УДК 519.872

СИСТЕМА ОБСЛУЖИВАНИЯ С РЕЗЕРВИРОВАНИЕМ ПРИБОРОВ ДЛЯ ПОВТОРНЫХ ЗАПРОСОВ

А. Н. Дудин, С. А. Дудин

Белорусский государственный университет,
Минск, Беларусь, dudin@bsu.by, dudins@bsu.by

Рассматривается система типа *MAP/PH/N* с возможностью повторных обслуживаний. Считается, что после обслуживания запрос может покинуть систему или отправиться на орбиту и совершать попытки получить повторное обслуживание. В зависимости от числа запросов на орбите некоторое число приборов может быть зарезервировано для обслуживания повторных запросов. Поведение системы описано многомерной цепью Маркова. Найдены основные характеристики производительности системы.

Ключевые слова: повторное обслуживание; резервирование приборов; стационарные вероятности; цепь Маркова; вероятность потери.

QUEUEING SYSTEM WITH SERVER RESERVATION FOR REPEAT CUSTOMERS

A. N. Dudin, S. A. Dudin

Belarusian State University,
Minsk, Belarus, dudin@bsu.by, dudins@bsu.by

A *MAP/PH/N*-type queuing system with the possibility of repeated services is considered. It is assumed that after the service, a request can leave the system or go to orbit and make attempts to receive service again. Depending on the number of requests in orbit, some number of servers can be reserved for repeated requests. The behavior of the system is described by a multidimensional Markov chain. The main characteristics of the system performance are obtained.

Keywords: repeated service; server reservation; stationary probabilities; Markov chain; loss probability.

1. Введение

Системы массового обслуживания с повторными вызовами характеризуются тем, что в них нет места для ожидания прибывающих запросов. В случае недоступности приборов в момент прибытия запрос может переместиться в виртуальное место, известное как орбита, и попытаться получить обслуживание через случайные интервалы времени. Монография [1] содержит примеры применения теории систем с повторными вызовами для анализа реальных систем. Обзоры исследований систем с повторными вызовами приведены в [2–4].

В отличие от классических систем с повторными вызовами в данной работе исследуется система с повторными обслуживаниями. Другими словами, на орбиту идут не запросы, которые не получили обслуживание из-за занятости всех приборов в момент их поступления, а запросы, получившие обслуживание. Считаем, что данные запросы являются своего рода постоянными клиентами и они более ценны для системы, чем первичные запросы. Чтобы обеспечить приоритет для обслуживания таких запросов, предполагается резервирование части приборов для обслуживания исключительно повторных запросов. Резервирование является динамическим, то есть число зарезервированных приборов зависит от числа повторных запросов в системе.

Процесс поступления запросов описывается марковским входным потоком, что позволяет учитывать возможную корреляцию длин интервалов между моментами поступления запросов, свойственную входным потокам в современных системах и сетях. Также предполагается, что времена обслуживания первичных и вторичных запросов имеют распределение фазового типа, что позволяет учитывать коэффициент вариации времен обслуживания.

Для большей адекватности реальным системам в данной работе считается, что повторные запросы являются нетерпеливыми, то есть они могут покидать систему без обслуживания, а также ненастойчивыми, то есть они могут покидать систему после неудачной попытки попасть на обслуживания.

Целью исследования является разработка методов для нахождения стационарных характеристик систем с повторными обслуживаниями, что позволило бы определять оптимальные стратегии динамического резервирования приборов.

2. Математическая модель

Рассмотрим многолинейную систему массового обслуживания без буфера, состоящую из N приборов.

Предполагается, что входящий поток первичных запросов в систему является марковским входным потоком (MAP-потоком), который управляется цепью Маркова ν_t, $t \geq 0$, с конечным пространством состояний $\{1, 2, \ldots, W\}$. Интенсивности переходов этого процесса в пространстве состояний определяются квадратными матрицами D_0 и D_1 размера W. Элементы матрицы D_1 определяют интенсивности переходов, во время которых происходит поступление запроса. Недиагональные элементы матрицы D_0 определяют интенсивности переходов между состояниями, во время которых поступление запросов не происходит. Диагональные элементы матрицы D_0 отрицательны. Их модули определяют интенсивность выхода из соответствующего состояния. Сумма матриц $D_0 + D_1$ является генератором цепи Маркова ν_t, $t \geq 0$.

Средняя интенсивность поступления запросов λ определяется по формуле $\lambda = \boldsymbol{\theta} D_1 \boldsymbol{e}$, где $\boldsymbol{\theta}$ – вектор-строка стационарных вероятностей цепи Маркова ν_t, $t \geq 0$. Этот вектор является единственным решением системы $\boldsymbol{\theta}(D_0 + D_1) = \mathbf{0}$, $\boldsymbol{\theta}\boldsymbol{e} = 1$. Здесь и далее $\mathbf{0}$ – вектор-строка соответствующего размера, состоящая из нулей, а $\boldsymbol{e}$ – вектор-столбец соответствующего размера, состоящий из единиц.

Более подробную информацию о *MAP*-потоке можно найти в [5–7].

Если в момент прибытия все приборы заняты, запрос теряется, в противном случае он начинает обслуживание.

После обслуживания запрос покидает систему с вероятностью p, а с дополнительной вероятностью переходит на орбиту неограниченного размера (становится вторичным запросом), откуда пытается снова обслужиться независимо от других запросов на орбите через случайные промежутки времени. Если в произвольный момент времени количество запросов на орбите составляет i, $i > 0$, то общая интенсивность повторных попыток равняется $i\alpha$, $\alpha > 0$. Попытка считается успешной, если есть хотя бы один свободный прибор. Если попытка неуспешна, вторичный запрос покидает систему с вероятностью q, и с дополнительной вероятностью $1 - q$ остается на орбите.

Мы предполагаем, что некоторое количество приборов (до K) зарезервировано для обслуживания вторичных запросов. Оно зависит от количества запросов на орбите. Зафиксируем некоторые пороги $G_1, G_2, \ldots, G_K$. Если число запросов на орбите находится в интервале $[G_k, G_{k+1})$, то k, $k = \overline{0, K}$, приборов зарезервированы для вторичных запросов, $G_0 = 0$, $G_{K+1} = \infty$.

Время обслуживания первичного и вторичного запроса любым прибором имеет распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}_1, S)$ и $(\boldsymbol{\beta}_2, S)$ соответственно. Это распределение задается цепью Маркова m_t, $t \geq 0$, с пространством переходных состояний $\{1, 2, \ldots, M\}$ и поглощающим состоянием $M + 1$. Время, имеющее такое распределение, можно рассматривать как время блуждания некоторого запроса в сети, состоящей из M узлов. Стохастический вектор-строка $\boldsymbol{\beta}_l$, $l = 1, 2$, размерности M определяет начальное положение первичного или вторичного запроса в сети в момент начала блуждания. Квадратная матрица S размерности M является субгенератором, определяющим интенсивности переходов цепи $m_t,\ t \geq 0,$ внутри множества $\{1, 2, \ldots, M\}$. Переход в поглощающее состояние означает окончание обслуживания. Интенсивности перехода в поглощающие состояния определяются вектором-столбцом $\boldsymbol{S}_0 = -S\boldsymbol{e}$. Среднее время обслуживания первичного запроса определяется как $b_1^{(1)} = \boldsymbol{\beta}_1(-S)^{-1}\boldsymbol{e},$ среднее время обслуживания вторичного запроса вычисляется как $b_1^{(2)} = \boldsymbol{\beta}_2(-S)^{-1}\boldsymbol{e}.$

Вторичные запросы могут быть нетерпеливыми и покидать орбиту и систему без обслуживания, независимо друг от друга, через экспоненциально распределенное с параметром γ, $\gamma > 0$, время.

3. Процесс состояний системы и его генератор

Поведение исследуемой системы описывается регулярной неприводимой цепью Маркова с непрерывным временем

$$\xi_t = \{i_t, n_t, \nu_t, m_t^{(1)}, \ldots, m_t^{(M)}\},\ \ t \geq 0,$$

где

i_t – количество вторичных запросов, $i_t \geq 0,$

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

ν_t – состояние управляющего процесса MAP-потока первичных запросов, $\nu_t = \overline{1, W},$

$m_t^{(l)}$ – количество приборов на l-й фазе обслуживания, $m_t^{(l)} = \overline{0, n_t},$ $l = \overline{1, M},$ $\sum_{l=1}^{M} m_t^{(l)} = n_t,$ в момент времени $t, t \geq 0.$

Теорема 1. Инфинитезимальный генератор Q цепи Маркова $\xi_t,\ t \geq 0,$ имеет блочно-тридиагональную структуру

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

где его ненулевые блоки $Q_{i,j},\ |i - j| \leq 1,\ i, j \geq 0,$ определяются как

• диагональные блоки $Q_{i,i}$ имеют блочно-тридиагональную структуру $Q_{i,i} = (Q_{i,i}^{(n,n')})$, $|n-n'| \le 1$, $n,n' = \overline{0,N}$, с ненулевыми блоками:

$$Q_{i,i}^{(n,n)} = D_0 \oplus (A_n + \Delta_n) + \delta_{n \ge N-k} D_1 \otimes I_{T_n} -$$
$$-i(\delta_{n<N}\alpha + q\delta_{n=N}\alpha + \gamma) I_{WT_n}, n = \overline{0,N}, i \in [G_k, G_{k+1}), k = \overline{0,K},$$
$$Q_{i,i}^{(n,n+1)} = D_1 \otimes P_n(\boldsymbol{\beta}_1), n = \overline{0, N-k-1}, i \in [G_k, G_{k+1}), k = \overline{0,K},$$
$$Q_{i,i}^{(n,n-1)} = pI_W \otimes L_n, n = \overline{1,N}, i \ge 0;$$

• наддиагональные блоки $Q_{i,i+1}$ имеют только ненулевые поддиагональные блоки $Q_{i,i+1}^{(n,n-1)}$, определяемые как

$$Q_{i,i+1}^{(n,n-1)} = (1-p) I_W \otimes L_n, n = \overline{1,N}, i \ge 0;$$

• поддиагональные блоки $Q_{i,i-1}$ имеют ненулевые наддиагональные блоки $Q_{i,i-1}^{(n,n+1)}$, заданные формулой

$$Q_{i,i-1}^{(n,n+1)} = i\alpha I_W \otimes P_n(\boldsymbol{\beta}_2), n = \overline{0, N-1}, i \ge 1,$$

и диагональные блоки

$$Q_{i,i-1}^{(n,n)} = i(\gamma + \delta_{n=N} q\alpha) I_{WT_n}, n = \overline{0,N}, i \ge 1.$$

Здесь

$\otimes$ и $\oplus$ – символы Кронекеровых произведения и суммы матриц, см., например, [8];

I – единичная матрица, а O – нулевая матрица, размерность которой указывается нижним индексом при необходимости;

$$\delta_a = \begin{cases} 1, & \text{если } a \text{ верно;} \\ 0, & \text{в обратном случае.} \end{cases}$$

Вспомогательные матрицы L_n, $n = \overline{1,N}$, Δ_n, A_n, $n = \overline{0,N}$, и $P_n(\boldsymbol{\beta}_l)$, $l = 1,2$, $n = \overline{0, N-1}$, вычисляются с помощью алгоритмов для матриц с тем же названием и смыслом, представленных в [9].

Размер T_n матриц, представленных выше, определяется как

$$T_n = \frac{(n+M-1)!}{n!(M-1)!}, \ n = \overline{1,N}.$$

Доказательство теоремы проводится путем анализа интенсивностей всех возможных переходов цепи Маркова $\xi_t, \ t \geq 0$, за бесконечно малый интервал времени.

Следствие 1. Цепь Маркова $\xi_t, \ t \geq 0$, принадлежит классу асимптотически квазитеплицевых цепей Маркова (см. [10]).

Используя результаты [11], легко показать, что цепь Маркова $\xi_t, \ t \geq 0$, является эргодической, если вторичные запросы являются нетерпеливыми ($\gamma > 0$) и / или ненастойчивыми ($q > 0$). Значит стационарные вероятности состояний цепи

$$\pi(i,n,\nu,m^{(1)},\ldots,m^{(M)}) = \lim_{t\to\infty} P\{i_t = i, n_t = n, \nu_t = \nu, m_t^{(1)} = m^{(1)},\ldots,m_t^{(M)} = m^{(M)}\},$$
$$i \geq 0, n = \overline{0,N}, \nu = \overline{1,W}, m^{(l)} = \overline{0,n}, \ l = \overline{1,M}, \ \sum_{l=1}^{M} m^{(l)} = n,$$

существуют для всех возможных значений параметров системы.

Сформируем векторы-строки $\boldsymbol{\pi}_i$ из этих вероятностей, перенумерованных в прямом лексикографическом порядке компонент n, ν и обратном лексикографическом порядке компонент $m^{(1)},\ldots,m^{(M)}$.

Хорошо известно, что эти векторы удовлетворяют следующей системе линейных алгебраических уравнений:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)Q = \mathbf{0}, \ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)\boldsymbol{e} = 1,$$

где Q – инфинитезимальный генератор цепи Маркова $\xi_t, \ t \geq 0$. Для решения этой системы рекомендуется использовать эффективный и численно устойчивый алгоритм, разработанный в [12].

4. Характеристики производительности

Среднее количество вторичных запросов рассчитывается по формуле

$$L^{retr} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\boldsymbol{e}.$$

Среднее количество занятых приборов рассчитывается по формуле

$$N^{serv}=\sum_{i=0}^{\infty}\sum_{n=1}^{N}n\boldsymbol{\pi}(i,n)\boldsymbol{e}.$$

Среднее общее количество запросов в системе (первичных и вторичных) рассчитывается по формуле

$$L^{sys}=\sum_{i=0}^{\infty}\sum_{n=0}^{N}(i+n)\boldsymbol{\pi}(i,n)\boldsymbol{e}=L^{retr}+N^{serv}.$$

Средняя интенсивность выходного потока успешно обслуженных запросов рассчитывается по формуле

$$\mu^{out}=\sum_{i=0}^{\infty}\sum_{n=1}^{N}\boldsymbol{\pi}(i,n)(I_W\otimes L_n)\boldsymbol{e}.$$

Вероятность того, что произвольная повторная попытка будет успешной, рассчитывается как

$$P^{suc-ret}=\frac{\sum_{i=1}^{\infty}\sum_{n=0}^{N-1}i\boldsymbol{\pi}(i,n)\boldsymbol{e}}{\sum_{i=1}^{\infty}\sum_{n=0}^{N}i\boldsymbol{\pi}(i,n)\boldsymbol{e}}.$$

Вероятность потери первичного запроса по прибытии рассчитывается как

$$P^{primary-loss}=\frac{1}{\lambda}\sum_{k=0}^{K}\sum_{i=G_k}^{G_{k+1}-1}\sum_{n=N-k}^{N}\boldsymbol{\pi}(i,n)(D_1\otimes I_{T_n})\boldsymbol{e}.$$

Вероятность потери вторичного запроса из-за нетерпеливости рассчитывается как

$$P^{retr-imp-loss}=\frac{1}{(1-p)\mu^{out}}\sum_{i=1}^{\infty}i\gamma\boldsymbol{\pi}_i\boldsymbol{e}.$$

Вероятность потери вторичного запроса из-за нетерпеливости рассчитывается как

$$P^{retr-non-pers-loss} = \frac{q\sum_{i=1}^{\infty} i\alpha\boldsymbol{\pi}(i,N)\boldsymbol{e}}{(1-p)\mu^{out}}.$$

Вероятность потери вторичного запроса рассчитывается как

$$P^{retr-loss} = P^{retr-imp-loss} + P^{retr-non-pers-loss}.$$

Вероятность потери произвольного запроса рассчитывается как

$$P^{loss} = \frac{1}{\lambda+(1-p)\mu^{out}}(\lambda P^{primary-loss} + (1-p)\mu^{out}P^{retr-loss}) =$$

$$= 1 - \frac{\mu^{out}}{\lambda+(1-p)\mu^{out}}.$$

5. Заключение

Исследована система массового обслуживания с резервированием приборов для запросов, осуществляющих повторное обслуживание. В систему поступает коррелированный входной поток первичных запросов. После обслуживания запрос может покинуть систему или отправиться на орбиту с целью повторения обслуживания через некоторое время. Запросы на орбите являются ненастойчивыми и нетерпеливыми. Поведение системы описано многомерной цепью Маркова с пространственно-неоднородными переходами. Найден инфинитезимальный генератор этой цепи, получены формулы для вычисления основных характеристик производительности системы. Результаты работы могут быть использованы для определения оптимальной политики резервирования приборов в зависимости от числа повторных запросов в системе.

Библиографические ссылки

1. *Artalejo J. R., Gomez-Corral A.* Retrial Queueing Systems: A Computational Approach. Springer-Verla, 2008.

2. *Falin G.* A survey of retrial queues // Queueing systems. 1990. Vol. 7, iss. 2. P. 127–167.

3. *Templeton J. G., Falin G. I.* Retrial queues. Routledge, 2023.

4. *Gomez-Corral A.* A bibliographical guide to the analysis of retrial queues through matrix analytic techniques // Annals of Operations Research. 2006. Vol. 141, iss. 1. P. 163–191.

5. *Chakravarthy S. R.* Introduction to Matrix-Analytic Methods in Queues 1: Analytical and Simulation Approach – Basics // New York : ISTE Ltd, London and John Wiley and Sons, 2022.

6. *Dudin A. N., Klimenok V. I., Vishnevsky V. M.* The theory of queuing systems with correlated flows // Springer Nature, 2020.

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

8. *Graham A.* Kronecker products and matrix calculus with applications. Courier Dover Publications, 2018.

9. Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users / C. Kim [et al.] // IEEE Access. 2021. Vol. 9. P. 106933–106946.

10. *Klimenok V. I., Dudin A. N.* Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory // Queueing Systems. 2006. Vol. 54. P. 245–259.

11. Stability of queueing systems with impatience, balking and non-persistence of customers / A. N. Dudin [et al.] // Mathematics. 2024. Vol. 12, iss. 14. Article no. 2214.

12. *Dudin S., Dudina O.* Retrial multi-server queuing system with *PHF* service time distribution as a model of a channel with unreliable transmission of information // Applied Mathematical Modelling. 2019. Vol. 65. P. 676–695.