Информационные процессы, Том 3, № 3, 2013, стр. 125–140.

Стационарные характеристики системы $M_2|G|1|r$ с гистерезисной политикой управления интенсивностью входящего потока[1]

А.В. Печинкин, Р.В. Разумчик

Институт проблем информатики РАН, Москва, Россия

Поступила в редколлегию 5.07.2013

Аннотация—Рассматривается система массового обслуживания конечной ёмкости с одним прибором, пуассоновским входящим потоком, произвольным распределением времени обслуживания и гистерезисной политикой управления интенсивностью входящего потока. Предложен аналитический метод нахождения стационарного распределения числа заявок в системе. Представлены некоторые результаты численных расчётов, проведённых на основе полученных соотношений.

КЛЮЧЕВЫЕ СЛОВА: система массового обслуживания, перегрузки, управление интенсивностью входящего потока, гистерезисная политика.

1. ВВЕДЕНИЕ

Одним из способов предотвращения различного рода перегрузок в информационно-телекоммуникационных системах, моделируемых с помощью систем массового обслуживания (СМО), является использование порогового управления нагрузкой [1]. В качестве механизма часто применяется гистерезисное управление с несколькими типами порогов для контроля перегрузок. Управление может осуществляться как входящим потоком и его параметрами, так и процессом обслуживания. В силу практической важности, данной проблематике посвящено большое число отечественных и зарубежных публикаций. В качестве примера можно привести работы [2]–[13]. В частности, достаточно подробный обзор результатов, полученных в этом направлении до конца прошлого столетия, можно найти, например, в [2] и [3].

В настоящей работе рассматривается СМО, в которой управление осуществляется только параметрами входящих потоков, но при этом гистерезисное управление состоит из двух петель, соответствующих различным уровням принимаемой в систему нагрузки. Ниже остановимся на обзоре некоторых наиболее близких к данной тематике работ.

В [14] рассматривалась СМО конечной ёмкости $MMPP|G|1|K$ с двумя порогами L_1 и L_2, $0 < L_1 \leq L_2 < K$ и одной петлей гистерезиса. Таким образом, система может находиться в двух состояний: недогруженном и перегруженном. Управление входящим потоком осуществляется по моментам изменения числа заявок в системе. В качестве метода исследования использовался метод введения вспомогательной переменной — прошедшего времени обслуживания. Получены выражения для расчёта совместного стационарного распределения числа заявок в системе, состояния системы и прошедшего времени обслуживания. В работе [15] с помощью того же метода исследована практически аналогичная СМО (с одной петлей гистерезиса) с тем лишь отличием, что управление входящим потоком осуществляется по моментам окончания обслуживания заявки на приборе. Система бесконечной ёмкости с k входящими

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований (гранты № 11-07-00112, № 12-07-00108 и № 12-07-00223)

$MMPP$ потоками различных типов, фазовым распределением времени обслуживания, r порогами (без петель гистерезиса) и нетерпеливыми заявками рассмотрена в [16]. В зависимости от того, между какими пороговыми значениями находится загрузка системы в момент поступления новой заявки, принимается решение о постановке новой заявки в очередь или её отбрасывании. Характеристики входящего потока зависят от того, была ли принята к обслуживанию новая заявка или нет. В работе найдено стационарное распределение вероятностей состояний системы. Система обслуживания $M_Q|M|m$ с повторными вызовами, двумя порогами и одной петлей гистерезиса рассмотрена в [17], где получен алгоритм рекуррентного типа для приближённого расчёта стационарных характеристик системы. В работе [18] рассмотрен класс систем обслуживания с входящим потоком, описываемым процессом Леви, произвольным обслуживанием и одной петлей гистерезиса, определяемой двумя порогами, для которых найдено стационарное распределение процесса работы. Система $M|G|1$ бесконечной ёмкости с инверсионным порядком обслуживания, вероятностным приоритетом и гистерезисной политикой изучена в [19]. Предполагается, что система может функционировать в двух режимах, каждый из которых характеризуется своей интенсивностью входящего потока и функцией распределения времени обслуживания заявки. С помощью метода введения дополнительной переменной (остаточного времени обслуживания) и некоторых результатов теории эргодических процессов получены выражения для основных стационарных вероятностно-временных характеристик этой системы. Ещё одно исследование СМО $M|G|1$ с двухпороговой гистерезисной политикой, но конечной емкости и двумя петлями гистерезиса, предпринято в [20] и [21]. В этих работах предполагается, что управление осуществляется только входящим потоком по моментам окончания обслуживания заявок на приборе. С помощью метода вложенной цепи Маркова находятся стационарные характеристики производительности системы. Заметим, что в работе [8], посвящённой исследованию двухпороговой системы $M^x|G|1$ с одной петлей гистерезиса также допускается управление интенсивностью входящего потока.

Метод, применяемый в настоящей работе для исследования СМО $M_2|G|1$ конечной ёмкости с гистерезисной политикой управления интенсивностью входящего потока (см. рис. 1), позволяет эффективно “обсчитывать” петли гистерезиса и получать рациональные с вычислительной точки зрения алгоритмы для расчёта стационарных вероятностных характеристик системы. Основное отличие рассматриваемой здесь СМО от СМО, изученных в [20] и [21], заключается в том, что переключение режима функционирования системы осуществляется не в моменты окончания обслуживания заявки на приборе, а в моменты изменения числа заявок в системе, что более естественно для практических применений. Заметим, что используемый в данной работе метод допускает обобщение на семейство гистерезисных петель, расположенных по тому же принципу, что и на рис. 1.

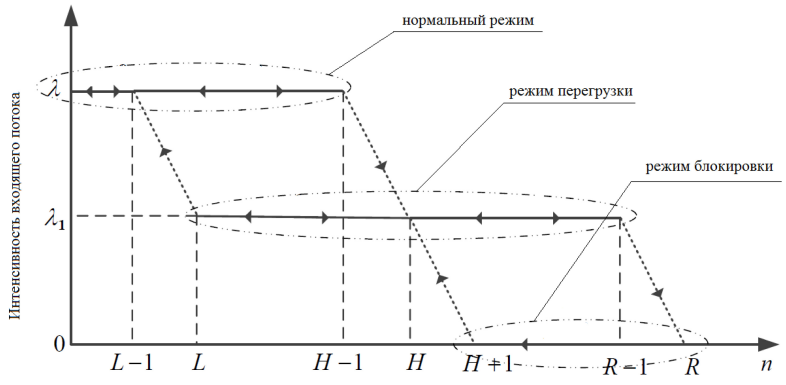

Рис. 1. Схема гистерезисной политики управления интенсивностью входящего потока

В заключение отметим, что, согласно [1], СМО с гистерезисным управлением могут служить адекватными моделями для оценки качества функционирования SIP-серверов с пороговым управлением в условиях перегрузок. Подробное описание применения подобного типа СМО к моделированию SIP-серверов, работающих в условиях перегрузки, можно найти в [1], [22] и [23].

2. ОПИСАНИЕ СИСТЕМЫ

Рассмотрим однолинейную СМО конечной ёмкости R, в которую поступают независимые пуассоновские потоки заявок двух типов, причём $\lambda_k, \quad k = 1, 2$, — интенсивность потока k-го типа. Через $\lambda = \lambda_1 + \lambda_2$ далее будем обозначать суммарную интенсивность этих потоков. Функция распределения времени обслуживания (длины) заявок каждого типа равна $B(x)$. Будем предполагать для простоты изложения, что существует плотность $b(x) = B'(x)$. Через $\beta(s) = \int_0^\infty e^{-sx} b(x)\,dx$ обозначим преобразование Лапласа–Стилтьеса (ПЛС) длины заявки, а через $b = \int_0^\infty x\,b(x)\,dx$ — среднюю длину заявки. Положим также $\tilde{\beta}^{(n)}(s) = \int_0^\infty x^n e^{-sx}[1 - B(x)]\,dx, \quad n \geq 0$.

В системе реализован двухпороговый гистерезисный механизм управления поступающей нагрузкой, действующий следующим образом (рис. 1). Имеется два числа L и H, для которых справедливы неравенства $0 < L < H < R$. С момента поступления в систему первой заявки и до того момента, когда в системе впервые окажется H заявок, система функционирует в нормальном режиме и к обслуживанию принимаются заявки обоих типов. Но как только в системе окажется H заявок, система переходит в режим перегрузки, прекращается приём заявок второго типа и принимаются лишь заявки первого типа. Это продолжается до того момента, когда в системе станет либо $(L-1)$, либо R заявок. В первом случае система переходит в нормальный режим функционирования и снова начинают приниматься заявки обоих типов. Во втором случае система переходит в режим блокировки и прекращается приём всех заявок (заявки только обслуживаются) до тех пор, пока в системе снова не окажется H заявок. Тогда система переходит в режим перегрузки и снова начинается приём заявок первого типа. Эта процедура продолжается и далее.

Будем предполагать, что выполнено условие $b < \infty$, необходимое и достаточное для существования стационарного режима функционирования рассматриваемой системы. Будем считать также, что выполнены неравенства $L \geq 2, \quad H - L \geq 2$ и $R - H \geq 2$. Эти предположения вводятся только для того, чтобы не рассматривать случаи, расчётные формулы для которых несколько отличаются от приводимых здесь, и нисколько не умоляют общности полученных результатов.

Введём процесс $\{X(t), \quad t \geq 0\}$, описывающий функционирование рассматриваемой системы. Пусть $\xi(t)$ — общее число заявок в системе в момент t, $\eta(t)$ — прошедшее время обслуживания, т.е. то время, которое уже обслуживалась заявка, находящаяся в момент t на приборе, $\nu(t)$ — состояние системы в момент t. Процесс $X(t)$ определим как $\{X(t) = (\xi(t), \eta(t), \nu(t)), \quad t \geq 0\}$. Если $\xi(t) = 0$, то компоненты $\eta(t)$ и $\nu(t)$ опускаются, и, следовательно, $X(t) = \xi(t)$. Можно показать (см., например, [24, с. 235]), что процесс $X(t)$ является марковским. Множество состояний процесса $X(t)$ записывается в виде $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \mathcal{X}_2$, где

$$\mathcal{X}_0 = \{0\} \cup \{(n, x, 0) : 0 < n \leq H - 1, \ x \geq 0\},$$
$$\mathcal{X}_1 = \{(n, x, 1) : L \leq n \leq R - 1, \ x \geq 0\},$$
$$\mathcal{X}_2 = \{(n, x, 2) : H + 1 \leq n \leq R, \ x \geq 0\}.$$

Для произвольного момента t нахождение процесса $X(t)$ в состоянии (0) означает, что система свободна от заявок, а в состоянии множества $\mathcal{X}_s$ соответствует ситуации, когда в системе

имеется n заявок, обслуживание находящейся на приборе заявки продолжается время x и система находится в нормальном режиме (при $s = 0$), в режиме перегрузки (при $s = 1$) и в режиме блокировки (при $s = 2$).

Как уже упоминалось выше, при $b < \infty$ существуют предельные (стационарные) вероятности процесса $X(t)$, которые обозначим через p_0 и $P_{ns}(x)$. Можно также показать, что при всех допустимых значениях n и s существуют плотности $p_{ns}(x) = P'_{ns}(x)$. Поясним, что представляет собой каждая из функций $p_{ns}(x)$. Так, $p_{n0}(x)$, $n = \overline{1, H-1}$, — плотность стационарной вероятности того, что в системе находится n заявок, заявка на приборе имеет (обслуженную) длину меньше x и система находится в нормально режиме (т.е. принимаются заявки обоих типов); $p_{n1}(x)$, $n = \overline{H, R-1}$, — плотность стационарной вероятности того, что в системе находится n заявок, заявка на приборе имеет (обслуженную) длину меньше x и система находится в режиме перегрузки (принимаются только заявки первого типа); наконец $p_{n2}(x)$, $n = \overline{H+1, R}$, — плотность стационарной вероятности того, что в системе находится n заявок, заявка на приборе имеет (обслуженную) длину меньше x и система находится в режиме блокировки (поступающие заявки любого типа не принимаются).

В настоящей работе для нахождения стационарных плотностей $p_{ns}(x)$ применяется метод анализа СМО, основанный на исключении отдельных временных интервалов и позволяющий свести решение исходной задачи к последовательному решению простых уравнений. Этот метод был с успехом применён в работах [25]–[29] для анализа других СМО с гистерезисной политикой и показал свою эффективность при расчёте показателей их производительности.

3. ВСПОМОГАТЕЛЬНЫЕ ФУНКЦИИ

Прежде чем выписать и решить уравнения, которым удовлетворяют стационарные плотности вероятностей $p_{ns}(x)$, необходимо ввести несколько вспомогательных функций.

Пусть в начальный момент в системе имеется n, $n = \overline{H+1, R-1}$, заявок, причём в систему принимаются только заявки первого типа (система находится в режиме перегрузки), а заявка на приборе имеет (обслуженную) длину x. Обозначим через $\alpha_n(x)$ вероятность того, что до момента, когда в системе впервые останется $(n-1)$ заявок, в ней никогда не будет R заявок.

Выведем уравнение для вероятности $\alpha_{R-1}(x)$. Если в начальный момент времени в системе имеется $(R-1)$ заявок, она находится в режиме перегрузки и заявка на приборе имеет (обслуженную) длину x, то число заявок в системе станет равным $(R-2)$ раньше, чем R, в следующих случаях:

- за “малое” время Δ закончится обслуживание заявки (вероятность этого события равна $b(x)\Delta/[1-B(x)] + o(\Delta)$);
- за “малое” время Δ не поступят заявки (с вероятностью $1 - \lambda_1\Delta + o(\Delta)$), не окончится обслуживание заявки на приборе (с вероятностью $[1-B(x+\Delta)]/[1-B(x)] + o(\Delta)$), а затем до того момента, когда в системе впервые останется $(R-2)$ заявок, в ней никогда не будет R заявок, но теперь уже при условии, что в системе $(R-1)$ заявок, система находится в режиме перегрузки, однако обслуженная длина заявки на приборе равна $x + \Delta$ (вероятность этого события равна $\alpha_{R-1}(x+\Delta)$).

Поскольку остальные события имеют вероятность $o(\Delta)$, по формуле полной вероятности имеем

$$\alpha_{R-1}(x) = \frac{b(x)\Delta}{1-B(x)} + (1-\lambda_1\Delta)\frac{1-B(x+\Delta)}{1-B(x)}\alpha_{R-1}(x+\Delta) + o(\Delta). \tag{1}$$

В дальнейшем будет удобно работать не с вероятностями $\alpha_n(x)$, а с функциями

$$q_n(x) = [1-B(x)]\alpha_n(x), \quad n = \overline{H+1, R-1}. \tag{2}$$

Тогда, подставляя в (1) выражение для $\alpha_{R-1}(x)$ через $q_{R-1}(x)$, получаем

$$q_{R-1}(x) - q_{R-1}(x+\Delta) = -\lambda_1 \Delta q_{R-1}(x+\Delta) + b(x)\Delta + o(\Delta),$$

откуда, деля на Δ и устремляя Δ к нулю, приходим к дифференциальному уравнению

$$-q'_{R-1}(x) = -\lambda_1 q_{R-1}(x) + b(x), \tag{3}$$

решение которого имеет вид

$$q_{R-1}(x) = ce^{\lambda_1 x} + \int_x^\infty b(y)e^{-\lambda_1(y-x)}dy,$$

где c — произвольная постоянная. Поскольку $|\alpha_{R-1}(x)| \le 1$ для всех $x \ge 0$, то $c = 0$ и

$$q_{R-1}(x) = \int_x^\infty b(y)e^{-\lambda_1(y-x)}dy.$$

Перейдем теперь к выводу уравнения для вероятности $\alpha_n(x)$, $n = \overline{H+1, R-2}$. Если в начальный момент времени в системе имеется n, $n = \overline{H+1, R-2}$, заявок, она находится в режиме перегрузки и заявка на приборе имеет (обслуженную) длину x, то число заявок в системе станет равным $(n-1)$ раньше, чем R, в следующих случаях:

- за "малое" время Δ закончится обслуживание заявки (вероятность этого события равна $b(x)\Delta/[1-B(x)] + o(\Delta)$);
- за "малое" время Δ не поступит новая заявка (с вероятностью $1-\lambda_1\Delta+o(\Delta)$), не окончится обслуживание заявки на приборе (с вероятностью $[1-B(x+\Delta)]/[1-B(x)]+o(\Delta)$), а затем до того момента, когда в системе впервые останется $(n-1)$ заявок, в ней никогда не будет R заявок, но теперь уже при условии, что в системе n заявок, система находится в режиме перегрузки, однако обслуженная длина заявки на приборе равна $x+\Delta$ (вероятность этого события равна $\alpha_n(x+\Delta)$);
- за "малое" время Δ поступит одна заявка (с вероятностью $\lambda_1\Delta + o(\Delta)$), затем до того момента, когда в системе впервые останется n заявок, в ней никогда не будет R заявок при условии, что сейчас в системе $n+1$ заявок, обслуженная длина заявки на приборе равна $x+\Delta$ и система находится в режиме перегрузки (вероятность этого события равна $\alpha_{n+1}(x+\Delta)$) и, наконец, до того момента, когда число заявок в системе станет равным $(n-1)$, в ней никогда не будет R заявок при условии, что теперь в системе n заявок, она находится в режиме перегрузки и обслуженная длина заявки на приборе равна нулю (вероятность этого события равна $\alpha_n(0)$).

Остальные события имеют вероятность $o(\Delta)$, поэтому по формуле полной вероятности

$$\alpha_n(x) = \frac{b(x)\Delta}{1-B(x)} + (1-\lambda_1\Delta)\frac{1-B(x+\Delta)}{1-B(x)}\alpha_n(x+\Delta) + \\ + \lambda_1\Delta\alpha_{n+1}(x+\Delta)\alpha_n(0) + o(\Delta), \quad n = \overline{H+1, R-2},$$

откуда, учитывая (2) и осуществляя предельный переход при $\Delta \to 0$, получаем уравнение

$$-q'_n(x) = -\lambda_1 q_n(x) + b(x) + \lambda_1 q_{n+1}(x)\alpha_n(0), \quad n = \overline{H+1, R-2}. \tag{4}$$

Нетрудно убедиться, что решение уравнения (4) с учётом неравенства $|\alpha_n(x)| \le 1$ для всех $x \ge 0$ имеет вид

$$q_n(x) = \int_x^\infty [b(y) + \lambda_1 \alpha_n(0) q_{n+1}(y)] e^{-\lambda_1 (y-x)} dy, \quad n = \overline{H+1, R-2}. \tag{5}$$

В (5) присутствует неизвестный множитель $\alpha_n(0)$. Для его нахождения воспользуемся (2) при $x = 0$. В результате из (5) получаем

$$q_n(0) = \alpha_n(0) = \int_0^\infty [b(y) + \lambda_1 \alpha_n(0) q_{n+1}(y)] e^{-\lambda_1 y} dy,$$

откуда находим следующее выражение для $\alpha_n(0)$:

$$\alpha_n(0) = \frac{\beta(\lambda_1)}{1 - \lambda_1 \int\limits_0^\infty q_{n+1}(y) e^{-\lambda_1 y} dy}, \quad n = \overline{H+1, R-2}.$$

Последняя формула вместе с (5) и (2) позволяет последовательно вычислять значения вероятности $\alpha_n(x)$ для всех $x \ge 0$ и $n = \overline{H+1, R-2}$.

Обратимся теперь к вычислению ещё одной вспомогательной функции.

Предположим, что в начальный момент в системе имеется n, $n = \overline{L, H-1}$, заявок, в систему принимаются заявки любого типа (т.е. система функционирует в нормальном режиме), а заявка на приборе имеет (обслуженную) длину x. Обозначим через $\alpha_n(x)$ вероятность того, что до того момента, когда в системе впервые останется $(n-1)$ заявок, в ней никогда не будет H заявок.

Вводя обозначение

$$q_n(x) = [1 - B(x)] \alpha_n(x), \quad n = \overline{L, H-1}, \tag{6}$$

рассматривая, как и ранее, возможные переходы за "малое" время Δ, осуществляя предельный переход при $\Delta \to 0$ и затем решая получившиеся уравнения, приходим к следующим формулам для $q_n(x)$:

$$q_{H-1}(x) = \int_x^\infty b(y) e^{-\lambda(y-x)} dy,$$

$$q_n(x) = \int_x^\infty [b(y) + \lambda \alpha_n(0) q_{n+1}(y)] e^{-\lambda(y-x)} dy, \quad n = \overline{L, H-2}. \tag{7}$$

В формуле (7) фигурирует неизвестный множитель $\alpha_n(0)$, который находится путем подстановки $x = 0$ в (7) с учётом (6). В результате получаем

$$q_n(0) = \alpha_n(0) = \frac{\beta(\lambda)}{1 - \lambda \int\limits_0^\infty q_{n+1}(y) e^{-\lambda y} dy}, \quad n = \overline{L, H-2}.$$

Полученная формула вместе с (6) и (7) позволяет последовательно вычислять значения вероятности $\alpha_n(x)$ для всех $x \ge 0$ и $n = \overline{L, H-2}$.

4. СТАЦИОНАРНЫЕ ПЛОТНОСТИ ВЕРОЯТНОСТЕЙ СОСТОЯНИЙ

Приступим к нахождению стационарных плотностей вероятности $p_{ns}(x)$.

Начнём со случая $s = 0$ и $n = \overline{1, L-1}$. Хотя в этом случае приводимые здесь формулы хорошо известны, для понимания дальнейших выкладок полезно их повторить с использованием метода исключения состояний.

Для того чтобы система оказалась в состоянии $(n, x, 0)$, необходимо выполнение следующих условий:

- система находилась в состоянии $(n, x-\Delta, 0)$, за "малое" время Δ не поступали новые заявки и не окончилось обслуживания заявки, находящейся на приборе;
- система находилась в состоянии $(n-1, x-\Delta, 0)$, за "малое" время Δ поступила одна заявка и не окончилось обслуживания заявки, находящейся на приборе (заметим, что если $n = 1$, то последнее событие произойти не может, поскольку прошедшее время обслуживания любой поступающей в свободную систему заявки равно нулю).

Так как остальные события имеют вероятность $o(\Delta)$, то по формуле полной вероятности имеем

$$p_{n0}(x) = p_{n0}(x-\Delta)(1-\lambda\Delta)\frac{1-B(x)}{1-B(x-\Delta)} + \\ + u(n-1)p_{n-1,0}(x-\Delta)\lambda\Delta\frac{1-B(x)}{1-B(x-\Delta)} + o(\Delta), \quad n = \overline{1, L-1}, \tag{8}$$

где $u(x)$ — функция Хевисайда.

В дальнейшем будет удобнее работать не с плотностями $p_{ns}(x)$, а с функциями

$$r_{ns}(x) = \frac{p_{ns}(x)}{1-B(x)}. \tag{9}$$

Тогда после предельного перехода при $\Delta \to 0$ получаем следующую систему уравнений:

$$r'_{10}(x) = -\lambda r_{10}(x), \qquad r'_{n0}(x) = -\lambda r_{n0}(x) + \lambda r_{n-1,0}(x), \quad n = \overline{2, L-1},$$

решение которой имеет вид

$$r_{n0}(x) = e^{-\lambda x}\sum_{i=1}^{n}\frac{(\lambda x)^{i-1}}{(i-1)!}c_{n+1-i}, \quad n = \overline{1, L-1}, \tag{10}$$

где c_n — некоторые постоянные, определяемые из граничных условий.

Для нахождения постоянной c_n исключим из рассмотрения все те моменты времени, когда в системе больше n заявок. Тогда если исходная система находилась в состоянии $(n, x, 0)$ и поступила новая заявка, то произошёл переход в состояние $(n+1, x, 0)$. Однако в системе с исключёнными состояниями при тех же условиях переход произойдет в состояние $(n, 0, 0)$. Поскольку уравнения для стационарных плотностей вероятности $p_{ns}(x)$ одинаковы для обеих систем, то, применяя для исходной системы уравнения для системы с исключёнными состояниями, получаем

$$p_{10}(0) = \lambda p_0 + \int_0^\infty \lambda p_{10}(x)\,dx,$$

$$p_{n0}(0) = \int_0^\infty \lambda p_{n0}(x)\,dx, \ n = \overline{2, L-1},$$

или

$$c_1 = \frac{\lambda}{1-\lambda\tilde{\beta}^{(0)}(\lambda)}p_0,$$

$$c_n = \frac{\lambda}{1-\lambda\tilde{\beta}^{(0)}(\lambda)}\sum_{i=1}^{n-1}\frac{\lambda^i}{i!}\tilde{\beta}^{(i)}(\lambda)c_{n-i}, \quad n=\overline{2,L-1}.$$

Перейдем к нахождению стационарных плотностей вероятности $p_{n0}(x)$ при $n=\overline{L,H-1}$. Уравнение, которому удовлетворяют данные плотности, полностью совпадают с уравнением (8) при $n=\overline{2,L-1}$ и поэтому приводятся без пояснений:

$$p_{n0}(x) = p_{n0}(x-\Delta)(1-\lambda\Delta)\frac{1-B(x)}{1-B(x-\Delta)} + p_{n-1,0}(x-\Delta)\lambda\Delta\frac{1-B(x)}{1-B(x-\Delta)} + o(\Delta), \quad n=\overline{L,H-1}.$$

Из этого уравнения после подстановки (9) и несложных преобразований имеем

$$r'_{n0}(x) = -\lambda r_{n0}(x) + \lambda r_{n-1,0}(x), \quad n=\overline{L,H-1}.$$

Решая данную систему для $n=L,L+1$ и применяя затем метод математической индукции, получаем

$$r_{n0}(x) = e^{-\lambda x}\sum_{i=1}^{n}\frac{(\lambda x)^{i-1}}{(i-1)!}c_{n+1-i}, \quad n=\overline{L,H-1}, \tag{11}$$

где c_n — постоянные, которые вычислим сейчас из граничных условий.

Для нахождения c_n заметим сначала, что процесс $X(t)$ никогда не посещает состояния $(H-1,0,0)$, и поэтому

$$c_{H-1} = p_{H-1,0}(0) = 0.$$

Теперь воспользуемся тем же самым методом исключения состояний и выкинем из рассмотрения те моменты времени, когда в рассматриваемой системе находится более n заявок. Тогда если исходная система находилась в состоянии $(n,x,0)$ и поступила новая заявка, то произошёл переход в состояние $(n+1,x,0)$. Однако, в отличие от ранее рассмотренного случая, в системе с исключёнными состояниями при поступлении новой заявки переход из состояния $(n,x,0)$ в состояние $(n,0,0)$ произойдёт только с вероятностью $\alpha_{n+1}(x)$. Поэтому для для случая $n=\overline{L+1,H-2}$ граничное условие имеет вид

$$p_{n0}(0) = c_n = \int_0^\infty \lambda\alpha_{n+1}(x)p_{n0}(x)\,dx, \quad n=\overline{L,H-2}.$$

Отсюда с помощью подстановок (6) и (9), а также с учётом (11) находим выражение для коэффициента c_n, $n=\overline{L,H-2}$,

$$c_n = \frac{\lambda}{1-\lambda q_{n,0}}\sum_{i=1}^{n-1}q_{n,i}c_{n-i}, \quad n=\overline{L,H-2},$$

где

$$q_{n,i} = \int_0^\infty e^{-\lambda x}q_{n+1}(x)\frac{(\lambda x)^i}{i!}dx, \quad n=\overline{L,H-2}, \quad i=\overline{0,n-1}.$$

Найдём стационарную плотность вероятности $p_{n1}(x)$, $n=\overline{L,H-1}$. Так как в данном случае система находится в режиме перегрузки, то в соответствии с гистерезисной стратегией

интенсивность поступления заявок снижается до λ_1. Система уравнений, которой удовлетворяют плотности $p_{n1}(x)$ полностью совпадает (с учётом замены λ на λ_1) с системой уравнений, которым удовлетворяют плотности $p_{n0}(x)$, $n = \overline{1, L-1}$, т.е. имеет вид

$$p_{n1}(x) = p_{n1}(x-\Delta)(1-\lambda_1\Delta)\frac{1-B(x)}{1-B(x-\Delta)} + \\ + u(n-L)p_{n-1,1}(x-\Delta)\lambda_1\Delta\frac{1-B(x)}{1-B(x-\Delta)} + o(\Delta), \quad n = \overline{L, H-1}. \tag{12}$$

Подставляя в предыдущее равенство выражение для $p_{n1}(x)$ через $r_{n1}(x)$ согласно (9), получаем систему уравнений

$$r'_{L1}(x) = -\lambda_1 r_{L1}(x), \qquad r'_{n1}(x) = -\lambda_1 r_{n1}(x) + \lambda_1 r_{n-1,1}(x), \quad n = \overline{L+1, H-1}.$$

Решая данную систему для $n = L, L+1$ и т.д., имеем

$$r_{n1}(x) = e^{-\lambda_1 x}\sum_{i=L}^{n}\frac{(\lambda_1 x)^{i-L}}{(i-L)!}\tilde{c}_{n+L-i}, \quad n = \overline{L, H-1}, \tag{13}$$

где $\tilde{c}_n$ — некоторые постоянные, причём $\tilde{c}_n = r_{n1}(0) = p_{n1}(0)$.

Для определения $\tilde{c}_n$ воспользуемся тем же приёмом, что и раньше.

Начнём со случая $n = L$. Исключим из рассмотрения все те моменты времени, в которые число заявок в системе больше $(L-1)$ и система находится в нормальном режиме или в системе больше L заявок и система находится в режиме перегрузки или блокировки. Тогда в системе с исключёнными состояниями в состояние $(L, 0, 1)$ можно попасть либо из любого состояния $(L, x, 1)$, либо из любого состояния $(L, x, 0)$. В первом случае это происходит с вероятностью единица. Во втором случае сначала нужно достичь любого состояния $(H, y, 1)$ (это происходит мгновенно с вероятностью $[1-\alpha_L(x)]$), а затем уже (также мгновенно) с вероятностью единица попасть в состояние $(L, 0, 1)$. Заменяя уравнение для граничного состояния исходной системы уравнением для граничного состояния системы с исключёнными состояниями, получаем

$$p_{L1}(0) = \tilde{c}_L = \lambda\tilde{\alpha} + \lambda_1\int_0^\infty p_{L1}(x)\,dx,$$

где

$$\tilde{\alpha} = \int_0^\infty [1-\alpha_L(x)]p_{L-1,0}(x)\,dx.$$

Проводя аналогичные рассуждения для случая $n = \overline{L+1, H-1}$, получаем следующие граничные условия:

$$p_{n1}(0) = \tilde{c}_n = \lambda\tilde{\alpha} + \lambda_1\int_0^\infty p_{n1}(x)\,dx, \ n = \overline{L, H-1}.$$

Используя подстановку (9) и с учётом (13), из последнего равенства находятся выражения для всех коэффициентов c_n, $n = \overline{L, H-1}$:

$$\tilde{c}_L = \frac{\lambda\tilde{\alpha}}{1-\lambda_1\tilde{\beta}^{(0)}(\lambda_1)},$$

$$\tilde{c}_n = \tilde{c}_L + \frac{1}{1-\lambda_1\tilde{\beta}^{(0)}(\lambda_1)}\sum_{i=1}^{n-L}\frac{\lambda_1^{i+1}}{i!}\tilde{\beta}^{(i)}(\lambda_1)c_{n-i}, \quad n = \overline{L+1, H-1}.$$

Обратимся к нахождению стационарной плотности вероятности $p_{H1}(x)$. Проводя аналогичные рассуждения, нетрудно убедиться, что плотность $p_{H1}(x)$ удовлетворяет уравнению

$$p_{H1}(x) = p_{H1}(x-\Delta)(1-\lambda_1\Delta)\frac{1-B(x)}{1-B(x-\Delta)} + \\ + p_{H-1,0}(x-\Delta)\lambda\Delta\frac{1-B(x)}{1-B(x-\Delta)} + p_{H-1,1}(x-\Delta)\lambda_1\Delta\frac{1-B(x)}{1-B(x-\Delta)} + o(\Delta),$$

которое после подстановки (9) приводится к виду

$$r'_{H1}(x) = -\lambda_1 r_{H1}(x) + \lambda r_{H-1,0}(x) + \lambda_1 r_{H-1,1}(x).$$

Решения этого уравнения имеет вид

$$r_{H1}(x) = e^{-\lambda_1 x}\left(c_H + \lambda\int\limits_0^x e^{\lambda_1 y} r_{H-1,0}(y)\,dy + \lambda_1\int\limits_0^x e^{\lambda_1 y} r_{H-1,1}(y)\,dy\right), \tag{14}$$

где $c_H = r_{H1}(0) = p_{H1}(0)$ — некоторая постоянная, которая находится из граничного условия. Воспользовавшись тем же самым приёмом, что и в предыдущих случаях, получаем следующее соотношение для плотности $p_{H1}(0)$:

$$p_{H1}(0) = c_H = \int\limits_0^\infty \lambda_1 p_{H1}(x)\,dx.$$

Подставляя в последнее равенство выражение для $p_{H1}(x)$ через $r_{H1}(x)$ согласно (9), а затем выражение для $r_{H1}(x)$ в соответствии с (14), после преобразований получаем, что c_H вычисляется по формуле

$$c_H = \frac{\lambda_1}{1-\lambda_1\tilde{\beta}^{(0)}(\lambda_1)}\left[\lambda\int\limits_0^\infty [1-B(x)]e^{-\lambda_1 x}dx\int\limits_0^x e^{\lambda_1 y} r_{H-1,0}(y)\,dy + \right. \\ \left. + \lambda_1\int\limits_0^\infty [1-B(x)]e^{-\lambda_1 x}dx\int\limits_0^x e^{\lambda_1 y} r_{H-1,1}(y)\,dy\right].$$

Уравнение для стационарной плотности вероятности $p_{n1}(x)$ при $n = \overline{H+1, R-1}$ полностью повторяет (с учётом замены λ на λ_1) уравнение для стационарной плотности $p_{n0}(x)$ при $n = \overline{L, H-1}$. Поэтому без пояснений приводим дифференциальное уравнение, которому удовлетворяет $p_{n1}(x)$ с учётом подстановки (9):

$$r'_{n1}(x) = -\lambda_1 r_{n1}(x) + \lambda_1 r_{n-1,1}(x), \quad n = \overline{H+1, R-1}.$$

Решение этой системы имеет вид

$$r_{n1}(x) = e^{-\lambda_1 x}\left(c_n + \lambda_1\int\limits_0^x e^{\lambda_1 y} r_{n-1,1}(y)\,dy\right), n = \overline{H+1, R-1}, \tag{15}$$

где c_n — некоторые постоянные. Рассуждая так же, как и при выводе граничных условий для постоянных c_n при $n = \overline{L, H-1}$, находим

$$p_{R-1,1}(0) = c_{R-1,1} = 0, \qquad p_{n1}(0) = c_n = \int\limits_0^\infty \lambda_1\alpha_{n+1}(x)p_{n1}(x)\,dx, \quad n = \overline{H+1, R-2},$$

откуда после подстановки $p_{n1}(x)$ через $r_{n1}(x)$ согласно (9), а также $\alpha_n(x)$ через $q_n(x)$ согласно (2) и приведения подобных слагаемых получаем следующую формулу для расчёта коэффициентов c_n:

$$c_n = \frac{\lambda_1^2 \alpha_n}{1 + \alpha_n - q_n(0)} \left(\int_0^\infty e^{-\lambda_1 x} q_{n+1}(x)\, dx \int_0^x e^{\lambda_1 y} r_{n-1,1}(y)\, dy \right), \quad n = \overline{H+1, R-2}.$$

Наконец, найдем стационарные плотности вероятности $p_{n2}(x)$. Уравнения, которым удовлетворяют эти плотности, имеют вид

$$p_{R2}(x) = p_{R2}(x-\Delta)\frac{1-B(x)}{1-B(x-\Delta)} + p_{D-1,1}(x-\Delta)\lambda_1 \Delta \frac{1-B(x)}{1-B(x-\Delta)} + o(\Delta), \tag{16}$$

$$p_{n2}(x) = p_{n2}(x-\Delta)\frac{1-B(x)}{1-B(x-\Delta)} + o(\Delta), \quad n = \overline{H+1, R-1}. \tag{17}$$

Начнём с решения уравнения (16). Используя подстановку (9), получаем уравнение

$$r'_{R2}(x) = \lambda_1 r_{R-1,1}(x),$$

решение которого имеет вид

$$r_{R2}(x) = \tilde{c}_R + \lambda_1 \int_0^x r_{R-1,1}(y)\, dy,$$

где $\tilde{c}_R$ — некоторая постоянная, причём $p_{R2}(0) = \tilde{c}_R$. Процесс $X(t)$ никогда не попадает в состояние $(R, 0, 2)$ и, значит, $p_{R2}(0) = 0$, откуда следует, что $\tilde{c}_R = 0$ и

$$r_{R2}(x) = \lambda_1 \int_0^x r_{R-1,1}(y)\, dy.$$

Для решения уравнения (17) используем в (17) подстановку (9). Получаем уравнение

$$r'_{n2}(x) = 0, \quad n = \overline{H+1, R-1},$$

откуда немедленно следует, что $r_{n2}(x) = \tilde{c}_n$, $n = \overline{H+1, R-1}$, где $\tilde{c}_n$ — некоторая постоянная. Поскольку постоянные $\tilde{c}_n$ находятся с помощью того же самого приёма, что и раньше, сразу же запишем

$$p_{n2}(0) = \tilde{c}_n = \lambda_1 \int_0^\infty [1 - \alpha_{H+1}(x)] p_{H1}(x)\, dx, \quad n = \overline{H+1, R-1}.$$

5. ПОКАЗАТЕЛИ ПРОИЗВОДИТЕЛЬНОСТИ

Выпишем выражения для некоторых характеристик производительности системы. Стационарная вероятность немедленного обслуживания (обслуживания без ожидания) равна

$$p_0 = \left(1 + \sum_{n,s} \int_0^\infty p_{ns}(x)\, dx \right)^{-1}.$$

Стационарные вероятности π_1 потери заявок первого и π_2 второго типа имеют вид

$$\pi_1 = \sum_{n=H+1}^{R} \int_0^\infty p_{n2}(x)\,dx, \quad \pi_2 = \pi_1 + \sum_{n=L}^{R-1} \int_0^\infty p_{n1}(x)\,dx \ .$$

Стационарная средняя длина очереди в системе определяется формулой

$$Q = \sum_{n=1}^{H-1}(n-1)P_{n0} + \sum_{n=L}^{R-1}(n-1)P_{n1} + \sum_{n=H+1}^{R}(n-1)P_{n2}, \quad P_{ns} = \int_0^\infty p_{ns}(x)\,dx \ ,$$

а стационарное среднее число заявок в системе равно $N = Q + 1 - p_0$.

Обслуженная нагрузка определяется выражением

$$\lambda^* = (1-\pi_1)\lambda_1 + (1-\pi_2)\lambda_2,$$

а стационарное среднее время ожидания начала обслуживания заявки равно

$$V = \frac{Q}{\lambda^*}.$$

6. ЧИСЛЕННЫЕ РАСЧЁТЫ

Приведем несколько результатов численных расчётов, выполненных на основе полученных аналитических соотношений. Эти соотношения были преобразованы и доведены до вида, упрощающего численные расчёты. Сами преобразования здесь не приводятся в силу их громоздкости.

В качестве функций распределения времени обслуживания заявки были выбраны следующие функции:

1) $B_1(x) = 1 - e^{-x}, \quad x > 0;$
2) $B_2(x) = 1 - e^{-2x} - 2xe^{-2x}, \quad x > 0;$
3) $B_3(x) = 1 - \frac{2}{3}e^{-2x} - \frac{1}{3}e^{-\frac{x}{2}}, \quad x > 0.$

Как видно, средние значения у этих функции совпадают (и равны 1), тогда как коэффициенты вариации удовлетворяют неравенству $c_2 = 1/\sqrt{2} < c_1 = 1 < c_3 = \sqrt{3/2}$. Здесь c_i — коэффициент вариации для функции $B_i(x)$.

Все расчёты производились для значений порогов $L = 12, \ H = 18, \ R = 30$ и двух различных значений интенсивности суммарного входящего потока $\lambda = 2$ и $\lambda = 3$. Таким образом, нагрузка на систему в первом случае равнялась 2, а во втором — 3.

Наиболее наглядное представление результатов получается, если воспользоваться следующим приемом [25]. Пусть дана общая интенсивность входящего в систему потока λ. Выберем число $q \in (0,1)$ и положим $\lambda_1 = (1-q)\lambda$ и $\lambda_2 = q\lambda$. Тогда q — вероятность сброса поступающей заявки в случае, когда система находится в режиме перегрузки.

Ниже на рис. 2 представлена зависимость стационарных вероятностей потерь π_1 и π_2 от вероятности сброса q для каждой из функций распределения $B_i(x)$.

На рис. 3 представлена зависимость стационарного среднего числа заявок в системе N от вероятности сброса q для каждой из функций распределения $B_i(x)$.

На рис. 4 для случая $\lambda = 2$ представлены стационарные распределения общего числа заявок в системе для каждой из функций распределения $B_i(x)$ в зависимости от вероятности сброса q.

Не останавливаясь на подробном анализе полученных результатов, отметим только, что для рассмотренных примеров вычисленные характеристики относительно мало зависят от выбранных распределений.

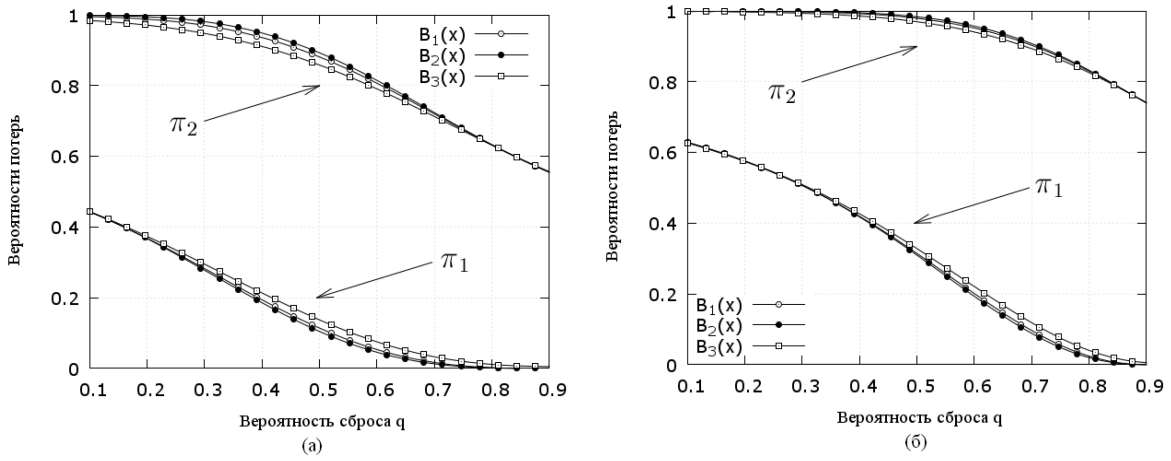

Рис. 2. Зависимость вероятностей потерь π_1 и π_2 от вероятности сброса q. Случай (а): $\lambda = 2$. Случай (б): $\lambda = 3$.

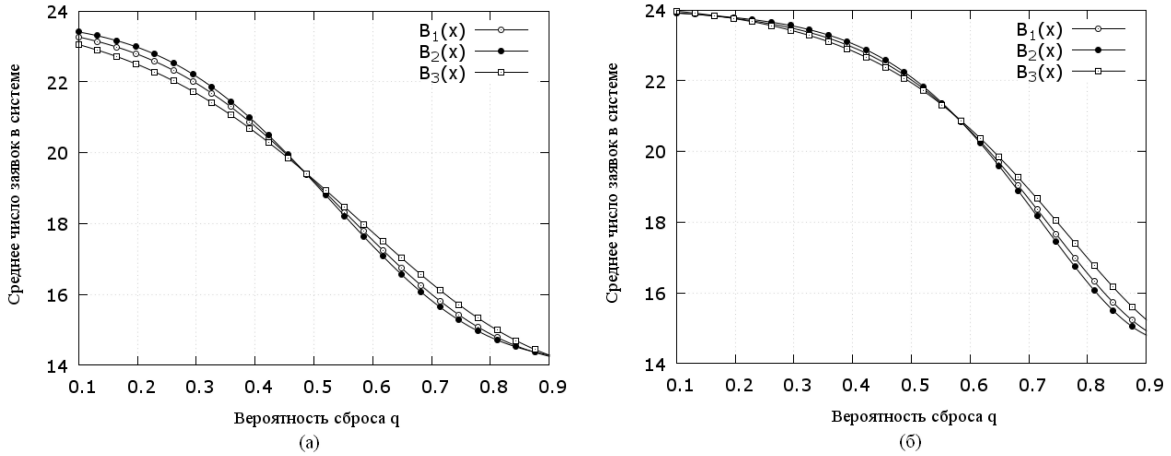

Рис. 3. Зависимость среднего числа заявок в системе от вероятности сброса q. Случай (а): $\lambda = 2$. Случай (б): $\lambda = 3$.

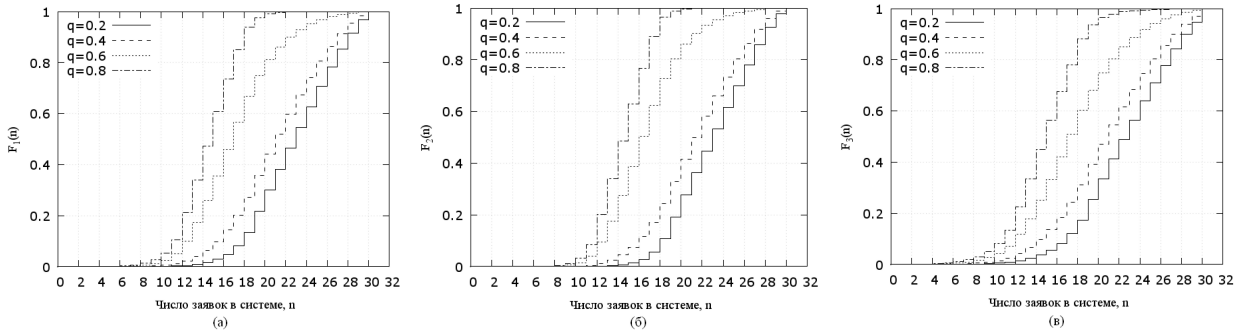

Рис. 4. Распределение числа заявок в системе в зависимости от распределения времени обслуживания при $\lambda = 2$. Случай (а): $F_1(n)$. Случай (б): $F_2(n)$. Случай (в): $F_3(n)$.

Все полученные аналитические результаты были проверены путём сравнения с результатами работы имитационной модели, написанной на языке GPSS.

7. ЗАКЛЮЧЕНИЕ

Таким образом, в настоящей статье получены соотношения, позволяющие производить эффективный расчёт стационарного распределения марковского процесса, описывающего функционирование СМО конечной ёмкости с одним прибором, пуассоновским входящим потоком, произвольным распределением времени обслуживания и гистерезисной политикой управления интенсивностью входящего потока. Представлены результаты тестовых расчётов.

Полученные соотношения могут быть непосредственно применены для решения оптимизационных задач, связанных с гистерезисным механизмом управления.

СПИСОК ЛИТЕРАТУРЫ

1. Абаев П.О., Гайдамака Ю.В., Самуйлов К.Е. Гистерезисное управление сигнальной нагрузкой в сети SIP-серверов. *Вестник Российского университета дружбы народов. Математика. Информатика. Физика*, 2011, № 4, стр. 54–71.

2. Dshalalow J.H. Queueing systems with state dependent parameters. In: *Frontiers in Queueing: Models and Applications in Science and Engineering*, 1997, pp. 61–116.

3. Kitaev M.Yu., Rykov V.V. *Controlled Queueing Systems*. New York: CRC-Press, 1995.

4. Bekker R., Boxma O.J. An M/G/1 queue with adaptable service speed. *Stochastic Models*, 2007, vol. 23, issue 3, pp. 373–396.

5. Chydzinski A. The oscillating queue with finite buffer. *Performance Evaluation*, 2004, vol. 57, no. 3, pp. 341–355.

6. Chydzinski A. The M/G-G/1 Oscillating Queueing System. *Queueing Systems*, 2002, vol. 42, issue 3, pp. 255–268.

7. Горцев А.М. Система массового обслуживания с произвольным числом резервных каналов и гистерезисным управлением включением и выключением резервных каналов. *Автоматика и телемеханика*, 1977, № 10, стр. 30–37.

8. Dudin A. Optimal control for an $M^x|G|1$ queue with two operation modes. *Probability in the Engineering and Informational Sciences*, 1997, vol. 11, no. 2, pp. 255–265.

9. Жерновый К.Ю., Жерновый Ю. В. Система $M^\theta|G|1|m$ с двухпороговой гистерезисной стратегией переключения интенсивности обслуживания. *Информационные процессы*, 2012, том 12, № 2, стр. 127–140.

10. Nishimura S., Jiang Y. An $M|G|1$ vacation model with two service modes. *Probability in the Engineering and Informational Sciences*, 1995, vol. 9, № 3, pp. 355–374.

11. Жерновый К.Ю., Жерновый Ю. В. Система $M^\theta|G|1$ гистерезисным переключением интенсивности обслуживания. *Информационные процессы*, 2012, том 12, № 3, стр. 176–190.

12. Choi B.D., Choi D.I. The queueing system with queue length dependent service times and its application to cell discarding scheme in ATM networks. *IEE Proceedings in Communication*, 1996, vol. 143, pp. 5–11.

13. Сегхайер А., Цитович И.И. Об интервальной модели для процесса рождения и гибели с гистерезисом. *Информационные процессы*, 2012, том 12, № 1, стр. 117–126.

14. Gyemin L., Jongwoo J. Analysis of an $MMPP|G|1|K$ finite queue with two-level threshold overload control. *Communications of the Korean Mathematical Society*, 1999, vol. 14, no. 4, pp. 805–814.

15. Doo Il Choi. Analysis of a queueing system with overload control by arrival rates. *Journal of Applied Mathematics and Computing*, 2005, vol. 18, no. 1–2, pp. 455–464.

16. Van Houdt B. Analysis of the adaptive $MMAP(K)|PH(K)|1$ queue: A multi-type queue with adaptive arrivals and general impatience. *European Journal of Operational Research*, 2012, vol. 220, no. 3, pp. 695–704.

17. Усар И., Макушенко И. Гистерезисная стратегия для системы с повторными вызовами. *Материалы международной научной конференции "Современные вероятностными методы анализа и оптимизации инфомрационно-телекоммункиационных сетей"*. Минск: РИВШ, 2011, стр. 253–257.

18. Bekker R. Queues with Levy input and hysteretic control. *Queueing Systems*, 2009, vol. 63, issue 1, pp. 281–299.

19. Милованова Т.А., Печинкин А.В. Стационарные характеристики системы обслуживания с инверсионным порядком обслуживания, вероятностным приоритетом и гистерезисной политикой. *Информатика и ее применения*, 2013, том 7, вып. 1, стр. 26–38.

20. Gaidamaka Yu., Samouylov K., Sopin E. Analysis of M/G/1 queue with hysteretic load control. *Proc. XXX International Seminar on Stability Problems for Stochastic Models and VI International Workshop Applied Problems in Theory of Probabilities and Mathematical Statistics Related to Modeling of Information Systems.* Moscow: Institute of Informatics Problems, RAS, 2012, pp. 87–89.

21. Sopin E. Analysis of M|G|1|r Queue with Batch Arrival and Hysteretic Overload Control. *Вестник Российского университета дружбы народов*, 2013, № 2, pp. 38–44.

22. Abaev P., Gaidamaka Yu., Samouylov K. Queuing Model for Loss-Based Overload Control in a SIP Server Using a Hysteretic Technique. In: *Lecture Notes in Computer Science*, Heidelberg, Springer-Verlag, 2012, vol. 7469, pp. 371–378.

23. Abaev P., Gaidamaka Yu., Samouylov K. Modeling of Hysteretic Signaling Load Control in Next Generation Networks. In: *Lecture Notes in Computer Science*, Heidelberg, Springer-Verlag, 2012, vol. 7469, pp. 440–452.

24. Bocharov P., D'Apice C., Pechinkin A., Salerno S. *Queueing Theory.* Utrecht: VSP Publishing, 2003.

25. Abaev P., Gaidamaka Yu., Pechinkin A., Razumchik R., Shorgin S. Simulation of overload control in SIP server networks. In: *Proc. of the 26th European Conference on Modelling and Simulation*, 2012, pp. 533–539.

26. Abaev P., Pechinkin A., Razumchik R. Analysis of queueing system with constant service time for sip server hop-by-hop overload control. *Lecture Notes in Communications in Computer and Information Science*, 2013, pp. 1–10.

27. Abaev P., Pechinkin A., Razumchik R. On analytical model for optimal sip server hop-by-hop overload control. *Proc. of the 4th International Congress on Ultra Modern Telecommunications and Control Systems*, 2012, pp. 303–308.

28. Pechinkin A., Razumchik R. Approach for analysis of finite $M_2|M_2|1|R$ with hysteric policy for sip server hop-by-hop overload control. *Proc. of the 27th European Conference on Modelling and Simulation*, 2013, pp. 573–579.

29. Abaev P., Razumchik R. Queuing Model for SIP Server Hysteretic Overload Control with Bursty Traffic. *Proc. of the 13th International Conference on Next Generation Wired/Wireless Networking*, 2013, (accepted).

Stationary characteristics of $M_2|G|1|r$ queue with hysteric control policy of arrival rate

Pechinkin A.V., Razumchik R.V.

Consideration is given to finite $M_2|G|1|r$ queue with bi-level hysteretic input load control. Considered control mechanism implies that system may be in three states (normal, overloaded, blocking). Whenever number of customer in the system changes, decision is made whether input flow rate should be adjusted of not. New approach that allows fast computation of joint stationary probability distribution of the system's state, total number of customers in the system and elapsed service time is proposed. Numerical example that illustrates the applicability of obtained results is given.

KEYWORDS: queueing system, hysteresis, thresholds, arrival rate control.