Труды Карельского научного центра РАН
№ 5. 2012. С. 68–77

УДК 519.872.1

СИСТЕМА С ПОВТОРНЫМИ ВЫЗОВАМИ: ОЦЕНИВАНИЕ ВЕРОЯТНОСТИ БЛОКИРОВКИ ВЫЗОВА НА КОНЕЧНОМ ИНТЕРВАЛЕ

Е. В. Морозов, Р. С. Некрасова

Институт прикладных математических исследований Карельского научного центра РАН

В статье рассматривается регенеративное оценивание вероятности занятости односерверной системы с повторными вызовами в переходном режиме. Первичная заявка, заставшая сервер занятым, поступает на орбиту бесконечной емкости, образуя очередь для повторной попытки попасть на сервер. Обсуждаются условия стационарности такой системы. Оценивание применяется как в области стационарности, так и в области нестационарности. В последнем случае рассматриваемая оценка сходится к вероятности занятости сервера в некоторой вспомогательной системе с потерями. Основное внимание в статье уделено оцениванию вероятности блокировки заявки на конечном интервале, т. е. в переходном режиме системы. Проведено сравнение эффективности стандартной выборочной оценки и альтернативной оценки. Приведены результаты численного моделирования и сравнения эффективности этих оценок.

Ключевые слова: система с повторными вызовами, условие стационарности, вероятность занятости, оценивание на конечном интервале.

E. V. Morozov, R. S. Nekrasova. A RETRIAL QUEUEING SYSTEM: ESTIMATION OF THE BLOCKING PROBABILITY IN A FINITE TIME INTERVAL

The paper deals with regeneration estimation of busy probability in a single server retrial queueing system in transient regime. Primary customer who finds the server busy joins the infinite capacity orbit to wait as in a queue for retrial. Stability conditions of such a system are discussed. Estimation is applied both to the stability and the instability domains of this system. The focus in the paper is on estimation of the blocking probability over a finite interval, that is in transient regime. Standard sample mean estimate and the alternative estimate based on the knowledge of the stationary blocking probability were compared.

Key words: Retrial queuing system, stability condition, busy probability, estimation on finite interval.

Введение

Рассмотрим односерверную систему с повторными вызовами Σ без буфера с пуассоновским входным потоком с интенсивностью λ и (произвольным) временем обслуживания S со средним $\mathsf{E}S := 1/\mu$. Заявки, поступающие в систему, когда сервер занят, уходят на орбиту бесконечного объема (блокируются),

образуя "очередь" в порядке поступления на орбиту. Первая заявка в этой очереди делает попытку попасть на сервер через экспоненциально распределенное время с интенсивностью μ_0, и в случае неудачи (мгновенно) возвращается на орбиту. Заметим, что первая заявка может, в частности, снова возвращаться в начало очереди для следующей попытки. (Это не меняет распределения числа заявок на орбите и в системе в целом.) Тогда без ограничения общности можно считать, что неудачных попыток не существует и (первая) заявка на орбите делает попытку лишь при освобождении сервера. Однако ниже мы рассматриваем также и поток неудачных попыток. Рассматриваемая система с постоянной скоростью ухода заявок с орбиты радикально отличается от классических систем с повторными вызовами, где интенсивность орбитальных заявок растет с ростом их числа, поскольку заявки делают повторные попытки независимо. По этой причине и условия стационарности рассматриваемой и классической систем принципиально различаются. В частности, условие стационарности классической системы с повторными вызовами совпадает с условием стационарности стандартной системы с неограниченным буфером и (индивидуальная) интенсивность ухода заявки с орбиты не играет в нем никакой роли. Напротив, в системе с постоянной скоростью ухода с орбиты эта интенсивность существенным образом влияет на область стационарности.

Таким образом, в рассматриваемой системе сама орбита может рассматриваться как односерверная система обслуживания вида $G/M/1$, где в обозначении G входного потока на орбиту отражена его сложная структура, в частности, что он не является процессом восстановления. Этот поток является суперпозицией потока первичных блокируемых заявок и потока неудачных попыток орбитальных заявок попасть на сервер. Заметим, что источником нестационарности рассматриваемой системы может быть только неограниченно растущая орбита.

Отметим, что анализ таких нестандартных систем с повторными вызовами мотивирован поведением некоторых сетевых протоколов (подробнее см., например, [7–9]).

В ряде предшествующих работ найдены условия стационарности некоторых из описанных систем с повторными вызовами. Впервые такая модель вида $M/G/1/0$ (без буфера) исследована в работе [11], где также получено необходимое и достаточное условие стационарности. Позднее в работе [3] была исследована двухсерверная система вида $M/M/2/0$, а также получено необходимое и достаточное условие ее стационарности (условие эргодичности марковского процесса, описывающего ее динамику). В работе [13] рассмотрен стационарный режим системы с повторными вызовами вида $M/M/m/n$ (с буфером размера n) и приведен критерий стационарности для системы $M/M/1/1$. Также условия эргодичности многосерверной системы вида $M/M/m/0$ были получены в [4]. Наконец, в недавней работе [7] найдено достаточное условие стационарности немарковской m-серверной системы с повторными вызовами вида $GI/G/m/n$, у которой лишь интервалы между повторными попытками остаются показательными (с интенсивностью μ_0).

Данная статья организована следующим образом. Вначале приведены условия стационарности некоторых из описанных систем с повторными вызовами, а также явное выражение для стационарной вероятности блокировки P_b в системе вида $M/M/1/0$. (По свойству PASTA эта вероятность совпадает также с вероятностью занятости системы.) Затем описана регенеративная структура, включая квази-регенерацию, используемую для оценки вероятности блокировки в нестационарном режиме. Далее рассмотрена стандартная оценка вероятности блокировки, а также альтернативная оценка этой вероятности на конечном интервале. В заключение приведены результаты численного моделирования и сравнение эффективности рассмотренных оценок для систем вида $M/G/1/0$. Отметим, что результаты моделирования показали преимущество альтернативной оценки.

Условия стационарности системы с повторными вызовами

Рассмотрим простейшую систему с повторными вызовами вида $M/M/1/0$ без буфера. Состояние системы в момент t описывается двухмерным марковским процессом $Y(t) = (N(t), \nu(t))$, где $N(t)$ – число заявок на орбите, а $\nu(t) \in \{0, 1\}$ – число заявок в системе. Легко увидеть, что марковский процесс $\{Y(t)\}_{t\geqslant 0}$ – неприводимый со множеством состояний $\mathbf{S} = \mathbb{Z}_+ \times \{0, 1\}$. Обозначим его предельное распределение

$$\mathsf{P}_{ij} = \lim_{t\to\infty} \mathsf{P}\{N(t) = i; \nu(t) = j\}, \ \ (i, j) \in \mathbf{S},$$

когда оно существует. Заметим, что поток заявок, поступающих на сервер, складывается из двух (вообще говоря, зависимых) потоков:

потока первичных заявок (λ-потока) и потока вторичных заявок с интенсивностью $\tilde{\mu}_0$. При этом интенсивность $\tilde{\mu}_0$ в момент t определяется на событии $\{N(t) = i\}$ как $\tilde{\mu}_0 = \mu_0(1 - \delta_{i0})$, где δ_{i0} – символ Кронекера. (Очевидно, $\tilde{\mu}_0 \leqslant \mu_0$.) В работе [4] показано, что критерий стационарности имеет вид:

$$\rho_1 := \rho\frac{\lambda + \mu_0}{\mu_0} < 1, \tag{1}$$

где $\rho := \lambda/\mu$, а стационарные вероятности состояний равны

$$\mathsf{P}_{i0} = \frac{\lambda}{\lambda + \tilde{\mu}_0}(1 - \rho_1)\rho_1^i, \ \ i \geqslant 0, \tag{2}$$

$$\mathsf{P}_{i1} = \rho(1 - \rho_1)\rho_1^i, \ \ i \geqslant 0. \tag{3}$$

Очевидно, стационарная вероятность занятости сервера ввиду (3) равна

$$\mathsf{P}_b = \sum_{i=0}^{\infty} \mathsf{P}_{i1} = \rho, \tag{4}$$

и, таким образом, критерий стационарности (1) принимает вид:

$$(\lambda + \mu_0)\mathsf{P}_b < \mu_0. \tag{5}$$

В работе [7] получено следующее достаточное условие стационарности (верное для широкого класса систем с повторными вызовами):

$$(\lambda + \mu_0)\mathsf{P}_{loss} < \mu_0, \tag{6}$$

где P_{loss} – стационарная вероятность потери заявки в мажорирующей односерверной системе с потерями с пуассоновским входным потоком с параметром $\lambda + \mu_0$ и тем же распределением времени обслуживания, что и в исходной системе. При этом в системе $M/G/1/0$ вероятность P_{loss} определяется по формуле Эрланга

$$\mathsf{P}_{loss} = \frac{\lambda + \mu_0}{\lambda + \mu_0 + \mu}. \tag{7}$$

Легко проверить, что условия (5) и (6) эквивалентны. Заметим, что условие (1), которое можно записать в виде

$$\rho < \frac{\mu_0}{\lambda + \mu_0} < 1, \tag{8}$$

сводится к классическому условию $\rho < 1$ при $\mu_0 \to \infty$. Этот результат объясняется тем, что с ростом интенсивности μ_0 заявки все меньше времени проводят на орбите и рассматриваемая система сближается с классической системой с буфером неограниченного объема.

Регенеративная структура системы с повторными вызовами

Одно из основных преимуществ регенеративного метода состоит в том, что он применим к широкому классу немарковских процессов. Это, в частности, позволяет уменьшить размерность основного процесса, описывающего динамику системы. Например, состояние введенной выше системы с повторными вызовами (и даже более общей системы вида $GI/G/m/n$) в момент t может быть также описано с помощью скалярного (непрерывного справа) немарковского процесса $\{X(t) := N(t) + \nu(t), t \geqslant 0\}$, где $\nu(t) \in \{0, n + m\}$. Пусть λ-заявки поступают в моменты $\{t_n\}$, и пусть $X(t_n^-) := X_n$. Очевидно, система (процесс $X := \{X(t)\}$) регенерирует каждый раз, когда λ-заявка поступает в пустую систему (пустой сервер и пустая орбита). Положим $T_0 := 0$ и определим моменты регенерации процесса X (и других процессов в системе) следующим образом:

$$T_{k+1} = \inf_i(t_i > T_k : X_i = 0), \ k \geqslant 0.$$

Пусть $T =_{st} T_k - T_{k-1}$, $k \geqslant 1$ – типичный период регенерации ($=_{st}$ означает стохастическое равенство). Обозначим $\beta_k = A(T_k^-)$ и заметим, что моменты $\{\beta_k\}$ удовлетворяют рекурсии

$$\beta_{k+1} = \inf_i\{i > \beta_k : X_i = 0\} \ (\beta_0 = 0) \tag{9}$$

и являются моментами регенерации (в дискретном времени), причем $T_k = t_{\beta_k}$. Ниже мы будем рассматривать более специальную систему с повторными вызовами вида $M/G/1/0$.

Стационарный режим. Предположим, что выполнено условие стационарности системы. Иными словами, процесс X – положительно-возвратный, т. е. $\mathsf{E}T < \infty$. Пусть $A(t)$ и $H(t)$ – число первичных λ-заявок, поступивших в систему и блокированных за время $[0, t]$, соответственно, а A и H – число приходов и блокировок, соответственно, на (типичном) цикле регенерации, так что $A =_{st} \beta_k - \beta_{k-1}$, $k \geqslant 1$. Пусть I_k есть индикатор блокировки k-й первичной заявки, тогда, в частности, $H(t) = \sum_{k=1}^{A(t)} I_k$. Процесс $\{I_k\}_{k\geqslant 0}$ регенерирует в моменты $\{\beta_k\}_{k\geqslant 0}$, и является положительно-возвратным, т. е. $\mathsf{E}A < \infty$. Ввиду свойства (пуассоновского) входного потока существует слабый предел $I_k \Rightarrow I$, при $k \to \infty$, и более того $\mathsf{P}(I_k = 1) = \mathsf{E}I_k \to \mathsf{E}I$, где, ввиду свойства PASTA, стационарная вероятность блокировки заявки $\mathsf{E}I$ совпадает с вероятностью занятости сервера. Суммируя ска-

занное и используя теорию регенерации, получаем с вероятностью 1 (с в. 1)

$$\lim_{t\to\infty}\frac{H(t)}{A(t)}=\frac{\mathsf{E}H}{\mathsf{E}A}=\mathsf{P}_b=\rho. \qquad (10)$$

Заметим, что равенство $\mathsf{P}_b=\rho$ (см. (4)) верно не только для системы вида $M/M/1/0$, но и для более общей системы с повторными вызовами вида $M/G/1/0$. Доказательство этого результата дано ниже, см. (22). (Это равенство также имеет место в классической системе $M/G/1$ без орбиты и с *неограниченным буфером* [2].)

Нестационарный режим. В этом случае число заявок на орбите неограниченно растет и применение классической регенерации становится невозможным. В таком случае можно использовать *квази-регенерации*, когда поступающая в систему заявка (первичная или вторичная) встречает пустой сервер, а орбита может быть не пустой. Определение квази-регенераций см. в [1, 6, 7]. Поток повторных заявок сближается с пуассоновским потоком с интенсивностью μ_0, и со временем исходная система начинает вести себя как система Эрланга с потерями с пуассоновским потоком с параметром $\lambda+\mu_0$ и стационарной вероятностью потери P_{loss}, удовлетворяющей (7). Другими словами, квази-регенерации становятся классическими регенерациями, но для процесса очереди, рассматриваемого изолированно. Заметим, что по свойству PASTA вероятность P_{loss} совпадает со стационарной занятостью сервера (в системе с потерями).

Оценки среднего на конечном интервале

Выборочная оценка, как правило, хорошо аппроксимирует оцениваемый параметр лишь при большом числе наблюдений (большом t). В то же время процесс моделирования часто весьма затратен и поэтому ограничивается некоторым фиксированным интервалом $[0,\,t]$.

Рассмотрим общий случай оценивания предельного среднего по времени от (измеримой) функции $f[X(t)]$ регенерирующего процесса $\{X(t)\}$ на основе оценки

$$r(t)=\frac{1}{t}\int_0^t f[X(u)]du. \qquad (11)$$

(В дискретном времени интеграл заменяется на сумму.) Предположим, что с в. 1 существует предел $r(t)\to r$, при $t\to\infty$ и требуется построить оценку нестационарного среднего $\mathsf{E}[r(t)]$ при моделировании на конечном интервале $[0,t]$. Заметим, что такое оценивание допредельной характеристики $\mathsf{E}[r(t)]$ в *переходном режиме* можно формально проводить и в условиях нестационарности. Предположим также, что $\mathsf{E}[r(t)]\to r$, для чего достаточно, например, равномерной интегрируемости семейства $\{r(t),\,t\geqslant 0\}$. В частности, последнее имеет место при оценивании вероятности занятости сервера P_b с использованием оценки

$$r(t)=\frac{1}{t}\int_0^t I(\nu(u)=0)du,$$

где, напомним, $\nu(t)$ есть суммарное число заявок на сервере и в буфере (если он существует) в момент t, а $r(t)\leqslant 1$.

Из сказанного выше следует, что важно иметь оценки среднего $\mathsf{E}r(t)$, эффективные при больших t.

Стандартная оценка для $\mathsf{E}[r(t)]$ является выборочным средним из m независимых траекторий $r_j(t)$ процесса на интервале $[0,t]$, именно

$$\hat r_m(t):=\frac{1}{m}\sum_{j=1}^m r_j(t). \qquad (12)$$

По усиленному закону больших чисел (УЗБЧ), при каждом t с в. 1

$$\hat r_m(t)\to\mathsf{E}r(t),\quad m\to\infty.$$

Регенеративная структура процесса $\{f[X(t)]\}$ позволяет получить следующий асимптотический результат для дисперсии оценки $\hat r_m(t)$ при $t\to\infty$ и фиксированном m [12].

Теорема 1. *Для любого фиксированного m,*

$$t\mathsf{D}[\hat r_m(t)]\to\frac{C_1}{m},\ \ t\to\infty, \qquad (13)$$

где C_1 – явно заданная константа.

Таким образом, дисперсия стандартной оценки среднего $\mathsf{E}r(t)$ убывает со скоростью $1/t$ с ростом t.

В работе [12] также предложена альтернативная оценка $\mathsf{E}r(t)$, построенная по остаточной длине цикла регенерации в момент t в предположении, что *предельное значение r известно*. Ситуация, когда r известно, а вычисление среднего на конечном интервале $\mathsf{E}r(t)$ вызывает большие сложности, встречается при моделировании довольно часто. Типичными примерами являются формулы Литтла, Поллачека-Хинчина, стационарное среднее незавершенное время восстановления и т. д. Как показывает дальнейший анализ, альтернативная оценка вероятности блокировки имеет существенно меньшую дисперсию чем стандартное выборочное среднее. Сохраним обозначение $\{T_n\}_{n\geqslant 0}$ моментов регенерации рассматриваемого процесса

$\{X(t)\}_{t\geqslant 0}$ и определим *незавершенное время регенерации* в момент t как

$$\mathcal{T}(t) := T_{N(t)+1} - t, \tag{14}$$

где $N(t)$ – число регенераций в интервале $(0, t]$, а также *"перескок" процесса накопления* на текущем в момент t цикле регенерации как:

$$\mathcal{I}(t) := \int_t^{T_{N(t)+1}} f[X(u)]du. \tag{15}$$

(Типичный процесс накопления имеет вид интеграла от регенерирующего процесса [14].)

В [12] предложена следующая взвешенная оценка (статистика) величины r

$$R(t) := r + r\frac{\mathcal{T}(t)}{t} - \frac{\mathcal{I}(t)}{t}, \tag{16}$$

которая имеет прозрачную физическую интерпретацию. Тогда выборочная оценка величины $\mathsf{E}r(t)$ *по остаточной длине цикла* принимает вид:

$$\hat{R}_m(t) := \frac{1}{m}\sum_j^m R_j(t), \tag{17}$$

где $R_j(t)$ – j-я (независимая) реализация статистики $R(t)$. Как показано в [12], оценка $\hat{R}_m(t)$ при всех t, m является несмещенной, т. е. $\mathsf{E}\hat{R}_m(t) = \mathsf{E}r(t)$. При исследовании эффективности, необходимо учитывать, что оценка по остаточной длине цикла требует дополнительных вычислений после момента t до окончания текущего цикла. Однако величины $\mathcal{T}(t)$ и $\mathcal{I}(t)$ (при определенных моментных требованиях, см. Теорема 2 ниже) асимптотически ведут себя как $o(t)$ при больших t и мало влияют на объем вычислений. Из теории процессов накопления следует, что если $\mathsf{E}T < \infty$, то $\mathsf{E}\mathcal{T}(t) = o(t)$, а если $\mathsf{E}|Y| < \infty$, где $Y = \int_0^T f[X(u)]du$, то $\mathcal{I}(t) = o(t)$ с в. 1 [14]. Имеет место следующий результат [12].

Теорема 2. *Если* $\mathsf{E}[T^3] < \infty$ *и* $E[TY^2] < \infty$, *то для любого фиксированного* m

$$t^2\mathsf{D}[\hat{R}_m(t)] \to \frac{C_2}{m}, \;\; t \to \infty, \tag{18}$$

где C_2 *– явно заданная константа.*

Таким образом, дисперсия оценки $\hat{R}_m(t)$ сходится к нулю в t раз быстрее, чем дисперсия стандартной оценки $\hat{r}_m(t)$. Следовательно, оценка $\hat{R}_m(t)$ может существенно повысить эффективность регенеративного моделирования. (Дополнительные аргументы в пользу этой оценки можно найти в [12].) Ниже этот результат установлен при оценивании вероятности блокировки в рассмотренной выше системе с повторными вызовами в интервале $[0, t]$ при фиксированном t. Для оценивания в *стационарном режиме* рассмотрим соответствующий аналог величины $r(t)$ выше:

$$\frac{H(t)}{A(t)} = \frac{\sum_{i=1}^{k(t)} I_i}{k(t)} := \theta[k(t)], \tag{19}$$

где I_i есть индикатор блокировки i-й заявки, а $k(t) = \max\limits_i\{i : t_i \leqslant t\}$ есть число первичных заявок, поступивших в интервале $[0, t]$. Несмещенной и сильно состоятельной оценкой $\mathsf{E}\theta[k(t)]$ является выборочное среднее

$$\hat{\theta}_m(t) := \frac{1}{m}\sum_{j=1}^m \theta_j[k_j(t)], \tag{20}$$

где $\theta_j[k_j(t)]$ есть j-я (независимая) реализация $\theta[k(t)]$ в интервале $[0, t]$ (которая включает случайное число наблюдений $k_j(t)$.) Отметим, что число регенераций в интервале $[0, t]$, содержащем $k(t)$ первичных заявок, равно $N(t) := \max\limits_i\{i : \beta_i \leqslant k(t)\}$. Введем следующие аналоги величин $\mathcal{T}(t)$, $\mathcal{I}(t)$, соответственно, для оценки вероятности блокировки

$$\mathcal{T}_{k(t)} := \beta_{N(t)+1} - k(t), \;\; \mathcal{I}_{k(t)} := \sum_{i=k(t)+1}^{\beta_{N(t)+1}} I_i. \tag{21}$$

Напомним, что $\mathsf{P}_b = \rho$ есть стационарная вероятность занятости системы с повторными вызовами вида $M/M/1/0$, см. (4). Используя иной подход, покажем, что и в системе с повторными вызовами вида $M/G/1/0$ с произвольным временем обслуживания S стационарная вероятность занятости

$$\mathsf{P}_b = \lambda\mathsf{E}S := \rho. \tag{22}$$

Именно, пусть $V(t)$ – поступившая, а $D(t)$ – обслуженная нагрузка в системе в интервале $[0, t]$, и пусть $W(t)$ – незавершенная в момент t работа в системе (в сервере и на орбите). Имеет место следующее уравнение баланса

$$V(t) = W(t) + D(t), \;\; t \geqslant 0. \tag{23}$$

Если система стационарна, то с в. 1 $W(t) = o(t)$, $t \to \infty$ [14]. Поскольку

$$\frac{V(t)}{t} \to \rho, \;\; \frac{D(t)}{t} \to \mathsf{P}_b$$

с в. 1, то из (23) следует соотношение (22). (Более детальный анализ см. в [2].) С учетом этих замечаний рассмотрим статистику

$$\Theta[k(t)] = \mathsf{P}_b + \frac{\mathsf{P}_b\mathcal{T}_{k(t)}}{k(t)} - \frac{\mathcal{I}_{k(t)}}{k(t)}, \tag{24}$$

которая приводит к такой (альтернативной) оценке величины $\mathsf{E}\theta[k(t)]$:

$$\hat{\Theta}_m(t) := \frac{1}{m}\sum_{j=1}^{m}\Theta_j[k_j(t)], \qquad (25)$$

где $\Theta_j[k_j(t)]$ есть j-я (независимая) реализация $\Theta[k(t)]$. (Отметим, что несмещенность оценки $\hat{\Theta}_m(t)$ не удается доказать по аналогии с $\hat{R}_m(t)$ (см. [12]), поскольку $k(t)$ является случайной величиной.)

Подчеркнем, что оценки $\hat{\Theta}_m(t)$ и $\hat{\theta}_m(t)$ строятся методом регенеративного моделирования, в котором удобнее использовать фиксированное число приходов k, чем фиксированный интервал $[0,\, t]$. Этот подход и реализован при численном моделировании в данной работе. Именно, (в очевидных обозначениях) каждая из m траекторий процессов $\{\theta(k)\}$ и $\{\Theta(k)\}$ строится для фиксированного числа наблюдений k. Такая аппроксимация величины $\mathsf{E}r(t)$ вполне приемлема, если значения k и t достаточно велики. Имеет место следующий результат, доказываемый аналогично равенству $\mathsf{E}\hat{R}_m(t) = \mathsf{E}r(t)$ в [12]. Обозначим $\theta(k) = \sum_{i=1}^{k} I_i/k$, см. (19).

Теорема 3. *Оценки вероятности блокировки*

$$\hat{\theta}_{m,k} := \frac{1}{m}\sum_{j}^{m}\theta_j(k) \;\; u \;\; \hat{\Theta}_{m,k} := \frac{1}{m}\sum_{j}^{m}\Theta_j(k),$$

построенные по k наблюдениям (в каждой из m независимых траекторий $\{\theta_j(k)\}$ и $\{\Theta_j(k)\}$) являются несмещенными,

$$\mathsf{E}[\hat{\theta}_{m,k}] = \mathsf{E}[\hat{\Theta}_{m,k}] = \mathsf{E}[\theta(k)],$$

при всех k и m.

Доказательство. Из определения $\hat{\theta}_{m,k}$ очевидным образом следует

$$\mathsf{E}[\hat{\theta}_{m,k}] \;\; = \;\; \frac{1}{m}\sum_{j=1}^{m}\mathsf{E}[\theta_j(k)] = \mathsf{E}[\theta(k)]. \quad (26)$$

Рассмотрим величину

$$\begin{aligned} k\theta(k) \;\; &= \;\; \sum_{i=1}^{k} I_i = \sum_{i=1}^{\beta_{n(k)+1}} I_i - \sum_{i=k+1}^{\beta_{n(k)+1}} I_i \\ &= \;\; \sum_{i=1}^{n(k)+1} H_i - \mathcal{I}_k = \sum_{i=1}^{n(k)+1} H_i \qquad (27) \\ &- \;\; \mathsf{P}_b \sum_{i=1}^{n(k)+1} A_i + \mathsf{P}_b\beta_{n(k)+1} - \mathcal{I}_k, \end{aligned}$$

где $n(k)$ – число регенераций при k наблюдениях, H_i – число блокировок заявок на i-м цикле регенерации, а $A_i = \beta_i - \beta_{i-1}$. Напомним, что $\mathsf{P}_b = \mathsf{E}H/\mathsf{E}A$, см. (10). Тогда, по тождеству Вальда:

$$\mathsf{E}\Big[\sum_{i=1}^{n(k)+1} H_i - \mathsf{P}_b \sum_{i=1}^{n(k)+1} A_i\Big] = 0.$$

Таким образом, взяв математическое ожидание в (27), получаем

$$\begin{aligned} k\mathsf{E}[\theta(k)] \;\; &= \;\; \mathsf{E}[\mathsf{P}_b(k+\mathcal{T}_k) - \mathcal{I}_k] \\ &= \;\; k\mathsf{E}[\mathsf{P}_b + \mathsf{P}_b\frac{\mathcal{T}_k}{k} - \frac{\mathcal{I}_k}{k}]. \qquad (28) \end{aligned}$$

Поэтому $\mathsf{E}[\theta(k)] = \mathsf{E}[\Theta(k)]$, и кроме того, очевидно, $\mathsf{E}[\hat{\Theta}_{m,k}] = \mathsf{E}[\Theta(k)]$. □

Рассмотрим *нестационарный режим* той же системы с повторными вызовами. Пусть $\{z_i,\, i \geqslant 0\}$ есть моменты обращения заявок (*как первичных, так и вторичных*) к серверу. Обозначим $\nu_i = \nu(z_i^-)$ и определим рекурсивно моменты *квази-регенерации* таким образом

$$\alpha_{k+1} = \inf_i(i > \alpha_k : \nu_i = 0),\; k \geqslant 0, \;\; (\alpha_0 = 0).$$

Теперь оценка для вероятности блокировки в переходном режиме имеет тот же вид, что и (19)

$$\theta[k(t)] := \frac{\sum_{i=1}^{k(t)} I_i}{k(t)},$$

где, однако, I_i есть индикатор *неудачной попытки* обращения к серверу, а $k(t)$ – общее число попыток в $[0,\, t]$.

Как показано в [1], P_b в данном случае совпадает со стационарной вероятностью потери P_{loss} в системе с потерями $M/G/1/0$ с интенсивностью входного потока $\lambda+\mu_0$, см. (7). Ввиду свойства PASTA P_{loss} также равна предельной вероятности блокировки заявки. Интересно отметить, что это происходит в условиях *неограниченного роста орбиты*. Такую систему можно назвать *частично устойчивой*.

Теорема 4. *В системе с повторными вызовами вида $M/M/1/0$ в стационарном режиме интенсивность потока попыток с орбиты на сервер равна*

$$\tilde{\mu}_0 = \rho^2(\mu + \lambda + \mu_0). \qquad (29)$$

Доказательство. Определим (потенциальное) общее число $\Lambda_0(t)$ попыток попасть на сервер, порожденных *блокированными заявками*, поступившими в интервале $[0,\, t]$, и пусть

γ_n есть число неудачных попыток n-й *повторной* заявки. Напомним, что общее число таких заявок в интервале $[0, t]$ есть $H(t)$. Для определенности будем считать, что блокируемая первичная заявка присоединяется к концу очереди заявок, находящихся на орбите. Заметим, что величины $\{\gamma_n\}$ – н.о.р. и что

$$p := \frac{\lambda}{\lambda + \mu_0}$$

есть вероятность того, что повторная заявка не попадет на освободившийся после обслуживания сервер, который в этом случае будет занят первичной заявкой. Поскольку вероятность не менее i таких неудачных попыток равна p^i, а среднее число неудачных попыток в течение (экспоненциального) времени обслуживания равно μ_0/μ, то нетрудно видеть, что

$$\mathsf{E}\gamma = \frac{\mu_0}{\mu}\sum_{i=0}^{\infty} p^i = \frac{\lambda + \mu_0}{\mu}. \qquad (30)$$

В области стационарности $H(t)/t \to \lambda \mathsf{P}_b = \lambda\rho$. Кроме того, все повторные заявки в конце концов попадают на сервер, и поэтому общее число попыток у n-й повторной заявки равно $\gamma_n + 1$. Таким образом, стационарная интенсивность потока повторных попыток $\tilde{\mu}_0$ определяется как предел

$$\begin{aligned}\tilde{\mu}_0 : &= \lim_{t\to\infty}\frac{\Lambda_0(t)}{t} = \lim_{t\to\infty}\frac{\sum_{i=0}^{H(t)}(1+\gamma_i)}{t}\\ &= \lim_{t\to\infty}\frac{\sum_{i=0}^{H(t)}(1+\gamma_i)}{H(t)}\,\frac{H(t)}{t}\\ &= \lambda\mathsf{P}_b[1 + \frac{\lambda+\mu_0}{\mu}],\end{aligned}$$

где $\gamma_0 := 0$. Альтернативный вывод (29) опирается на равенство $\tilde{\mu}_0 = \mu_0 q$, где $q = \mathsf{P}$(орбита не пуста)$=\sum_{i>0}(\mathsf{P}_{0i} + \mathsf{P}_{i1})$, и явный вид вероятностей (2), (3). □

Напомним доказанный выше результат (22) о том, что в системе вида $M/G/1/0$ стационарная вероятность простоя сервера

$$\mathsf{P}_0 := 1 - \mathsf{P}_b = 1 - \rho$$

и что условие (5) является *критерием стационарности* такой системы, см. [7]. Обозначим через (ν, N) основной стационарной процесс, т. е. $(\nu(t), N(t)) \Rightarrow (\nu, N)$. Перепишем условие (5) в виде

$$\rho + \rho\frac{\lambda}{\mu_0} < 1. \qquad (31)$$

В соответствии с общим принципом получения критерия стационарности, левая часть неравенства (31) должна быть равна предельной доле времени (вероятности) того, что *система не пуста*. Отсюда легко следует, что величина $\rho\lambda/\mu_0$ должна быть равна стационарной вероятности простоя сервера при непустой орбите. Иными словами, это *доля мощности, потерянной сервером из-за простоев при непустой орбите*. Поскольку эта вероятность положительна, то дисциплина в данной системе не является *консервативной*. Этот результат подтверждается в системе с повторными вызовами вида $M/M//1/0$, поскольку (см. (2))

$$\begin{aligned}&\mathsf{P}(\nu = 0,\, N > 0)\\ &= \mathsf{P}(\nu = 0) - \mathsf{P}(\nu = 0,\, N = 0)\\ &= 1 - \rho - \sum_{i>0}\mathsf{P}_{i0} = \rho\frac{\lambda}{\mu_0}.\end{aligned}$$

Результаты численного моделирования

В данном разделе представлены результаты имитационного моделирования и оценивания вероятности блокировки P_b в системах $M/M/1/0$ и $M/Pareto/1/0$ в стационарном и нестационарном режимах. Эти (точечные) оценки дают очень хорошее согласие с полученными выше теоретическими результатами (см. близость величин P_b, $\hat{\theta}_m(t)$ и $\hat{\Theta}_m(t)$ в таблицах 1, 2).

В качестве исходных характеристик системы задаются величины λ, μ_0, μ. На основе условия (5) в работе [1] введена *мера стационарности*

$$\Gamma \;:=\; 1/\mathsf{E}S - \Big(\frac{\lambda^2}{\mu_0} + \lambda\Big) \equiv \mu - \Big(\frac{\lambda^2}{\mu_0} + \lambda\Big),$$

значение которой положительно в области стационарности (и растет по мере "углубления" в область стационарности). Заметим, что при больших $\Gamma > 0$ реальная интенсивность орбитальных заявок $\tilde{\mu}_0 \ll \mu_0$, и заданный параметр μ_0 не является показателем качества обслуживания. Ниже рассматривается следующая оценка интенсивности $\tilde{\mu}_0$ потока заявок с орбиты на сервер

$$\hat{\mu}_0(t) = \frac{\sum_{i=1}^{k(t)} I_i}{t},$$

где индикатор $I_i = 1$, если i-я попытка обращения к серверу неудачна, а $k(t)$ – общее число попыток за время наблюдения $[0, t]$. При этом, в стационарном режиме с в. 1, $\hat{\mu}_0(t) \to \tilde{\mu}_0$ при $t \to \infty$. Можно также показать (см. [1]), что в нестационарном режиме (т. е. при $\Gamma < 0$)

Таблица 1. Оценивание P_b в системе вида $M/M/1/0,\ \lambda = 1,\ m = 50$

μ_0	$\hat{\mu}_0(t)$	μ	$\hat{\rho}(t)$	Γ	P_b	$\hat{\theta}_m(t)$	$\hat{\Theta}_m(t)$	$VR(t)$
4,00	0,150	10,000	0,120	8,750	0,100	0,101	0,100	0,003
14,000	0,590	6,000	0,270	4,930	0,167	0,167	0,167	0,001
0,600	0,510	3,00	0,500	0,330	0,333	0,333	0,331	0,221
0,050	0,050	20,000	0,050	-1,000	0,050	0,051	0,050	0,001
0,200	0,200	1,000	1,200	-5,000	0,545	0,546	0,545	0,001
0,100	0,100	0,100	11,010	-10,900	0,917	0,917	0,917	0,004

Таблица 2. Оценивание P_b в системе вида $M/Pareto/1/0,\ \lambda = 0,5,\ m = 50$

μ_0	$\hat{\mu}_0(t)$	α	$\hat{\rho}(t)$	Γ	P_b	$\hat{\theta}_m(t)$	$\hat{\Theta}_m(t)$	$VR(t)$
10,000	2,810	10,000	3,680	0,370	0,556	0,565	0,552	0,011
5,000	2,140	5,000	3,300	0,250	0,625	0,629	0,620	0,025
0,200	0,200	40,00	0,720	-0,780	0,418	0,417	0,418	0,000
0,100	0,100	2,100	1,150	-2,480	0,534	0,532	0,534	0,002

$\hat{\mu}_0(t) \to \mu_0$ с в. 1. Кроме того, используя аргументы теории регенерации, можно показать, что $\mathsf{E}\hat{\mu}_0(t) \to \mu_0$ на границе области стационарности, т. е. при $\Gamma = 0$. В целом, коэффициент

$$\hat{\rho}(t) = (\lambda + \hat{\mu}_0(t))\mathsf{E}S$$

выражает загрузку сервера (в числе попыток в единицу времени) на конечном интервале $[0,\, t]$. Эффективность оценок $\hat{\theta}_m(t)$ и $\hat{\Theta}_m(t)$ выражается величиной отношения их дисперсий

$$VR(t) := \frac{\hat{D}[\hat{\Theta}_m(t)]}{\hat{D}[\hat{\theta}_m(t)]},$$

где

$$\hat{D}[\hat{\Theta}_m(t)] = \frac{1}{m-1}\sum_{j=1}^{m}\Big[\Theta_j[k_j(t)] - \hat{\Theta}_m(t)\Big]^2,$$

$$\hat{D}[\hat{\theta}_m(t)] = \frac{1}{m-1}\sum_{j=1}^{m}\Big[\theta_j[k_j(t)] - \hat{\theta}_m(t)\Big]^2.$$

Подчеркнем, что в области стационарности оценки строились на конечном интервале $[0, t]$ по числу $k(t) = \max\limits_i\{i : t_i \leqslant t\}$ приходов первичных заявок, а в области нестационарности по числу $k(t) = \max\limits_i\{i : z_i \leqslant t\}$ попыток обращения к серверу. Ниже приведены оценки вероятности занятости $\hat{\theta}_m(t)$ и $\hat{\Theta}_m(t)$ на основе фиксированного числа реализаций $m = 50$. (Эксперименты при значениях $m = 25$, $m = 100$ также дали аналогичные результаты.)

В табл. 1 представлены результаты для системы вида $M/M/1/0$ при значении параметра потока первичных заявок $\lambda = 1$.

В глубине области стационарности (строки 1–2 табл. 1) дисперсия альтернативной оценки существенно меньше, чем дисперсия стандартной оценки, $VR(k) \ll 1$. Внутри области стационарности вблизи ее границы (строка, соответствующая $\Gamma = 0,330$) наблюдается снижение эффективности альтернативной оценки, однако ее дисперсия не превосходит дисперсию стандартной оценки. Внутри области нестационарности ($\Gamma < 0$) альтернативная оценка также эффективнее стандартной и близка к точному значению.

Результаты табл. 2 показывают, что в системе с повторными вызовами типа $M/Pareto/1/0$ эффективность альтернативной оценки вероятности P_b также превышает эффективность стандартной оценки.

Как легко проверить, оценки $\hat{\mu}_0$ (3 первые строки в столбце 2 табл. 1) очень близки к значениям интенсивности $\tilde{\mu}_0$ из формулы (29), равным 0,15; 0,58 и 0,50, соответственно.

На рис. 1 представлена зависимость величин $t\hat{D}[\hat{\theta}_m(t)]$ и $t^2\hat{D}[\hat{\Theta}_m(t)]$ от времени наблюдения t для системы вида $M/Pareto/1/0$.

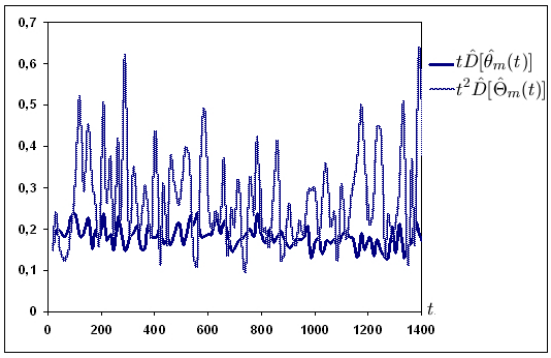

Рис. 1. Дисперсии оценок в системе $M/Pareto/1/0$ при $\lambda = 0,5$, $\mu_0 = 0,2$, $\alpha = 40$

Визуальный анализ рис. 1 показывает, что

$$t\,\hat{D}[\hat{\theta}_m(t)] \sim t^2\,\hat{D}[\hat{\Theta}_m(t)] \cdot C,$$

для некоторой постоянной C, что соответствует утверждениям Теорем 1, 2. Заметим также, что полученные выводы хорошо согласуются с результатами из [12] для систем с потерями.

Заключение

Численные результаты имитационного моделирования показали, что при оценивании вероятности занятости сервера на конечном интервале альтернативная оценка по остаточной длине цикла существенно эффективнее, чем стандартная, что согласуется с результатами из работы [12]. Эта эффективность выражается в уменьшении дисперсии оценки и проявляется как в области стационарности, так и в области нестационарности систем с повторными вызовами вида $M/M/1/0$ и $M/Pareto/1/0$. Однако необходимо учитывать, что альтернативная оценка строится на основе известного точного значения, которое не всегда известно.

Работа выполнена при поддержке РФФИ, проект 10-07-00017 и при поддержке Программы стратегического развития на 2012–2016 гг. «Университетский комплекс ПетрГУ в научно-образовательном пространстве Европейского Севера: стратегия инновационного развития».

Литература

1. *Морозов Е. В., Некрасова Р. С.* Оценивание вероятности блокировки в системе с повторными вызовами и постоянной скоростью возвращения заявок с орбиты // Труды Карельского научного центра РАН. Сер. Математическое моделирование и информационные технологии. Вып. 2. 2011. № 5. С. 63–74.

2. *Морозов Е. В., Некрасова Р. С.* Об оценивании вероятности переполнения конечного буфера в регенеративных системах обслуживания // Информатика и ее применения. 2012. Т. 6, № 3. С. 90–98.

3. *Artalejo J. R.* Stationary analysis of the characteristics of the M/M/2 queue with constant repeated attempts // Operation search. 1996. Vol. 33. P. 83–95.

4. *Artalejo J. R., Gómez-Corral A., Neuts M. F.* Analysis of multiserver queues with constant retrial rate // European Journal of Operational Research. 2001. Vol. 135. P. 569–581.

5. *Asmussen S.* Applied Probability and Queues. Springer, 2002. P. 476.

6. *Avrachenkov K., Goricheva R. S., Morozov E. V.* Verification of stability region of a retrial queuing system by regenerative method // Proceedings of the Intenational Conference "Modern Probabilistic Methods for Analysis and optimization of Information and Telecommunication Networks". Minsk, 2011. P. 22–28.

7. *Avrachenkov K., Morozov E. V.* Stability analysis of GI/G/c/K Retrial Queue with Constant Retrial Rate // INRIA (Sophia Antipolis), Research Report. 2010. N 7335. Available online at `http://hal.inria.fr/inria-00499261/en/`

8. *Avrachenkov K., Yechiali U.* Retrial networks with finite buffers and their application to Internet data traffic // Probability in the Engineering and Informational Sciences. 2008. Vol. 22. P. 519–536.

9. *Avrachenkov, K., Yechiali U.* On tandem blocking queues with a common retrial queue // Computers and Operations Research. 2010. N 37(7). P. 1174–1180.

10. *Billingsley P.* Convergence of probability measures 2nd edt. New-York: John Wiley, 1999. P. 296.

11. *Fayolle G.* A simple telephone exchange with delayed feedback/ In O. J. Boxma, J. W. Cohen and H. C. Tijms (eds.) // Teletraffic Analysis and Computer Performance Evaluation. 1986. Vol. 7. P. 245–253.

12. *Kang W., Whitt W., Shahabuddin P.* Exploting Regenerative Structure to estimate finite time averages via simulation // ACM. 2006. P. 1–38.

13. *Ramalhoto M. F., Gómez-Corral A.* Some decomposition formulae for M/M/r/r+d queues with constant retrial rate // Stochastic Models. 1998. Vol. 14. P. 123–145.

14. *Smith W. L.* Regenerative stochastic processes // Proc. Royal Soc. Ser. A. 1955. Vol. 232. P. 6–31.

СВЕДЕНИЯ ОБ АВТОРАХ:

Морозов Евсей Викторович
ведущий научный сотрудник, д. ф.-м. н.
Институт прикладных математических исследований Карельского научного центра РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185610,
эл. почта: emorozov@karelia.ru
тел.: (8142) 763370

Некрасова Руслана Сергеевна
аспирантка
Институт прикладных математических исследований Карельского научного центра РАН
ул. Пушкинская, 11, Петрозаводск, Республика Карелия, Россия, 185610,
эл. почта: ruslana.nekrasova@mail.ru
тел.: (8142) 763370

Morozov, Evsey
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences
11 Pushkinskaya St., 185610 Petrozavodsk, Karelia, Russia
e-mail: emorozov@karelia.ru
tel.: (8142) 763370

Nekrasova, Ruslana
Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences
11 Pushkinskaya St., 185610 Petrozavodsk, Karelia, Russia
e-mail: ruslana.nekrasova@mail.ru
tel.: (8142) 763370