{loss}} = 0{,}579\,08$ при $K = 10$ и $P_{\text{loss}} = 0{,}611\,01$ при $K = 30$. Рост общей вероятности потери произвольного пользователя с увеличением емкости буфера объясняется следующим образом. При слишком большой емкости буфера и малом числе приборов многие пользователи присоединяются к очереди и не дожидаются начала обслуживания, фактически уходя из системы из-за нетерпеливости, при этом оставляя свой талон. В данном случае длина очереди не уменьшается, а система вынуждена тратить ресурсы на обслуживание большого числа неактивных пользователей. Другими словами, если количество пользователей в очереди достаточно велико и высока вероятность того, что пришедший пользователь не дождется обслуживания, имеет смысл отказать в обслуживании сразу. В рассматриваемом примере вероятность P_{loss} принимает минимальное значение 0,040 371 при $N = 5$ и $K = 18$. С точки зрения минимизации общей вероятности потери произвольного пользователя данные параметры являются оптимальными. Однако в реальных системах содержание каждого прибора требует затрат. К тому же штрафы за потерю пользователя из-за разных причин могут отличаться. Например, репутационные потери системы в случае, когда пользователь не захотел ждать обслуживания по приходу в систему, отличаются от потерь в случае, когда

пользователь ждал обслуживания, но так и не дождался. Чтобы учесть данные аспекты, предположим, что качество функционирования системы задается следующим экономическим критерием:

$$E = E(N, K) = a\lambda_{\text{out-active}} - b\lambda_{\text{out-inactive}} - c_1\lambda P_{\text{balk-loss}} - c_2\lambda P_{\text{full-buffer-loss}} - c_3\lambda P_{\text{imp-loss}} - d_1K - d_2N.$$

Здесь a – прибыль, получаемая системой за обслуживание одного активного пользователя; b – затраты системы на обслуживание неактивного пользователя; c_1, c_2 и c_3 – штрафы, уплачиваемые системой за потерю пользователя из-за отказа присоединяться к длинной очереди, переполненности буфера и нетерпеливости соответственно; d_1 – плата за содержание одной единицы буферного пространства, а d_2 – плата за использование одного прибора. Экономический критерий $E(N, K)$ определяет среднюю прибыль, получаемую системой в единицу времени. Наша цель – найти оптимальные значения числа приборов и емкости буфера, при которых средняя прибыль системы была бы максимальной.

Зафиксируем следующие значения стоимостных коэффициентов: $a = 5$; $b = 0{,}5$; $c_1 = 1$; $c_2 = 1{,}2$; $c_3 = 1{,}4$; $d_1 = 0{,}001$; $d_2 = 0{,}2$.

Зависимость значений экономического критерия $E(N, K)$ от параметров N и K представлена на рис. 12.

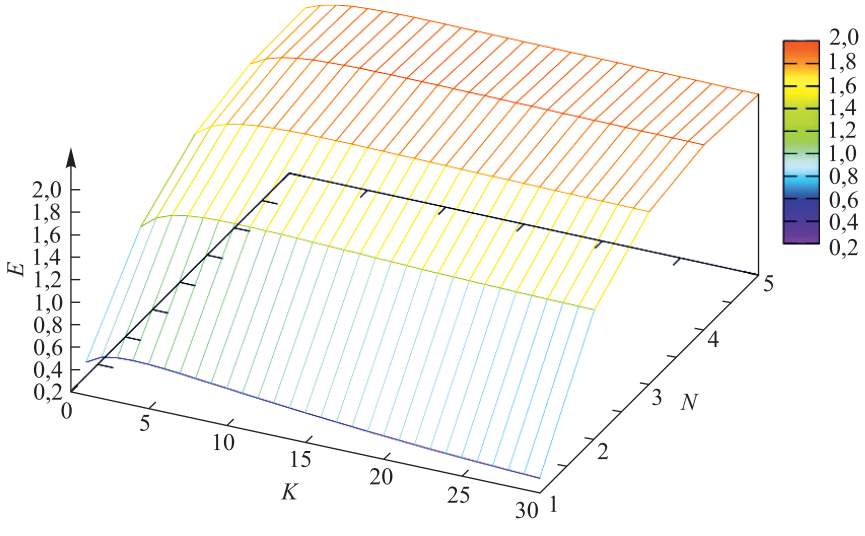

Рис. 12. Зависимость экономического критерия $E(N, K)$ от параметров N и K
Fig. 12. Dependence of the economic criterion $E(N, K)$ on parameters N and K

Как видно из рис. 12, значения экономического критерия $E(N, K)$ ведут себя немонотонно как по N, так и по K. Оптимальное значение критерия качества составляет 1,864 15 и достигается при $N = 4$ и $K = 10$. Другими словами, прибыль описанной системы будет максимальной, если мы зафиксируем число приборов, равное 4, и емкость буфера, равную 10.

Замечание 3. Вычисления проводились на персональном компьютере с процессором Intel Core i7-8700 (CPU, 16 RAM) с использованием программы *Wolfram Mathematica* (версия 13.2). Время расчетов в данном численном эксперименте составило 232 с на 150 различных вариантов пар (N, K), или в среднем 1,57 с на 1 пару.

Заключение

В настоящей работе исследована СМО с коррелированным входным потоком и нетерпеливыми запросами, моделирующая функционирование систем с электронной очередью. Найдено стационарное распределение состояний системы, вычислены основные характеристики производительности. Приведен численный эксперимент, иллюстрирующий зависимость характеристик производительности от числа приборов и емкости буфера. Полученные результаты могут быть использованы на практике для выбора оптимальных параметров системы с точки зрения заданного критерия качества.

Библиографические ссылки / References

1. Sun B, Dudin A, Dudin S. Queueing system with impatient customers, visible queue and replenishable inventory. *Applied and Computational Mathematics.* 2018;17(2):161–174.

2. Dudin A, Dudina O, Dudin S, Gaidamaka Y. Self-service system with rating dependent arrivals. *Mathematics.* 2022;10(3):297. DOI: 10.3390/math10030297.

3. Garnett O, Mandelbaum A, Reiman M. Designing a call center with impatient customers. *Manufacturing & Service Operations Management.* 2002;4(3):208–227. DOI: 10.1287/msom.4.3.208.7753.

4. Wang K, Li N, Jiang Z. Queueing system with impatient customers: a review. In: *Proceedings of 2010 IEEE International conference on service operations and logistics, and informatics; 2010 July 15–17; QingDao, China.* [S. l.]: IEEE; 2010. p. 82–87. DOI: 10.1109/SOLI.2010.5551611.

5. Xu SH, Gao L, Ou J. Service performance analysis and improvement for a ticket queue with balking customers. *Management Science.* 2007;53(6):971–990. DOI: 10.1287/mnsc.1060.0660.

6. Hanukov G, Hassoun M, Musicant O. On the benefits of providing timely information in ticket queues with balking and calling times. *Mathematics*. 2021;9(21):2753. DOI: 10.3390/math9212753.

7. Jennings OB, Pender J. Comparisons of ticket and standard queues. *Queueing Systems.* 2016;84(1–2):145–202. DOI: 10.1007/s11134-016-9493-y.

8. Xiao L, Xu SH, Yao DD, Zhang H. Optimal staffing for ticket queues. *Queueing Systems*. 2022;102(1–2):309–351. DOI: 10.1007/s11134-022-09854-8.

9. Kim C, Dudin A, Dudin S, Dudina O. Analysis of *MAP*/*M*/*1*/*K* ticket queue with users balking and reneging and service of no-show users. In: Vicario E, Bandinelli R, Fani V, Mastroianni M, editors. *Proceedings of the 37th ECMS International conference on modelling and simulation, ECMS 2023; 2023 June 20–23; Florence, Italy.* Saarbrücken: Digitaldruck Pirrot; 2023. p. 26–32 (Communications of the ECMS; volume 37, issue 1).

10. Chakravarthy SR. *Introduction to matrix-analytic methods in queues. Volume 1, Analytical and simulation approach – basics.* London: ISTE; 2022. XV, 341 p. (Limnios N, editor. Mathematics and statistics series). Co-published bu the John Wiley & Sons. DOI: 10.1002/9781394165421.

11. Chakravarthy SR. *Introduction to matrix-analytic methods in queues. Volume 2, Analytical and simulation approach – queues and simulation.* London: ISTE; 2022. XV, 415 p. (Limnios N, editor. Mathematics and statistics series). Co-published bu the John Wiley & Sons. DOI: 10.1002/9781394174201.

12. Dudin AN, Klimenok VI, Vishnevsky VM. *The theory of queuing systems with correlated flows*. Cham: Springer; 2020. XXII, 410 p. DOI: 10.1007/978-3-030-32072-0.

13. 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.

14. Kim C, Dudin A, Dudina O, Dudin S. Tandem queueing system with infinite and finite intermediate buffers and generalized phase-type service time distribution. *European Journal of Operational Research.* 2014;235(1):170–179. DOI: 10.1016/j.ejor.2013.12.012.

15. Kim C, Dudin A, Dudin S, Dudina O. Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users. *IEEE Access.* 2021;9:106933–106946. DOI: 10.1109/ACCESS.2021.3100561.

Получена 16.04.2024 / исправлена 20.06.2024 / принята 20.06.2024.
Received 16.04.2024 / revised 20.06.2024 / accepted 20.06.2024.