УДК 519.872

АНАЛИЗ НОВОЙ ДИСЦИПЛИНЫ РАСПРЕДЕЛЕНИЯ РЕСУРСА МЕЖДУ ЭЛАСТИЧНЫМИ И НЕЭЛАСТИЧНЫМИ ЗАПРОСАМИ

А. Н. Дудин[1)], **О.С. Дудина**[2)]

[1), 2)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь, dudin@bsu.by, dudina@bsu.by*

Исследуется система массового обслуживания с двумя типами запросов и гибридной дисциплиной обслуживания, описанная в статье [1]. Выписан инфинитезимальный генератор многомерной цепи Маркова, моделирующей поведение системы. Найдено условие существования стационарного режима. Получены формулы для нахождения основных характеристик производительности системы.

Ключевые слова: гибридная дисциплина обслуживания; условие эргодичности; асимптотически-квазитеплицевые цепи Маркова.

Процесс изменения состояний системы

Пусть фиксированы целые числа M и N, задающие пропускную способность системы и число неэластичных запросов, которые могут одновременно обслуживаться в системе массового обслуживания. Тогда поведение рассматриваемой системы можно описать регулярной неприводимой цепью Маркова (ЦМ) с непрерывным временем $\xi_t = \{i_t, n_t, \nu_t, h_t\}, t \geq 0,$ где в момент времени $t, t \geq 0, i_t$ – число запросов второго типа в системе, $i_t \geq 0$; n_t – число запросов первого типа в системе, $n_t = \overline{0, N}$; ν_t – состояние управляющего процесса потока поступления запросов первого типа MAP_1, $\nu_t = \overline{1, W}$; h_t – состояние управляющего процесса потока поступления запросов второго типа MAP_2, $h_t = \overline{1, V}$.

Перенумеруем состояния ЦМ $\xi_t, t \geq 0,$ в лексикографическом порядке компонент. Совокупность состояний, имеющих значение i счетной компоненты i_t, будем называть уровнем i ЦМ ξ_t.

Обозначим инфинитезимальный генератор этой цепи как Q. Матрица Q содержит интенсивности всех возможных переходов рассматриваемой цепи за бесконечно малый интервал времени. Путем внимательного анализа интенсивностей этих переходов и дальнейшим формированием из них блочных матриц, мы доказали следующее утверждение.

Теорема 1. Инфинитезимальный генератор Q ЦМ $\xi_t, t \geq 0,$ имеет блочно-трехдиагональную структуру, где ненулевые блоки $Q_{i,j}, |i - j| \leq 1,$ определяются следующим образом:

$$Q_{i,i} = \begin{pmatrix} Q_{i,i}^{(0,0)} & Q_{i,i}^{(0,1)} & O & \dots & O & O \\ Q_{i,i}^{(1,0)} & Q_{i,i}^{(1,1)} & Q_{i,i}^{(1,2)} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \dots & Q_{i,i}^{(N,N-1)} & Q_{i,i}^{(N,N)} \end{pmatrix}, \; i \geq 0,$$

$$Q_{0,0}^{(n,n)} = D_0 \oplus H_0 - n\alpha I_{WV}, 0 \leq n < N,$$

$$Q_{0,0}^{(N,N)} = D_0 \oplus H_0 + D_1 \otimes I_V - N\alpha I_{WV},$$

$$Q_{i,i}^{(n,n)} = D_0 \oplus H_0 - (i\gamma + \mu_n + n\alpha) I_{WV}, i > 0, 0 \leq n < N,$$

$$Q_{i,i}^{(N,N)} = D_0 \oplus H_0 - (i\gamma + \mu_N + N\alpha) I_{WV} + D_1 \otimes I_V, i > 0,$$

$$Q_{i,i}^{(n,n+1)} = D_1 \otimes I_V, i \geq 0, 0 \leq n < N,$$

$$Q_{i,i}^{(n,n-1)} = n\alpha I_{WV}, i \geq 0, 0 < n \leq N,$$

$$Q_{i,i-1} = \mathrm{diag}\{\mu_0, \mu_1, \dots, \mu_N\} \otimes I_{WV} + i\gamma I_{(N+1)WV}, i > 0,$$

$$Q_{i,i+1} = I_{(N+1)W} \otimes H_1, i \geq 0.$$

Здесь I – единичная матрица; O – нулевая матрица соответствующей размерности; $\otimes(\oplus)$ – это символ кронекерова произведения (суммы) матриц; $\mathrm{diag}\{\dots\}$ – это диагональная матрица с диагональными элементами, приведенными в скобках.

Условие эргодичности

Одним из важных этапов исследования модели СМО, определяемой случайным процессом с неограниченным пространством состояний, является определение условий существования стационарного режима работы системы.

Сначала рассмотрим случай, когда запросы второго типа являются нетерпеливыми, то есть параметр γ строго больше нуля. В этом случае можно проверить, что исследуемая ЦМ ξ_t принадлежит классу асимптотически квазитеплицевых ЦМ, см. [2]. Используя результаты [2], можно формально доказать интуитивно очевидный факт: если запросы второго типа нетерпеливы ($\gamma > 0$), то стационарное распределение состояний си-

стемы существует при любых значениях системных параметров. Это доказательство достаточно простое и здесь опускается.

Далее мы рассматриваем случай, когда запросы второго типа являются терпеливыми, то есть $\gamma = 0,$ и каждый запрос второго типа должен получать полное обслуживание в системе. В этом случае блоки генератора $Q_{i,i-1}, Q_{i,i}$ и $Q_{i,i+1}$ не зависят от i и имеют вид:

$$Q_{i,i} = Q^0 = (Q_0^{n,m})_{n=\overline{0,N},m=\overline{0,N}}, Q_0^{(n,n)} = D_0 \oplus H_0 - (\mu_n + n\alpha)I_{WV}, 0 \le n < N,$$

$$Q_0^{(N,N)} = D_0 \oplus H_0 - (\mu_N + N\alpha) + D_1 \otimes I_V, Q_0^{(n,n+1)} = D_1 \otimes I_V, 0 \le n < N,$$

$$Q_0^{(n,n-1)} = n\alpha I_{WV}, 0 < n \le N,$$

$$Q_{i,i-1} = Q^- = \mathrm{diag}\{\mu_0, \mu_1, \ldots, \mu_N\} \otimes I_{WV}, i > 0,$$

$$Q_{i,i+1} = Q^+ = I_{(N+1)W} \otimes H_1.$$

Имея такую форму блоков генератора, можно заключить, что в случае $\gamma = 0$ рассматриваемая ЦМ $\xi_t, t \ge 0,$ принадлежит классу квазитеплицевых ЦМ с непрерывным временем (или ЦМ типа $M/G/1$), см., например, [3, 4]. В этом случае необходимое и достаточное условие существования стационарного режима можно записать следующим образом:

$$\boldsymbol{y}Q^+\mathbf{e} < \boldsymbol{y}Q^-\mathbf{e}, \qquad (1)$$

где вектор y является единственным решением следующей системы

$$\boldsymbol{y}(Q^- + Q^0 + Q^+) = \mathbf{0}, \boldsymbol{y}\mathbf{e} = 1. \qquad (2)$$

Легко видеть, что матрица $Q^- + Q^0 + Q^+$ имеет блочно-трехдиагональный вид с наддиагональными элементами, равными $D_1 \otimes I_V$, диагональными элементами $\tilde{D}_0$, $\tilde{D}_0 - \alpha I, \ldots, \tilde{D}_0 - (N-1)\alpha I$, $\tilde{D}_0 - N\alpha I + D_1 \otimes I_V$, где $\tilde{D}_0 = D_0 \otimes I_V + I_W \otimes H(1),$ и поддиагональными элементами $\alpha I, 2\alpha I, \ldots, N\alpha I.$ Следовательно, решение системы (2) может быть найдено в виде $\boldsymbol{y} = (\boldsymbol{x} \otimes \boldsymbol{\theta}_2)$, где $\boldsymbol{\theta}_2$ – вектор стационарного распределения состояний MAP_2, а $\boldsymbol{x}$ – вектор стационарного распределения двухкомпонентной ЦМ $\zeta_t = \{n_t, \nu_t\}, t \ge 0,$ с непрерывным временем, описывающей работу системы типа $MAP/M/N/0$ с MAP-потоком, заданным матрицами D_0 и D_1, и экспоненциально распреде-

ленным с параметром α временем обслуживания. Здесь n_t определяет количество обслуживаемых запросов первого типа, а ν_t определяет состояние управляющего процесса потока MAP_1. Следовательно, инвариантный вектор вероятности $\boldsymbol{x}$ процесса ζ_t задает стационарное распределение числа запросов первого типа в исследуемой системе при условии игнорирования поступления запросов второго типа. Этот вектор определяется как

$$\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N), \ \ \boldsymbol{x}_n = (x_{n,1}, x_{n,2}, \ldots, x_{n,W}),$$

$$x_{n,\nu} = \lim_{t\to\infty} P\{n_t = n, \nu_t = \nu\}, n = \overline{0,N}, \nu = \overline{1,W},$$

и его легко найти любым стандартным методом, см., например, [4].

Найдя вектор $\boldsymbol{x}$ и подставив вектор $\boldsymbol{y}$ в виде $\boldsymbol{y} = (\boldsymbol{x} \otimes \boldsymbol{\theta}_2)$ в неравенство (1), получим, что условие эргодичности можно записать как

$$\theta_2 H_1 \mathbf{e} < \sum_{n=0}^{N} \mu_n \boldsymbol{x}_n \mathbf{e}.$$

Поскольку $\lambda_2 = \theta_2 H_1 \mathbf{e},$ неравенство (1) принимает вид

$$\lambda_2 < \sum_{n=0}^{N} \mu_n \boldsymbol{x}_n \mathbf{e}. \qquad (3)$$

Таким образом, мы доказали следующую теорему.

Теорема 2. Если запросы второго типа абсолютно терпеливы ($\gamma = 0$), то условием эргодичности ЦМ $\xi_t, t \ge 0,$ является выполнение неравенства (3), где векторы вероятностей $\boldsymbol{x}_n, n = \overline{0,N},$ определяются как векторы совместного распределения количества запросов первого типа в системе.

В дальнейшем будем считать, что ЦМ $\xi_t, t \ge 0,$ эргодична. Тогда существуют стационарные вероятности

$$\pi(i,n,\nu,h) = = \lim_{t\to\infty} P\{i_t = i, n_t = n, \nu_t = \nu, h_t = h\}.$$

Стандартным образом из этих вероятностей формируем векторы-строки

$$\boldsymbol{\pi}(i,n,\nu) = (\pi(i,n,\nu,1), \pi(i,n,\nu,2), \ldots, \pi(i,n,\nu,V)), i \ge 0, n = \overline{0,N}, \nu = \overline{1,W},$$

$$\boldsymbol{\pi}(i,n) = (\boldsymbol{\pi}(i,n,1), \boldsymbol{\pi}(i,n,2), \ldots, \boldsymbol{\pi}(i,n,W)), i \ge 0, n = \overline{0,N},$$

$$\boldsymbol{\pi}_i = (\pi(i,1), \pi(i,2), \ldots, \pi(i,N)), i \geq 0.$$

Как известно, векторы стационарных вероятностей можно найти как решение системы уравнений равновесия $(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)Q = \mathbf{0}, (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)\mathbf{e} = 1.$

В случае $\gamma = 0,$ решение такой бесконечной системы уравнений имеет матрично-геометрический вид, см., например, [3] и [4].

В более важном и практически актуальном случае $\gamma > 0$ решение этой системы представляет собой нетривиальную задачу. Это связано с тем, что интенсивности переходов ЦМ $\xi_t, t \geq 0,$ зависят от уровня ЦМ. Для решения полученной системы уравнений, мы рекомендуем использовать эффективный и численно устойчивый алгоритм, разработанный в [5].

Найдя векторы $\boldsymbol{\pi}_i, i \geq 0,$ стационарных вероятностей, можно подсчитать значения различных характеристик производительности системы.

Характеристики производительности

Среднее количество запросов первого (второго) типа на обслуживании

$$N_{serv-1} = \sum_{i=0}^{\infty} \sum_{n=1}^{N} n\boldsymbol{\pi}(i,n)\mathbf{e}, \; N_{serv-2} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}.$$

Среднее количество запросов в системе $L = \sum_{i=0}^{\infty} \sum_{n=0}^{N} (i+n)\boldsymbol{\pi}(i,n)\mathbf{e}.$

Вероятность того, что в произвольный момент в системе нет запросов первого (второго) типа $P_{idle}^{(1)} = \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,0)\mathbf{e}, P_{idle}^{(2)} = \sum_{n=0}^{N} \boldsymbol{\pi}(0,n)\mathbf{e}.$

Интенсивность выходного потока обслуживаемых запросов первого (второго) типа $\lambda_{out-1} = \sum_{i=0}^{\infty} \sum_{n=1}^{N} n\alpha\boldsymbol{\pi}(i,n)\mathbf{e} = \alpha N_{serv-1}, \lambda_{out-2} = \sum_{i=1}^{\infty} \sum_{n=0}^{N} \mu_n \boldsymbol{\pi}(i,n)\mathbf{e}.$

Вероятность того, что произвольный запрос первого типа будет потерян, $P_{loss-1} = \frac{1}{\lambda_1} \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,N)(D_1 \otimes I_V)\mathbf{e} = 1 - \frac{\lambda_{out-1}}{\lambda_1}.$

Вероятность того, что произвольный запрос второго типа будет потерян, $P_{loss-2} = \frac{1}{\lambda_2} \sum_{i=1}^{\infty} i\gamma\boldsymbol{\pi}_i\mathbf{e} = 1 - \frac{\lambda_{out-2}}{\lambda_2}.$

Средняя пропускная способность B, выделенная для обслуживания произвольного запроса второго типа в произвольный момент, $B = \frac{1}{1 - P_{idle}^{(2)}} \sum_{i=1}^{\infty} \sum_{n=0}^{N} \frac{M - nX}{i} \boldsymbol{\pi}(i,n)\mathbf{e}.$

Вероятность того, что в произвольный момент меньше или равно T Мбит/с выделяется для обслуживания каждого запроса второго типа

$$P_{lessT} = \frac{1}{1 - P_{idle}^{(2)}} \sum_{i=1}^{\infty} \sum_{n=\max\{0, \lceil \frac{M - iT}{X} \rceil\}}^{N} \boldsymbol{\pi}(i,n)\mathbf{e},$$

где $\lceil a \rceil$ – минимальное натуральное число, не меньшее значения a.

Заключение

Рассмотрена СМО с двумя типами запросов. Запросы первого типа (неэластичный трафик) обслуживаются как в стандартной многолинейной СМО с потерями. Оставшаяся часть пропускной способности системы, а также пропускная способность, выделенная для обслуживания этих запросов, но не используемая ими в настоящее время, отдается запросам второго типа (эластичный трафик), которые получают обслуживание в соответствии с дисциплиной разделения процессора. В качестве возможного расширения результатов этого исследования на основе представленной здесь методологии, планируется проанализировать модель с фазовыми распределениями требуемого времени обслуживания и групповыми марковскими входными процессами. Используя результаты для так называемого обобщенного фазового распределения, см. [6], полученные результаты можно распространить на модель, когда имеется более двух типов трафика, отличающихся требуемой интенсивностью обслуживания.

Библиографические ссылки

1. *Дудин А. Н., Дудин С. А.* Новая модель распределения ресурса между эластичными и неэластичными запросами. // Сборник трудов Международной научной конференции «Теория вероятностей, математическая статистика и приложения» (Беларусь, Минск, 22-24 апреля 2024 г.). 2024. P. 1-6.

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

3. *Neuts M. F.* Structured stochastic matrices of M/G/1 type and their applications. CRC Press, 2021.

4. *Dudin A. N., Klimenok V. I., Vishnevsky V. M.* The theory of queuing systems with correlated flows. Cham : Springer, 2020. T. 430.

5. *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. V. 65. P. 676-695.

6. *Dudin A., Kim C., Dudina O., Dudin S.* Multi-server queueing system with a generalized phase-type service time distribution as a model of call center with a call-back option // Annals of Operations Research. 2016. V. 239. P. 401-428.