ИССЛЕДОВАНИЕ ХАРАКТЕРИСТИК ПРИОРИТЕТНОЙ МУЛЬТИСЕРВИСНОЙ СИСТЕМЫ MMAP/PH/M/N С ИСПОЛЬЗОВАНИЕМ МЕТОДА МОНТЕ-КАРЛО[1]

Вишневский В. М.[2]
(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Москва*)
Клименок В. И.[3]
(*Белорусский государственный университет, Минск*)
Ларионов А. А.[4], **Мухтаров А. А.**[5], **Соколов А. М.**[6]
(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Москва*)

Представлены результаты исследования приоритетной многолинейной системы массового обслуживания (СМО) с маркированным марковским входным потоком (MMAP), обслуживанием фазового типа PH и очередью конечной ёмкости. Приоритетные классы трафика различаются вероятностью присоединения к очереди, зависящей от количества заявок в ней, и PH-распределением времени обслуживания. Если очередь заполнена, заявка не присоединяется к системе. Для частного случая такой СМО с двумя классами трафика разработана и исследована аналитическая модель, а также предложен алгоритм вычисления стационарных вероятностей состояния системы, вероятностей потерь, среднего числа заявок в системе и других характеристик. Для общего случая системы с K-классами построена имитационная модель, исследованы характеристики системы.

Ключевые слова: многолинейная система массового обслуживания, метод Монте-Карло, стационарный режим.

---

[1] *Исследование выполнено за счет гранта Российского научного фонда №23-29-00795, https://rscf.ru/project/23-29-00795.*

[2] *Владимир Миронович Вишневский, д.т.н., профессор* (*vishn@inbox.ru*).

[3] *Валентина Ивановна Клименок, д.ф-м.н., профессор* (*klimenok@bsu.by*).

[4] *Андрей Алексеевич Ларионов, к.т.н.* (*larioandr@gmail.com*).

[5] *Амир Амангельдыевич Мухтаров, к.т.н.* (*mukhtarov.amir.a@gmail.com*).

[6] *Александр Михайлович Соколов, аспирант ИПУ РАН,* (*aleksandr.sokolov@phystech.edu*).

1. Введение

Системы массового обслуживания с приоритетами, являющиеся важным разделом теории очередей, эффективно используются при анализе реальных технических и социальных систем [1, 2, 7, 9, 19, 20, 22]. Примерами систем с приоритетным трафиком являются цифровое телевидение, в котором передача синхронизирующих сигналов имеет более высокий приоритет, чем передача видео, оптические сети нового поколения с поддержкой приоритетных классов обслуживания (QoS), в частности – системы Интернета вещей (IoT) [2, 7, 20, 22], информационные сервисы с различными категориями пользователей [19], а также любые социальные системы с клиентами разного типа, например, больницы. В качестве примера последних можно привести исследования [1, 9], в которых изучается эффективность алгоритмов распределения мест в очереди на пересадку органов для пациентов в зависимости от тяжести их заболевания. Кроме того, приоритетные системы используются для предоставления пользователям приоритетного входа в систему при работе с различными сервисами. Так в статье [19] приведён анализ влияния приоритетов на загруженность веб-сервисов, время проведения транзакции и другие характеристики, исследована производительность различных баз данных для пользователей с разными приоритетами доступа.

Как известно, трафик в современных компьютерных сетях является коррелированным [11], и для его моделирования необходимо использовать марковские входные потоки MAP[4]. Естественным обобщением MAP-потоков на случай приоритетного неоднородного трафика являются маркированные марковские входные потоки (Marked MAP, $MMAP$) [4], которые позволяют описывать коррелированные поступления заявок для произвольного числа классов трафика.

Системы с $MMAP$-потоками слабо исследованы в мировой литературе по сравнению с классическими MAP-потоками [3, 4, 17]. В статье [10] анализируется размер очереди приоритетной системы $MMAP/MAP/1$. Исследование условий ста-

ционарного режима многолинейной СМО с входящим $MMAP$-потоком представлено в статье [9]. Однако в ней отсутствует описание алгоритма расчёта стационарных вероятностей состояний и других характеристик производительности системы. В недавних работах [5, 13] была исследована проблема поиска стационарного решения для приоритетных систем с входным $MMAP$-потоком для случая одного обслуживающего прибора. В статье [13] исследуется очередь с относительным приоритетами, в [5] – с абсолютным приоритетом.

В статье [14] рассмотрена сложная многолинейная СМО с входящим $MMAP$-потоком с двумя классами приоритетов и отсутствием буфера. Настоящая работа является развитием и обобщением этой статьи. Принципиальным отличием настоящей статьи от [14] является наличие буфера конечного размера и произвольного числа классов заявок. Такое обобщение, с одной стороны, значительно усложняет математический анализ даже для двух классов приоритетов, но, с другой стороны, расширяет область практического применения рассматриваемой модели.

Формальная постановка задачи и описание модели приведены в разделе 2. В разделе 3 приводится аналитическое решение для частного случая рассматриваемой модели. Дано описание многомерной марковской цепи, алгоритма расчёта вероятностей стационарного состояния и других характеристик системы. В разделе 4 приведены результаты численного исследования основных характеристик рассматриваемой многолинейной СМО, включая вероятность потери заявок и времени пребывания заявок в системы.

2. Постановка задачи

Рассматривается N-линейная система массового обслуживания с буфером размера R. Заявки разных типов поступают в $MMAP$-потоке под управлением неприводимой цепи Маркова с непрерывным временем ν_t, $t \geqslant 0$, которая принимает значения в множестве $\{0, 1, 2, \ldots, W\}$. Процесс ν_t пребывает в состоянии ν в течение экспоненциально распределенного времени

с параметром λ_ν, $\nu = \overline{0, W}$, после чего с вероятностью $p_k(\nu, \nu')$ переходит в состояние ν' и генерируется заявка k-го типа, $k \in \{1, 2, \ldots, K\}$, или с вероятностью $p_0(\nu, \nu')$ цепь переходит в состояние ν' без генерации заявки, причем $p_0(\nu, \nu) = 0$. Процесс ν_t называется управляющим процессом $MMAP$-потока. Для указанных вероятностей выполняются естественные ограничения: $\sum\limits_{k=1}^{K} \sum\limits_{\nu'=0}^{W} p_k(\nu, \nu') = 1,\ \nu, \nu' = \overline{0, W}$.

Таким образом, $MMAP$-поток задается:

– размерностью пространства состояний управляющего процесса, $W + 1$;

– количеством классов заявок, K;

– интенсивностями времен пребывания управляющего процесса в соответствующих состояниях, λ_ν, $\nu = \overline{0, W}$;

– вероятностями переходов, $p_k(\nu, \nu')$, $k = \overline{1, K}$, $\nu, \nu' = \overline{0, W}$.

Всю информацию о $MMAP$ удобно хранить в виде набора матриц D_k, $k = \overline{1, K}$, порядка $(W + 1) \times (W + 1)$, элементы которых определяются как

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

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

Нетрудно видеть, что элементами матриц D_k, $k = \overline{1, K}$, являются интенсивности переходов процесса ν_t, сопровождающиеся генерацией заявки k-го типа. Аналогичный смысл имеют недиагональные элементы матрицы D_0, а диагональные элементы этой матрицы – это взятые с противоположным знаком интенсивности выхода процесса ν_t из соответствующих состояний.

Естественное требование к матрицам D_k, $k = \overline{1, K}$, состоит в том, что не все они нулевые. При выполнении этого требования матрица D_0 является невырожденной и, более того, устойчивой, так как все ее собственные значения имеют отрицательную действительную часть.

Матрицы D_k, $k = \overline{1,K}$, можно задавать их матричной производящей функцией $D(z) = \sum\limits_{k=0}^{K} D_k z^k$, $|z| < 1$. Отметим, что значение этой функции в точке $z = 1$ – матрица $D(1)$ – является инфинитезимальным генератором управляющего процесса ν_t, $t \geqslant 0$. Стационарное распределение этого процесса, представленное в виде вектора-строки $\boldsymbol{\theta}$, определяется как решение системы линейных алгебраических уравнений $\boldsymbol{\theta} D(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$. Здесь и далее $\mathbf{e}$ – вектор-столбец, состоящий из единиц.

Интенсивность λ_k поступления заявок k-го типа в $MMAP$-потоке задается формулой

$$\lambda_k = \boldsymbol{\theta} D_k \mathbf{e},\ k = \overline{1,K},$$

а суммарная интенсивность поступления заявок λ – по формуле

$$\lambda = \boldsymbol{\theta} \sum_{k=1}^{K} D_k \mathbf{e}.$$

Дисперсия v_k длин интервалов между моментами поступления групп заявок k-го типа вычисляется по формуле

$$v^{(k)} = \frac{2\boldsymbol{\theta}(-D_0 - \sum\limits_{l=1,l\neq k}^{K} D_l)^{-1}\mathbf{e}}{\lambda_k} - \left(\frac{1}{\lambda_k}\right)^2,\ k = \overline{1,K}.$$

Коэффициент корреляции $c_{cor}^{(k)}$ длин двух соседних интервалов между моментами поступления групп заявок k-го типа вычисляется по формуле

$$c_{cor}^{(k)} = \left[\frac{\boldsymbol{\theta}(D_0 + \sum\limits_{l=1,l\neq k}^{K} D_l)^{-1}\mathbf{e}}{\lambda_k} D_k (D_0 + \sum_{l=1,l\neq k}^{K} D_l)^{-1}\mathbf{e} - \left(\frac{1}{\lambda_k}\right)^2\right] v_k^{-1},$$

$$k = \overline{1,K}.$$

Более подробное описание $MMAP$ можно найти, например, в [9]. Отметим, что стационарный пуассоновский поток является частным случаем $MMAP$-потока при $W = 0$, $K = 1$, $D_0 = -\lambda$, $D_1 = \lambda$.

В общем случае заявки разных типов отличаются приоритетами и параметрами PH – распределения времени обслуживания. Более подробно эти отличия будут описаны ниже. Полагаем, что все обслуживающие приборы одинаковы и независимы друг от друга. Время обслуживания любым прибором заявки k-го, $k = 1, .., K$, типа имеет фазовое распределение (PH–Phase type distribution), которое задается парой $(\boldsymbol{\beta}_k, S_k)$. Здесь $\boldsymbol{\beta}_k$–вектор-строка порядка M_k, а S_k – квадратная матрица порядка M_k. Таким образом заданное время обслуживания интерпретируется как время, за которое некоторая управляющая цепь Маркова $m_t^{(k)}$, $t \geqslant 0$, с пространством состояний $\{1, \ldots, M_k, M_k + 1\}$ достигнет единственного поглощающего состояния $M_k + 1$. Переходы цепи $m_t^{(k)}$, $t \geqslant 0$, в пространстве несущественных состояний $\{1, \ldots, M_k\}$ задаются субгенератором S_k, а интенсивности переходов в поглощающее состояние задаются вектором $\boldsymbol{S}_0^{(k)} = -S_k\mathbf{e}$. В момент начала обслуживания состояние процесса $m_t^{(k)}$, $t \geqslant 0$, выбирается из пространства состояний $\{1, \ldots, M_k\}$ на основании вероятностного вектора-строки $\boldsymbol{\beta}_k$. Интенсивности обслуживания задаются как $\mu_k = -(\boldsymbol{\beta}_k S_k^{-1}\mathbf{e})^{-1}$.

Предполагаем, что заявки типа $k' < k$, $k, k' = \overline{1, K}$ обладают относительным приоритетом. Это означает, что более приоритетные заявки стоят в буфере впереди всех менее приоритетных заявок. Поступающая заявка типа k', заставшая все приборы занятыми и в очереди $i, i = \overline{0, R-1}$, заявок, с вероятностью $q_i^{(k')}$ становится впереди всех неприоритетных заявок типов $k = k' + 1, k' + 2, \ldots, K$ и в конце очереди приоритетных заявок типов $k = 1, 2, \ldots, k'$. С вероятностью $1 - q_i^{(k')}$ новая заявка решает не присоединяться к очереди и уходит из системы навсегда. Если заявка любого типа, поступающая в систему, застает систему полностью занятой, то она покидает систему навсегда. Для упрощения записи в случае $K = 2$ типов заявок будем обозначать

вероятности присоединения к очереди длины i приоритетных заявок класса $k = 1$ как $q_i \equiv q_i^{(1)}$, а неприоритетных заявок класса $k = 2$ – как $f_i \equiv q_i^{(2)}$.

Нашей целью является расчет стационарных вероятностей системы и характеристик производительности, включая вероятности потерь и время отклика системы.

3. Система $MMAP/PH/M/N$ с двумя классами заявок

3.1. Цепь Маркова, описывающая состояние системы

Пусть в момент времени t:

– i_t – количество заявок в буфере, $i_t = \overline{0, R}$;

– k_t – количество приоритетных заявок в буфере, $k_t = \overline{0, i_t}$;

– n_t – количество занятых приборов, $n_t = \overline{0, N}$;

– r_t – количество приборов, занятых обслуживанием приоритетных заявок, $r_t = \overline{0, n_t}$;

– $n_t^{(m_t^{(1)})}$ – количество приборов, обслуживающих приоритетные заявки на фазе $m_t^{(1)}$, $n_t^{(m_t^{(1)})} = \overline{0, r_t}$, $m_t^{(1)} = \overline{1, M_1}$;

– $\tilde{n}_t^{(m_t^{(2)})}$ – количество приборов, обслуживающих неприоритетные заявки на фазе $m_t^{(2)}$, $\tilde{n}_t^{(m_t^{(2)})} = \overline{0, n_t - r_t}$, $m_t^{(2)} = \overline{1, M_2}$;

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

Процесс изменения состояний системы описывается регулярной неприводимой цепью Маркова ξ_t с непрерывным временем

$$\xi_t = \{(i_t, k_t, r_t, n_t, \nu_t, n_t^{(1)}, n_t^{(2)}, \ldots, n_t^{(M_1)}, \tilde{n}_t^{(1)}, \tilde{n}_t^{(2)}, \ldots, \tilde{n}_t^{(M_2)}\}, \; t \geqslant 0,$$

и пространством состояний

$$\Omega = \{(i, n, r, \nu, n^{(1)}, n^{(2)}, \ldots, n^{(M_1)}, \tilde{n}^{(1)}, \tilde{n}^{(2)}, \ldots, \tilde{n}^{(M_2)}),$$

$$i = 0, n = \overline{0, N}, r = \overline{0, n}, \nu = \overline{0, W}$$

$$n^{(m)} = \overline{0, r}, m = \overline{1, M_1}, \tilde{n}^{(\tilde{m})} = \overline{0, n - r}, \tilde{m} = \overline{1, M_2}\} \bigcup$$

$$\bigcup\{(i,k,n,r,\nu,n^{(1)},n^{(2)},\ldots,n^{(M_1)},\tilde{n}^{(1)},\tilde{n}^{(2)},\ldots,\tilde{n}^{(M_2)}),$$

$$i=\overline{1,R},k=\overline{0,i},n=N,r=\overline{0,n},\nu=\overline{0,W},n^{(m^{(1)})}=\overline{0,r},$$

$$m^{(1)}=\overline{1,M_1},\tilde{n}^{(m^{(1)})}=\overline{0,N-r},m^{(2)}=\overline{1,M_2},$$

$$\sum_{m^{(1)}=1}^{M_1} n^{(m^{(1)})}=r,\ \sum_{m^{(2)}=1}^{M_2}\tilde{n}^{(m^{(2)})}=N-r\}$$

Количество векторов пространства состояний при $i=0$ есть

$$K_0=(W+1)\sum_{n=0}^{N}\sum_{r=0}^{n}C_{r+M_1-1}^{M_1-1}C_{n-r+M_2-1}^{M_2-1}$$

и для любого фиксированного $i=\overline{1,R}$ количество векторов есть

$$K=(i+1)(W+1)\sum_{r=0}^{N}C_{r+M_1-1}^{M_1-1}C_{N-r+M_2-1}^{M_2-1}.$$

В дальнейшем будем использовать следующие обозначения:

– I (O) – тождественная (нулевая) матрица подходящего порядка;

– $\bar{W}=W+1$;

– $\otimes$ ($\oplus$) – символ кронекерова произведения (суммы) матриц, см., например, [8];

– $C_n^l=\frac{n!}{l!(n-l)!}$;

– $d_r^{(1)}=C_{r+M_1-1}^{M_1-1}$;

– $d_r^{(2)}=C_{r+M_2-1}^{M_2-1}$;

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

– $diag^{+}\{a_1,a_2,...,a_n\}$ – квадратная блочная матрица, у которой наддиагональные блоки равны элементам, перечисленным

в скобках, а остальные блоки нулевые, т.е. это матрица вида

$$\begin{pmatrix} 0 & a_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & a_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & a_n \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix};$$

– $diag^{-}\{a_1, a_2, ..., a_n\}$ – квадратная блочная матрица, у которой поддиагональные блоки равны элементам, перечисленными в скобках, а остальные блоки нулевые, т.е. это матрица вида

$$\begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ a_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & a_2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_n & 0 \end{pmatrix};$$

– $\bar{q}_i = 1 - q_i$, $\bar{f}_i = 1 - f_i$;

– $\mathbf{n}_t^{(1)} = \{n_t^{(1)}, n_t^{(2)}, \dots, n_t^{(M_1)}\}$ – процесс обслуживания заявок 1 типа;

– $\mathbf{n}_t^{(2)} = \{\tilde{n}_t^{(1)}, \dots, \tilde{n}_t^{(M_2)}\}$ – процесс обслуживания заявок 2 типа;

С двумя последними обозначениями рассматриваемая цепь Маркова ξ_t, $t \geqslant 0$, может быть записана в виде

$$\xi_t = \{(i_t, k_t, r_t, n_t, \nu_t, \mathbf{n}_t^{(1)}, \mathbf{n}_t^{(2)}\},\ t \geqslant 0.$$

В дальнейшем будем предполагать, что состояния цепи Маркова ξ_t, $t \geqslant 0$, упорядочены следующим образом: первые четыре компоненты $\{i_t, k_t, r_t, \nu_t\}$ упорядочены в прямом лексикографическом порядке, а состояния каждого из процессов $\mathbf{n}_t$ и $\tilde{\mathbf{n}}_t$ упорядочены в обратном лексикографическом порядке. Упорядочение в обратном лексикографическом порядке требуется для дальнейшего описания интенсивностей переходов процессов $\mathbf{n}_t$ и $\tilde{\mathbf{n}}_t$ с использованием введенных в статьях [21, 18] матриц $P_i(\cdot)$, $A_i(\cdot,\cdot)$, и $L_i(\cdot,\cdot)$.

Дадим краткое объяснение вероятностных значений этих матриц. Введем в рассмотрение матрицы $\tilde{S}_l = \begin{pmatrix} 0 & O \\ \boldsymbol{S}_0^{(l)} & S_l \end{pmatrix}$, $l = 1, 2$, тогда:

– $L_k(n, \tilde{S}_l)$ – матрица порядка $C_{n-k+M_l-1}^{M_l-1} \times C_{n-k+M_l-2}^{M_l-1}$. Матрица содержит интенсивности переходов процесса $\mathbf{n}_t^{(l)}$, приводящих к освобождению одного из $n-k$ приборов, которые обслуживают заявки l-го типа (k есть число свободных приборов, n – общее число свободных приборов и приборов, которые обслуживают заявки l-го типа);

– $P_n(\boldsymbol{\beta}_l)$ – матрица порядка $C_{n+M_l-1}^{M_l-1} \times C_{n+M_l}^{M_l-1}$. Матрица содержит интенсивности переходов процесса $\mathbf{n}_t^{(l)}$, приводящих к увеличению числа приборов, которые обслуживают заявки l-го типа, с n до $n+1$;

– $A_n(k, \tilde{S}_l)$ – матрица порядка $C_{n+M_l-1}^{M_l-1} \times C_{n+M_l}^{M_l-1}$. Матрица содержит интенсивности переходов процесса $\mathbf{n}_t^{(l)}$ в его пространстве состояний без увеличения или уменьшения числа приборов, которые обслуживают заявки l-го типа (n есть число приборов, которые обслуживают заявки l-го типа, k – общее число свободных приборов и приборов, которые обслуживают заявки l-го типа).

В дальнейшем полагаем $L_0(0) = A_0(\cdot) = P_{-1}(\cdot) = 0$. Алгоритм вычисления матриц $P_i(\cdot)$, $A_i(\cdot,\cdot)$, и $L_i(\cdot,\cdot)$ следует из результатов В. Рамасвами и Д. Лукантони, опубликованных в статьях [18, 21]. Четко по шагам этот алгоритм описан в [6]. Соответствующее описание мы приводим далее в разделе 3.4.

Теорема 1. Инфинитезимальный генератор цепи Маркова ξ_t имеет следующую блочную структуру:

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

где ненулевые блоки T, $Q_{i,i'}$ определяются следующим образом.

1. Блок $T = (T_{n,n'})_{n,n'=\overline{0,N}} + \Delta^{(0)}$ есть блочная трехдиагональная матрица, где

а) $T_{n,n}, n = \overline{0,N-1}$, – квадратная матрица порядка $\bar{W}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)}$,

$$T_{n,n} = diag\{D_0 \oplus A_r(N-n+r, S_1) \oplus A_{n-r}((N-r, S_2),\ r = \overline{0,n}\},$$

$$n = \overline{0,N-1},$$

б) $T_{N,N}$ – квадратная матрица порядка $\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$,

$$T_{N,N} = diag\{\bar{D} \oplus A_r(r, S_1) \oplus A_{N-r}((N-r, S_2),\ r = \overline{0,N}\},$$

$$\bar{D} = D_0 + \bar{q}_0 D_1 + \bar{f}_0 D_2,$$

в) $T_{n,n-1}$ – матрица порядка $\bar{W}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} \times \bar{W}\sum\limits_{r=0}^{n-1} d_r^{(1)} d_{n-r-1}^{(2)}$,

$$T_{n,n-1} = \left(\begin{array}{c|c} diag\{I_{\bar{W}d_r^{(1)}} \otimes L_{N-n}(N-r, \tilde{S}_2),\ r = \overline{0,n-1}\} & \\ \hline O_{\bar{W}d_n^{(1)} \times \bar{W}\sum\limits_{r=0}^{n-2} d_r^{(1)} d_{n-r-1}^{(2)}} & I_{\bar{W}} \otimes L_{N-n}(N, \tilde{S}_1) \end{array} \right) +$$

$$+ \left(\begin{array}{c} diag^{-}\{I_{\bar{W}} \otimes L_{N-n}(N-n+r, \tilde{S}_1) \otimes I_{d_{n-r}^{(2)}}\ r = \overline{1,n-1}\} \\ \hline O_{\bar{W}d_n^{(1)} \times \bar{W}\sum\limits_{r=0}^{n-1} d_r^{(1)} d_{n-r-1}^{(2)}} \end{array} \right),$$

$$n = \overline{1,N},$$

г) $T_{n,n+1}$ – матрица порядка $\bar{W}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} \times \bar{W}\sum\limits_{r=0}^{n+1} d_r^{(1)} d_{n-r+1}^{(2)}$,

$$T_{n,n+1} = \left(diag\{D_2 \otimes I_{d_r^{(1)}} \otimes P_{n-r}(\boldsymbol{\beta}_2), r = \overline{0,n}\} \mid O_{\bar{W}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} \times \bar{W} d_{n+1}^{(1)}} \right) +$$

$$+ \left(\begin{array}{c|c} O_{\bar{W}\sum\limits_{r=0}^{n-1} d_r^{(1)} d_{n-r}^{(2)} \times \bar{W}\sum\limits_{r=0}^{n+1} d_r^{(1)} d_{n-r+1}^{(2)}} & \\ \hline O_{\bar{W}d_n^{(1)} \times \bar{W}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r+1}^{(2)}} & |D_1 \otimes P_n(\boldsymbol{\beta}_1) \end{array} \right) +$$

$$+\left(\begin{array}{c} diag^{+}\{D_1 \otimes P_r(\boldsymbol{\beta}_1) \otimes I_{d_{n-r}^{(2)}},\ r=\overline{0,n-1}\} \mid O_{\bar{W}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} \times \bar{W} d_{n+1}^{(1)}} \end{array} \right),$$

$$n=\overline{0,N-1},$$

2. Блок $Q_{0,1}$ – это матрица порядка $\bar{W}\sum\limits_{n=0}^{N}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} \times 2\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$, которая имеет вид:

$$Q_{0,1}=\left(\begin{array}{c} O_{\bar{W}\sum\limits_{n=0}^{N-1}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} \times 2\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}} \\ \hline H_1 \quad \mid \quad H_2 \end{array} \right)$$

где квадратные матрицы H_1, H_2 порядка $\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$ имеют вид

$$H_1=f_0 diag\{D_2 \otimes I_{d_r^{(1)} d_{N-r}^{(2)}},\ r=\overline{0,N}\},$$

$$H_2=q_0 diag\{D_1 \otimes I_{d_r^{(1)} d_{N-r}^{(2)}},\ r=\overline{0,N}\},$$

3. Блок $Q_{1,0}$ – это матрица порядка $2\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} \times \bar{W}\sum\limits_{n=0}^{N}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)}$, которая имеет вид:

$$Q_{1,0}=\left(\begin{array}{cc} O_{\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} \times \bar{W}\sum\limits_{n=0}^{N-1}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)}} & F_1 \\ O_{\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} \times \bar{W}\sum\limits_{n=0}^{N-1}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)}} & F_2 \end{array} \right),$$

где квадратные матрицы F_1, F_2 порядка $\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$ имеют вид

$$F_1=diag\{I_{\bar{W} d_r^{(1)}} \otimes L_0(N-r,\tilde{S}_2) P_{N-r-1}(\boldsymbol{\beta}_2), r=\overline{0,N-1}, O_{\bar{W} d_N^{(1)} d_0^{(2)}}\}+$$

$$+diag^{-}\{I_{\bar{W}} \otimes L_0(r,\tilde{S}_1) \otimes P_{N-r}(\boldsymbol{\beta}_2), r=\overline{1,N}\},$$

$$F_2 = diag\{O_{\bar{W}d_N^{(2)}}, I_{\bar{W}} \otimes L_0(r, \tilde{S}_1)P_{r-1}(\boldsymbol{\beta}_1) \otimes I_{d_{N-r}^{(2)}}, r = \overline{1, N}\}+$$

$$+diag^{+}\{I_{\bar{W}} \otimes P_r(\boldsymbol{\beta}_1) \otimes L_0(N-r, \tilde{S}_2), r = \overline{0, N-1}\}.$$

4. Блок $Q_{i,i-1}, i = \overline{2, R}$, порядка $(i+1)\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} \times i\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$ имеет вид

$$Q_{i,i-1} = \left(\begin{array}{cc} F_1 & O_{\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} \times (i-1)\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}} \\ \hline I_i \otimes F_2 & \end{array} \right), i = \overline{2, R},$$

5. Блок $Q_{i,i}, i = \overline{1, R-1}$, – квадратная матрица порядка $(i+1)\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$, которая имеет вид:

$$Q_{i,i} = I_{i+1} \otimes T_{N,N} + \Delta^{(i)}, \; i = \overline{1, R-1}.$$

6. Блок $Q_{R,R}$ – квадратная матрица порядка $(R+1)\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$, которая имеет вид:

$$Q_{R,R} = I_{R+1} \otimes \hat{T}_{N,N} + \Delta^{(R)},$$

где матрица $\hat{T}_{N,N}$ получена из матрицы $T_{N,N}$ путем замены кронекеровых множителей $\bar{D}$ на множители $D(1)$.

7. Блок $Q_{i,i+1}$ порядка $(i+1)\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} \times (i+2)\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$ имеет вид:

$$Q_{i,i+1} =$$

$$= \left(I_{i+1} \otimes diag\{f_i D_2 \otimes I_{d_r^{(1)} d_{N-r}^{(2)}}, r = \overline{0, N}\} \mid O_{(i+1)\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} \times \bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}} \right) +$$

$$+\hat{I}_{i+1} \otimes diag\{q_i D_1 \otimes I_{d_r^{(1)} d_{N-r}^{(2)}}, r = \overline{0, N}\}, i = \overline{0, R-1},$$

где $\hat{I}_{i+1}$ – матрица порядка $(i+1)\times(i+2)$ вида

$$\begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Здесь матрицы $\Delta^{(i)}, i = \overline{0,R}$– это диагональные матрицы, образованные таким образом, что выполняется равенство $Q\mathbf{e} = \mathbf{0}^{\top}$. Поясним более подробно, как можно вычислить матрицы $\Delta_i, i = \overline{0,R}$. Пусть $i = 0$. Вычислим вектор-столбец $T\mathbf{e} + Q_{0,1}\mathbf{e}$. Тогда элементы этого вектора, взятые со знаком минус, являются диагональными элементами матрицы Δ_0. Аналогично, чтобы сформировать матрицу Δ_i для $i = \overline{1,R}$, нужно вычислить вектор-столбец $Q_{i,i-1}\mathbf{e} + Q_{i,i}\mathbf{e} + Q_{i,i+1}\mathbf{e}$. Тогда элементы этого вектора, взятые со знаком минус, являются диагональными элементами матрицы Δ_i.

3.2. Стационарное распределение

Пусть $\mathbf{p}$ является вектором-строкой стационарного распределения вероятностей состояний цепи. Этот вектор определяется как единственное решение системы линейных алгебраических уравнений

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

В случае большой размерности системы (1) для ее решения целесообразно использовать алгоритм, описанный в [15]. Для этого представим вектор $\mathbf{p}$ как $\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, \dots, \mathbf{p}_R)$, где вектор $\mathbf{p}_0$ имеет порядок $\bar{W}\sum\limits_{n=0}^{N}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)}$, а вектор $\mathbf{p}_i$ имеет порядок $(i+1)\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$, $i = \overline{1,R}$.

1. Находим матрицы $G_{R-1}, G_{R-2}, \dots, G_0$, из уравнения обратной рекурсии:

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

где полагаем

$$G_{R-1} = (-Q_{R,R})^{-1} Q_{R,R-1}.$$

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

$$\overline{Q}_{R,R} = Q_{R,R},$$

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

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

3. Находим матрицы Φ_i из рекуррентных соотношений:

$$\Phi_0 = I, \Phi_i = \Phi_{i-1} \bar{Q}_{i-1,i} \left(-\bar{Q}_{i,i}\right)^{-1},\ i = \overline{1, R}.$$

4. Вычисляем вектор $\mathbf{p}_0$ как единственное решение СЛАУ:

$$\mathbf{p}_0(-\bar{Q}_{0,0}) = 0, \tag{2}$$

$$\mathbf{p}_0(\mathbf{e} + \sum_{i=1}^{R} \Phi_i \mathbf{e}) = 1. \tag{3}$$

5. Вычисляем векторы $\mathbf{p}_i$ по формулам $\mathbf{p}_i = \mathbf{p}_0 \Phi_i, i = \overline{1, R}$.

Замечание 1. При решении системы линейных алгебраических уравнений (2)–(3) надо знать, что система (2) имеет ранг на единицу меньший ее размерности. Однако, заменив одно из уравнений, например, первое, этой системы на уравнение (3), получим систему, которая имеет единственное решение. Такую модификацию надо сделать, прежде чем решать систему.

Замечание 2. Чтобы определить стационарное распределение, надо решить систему линейных алгебраических уравнений (1) или, расписывая по блокам $Q_{i,j}$,

$$\sum_{j=max\{i-1,0\}}^{min\{i+1,R\}} \mathbf{p}_i Q_{i,j} = \mathbf{0}, i = \overline{0, R},\ \sum_{i=0}^{R} \mathbf{p}_i \mathbf{e} = 1. \tag{4}$$

Однако ранг этой системы равен $\bar{W} \sum\limits_{n=0}^{N} \sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} +$ $+ \sum\limits_{i=1}^{R} (i+1) \bar{W} \sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$ и при больших значениях R, N и размерностях $MMAP$ и PH ранг этой системы становится настолько большим, что не представляется возможностей решить эту систему непосредственно, например, используя обратную матрицу.

Поэтому представляется целесообразным использовать описанный выше алгоритм, в котором максимальный размер блока, используемого в вычислениях, равен $(R+1)\bar{W}\sum\limits_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}$. Однако при малых значениях упомянутых параметров систему (1) (или (4)) можно использовать для проверки решения, полученного при помощи вышеприведенного алгоритма.

3.3. Характеристики производительности системы

Рассчитав стационарное распределение, можно найти ряд важных стационарных характеристик производительности системы. Приведем некоторые из них.

– Вектор вероятностей того, что буфер пустой и заняты n приборов:

$$\mathbf{p}_0(n) = \mathbf{p}_0 U(n),$$

где матрица $U(n)$ имеет вид:

$$U(n) = \begin{pmatrix} O & & \\ \bar{W}\sum\limits_{l=0}^{n-1}\sum\limits_{m=0}^{l} d_m^{(1)} d_{l-m}^{(2)} \times \bar{W}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} & & \\ & I & \\ & \bar{W}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} & \\ O & & \\ \bar{W}\sum\limits_{l=n+1}^{N}\sum\limits_{m=0}^{l} d_m^{(1)} d_{l-m}^{(2)} \times \bar{W}\sum\limits_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} & & \end{pmatrix}, \; n = \overline{0, N}.$$

– Вероятность того, что буфер пустой, заняты n приборов и r из них заняты приоритетными заявками:

$$p_0(n, r) = \mathbf{p}(n)\mathbf{u}(n, r), \; n = \overline{0, N}, r = \overline{0, n},$$

где вектор-столбец $\mathbf{u}(n, r)$ имеет вид:

$$\mathbf{u}(n, r) = \begin{pmatrix} \mathbf{0}^T & \\ \bar{W}\sum\limits_{l=0}^{r-1} d_l^{(1)} d_{n-l}^{(2)} & \\ \mathbf{e}_{\bar{W} d_r^{(1)} d_{n-r}^{(2)}} & \\ \mathbf{0}^T & \\ \bar{W}\sum\limits_{l=r+1}^{n} d_l^{(1)} d_{n-l}^{(2)} & \end{pmatrix}.$$

– Вероятность того, что буфер пустой и заняты n приборов:

$$p_0(n) = \sum_{r=0}^{n} p_0(n,r),\ n = \overline{0,N}.$$

– Вероятность того, что в буфере находятся i заявок, из них k приоритетных, и r приоритетных заявок находятся на обслуживании:

$$p_i(k,r) = \mathbf{p}_i \begin{pmatrix} \mathbf{0}^T \\ k\bar{W} \sum\limits_{n=0}^{N} d_n^{(1)} d_{N-n}^{(2)} \\ \mathbf{u}(N,r) \\ \mathbf{0}^T \\ (i-k)\bar{W} \sum\limits_{n=0}^{N} d_n^{(1)} d_{N-n}^{(2)} \end{pmatrix},\ i = \overline{1,R},\ k = \overline{0,i},\ r = \overline{0,N}.$$

– Вероятность того, что в буфере находятся i заявок, из них k приоритетных:

$$p_i(k) = \sum_{r=0}^{N} p_i(k,r),\ i = \overline{1,R},\ k = \overline{0,i}.$$

– Вероятность того, что в буфере находятся i заявок:

$$p_i = \mathbf{p}_i \mathbf{e}, i = \overline{1,R}.$$

Должно выполняться равенство $p_i = \sum\limits_{k=0}^{i} p_i(k)$.

– Среднее число заявок в системе:

$$\bar{N}_{system} = \sum_{n=1}^{N} n p_0(n) + \sum_{i=1}^{R} (i+N) p_i.$$

– Среднее число приоритетных заявок в системе:

$$\bar{N}_{system}^{(1)} = \sum_{n=1}^{N} \sum_{r=1}^{n} r p_0(n,r) + \sum_{n=1}^{N} \sum_{i=1}^{R} \sum_{k=1}^{i} (r+k) p_i(k,r).$$

– Среднее число занятых приборов:

$$\bar{N}_{servers} = \sum_{n=1}^{N} np_0(n) + N\sum_{i=1}^{R} p_i.$$

– Среднее число приборов, занятых обслуживанием приоритетных заявок:

$$\bar{N}^{(1)}_{servers} = \sum_{n=1}^{N}\sum_{r=1}^{n} rp_0(n,r) + \sum_{i=1}^{R}\sum_{k=0}^{i}\sum_{r=1}^{N} rp_i(k,r).$$

– Среднее число заявок в буфере:

$$\bar{N}_{queue} = \sum_{i=1}^{R} ip_i.$$

3.4. Вычисление матриц $P_i(\cdot)$, $A_i(\cdot,\cdot)$, и $L_i(\cdot,\cdot)$.

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

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

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

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

$$L_i^{(w)}(T_j) = \begin{pmatrix} t_0 & O & O & \cdots & O \\ l_0 & t_1 & O & \cdots & O \\ O & l_1 & t_2 & O & O \\ \vdots & \vdots & \ddots & \ddots & O \\ \vdots & \vdots & \ddots & l_{N-i-2} & t_{N-i-1} \\ O & \cdots & O & \cdots & l_{N-i-1} \end{pmatrix},$$

где блоки матрицы равны $t_k = (N-i-k)t^j_{r_j-w,1}I$, $l_k = L^{(w-1)}_{N-1-k}(T_j)$, где $k \in \{0,\ldots,N-i-1\}$, $w \in \{1,\ldots,r_j-2\}$,

$i \in \{0, \ldots, N-1\}$, $j \in \{1, \ldots, M-2\}$, где $t^j_{k,l}$ есть (k,l)-й элемент матрицы T_j и r_j есть число строк матрицы T_j.

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

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

$$U_i^{(w)}(T_j) = \begin{pmatrix} t_0 & u_0 & O & \cdots & O \\ O & t_1 & u_1 & O & O \\ \vdots & \vdots & \ddots & \ddots & O \\ \vdots & \vdots & \ddots & \ddots & O \\ \vdots & \vdots & t_{N-i-1} & u_{N-i-1} & O \\ O & O & O & t_{N-i} & u_{N-i} \end{pmatrix},$$

элементами матрицы являются блоки: $t_k = t^j_{1,r_j-w} I$, $u_k = U_{N-k}^{(w-1)}(T_j)$, где $k \in \{0, \ldots, N-i\}$.

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

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

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

$$i \in \{1, \ldots, N\},$$

$$A_i^{(j)} = \begin{pmatrix} O & u_0 & O & \cdots & O \\ l_0 & a_0 & u_1 & \cdots & O \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ O & \ddots & l_{i-2} & a_{i-2} & u_{i-1} \\ O & \cdots & \cdots & l_{i-1} & a_{i-1} \end{pmatrix},$$

где элементы данной матрицы: $l_k = L_{N-k-1}(N, T_j)$, $a_k = A_{k+1}^{j-1}$, $u_k = \frac{(i-k)U_{N-k}(N,T_j)}{N-k}$, где $k \in \{0, \ldots, i-1\}$,

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

8. При вычислении матриц $L_i(N, \tilde{S})$ мы идем на шаг 3, где игнорируем матрицы T_j и вместо них рассматриваем одну матрицу $\tilde{S}$. Также заменяем M на $\tilde{M} = M + 1$, поскольку порядок матрицы $\tilde{S}$ на единицу больше размерности матрицы S.

Искомые матрицы $L_i(N, \tilde{S})$ вычисляем следующим образом: $L_i(N, \tilde{S}) = L_i^{(M-1)}(\tilde{S})$, $i \in \{0, \ldots, N-1\}$, $L_N(N, \tilde{S}) = O_{1\times 1}$. Здесь в верхнем индексе уже имеется в виду, что M – порядок исходной матрицы S.

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

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

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

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

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

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

4. Использование метода Монте-Карло для расчёта характеристик приоритетной системы

Аналитическое решение построено только для приоритетных систем с двумя классами трафика. Для систем в общем

виде практически невозможно построить генератор цепи Маркова, найти стационарное распределение и вычислить все необходимые характеристики производительности системы. При числе классов большем двух оценки характеристик удается найти только с помощью численных методов. В настоящей статье для расчета характеристик используется имитационная модель СМО, реализующая метод Монте-Карло. Алгоритм для расчета аналитической модели и интерфейс имитационной модели были реализованы на языке Python с использованием библиотеки numpy. Ядро имитационной модели было разработано на языке C++, а для того чтобы иметь возможность работать с ней из Python, использовалась библиотека Cython. Исходный код доступен на GitLab по адресу https://gitlab.com/lab69/priority-queues.

4.1. Влияние входных параметров на время расчёта аналитической модели

На практике возможность получения численных результатов даже для двухприоритетной системы ($K = 2$) с помощью аналитической модели оказывается ограниченной из-за быстрого роста порядка матрицы инфинитезимального генератора Q. Порядок матрицы Q можно оценить по ее диагональным элементам. Для этого просуммируем порядки каждого из блоков матрицы Q: $T, Q_{1,1}, Q_{2,2}, ..., Q_{R,R}$. Обозначим $P(T)$ - порядок матрицы T, а $P_{Q_{sum}}$ – суммарный порядок матриц $Q_{1,1}, Q_{2,2}, .., Q_{R,R}$:

$$P_T = \bar{W} \sum_{n=0}^{N-1} \sum_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} + \bar{W} \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)},$$

$$P_{Q_{sum}} = \bar{W} \sum_{i=1}^{R-1} (i+1) \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} + (R+1)\bar{W} \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} =$$

$$= \bar{W} \sum_{i=1}^{R} (i+1) \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} = \frac{R(R+3)}{2} \bar{W} \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)}.$$

Суммарный порядок P генератора Q будет равен

$$P = P_T + P_{Q_{sum}} = \bar{W} \sum_{n=0}^{N-1} \sum_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} + \bar{W} \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} +$$

$$+ \frac{R(R+3)}{2} \bar{W} \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} =$$

$$= \bar{W} \sum_{n=0}^{N-1} \sum_{r=0}^{n} d_r^{(1)} d_{n-r}^{(2)} + \frac{(R+1)(R+2)}{2} \bar{W} \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)},$$

где

$$d_r^{(1)} = C_{r+M_1-1}^{M_1-1} = \frac{(r+M_1-1)!}{(M_1-1)!r!},$$

$$d_r^{(2)} = C_{r+M_2-1}^{M_2-1} = \frac{(r+M_2-1)!}{(M_2-1)!r!}.$$

$$d_{N-r}^{(2)} = C_{N-r+M_2-1}^{M_2-1} = \frac{(N-r+M_2-1)!}{(M_2-1)!(N-r)!}.$$

Приведем оценку сверху данного выражения. Максимум первого слагаемого достигается при $n = N$, заменим все слагаемые максимальным значением:

$$P \leqslant \bar{W}(N+1) \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} + \frac{(R+1)(R+2)}{2} \bar{W} \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} =$$

$$= \frac{(R+1)(R+2)+2N+2}{2} \bar{W} \sum_{r=0}^{N} d_r^{(1)} d_{N-r}^{(2)} \leqslant$$

$$\leqslant \frac{(R+1)(R+2)+2N+2}{2} \bar{W}(N+1) \max(d_r^{(1)} d_{N-r}^{(2)})$$

Исходя из данного выражения, ограничивающего размер инфинитезимального генератора сверху, видим, что порядок квадратично зависит от размера буфера $P \sim O(R^2)$, зависит линейно

от порядка входного $MMAP$-потока $P \sim O(W)$. Оценим вклад множителя $\max(d_r^{(1)} d_{N-r}^{(2)})$. Для этого рассмотрим функцию $p(r)$:

$$p(r) = d_r^{(1)} d_{N-r}^{(2)} = \frac{(r+M)!(N-r+M)!}{(M!)^2 r!(N-r)!}.$$

Для оценки факториалов воспользуемся формулой Стирлинга:

$$p(r) = d_r^{(1)} d_{N-r}^{(2)} \approx \frac{(r+M)^{r+M+\frac{1}{2}} (N-r+M)^{N-r+M+\frac{1}{2}}}{M^{2M+1} r^{r+\frac{1}{2}} (N-r)^{N-r+\frac{1}{2}}} =$$

$$= \frac{e^{(r+M+\frac{1}{2})\ln(r+M)} e^{(r+M+\frac{1}{2})\ln(N-r+M)}}{M^{2M+1} e^{(r+\frac{1}{2})\ln(r)} e^{(N-r+\frac{1}{2})\ln(N-r)}}.$$

Найдем точки экстремума $p(r)$. Вычислим производную числителя, обозначим $a(r) = e^{(r+M+\frac{1}{2})\ln(r+M)}$ и $b(r) = e^{(N-r+M+\frac{1}{2})\ln(N-r+M)}$, тогда

$$a'(r) = a(r)\left(\ln(r+M) + \frac{r+M+\frac{1}{2}}{r+M}\right)$$

$$b'(r) = b(r)\left(-\ln(N-r+M) - \frac{(N-r+M+\frac{1}{2})}{N-r+M}\right)$$

$$u'(r) = (a(r)b(r))' = a'(r)b(r) + a(r)b'(r) =$$

$$= a(r)b(r)\left(\ln\frac{(r+M)}{(N-r+m)} + \frac{1}{2(r+M)} - \frac{1}{2(N-r+M)}\right)$$

Вычислим производную знаменателя. Пусть $c(r) = e^{(r+\frac{1}{2})\ln(r)}$ и $d(r) = e^{(N-r+\frac{1}{2})\ln(N-r)}$, тогда

$$v'(r) = (c(r)d(r))' = c(r)'d(r) + c(r)d(r)' =$$

$$= c(r)d(r)\left(\ln r + \frac{r+\frac{1}{2}}{r} - \ln(N-r) - \frac{(N-r+\frac{1}{2})}{N-r}\right) =$$

$$= c(r)d(r)\left(\ln\frac{r}{N-r} + \frac{1}{2(r)} - \frac{1}{2(N-r)}\right).$$

Вычисляем производную по формуле:

$$p'(r) = \frac{u(r)'v(r) - u(r)v(r)'}{v^2(r)} =$$

$$= \frac{a(r)b(r)c(r)d(r)\left(\ln\frac{(r+M)(N-r)}{(N-r+M)r} + \frac{(2r-N)M(M+N)}{(r+M)r(N+M-r)(N-r)}\right)}{(c(r)d(r))^2} =$$

$$= \frac{a(r)b(r)c(r)d(r)(f(r)+g(r))}{(c(r)d(r))^2},$$

где $f(r) = \ln\frac{(r+M)(N-r)}{(N-r+M)r}$ и $g(r) = \frac{(2r-N)M(M+N)}{(r+M)r(N+M-r)(N-r)}$. Приравняем производную к 0 и найдем точку экстремума. Так как функции $a(r), b(r), c(r), d(r)$ положительны на интервале $(0, N)$, то производная рана 0, когда $f(r) + g(r) = 0$. Получаем

$$\ln\frac{(r+M)(N-r)}{(N-r+M)r} + \frac{(2r-N)M(M+N)}{(r+M)r(N+M-r)(N-r)} = 0.$$

Обратим внимание, что $r = \frac{N}{2}$ – корень уравнения, так как $f(\frac{N}{2}) = 0$ и $g(\frac{N}{2}) = 0$. Докажем, что больше корней у данного уравнения на интервале $(0, N)$ нет. Функция $f(r)$ на интервале $(0, \frac{N}{2})$ отрицательна, на интервале $(\frac{N}{2}, N)$ положительна и $f(\frac{N}{2}) = 0$. Аналогичными свойствами обладает функция $g(r)$. Таким образом, на $(0, \frac{N}{2})$ функция $h(r) < 0$, в точке $\frac{N}{2}$ равна нулю и на интервале $(\frac{N}{2}, N)$ функция $h(r) > 0$. Единственной точкой экстремума является $\frac{N}{2}$.

Значение в точке $r = \frac{N}{2}$ является локальным минимумом, так как $h(r)$ меняет знак с минуса на плюс. Таким образом, максимум будем искать на краевых точках интервала $(0, N)$. $p(0) = p(N) = \frac{(N+M)^{(N+M+\frac{1}{2})}}{M^{M+\frac{1}{2}} N^{N+\frac{1}{2}}}$ – значения функции в точках $0, N$ равны. Оценим влияние N и M на величину данного выражения и оценим $O(N, M)$. Подставляя максимум в формулу для порядка генератора, получим оценку сверху:

$$\max(d_r^{(1)} d_{N-r}^{(2)}) = \frac{(N+M)^{(N+M+\frac{1}{2})}}{M^{M+\frac{1}{2}} N^{N+\frac{1}{2}}},$$

$$P \leqslant \frac{(R+1)(R+2)+2N+2}{2} \bar{W}(N+1) \frac{(N+M)^{(N+M+\frac{1}{2})}}{M^{M+\frac{1}{2}} N^{N+\frac{1}{2}}}$$

В качестве иллюстрации влияния входных параметров на размер системы в таблицеЁ1 приведено время, требуемое для $MMAP$-потока порядка 5, PH-распределения порядка 5 и очереди емкости 1, на компьютере, имеющем процессор Intel Core i7-9750H и 16 гб ОЗУ. При числе обслуживающих приборов $N = 5$ не хватило ресурсов персонального компьютера для получения численного результата.

Таблица 1. Зависимость времени построения генератора от количества приборов

Число приборов	**Время, сек**
1	0,027
2	1,184
3	34,485
4	1155,581
5	Недостаточно памяти

Из рис. 1 видно, что размер матрицы генератора, описывающего процесс обработки заявок в приоритетной системе, экспоненциально зависит от порядка матрицы генератора $MMAP$-распределения и от количества обслуживающих приборов, кро-

ме того линейно зависит от порядка матрицы PH-распределения и квадратично – от размера буфера.

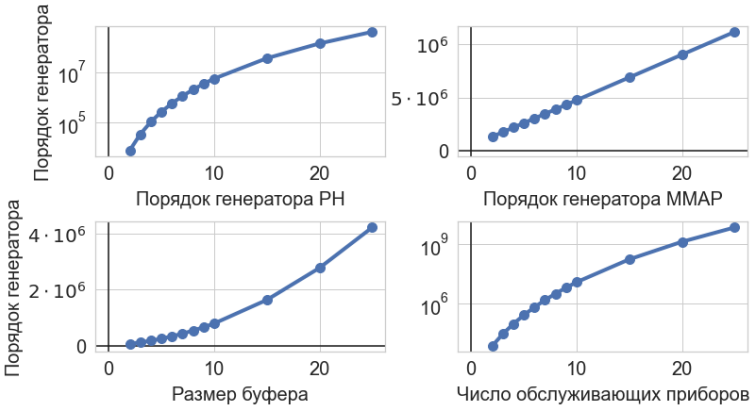

Рис. 1. Зависимость размера матрицы генератора в зависимости от различных входных параметров системы

4.2. Численные результаты

Численный эксперимент состоял из двух этапов. На первом этапе мы произвели валидацию имитационной модели, сравнив результаты ее расчетов с аналитическими расчетами на одних и тех же входных параметрах. На втором этапе мы исследовали влияние различных характеристик СМО на вероятность потери, время отклика и среднее число заявок в системе для заявок приоритетных и неприоритетных классов.

Для валидации были использованы матрицы небольшой размерности. Итоговые характеристики можно было рассчитать вручную. С помощью этих примеров валидировалась аналитическая модель, а уже с помощью аналитической модели валидировалась имитационная модель на различных примерах с ограничением на количество классов, $K \leqslant 2$.

Для сравнения результатов расчетов, полученных аналитически и с помощью имитационного моделирования, на вход подавались одинаковые входные данные. Так как число клас-

сов в аналитической модели $K = 2$, $MMAP$-поток задается матрицами D_0, D_1, D_2, а обслуживание задается двумя PH-распределениями с матрицами S_1, S_2 и векторами начальных состояний β_1, β_2. Вероятности присоединения задаются двумя векторами f, q для приоритетных и неприоритетных заявок соответственно. Пример входных параметров:

$$D_0 = \begin{pmatrix} -22 & 1 & 7 \\ 0 & -23 & 5 \\ 0 & 0 & -2 \end{pmatrix}, D_1 = D_2 = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 3 & 5 \\ 1 & 0 & 0 \end{pmatrix},$$

$$S_1 = S_2 = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \beta_1 = \beta_2 = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{pmatrix}$$

$$R = 5, N = 2, f = q = [1, 1, 1, 1, 1], q = [1, 1, 1, 1, 1].$$

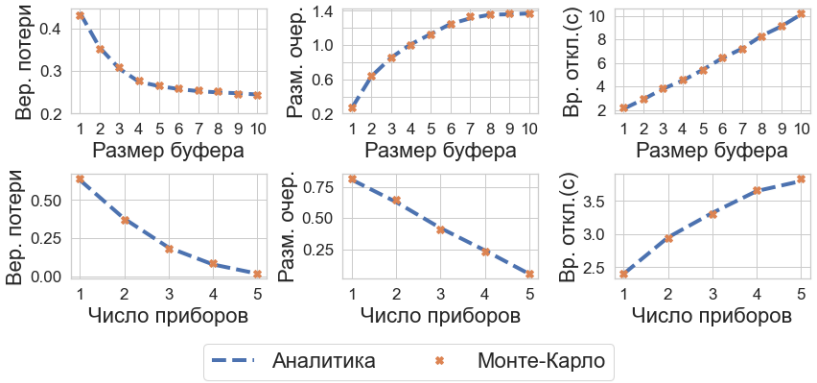

Рис. 2. Валидация имитационной модели. Сравнение результатов выполнения имитационной модели и аналитического решения

На рис. 2 показан результат сравнения работы имитационной модели и аналитических расчетов в зависимости от изменяющихся параметров: размера очереди и числа обслуживающих

приборов. Точность результата, полученного с помощью метода Монте-Карло зависит от количества сгенерированных заявок. В данном случае мы запускали симуляцию на десяти миллионах заявок, погрешность составила не более одного процента для каждого параметра.

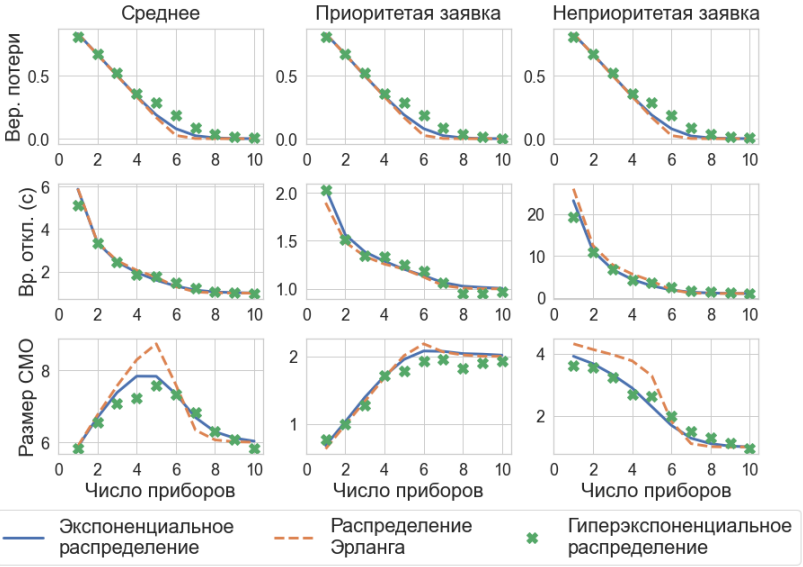

Рис. 3. Зависимость различных параметров системы от количества обслуживающих приборов

На втором этапе мы исследовали систему с $K = 5$ классами. В качестве входящего потока мы использовали $MMAP$, в котором интенсивность поступления заявок для приоритетных заявок в два раза больше чем для неприоритетных ($\lambda_1 = 2, \lambda_2 = 1$), коэффициент вариации $c_v = 0{,}5$, коэффициент асимметрии $\gamma = 0{,}5$, лаг распределения $\rho_1 = 0{,}1$. В качестве PH-распределений времени обработки заявок были исследованы распределение Эрланга с параметрами $\lambda = 1, c_v = 0{,}1$, экспоненциальное с параметром $\lambda = 1$ и гиперэкспоненциальное распределение с параметрами $\lambda = 1, c_v = 25$. Для восстановления $MMAP$ и построения

матриц D_i и восстановления PH по трем моментам мы использовали метод, описанный в статье[12]. Остальные параметры системы варьировались в ходе эксперимента. Также в данном эксперименте заявки любого типа всегда присоединялись к очереди, если там были свободные места, т.е. вероятность присоединения заявки всегда равна единицы в независимости от класса заявки и от числа заявок в очереди.

На рис. 3 и рис. 4 показаны зависимости вероятности потери, времени отклика системы и размера системы для заявок наиболее и наименее приоритетного класса в зависимости от числа обслуживающих приборов и от размера буфера соответственно. Также приведены результаты для усредненного значения по всем классам. При исследовании зависимости характеристик системы от числа обслуживающих приборов размер буфера был равен 5, а при исследовании зависимости от размера буфера, число обслуживающих приборов также было равно 5.

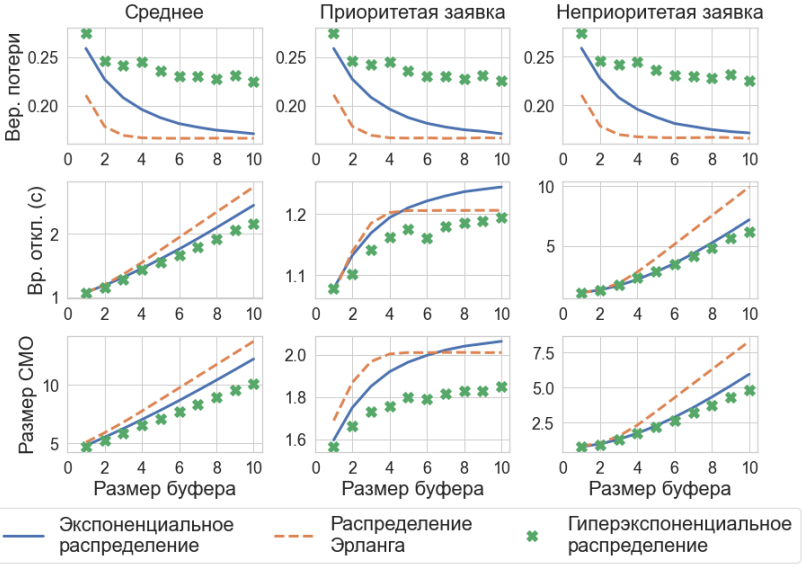

Рис. 4. Зависимость различных параметров системы от размера буфера

Рис. 5. Зависимость различных параметров системы от c_v

Стоит отметить, что наибольшая вероятность потери наблюдается при числе приборов меньше либо равно 6 (рис. 3), что равно среднему числу заявок, поступающих в секунду. При числе сервером большем 6 вероятность потери значительно уменьшается. Гиперэкспоненциальное распределение демонстрирует наибольшую вероятность потери для заявок для всех классов. Чуть больший размер системы наблюдается для распределения Эрланга, особенно это выражено при малом количестве обслуживающих приборов. Время отклика для всех распределений практически одинаковое при любом числе обслуживающих проборов.

На рис. 4 показаны различные характеристики системы в зависимости от размера буфера для различных PH-распределений. Наибольшая вероятность потерь наблюдается для гиперэкспоненциального распределения для всех видов заявок, при этом время отклика системы минимально. Противоположный результат демонстрирует распределение Эрланга, для которого вероятность потери минимальна, а время отклика максимально.

На рис. 5 показаны зависимости вероятности потери, времени отклика и размера системы от коэффициента вариации (c_v), когда число обслуживающих приборов равно 5 и размер буфера также равен 5, остальные параметры системы такие же как и в предыдущих экспериментах. Результаты показали, что для системы такого вида вероятность потери приоритетной заявки практически совпадает с вероятностью потери неприоритетной заявки. Время отклика для приоритетной заявки практически меньше в 2 раза при всех значениях c_v. В среднем число неприоритетных заявок в 1,5 раза больше чем приоритетных.

5. Заключение

В данной работе мы представили результаты исследования приоритетной мультисервисной СМО с входящим $MMAP$-потоком, буфером ограниченной длины, произвольным числом приоритетных классов и временем обслуживания, распределенного по PH. Для случая, когда в системе предусмотрено два класса ($K = 2$), в работе найдено стационарное распределение, получены формулы для вычисления различных характеристик СМО: среднее число заявок в системе, среднее число занятых приборов, среднее число заявок в буфере и других. Для оценок характеристик приоритетных СМО с большим числом классов в работе использовалась имитационная модель, использующая метод Монте-Карло, написанная на языке C++. Для валидации имитационной модели мы использовали математическую модель. Реализация расчетов математической модели выполнена на языке Python. В работе приведен пример СМО с пятью приоритетными классами, для которой исследованы вероятность потери, время отклика и количество заявок в системе для приоритетных и неприоритетных заявок.

Литература

1. AKAN M. et al. *A broader view of designing the liver allocation system* // Oper. Res. – 2012. – Vol. 60, No. 4. – P. 757–770.
2. AWAN I., YOUNAS M., NAVEED W. *Modelling QoS in IoT applications* // Proc. 2014 Int. Conf. Network-Based Inf. Syst. NBiS-2014. – 2014. – P. 99–105.
3. BOCHAROV P.P., D'APICE C., PECHINKIN A.V. *Queueing Theory.* – Berlin, Boston: De Gruyter, 2003.
4. DUDIN A.N., KLIMENOK V.I., VISHNEVSKY V.M. *The theory of queuing systems with correlated flows.* // Springer Publishing Company, Inc., 2019. – Iss. 1. – P. 1–410.
5. DUDIN S. et al. *Improvement of the fairness of non-preemptive priorities in the transmission of heterogeneous traffic* // Mathematics. – 2020. – Vol. 8, No. 6.

6. DUDINA O., KIM C., DUDIN S. *Retrial queuing system with Markovian arrival flow and phase-type service time distribution* // Comput. Ind. Eng. – 2013. – Vol. 66, No. 2. – P. 360–373.
7. EMARA M., ELSAWY H., BAUCH G. Prioritized Multistream Traffic in Uplink IoT Networks: Spatially Interacting Vacation Queues // IEEE Internet of Things Journal. – 2021. – Vol. 8(3). – P. 1477–1491. – DOI: https://doi.org/10.1109/JIOT.2020.3012515.
8. GRAHAM A. *Kronecker Products and Matrix Calculus with Applications.* – Courier Dover Publications, 2018.
9. HE Q.M., XIE J., ZHAO X. *Priority queue with customer upgrades* // Nav. Res. Logist. – 2012. – Vol. 59, No. 5. – P. 362–375.
10. HORVATH G. *Efficient analysis of the queue length moments of the MMAP/MAP/1 preemptive priority queue* // Perform. Eval. – 2012. – Vol. 69, No. 12. – P. 684–700.
11. HEYMAN D.P., LUCANTONI D. *Modeling Multiple IP Traffic Streams with Rate Limits* // IEEE/ACM Trans. Netw. – 2003. – Vol. 11, No. 6. – P. 948–958.
12. JOHNSON M.A., TAAFFE M.R. *Matching moments to phase distributions: Mixtures of erlang distributions of common order* // Commun. Stat. Stoch. Model. – 1989. – Vol. 5, No. 4. – P. 711–743.
13. KLIMENOK V. et al. *Queuing system with two types of customers and dynamic change of a priority* // Mathematics. – 2020. – Vol. 8, No. 5.
14. KLIMENOK V., DUDIN A., VISHNEVSKY V. *Priority multi-server queueing system with heterogeneous customers* // Mathematics. – 2020. – Vol. 8, No. 9.
15. KLIMENOK V. et al. *Lack of invariant property of the erlang loss model in case of MAP input* // Queueing Syst. – 2005. – Vol. 49, No. 2. – P. 187–213.
16. LUCANTONI D.M. *New results on the single server queue with a batch markovian arrival process* // Commun. Stat. Stoch.

Model. – 1991. – Vol. 7, No. 1. – P. 1–46.
17. NEUTS M.F. *Matrix-Geometric Solutions to Stochastic Models* // DGOR / Eds: H. Steckhan et al. – Berlin, Heidelberg: Springer Berlin Heidelberg, 1984. – P. 425.
18. LUCANTONI D.M. *Algorithms for the Multi-Server Queue with Phase Type Service* // Commun. Stat. Stoch. Model. – 1985. – Vol. 1, No. 3. – P. 393–417.
19. MCWHERTER D.T. et al. *Priority mechanisms for OLTP and transactional Web applications* // Proc. Int. Conf. Data Eng. – 2004. – Vol. 20. – P. 535–546.
20. MURALIDHARAN S., ROY A., SAXENA N. *MDP-IoT: MDP based interest forwarding for heterogeneous traffic in IoT-NDN environment* // Futur. Gener. Comput. Syst. – 2018. – Vol. 79. – P. 892–908.
21. RAMASWAMI V. *Independent Markov Processes in Parallel* // Commun. Stat. Stoch. Model. – 1985. – Vol. 1, No. 3. – P. 419–432.
22. TACHIBANA T., FURUICHI T., MINENO H. *Implementing and evaluating priority control mechanism for heterogeneous remote monitoring IOT system* // ACM Int. Conf. Proceeding Ser. – 2016. – Vol. 28 – November. – P. 239–244.

PERFORAMNCE EVALUATION OF THE PRIORITY MULTI-SERVICE SYSTEM MMAP/PH/M/N USING THE MONTE CARLO METHOD.

Vladimir Vishnevsky, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Doc.Sc., professor (vishn@inbox.ru).
Valentina Klimenok, Department of Applied Mathematics and Computer Science, Belarusian State University, Minsk, Doc.Sc., professor (klimenok@bsu.by).
Andrey Larionov, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Cand.Sc. (larioandr@gmail.com).
Amir Mukhtarov, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Cand.Sc. (mukhtarov.amir.a@gmail.com).

Aleksandr Sokolov, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, post-graduate student (aleksandr.sokolov@phystech.edu).

Abstract: In this paper, we present the results of a study of a priority multiline queuing system with a marked Markov arrival process (MMAP), phase-type service time (PH), and a buffer with finite capacity. Priority traffic classes differ in the probability of joining the queue, which depends on the number of customers in buffer, and in the service time PH distribution. If the buffer is full, customers don't join the system. An analytical model has been developed and studied for a particular case of a queueing system with two priority classes. We present an algorithm for calculating stationary probabilities of the system state, loss probabilities, the average number of customers in the queue, and other performance characteristics for this particular case. For the general case of a system with K-classes, a simulation model is constructed, with the help of which various characteristics of the system are studied.

Keywords: multiservice queueing systems, Monte Carlo method, stationary mode.

УДК 519.6
ББК 22.1
DOI: 10.25728/ubs.2023.103.1

Статья представлена к публикации членом редакционной коллегии М.Ф. Караваем.

Поступила в редакцию 17.04.2023.
Дата опубликования 31.05.2023.