ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

УДК 001.891.573
https://doi.org/10.37661/1816-0301-2020-17-1-29-38

Поступила в редакцию 03.12.2019
Received 03.12.2019

Принята к публикации 16.01.2020
Accepted 16.01.2020

Анализ системы обслуживания с повторными вызовами, неоднородными приборами и марковским процессом поступления

Лю Мэй

Белорусский государственный университет, Минск, Беларусь
E-mail: liumei19910101@126.com

Аннотация. Анализируется многолинейная система массового обслуживания с повторными попытками и разнородными приборами. Запросы поступают в систему в соответствии с марковским процессом прибытия. Прибывающие первичные запросы и запросы, которые повторяют попытки попасть на обслуживание с орбиты, занимают свободный прибор с самой высокой скоростью обслуживания, если таковой имеется. В противном случае, если все приборы заняты, запросы переходят на орбиту бесконечной емкости, с которой осуществляют повторные попытки попасть на обслуживание. Общая интенсивность потока повторных попыток бесконечно возрастает с увеличением числа запросов на орбите. Время обслуживания запроса имеет экспоненциальное распределение с интенсивностью, зависящей от номера прибора. Поведение системы описывается многомерной цепью Маркова с непрерывным временем, которая принадлежит классу асимптотически квазитеплицевых цепей Маркова. Это позволяет вывести простое и прозрачное условие эргодичности и вычислить стационарное распределение вероятностей состояний цепи. Представленные численные результаты иллюстрируют динамику некоторых показателей эффективности системы и важность учета корреляции в процессе поступления запросов.

Ключевые слова: система массового обслуживания, разнородные приборы, марковский процесс прибытия, повторные вызовы, асимптотически квазитеплицевая цепь Маркова

Для цитирования. Мэй, Лю. Анализ системы обслуживания с повторными вызовами, неоднородными приборами и марковским процессом поступления / Лю Мэй // Информатика. – 2020. − Т. 17, № 1. – С. 29–38. https://doi.org/10.37661/1816-0301-2020-17-1-29-38

Analysis of retrial queue with heterogeneous servers and Markovian arrival process

Liu Mei

Belarusian State University, Minsk, Belarus
E-mail: liumei19910101@126.com

Abstract. Multi-server retrial queueing system with heterogeneous servers is analyzed. Requests arrive to the system according to the Markovian arrival process. Arriving primary requests and requests retrying from orbit occupy an available server with the highest service rate, if there is any available server. Otherwise, the requests move to the orbit having an infinite capacity. The total retrial rate infinitely increases when the number of requests in orbit increases. Service periods have exponential distribution. Behavior of the system is described by multi-dimensional continuous-time Markov chain which belongs to the class of asymptotically quasi-toeplitz Markov chains. This allows to derive simple and transparent ergodicity condition and compute the stationary probabilities distribution of chain states. Presented numerical results illustrate the dynamics of some system effectiveness indicators and the importance of considering of correlation in the requests arrival process.

Keywords: retrial queue, heterogeneous servers, markovian arrival process, retrials, asymptotically quasi-toeplitz Markov chain

For citation. Mei Liu. Analysis of retrial queue with heterogeneous servers and Markovian arrival process. *Informatics,* 2020, vol. 17, no. 1, pp. 29–38 (in Russian). https://doi.org/10.37661/1816-0301-2020-17-1-29-38

Введение. Теория систем массового обслуживания с повторными вызовами является важной частью теории очередей, которая учитывает влияние повторных попыток попасть на обслуживание. Пропускная способность системы ограничена, и некоторые запросы не могут быть приняты к обслуживанию сразу по прибытии из-за временной недоступности пропускной способности. В отличие от очередей с буферами, где такие запросы помещаются в буфер и затем выбираются для обслуживания в соответствии с некоторыми дисциплинами, и очередей с потерями, в которых запросы теряются, в системах с повторными попытками данные запросы перемещаются в некое виртуальное место, называемое орбитой, и пытаются получить доступ к прибору через случайные промежутки времени. Ввиду своей высокой практической значимости системы обслуживания с повторными вызовами привлекают большое внимание исследователей. Область применения теории очередей с повторными вызовами включает в себя различные телекоммуникационные системы с дисциплинами множественного доступа, базы данных, центры обработки вызовов и т. д. Современное состояние исследований в области систем обслуживания с повторными вызовами изложено в работах [1, 2].

Из-за неоднородного по состоянию поведения цепей Маркова, которые описывают поведение систем обслуживания с повторными вызовами, их анализ значительно более сложен, чем анализ очередей с буферами или потерями. Наибольшие трудности возникают при анализе многолинейных систем обслуживания с повторными вызовами даже в самых простых предположениях о процессах поступления, обслуживания и повторных попыток. Например, система $M/M/N$ с классической политикой повторных попыток исследована в книге [2]. Трудности существенно возрастают, если вводятся более реалистичные предположения о процессах прибытия и обслуживания. В работе [3] изучается система обслуживания с повторными вызовами типа $BMAP/PH/N$. Здесь $BMAP$ (англ. batch Markovian arrival process) обозначает групповой марковский процесс прибытия, представленный в статье [4] как потенциально полезный дескриптор коррелированных групповых потоков в современных телекоммуникационных сетях. Дополнительная информация о $BMAP$ и результаты исследования систем с таким потоком даны в [5, 6]. В настоящей статье предполагается, что процесс поступления – это марковский процесс прибытия MAP, который является частным случаем $BMAP$, когда не допускается групповое поступление. Аббревиатура PH (англ. phase type distribution) обозначает распределение фазового типа [7]. Этот класс распределений довольно широк и включает, в частности, экспоненциальное, эрланговское и коксовское распределения, а также их варианты.

При рассмотрении многолинейных систем обычно предполагается, что приборы однородны и произвольный незанятый прибор задействуется с равной вероятностью для обслуживания, когда приходит новый запрос. Гораздо меньше исследованы очереди с разнородными приборами, которые являются более интересными объектами для исследования. Часто возникают довольно нетривиальные проблемы оптимизации, связанные с назначением приборов прибывающим запросам в зависимости от соотношения средних скоростей обслуживания и затрат на их использование. Проблема оптимального распределения запросов между гетерогенными серверами с целью минимизации среднего числа запросов в обычной системе обслуживания (без учета повторных вызовов) рассматривалась в работах [8–14]. Было показано, что оптимальная стратегия принадлежит к классу монотонных стратегий, т. е. пороговых стратегий, которые используют медленный прибор только тогда, когда длина очереди превышает определенный порог. В статье [15] показано, что для систем с повторными вызовами и классической политикой повторных попыток пороговая стратегия также оптимальна, и предложен алгоритм, который позволяет построить оптимальные стратегии для широкого класса систем массового обслуживания. Аналогичный анализ приведен в [15] для случая постоянной (не зависящей от числа запросов на орбите) частоты повторных попыток.

Многолинейные системы с повторными вызовами, в которых серверы являются однородными, а вновь прибывающий запрос выбирает произвольный незанятый прибор с равной вероятностью и обращается к какому-то конкретному прибору, рассматривались, например, в работах [16, 17].

В настоящей статье анализируется многолинейная система с повторными попытками типа $MAP/\hat{M}_N/N$. Символы $\hat{M}_N$ означают, что распределение времени обслуживания является экспоненциальным с различной скоростью на разных приборах. Предполагается, что приборы пронумерованы в порядке уменьшения скорости обслуживания, т. е. прибор с номером 1 самый быстрый, а прибор с номером N – самый медленный. Согласно известным результатам о структуре оптимального управления (см., например, [15]) лицо, принимающее решение, имеет возможность наблюдать за количеством запросов на орбите и активирует новый, более медленный сервер, если это число превышает определенный порог. Предположим следующее: а) число запросов на орбите не наблюдается, что имеет место в большинстве реальных систем, потому что орбита – это виртуальное место и в действительности ожидающие запросы расположены в некоторой, возможно очень широкой, области; б) дисциплина обслуживания является консервативной. Это означает, что если запрос с орбиты делает попытку попасть на обслуживание и не все приборы заняты, то такой запрос будет принят к обслуживанию. Проблема выбора конкретного прибора из множества доступных в данный момент приборов довольно сложная. Ее решению должно предшествовать формулирование некоторого экономического критерия, включающего, например, затраты на ожидание запросов на орбите (время пребывания в системе) и затраты на использование доступных приборов в единицу времени. В рамках данной статьи экономические аспекты не учитываются (это планируется сделать в дальнейших исследованиях), а изучается следующая дисциплина выбора прибора: в первую очередь задействуется самый быстрый из свободных приборов. Смена прибора в процессе обслуживания любого запроса не допускается.

Описание модели. Рассматривается система массового обслуживания с N приборами. Первичные запросы приходят в систему в соответствии с MAP. Поступление запросов в MAP-потоке присходит под управлением неприводимой цепи Маркова $\nu_t,\, t \ge 0$, с непрерывным временем и конечным пространством состояний $\{0,1,\ldots,W\}$. Поведение MAP-потока полностью характеризуется квадратными матрицами (D_0, D_1) порядка $\overline{W} = W+1$. При этом матрица $D(1) = D_0 + D_1$ является инфинитезимальным генератором цепи Маркова ν_t. Интенсивность λ поступления запросов в MAP определяется как $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$, где $\boldsymbol{\theta}$ – вектор-строка стационарного распределения цепи Маркова $\nu_t,\, t \ge 0$. Вектор $\boldsymbol{\theta}$ будет единственным решением системы линейных алгебраических уравнений $\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$. Коэффициент c_{var} вариации интервалов между прибытием клиентов определяется как $c_{var} = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1$, коэффициент корреляции c_{cor} последовательных интервалов между прибытиями – как $c_{cor} =$ $= (\lambda\boldsymbol{\theta}(-D_0)^{-1} D_1 (-D_0)^{-1}\mathbf{e} - 1)/c_{var}^2$.

Распределение времени обслуживания предполагается экспоненциальным. Приборы имеют разные скорости обслуживания: $\mu_1, \mu_2, \ldots, \mu_N$.

Предположение 1. Серверы нумеруются таким образом, что выполняются неравенства $\mu_1 > \mu_2 > \ldots > \mu_N$. В будущем полученные результаты могут быть использованы для решения проблемы оптимальной нумерации серверов с учетом не только скоростей работы приборов, но и затрат на их использование.

Если входящий запрос застает все приборы незанятыми, он обслуживается на первом приборе. Если первый прибор занят, то запрос ищет незанятый сервер с минимальным номером. Найдя такой прибор, запрос начинает обслуживание на нем. Если же все приборы заняты, то запрос уходит на орбиту. Число мест на орбите (емкость орбиты) не ограничено. Запросы генерируют повторные попытки попасть на обслуживание до тех пор, пока их не обслужат. Предположительно, общий поток повторных попыток таков, что вероятность генерации повторной попытки в интервале $(t, t+\Delta t)$ равна $\alpha_i \Delta t + o(\Delta t)$, когда количество запросов на орбите равно $i,\ i>0,\ \alpha_0 = 0.$ Явная зависимость интенсивностей α_i от i не фиксируется. Предполагается, что интенсивность потока повторных попыток возрастает неограниченно: $\lim\limits_{i\to\infty}\alpha_i = \infty.$ Это справедливо, в частности, для классической стратегии повторов, где $\alpha_i = i\alpha,$ и линейной стратегии, где $\alpha_i = i\alpha + \gamma.$

Целью работы является получение условия существования стационарного распределения вероятностей состояния системы, нахождение этого распределения и краткий анализ проблемы оптимизации работы системы.

Процесс состояний системы. Введем следующие обозначения:

i_t – количество запросов на орбите, $i_t \geq 0;$

$\xi_t^{(n)}$ – состояние n-го прибора, $n = \overline{1, N}$:

$$\xi_t^{(n)} = \begin{cases} 0, & \text{если } n\text{-й прибор свободен,} \\ 1, & \text{если } n\text{-й прибор занят;} \end{cases}$$

v_t – состояние управляющего процесса MAP, $v_t = \overline{0, W},$

в произвольный момент времени $t,\ t > 0.$

Рассмотрим многомерный стохастический процесс с непрерывным временем:

$$\zeta_t = \{i_t, \xi_t^{(1)}, \ldots, \xi_t^{(N)}, \nu_t\},\ t \geq 0.$$

Видно, что данный процесс является неприводимой цепью Маркова. Предположим, что существуют стационарные вероятности этой цепи Маркова:

$$\pi(i, r^{(1)}, \ldots, r^{(N)}, \nu) = \lim_{t\to\infty} P\{i_t = i,\ \xi_t^{(1)} = r^{(1)}, \ldots, \xi_t^{(N)} = r^{(N)},\ \nu_t = \nu\},\ i \geq 0,\ r^{(n)} = \overline{0,1},\ n = \overline{0, N},\ \nu = \overline{0, W}.$$

Условие существования стационарных вероятностей будет приведено ниже.

Пронумеровав состояния цепи Маркова $\zeta_t, t \geq 0,$ в лексикографическом порядке, сформируем вектор-строку

$$\boldsymbol{\pi}(i, r^{(1)}, \ldots, r^{(N)}) = (\pi(i,\ r^{(1)}, \ldots, r^{(N)}, 0), \ldots, \pi(i, r^{(1)}, \ldots, r^{(N)}, W))$$

стационарных вероятностей $\pi(i,\ r^{(1)}, \ldots, r^{(N)}, \nu)$ и вектор-строку $\boldsymbol{\pi}_i,$ состоящую из векторов $\boldsymbol{\pi}(i, r^{(1)}, \ldots, r^{(N)}), i \geq 0$. Отметим, что размер векторов $\boldsymbol{\pi}_i$ рассчитывается как $K = (W+1)2^N$. Определим также бесконечномерный вектор вероятностей $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$.

Дополнительно введем следующие обозначения:

I – тождественная матрица соответствующего размера (при необходимости размер определяется нижним индексом);

O_n – нулевая матрица размера n;

$\otimes$ и $\oplus$ – символы Кронекерова произведения и суммы матриц, $S^{\otimes l} = \underbrace{S \otimes \cdots \otimes S}_{l},\ l \geq 1;$

J – квадратная матрица размера 2^N, определяемая как $J = \mathrm{diag}\{0, \ldots, 0, 1\}$;

$\mathrm{diag}\{\ldots\}$ – диагональная матрица с диагональными элементами, указанными в скобках;

$\bar{I} = (I - J) \otimes I_{\bar{W}}$;

E^0 – матрица размера 2^N со всеми нулевыми элементами, кроме элементов $\left(E^0\right)_{r,r} = -\sum\limits_{k=1}^{N} n_k \mu_k$ для $r = \sum\limits_{k=1}^{N} n_k 2^{N-k}$;

E^- – матрица размера 2^N со всеми нулевыми элементами, кроме элементов $\left(E^-\right)_{r,r'} = \mu_l$ для $r = \sum_{k=1}^{N} n_k 2^{N-k}$, $r' = \sum_{k=1, k\neq l}^{N} n_k 2^{N-k}$, $l = \arg\{n_l = 1\}$;

E^+ – матрица размера 2^N со всеми нулевыми элементами, кроме элементов $\left(E^+\right)_{r,r'} = 1$ для $r = \sum_{k=1}^{q-1} 2^{N-k} + \sum_{k=q+1}^{N} n_k 2^{N-k}$, $r' = \sum_{k=1}^{q} n_k 2^{N-k} + \sum_{k=q+1}^{N} n_k 2^{N-k}$, $q = \arg\min_q \left\{n_q = 0\right\}$;

$\tilde{I} = E^+ \otimes I_{\bar{W}}$.

Лемма. Если существует вектор $\boldsymbol{\pi}$ стационарных вероятностей цепи Маркова ζ_t, $t \geq 0$, то он удовлетворяет уравнению равновесия

$$\boldsymbol{\pi}\mathbf{Q} = \mathbf{0},$$

где $\mathbf{0}$ – бесконечная вектор-строка, состоящая из нулей, а матрица $\mathbf{Q}$, которая является генератором цепи ζ_t, $t \geq 0$, имеет следующую структуру:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{00} & \mathbf{Q}_{01} & 0 & 0 & \cdots \\ \mathbf{Q}_{10} & \mathbf{Q}_{11} & \mathbf{Q}_{12} & 0 & \cdots \\ 0 & \mathbf{Q}_{21} & \mathbf{Q}_{22} & \mathbf{Q}_{23} & \cdots \\ 0 & 0 & \mathbf{Q}_{32} & \mathbf{Q}_{33} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{1}$$

Блоки $\mathbf{Q}_{ij}$, $i,\ j \geq 0$, $j = \left\{\max\left\{0, i-1\right\}, i,\ i+1\right\}$, матрицы $\mathbf{Q}$ имеют размер K и определяются следующим образом:

$$\mathbf{Q}_{i,i+1} = J \otimes D_1,\ \mathbf{Q}_{i,i-1} = \alpha_i \tilde{I},\ \mathbf{Q}_{i,i} = \mathbf{Q}_{ii}{}^{0} + \mathbf{Q}_{ii}{}^{-} + \mathbf{Q}_{ii}{}^{+},$$

где $\mathbf{Q}_{ii}{}^{0} = I_{2^N} \otimes D_0 + E^0 \otimes I_{\bar{W}} - \alpha_i \bar{I}, \mathbf{Q}_{ii}{}^{-} = E^- \otimes I_{\bar{W}}, \mathbf{Q}_{ii}{}^{+} = E^+ \otimes D_1$.

Доказательство. Пронумеруем все возможные комбинации состояний $(r^{(1)}, \ldots, r^{(N)})$ таким образом, чтобы комбинация $\left\{n_1, n_2, \ldots, n_N\right\}$, $n_k = 0,1$, $k = \overline{1, N}$, имела порядковый номер $\sum_{k=1}^{N} n_k 2^{N-k}$.

Матрица E^0 является диагональной, и ее диагональные элементы $\left(E^0\right)_{r,r}$, взятые с противоположным знаком, задают интенсивности выхода цепи Маркова из состояния $(i, n_1, \ldots, n_N, \nu)$. Такой выход возможен за счет окончания обслуживания на одном из приборов, интенсивность выхода равна $\sum_{k=1}^{N} n_k \mu_k$.

Матрица E^- задает интенсивности переходов цепи Маркова, когда один из занятых приборов заканчивает обслуживание. При этом учитываются изменения номера комбинации из-за изменения соответствующего ненулевого значения компоненты n_l на нулевое значение. Интенсивность такого перехода равна μ_l.

Матрица E^+ задает вероятности переходов цепи Маркова, когда первый свободный прибор начинает обслуживание. При этом учитывается изменение номера комбинации из-за замены соответствующего нулевого значения компоненты n_q на ненулевое значение.

Таким образом, справедливость леммы подтверждена.

Следствие. *Марковская цепь ζ_t принадлежит классу асимптотически квазитеплицевых цепей Маркова.*

Доказательство. Согласно определению асимптотически квазитеплицевых цепей Маркова, приведенному в работе [18], необходимо доказать существование пределов

$$Y_0 = \lim_{i\to\infty} R_i^{-1}\mathbf{Q}_{i,i-1},\ Y_2 = \lim_{i\to\infty} R_i^{-1}\mathbf{Q}_{i,i+1},\ Y_1 = \lim_{i\to\infty} R_i^{-1}\mathbf{Q}_{ii} + I,$$

где R_i – диагональная матрица с диагональными элементами, определенными как модули соответствующих диагональных элементов матрицы $\mathbf{Q}_{ii}$, $i \ge 0$. Легко проверить, что R_i является матрицей с диагональными блоками $T_i^{(n)}$, $n = \overline{0,N}$:

$$T_i^{(n)} = \begin{cases} \Lambda \oplus Z_n + \alpha_i I_{\bar{W}\cdot 2^{N-n-1}},\ n = \overline{0, N-1}; \\ \Lambda + \sum_{k=1}^{N} \mu_k I_{\bar{W}},\ n = N, \end{cases}$$

где Λ, Z_n – диагональные матрицы с диагональными элементами, определяемыми диагональными элементами матриц $-D_0$, Δ_{N-n-1}, $n = \overline{0,N}$.

С помощью прямых расчетов можно убедиться в справедливости следующих формул:

$$Y_0 = \tilde{I}_\beta,\ Y_1 = \begin{pmatrix} O & O & \cdots & O & O \\ O & O & \cdots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \cdots & O & O \\ \Gamma_1 & \Gamma_2 & \cdots & \Gamma_N & \Psi \end{pmatrix},\ Y_2 = \begin{pmatrix} O & O & \cdots & O & O \\ O & O & \cdots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \cdots & O & O \\ O & O & \cdots & O & \Phi \end{pmatrix},$$

где

$$\Gamma_n = \mu_n \mathbf{b}_1^{\otimes(N-n)} \otimes C,\ n = \overline{1,N},\ C = (\Lambda + \sum_{k=1}^{N} \mu_k I_{\bar{W}})^{-1};$$

$$\Psi = C(D_0 - \sum_{k=1}^{N} \mu_k I_{\bar{W}}) + I,\ \Phi = CD_1.$$

Следствие доказано.

Условие эргодичности и условие неэргодичности цепи.

Теорема. *Цепь Маркова ζ_t эргодична, если выполняется неравенство*

$$\lambda < \sum_{k=1}^{N} \mu_k, \tag{2}$$

и неэргодична, если выполняется неравенство

$$\lambda > \sum_{k=1}^{N} \mu_k, \tag{3}$$

где λ – фундаментальная скорость MAP.

Доказательство теоремы следует из работы [18] с учетом полученного вида матриц Y_0, Y_1, Y_2. Условие эргодичности (2) интуитивно понятно. Обычно условие эргодичности состоит в том, что в перегруженной системе скорость поступления запросов меньше, чем скорость обслуживания. В рассматриваемой модели, когда она перегружена, т. е. на орбите находится очень много запросов, все приборы непрерывно заняты. Таким образом, общая скорость обслуживания равна $\sum_{k=1}^{N} \mu_k$.

В дальнейшем будем считать, что условие (2) выполнено. Тогда существуют векторы стационарных вероятностей $\boldsymbol{\pi}_i$, $i \geq 0$, определенные выше. Они удовлетворяют системе уравнений равновесия $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$ и условию нормировки $\boldsymbol{\pi}\mathbf{e} = 1$. Эта система бесконечна, и ее решение довольно сложное. Система может быть решена с использованием алгоритма, разработанного в [18], и более эффективного алгоритма, предложенного в [19].

Показатели эффективности работы системы. Среднее количество запросов на орбите вычисляется как

$$L_{orbit} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i \mathbf{e}\,.$$

Вероятность того, что n-й прибор в произвольный момент занят, рассчитывается по формуле

$$P_{busy}^{(n)} = \sum_{i=0}^{\infty} \sum_{(r^{(1)},\ldots,r^{(N)})\in\mathcal{R},\, r^{(n)}=1} \boldsymbol{\pi}(i,\, r^{(1)},\, \ldots,\, r^{(N)})\mathbf{e},\ \ n = \overline{1, N}\,.$$

Предположим, что функционал качества работы системы, задающий ее средние расходы в единицу времени, определяется как

$$E = aL_{orbit} + \sum_{n=1}^{N} c_n P_{busy}^{(n)},$$

где a – штраф за единицу времени ожидания одного запроса на орбите, c_n – штраф за единицу времени использования n-го прибора.

Предположение 2. В дополнение к сделанному выше (без ограничения общности) предположению 1, что приборы перенумерованы в таком порядке, что интенсивности обслуживания удовлетворяют неравенствам $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_N$, предположим выполнение и соотношения $\frac{c_1}{\mu_1} \leq \frac{c_2}{\mu_2} \leq \cdots \leq \frac{c_N}{\mu_N}$. Предположение 2 выглядит обоснованным в практических ситуациях, поскольку оно означает, что более быстрый сервер имеет также меньшую стоимость эксплуатации на единицу скорости обслуживания. При выполнении предположения 2 численные эксперименты подтверждают справедливость следующего правила.

Правило c/μ: прибор с меньшим значением отношения c/μ должен иметь меньший номер, т. е. более высокий приоритет на обслуживание поступающего запроса.

Численные результаты. Чтобы проиллюстрировать результаты выполнения алгоритмов расчета стационарных вероятностей и показателей эффективности, а также показать влияние корреляции на характеристики системы в процессе поступления, рассмотрим кратко следующий пример.

Пусть изначально входящий поток MAP характеризуется матрицами

$$D_0 = \begin{pmatrix} -1,35164 & 0 \\ 0 & -0,04387 \end{pmatrix},\ D_1 = \begin{pmatrix} 1,34265 & 0,00899 \\ 0,02443 & 0,01944 \end{pmatrix}$$

и имеет интенсивность $\lambda = 1$. В дальнейшем интенсивность потока будет варьироваться путем умножения матриц D_0 и D_1 на соответствующую константу.

Процесс поступления имеет коэффициент корреляции длин двух последовательных интервалов между поступлениями запросов $c_{cor} = 0,2$ и коэффициент вариации длин интервалов между прибытиями запросов $c_{var} = 13,4$.

Представим также результаты расчета для модели, в которой поток прихода запросов определяется как стационарный пуассоновский процесс с той же интенсивностью поступления.

Предположим, что общее количество серверов $N = 3$, а скорости обслуживания на приборах $\mu_1 = 4,\ \mu_2 = 3$ и $\mu_3 = 1$ соответственно. Интенсивности выполнения повторных попыток определяются как $\alpha_0 = 0,\ \alpha_i = i\alpha,\ \alpha = 1,\ i > 0$. Штраф за единицу времени ожидания одного запроса на орбите $a = 7$, штрафы за единицу времени использования приборов соответственно $c_1 = 10,\ c_2 = 8,\ c_3 = 5$. При этом $\frac{c_1}{\mu_1} = 2{,}5,\ \ \frac{c_2}{\mu_2} = 2{,}67,\ \ \frac{c_3}{\mu_3} = 5.$

В табл. 1 комбинация (1, 2, 3) дает минимальное значение критерия *E* при любом значении λ. С ростом величины λ значения критерия *E* при различных комбинациях нумерации приборов сближаются, поскольку возрастает вероятность того, что все приборы заняты и возможность их выбора исчезает.

Таблица 1

Значения критерия *E* при различных комбинациях нумерации приборов и различных значениях λ

Комбинация	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	$\lambda = 5$	$\lambda = 6$	$\lambda = 7$
(1, 2, 3)	2,69	6,42	12,90	26,07	57,22	141,55	425,88
(1, 3, 2)	3,02	7,10	13,88	27,32	58,69	143,17	427,62
(2, 3, 1)	3,22	7,52	14,62	28,54	60,52	145,55	430,60
(2, 1, 3)	2,79	6,63	13,30	26,78	58,33	143,03	427,76
(3, 1, 2)	3,74	7,85	14,75	28,47	60,23	145,07	429,92
(3, 2, 1)	3,81	8,07	15,20	29,24	61,40	146,61	431,85

В табл. 2 приведены значения критерия *E* с *M* стационарным пуассоновским процессом поступления запросов при различных комбинациях нумерации приборов и различных значениях λ. Пусть здесь изначально *M* – это входящий поток, характеризующийся матрицами $D_0 = -\lambda$ и $D_1 = \lambda$.

Таблица 2

Значения критерия *E* с *M* стационарным пуассоновским процессом поступления запросов

Комбинация	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	$\lambda = 5$	$\lambda = 6$	$\lambda = 7$
(1, 2, 3)	2,62	5,79	10,27	17,40	30,15	57,22	141,38
(1, 3, 2)	2,92	6,49	11,29	18,72	31,77	59,13	143,57
(2, 3, 1)	3,13	6,89	11,94	19,73	33,31	61,47	147,27
(2, 1, 3)	2,73	6,00	10,61	17,95	31,04	58,63	143,72
(3, 1, 2)	3,83	7,44	12,27	19,82	33,13	60,96	146,24
(3, 2, 1)	3,90	7,63	12,63	20,44	34,12	62,48	148,65

Согласно данным табл. 3 и 4 при невыполнении предположения 2 правило c/μ может перестать действовать. Жирным шрифтом выделены минимальные значения критерия *E*.

Таблица 3

Значения критерия *E* при различных комбинациях нумерации приборов и штрафах за единицу времени использования первого прибора при $\lambda = 1$

Комбинация	$c_1 = 10{,}8,\ c_2 = 8,\ c_3 = 5$	$c_1 = 10{,}9,\ c_2 = 8,\ c_3 = 5$	$c_1 = 11,\ c_2 = 8,\ c_3 = 5$
(1, 2, 3)	**2,838**	2,857	2,876
(1, 3, 2)	3,172	3,191	3,210
(2, 3, 1)	3,242	3,244	3,247
(2, 1, 3)	2,845	**2,851**	**2,858**
(3, 1, 2)	3,826	3,838	3,849
(3, 2, 1)	3,838	3,841	3,845

Таблица 4

Значения критерия E при различных комбинациях нумерации приборов и штрафах за единицу времени использования второго прибора при $\lambda = 1$

Комбинация	$c_1 = 10,\ c_2 = 15{,}5,\ c_3 = 5$	$c_1 = 10,\ c_2 = 15{,}6,\ c_3 = 5$	$c_1 = 10,\ c_2 = 16,\ c_3 = 5$
(1, 2, 3)	**3,200**	3,207	3,235
(1, 3, 2)	3,201	**3,203**	**3,213**
(2, 3, 1)	4,974	4,997	5,091
(2, 1, 3)	4,537	4,560	4,653
(3, 1, 2)	4,001	4,014	4,028
(3, 2, 1)	4,882	4,897	4,954

Заключение. В статье проанализирована система обслуживания с повторными вызовами, разнородными приборами и *MAP*-процессом поступления запросов. Полученные результаты могут найти применение при решении различных задач оптимизации, связанных, в частности, с порядком использования имеющихся обслуживающих приборов, а также в случае фазового распределения времен обслуживания запросов на приборах системы и в случае системы с ненадежными приборами.

References

1. Artalejo J. R., Gomez-Corral A. *Retrial Queueing Systems: a Computational Approach*. Springer, Berlin – Heidelberg, 2008, 318 p.

2. Falin G. I., Templeton J. G. C. *Retrial Queues*. Chapman & Hall, London, 1997, 328 p.

3. Breuer L., Dudin A. N., Klimenok V. I. A retrial $BMAP/PN/N$ system. *Queueing Systems*, 2002, vol. 40, pp. 433–457.

4. Lucantoni D. New results on the single server queue with a batch Markovian arrival process. *Communication in Statistics-Stochastic Models*, 1991, vol. 7, pp. 1–46.

5. Chakravarthy S. R. The batch Markovian arrival process: a review and future work. In Krishnamoorthy A., Raju N., Ramaswami V. (eds.). *Advances in Probability Theory and Stochastic Processes*, Notable Publications Inc., New Jersey, 2001, pp. 21–29.

6. Vishnevskii V. M., Dudin A. N. Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks. *Automation and Remote Control*, 2017, vol. 78, pp. 1361–1403.

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

8. Efrosinin D. V. *Controlled Queueing Systems with Heterogeneous Servers.* Trier University, Germany, 2004, 229 p.

9. Lin W., Kumar P. R. Optimal control of a queueing system with two heterogeneous servers. *IEEE Transactions on Automatic Control*, 1984, vol. 29, pp. 696–703.

10. Luh H. P., Viniotis I. *Optimality of Threshold Policies for Heterogeneous Server Systems*. Raleign, North Carolina State University, 1990.

11. Nobel R., Tijms H. C. Optimal control of a queueing system with heterogeneous servers. *IEEE Transactions on Automatic Control*, 2000, vol. 45, no. 4, pp. 780–784.

12. Rosberg Z., Makowski A. M. Optimal routing to parallel heterogeneous servers-small arrival rates. *Transactions on Automatic Control*, 1990, vol. 35, no. 7, pp. 789–796.

13. Rykov V. V. Monotone control of queueing systems with heterogeneous servers. *Queueing Systems*, 2001, vol. 37, pp. 391–403.

14. Rykov V. V., Efrosinin D. V. Numerical analysis of optimal control polices for queueing systems with heterogeneous servers. *Information Processes*, 2002, vol. 2, no. 2, pp. 252–256.

15. Efrosinin D., Breuer L. Threshold policies for controlled retrial queues with heterogeneous servers. *Annals of Operations Research*, 2006, vol. 41, no. 1, pp. 139–162.

16. Falin G. Stability of the multiserver queue with addressed retrials. *Annals of Operations Research*, 2012, vol. 196, no. 1, pp. 241–246.

17. Mushko V. V. Multiserver queue with addressed retrials. *Annals of Operations Research*, 2006, vol. 141, pp. 283–301.

18. Klimenok V., Dudin A. Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems*, 2006, vol. 54, no. 4, pp. 245–259.

19. Dudin S., Dudina O. Retrial multi-server queueing system with PHF service time distribution as a model of a channel with unreliable transmission of information. *Applied Mathematical Modelling*, 2019, vol. 65, pp. 676–695.

Информация об авторе

Лю Мэй, аспирантка кафедры теории вероятностей и математической статистики факультета прикладной математики и информатики, Белорусский государственный университет, Минск, Беларусь.
E-mail: liumei19910101@126.com

Information about the author

Liu Mei, Postgraduate Student of Department of Probability Theory and Mathematical Statistics of Faculty of Applied Mathematics and Computer Science, Belarusian State University, Minsk, Belarus.
E-mail: liumei19910101@126.com