Вестник РУДН Серия *Математика. Информатика. Физика.* № 3 (1). 2010. С. 34–46

Теория массового обслуживания и сети телекоммуникаций

УДК 519.21

Стационарные характеристики многоканальной неоднородной системы с FCFS орбитой и пороговым управлением

Д. В. Ефросинин

Кафедра теории вероятностей и математической статистики
Российский университет дружбы народов
ул. Миклухо-Маклая, 6, 117198 Москва, Россия

В работе рассматривается марковская система массового обслуживания с несколькими неоднородными приборами. Включение приборов производится в соответствии с пороговой политикой управления. Заявки, которые не получили обслуживания, направляются на орбиту, где они через случайное время повторяют попытку занять прибор. Заявки на орбите формируют очередь с FCFS дисциплиной обслуживания. Система описывается обобщённым процессом размножения и гибели с большим числом пограничных состояний. В работе проводится анализ системы в стационарном режиме, выводятся матричные выражения для вычисления средних характеристик производительности системы, а также формулы оптимальных пороговых уровней.

Ключевые слова: СМО с повторными заявками, управляемая очередь, стационарный режим, условие эргодичности, характеристики производительности.

1. Введение

Различные типы одноканальных и многоканальных систем массового обслуживания с повторными заявками широко применяются для моделирования локальных компьютерных сетей, коммуникационных протоколов и телефонных систем. Подробный обзор литературы по СМО с повторными заявками может быть найден, например, в работах [1, 2]. Главным отличием этих систем от классических моделей является то, что заявка, получившая отказ на обслуживание, направляется на орбиту, откуда через случайное время она повторяет попытку занять прибор. Большинство систем с повторными заявками имеют *классическую дисциплину повторных заявок*. В этом случае время до повторного поступления заявки с орбиты имеет экспоненциальное распределение с параметром, зависящим от числа заявок на орбите.

Тем не менее некоторые компьютерные сети не удовлетворяют этому предположению. Иногда серверу необходимо проверять доступность передающего устройства или перед началом обслуживания необходима задержка для поиска заявки сервером. В данном случае речь идёт об FCFS орбите с *дисциплиной постоянных повторов*, когда интенсивность повторного поступления заявки не зависит от состояния орбиты. Дисциплина постоянных повторов выбрана для изучаемой в статье системы.

Пример практического применения таких систем в коммуникационных сетях можно найти, например, в работах [3, 4], где изучалась стабильность протоколов ALOHA и CSMA/CD.

Многоканальные системы с повторными заявками подробно исследовались для случая однородных приборов, см., например [5, 6]. Однако для неоднородных приборов количество статей довольно ограничено, см. например [7, 8], несмотря на то, что такие системы активно используются в качестве моделей для многих реальных систем. Примером таких систем могут служить группы серверов с процессорами различной производительности, узлы телекоммуникационных сетей с гибридными каналами связи и т.д. Принимая во внимание обширную область

применения СМО с неоднородными приборами, а также наличие значительных пробелов в области их теоретического исследования, целью данной статьи выбраны именно такие системы.

Заметим, для СМО с неоднородными приборами необходимо определять порядок включения приборов. В тех немногочисленных работах по системам с повторными заявками и неоднородными приборами, которые удалось найти, рассматривались только эвристические политики, такие как, например, включение самого быстрого прибора и случайный выбор прибора. Однако, как было доказано в [9] и численно подтверждено в [10], оптимальной политикой управления, минимизирующей среднее число заявок в системе без повторных заявок, является пороговая политика. Для систем с классической политикой повторных заявок аналогичный результат подтверждён в [11], поэтому в данной работе рассматривается СМО именно с пороговой дисциплиной обслуживания, которая дополнительно сравнивается с указанными выше эвристическими политиками управления. Алгоритмы вычисления характеристик производительности для системы без повторных заявок представлен в работе [12].

Проведённый анализ СМО с дисциплиной постоянных повторов и неоднородными приборами включает следующие результаты:

(а) Система моделируется с помощью обобщённого ПРГ с инфинитезимальной блочной трёхдиагональной матрицей.
(б) Выводится условие существования стационарного режима.
(в) Вычисляются значения стационарных вероятностей состояний и средние характеристики производительности системы.
(г) Формулируется задача оптимизации и вычисляются оптимальные пороговые уровни.

Статья организована следующим образом. В разделе 2 описывается математическая модель системы. Стационарное распределение состояний системы, а также формулы основных средних характеристик системы приводятся в разделе 3. Задача оптимизации формулируется в разделе 4. Раздел 5 содержит численные эксперименты и их краткое обсуждение. Раздел 6 заключает данную статью.

Далее в работе использованы обозначения $\boldsymbol{e}(n)$, $\boldsymbol{e}_j(n)$ и I_n соответственно для единичного вектор-столбца размера n, нулевого вектор-столбца размера n с 1 на j-м месте (начиная с 0), единичной матрицы размера n. Если указание размера не имеет большого значения, будем опускать аргумент в скобках, тогда $\boldsymbol{e}$ обозначает единичный вектор соответствующего размера. Далее, $\mathrm{diag}\,(A_1, \ldots, A_n)$ обозначает диагональную матрицу с блочными элементами $A_1, \ldots, A_n$, $\mathrm{diag}^{+}(A_1, \ldots, A_n)$ и $\mathrm{diag}^{-}(A_1, \ldots, A_n)$ — соответственно, нулевые матрицы с над- и поддиагональными блочными элементами $A_1, \ldots, A_{n-1}$. Будем использовать также стандартное обозначение $\mathbf{1}_{\{A\}}$ для индексной функции, принимающей значение 1 при выполнении события A и 0 — в противном случае. Верхний символ t используется для транспонированных векторов.

2. Математическая модель

Рассмотрим систему $M|M|K$, в которой новые заявки формируют пуассоновский поток с интенсивностью $\lambda > 0$, $K > 1$ неоднородных приборов обслуживают заявки с интенсивностями $\mu_1 > \ldots > \mu_K > 0$, и задана политика управления включением приборов посредством пороговых уровней $1 = q_1 \leqslant \ldots \leqslant q_K < q_{K+1} = \infty$, представляемые в виде $(q_2, \ldots, q_K)$. Согласно пороговой политике управления новая заявка в момент поступления направляется сразу на свободный прибор с индексом $i = \overline{1, k-1}$, имеющим наибольшую интенсивность обслуживания, если число ожидающих обслуживания заявок q находится в интервале $q = \overline{q_{k-1} - 1, q_k - 2}$, $k = \overline{2, K}$. Если при данном числе ожидающих заявок все приборы заняты, то поступившая заявка направляется на орбиту с бесконечной ёмкостью. Заявка на орбите повторяет попытку занять прибор через экспоненциально распределённое время с параметром $\gamma > 0$. Заявки на орбите образуют FCFS очередь. В момент поступления повторной заявки включается свободный

прибор с индексом $i = \overline{1, k-1}$, имеющим наибольшую интенсивность обслуживания, если $q = \overline{q_{k-1}, q_k - 1}$, $k = \overline{2, K}$. В противном случае попытка считается неудачной, и заявка возвращается на первое место на орбите, где ожидает новой попытки. Случайные времена между поступлениями новых и повторных заявок, а также времена обслуживания на приборах предполагаются взаимно независимыми.

Пусть $Q(t)$ обозначает число заявок на орбите и $D(t) = (D_1(t), \ldots, D_K(t))-$ вектор состояний приборов в момент времени t, причём

$$D_k(t) = \begin{cases} 0, & \text{если прибор } k \text{ свободен,} \\ 1, & \text{если прибор } k \text{ занят,} \end{cases} \qquad k = \overline{1, K}.$$

Очевидно, что $K+1$-мерный процесс

$$\{X(t)\}_{t \geqslant 0} = \{Q(t), D(t)\}_{t \geqslant 0} \tag{1}$$

представляет собой неприводимую регулярную цепь Маркова с непрерывным временем, определённую на пространстве состояний $E = \{x = (q, d); q \geqslant 0,\ d \in D\}$. Здесь $D = \{(d_1, \ldots, d_K); d_k \in \{0, 1\},\ k = \overline{1, K}\}$ представляет собой множество из 2^K элементов, обозначающих набор состояний приборов. Переменные q, d и d_k обозначают соответственно число заявок на орбите и вектор состояний приборов и состояние прибора с индексом k. Далее также будем использовать обозначения $q(x)$, $d(x)$ и $d_k(x)$, если необходимо подчеркнуть принадлежность данных элементов к определённому состоянию $x \in E$.

Перенумеруем состояния цепи Маркова в лексикографическом порядке и далее считаем $\mathbf{i}$ состоянием цепи Маркова, учитывая, что это макросостояние, состоящее из 2^K состояний $\mathbf{i} = \{(i, d); d \in D\}$, $i \geqslant 0$. Порядковый номер состояния приборов d для некоторого фиксированного i обозначим через $\#d = \overline{0, 2^K - 1}$.

3. Стационарное распределение состояний системы

Пусть при некотором условии, которое будет определено позже, существует стационарный режим цепи Маркова $\{X(t)\}_{t \geqslant 0}$.

Обозначим через $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$ макро-вектор строку стационарных вероятностей состояний системы, где $\boldsymbol{\pi}_i$ обозначает суб-вектор, состоящий из вероятностей π_x с $q(x) = i$, соответствующий макросостоянию $\mathbf{i}$,

$$\pi_x = \lim_{t \to \infty} \mathbb{P}\{X(t) = x\}.$$

Для вычисления стационарного распределения состояний системы воспользуемся аппаратом матрично-аналитических решений, принимая во внимание специальною структуру граничных состояний в соответствии с заданной пороговой политикой управления.

Лемма 1. *Если выполнено условие существования стационарного режима, то вектор $\boldsymbol{\pi}$ является единственным решением системы*

$$\boldsymbol{\pi}\Lambda = \mathbf{0},\ \boldsymbol{\pi}\boldsymbol{e} = 1,$$

где Λ является инфинитезимальной матрицей цепи Маркова $\{X(t)\}_{t \geqslant 0}$, которая зависит от пороговых уровней q_j, $j = \overline{1, K}$ и имеет трёхдиагональную блочную структуру (2)

$$\Lambda = \operatorname{diag}(Q_{1,0}, \underbrace{Q_{1,1}, \ldots, Q_{1,1}}_{q_2 - 2}, Q_{1,2}, \ldots, \underbrace{Q_{1,2j-1}, \ldots, Q_{1,2j-1}}_{q_{j+1} - q_j - 1}, Q_{1,2j}, \ldots, Q_{1,2K-1}, \ldots) +$$

$$+\operatorname{diag}^{+}(\underbrace{Q_{0,1},\ldots,Q_{0,1}}_{q_2-1},\ldots,\underbrace{Q_{0,j},\ldots,Q_{0,j}}_{q_{j+1}-q_j},\ldots,Q_{0,K},\ldots)+$$
$$+\operatorname{diag}^{-}(\underbrace{Q_{2,1},\ldots,Q_{2,1}}_{q_2-1},\ldots,\underbrace{Q_{2,j},\ldots,Q_{2,j}}_{q_{j+1}-q_j},\ldots,Q_{2,K},\ldots),\quad j=\overline{1,K-1},\quad (2)$$

где

$$(Q_{1,0}+Q_{0,1})\boldsymbol{e}=\mathbf{0},\quad (Q_{2,j}+Q_{1,2j-1}+Q_{0,j})\boldsymbol{e}=\mathbf{0},\quad j=\overline{1,K},$$
$$(Q_{2,j-1}+Q_{1,2j-2}+Q_{0,j})\boldsymbol{e}=\mathbf{0},\quad j=\overline{2,K}.$$

Блоки $Q_{1,j}$, $Q_{2,j}$ и $Q_{0,j}$ представляют собой квадратные матрицы размера 2^K. Блоки $Q_{1,j}$, $j=\overline{0,2K-1}$ имеют трёхдиагональную блочную структуру (3).

$$Q_{1,j}=\begin{pmatrix} A_{1,0}^{(j)} & A_{0,1}^{(j)} & 0 & 0 & \ldots & 0\\ A_{2,1} & A_{1,1}^{(j)} & A_{0,2}^{(j)} & 0 & \ldots & 0\\ 0 & A_{2,2} & A_{1,2}^{(j)} & A_{0,3}^{(j)} & \ldots & 0\\ \ddots & \ddots & \ddots & \ddots & \ddots & \\ 0 & \ldots & 0 & A_{2,K-1} & A_{1,K-1}^{(j)} & A_{0,K}^{(j)}\\ 0 & \ldots & 0 & 0 & A_{2,K} & A_{1,K}^{(j)} \end{pmatrix},\quad j=\overline{0,2K-1}. \quad (3)$$

Прямоугольные блоки $A_{2,l}$, $l=\overline{1,K}$, имеют размер $\binom{K}{l}\times\binom{K}{l-1}$ и содержат интенсивности обслуживания в зависимости от набора занятых приборов:

$$A_{2,l}=B_{l-1,l},\ l=\overline{1,K},$$

где

$$B_{i,l}=\begin{pmatrix} B_{i-1,l} & & & & I_{S_{i,l}}\mu_{l-i}\\ 0 & B_{i-1,l+1} & & & I_{S_{i,l+1}}\mu_{l+1-i}\\ \ddots & \ddots & \ddots & \ddots & \\ 0 & \ldots & 0 & B_{i-1,K} & I_{S_{i,K}}\mu_{K-i} \end{pmatrix},\quad i=\overline{1,K-1},$$

$$B_{0,l}=(\mu_l,\mu_{l+1},\ldots,\mu_K)^t,\quad S_{i,l}=\binom{K-l+i}{i}.$$

Диагональные блоки $A_{1,l}^{(j)}$, $l=\overline{0,K}$, содержат интенсивности пребывания в соответствующих состояниях и представляют собой диагональные матрицы размера $\binom{K}{l}\times\binom{K}{l}$ с элементами, зависящими от числа заявок на орбите:

$$A_{1,l}^{(j)}=-\left(\lambda+\sum_{k=1}^{K}d_k(x)\mu_k+\gamma\sum_{k=1}^{K}\mathbf{1}_{\{j\in\{2k-1,2k\},\,d_i(x)=1,\,i=\overline{1,k-1},\,d_k(x)=0\}}\right)I_{S_{l,l}}.$$

Порядковый номер состояния $\#x$ определяет номер строки и столбца соответствующего элемента.

Прямоугольные блоки $A_{0,l}^{(j)}$, $l=\overline{1,K}$ размера $\binom{K}{l}\times\binom{K}{l-1}$ содержат интенсивности поступления новых требований в зависимости от числа заявок

на орбите:

$$A_{0,l}^{(j)} = \\ = \lambda \sum_{k=1}^{K} \mathbf{1}_{\{j\in\{2k-2,2k-1\},\, d_i(x)=d_i(y)=1,\, i=\overline{1,k-1},\, d_i(x)=d_i(y),\, i=\overline{k+1,K},\, d_k(x)=1,\, d_k(y)=0\}} \times \\ \times J_{S_{l,l}\times lS_{l,l}/S_{1,l}},$$

где матрица J состоит из единиц, а порядковые номера состояний $\#x$ и $\#y$ определяют соответственно столбец и строку элемента.

Блоки $Q_{2,j}$ и $Q_{0,j}$, $j=\overline{1,K}$ имеют вид:

$$Q_{2,j} = \frac{\gamma}{\lambda} \begin{pmatrix} 0 & A_{0,1}^{(j)} & 0 & 0 & \dots & 0 \\ 0 & 0 & A_{0,2}^{(j)} & 0 & \dots & 0 \\ 0 & 0 & 0 & A_{0,3}^{(j)} & \dots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \\ 0 & \dots & 0 & 0 & 0 & A_{0,K}^{(j)} \\ 0 & \dots & 0 & 0 & 0 & 0 \end{pmatrix}, \quad j=\overline{1,K},$$

$$Q_{0,j} = \lambda \left(1 - \sum_{k=1}^{K} \mathbf{1}_{\{j\in\{2k-1,2k\},\, d_i(x)=1, i=\overline{1,k-1},\, d_k(x)=0\}} \right) I_{2^K}, \quad j=\overline{1,K}.$$

Доказательство. Обозначим через $A_{q_k}(x)$ и $\overline{A}_{q_k}(x)$, $k=\overline{1,K}$, следующие события и их дополнения

$$A_{q_k}(x) = \{q(x) \geqslant q_k,\, d_i(x)=1,\, i=\overline{1,k-1},\, d_k(x)=0\},$$
$$\overline{A}_{q_k}(x) = \{q(x) \leqslant q_k-1,\, d_i(x)=1,\, i=\overline{1,k-1},\, d_k(x)=0\}.$$

Анализируя переходы цепи Маркова $\{X(t)\}_{t\geqslant 0}$, с учётом введённых обозначений, получим следующую систему уравнений равновесия для состояний $x\in E$,

$$\left(\lambda + \sum_{k=1}^{K} d_k(x)\mu_k + \gamma \sum_{k=1}^{K} \mathbf{1}_{\{A_{q_k}(x)\}} \right) \pi_x = \\ = \lambda \left(\sum_{k=1}^{K} \pi_{x-\boldsymbol{e}_k} \mathbf{1}_{\{A_{q_k-1}(x-\boldsymbol{e}_k)\}} + \pi_{x-\boldsymbol{e}_0} \mathbf{1}_{\{\overline{A}_{q_k-1}(x-\boldsymbol{e}_0)\}} \right) + \\ + \sum_{k=1}^{K} (1-d_k(x))\mu_k \pi_{x+\boldsymbol{e}_k} + \gamma \sum_{k=1}^{K} d_k(x) \pi_{x+\boldsymbol{e}_0-\boldsymbol{e}_k} \mathbf{1}_{\{A_{q_k}(x+\boldsymbol{e}_0-\boldsymbol{e}_k)\}}.$$

Далее, группируя интенсивности соответствующих переходов в блочные матрицы, получим искомые матричные соотношения. □

Теорема 1. *Необходимое и достаточное условие существования стационарного режима цепи Маркова $\{X(t)\}_{t\geqslant 0}$ задаётся следующим неравенством*

$$\rho = \frac{\lambda}{\gamma} \left(\sum_{i=1}^{K-1} \prod_{j=1}^{K-i} L_{K-j}\, \boldsymbol{e} \right)^{-1} < 1, \tag{4}$$

где выражение в скобках представляет собой скаляр. Здесь вектор-строка L_{K-1}, а также матрицы L_i, $i = \overline{1, K-2}$ удовлетворяют соотношениям

$$\begin{aligned} L_0 &= -A_{2,1}\left(A_{1,0}^{(2K-1)}\right)^{-1}, \\ L_i &= -A_{2,i+1}\left(\frac{\lambda+\gamma}{\lambda}L_{i-1}A_{0,i}^{(2K-1)} + A_{1,i}^{(2K-1)}\right)^{-1}, \quad i = \overline{1, K-1}. \end{aligned} \tag{5}$$

Доказательство. Из теории обобщённых ПРГ (см., например [13] известно, что необходимое и достаточное условие существования стационарного режима цепи Маркова $\{X(t)\}_{t\geqslant 0}$ представляется в виде

$$\mathbf{p}\,Q_{0,K}\boldsymbol{e}(2^K) < \mathbf{p}\,Q_{2,K}\boldsymbol{e}(2^K), \tag{6}$$

где вектор $\mathbf{p}$ имеет размер 2^K и удовлетворяет системе $\mathbf{p}\,(Q_{0,K}+Q_{1,2K-1}+Q_{2,K}) = \mathbf{0}$ и $\mathbf{p}\boldsymbol{e} = 1$. Матрицы $Q_{0,K}, Q_{1,2K-1}, Q_{2,K}$ являются матрицами интенсивностей переходов однородной части ПРГ, причём

$$Q_{0,K} + Q_{1,2K-1} + Q_{2,K} =$$
$$= \begin{pmatrix} A_{1,0}^{(2K-1)} & \frac{\lambda+\gamma}{\lambda}A_{0,1}^{(2K-1)} & 0 & 0 & \dots & 0 \\ A_{2,1} & A_{1,1}^{(2K-1)} & \frac{\lambda+\gamma}{\lambda}A_{0,2}^{(2K-1)} & 0 & \dots & 0 \\ 0 & A_{2,2} & A_{1,2}^{(2K-1)} & \frac{\lambda+\gamma}{\lambda}A_{0,3}^{(2K-1)} & \dots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \\ 0 & \dots & 0 & A_{2,K-1} & A_{1,K-1}^{(2K-1)} & \frac{\lambda+\gamma}{\lambda}A_{0,K}^{(2K-1)} \\ 0 & \dots & 0 & 0 & A_{2,K} & A_{1,K}^{(2K-1)} + \lambda \end{pmatrix},$$
$$j = \overline{0, 2K-1}.$$

Макро-вектор $\mathbf{p} = (\mathbf{p}_0, \dots, \mathbf{p}_K)$ состоит из векторов $\mathbf{p}_i = (p_{(d_1,\dots,d_K)} : \sum\limits_{j=1}^{K} d_j = i)$, $i = \overline{0, K}$, размера $\binom{K}{i}$, компоненты которых упорядочены в лексикографическом порядке. С помощью обратной подстановки нетрудно получить решение системы для вектора $\mathbf{p}$ в виде

$$\mathbf{p}_i = \left[\sum_{i=0}^{K-1}\prod_{j=1}^{K-i} L_{K-j}\,\boldsymbol{e} + 1\right]^{-1} \prod_{j=1}^{K-i} L_{K-j}, \quad i = \overline{0, K},$$

где матрицы L_i, $i = \overline{0, K-1}$ удовлетворяют соотношениям (5). Подставляя последние соотношения для элементов вектора $\mathbf{p}$ в неравенство (6) и учитывая, что $Q_{0,K}\boldsymbol{e}(2^K) = \lambda\boldsymbol{e}_{2^K-1}(2^K)$ и $Q_{2,K}\boldsymbol{e}(2^K) = \gamma(\boldsymbol{e}(2^K) - \boldsymbol{e}_{2^K-1}(2^K))$, получим условие (4). □

Теорема 2. *Векторы $\boldsymbol{\pi}_i$, $i \geqslant 0$ стационарных вероятностей состояний системы вычисляются по формулам*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_{q_K}\prod_{j=1}^{q_K-i} M_{q_K-j},\ i = \overline{0, q_K-1}, \quad \boldsymbol{\pi}_i = \boldsymbol{\pi}_{q_K}R^{i-q_K},\ i \geqslant q_K, \tag{7}$$

где матрицы M_i удовлетворяют соотношениям

$$
\begin{gathered}
M_0 = -Q_{2,1}Q_{1,0}^{-1}, \\
M_i = -Q_{2,j-1}\left(M_{i-1}Q_{0,j-1} + Q_{1,2j-3}\right)^{-1},\ i = \overline{q_{j-1}, q_j - 2}, \\
M_{q_j-1} = -Q_{2,j}\left(M_{q_j-2}Q_{0,j-1} + Q_{1,2j-2}\right)^{-1}
\end{gathered} \tag{8}
$$

для $j = \overline{2, K}$. $\boldsymbol{\pi}_{q_K}$ *является единственным решением системы уравнений*

$$
\begin{gathered}
\boldsymbol{\pi}_{q_K}\left[\sum_{i=0}^{q_K-1}\prod_{j=1}^{q_K-i} M_{q_K-j} + (I-R)^{-1}\right]\boldsymbol{e} = 1, \\
\boldsymbol{\pi}_{q_K}\left(M_{q_K-1}Q_{0,K} + Q_{1,2K-1} + RQ_{2,K}\right) = \mathbf{0}.
\end{gathered} \tag{9}
$$

Матрица R *является минимальным неотрицательным решением квадратичного матричного уравнения*

$$
R^2Q_{2,K} + RQ_{1,2K-1} + Q_{0,K} = 0. \tag{10}
$$

Доказательство. Очевидно, что приведённые в формулах (8) обратные матрицы существуют. Вероятности состояний ниже порогового уровня q_K можно выразить в виде $\boldsymbol{\pi}_i = \boldsymbol{\pi}_{i+1}M_i,\ i = \overline{0, q_K - 1}$, используя обратную подстановку. Матрично-аналитическое решение для состояний выше порогового уровня q_K задаётся в виде $\boldsymbol{\pi}_i = \boldsymbol{\pi}_{q_K}R^{i-q_K},\ i \geqslant q_K$, где матрица R удовлетворяет квадратичному уравнению (10). Это уравнение обычно решается используя рекуррентную процедуру: $R_0 = 0,\ R_{n+1} = -\left(Q_{0,K}Q_{1,2K-1}^{-1} + R_n^2Q_{2,K}Q_{1,2K-1}^{-1}\right)$.

Из условия нормировки вместе с граничным уравнением (9) получаем выражение для вероятности $\boldsymbol{\pi}_{q_K}$, которое может быть подставлено в выражения для остальных вероятностей состояний системы $\boldsymbol{\pi}_i$. □

Найдя векторы стационарных вероятностей состояний $\boldsymbol{\pi}_i,\ i \geqslant 0$, можно вычислить различные характеристики производительности системы.

Утверждение 1. Вероятность загрузки системы:

$$
\bar{U} = 1 - \boldsymbol{\pi}_0\boldsymbol{e}_0. \tag{11}
$$

Вероятность загрузки j-го прибора:

$$
\bar{U}_j = \left[\sum_{i=0}^{q_K-1}\boldsymbol{\pi}_i + \boldsymbol{\pi}_{q_K}(I-R)^{-1}\right]\sum_{\substack{\#d=1 \\ d_j=1}}^{2^K-1}\boldsymbol{e}_{\#d}. \tag{12}
$$

Среднее число занятых приборов:

$$
\bar{C} = \sum_{j=1}^{K}\bar{U}_j. \tag{13}
$$

Среднее число заявок на орбите:

$$
\bar{Q} = \left[\sum_{i=0}^{q_K-1} i\boldsymbol{\pi}_i + \boldsymbol{\pi}_{q_K}(R(I-R)^{-1} + q_KI)(I-R)^{-1}\right]\boldsymbol{e}. \tag{14}
$$

Среднее число заявок в системе:

$$
\bar{N} = \bar{C} + \bar{Q}. \tag{15}
$$

Среднее время ожидания и пребывания заявки в системе:

$$\overline{W} = \frac{\overline{Q}}{\lambda}, \quad \overline{T} = \frac{\overline{N}}{\lambda}. \tag{16}$$

Вероятность блокировки (поступившая заявка отправляется на орбиту):

$$P_{\text{blocking}} = \sum_{k=2}^{K} \sum_{i=q_{k-1}-1}^{q_k-2} \sum_{\substack{\#d=1 \\ d_j=1, j=\overline{1,k-1}}}^{2^K-1} \boldsymbol{\pi}_i \boldsymbol{e}_{\#d} + (\boldsymbol{\pi}_{q_K-1} + \boldsymbol{\pi}_{q_K}(I-R)^{-1})\boldsymbol{e}_{2^K-1}. \tag{17}$$

Среднее время занятости:

$$\overline{L} = \frac{1}{\lambda}\left(\frac{1}{\boldsymbol{\pi}_0 \boldsymbol{e}_0} - 1\right). \tag{18}$$

4. Задача оптимизации

Средние характеристики производительности системы, представленные в разделе 2, зависят от пороговых уровней $q_k, k = \overline{2, K}$. Возникает естественный вопрос вычисления оптимальных значений q_k^*. Одной из наиболее важных характеристик системы является среднее число заявок в системе $\bar{N}$, которую необходимо минимизировать. Таким образом, в данном разделе решается следующая задача оптимизации: $\bar{N} = \bar{N}(\lambda, \mu_1, \ldots, \mu_K, \gamma, q_2, \ldots, q_K) \to \min_{q_2,\ldots,q_K}$.

В работах [9, 14] представлены алгоритмы вычисления оптимальных пороговых уровней с помощью двухшаговой итерационной процедуры динамического программирования. Также может быть использован простой метод перебора. Для эквивалентной задачи расписания удаётся получить аналитические результаты.

Теорема 3. *Для задачи расписания существуют оптимальные пороговые уровни*

$$q_k^* = \left\lfloor \frac{\gamma}{\sum_{j=1}^{k-1} \mu_j + \gamma} \left(\frac{\sum_{j=1}^{k-1} \mu_j}{\mu_k} - (k-1) \right) \right\rfloor, \quad k = \overline{1, K}, \tag{19}$$

имеющие следующий смысл: если $q = \overline{q_{k-1}^*, q_k^* - 1}$, тогда заявка при повторной попытке занимает самый быстрый из свободных приборов с индексом $j = \overline{1, k-1}$. Если свободных приборов с данными индексами нет, тогда повторная заявка направляется обратно на орбиту.

Доказательство. Пусть свободные приборы с индексами $j = \overline{1, k-1}$ необходимо включать при поступлении повторной заявки. Определим число заявок на орбите q_k, при котором очередная повторная заявка при всех занятых приборах с индексами $j = \overline{1, k-1}$, должна быть направлена на прибор с индексом k. Для этого необходимо, чтобы сумма среднего времени пребывания $k-1$-ой заявки на занятых приборах до момента изменения состояния и среднего времени ожидания q_k заявок на орбите превышало среднее время обслуживания на приборе k, т.е.

$$\frac{k-1}{\sum_{j=1}^{k-1} \mu_j} + q_k \left(\frac{1}{\gamma} + \frac{1}{\sum_{j=1}^{k-1} \mu_j} \right) \geqslant \frac{1}{\mu_k}.$$

Таким образом, прибор с индексом k необходимо включать при повторной попытке поступления заявки с орбиты всякий раз, когда первые $k-1$ прибора заняты, а число заявок на орбите удовлетворяет неравенству $q_k \geqslant q_k^*$. □

В общем случае, когда $\lambda > 0$, можно получить приближенные формулы для вычисления оптимальных пороговых уровней. Для этого подбираются асимптотические гиперповерхности в виде функций от параметров $\frac{\mu_k}{\lambda}, k = \overline{2,K}$ и $\frac{\gamma}{\lambda}$ для границ областей, где оптимальный порог k-го прибора равняется q_k^* и q_k^*+1. Пример областей оптимальности пороговых уровней $q_2^* = 1,2,3,4$, в зависимости от параметров системы, а также поверхности для границ между областями изображены на рис. 1. Структура формулы (19) позволяет сделать предположение, что $q_k^* = \frac{\gamma}{\gamma+\sum_{j=1}^{k-1}\mu_j}\hat{q}_k^*$, где $\hat{q}_k^*$ обозначает оптимальный пороговый уровень стандартной системы $M/M/K$ с неоднородными приборами. В следующем утверждении приводятся формулы, полученные эмпирическим путём.

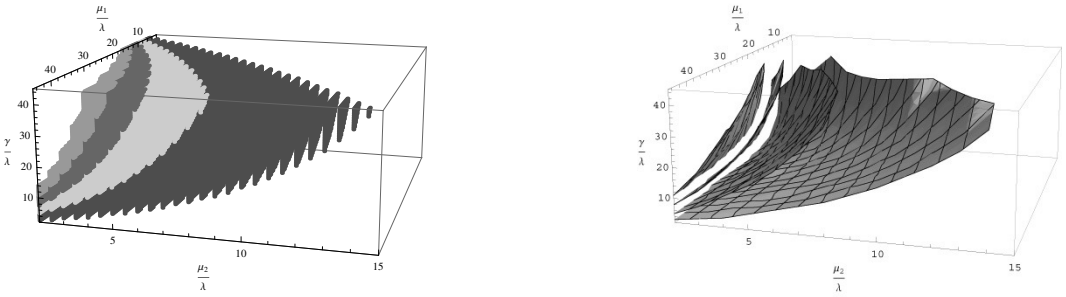


Рис. 1. **Области оптимальности пороговых уровней $q_2^* = 1,2,3,4$ и соответствующие границы**

Утверждение 2. Используя функциональное асимптотическое представление границ между областями оптимальности с пороговыми уровнями q_k^* и q_k^*+1, получены следующие приближенные формулы, дающие ошибку производительности не более 5 процентов

$$q_k^* \approx \tilde{q}_k^* = \\ = \left\lfloor \frac{\gamma}{\gamma+\sum_{j=1}^{k-1}\mu_j}\left(\frac{\sum_{j=1}^{k-1}\mu_j - \lambda + \sqrt{(\sum_{j=1}^{k-1}\mu_j-\lambda)^2 + 4(k-1)\mu_k\lambda}}{2\mu_k} - (k-1)\right)\right\rfloor. \tag{20}$$

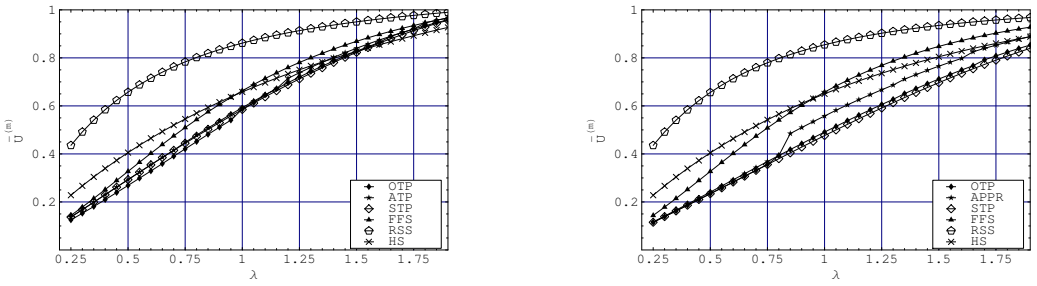
5. Численные эксперименты

В данном разделе вычисляются характеристики системы $M/M/3$ для оптимальной пороговой политики управления (*Optimal Threshold Policy* — OTP) с уровнями (q_2^*, q_3^*). Для проведения сравнительного анализа исследуются также другие политики управления и однородная система.

Приближенная пороговая политика (*Approximated Threshold Policy* — ATP) задаётся пороговыми уровнями (20). Пороговая политика расписания (*Scheduling Threshold Policy* — STP) задаётся пороговыми уровнями (19) и представляет собой неплохую аппроксимацию в случае $\lambda < \mu_K$. Политика включения самого быстрого свободного прибора (*Fastest Free Server* — FFS) предписывает использование самого быстрого свободного прибора всякий раз при поступлении новой или повторной заявки. Очевидно, что эта политика является частным случаем пороговой политики, если $(q_2, q_3) = (1,1)$. Политика случайного выбора прибора (*Random Server Selection* — RSS) выбирает свободные приборы равновероятно.

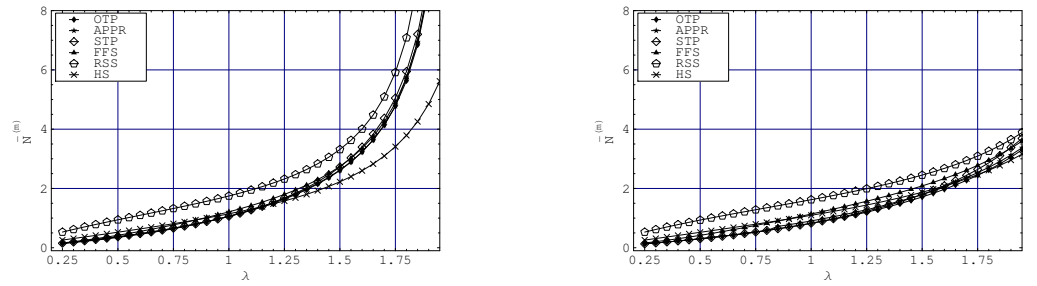
Неуправляемая система с однородными приборами (*Homogeneous Servers* — HS) имеет общую интенсивность обслуживания $\mu = \sum_{k=1}^{K} \mu_k$. С помощью программного пакета *Mathematica* созданы процедуры для вычисления вектора стационарных вероятностей и средних характеристик производительности.

На рис. 2–5 показаны соответственно графики функций загрузки системы, среднего числа заявок в системе, среднего времени ожидания на орбите и среднего времени пребывания заявок в системе для фиксированных интенсивностей обслуживания $\mu_1 = 2,2,\ \mu_2 = 0,5,\ \mu_3 = 0,2$, интенсивностей повторных заявок $\gamma = 2,5$ (см. рис., обозначенные «а») и $\gamma = 18,5$ (см. рис., обозначенные «б»), для различных значений параметра λ, где $\lambda \in [0,05; 1,90]$ с шагом $0,1$.



(а) (б)

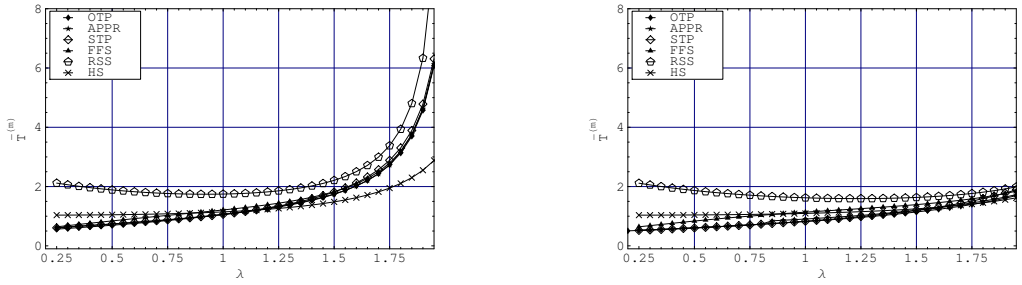
Рис. 2. **Загрузка системы для (а)** $\gamma = 2,5$ **(б)** $\gamma = 18,5$



(а) (б)

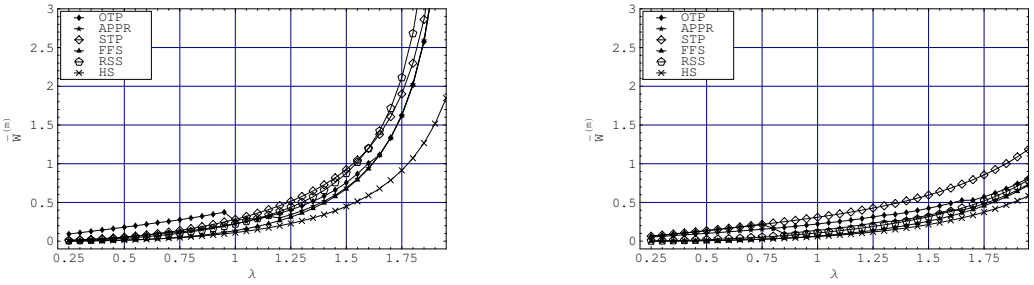
Рис. 3. **Среднее число заявок в системе для (а)** $\gamma = 2,5$ **(б)** $\gamma = 18,5$

Из представленных графиков видно, что максимальная загрузка системы приходится на систему с RSS политикой управления, в то время как при пороговых политиках загрузка принимает наименьшие значения. Загрузка для систем FFS и HS принимает промежуточные значения. Среднее число заявок и среднее время пребывания являются наиболее интересными характеристиками. Как и ожидалось, для управляемых систем наибольшее значение эти характеристики принимают для системы RSS, а наименьшее — для OTP. На графиках видно, что приближенная политика ATP довольно хорошо аппроксимирует оптимальную политику. При малых значениях параметра λ политика STP также практически совпадает с OTP. При больших значениях λ различие между политиками значительно сокращается. Так как при увеличении загрузки пороговые уровни OTP убывают, то политика FSS может рассматриваться в данном случае как



(а) (б)

Рис. 4. **Среднее время пребывания в системе для (а)** $\gamma = 2,5$ **(б)** $\gamma = 18,5$



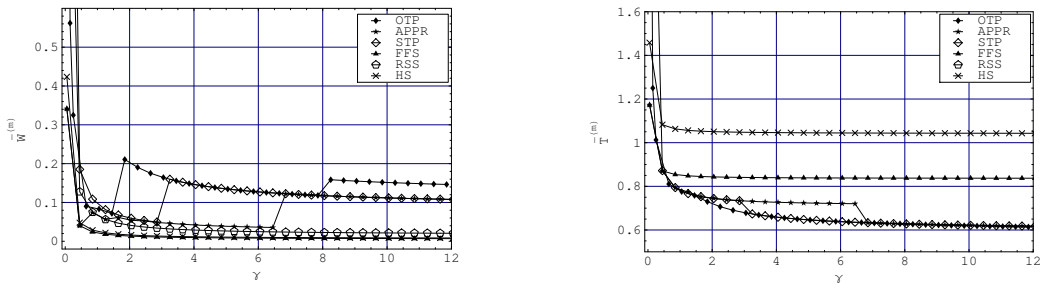
(а) (б)

Рис. 5. **Среднее время ожидания на орбите для (а)** $\gamma = 2,5$ **(б)** $\gamma = 18,5$

квази-оптимальная. Заметим также, что неуправляемая система HS имеет преимущество перед оптимальной политикой для больших значений интенсивности поступления λ и малых значений интенсивности повторов θ.

Графики для среднего времени ожидания интересны тем, что здесь имеет место обратная картина по сравнению с предыдущими графиками для среднего числа заявок. Максимальное значение данная характеристика принимает для пороговых политик управления, а также для системы RSS при больших значениях λ. Это не противоречит оптимальности пороговой политики, так как она по определению минимизирует среднее число заявок или среднее время пребывания заявок в системе. Для пороговых политик OTP и ATP функция среднего времени ожидания имеет скачкообразную структуру. Скачки означают, что пороговые уровни уменьшаются с увеличением интенсивности поступления, что, в свою очередь, приводит к значительному уменьшению времени ожидания. Наименьшее значение среди управляемых систем среднее время ожидания имеет система FFS. Неуправляемая система HS также характеризуется малым значением времени ожидания.

На рис. 6 показаны средние значения времени ожидания на орбите и среднего времени пребывания заявок в системе при фиксированных параметрах $\mu_1 = 2,2$, $\mu_2 = 0,5$, $\mu_3 = 0,2$, $\lambda = 0,5$ и различных значений параметра γ, где $\gamma \in [0,05; 12,0]$ с шагом 0.1. В данном случае графики функций для пороговых дисциплин также имеют скачкообразную структуру и сходятся к средним значениям для стандартной системы без повторов с увеличением значений параметра γ. Из рис. 6 видно, что преимущество оптимальной пороговой политики возрастает с увеличением γ.



(а) (б)

Рис. 6. **(а) Среднее время ожидания и (б) среднее время пребывания заявки**

6. Заключение

Исследована модель системы обслуживания с постоянной интенсивностью повторных заявок, неоднородными приборами и пороговой политикой управления. В результате проведённого вероятностного анализа решена проблема нахождения средних характеристик системы в стационарном режиме и оптимальных пороговых уровней. При увеличении числа приборов эффективность приведённых алгоритмов падает. В этом случае имеет смысл объединять приборы в группы по величине их интенсивностей. Таким образом получается небольшое число неоднородных групп однородных приборов. Подобные системы будут в дальнейшем объектом исследований автора.

Литература

1. *Artalejo J. R., Gomez-Corral A.* Retrial Queueing Systems. — Springer, 2008.
2. *Falin G. I., Templelton J. G. C.* Retrial Queues. — Chapman and Hall, 1997.
3. *Choi B. D., Rhee K. H., Park K. K.* The M/G/1 Retrial Queue with Retrial Rate Control Policy // Probability in the Engineering and Informational Sciences. — 1993. — Vol. 7. — Pp. 29–46.
4. *Choi B. D., Shin Y. W., Ahn W. C.* Retrial Queues with Collision Arising from Unslotted CSMA/CD Protocol // Queueing Systems. — 1992. — Vol. 11. — Pp. 335–356.
5. *Artalejo J. R.* Stationary Analysis of the Characteristics of the M/M/2 Queue with Constant Repeated Attempts // Opsearch. — 1996. — Vol. 33. — Pp. 83–95.
6. *Artalejo J. R., Gomez-Corral A., Neuts M. F.* Analysis of Multiserver Queues with Constant Retrial Rate // European Journal on Operations Research. — 2001. — Vol. 135. — Pp. 569–581.
7. *Pourbabai B.* Markovian Queueing Systems with Retrials and Heterogeneous Servers // Computers and Mathematics with Applications. — 1987. — Vol. 13. — Pp. 917–923.
8. *Sztrik J., Roszik J.* Performance Analysis of Finite-Source Retrial Queueing Systems with Nonreliable Heterogeneous Servers // Journal of Mathematical Sciences. — 2007. — Vol. 146. — Pp. 6033–6038.
9. *Рыков В. В.* Об условиях монотонности оптимальных политик управления системами массового обслуживания // Автоматика и телемеханика. — 1999. — Т. 9. — С. 92–106.
10. *Рыков В. В., Ефросинин Д. В.* Численное исследование оптимального управления системой с неоднородными приборами // Автоматика и телемеханика. — 2003. — Т. 2. — С. 389–407.

11. *Efrosinin D., Breuer B.* Threshold Policies for Controlled Retrial Queues with Heterogeneous Servers // Annals of Operation Research. — 2006. — Vol. 141. — Pp. 139–162.
12. *Ефросинин Д. В., Рыков В. В.* К анализу характеристик производительности СМО с неоднородными приборами // Автоматика и телемеханика. — 2008. — Т. 1. — С. 64–82.
13. *Neuts M. F.* Matrix-Geometric Solutions in Stochastic Models. — The John Hopkins University Press, 1981.
14. *Efrosinin D.* Controlled Queueing Systems with Heterogeneous Servers. Dynamic Optimization and Monotonicity Properties. — VDM Verlag, 2008.

UDC 519.21

Stationary Characteristics of the Multichannel Queue with FCFS Orbit And Threshold Control

D. V. Efrosinin

Department of Probability Theory and Mathematical Statistics
Peoples' Friendship University of Russia
6, Miklukho–Maklaya str., Moscow, 117198, Russia

In the paper we deal with a Markovian queueing system with heterogeneous servers and constant retrial rate. The system operates under a threshold policy. The system is described by quasi-birth-and-death process with infinitesimal matrix depending on the threshold levels. Using a matrix-geometric approach we perform a stationary analysis of the system, derive expressions for the mean performance measures and formulas for optimal threshold levels.

Key words and phrases: retrial queue, controllable queueing system, stationary regime, ergodicity condition, threshold control policy, waiting time distribution.