Математическая Теория Игр и её Приложения, т.11, в.3, с. 31–52

УДК 519.25, 004.75, 519.216.5
ББК 22.18

МНОГОКАНАЛЬНАЯ ТАНДЕМНАЯ СИСТЕМА ОБСЛУЖИВАНИЯ СО СПЕЦИАЛИЗИРОВАННЫМ ПРИБОРОМ С КАНИКУЛАМИ

Сину Лал Т.С.*
Ачьюта Кришнамурти
Варгиз Ц. Джошуа
Колледж Коттаям
Математический факультет
686001, Керала, Индия
e-mail: sinulal@cmscollege.ac.in

Рассматривается тандемная система обслуживания, состоящая из двух пулов обслуживающих устройств. Первый пул включает в себя c одинаковых обслуживающих устройств, работающих параллельно, а второй пул состоит из единственного специализированного обслуживающего устройства. Время обслуживание на каждом устройстве первого пула имеет показательное распределение, а на специализированном устройстве второго пула – распределение фазового типа. Заявки поступают на первый пул в соответствии с марковским входящим процессом. В случае наличия свободного устройства поступившая заявка сразу начинает обслуживаться, иначе ожидает в буфере неограниченной емкости. После завершения обслуживания на первом пуле заявки могут либо перейти в конечный

©2019 С. Лал Т.С., А. Кришнамурти, В.Ц. Джошуа

* Работа частично поддержана Kerala State Council for Science Technology and Environment (KSCSTE), Kerala, India, KSCSTE Research Fellowship 2015 (No 001/FSHP-MAIN/2015/KSCSTE)

буфер ожидания второго пула с вероятностью p, либо с вероятностью $1-p$ покинуть систему. При этом в случае если буфер ожидания второго пула оказывается заполненным, заявка будет потеряна. Специализированное устройство второго пула включается только тогда, когда число заявок, ожидающих обслуживания в буфере, превысит некоторое пороговое значение, и остается включенным до тех пор, пока буфер не станет пустым. В настоящей работе с помощью матрично-аналитического метода получено условие стационарности данной системы, а также получено выражение для стационарного распределения. Также вычислены некоторые важные характеристики качества обслуживания. Рассматриваемая в работе модель системы обслуживания описывает процедуру обслуживания пациентов в поликлинике, где первому пулу соответствует терапевтическое отделение, а второму – специализированные отделения, профильные врачи которого начинают работу только тогда, когда наберется достаточное число пациентов в очереди.

Ключевые слова: специализированнное устройство, дисциплина с каникулами, марковский входящий процесс.

Поступила в редакцию: 19.07.18 *После доработки*: 23.08.19 *Принята к публикации*: 30.09.19

1. Введение

Тандемные системы обслуживания постоянно встречаются в повседневной жизни. Такого рода модели являются предметом интенсивных исследований ввиду многочисленных приложений в самых различных областях науки и техники. Учесть все требования клиентов в односерверной системе обслуживания весьма сложно. Ограничения доступности сервера могут привести к бесконечно долгому ожиданию клиента в очереди на обслуживание. Кроме того, требования клиентов могут быть различными, ввиду чего предпочтительно иметь модель обслуживания с экономически эффективной стратегией, которая позволяла бы обслуживать всех клиентов с учетом их потребностей. В данных условиях находят свое применение многосерверные тандемные системы обслуживания.

В данной работе рассматривается тандемная система обслуживания с двумя пулами обслуживающих устройств. Такие системы опи-

сывают процесс обслуживания пациентов в больнице. В качестве первого пула устройств выступает терапевтическое отделение, в качестве второго – специализированные отделения с профильными врачами, например хирургическое отделение. Пациенты обращаются в больницу с различными заболеваниями. Одним пациентам достаточно медицинской консультации терапевта, другим необходимо углубленное обследование и, возможно, лечение, например, хирургическое вмешательство, трансплантация органов и т. п. Такие пациенты направляются в специализированные отделения больницы к профильным врачам. При этом в целях оптимизации врач-специалист начинает прием только после того, как к нему в очередь набралось некоторое достаточное число пациентов. Прием продолжается до тех пор, пока все пациенты не будут обслужены.

Ситуация, рассматриваемая выше, характерна и для других предметных областей, таких, например, как работа колл-центров, системы спецсвязи, антивирусные системы и др. Во избежание потерь клиентов мы предполагаем наличие очереди бесконечной вместимости перед первым пулом. Так, в больницах пациентам разрешается ожидать в приемных отделениях перед посещением специалиста. Однако вместимость отделения специалиста ограничена ввиду экономических соображений, поэтому мы предполагаем наличие очереди конечной вместимости перед вторым пулом. Клиенты направляются в очередь второго пула в соответствии с вероятностным критерием, если доступно место для ожидания.

В отличие от системы, изучаемой в работе [10], рассматриваемая в данной статье система имеет ряд дополнительных свойств. Так, например, отличием является бесконечная очередь перед первым пулом, вводимая в рассмотрение для того, чтобы избежать потерь клиентов и иметь возможность обслуживать клиентов в порядке их поступления. В работе [8] Гомез-Коррел и Мартос детально исследуют тандемную систему обслуживания с двумя пулами обслуживающих устройств с блокировкой. Главной особенностью процесса обслуживания в такой тандемной системе является наличие блокировки, которая происходит в соответствии с механизмом блокирования после обслуживания, позволяющим избежать потерь клиентов после первого этапа обслуживания. Бауман и Зандман [1] исследуют

многосерверную тандемную систему с марковским входящим процессом, временем обслуживания фазового типа и конечными буферами как систему с потерями вследствие ограничения вместимости очередей на всех пулах. Ким, Дудин А., Дудин С., Дудина О. [12] анализируют многосерверную систему обслуживания с приоритетами и гистерезисной стратегией управления числом активных серверов. Ким, Клименок, Дудин [13] рассматривают тандемную систему обслуживания с многосерверными пулами и конечными промежуточными буферами как модель колл-центров с повторными попытками клиентов.

Марковский входящий процесс (MAP) является популярной моделью входного потока, позволяющей учитывать различные варианты потока входящих заявок и зависимости между ними. MAP является одним из наиболее общих классов стохастических процессов со счетным пространством состояний (так называемых считающих процессов) и включает в себя многие процессы поступления заявок, такие как пуассоновский, процесс восстановления фазового типа и пуассоновский процесс, управляемый марковской цепью (MMPP). MAP также может применяться как средство приближения произвольного точечного процесса с заданной точностью. В настоящей модели клиенты поступают в систему согласно MAP-процессу. На втором пуле тандема имеется всего один сервер с распределением времени обслуживания фазового типа. Марковская цепь, которая моделирует систему в предположениях о входном процессе, сценарии обслуживания и стратегии каникул, является независимым от уровня квазипроцессом рождения и гибели (LIQBD), где уровни процесса определяются числом клиентов в первом пуле. Первый пул обслуживания является системой типа MAP/M/c, второй – G/PH/1/N.

Существует множество причин использовать пороговые стратегии включения серверов в системах обслуживания. Основная из них – высокая стоимость запуска и использования сервера. Ибе и Кейлсон [9] исследуют многосерверную систему обслуживания с набором порогов для включения серверов. В этой работе также определяется множество обратных пороговых значений, таких, что когда число клиентов становится ниже этих значений, серверы выключаются по одному. Чоу и Голубчик [3] проводят анализ системы, в которой

N классов клиентов используют пул из K серверов. Совместное использование серверов основывается на динамической стратегии размещения, управляемой набором порогов с гистерезисным поведением. Луи и Голубчик [15] также описывают многосерверную систему обслуживания с гистерезисным пороговым управлением и получают алгоритмические и аналитические решения для различных случаев.

Методы, основанные на матрично-аналитическом подходе в некоторых случаях являются единственным средством анализа сложных систем обслуживания. Данный подход был впервые предложен Ньютсом в семидесятых годах прошлого века. Преимуществом матрично-аналитического метода является возможность исследования модели аналитически, а также применимость численных методов. Более подробно метод представлен в работах [17, 18]. В работе [2] Чакраварти, Кришнамурти и Джошуа исследуют многосерверную систему обслуживания с марковским входящим потоком, при этом стационарное распределение системы найдено с помощью матрично-аналитического метода. В статье [4] Дипак, Джошуа и Кришнамурти исследуют вопросы организации системы обслуживания с конечной очередью таким образом, чтобы минимизировать потери клентов. В работах [5, 16] также рассматриваются системы обслуживания с потерями клиентов. Данья, Кришнамурти и Джошуа в работе [5] исследовали систему параллельного обслуживания, основанную на токенах, в которой целью является минимизация времени ожидания приоритетных клиентов. В статьях [5, 16] матрично-аналитические методы использованы для нахождения стационарного распределения процесса обслуживания в системе. В настоящей статье анализ проводится также с помощью матрично-аналитического метода.

Статья организована следующим образом: во втором разделе представлено подробное описание математической модели и условие устойчивости системы. В Разделе 3 получено стационарное распределение системы. Время ожидания клиента в очереди второго пула получено в Разделе 4. Ключевые характеристики производительности системы представлены в Разделе 5. Раздел 6 содержит результаты численных экспериментов.

В работе используются следующие обозначения и сокращения:

- $\oplus$ и $\otimes$ обозначают сумму и произведение Кронекера соотвест-

венно.

- $diag(A_1, A_2, \ldots, A_m)$ – блочно-диагональная матрицами с блоками $A_1, A_2, \ldots, A_m$ по главное диагонали.
- 0_{mn} – матрица из нулей порядка $m \times n$; 0_a – квадратная матрица из нулей порядка a.
- $[A_{ij}]$ – блочная матрица с блоками A_{ij}.
- $\mathbf{e_m}$ – вектор-столбец длины m из единиц.
- (α, S) – распределение фазового типа времени обслуживания на специализированном устройстве; S^0 – вектор-столбец, такой, что $S\mathbf{e}+S^0 = 0$.

2. Математическая модель

Рассматривается система обслуживания, имеющая два пула обслуживающих устройств, работающих последовательно. В первом пуле имеется c идентичных серверов, времена обслуживания на которых имеют показательное распределение с параметром μ. Клиенты попадают в первый пул в соответствии с марковским входящим процессом, управляемым вложенным процессом $\psi(t), t \geq 0$. Процесс $\psi(t)$ является неприводимой цепью Маркова с непрерывным временем и пространством состояний $\{1, 2, \ldots, a\}$. Интенсивности переходов цепи, не сопровождающихся приходом клиента и сопровождающихся одним приходом определяются, соответственно, матрицами D_0 и D_1. Матрица $D = D_0 + D_1$ является инфинитезимальной образующей цепи $\psi(t)$. Стационарное распределение ζ, соответствующее $\psi(t)$, может быть получено как единственное решение следующей системы:

$$\zeta D = \mathbf{0},$$

$$\zeta e_a = 1.$$

Здесь e_a есть вектор-столбец единиц, а $\mathbf{0}$ есть вектор-строка нулей соответствующей размерности. Средняя интенсивность входного потока (фундаментальная интенсивность марковского входного процесса) λ^* определяется формулой

$$\lambda^* = \zeta D_1 e_a.$$

Коэффициент вариации, C_{var}, времен между приходами может быть получен следующим образом:

$$C_{var}^2 = 2\lambda^*\zeta(-D_0)^{-1}e_a - 1.$$

Коэффициент корреляции времен между приходами определяется как

$$C_{cor} = \lambda^*\zeta(-D_0)^{-1}D_1(-D_0)^{-1}e_a - 1)/C_{var}^2.$$

Клиент, попадающий в первый пул, может начать обслуживание при наличии свободного сервера на первом пуле, в противном случае клиент ожидает в неограниченной очереди перед первым пулом, таким образом все входящие клиенты будут в конечном счете обслужены на первом пуле. На втором (односерверном) пуле случайная длительность обслуживания имеет распределение фазового типа. Время обслуживания имеет неприводимое представление $(\alpha, \mathbf{S})$ порядка r. Между первым и вторым пулом имеется очередь конечной емкости. Клиент, покидающий первый пул, с вероятностью p занимает место (при наличии) в очереди перед вторым пулом, либо покидает систему с дополнительной вероятностью $1-p$. Если очередь перед вторым пулом заполнена, все клиенты, покидающие первый пул, покидают систему. Сервер на втором пуле активируется, когда число клиентов в очереди второго пула достигнет порогового значения T. В этот момент сервер немедленно активируется и начинает последовательное обслуживание клиентов до опустошения очереди.

Процесс обслуживания в системе может быть представлен в форме четырехмерной неприводимой цепи Маркова с непрерывным временем

$$\chi(t) = \{(N(t), n(t), s(t), a(t)), t \geq 0\},$$

где $N(t)$ есть число клиентов в первом пуле (включая находящихся на обслуживании), $n(t)$ есть число клиентов во втором пуле. Если сервер второго пула активен в момент времени t, то $n(t)$ включает как клиента, получающего обслуживание на специализированном сервере, так и клиентов, ожидающих в очереди второго пула. Если сервер второго пула находится на каникулах в момент времени t, то $n(t)$ есть только число ожидающих в очереди второго пула клиентов. $s(t)$ представляет фазу обслуживания на специализированном сервере и $a(t)$ есть фаза входного потока в момент t.

Пространство состояний, S, процесса обслуживания $\chi(t)$ состоит из множеств $S = S_1 \cup S_2$, где

$$S_1 = (\mathbb{Z}_+ \cup \{0\}) \times \{1, ..., N+1\} \times \{1, 2, ..., r\} \times \{1, 2, ..., a\},$$

$$S_2 = (\mathbb{Z}_+ \cup \{0\}) \times \{0, 1, ...T-1\} \times \{1, 2, ..., a\}.$$

Состояния во множестве S_1 соответствуют состояниям процесса, при которых специализированный сервер активен, в то время как S_2 соответствует каникулам этого сервера. Вторая компонента процесса, $n(t)$, на множестве S_1 принимает значения $1, \ldots, N+1$, поскольку обслуживание на втором пуле ведется до исчерпания очереди. Следовательно, $n(t)$ может быть менее T на цикле занятости специализированного сервера. Уровни в пространстве состояний являются неотрицательными целыми числами. Каждый уровень состоит из $a(T + (N+1)r)$ состояний, упорядоченных лексикографически. Процесс $\chi(t)$ таким образом является независимым от уровня обобщенным процессом рождения-гибели с пространством состояний S.

Инфинитезимальная образующая процесса $\chi(t)$ имеет вид

$$\mathbf{Q} = \begin{pmatrix} B_{00} & B_0 & & & & & \\ B_{10} & B_{11} & B_0 & & & & \\ & B_{20} & B_{21} & B_0 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & B_2 & B_1 & B_0 & \\ & & & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Начиная с уровня c, образующая $\mathbf{Q}$ имеет повторяющуюся блочно-трехдиагональную структуру. Указанные блоки определяются следующим образом.

$$B_0 = diag(D_1, I_{r+1} \otimes D_1, I_{r+1} \otimes D_1, \ldots, I_{r+1} \otimes D_1, I_r \otimes D_1, \ldots, I_r \otimes D_1).$$

В блоке B_0 первый подблок D_1 соответствует состояниям системы, в которых $N(t)$ и $n(t)$ равны нулю. При $1 \leq n(t) \leq T-1$, состояния системы могут соответствовать как рабочему режиму, так и каникулам. Поэтому для фиксированного $n(t)$, при $1 \leq n(t) \leq T-1$, имеется $a(r+1)$ состояний, а соответствующие блоки B_0 имеют вид $I_{r+1} \otimes D_1$.

Здесь I_m есть единичная матрица порядка m. Для определения матриц B_{1i} при $i = 0, 1, 2, ..., c-1$, введем матрицы A_1 и A_2 следующим образом:

$$A_1 = \begin{pmatrix} D_0 & & & & & & & & \\ V_0 & U_0 & & & & & & & \\ & V_1 & U_0 & & & & & & \\ & & \ddots & \ddots & & & & & \\ & & & V_1 & U_0 & & & & \\ & & & & V_* & U_1 & & & \\ & & & & & V & U_1 & & \\ & & & & & & \ddots & \ddots & \\ & & & & & & & V & U_1 \end{pmatrix},$$

$$A_2 = \mu I_\kappa, \kappa = a(T + (N+1)r),$$

где $U_0 = diag(D_0, S \oplus D_0)$, $U_1 = S \oplus D_0$,

$$V_0 = \begin{pmatrix} 0_a \\ S^0 \otimes I_a \end{pmatrix},$$

$$V_1 = diag(0_a, (\alpha \otimes S) \otimes I_a),$$

$$V_* = \begin{pmatrix} 0_{a1} \otimes I_a & (\alpha \otimes S) \otimes I_a \end{pmatrix},$$

$$V = \begin{pmatrix} (\alpha \otimes S) \otimes I_a \end{pmatrix}.$$

В указанных подблоках, 0_a суть квадратная нулевая матрица порядка a, а 0_{mn} есть прямоугольная нулевая матрица размерности $m \times n$. Наконец, определим $B_{00} = A_1$, и при $i = 1, 2, ..., c-1$, имеем $B_{i1} = A_1 - iA_2$.

Определим

$$B_* = \begin{pmatrix} M_1^* & M_2^* & & & & & & & & \\ & M_1 & M_2 & & & & & & & \\ & & \ddots & \ddots & & & & & & \\ & & & M_1 & M_2 & & & & & \\ & & & & M_1 & M_2^{\#} & & & & \\ & & & & & M_3 & M_4 & & & \\ & & & & & & \ddots & \ddots & & \\ & & & & & & & M_3 & M_4 & \\ & & & & & & & & M & \end{pmatrix}.$$

B_2 есть блочная матрица с ненулевыми блоками лишь на главной и надглавной диагоналях. В матрице B_2 ненулевые подблоки определяются следующим образом:

$$M_1^* = q\mu I_a,$$

$$M_2^* = \begin{pmatrix} p\mu I_a & 0_{1r} \otimes 0_a \end{pmatrix},$$

$$M_2^{\#} = \begin{pmatrix} p\mu\alpha \otimes I_a \\ p\mu I_{ar} \end{pmatrix},$$

$$M_1 = \begin{pmatrix} q\mu I_{(a+1)r} \end{pmatrix}, \quad M_2 = \begin{pmatrix} p\mu \otimes I_{(a+1)r} \end{pmatrix},$$

$$M_3 = \begin{pmatrix} p\mu I_{ar} \end{pmatrix}, \quad M_4 = \begin{pmatrix} p\mu \otimes I_{ar} \end{pmatrix}.$$

Наконец, при $i = 1, 2, ..., c-1$, $B_{i0} = iB_*$, $B_2 = cB_*$ и $B_1 = A_1 - cA_2$.

Теорема 2.1. *Критерий стационарности рассматриваемого обобщенного процесса рождения-гибели имеет вид*

$$\pi_0 D_1 + \sum_{i=1}^{T-1} \pi_i I_{r+1} \otimes D_1 + \sum_{i=T}^{N+1} I_r \otimes D_1 < c\mu.$$

Доказательство. Пусть $B = B_0 + B_1 + B_2$. Имеем

$$B = \begin{pmatrix}
\Omega_1^* & M_2^* & & & & & & & \\
V_0 & \Omega_1 & M_2 & & & & & & \\
 & V_1 & \ddots & \ddots & & & & & \\
 & & V_1 & \Omega_1 & M_2 & & & & \\
 & & & V_* & \Omega_1 & M_2^{\#} & & & \\
 & & & & V & \Omega_2 & M_4 & & \\
 & & & & & \ddots & \ddots & \ddots & \\
 & & & & & & V & \Omega_2 & M_4 \\
 & & & & & & & V & \Omega
\end{pmatrix}.$$

Тогда B есть инфинитезимальная образующая. B имеет блочно-трехдиагональную структуру, в которой подблоки имеют следующий вид: блоки V_0, V_1, V_*, V те же самые, что в определении матрицы A_1. Диагональные подблоки определяются следующим образом: $\Omega_1^* = M_1^* - cp\mu I_a$, $\Omega_1 = M_1 - cp\mu I_{(a+1)r}$, $\Omega_2 = M_3 - cp\mu I_{ar}$ и $\Omega = M - cp\mu I_{ar}$.

Пусть $\pi = (\pi_0, \pi_1, \dots \pi_{N+1})$ есть соответствующий вектор стационарных вероятностей состояний системы, где $\pi_0 = (\pi_{01}, \pi_{02}, \dots, \pi_{0a})$, $\pi_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{i(a+1)r})$ для $i = 1, 2, \dots, T-1$, и $\pi_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{ira})$ для $i = T, \dots, N+1$.

Для нахождения π решается следующая система уравнений: $\pi B = 0$ и $\pi e_{a(T+(N+1)r)} = 1$, где $e_{a(T+(N+1)r)}$ есть вектор-столбец из единиц размерности $a(T + (N + 1)r)$. Указанная система может быть преобразована к виду

$$\begin{aligned}
&\pi_0(M_1^* - cp\mu I_a) + \pi_1(V_0) = 0,\\
&\pi_1(M_1^*) + \pi_2(M_1 - cp\mu I_{(a+1)r}) + \pi_3 V_1 = 0,\\
&\vdots\\
&\pi_{T-2}(M_2) + \pi_{T-1}(M_1 - cp\mu I_{(a+1)r}) + \pi_T V_* = 0,\\
&\pi_{T-1}(M_2) + \pi_T(M_1 - cp\mu I_{(a+1)r}) + \pi_{T+1} V = 0,\\
&\pi_T(M_2^{\#}) + \pi_{T+1}(M_3 - cp\mu I_{ar}) + \pi_{T+2} V = 0,\\
&\vdots\\
&\pi_{N-1}(M_4) + \pi_{N-1}(M_3 - cp\mu I_{ar}) + \pi_N V = 0,\\
&\pi_N(M_4) + \pi_{N+1}(M - cp\mu I_{ar}) = 0.
\end{aligned}$$

Поскольку матрица $cp\mu I_{ar}$ является обратимой, неизвестный вектор

может быть выражен через π_{N+1}, и система может быть легко решена с использованием балансового соотношения $\pi e_{(N+1)ra} = 1$. Условие стационарности обобщенного процесса рождения-гибели, полученное в [17], имеет вид критерия $\pi B_0 e < \pi B_2 e$, где e есть вектор-столбец из единиц размерности, совпадающей с числом строк матрицы B_1. Имеем

$$\pi B_0 e = \pi_0 D_1 + \sum_{i=1}^{T-1} \pi_i I_{r+1} \otimes D_1 + \sum_{i=T}^{N+1} I_r \otimes D_1,$$

$$\pi B_2 e = c\mu.$$

Указанные равенства завершают доказательство. □

Следствие 2.1. *Для пуассоновского входного потока интенсивности λ, условие стационарности принимает вид $\lambda < c\mu$.*

Доказательство. Поскольку в этом случае $D_0 = [-\lambda]$ и $D_1 = [\lambda]$, утверждение очевидно. □

Следствие 2.2. *Если клиенты попадают в систему в моменты эрланговского процесса восстановления с m фазами, при этом время между фазами распределено показательно с параметром λ, то условие стационарности примет вид*

$$\lambda\Big(\sum_{i=0}^{T-1}\sum_{j=1}^{a+1}\pi_{i(jm)} + \sum_{i=T}^{N+1}\sum_{l=1}^{r}\pi_{ilm}\Big) < c\mu.$$

Доказательство. Матричное представление эрланговского процесса восстановления имеет следующую форму: $D_0 = \lambda\mathcal{J}(m)$ и $D_1 = \lambda\mathcal{K}(m)$, где

$$\mathcal{J}(m) = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & -1 & 1 & \\ & & & \ddots & 1 \\ & & & & -1 \end{pmatrix},$$

а также

$$\mathcal{K}(\boldsymbol{m}) = \begin{pmatrix} 0 & 0 & & & \\ & 0 & 0 & & \\ & & 0 & 0 & \\ & & & \ddots & 0 \\ 1 & & & & 0 \end{pmatrix}.$$

Тогда $\pi B_0 e$ примет вид $\lambda\Big(\sum_{i=0}^{T-1}\sum_{j=1}^{a+1}\pi_{i(jm)} + \sum_{i=T}^{N+1}\sum_{l=1}^{r}\pi_{ilm}\Big)$. □

3. Стационарное распределение

В предположении стационарности, существует вектор стационарных вероятностей состояний системы. Пусть $\eta = (\eta_0, \eta_1, \eta_2, ...)$ есть вектор стационарных вероятностей состояний цепи Маркова $\chi(t)$. Тогда η есть единственное решение системы $\eta Q = 0$ и $\eta e = 1$. Каждый вектор η_i есть вектор-строка размерности $(T + (N + 1)r)a$.

Преобразуем систему $\eta Q = 0$ и $\eta e = 1$ к виду

$$\begin{aligned} &\eta_0 B_{00} + \eta_1 B_{10} = 0, \\ &\eta_0 B_0 + \eta_1 B_{11} + \eta_2 B_{20} = 0, \\ &\vdots \\ &\eta_{c-1} B_0 + \eta_c B_1 + \eta_{c+1} B_2 = 0, \\ &\eta_i B_0 + \eta_{i+1} B_1 + \eta_{i+2} B_2 = 0, i \geq c. \end{aligned}$$

Из матрично-аналитического метода известно, что решение следует искать в виде $\eta_{c+i} = \eta_c R^i$, $i = 0, 1, 2 \ldots$, где R есть минимальное неотрицательное решение матричного квадратного уравнения

$$R^2 B_2 + R B_1 + B_0 = 0.$$

R при этом вычисляется алгоритмически, с использованием метода логарифмического спуска [14]. Далее, при $i = 1, 2, \ldots, c$, имеют место уравнения $\eta_i = \eta_{i-1} B_0 (B_{i1} + H_{i+1} B_{i+10})$, где $H_c = -B_0 (B_1 + R B_2)^{-1}$, и для $i = 1, 2, \ldots, c-1$, матрицы H_i определены следующим образом: $H_i = -B_0 (B_{i1} + H_{i+1} B_{i+10})$. Наконец, получим систему

$$\eta_0 (B_{00} + H_1 B_{10}) = 0.$$

Следовательно, η_0 является стационарным вектором цепи Маркова с конечным числом состояний и образующей $B_{00}+H_1B_{10}$. Вектор η далее вычисляется путем нормализации (деления) векторов η_i на нормализующий коэффициент $\sum_{i=0}^{\infty}\eta_i e$.

4. Распределение времени ожидания клиента в очереди второго пула

Среднее время ожидания клиента, который попадает на k-ю позицию в конечную очередь второго пула может быть получено путем исследования процесса $\xi(t)=\{(\tau(t),s(t))\geq 0\}$. Здесь $\tau(t)$ есть метка k-го в очереди клиента, а $s(t)$ суть фаза обслуживания. Пространство состояний процесса имеет следующий вид:

$$\Sigma=\{1,2,\ldots,N\}\times\{1,2,\ldots,r\}\cup\{\Delta\},$$

где $\{\Delta\}$ — поглощающее состояние. Образующая матрица имеет вид

$$Q^*=\begin{pmatrix}\mathcal{T} & \mathcal{T}^0\\ 0 & 0\end{pmatrix},$$

где

$$\mathcal{T}=\begin{pmatrix}A_{1r} & A_{2r} & & & \\ & A_{1r-1} & A_{2r-1} & & \\ & & \ddots & \ddots & \\ & & & A_{12} & A_{22}\\ & & & & A_{11}\end{pmatrix},\quad \mathcal{T}^0=\begin{pmatrix}0_{1aN}S^0\end{pmatrix}^T.$$

Определим соответствующие подблоки: при $i=N,N-1,\ldots,T$

$$A_{1i}=S,\quad A_{2i}=S^0\otimes\alpha\ ,$$

а при $i=T-1,T-2\ldots,2$,

$$A_{1i}=diag(0_1,S),\quad A_{2i}=diag(0_1,S\otimes\alpha),$$

$$A_{11}=diag(0_1,S\otimes\alpha).$$

Время ожидания отмеченного клиента, занявшего k-ю позицию в очереди, есть время до достижения цепью Маркова поглощающего

состояния. Поэтому время ожидания имеет распределение фазового типа с представлением $(\beta, \mathcal{T})$, где $\beta = (\alpha, 0, \ldots, 0)$. Функция распределения F_k указанного времени ожидания имеет вид

$$F_k(t) = 1 - \beta(\mathbf{exp}(\mathcal{T}\mathbf{t}))\mathbf{e}.$$

Среднее время ожидания отмеченного клиента задается формулой

$$E^k(w) = -\beta\mathcal{T}^{-1}e.$$

Наконец, среднее время ожидания произвольного попадающего в очередь клиента можно вычислить следующим образом:

$$E_w^{\#} = \sum_{i=0}^{\infty} z_{ik-1} E^k(w),$$

где $z_{ik-1} = (z_{ik-11}, z_{ik-12}, \ldots, z_{ik-1r})$, и $z_{ik-1j} = \sum_{l=1}^{a} \eta_{(ik-1jl)}, 1 \leq j \leq r$.

5. Характеристики производительности

- Среднее число клиентов в первом пуле:

$$EC1 = \sum_{j=o}^{\infty} j\eta_j e_{(T+(N+1)r)a}.$$

- Среднее число клиентов в очереди первого пула:

$$EQ1 = \sum_{j=c+1}^{\infty} (j-c)\eta_j e_{(T+(N+1)r)a}.$$

- Среднее число занятых серверов в первом пуле:

$$\begin{aligned} EBS1 &= \sum_{j=c+1}^{\infty} j\eta_j e_{(T+(N+1)r)a} - \sum_{j=c+1}^{\infty} (j-c)\eta_j e_{(T+(N+1)r)a} \\ &= c \sum_{j=c+1}^{\infty} \eta_j e_{(T+(N+1)r)a}. \end{aligned}$$

- Вероятность занятости всех серверов первого пула:

$$P_{B1} = \sum_{i=c}^{\infty} \eta_i e_{(T+(N+1)r)a}.$$

- Вероятность занятости хотя бы одного сервера первого пула:

$$P_{b1} = \sum_{i=1}^{\infty} \eta_i e_{(T+(N+1)r)a}.$$

- Среднее число клиентов в очереди второго пула:

$$EF_c = \sum_{i=0}^{\infty} \Big(\sum_{j=0}^{T-1} j\eta_{ij} e_{(a+1)r} + \sum_{j=T}^{N+1} j\eta_{ij} e_{ar} \Big).$$

- Среднее число клиентов во втором пуле:

$$ES2 = \sum_{i=0}^{\infty} \Big(\sum_{j=0}^{T-1} j\eta_{ij} (0_{a1} e_{(a+1)r}) + \sum_{j=T}^{N+1} j\eta_{ij} e_{ar} \Big).$$

- Вероятность занятости сервера второго пула:

$$P_{B2} = \sum_{j=0}^{\infty} \sum_{i=1}^{T-1} \eta_{ij} (0_{a1} e_{(a+1)r}) + \sum_{j=0}^{\infty} \sum_{i=T}^{N+1} \eta_{ij} e_{ar}.$$

  В определении P_{B2} и $ES2$, вектора η_i имеют различные длины. Это связано с тем, что для $1 \le n(t) \le T-1$ имеется два типа состояний, в одном случае специализированный сервер находится на каникулах, в другом работает. $0_{a1} e_{(a+1)r}$ есть матрица, состоящая из двух блоков столбцов. Первый блок 0_{a1} есть столбец из нулей размерности $a1$, соответствующий состояниям каникул специализированного сервера, в то время как блок $e_{(a+1)r}$ соответствует рабочему режиму указанного сервера.

- Среднее число клиентов в системе (сумма среднего числа клиентов в первом и втором пулах):

$$ECS = ES1 + ES2 = \sum_{j=o}^{\infty} j\eta_j e_{(T+(N+1)r)a} + \sum_{i=0}^{\infty} \Big(\sum_{j=0}^{T-1} j(\eta_{ij} (0_{a1} e_{(a+1)r})) + \sum_{j=T}^{N+1} j\eta_{ij} e_{ar} \Big).$$

- Средняя интенсивность выходящего с первой стадии потока:

$$\Lambda_0 = c\mu \sum_{i=c}^{\infty} \eta_{\mathbf{i}} \mathbf{e}_{(\mathbf{T}+(\mathbf{N}+\mathbf{1})\mathbf{r})\mathbf{a}} + \mu \sum_{\mathbf{i}=\mathbf{0}}^{\mathbf{c}-\mathbf{1}} \mathbf{i} \eta_{\mathbf{i}} \mathbf{e}_{(\mathbf{T}+(\mathbf{N}+\mathbf{1})\mathbf{r})\mathbf{a}}.$$

- Вероятность простоя системы: $P0 = \eta_{00} e_a$.

- Среднее время, проведенное клиентом в первом пуле: $ET1 = \frac{EC1}{\lambda}$

- Среднее число клиентов, покидающих систему после первого пула:

$$LN1 = q\Big(\sum_{i=1}^{c} i\mu\eta_i e_{(T+(N+1)r)a} + c \sum_{i=c+1}^{\infty} \eta_i e_{(T+(N+1)r)a} \Big).$$

- Вероятность заполненности очереди второго пула:

$$PBF = \sum_{i=0}^{\infty} \eta_{\mathbf{iN}+\mathbf{1}} \mathbf{e}_{\mathbf{ar}}.$$

- Среднее время ожидания клиента в системе:

$$E_w^* = \begin{cases} ET1, \text{если клиент покидает систему после первого пула,} \\ ET1 + p(1 - PBF)E_w^r, \text{ если идет в очередь второго пула.} \end{cases}$$

- Интенсивность потери клиентов после первого пула:

$$RL = PBF\Lambda_0.$$

- Интенсивность попадания во второй пул:

$$RP2 = \Lambda_0(1 - PBF).$$

6. Численный пример

Для иллюстрации характеристик системы, выбраны следующие исходные параметры.

$$D_0 = \begin{pmatrix} -2.75 & 1.6 \\ 1 & -3.4 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.5 & 0.65 \\ 1 & 1.4 \end{pmatrix},$$

$$S = \begin{pmatrix} -2.1 & 1.6 \\ 1 & -1.5 \end{pmatrix}, \quad S^0 = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}^T.$$

Фундаментальная интенсивность входного процесса $\lambda^* = 0.6250$, коэффициент вариации равен

$$C_{var} = 2\lambda^*\zeta(-D0)^{-1})e_a - 1 = 1.4395,$$

а коэффициент корреляции

$$C_{cor} = \lambda^*\zeta(-D0)^{-1}D1(-D0)^{-1}e_a) - 1/C_{var}^2 = 0.4632.$$

В данном случае марковский входной поток обладает положительно коррелированными временами между приходами. Характеристики производительности системы исследуем путем изменения остальных параметров. Так, в Таблице 1 исследуется влияние малых изменений интенсивности обслуживания на характеристики производительности системы. В Таблице 2 представлены результаты расчета указанных характеристик при изменении интенсивности обслуживания μ в большем диапазоне, по сравнению с Таблицей 1. В Таблице 3 представлено влияние параметра p на характеристики производительности системы.

На Рис. 1 (слева) видно, что увеличение μ приводит к снижению средней очереди. На Рис. 1 (справа) исследуется зависимость размера очереди от параметров p и μ. Видно, что число клиентов достигает максимума в точке Z диапазона изменений параметров p и μ. Это позволяет формулировать задачи управления занятостью очереди.

μ	EQ1	EC1	EBS1	EQ2	ES2	ECS	Λ_0	Po	ET1	LN1
5	0.022	0.2444	0.2224	0.1738	0.372	0.7902	1.1118	0.0445	0.391	0.4863
5.2	0.0201	0.2314	0.2113	0.1711	0.3819	0.7844	1.0987	0.0452	0.3702	0.475
5.4	0.0183	0.2194	0.201	0.1688	0.3915	0.7796	1.0857	0.0458	0.351	0.4643
5.6	0.0168	0.2084	0.1915	0.1667	0.4007	0.7757	1.0727	0.0464	0.3334	0.454
5.8	0.0155	0.1982	0.1827	0.1648	0.4097	0.7726	1.0598	0.047	0.3171	0.4443
6	0.0143	0.1888	0.1745	0.1631	0.4183	0.7701	1.047	0.0476	0.302	0.4349
6.2	0.0132	0.18	0.1668	0.1616	0.4266	0.7682	1.0344	0.0482	0.288	0.426
6.4	0.0122	0.1719	0.1597	0.1602	0.4347	0.7668	1.022	0.0487	0.275	0.4175
6.6	0.0113	0.1643	0.153	0.159	0.4425	0.7658	1.0097	0.0493	0.2629	0.4093
6.8	0.0105	0.1572	0.1467	0.1579	0.4501	0.7652	0.9976	0.0498	0.2515	0.4014

Таблица 1. Изменение характеристик производительности системы с увеличением интенсивности обслуживания μ.

μ	EQ1	EC1	EBS1	EF_c	ES2	ECS	Λ_0	Po	ET1	LN1
2	0.1731	0.7939	0.6208	0.3372	0.1667	1.2978	1.2416	0.0293	1.2703	0.7801
3	0.0709	0.4808	0.4099	0.2348	0.2493	0.9648	1.2296	0.0361	0.7692	0.6437
4	0.037	0.3311	0.2941	0.1937	0.3166	0.8414	1.1762	0.0408	0.5297	0.5528
5	0.022	0.2444	0.2224	0.1738	0.372	0.7902	1.1118	0.0445	0.391	0.4863
6	0.0143	0.1888	0.1745	0.1631	0.4183	0.7701	1.047	0.0476	0.302	0.4349
7	0.0098	0.1506	0.1408	0.1569	0.4574	0.765	0.9857	0.0503	0.241	0.3939
8	0.007	0.1232	0.1161	0.1533	0.491	0.7674	0.929	0.0526	0.1971	0.3602
9	0.0052	0.1027	0.0975	0.1512	0.52	0.7738	0.8773	0.0546	0.1643	0.332
10	0.004	0.087	0.083	0.1499	0.5453	0.7822	0.8302	0.0563	0.1392	0.3079
11	0.0031	0.0747	0.0716	0.1492	0.5676	0.7915	0.7874	0.0579	0.1195	0.2872
12	0.0025	0.0649	0.0624	0.1489	0.5874	0.8012	0.7484	0.0592	0.1038	0.2692
13	0.002	0.0569	0.0548	0.1488	0.6051	0.8108	0.7129	0.0605	0.091	0.2533
14	0.0017	0.0503	0.0486	0.1489	0.621	0.8202	0.6804	0.0616	0.0804	0.2392
15	0.0014	0.0447	0.0434	0.1491	0.6353	0.8292	0.6506	0.0626	0.0716	0.2266

Таблица 2. Изменение характеристик производительности системы с увеличением интенсивности обслуживания μ.

p	EQ1	EC1	EBS1	EQ2	ES2	ECS	Po	ET1	LN1
0.1	0.0525	0.4881	0.4356	0.1616	0.0159	0.6656	0.0251	0.7810	1.1768
0.2	0.0545	0.4971	0.4426	0.1643	0.0146	0.6760	0.0232	0.7954	1.1835
0.3	0.0565	0.5062	0.4497	0.1675	0.0136	0.6873	0.0215	0.8099	1.1897
0.4	0.0586	0.5155	0.4569	0.1712	0.0128	0.6995	0.0202	0.8248	1.1955
0.5	0.0609	0.5250	0.4641	0.1755	0.0120	0.7126	0.0190	0.8401	1.2010
0.6	0.0633	0.5349	0.4715	0.1803	0.0114	0.7266	0.0180	0.8558	1.2063

Таблица 3. Изменение характеристик производительности системы с увеличением вероятности ухода во второй пул p.

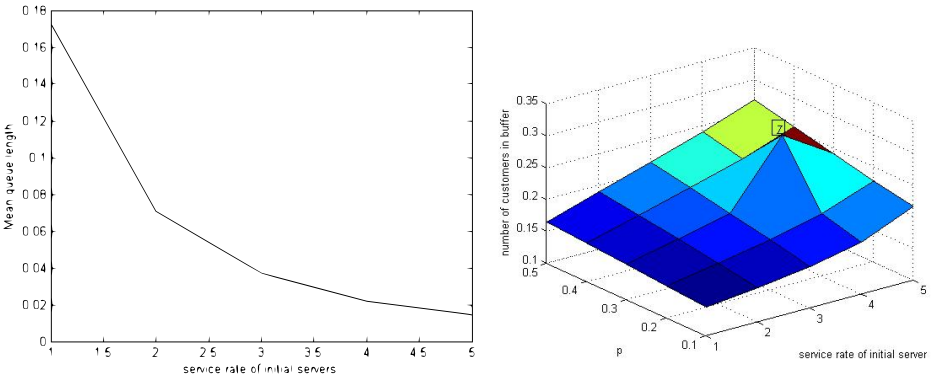

Рисунок 1. Влияние интенсивности μ на характеристику $EQ1$ (слева). Изменение числа клиентов в очереди в зависимости от значений параметров p и μ (справа).

СПИСОК ЛИТЕРАТУРЫ

1. Baumann H., Sandmann W. *Multi-server tandem queue with Markovian arrival process, phase-type service times and finite buffers* // European Journal of Operations Research, Elsevier. 2017. V. 256. No. 1. P. 187–195.

2. Chakravarthy S.R., Krishnamoorthy A., Joshua V.C. *Analysis of multi-server retrial queue with search of customers from the orbit* // Performance Evaluation, Elsevier. 2006. V. 63. No. 8. P. 776–798.

3. Chou C.F., Golubchik L., Lui J.C.S. *Multiclass multiserver threshold-based systems: A study of noninstantaneous server activation* // IEEE Transactions on Parallel and Distributed Systems. 2007. V.18. No.1 P. 96–110.

4. Deepak T.G., Joshua V.C., Krishnamoorthy A. *Queues with postponed work* // Sociedad de Estadistica e Investigacion operativa Top. 2004. V. 12. No. 2. P. 375–398.

5. Dhanya B., Krishnamoorthy A., Joshua V.C. *Token based parallel processing queueing system with priority* // Distributed Computer and Communication Network, Springer. 2017. Vol. 700. P. 231–139.

6. Gomez-Corral A. *A Tandem queue with blocking and Markovian arrival process* // Queueing Systems, Springer. 2002. Vol. 41. No. 4. P. 343–370.

7. Gomez-Corral A., Martos M.E. *Performance of two stage tandem queues with Blocking: The impact of several flows of signals* // Performance Evaluation, Elsevier. 2006. Vol. 63. No. 9–10. P. 910–938.

8. Gomez-Corral A., Martos M.E. *Matrix geometric approximations for tandem queues with blocking and repeated attempt* // Operations Research Letters, Elsevier. 2002. Vol. 30. No. 6. P. 360–374.

9. Ibe O.C., Keilson J. *Multi-server threshold queues with hysteresis* // Performance Evaluation, Elsevier. 1995. Vol. 21. No. 3. P. 185–213.

10. Kim C., Dudin A.N., Dudin S., Dudina O. *Tandem queueing system with impatient customers as a model of call center with interactive voice response* // Performance Evaluation. 2013. Vol. 70. No. 6. P. 440–453.

11. Kim C., Dudin A.N., Dudina O., Dudin S. *Tandem queueing system with infinite and finite intermediate buffers and generalized phase-type service time distribution* // European Journal of Operations Research, Elsevier. 2014. Vol. 235. No. 1. P. 170–179.

12. Kim C., Dudin A.N., Dudin S., Dudina O. *Hysterisis control by the number of active servers in Queueing System MMAP/PH/N with Priority Service* // Performance Evaluation, Elsevier. 2016. Vol. 101. P. 20–33.

13. Kim C., Klimenok V.I., Dudin A.N. *Priority tandem queueing system with retrials and reservation of channels as a model of call center* // Computers and Industrial Engineering, Elsevier. 2016. Vol. 96, P. 61–71.

14. Latouche G., Ramaswami V. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* SIAM. 1999.

15. Lui J.C.S., Golubchik L. *Stochastic coplement analysis of multiserver threshold queues with hysterisis* // Performance Evaluation, Elsevier. 1999. Vol. 35. No. 2. P. 19–48.

16. Mathew A.P., Krishnamoorthy A., Joshua V.C. *A Retrial queueing system with orbital search of customers lost from an offer zone* // Information Technologies and Mathematical Modeling (ITMM 2018), Springer. 2018. Vol. 912. P. 39–54.

17. Neuts M.F. *Matrix geometric solutions in stochastic models – an algorithmic approach.* John Hopkins University Press. 1981.

18. Perros H.G. *A Bibliography of papers on queueing networks with finite capacity queues* // Performance Evaluation, Elsevier. 1989. Vol. 10. No. 3. P. 255–260.

A MULTISERVER TANDEM QUEUE WITH A SPECIALIST SERVER OPERATING WITH A VACATION STRATEGY

Sinu Lal T.S., Department of Mathematics, CMS College Kottayam (sinulal@cmscollege.ac.in),
Achyutha Krishnamoorthy, Department of Mathematics, CMS College Kottayam,
Vargese C. Joshua, Department of Mathematics, CMS College Kottayam.

Abstract: In the queueing system considered, service is provided at two stations, station 1 and station 2 operating in tandem. The station 1 is a multi-server station with c identical servers working in parallel and station 2 is equipped with a single server called specialist server. The service times of each of the servers at station 1 follow an exponential distribution. The specialist server has phase type distributed service times. Customers arrive to the station 1 according to a Markovian arrival process. An arriving customer directly enters into service at station 1 if at least one of the servers is idle, otherwise joins an infinite queue. After receiving service at station 1 customers either proceed to station 2, or can exit the system. There is a finite buffer between two stations. When the buffer is not full, a customer coming out of the station 1 joins the buffer with a probability p or leaves out system with the complimentary probability $1-p$. If the buffer is full, then all the customers coming out of the station 1 are lost forever. The server at the station 2 will be turned on only if the number of customers in the buffer reaches a threshold. Once the server is turned on, the service will be rendered until the buffer is emptied. Stability condition for this system is established and stationary distribution is obtained using matrix analytic methods. Various performance measures are also calculated. Our model is motivated by a hospital situation where station 1 represent the causality clinic and specialist server represents an expert giving consultation at the request of a threshold number of patients.

Keywords: cloud computing, spot instance, mathematical modeling, full-information best-choice problem, Amazon EC2.