ИНФОРМАТИКА

2011 **июль-сентябрь** **№ 3**

УДК 519.872

С.А. Дудин, О.С. Тарамин

СИСТЕМА МАССОВОГО ОБСЛУЖИВАНИЯ M|PH|1|N С ПОВТОРНЫМИ ВЫЗОВАМИ СЕССИЙ

Рассматривается однолинейная система массового обслуживания с конечным буфером и сессионным поступлением запросов. Прием сессий в систему регулируется с помощью так называемых жетонов. Пул жетонов предполагается конечным. Если в момент поступления сессии свободные жетоны отсутствуют, она покидает систему или идет на орбиту. Находится совместное стационарное распределение числа сессий и запросов в системе, рассчитываются формулы для основных характеристик производительности. Проводится численный эксперимент, решается задача оптимизации пропускной способности системы.

Введение

Многие проблемы, возникающие в телекоммуникационных сетях, например проблемы маршрутизации, управления потоками, распределения ресурсов и управления памятью, могут быть успешно решены с помощью теории массового обслуживания. Обычно пользователь телекоммуникационной сети генерирует не один запрос (пакет данных), а целое множество запросов, и обслуживание данного пользователя заключается в передаче всех этих запросов. Поэтому при анализе систем массового обслуживания часто предполагается групповой приход запросов, а также то, что в момент прихода группы все запросы из нее поступают в систему одновременно.

Однако особенностью многих современных телекоммуникационных сетей, в частности IP-сетей, является то, что запросы, принадлежащие одной группе, поступают в систему не одновременно, а в течение некоторого случайного времени. Первый запрос группы прибывает в систему в момент прихода группы, в то время как остальные запросы прибывают один за одним через случайные интервалы времени. Для того чтобы отличать такой механизм прихода от стандартного группового прихода, будем называть группы запросов, приход которых распределен во времени, *сессиями*. Число запросов, которые поступят в составе сессии, является случайным, и оно неизвестно в момент поступления сессии. Такая ситуация возникает, например, при передаче видео- и мультимедиаинформации, ее детальный анализ проведен в работе [1], посвященной моделированию схемы маршрутизации SAPOR (Scheme of Alternative Packet Overflow Routing) в IP-сетях.

В [1] характеристики производительности схемы SAPOR были вычислены с помощью имитационного моделирования. В [2] описана и аналитически исследована первая модель системы массового обслуживания с сессионным поступлением запросов, пригодная как для вычисления характеристик производительности схемы SAPOR, так и для других систем с групповым поступлением запросов, распределенным во времени. С практической точки зрения важным недостатком моделей, исследованных в [1, 2], является предположение о том, что поступившая сессия теряется, если в момент ее прихода свободные жетоны в системе отсутствуют. На практике сессия, которая не была принята в систему, не уходит из нее навсегда, а повторяет попытки попасть на обслуживание через случайные интервалы времени. Механизм повторных вызовов характерен для многих телекоммуникационных сетей. Подробную информацию о текущем состоянии исследования систем массового обслуживания с повторными вызовами можно найти в [3–5].

В рассматриваемой в настоящей статье модели отмеченные выше недостатки устранены. Предполагается, что сессия, пришедшая в момент, когда все жетоны заняты, либо теряется, либо совершает повторные попытки попасть на обслуживание в систему. Из-за пространственной неоднородности процесса, описывающего поведение данной системы, ее исследование значительно сложнее по сравнению с исследованием аналогичной системы с буфером.

1. Математическая модель

Рассматривается однолинейная система массового обслуживания с буфером конечной размерности $N-1, N>0$. Время обслуживания запроса имеет распределение фазового типа (PH – phase type). Это значит, что время обслуживания запроса можно интерпретировать как время, в течение которого управляющий марковский процесс $\eta_t, t \ge 0,$ с конечным пространством состояний $\{1,\ldots,M,M+1\}$ достигнет единственного поглощающего состояния $M+1$ при условии, что начальное состояние этого процесса выбирается в множестве $\{1,\ldots,M\}$ согласно стохастическому вектору-строке $\boldsymbol{\beta}$. Интенсивности переходов процесса $\eta_t, t \ge 0,$ в множестве состояний $\{1,\ldots,M\}$ определяются субгенератором S (квадратной матрицей порядка M с отрицательными диагональными элементами и неотрицательными недиагональными элементами), а интенсивности переходов в поглощающее состояние определяются элементами вектор-столбца $\boldsymbol{S}_0 = -S\mathbf{e}$ с неотрицательными элементами, среди которых есть хотя бы один положительный. Здесь и далее $\mathbf{e}$ – вектор-столбец, состоящий из единиц.

Средняя интенсивность обслуживания μ задается формулой

$$\mu = (\boldsymbol{\beta}(-S)^{-1}\mathbf{e})^{-1}.$$

Запросы поступают в систему в составе сессий. Сессии поступают в соответствии со стационарным пуассоновским потоком со средней интенсивностью λ. По аналогии с [1] считаем, что принятие сессий на обслуживание регулируется с помощью жетонов, и будем полагать, что общее количество жетонов равно $K, K \ge 1$. Параметр K будет рассматриваться как управляющий параметр.

Если в момент прихода сессии есть свободный жетон, то сессия принимается в систему и число свободных жетонов уменьшается на единицу. В противном случае с вероятностью q, $0 \le q \le 1,$ сессия покидает систему, а с дополнительной вероятностью $1-q$ идет на орбиту.

Орбитой называют некоторую виртуальную область, попав в которую сессия делает повторные попытки попасть на обслуживание. Считаем, что сессия совершает повторные попытки через экспоненциально распределенное с параметром $\alpha, \alpha > 0,$ время независимо от других сессий, находящихся на орбите. Таким образом, если на орбите в момент времени t находится i сессий, то вероятность того, что будет совершена попытка начать обслуживание в течение малого интервала времени $(t, t+\Delta t),$ равна $i\alpha\Delta t + o(\Delta t), i > 0.$

Предполагается, что первый запрос сессии поступает в систему в момент прихода сессии и, если буфер не заполнен, этот запрос принимается в систему. Если буфер занят, запрос теряется.

После приема сессии следующий запрос из нее может прийти в течение экспоненциально распределенного времени с параметром γ. Если система незаполнена, этот запрос принимается в систему, иначе он теряется. Потеря запроса не влечет прекращения сессии.

Число запросов в сессии имеет геометрическое распределение с параметром $\theta,$ т. е. вероятность того, что сессия состоит из k запросов, равна $\theta^{k-1}(1-\theta), k \ge 1.$ Если в течение экспоненциально распределенного с параметром γ времени с момента прихода предыдущего запроса из принятой сессии новый запрос не поступает, это значит, что поступление сессии окончено. Жетон, который был выдан этой сессии, освобождается и переходит в пул свободных жетонов, а сессия считается завершенной.

Описанная система может использоваться при моделировании современных p2p (peer to peer) телекоммуникационных сетей (работающих, например, по протоколу BitTorrent, где пользователь может устанавливать несколько соединений с другими пользователями), в частности для определения оптимального числа соединений с другими пользователями. Применительно к BitTorrent λ – это средняя интенсивность установления новых соединений; α^{-1} – средний интервал, через который пользователь повторит попытку соединиться после того, как он был отсоединен; γ – средняя интенсивность в каждом соединении; $(1-\theta)^{-1}$ – среднее число пакетов, получаемых в каждом со-

единении; q – вероятность того, что отсоединенный пользователь не будет повторять попыток соединиться снова; μ – средняя пропускная способность сети пользователя.

2. Совместное распределение числа сессий и запросов в системе

В настоящей работе ставится задача нахождения основных характеристик производительности системы и демонстрации возможности максимизации пропускной способности системы за счет выбора оптимального значения параметра K.

Рассмотрим четырехмерный процесс

$$\xi_t = \{i_t, r_t, k_t, \eta_t\}, \quad t \geq 0,$$

где i_t, $i_t \geq 0$, – число сессий на орбите; r_t, $r_t = \overline{0,N}$, – число запросов в системе; k_t, $k_t = \overline{0,K}$, – число сессий, имеющих жетон; η_t, $\eta_t = \overline{1,M}$, – состояние процесса обслуживания в момент времени $t, t \geq 0$.

Процесс $\xi_t = \{i_t, r_t, k_t, \eta_t\}, t \geq 0$, является неприводимой, регулярной цепью Маркова с непрерывным временем.

Введем следующие обозначения:

I – единичная матрица, O – нулевая матрица, $\mathbf{0}_M$ – вектор-строка размерности M, состоящий из нулей, где $M \geq 1$. Если размерность вектора ясна из контекста, она может быть опущена;

$\mathbf{e}_M^{(m)}$ – вектор-столбец размерности $M+1$, состоящий из нулей, кроме m-й компоненты, равной единице;

$\gamma^- = \gamma(1-\theta)$, $\gamma^+ = \gamma\theta$;

C_K – диагональная матрица с диагональными элементами $\{0,1,\ldots,K\}$;

$A = C_K(\gamma^- E_{K+1}^- - \gamma I_{K+1})$;

E_l^+, E_l^-, $\hat{I}_l$ и $\overline{I}_l$ – квадратные матрицы порядка l, состоящие из нулей, кроме элементов $(E_l^+)_{k,k+1} = 1$, $(\hat{I}_l)_{k,k} = 1, k = \overline{0,l-2}$, $(E_l^-)_{k,k-1} = 1, k = \overline{1,l-1}$, $(\overline{I}_l)_{0,0} = 1$;

$\tilde{I}_l = I_l - \hat{I}_l$;

$\otimes$ – символ кронекерова произведения матриц.

Введем обозначение для стационарных вероятностей цепи Маркова ξ_t:

$$\pi(i,r,k,\eta) = \lim_{t\to\infty} P\{i_t = i, r_t = r, k_t = k, \eta_t = \eta\}, i \geq 0, r = \overline{0,N}, k = \overline{0,K}, \eta = \overline{1,M}. \quad (1)$$

Далее перенумеруем состояния цепи в лексикографическом порядке и сгруппируем вероятности, соответствующие состоянию i первой компоненты, в вектор-строки $\boldsymbol{\pi}_i$, $i \geq 0$.

Известно, что векторы $\boldsymbol{\pi}_i$, $i \geq 0$, удовлетворяют следующей системе линейных алгебраических уравнений (уравнений Чепмена – Колмогорова):

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)Q = \mathbf{0}, (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)\mathbf{e} = 1, \quad (2)$$

где Q является генератором цепи Маркова $\xi_t, t \geq 0$.

Чтобы получить условие существования пределов (1) (условие эргодичности) и найти решение системы (2), необходимо получить точный вид генератора Q.

Лемма. *Генератор Q цепи Маркова $\xi_t, t \geq 0$, имеет следующую блочную трехдиагональную структуру:*

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

где ненулевые блоки $Q_{i,j}$ *определяются как*

$$Q_{i,i} = I_{N+1} \otimes (A - \lambda \hat{I}_{K+1} - (1-q)\lambda \tilde{I}_{K+1} - i\alpha \hat{I}_{K+1}) \otimes I_M + (I_{N+1} - \overline{I}_{N+1}) \otimes I_{K+1} \otimes S +$$
$$+ E^-_{N+1} \otimes I_{K+1} \otimes (\boldsymbol{S}_0 \boldsymbol{\beta}) + (E^+_{N+1} + \tilde{I}_{N+1}) \otimes (\gamma^+ C_K + \lambda E^+_{K+1}) \otimes I_M, \quad i \geq 0;$$

$$Q_{i,i+1} = (1-q)\lambda I_{N+1} \otimes \tilde{I}_{K+1} \otimes I_M, i \geq 0;$$

$$Q_{i,i-1} = i\alpha (E^+_{N+1} + \tilde{I}_{N+1}) \otimes E^+_{K+1} \otimes I_M, i \geq 1.$$

Доказательство леммы состоит в анализе возможных переходов цепи Маркова $\xi_t, t \geq 0,$ за бесконечно малый интервал времени и группировке соответствующих интенсивностей переходов в матричные блоки.

Цепь Маркова $\xi_t, t \geq 0,$ принадлежит классу асимптотически квазителицевых цепей Маркова с непрерывным временем, который введен в рассмотрение и исследован в [6].

Теорема. *Цепь Маркова* $\xi_t, t \geq 0,$ *является эргодической, если выполнено следующее неравенство:*

$$K\gamma^- > \tilde{\lambda}, \tag{3}$$

где

$$\tilde{\lambda} = (1-q)\lambda.$$

Доказательство. В работе [6] условие эргодичности асимптотически квазителицевых цепей Маркова выражается в терминах матриц

$$Y_0 = \lim_{i \to \infty} R_i^{-1} Q_{i,i-1}, \; Y_1 = \lim_{i \to \infty} R_i^{-1} Q_{i,i} + I, \; Y_2 = \lim_{i \to \infty} R_i^{-1} Q_{i,i+1},$$

где R_i – диагональная матрица с положительными диагональными элементами, определенными как модули соответствующих диагональных элементов матрицы $Q_{i,i}$.

Можно убедиться, что для рассматриваемой модели

$$R_i = I_{N+1} \otimes (\gamma C_K + (i\alpha + \lambda)\hat{I}_{K+1} + \tilde{\lambda}\tilde{I}_{K+1}) \otimes I_M + (I_{N+1} - \overline{I}_{N+1}) \otimes I_{K+1} \otimes \hat{S} - \gamma^+ \tilde{I}_{N+1} \otimes C_K \otimes I_M, i \geq 0,$$

где $\hat{S}$ – диагональная матрица с диагональными элементами матрицы $S,$ взятыми с противоположным знаком.

Матрицы $Y_k, k = 0,1,2,$ порядка $(N+1)(K+1)M$ принимают вид

$$Y_0 = (E^+_{N+1} + \tilde{I}_{N+1}) \otimes E^+_{K+1} \otimes I_M, \; Y_2 = \operatorname{diag}\{B_0, \ldots, B_N\},$$

$$Y_1 = \begin{pmatrix} D_0 & T_0 & O & \ldots & O \\ H_1 & D_1 & T_1 & \ldots & O \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ O & \ldots & H_{N-1} & D_{N-1} & T_{N-1} \\ O & \ldots & O & H_N & D_N \end{pmatrix},$$

где блочные матрицы B_n, D_n, T_n, H_n имеют все нулевые блоки, кроме блоков

$$(B_n)_{K,K} = \tilde{\lambda} L_n, \; n = \overline{0,N};$$

$$(D_n)_{K,K-1} = K\gamma^- L_n, \; (D_n)_{K,K} = L_n(-(K\gamma + \tilde{\lambda}) + \delta_{n,0} S) + I, \; n = \overline{0,N-1};$$

$$(T_n)_{K,K} = K\gamma^+ L_n, \; n = \overline{0,N-1};$$

$$(H_n)_{K,K} = L_n(\boldsymbol{S}_0 \boldsymbol{\beta}), \; n = \overline{1,N}.$$

Здесь $L_n = ((\tilde{\lambda} + K\gamma)I_M + \delta_{n,0}\hat{S})^{-1}$, $n = \overline{0, N-1}$;

$L_N = ((\tilde{\lambda} + K\gamma^-)I_M + \hat{S})^{-1}$;

$\delta_{i,j}$ – символ Кронекера, равный единице, если $i = j$, и нулю в противном случае.

Условие эргодичности из [6] может быть переписано в следующем виде: стационарное распределение цепи Маркова $\xi_t, t \geq 0,$ существует, если выполняется неравенство

$$\boldsymbol{y}Y_0\mathbf{e} > \boldsymbol{y}Y_2\mathbf{e}, \tag{4}$$

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

$$\boldsymbol{y}(Y_0 + Y_1 + Y_2) = \boldsymbol{y}, \ \boldsymbol{y}\mathbf{e} = 1. \tag{5}$$

Прямой подстановкой в (5) можно убедиться в том, что решение этой системы имеет вид

$$\boldsymbol{y} = (\boldsymbol{y}^{(0)}, \boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)}), \tag{6}$$

где элементы $\boldsymbol{y}_0^{(l)}, \ldots, \boldsymbol{y}_K^{(l)}$ вектора $\boldsymbol{y}^{(l)}, l = \overline{0, N},$ являются векторами размера M и определяются следующим образом:

$$\boldsymbol{y}_r^{(l)} = \mathbf{0}, \ r = \overline{0, K-2}, \ l = \overline{0, N};$$

$$\boldsymbol{y}_{K-1}^{(l)} = K\gamma^- \boldsymbol{y}_K^{(l)} L_l, \ l = \overline{0, N}; \tag{7}$$

$$\boldsymbol{y}_K^{(0)} = \boldsymbol{y}_K^{(0)}(-K\gamma L_0 + I) + \boldsymbol{y}_K^{(1)} L_1(\boldsymbol{S}_0\boldsymbol{\beta}); \tag{8}$$

$$\boldsymbol{y}_K^{(l)} = K\gamma \boldsymbol{y}_K^{(l-1)} L_{l-1} + \boldsymbol{y}_K^{(l)}(L_l(-K\gamma^- I_M + S) + I) + \boldsymbol{y}_K^{(l+1)} L_{l+1}(\boldsymbol{S}_0\boldsymbol{\beta}); \tag{9}$$

$$\boldsymbol{y}_K^{(N)} = K\gamma \boldsymbol{y}_K^{(N-1)} L_{N-1} + \boldsymbol{y}_K^{(N)}(L_N(-K\gamma^- I_M + S) + I) + \boldsymbol{y}_{K-1}^{(N)}. \tag{10}$$

Суммируя равенства (8)–(10), после некоторых алгебраических преобразований получаем равенство

$$(\sum_{l=1}^{N} \boldsymbol{y}_K^{(l)} L_l)(S + \boldsymbol{S}_0\boldsymbol{\beta}) = \mathbf{0}. \tag{11}$$

Так как матрица $S + \boldsymbol{S}_0\boldsymbol{\beta}$ является генератором, из (11) следует, что

$$\sum_{l=1}^{N} \boldsymbol{y}_K^{(l)} L_l = c\boldsymbol{\delta}, \tag{12}$$

где $\boldsymbol{\delta}$ – стохастический вектор, который служит единственным решением системы

$$\boldsymbol{\delta}(S + \boldsymbol{S}_0\boldsymbol{\beta}) = \mathbf{0}, \ \boldsymbol{\delta}\mathbf{e} = 1,$$

а c – некоторая ненулевая константа.

Из (7) и (12) следует, что

$$\sum_{l=1}^{N} \boldsymbol{y}_{K-1}^{(l)} = K\gamma^- \sum_{l=1}^{N} \boldsymbol{y}_K^{(l)} L_l = cK\gamma^- \boldsymbol{\delta}. \tag{13}$$

С помощью подстановки вектора $\boldsymbol{y}$ вида (6) в неравенство (4) после некоторых преобразований получаем неравенство

$$\sum_{l=0}^{N} \boldsymbol{y}_{K-1}^{(l)} \mathbf{e} > \tilde{\lambda} \sum_{l=0}^{N} \boldsymbol{y}_K^{(l)} L_l \mathbf{e},$$

что, в свою очередь, эквивалентно

$$(\boldsymbol{y}_{K-1}^{(0)} + K\gamma^{-}c\boldsymbol{\delta})\mathbf{e} > \tilde{\lambda}(\boldsymbol{y}_{K}^{(0)}L_0 + c\boldsymbol{\delta})\mathbf{e}. \qquad (14)$$

Так как $\boldsymbol{y}_{K-1}^{(0)} = K\gamma^{-}\boldsymbol{y}_{K}^{(0)}L_0$, неравенство (14) перепишется в виде

$$K\gamma^{-}(\boldsymbol{y}_{K}^{(0)}L_0 + c\boldsymbol{\delta})\mathbf{e} > \tilde{\lambda}(\boldsymbol{y}_{K}^{(0)}L_0 + c\boldsymbol{\delta})\mathbf{e},$$

что эквивалентно неравенству (3). Теорема доказана.

Неравенство (3) имеет очевидный смысл: цепь Маркова ξ_t является эргодической, если интенсивность освобождения жетонов в результате окончания поступления сессий превосходит среднюю интенсивность поступления сессий на орбиту. Стоит заметить, что условие (3) может быть использовано для ограничения допустимого числа жетонов снизу.

Интересно то, что средняя интенсивность обслуживания μ не входит в условие эргодичности. Это можно объяснить наличием конечного буфера для запросов и потерями запросов в случае занятости буфера, что никак не влияет на число сессий в системе.

В дальнейшем предполагаем, что условие (3) выполнено. Система (2) имеет бесконечную размерность, и нахождение ее решения является нетривиальной задачей. Эффективный и численно устойчивый алгоритм для вычисления векторов стационарного распределения асимптотически квазитеплицевых цепей Маркова приведен в [6].

Поскольку генератор Q в данной модели имеет блочную трехдиагональную форму, в то время как в [6] допускается более сложная блочная верхняя хессенбергова форма, для вычисления стационарных вероятностей можно использовать более простую модификацию алгоритма, приведенного в [6]. Алгоритм включает следующие шаги:

Шаг 1. Вычисляем матрицы G_i с помощью рекурсии:

$$G_i = [-(Q_{i+1,i+1} + Q_{i+1,i+2}G_{i+1})]^{-1}Q_{i+1,i},\ i \geq 0. \qquad (15)$$

Шаг 2. Вычисляем неотрицательные матрицы $F_i, i \geq 1,$ по формуле

$$F_0 = I_{N+1},\ F_i = \prod_{l=1}^{i} Q_{l-1,l}[-(Q_{l,l} + Q_{l,l+1}G_l)]^{-1},\ i \geq 1.$$

Шаг 3. Находим вектор $\boldsymbol{\pi}_0$ как единственное решение системы:

$$\boldsymbol{\pi}_0(Q_{0,0} + Q_{0,1}G_0) = \mathbf{0},\quad \boldsymbol{\pi}_0\sum_{l=0}^{\infty}F_l\mathbf{e} = 1.$$

Шаг 4. Вычисляем векторы $\boldsymbol{\pi}_i$ по формуле

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0F_i, i \geq 1.$$

Приведенный выше алгоритм определяет простой путь для нахождения векторов стационарного распределения $\boldsymbol{\pi}_i, i \geq 0,$ цепи Маркова $\xi_t, t \geq 0.$ Рекурсия (15) является обратной. Следовательно, для начала вычислений необходимо знать матрицу G_∞. Как следует из теории асимптотически квазитеплицевых цепей Маркова [6], последовательность матриц G_i сходится при i, стремящемся к бесконечности, к матрице G – минимальному неотрицательному решению матричного уравнения

$$G = Y_0 + Y_1G + Y_2G^2,$$

которое может быть найдено методом последовательных приближений. Из этого следует, что для любого малого $\varepsilon_G > 0$ существует такое i_0, что норма матрицы $G_i - G$ будет меньше ε_G для любых $i,\ i \geq i_0$.

Для нахождения i_0 необходимо выполнить следующие шаги:

Шаг 1. Зафиксировать малые значения ε_G – точность вычисления матриц G_i.

Шаг 2. Зафиксировать произвольное положительное целое число i^*.

Шаг 3. Положить $G_{i^*+1} = G$ и вычислить G_{i^*} из рекурсии (15).

Шаг 4. Сравнить норму $G_{i^*+1} - G_{i^*}$ со значением ε_G.

Если эта норма меньше ε_G, можно положить $i_0 = i^*$. В противном случае, если норма матрицы $G_{i^*+1} - G_{i^*}$ не меньше ε_G, увеличиваем значение i^*, например, умножая его на некоторый множитель κ, $\kappa > 1$, и переходим к *шагу 3*.

Матрицы $F_i, i \geq 0,$ содержат неотрицательные элементы, и, так как предполагается, что рассматриваемая цепь Маркова является эргодической, норма матриц F_i стремится к нулю при i, стремящемся к бесконечности. Таким образом, рекурсивные вычисления матриц F_i могут быть остановлены, когда норма матрицы F_i станет меньше некоторого заданного уровня ε_F.

3. Характеристики производительности системы

Вычислив стационарные векторы $\boldsymbol{\pi}_i$, $i \geq 0,$ и вектор $\boldsymbol{\pi} = \sum_{i=0}^{\infty} \boldsymbol{\pi}_i$, можно найти различные характеристики производительности рассматриваемой системы.

Среднее число сессий на орбите L вычисляется как

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

Среднее число сессий в системе $\bar{K}$ вычисляется по формуле

$$\bar{K} = \boldsymbol{\pi} \sum_{k=1}^{K} k (\mathbf{e}_{N+1} \otimes \mathbf{e}_K^{(k)} \otimes \mathbf{e}_M).$$

Среднее число запросов в системе $\bar{N}$ определяется как

$$\bar{N} = \boldsymbol{\pi} \sum_{r=1}^{N} r (\mathbf{e}_N^{(r)} \otimes \mathbf{e}_{K+1} \otimes \mathbf{e}_M).$$

Среднее число запросов T, обслуженных системой за единицу времени (пропускная способность системы), находится по формуле

$$T = \boldsymbol{\pi} \sum_{r=1}^{N} (\mathbf{e}_N^{(r)} \otimes \mathbf{e}_{K+1} \otimes \boldsymbol{S}_0).$$

Вероятность потери произвольной сессии на входе в систему $P_s^{(loss)}$ задается как

$$P_s^{(loss)} = q \boldsymbol{\pi} (\mathbf{e}_{N+1} \otimes \mathbf{e}_K^{(K)} \otimes \mathbf{e}_M).$$

Вероятность потери произвольного запроса из принятой сессии $P_c^{(loss)}$ находится согласно выражению

$$P_c^{(loss)} = \frac{\sum_{i=0}^{\infty} \sum_{k=1}^{K} \sum_{\eta=1}^{M} k \pi(i, N, k, \eta)}{\sum_{i=0}^{\infty} \sum_{r=0}^{N} \sum_{k=1}^{K} \sum_{\eta=1}^{M} k \pi(i, r, k, \eta)}.$$

4. Проблема оптимизации и численный эксперимент

С экономической точки зрения главной характеристикой исследованной модели является пропускная способность системы T, так как она определяет прибыль, полученную за передачу информации. Если число жетонов K растет, пропускная способность системы T увеличивается. Однако с ростом числа жетонов увеличивается и вероятность потери произвольного запроса из принятой сессии $P_c^{(loss)}$. Для того чтобы защитить интересы пользователей, необходимо, чтобы выполнялись условия на качество обслуживания. Будем предполагать, что вероятность потери произвольного запроса из принятой сессии не превосходит заданный параметр ε.

Таким образом, возникает задача оптимизации

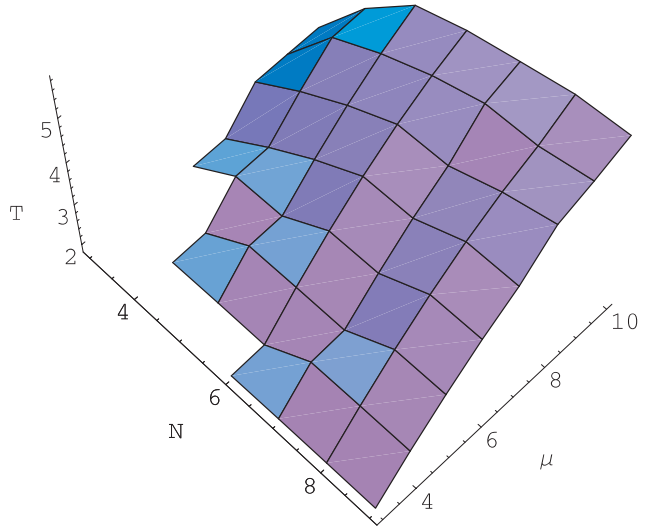
$$T = T(K) \to \max \tag{16}$$

при условии выполнения неравенств (3) и $P_c^{(loss)} = P_c^{(loss)}(K) \le \varepsilon$.

Значение ε зависит от толерантности системы к частичной потере запросов.

В численном эксперименте зафиксируем следующие параметры системы: $\lambda = 0,6$; $\alpha = 0,5$; $\beta = (0,7,\ 0,3)$; $S = \begin{pmatrix} -\mu & 0 \\ 0 & -\mu \end{pmatrix}$; $\gamma = 1$; $\theta = 0,9$; $q = 0,7$; $\varepsilon = 0,03$.

Из рисунка и таблицы видно, что оптимальная пропускная способность системы и оптимальное число сессий, которые могут находиться в системе одновременно, увеличиваются с ростом размерности системы и средней интенсивности обслуживания. Это очевидно, так как с ростом емкости буфера и ростом средней интенсивности обслуживания при фиксированных параметрах системы уменьшается вероятность потери запроса из принятой сессии $P_c^{(loss)}(K)$, поэтому ограничение $P_c^{(loss)}(K) \le \varepsilon$ остается справедливым для больших K (число сессий в системе). Пропускная способность $T(K)$ увеличивается с ростом емкости системы N и средней интенсивности обслуживания μ, поскольку уменьшается вероятность $P_c^{(loss)}$ и большее число запросов попадает на обслуживание. При увеличении числа сессий, которые могут находиться в системе одновременно (параметр K), пропускная способность системы также растет, так как увеличивается среднее число запросов, поступающих в систему за единицу времени, и прибор реже простаивает. Стоит отметить, что для малых значений μ и N ограничения задачи (16) несовместны и сама задача не имеет решения. Другими словами, в систему не может быть принята ни одна сессия при условии выполнения ограничений на качество обслуживания.



Зависимость средней пропускной способности системы T от средней интенсивности обслуживания μ и размерности системы N при оптимальном числе жетонов K^*

Оптимальное число жетонов K^* при фиксированных значениях интенсивности обслуживания μ и размерности системы N

K^*	$N=3$	$N=4$	$N=5$	$N=6$	$N=7$	$N=8$	$N=9$
$\mu=3$	–	–	–	2	2	2	2
$\mu=4$	–	2	2	2	3	3	3
$\mu=5$	–	2	3	3	3	4	4
$\mu=6$	2	3	3	4	4	5	5
$\mu=8$	3	4	5	6	6	7	8
$\mu=10$	3	5	7	8	9	11	13

Заключение

В статье исследована система массового обслуживания M|PH|1|N с повторными вызовами сессий. Найдено совместное распределение числа сессий и запросов в системе, получены формулы для вычисления основных характеристик производительности системы, численно решена задача оптимизации максимального числа сессий, которые могут обслуживаться в системе одновременно. Полученные в статье результаты могут быть использованы для моделирования и оптимизации процессов передачи информации в современных телекоммуникационных сетях.

Данная работа была частично поддержана Белорусским республиканским фондом фундаментальных исследований через грант № Ф11М-003.

Список литературы

1. Kist, A.A. A simple IP flow blocking model / A.A. Kist, B. Lloyd-Smith, R.J. Harris // Proc. 19th International Teletraffic Congress. – Beiging, 2005. – P. 355–364.

2. Lee, M.H. Queueing Model with Time-Phased Batch Arrivals / M.H. Lee, S. Dudin, V. Klimenok // Lecture Notes in Computer Science. – 2007. – Vol. 4516. – P. 716–730.

3. Artalejo, J.R. Retrial queueing systems: A computational approach / J.R. Artalejo, A. Gomez-Corral. – Berlin-Heidelberg : Springer, 2008. – 318 p.

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

5. Falin, G. Retrial queues / G. Falin, J. Templeton. – London : Chapman and Hall, 1997. – 328 p.

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

Поступила 25.05.11

Белорусский государственный университет,
Минск, пр. Независимости, 4
email: dudin85@mail.ru,
taramin@mail.ru

S.A. Dudin, O.S. Taramin

QUEUING SYSTEM M|PH|1|N WITH SESSION RETRIALS

A single-server queuing system with a finite buffer and a session arrival of customers is considered. An admission of session is regulated through so-called tokens. A pool of tokens is finite. A session which arrives when the tokens are not available leaves the system forever or joins the orbit. A steady state distribution of the number of sessions and customers in the system is analyzed; the formulas for calculation of the main performance measures are derived. The numerical results are presented; the optimization problem of throughput is numerically solved.