ТАНДЕМНАЯ СИСТЕМА ОБСЛУЖИВАНИЯ С УПРАВЛЯЕМОЙ СТРАТЕГИЕЙ ПОВТОРОВ

В.И. Клименок, О.С. Тарамин

Белорусский государственный университет, кафедра ТВиМС
пр. Независимости, д. 4, Минск, Беларусь
телефон: + 375172095486; e-mail: vklimenok@yandex.ru, taramin@bsu.by

Исследована тандемная система обслуживания, первую фазу которой образует однолинейная система с групповым марковским входящим потоком и повторными вызовами, вторую фазу – многолинейная система без буфера. Система имеет два режима работы, различающиеся стратегией повторных попыток. В зависимости от числа повторных вызовов используется либо централизованная, либо децентрализованная стратегия повторов.

Ключевые слова – групповой марковский поток, тандемная система обслуживания, управляемая стратегия повторных вызовов.

1 ВВЕДЕНИЕ

Тандемные системы обслуживания могут быть использованы как для моделирования реальных сетей, состоящих из последовательно расположенных узлов, так и для проверки алгоритмов декомпозиции сетей общего вида. Таким системам уделено большое внимание в литературе, так как практическая важность и математическая сложность таких систем делают их интересными для исследователей в области теории массового обслуживания и ее приложений.

В данной статье исследуется тандемная система с групповым марковским потоком (Batch Markovian Arrival Process – $BMAP$) и орбитой неограниченного объема для повторных вызовов на первой фазе. Система может работать в двух режимах, различающихся стратегией повторных попыток. Если число источников повторных вызовов не превышает некоторого порогового значения, то используется режим с централизованной стратегией, при которой всем источникам разрешается генерировать повторные вызовы таким образом, чтобы суммарная интенсивность повторов была постоянной. В противном случае используется режим с децентрализованной стратегией повторов, при которой каждый источник генерирует вызовы независимо от других источников. При этом интенсивность повторных попыток является возрастающей функцией числа повторных вызовов.

В статье получено условие существования стационарного режима в рассматриваемой системе, найдены стационарные распределения во вложенные и произвольные моменты времени, приведены формулы для нахождения основных характеристик производительности.

2 МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Рассматривается тандемная система обслуживания, состоящая из двух фаз. Первая фаза представлена однолинейной системой обслуживания с повторными вызовами. Заявки поступают в систему в $BMAP$-потоке. Поступление групп заявок в этом потоке возможно только в моменты скачков некоторой неприводимой цепи Маркова $\nu_t, t \ge 0$, с непрерывным временем и конечным пространством состояний $\{0,1,\ldots,W\}$, которая называется управляющим процессом $BMAP$-потока. Интенсивности переходов процесса $\nu_t, t \ge 0$, сопровождающиеся поступлением групп заявок размера $k \ge 1$ (не сопровождающиеся поступлением заявок – при $k = 0$), задаются элементами матриц D_k (недиагональными элементами матрицы D_0) или производящей функцией $D(z) = \sum_{k=0}^{\infty} D_k z^k, |z| \le 1$. При этом матрица $D(1)$ является неприводимым инфинитезимальным генератором цепи Маркова $\nu_t, t \ge 0$. Вектор-строка $\boldsymbol{\theta}$ стационарного распределения этой цепи является решением системы уравнений $\boldsymbol{\theta} D(1) = \mathbf{0}, \boldsymbol{\theta}\boldsymbol{e} = 1$, где $\boldsymbol{e}$ – вектор-столбец, состоящий из единиц. Интенсивность λ поступления заявок в стационарном $BMAP$-потоке имеет вид $\lambda = \boldsymbol{\theta} D'(z)|_{z=1} \boldsymbol{e}$, интенсивность λ_b поступления групп заявок определяется как $\lambda_b = \boldsymbol{\theta}(-D_0)\boldsymbol{e}$. Коэффициент вариации c_{var} длины интервала между моментами поступления последовательных групп определяется формулой $c_{var}^2 = 2\lambda_b \boldsymbol{\theta}(-D_0)^{-1}\boldsymbol{e} - 1$. Коэффициент корреляции c_{cor} длин двух соседних интервалов вычисляется следующим образом:

$$c_{cor} = (\lambda_b \boldsymbol{\theta}(-D_0)^{-1}(D(1) - D_0)(-D_0)^{-1}\boldsymbol{e} - 1)/c_{var}^2 .$$

Более подробное описание $BMAP$-потока можно найти в [1].

Если поступившая группа первичных заявок застает обслуживающий прибор свободным, то одна из заявок группы начинает обслуживаться, а остальные идут на орбиту неограниченного объема. Если прибор занят в момент поступления группы, то все заявки этой группы идут на орбиту, откуда совершают попытки попасть на обслуживание. Если в момент времени t число заявок на орбите равно $i, i > 0$, то вероятность повторной попытки с орбиты в интервале $(t, t+\Delta t)$ равна $\alpha_i \Delta t + o(\Delta t)$.

Предполагаем, что вид функции α_i, задающей суммарную интенсивность повторных попыток, может изменяться в моменты окончания обслуживания на первой фазе и зависит от числа вызовов на орбите в эти моменты. Если в момент окончания обслуживания число i вызовов на орбите меньше некоторого целого числа (порога) $J, J \geq 0$, то до следующего момента окончания обслуживания α_i определяется как $\alpha_i = \gamma > 0$ (если $i > 0$). В противном случае $\alpha_i = i\alpha, \alpha > 0$. Это означает, что стратегия повторов меняется динамически с централизованной на децентрализованную и наоборот в зависимости от числа повторных вызовов в системе.

Времена обслуживания заявок (первичных и повторных) на первой фазе являются независимыми случайными величинами с общим распределением $B(t)$, имеющим конечный первый момент $b_1 = \int_0^\infty t dB(t)$.

После обслуживания на первой фазе заявка поступает на вторую фазу рассматриваемой системы, которая представляет собой N-линейную систему обслуживания без буфера. Время обслуживания прибором второй фазы имеет показательное распределение с параметром μ. Для обслуживания на второй фазе заявке требуется случайное число приборов ϕ. Здесь ϕ – целочисленная случайная величина с распределением

$$q_n = P\{\phi = n\},\ q_n \geq 0,\ n = \overline{0, N},\ \sum_{n=0}^{N} q_n = 1.$$

Если в момент окончания обслуживания заявки на первом приборе на второй фазе отсутствует необходимое число свободных приборов, то с вероятностью p, $0 \leq p \leq 1$, заявка уходит из системы недообслуженной (получает отказ в обслуживании на второй фазе, теряется), а с дополнительной вероятностью ждет, пока освободится нужное число приборов. Период ожидания сопровождается блокировкой прибора первой фазы.

Введем следующие обозначения:

- I – тождественная матрица, I^* – $(N+1)\times(N+1)$-матрица, у которой все диагональные элементы равны -1, элементы первой поддиагонали равны 1, а остальные элементы нулевые;
- $\otimes$ и $\oplus$ – символы Кронекерова произведения и суммы матриц;
- Q_1 – квадратная матрица размерности $(N+1)$, диагональные элементы которой равны q_0, каждый элемент n-ой наддиагонали равен $q_n, n = \overline{1, N}$, а остальные элементы нулевые; $Q_2 = \mathrm{diag}\{\sum_{n=N-r+1}^{N} q_n, r = \overline{0, N}\}$ и Q_3 – квадратная матрица размерности $(N+1)$, у которой все элементы нулевые, кроме последнего столбца, равного $(q_N, q_{N-1}, \ldots, q_0)^T$.
- $R = \mathrm{diag}\{r, r = \overline{0, N}\}$; $\quad \overline{W} = W + 1$;
- $\hat{Q}_m = Q_m \otimes I_{\overline{W}}, m = \overline{1, 3}$; $\quad \overline{Q} = \hat{Q}_1 + p\hat{Q}_2$.

3 СТАЦИОНАРНОЕ РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ СОСТОЯНИЙ ВЛОЖЕННОЙ ЦЕПИ МАРКОВА

Пусть t_n – n-ый момент окончания обслуживания на первой фазе. Рассмотрим процесс $\xi_n = \{i_n, r_n, \nu_n\}, n \geq 1$, где i_n – число заявок на орбите в момент t_n, $i_n \geq 0$; r_n – число занятых приборов на второй фазе в момент $t_n - 0$, $r_n = \overline{0, N}$; ν_n – состояние управляющего процесса $BMAP$ в момент t_n, $\nu_n = \overline{0, W}$.

Процесс $\xi_n, n \geq 1$, является трехмерной цепью Маркова. Перенумеруем состояния этой цепи в лексикографическом порядке и сформируем $(N+1)\overline{W} \times (N+1)\overline{W}$-матрицы $P_{i,j}, i, j \geq 0$, вероятностей переходов цепи $\xi_n, n \geq 1$, из состояний со значением i первой компоненты в состояния со значением j этой компоненты.

Лемма 1. Матрица P одношаговых вероятностей переходов цепи Маркова $\xi_n, n \geq 1$, имеет блочную структуру $P = (P_{i,j})_{i,j \geq 0}$, где

$$P_{i,j} = 0,\ 0 \leq j < i-1, i \geq 0,$$

$$P_{0,j} = \sum_{k=1}^{j+1}[-Q(\Delta \oplus D_0)^{-1}\hat{D}_k + (1-p)F_k\hat{Q}_3\gamma A]\Omega_{j-k+1} +$$

$$+(1-p)\sum_{n=1}^{j}F_n\hat{Q}_3 A\sum_{k=1}^{j-n+1}\hat{D}_k\Omega_{j-n-k+1},\ j \geq 0,\ J \neq 0,$$

$$P_{i,i+l-1} = \overline{Q}A[\gamma\Omega_l + \sum_{k=1}^{l}\hat{D}_k\Omega_{l-k}] + (1-p)[\sum_{k=0}^{l}F_k\hat{Q}_3 A\gamma\Omega_{l-k} +$$

$$+\sum_{m=0}^{l-1}F_m\hat{Q}_3 A\sum_{k=1}^{l-m}\hat{D}_k\Omega_{l-m-k}],\ 0 < i < J, l \geq 0,$$

$$P_{i,j} = \overline{Q}A_i[i\alpha\Omega_{j-i+1} + \sum_{k=1}^{j-i+1}\hat{D}_k\Omega_{j-i-k+1}] +$$

$$+(1-p)[\sum_{n=0}^{j-i+1}F_n\hat{Q}_3 A_{i+n}(i+n)\alpha\Omega_{j-i-n+1} + \sum_{n=0}^{j-i}F_n\hat{Q}_3 A_{i+n} \times$$

$$\times \sum_{k=1}^{j-i-n+1}\hat{D}_k\Omega_{j-i-n-k+1}],\ j \ge max\{0, i-1\},\ i \ge J.$$

Здесь $P(n,t), n \ge 0,$ – коэффициенты матричного ряда

$$e^{D(z)t} = \sum_{n=0}^{\infty}P(n,t)z^n, |z| \le 1,\ \hat{D}_k = I_{N+1} \otimes D_k,\ k \ge 0,$$

$$\Omega_n = \int_0^{\infty} e^{\Delta t} \otimes P(n,t)dB(t),\ F_n = \int_0^{\infty} dF(t) \otimes P(n,t),\ n \ge 0,$$

$$F(t) = (F_{r,r'}(t))_{r,r'=\overline{0,N}},\ \text{где}\ F_{r,r'}(t) = 0\ \text{для}\ r \le r',$$

а при $r > r'$ $F_{r,r'}(t)$ – функция распределения с преобразованием Лапласа-Стилтьеса вида

$$f_{r,r'}(s) = \prod_{l=r'+1}^{r} l\mu(l\mu + s)^{-1},$$

$$Q = \overline{Q} + (1-p)\int_0^{\infty}(dF(t) \otimes e^{D_0 t})\hat{Q}_3,$$

$$A = \int_0^{\infty} e^{-\gamma t}e^{(\Delta \oplus D_0)t}dt = (\gamma I - \Delta \oplus D_0)^{-1},$$

$$A_i = \int_0^{\infty} e^{-i\alpha}e^{(\Delta \oplus D_0)t}dt = (i\alpha I - \Delta \oplus D_0)^{-1},\ i \ge 0,\ \ \Delta = \mu R I^*.$$

Из Леммы 1 следует, что зависимость матриц вероятностей переходов $P_{i,j}$ от значений i и j счетной компоненты не сводится к зависимости от $j-i$, но в то же время видно, что зависимость от i ослабевает при $i \to \infty$ и матрицы $P_{i,j}$ в пределе зависят от значений i, j только через их разность $j-i$. Данное обстоятельство, дополненное тем, что рассматриваемая цепь $\xi_n, n \ge 1$, является неприводимой и непериодической, позволяет отнести эту цепь к классу асимптотически квазитеплицевых цепей Маркова (см. статью [2]).

Теорема 1. Достаточным условием существования стационарного распределения цепи Маркова $\xi_n, n \ge 1$, является выполнение неравенства

$$\rho = \lambda(b_1 + (1-p)\boldsymbol{y}\int_0^{\infty} t dF(t) Q_3 \boldsymbol{e}) < 1, \quad (1)$$

где вектор $\boldsymbol{y}$ – единственное решение системы линейных алгебраических уравнений

$$\boldsymbol{y}Q^*\int_0^{\infty} e^{\Delta t}dB(t) = \boldsymbol{y},\ \ \boldsymbol{ye} = 1.$$

Здесь $Q^* = Q_1 + pQ_2 + (1-p)EQ_3$, E – квадратная матрица, у которой все поддиагональные элементы равны единице, а остальные – нулю.

Далее будем считать, что неравенство (1) выполняется. Пусть $\pi(i,r,\nu), i \ge 0, r = \overline{0,N}, \nu = \overline{0,W},$ – искомые стационарные вероятности. Введем обозначения для векторов-строк этих вероятностей:

$$\boldsymbol{\pi}(i,r) = (\pi(i,r,0), \pi(i,r,1), \dots, \pi(i,r,W));$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1), \dots, \boldsymbol{\pi}(i,N)),\ i \ge 0.$$

Для вычисления векторов $\boldsymbol{\pi}_i, i \ge 0,$ используется алгоритм (см. [2]), разработанный для вычисления стационарного распределения многомерных асимптотически квазитеплицевых цепей Маркова.

4 СТАЦИОНАРНОЕ РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ СОСТОЯНИЙ СИСТЕМЫ В ПРОИЗВОЛЬНЫЙ МОМЕНТ ВРЕМЕНИ

Состояние системы в произвольный момент времени t определим как $(i_t, r_t, \nu_t, m_t), t \ge 0$, где i_t – число заявок на орбите, r_t – число занятых приборов на второй фазе, ν_t – состояние управляющего процесса $BMAP$-потока, m_t –случайная величина, принимающая значения 0, 1, 2 в зависимости от того, свободен, занят или заблокирован прибор первой фазы в момент времени t.

Рассмотрим процесс $\xi_t = \{i_t, r_t, \nu_t, m_t\}, t \ge 0$. Введем обозначения для стационарных вероятностей этого процесса: $p(i,r,\nu,m) = \lim_{t\to\infty} P\{i_t = i, r_t = r, \nu_t = \nu, m_t = m\}$,

$i \ge 0, r = \overline{0,N}, \nu = \overline{0,W}, m = 0,1,2$. Пусть также $\boldsymbol{p}_i(m)$ вектор-строка вероятностей $p(i,r,\nu,m)$, упорядоченных в лексикографическом порядке компонент $(r,\nu), i \ge 0, m = 0,1,2$.

Теорема 2. Ненулевые векторы $\boldsymbol{p}_i(m)$ стационарных вероятностей процесса $\xi_t, t \ge 0$, выражаются через стационарное распределение вложенной цепи Маркова $\boldsymbol{\pi}_i, i \ge 0$, следующим образом:

$$\boldsymbol{p}_0(0) = -\tau^{-1}\boldsymbol{\pi}_0 Q(\Delta \oplus D_0)^{-1},$$

$$\boldsymbol{p}_i(0) = \tau^{-1}[\boldsymbol{\pi}_i\overline{Q}\hat{A}_i + (1-p)\sum_{l=0}^{i}\boldsymbol{\pi}_l F_{i-l}\hat{Q}_3\hat{A}_i],\ i \ge 1,$$

$$\boldsymbol{p}_i(1) = \tau^{-1}\{\sum_{l=0}^{i}\boldsymbol{\pi}_l[\overline{Q}\hat{A}_l\sum_{k=1}^{i-l+1}\hat{D}_k\hat{\Omega}_{i-l-k+1} +$$

$$+(1-p)\sum_{k=0}^{i-l}F_k\hat{Q}_3\hat{A}_{l+k}\sum_{m=1}^{i-l-k+1}\hat{D}_m\hat{\Omega}_{i-l-k-m+1}] +$$

$$+\sum_{l=0}^{i+1}\boldsymbol{\pi}_l[\overline{Q}\hat{A}_l\alpha_l\hat{\Omega}_{i-l+1}+(1-p)\sum_{k=0}^{i-l+1}F_k\hat{Q}_3\hat{A}_{l+k}\alpha_{l+k}\hat{\Omega}_{i-l-k+1}]\},$$

$$\boldsymbol{p}_i(2)=\tau^{-1}(1-p)\sum_{l=0}^{i}\boldsymbol{\pi}_l\sum_{k=0}^{i-l}(F_k+\delta_{0,k}I)\hat{Q}_2\times$$

$$\times\int_0^\infty e^{-\mu Rt}\otimes P(i-l-k,t)dt,\ i\ge 0,$$

где $\hat{A}_i=\begin{cases}A, & i<J\\ A_i, & i\ge J\end{cases}$, $\delta_{i,j}=\begin{cases}0, & i\ne j\\ 1, & i=j\end{cases}$,

$$\hat{\Omega}_n=\int_0^\infty e^{\Delta t}\otimes P(n,t)(1-B(t))dt,n\ge 0,$$

τ – средняя величина интервала между двумя соседними моментами окончания обслуживания на первой фазе:

$$\tau=b_1+(1-p)\sum_{i=0}^\infty\boldsymbol{\pi}_i(I_{N+1}\otimes\boldsymbol{e}_{\overline{W}})\int_0^\infty tdF(t)Q_3\boldsymbol{e}+$$

$$+\sum_{i=0}^\infty\boldsymbol{\pi}_i[\overline{Q}(\boldsymbol{e}_{N+1}\otimes I_{\overline{W}})(\alpha_i I-D_0)^{-1}+$$

$$+(1-p)\sum_{n=0}^\infty F_n\hat{Q}_3(\boldsymbol{e}_{N+1}\otimes I_{\overline{W}})(\alpha_{i+n}I-D_0)^{-1}]\boldsymbol{e}_{\overline{W}}.$$

Следствие 1. Векторы $\boldsymbol{p}_i$ стационарных вероятностей процесса $\{i_t^*,r_t,\nu_t\}$, где i_t^* – число заявок на первой фазе (на орбите и на приборе, включая заблокированную), вычисляются следующим образом: $\boldsymbol{p}_0=\boldsymbol{p}_0(0)$, $\boldsymbol{p}_i=\boldsymbol{p}_i(0)+\sum_{m=1}^{2}\boldsymbol{p}_{i-1}(m),\ i\ge 1$.

5 ХАРАКТЕРИСТИКИ ПРОИЗВОДИТЕЛЬНОСТИ

Вычислив стационарные распределения вероятностей состояний системы, можно найти различные характеристики ее производительности:

• Среднее число заявок на первой фазе в моменты окончания обслуживания и в произвольный момент времени: $L=\boldsymbol{\Pi}'(1)\boldsymbol{e}$ и $\overline{L}=\boldsymbol{P}'(1)\boldsymbol{e}$,

где $\boldsymbol{\Pi}(z)=\sum_{i=0}^\infty\boldsymbol{\pi}_i z^i$, $\boldsymbol{P}(z)=\sum_{i=0}^\infty\boldsymbol{p}_i z^i, |z|\le 1$.

• Векторы стационарного распределения числа занятых приборов на второй фазе в моменты окончания обслуживания на первой фазе и в произвольный момент времени: $\boldsymbol{r}=\boldsymbol{\Pi}(1)(I_{N+1}\otimes\boldsymbol{e}_{\overline{W}}),\overline{\boldsymbol{r}}=\boldsymbol{P}(1)(I_{N+1}\otimes\boldsymbol{e}_{\overline{W}})$.

• Среднее число занятых приборов на второй фазе в моменты окончания обслуживания на первой фазе и в произвольный момент времени:

$$N_{busy}=\boldsymbol{r}R\boldsymbol{e},\ \overline{N}_{busy}=\overline{\boldsymbol{r}}R\boldsymbol{e}.$$

• Вероятность P_{loss} того, что произвольная заявка уйдет из системы (потеряется) после обслуживания на первой фазе, и вероятность P_{block} того, что произвольная заявка вызовет блокировку прибора первой фазы: $P_{loss}=p\boldsymbol{\Pi}(1)\hat{Q}_2\boldsymbol{e}$, $P_{block}=(1-p)\boldsymbol{\Pi}(1)\hat{Q}_2\boldsymbol{e}$.

• Вероятности $P_{idle},P_{serve},P_{block}$ того, что прибор первой фазы свободен, занят или заблокирован:

$$P_{idle}=\sum_{i=0}^\infty\boldsymbol{p}_i(0)\,\boldsymbol{e},\ P_{serve}=\tau^{-1}b_1,\ P_{block}=1-P_{idle}-P_{serve}.$$

• Вероятность того, что произвольная заявка, поступившая в систему, сразу пойдет на обслуживание (не посетив орбиту): $P_{imm}=-\lambda^{-1}\sum_{i=0}^\infty\boldsymbol{p}_i(0)(\boldsymbol{e}_{N+1}\otimes D_0\boldsymbol{e})$.

• Вероятность того, что произвольная заявка, поступившая в систему, успешно обслужится на обеих фазах (не попадет на орбиту, не потеряется и не вызовет блокировку прибора первой фазы):

$$P_{success}=-\lambda^{-1}\sum_{i=0}^\infty\boldsymbol{p}_i(0)(I_{N+1}\otimes D_0\boldsymbol{e})\int_0^\infty e^{\Delta t}dB(t)Q_1\boldsymbol{e}.$$

Полученные в статье аналитические результаты были реализованы в виде модуля пакета прикладных программ «Sirius++» (см. работу [3]), который позволяет вычислять стационарные распределения системы во вложенные и произвольные моменты времени, а также различные характеристики ее производительности. Решена задача численной оптимизации, иллюстрирующая возможность улучшения качества функционирования системы за счет оптимального выбора порога переключения режимов.

ЛИТЕРАТУРА

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

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

[3] Software "Sirius++" for performance evaluations of modern communication networks / A.N. Dudin, V.I. Klimenok, G.V. Tsarenkov // Modelling and Simulation 2002: Proc. of the 16th European Simulation Multiconference, Darmstadt, June 3-5, 2002, Netherlands: SCS. – 2002. – P. 489-493.