Информационные процессы, Том 15, № 3, 2015, стр. 324–336

МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

Время пребывания в различных режимах системы обслуживания с неординарными пуассоновскими входящими потоками, рекуррентным обслуживанием и гистерезисной политикой[1]

А.В. Печинкин, Р.В. Разумчик

Федеральный исследовательский центр "Информатика и управление" Российской академии наук, Москва, Россия

Поступила в редколлегию 01.09.2015

Аннотация—Рассматривается модель SIP-сервера в виде однолинейной системы массового обслуживания конечной емкости R с двумя неординарными пуассоновскими потоками, рекуррентным обслуживанием и двухпороговой гистерезисной политикой упавления интенсивностью входящего потока, которая определяется двумя числами L и H ($0 < L < H < R$). Согласно рассматриваемой гистерезисной политике система в каждый момент времени может функционировать в одном из трех режимов: нормальном режиме, режиме перегрузки и режиме блокировки. В нормальном режиме система принимает заявки обоих потоков, в режиме перегрузки – заявки только от одного из потоков, а в режиме блокировки не принимает новых заявок. Предполагается, что переключение режимов работы системы может происходить только в моменты окончания обслуживания заявки на приборе. В работе предложен аналитический метод нахождения распределений (в терминах преобразований Лапласа-Стилтьеса) времен пребывания системы в каждом режиме функционирования, а также времени возвращения системы в режим нормальной работы. Приводятся некоторые результаты численных расчетов.

КЛЮЧЕВЫЕ СЛОВА: перегрузка сервера, система массового обслуживания, групповое поступление, гистерезисное управление нагрузкой.

1. ВВЕДЕНИЕ

Согласно [1], системы массового обслуживания (СМО) с гистерезисным управлением могут служить адекватными моделями для оценки качества функционирования SIP-серверов с пороговым управлением, функционирующих в условиях перегрузок. Достаточно подробное описание применения подобного типа СМО к моделированию SIP-серверов, работающих в условиях перегрузки, можно найти, например, в [2]. Вообще изучению СМО с различными видами гистерезисного управления посвящего достаточно много работ, а также написано несколько работ обзорного характера, из которых можно отметить [1], [3]–[12]. В настоящей работе анализируется СМО с гистерезисной политикой управления только параметрами входящих в СМО потоков. Гистерезисное управление подразумевает наличие нескольких "гистерезисных петель", соответствующих различным уровням принимаемой в систему нагрузки. В качестве примера рассмотрим двухпороговый гистерезисный механизм. Он действует следующим образом. Имеется два числа L и H, для которых справедливы неравенства $0 < L < H < R$.

[1] Работа выполнена при поддержке Российского фонда фундаментальных исследований (гранты 13-07-00223, 15-07-03007)

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. Заметим, что из приведенного описания следует, что переключение режима функционирования системы осуществляется в моменты изменения числа заявок в системе. Однако переключение может быть и в другие момент – моменты окончания обслуживания заявки на приборе. Оба этих правила переключения имеют свои преимущества и недостатки и выбор того или другого диктуется конкретными практическими приложениями.

В работе [13] авторами предложена СМО $M|G|1|R$ с групповым поступлением заявок и двухпороговой гистерезисной политикой в качестве модели SIP-сервера со встроенным механизмом управления интенсивностью входящего потока. Для этой модели при двух изложенных выше правилах переключения режима работы системы (либо в момент изменения числа заявок в системе, либо в моменты окончания обслуживания) предложены математические соотношения, позволяющие вычислять совместное стационарное распределение числа заявок в системе, состояния системы и прошедшего времени обслуживания заявки на приборе.

Заметим, что с точки зрения показателей качества функционирования интерес представляют и стационарные временные характеристики системы, а именно время перехода случайного процесса, описывающего функционирование системы, из множества состояний перегрузки и сброса нагрузки в множество состояний нормальной нагрузки. Для нахождения оптимальных значений параметров гистерезисной политики характеристики этой случайной величины такие, как её математическое ожидание или 95% квантиль, подлежат минимизации при заданных ограничениях на нагрузочные и структурные параметры системы. В настоящей работе предлагается аналитический метод нахождения основных стационарных временных характеристик рассмотренной в [13] СМО. Стоит отметить, что, в отличие от [13], полученные результаты справедливы лишь в том случае, когда смена режима работы системы происходит в момент окончания обслуживания заявки на приборе. Основными полученными результатами являются распределения (в терминах преобразований Лапласа-Стилтьеса) времен первого выхода из каждого режима системы, а также распределения времени возвращения системы в режим нормальной нагрузки.

Статья организована следующим образом. В следующем разделе приводится подробное описание системы с описание гистерезисной политики. В разделе 3 показано, как можно вычислять распределение времен первого выхода системы из каждого режима функционирования (нормального, перегрузки, блокировки). Раздел 4 посвящен нахождению распределения времени возвращения системы в режим нормальной загрузки. В последних двух разделах приводятся некоторые результаты численных расчетов и дается краткое обсуждение полученных результатов.

2. ОПИСАНИЕ СИСТЕМЫ

Рассмотрим однолинейную систему массового обслуживания с функцией распределения $B(x)$ времени обслуживания (длины) заявки и гистерезисной стратегией обслуживания за-

явок. Через $\beta(u) = \int\limits_0^\infty e^{-ux} dB(x)$ обозначим преобразование Лапласа–Стилтьеса (ПЛС) длины заявки, а через $b = \int\limits_0^\infty x\, dB(x)$ — среднюю длину заявки.

Опишем функционирование этой СМО.

В систему поступают независимые неординарные пуассоновские потоки заявок двух типов, причём λ_k, $k = 1, 2$, — интенсивность потока k-го типа. Через $\lambda_0 = \lambda_1 + \lambda_2$ будем обозначать суммарную интенсивность этих потоков. Вероятность того, что в поступающей группе заявок потока любого типа будет n, $n \geq 1$, заявок равна ω_k. Положим $\Omega_n = \sum\limits_{m=n}^{\infty} \omega_m$, $n \geq 1$, — вероятность того, что в поступающей группе заявок будет не менее n заявок.

Гистерезисная политика обслуживания определяется следующим образом. Система может работать в трех режимах: в нормальном режиме (режиме 0); в режиме перегрузки (режиме 1); в режиме сброса нагрузки (режиме 2). Выбор режима происходит в момент окончания или начала обслуживания заявки на приборе и определяется числами L, H и R, для которых справедливы неравенства $0 < L < H < R$.

Если в поступающей в свободную систему группе будет менее H заявок, то далее вплоть до окончания обслуживания заявки на приборе в систему будут приниматься все заявки (система будет работать в нормальном режиме — режиме 0). Если в этой группе будет от H до $(R-1)$ заявок, то до окончания обслуживания заявки на приборе в систему будут приниматься только заявки первого типа (система будет работать в режиме перегрузки, или режиме 1). Если же будет не менее R заявок, то прекратится прием заявок обоих типов и заявки будут только обслуживаются (режим сброса нагрузки, или режим 2), причем из поступившей группы в системе останется ровно R заявок.

Далее, если в момент окончания обслуживания заявки на приборе система работала в режиме 0 и сразу же после этого момента в системе оказалось менее H заявок, то система продолжит работу в режиме 0 до следующего момента освобождения прибора, если оказалось от H до $(R-1)$ заявок, то система переходит в режим 1 и, наконец, если оказалось R заявок, то система переходит в режим 2. В последнем случае, если за время обслуживания успело прийти более R заявок, то остается только R заявок, а лишние покидают систему.

Следующий случай: система работала в режиме 1 и в ней находилось от L до $(R-1)$ заявки. Тогда если останется $(L-1)$ заявка, то система переходит в режим 0, если окажется не менее R заявок, — в режим 2, причем, как и прежде, лишние заявки теряются, а в остальных случаях — продолжает работать в режиме 1.

Последний случай: система работала в режиме 2 (заявки любого типа не принимались) и в ней находилось от $(H+1)$ до R заявок. В этом случае, если останется H заявок, то система переходит в режим 1, иначе продолжает работать в прежнем режиме.

Будем предполагать, что выполнено условие $b < \infty$, необходимое и достаточное для существования стационарного режима функционирования рассматриваемой системы.

Будем считать также, что $H - L \geq 1$ 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 умоляет общности полученных результатов.

3. РАСПРЕДЕЛЕНИЕ ВРЕМЕН ПЕРВОГО ВЫХОДА

Будем искать распределения времен первого выхода системы из каждого режима функционирования в терминах ПЛС. Введем следующие обозначения:

– $V_n^0(u)$, $n = \overline{0, H-1}$, — ПЛС времени до того момента, когда система впервые выйдет из режима 0, при условии, что в начальный момент либо система была свободна ($n = 0$), либо в системе было n, $n = \overline{1, H-1}$, заявок и началось обслуживание заявки на приборе;
– $V_n^1(u)$, $n = \overline{L, R-1}$, — ПЛС времени до того момента, когда система впервые перейдет в режим 0, при условии, что в начальный момент в системе было n заявок, система работала в режиме 1 (принимались только заявки первого типа) и началось обслуживание заявки на приборе;
– $V_n^2(u)$, $n = \overline{H+1, R}$, — ПЛС времени до того момента, когда в система впервые перейдет в режим 0, при условии, что в начальный момент в системе было n заявок, система работала в режиме 2 (не принимались заявки любого типа) и началось обслуживание заявки на приборе.

В этом разделе ограничимся вычислением только ПЛС $V_n^0(u)$, $V_n^1(u)$ и $V_n^2(u)$, однако заметим, что полученные формулы пригодны и для нахождения соответствующих распределений путем численного обращения ПЛС с помощью известных методов.

Определим вспомогательные функции, которые понадобятся в дальнейшем. Обозначим через $\beta_k^s(u)$, $s = 0, 1$, $k \geq 0$, ПЛС времени обслуживания одной заявки и вероятность того, что за это время в режиме s в систему поступит k групп заявок, т. е.

$$\beta_k^s(u) = \int_0^\infty e^{-(\lambda_s+u)x} \frac{(\lambda_s x)^k}{k!}\, dB(x) = \frac{\lambda_s^k}{k!}\, \beta^{(k)}(\lambda_s + u), \quad s = 0, 1, \quad k \geq 0,$$

где через $\beta^{(k)}(u)$ обозначена k-я производная ПЛС $\beta(u)$, причем $\beta^{(0)}(u) = \beta(u)$. Кроме того, введем ω_i^k, $k \geq 0$, $i \geq 0$, — вероятность того, что в k группах поступит ровно i заявок. Очевидно, что распределение ω_i^k является k-кратной сверткой распределения ω_i и может быть рассчитано с помощью следующей рекуррентной формулы:

$$\omega_i^0 = \delta_i, \quad i \geq 0, \qquad \omega_i^k = \sum_{n=0}^{i} \omega_{i-n}^{k-1} \omega_n, \quad k \geq 1, \quad i \geq 0,$$

где δ_i — символ Кронекера. Наконец, обозначим через $\alpha_i^s(u)$, $s = 0, 1$, $i \geq 0$, ПЛС времени обслуживания одной заявки и вероятность того, что за это время в режиме s в систему поступит i заявок, через $A_i^s(u)$, $s = 0, 1$, $i \geq 0$, — не менее i заявок. Тогда $\alpha_i^s(u)$ и $A_i^s(u)$ выражаются через $\beta_k^s(u)$ и ω_i^k следующим образом:

$$\alpha_i^s(u) = \sum_{k=0}^{i} \beta_k^s(u) \omega_i^k, \quad A_i^s(u) = \sum_{k=i}^{\infty} \alpha_k^s(u), \quad s = 0, 1, \quad i \geq 0.$$

Для сокращения записи положим также

$$\alpha_i = \alpha_i^0(0), \quad A_i = A_i^0(0), \quad i \geq 0.$$

Величины α_i и A_i имеют простую вероятностную интерпретацию: α_i, $i \geq 0$, есть вероятность того, что за время обслуживания одной заявки в режиме 0 в систему поступит i заявок, а A_i, $i \geq 0$, — не менее i заявок.

3.1. Вычисление ПЛС $V_n^0(u)$

Для нахождения $V_n^0(u)$, $n = \overline{0, H-1}$, введем вектор

$$\vec{V}(u)^T = (V_0^0(u), V_1^0(u), V_2^0(u), \ldots, V_{H-1}^0(u))$$

размерности H. Воспользовавшись формулой полной вероятности и свойствами ПЛС, можно убедиться, что для $\vec{V}(u)$ справедливо уравнение

$$\vec{V}(u) = P(u)\vec{V}(u) + \vec{Q}(u), \tag{1}$$

где квадратная матрица

$$P(u) = \begin{pmatrix} p_{0,0}(u) & p_{0,1}(u) & p_{0,2}(u) & \dots & p_{0,H-1}(u) \\ p_{1,0}(u) & p_{1,1}(u) & p_{1,2}(u) & \dots & p_{1,H-1}(u) \\ p_{2,0}(u) & p_{2,1}(u) & p_{2,2}(u) & \dots & p_{2,H-1}(u) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{H-1,0}(u) & p_{H-1,1}(u) & p_{H-1,2}(u) & \dots & p_{H-1,H-1}(u) \end{pmatrix}$$

порядка H определяется выражением

$$P(u) = \begin{pmatrix} 0 & \frac{\lambda_0}{u+\lambda_0}\omega_1 & \frac{\lambda_0}{u+\lambda_0}\omega_2 & \dots & \frac{\lambda_0}{u+\lambda_0}\omega_{H-1} \\ \alpha_0^0(u) & \alpha_1^0(u) & \alpha_2^0(u) & \dots & \alpha_{H-1}^0(u) \\ 0 & \alpha_0^0(u) & \alpha_1^0(u) & \dots & \alpha_{H-2}^0(u) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \alpha_1^0(u) \end{pmatrix},$$

а вектор $\vec{Q}(u)$ размерности H имеет вид

$$\vec{Q}(u)^T = (q_0(u), q_1(u), q_2(u), \dots, q_{H-1}(u)) = \left(\frac{\lambda_0}{u+\lambda_0}\Omega_H, A_H^0(u), A_{H-1}^0(u), \dots, A_2^0(u)\right).$$

Алгоритм решения системы (1) заключается в следующем. Сначала из последнего уравнения этой системы находится значение $V_{H-1}^0(u)$:

$$V_{H-1}^0(u) = \frac{p_{H-1,H-2}(u)V_{H-2}^0(u) + q_{H-1}(u)}{1 - p_{H-1,H-1}(u)}. \tag{2}$$

Далее, вводя новый вектор

$$\vec{V}^*(u)^T = (V_0^0(u), V_1^0(u), V_2^0(u), \cdots, V_{H-2}^0(u))$$

размерности $(H-1)$ и подставляя в остальные уравнения этой системы вместо $V_{H-1}^0(u)$ его значение по формуле (2), получаем новое уравнение

$$\vec{V}^*(u) = P^*(u)\vec{V}^*(u) + \vec{Q}^*(u), \tag{3}$$

где коэффициенты квадратной матрицы

$$P^*(u) = \begin{pmatrix} p_{0,0}^*(u) & p_{0,1}^*(u) & p_{0,2}^*(u) & \dots & p_{0,H-2}^*(u) \\ p_{1,0}^*(u) & p_{1,1}^*(u) & p_{1,2}^*(u) & \dots & p_{1,H-2}^*(u) \\ p_{2,0}^*(u) & p_{2,1}^*(u) & p_{2,2}^*(u) & \dots & p_{2,H-2}^*(u) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{H-2,0}^*(u) & p_{H-2,1}^*(u) & p_{H-2,2}^*(u) & \dots & p_{H-2,H-2}^*(u) \end{pmatrix}$$

порядка $(H-1)$ определяются выражениями

$$p_{i,j}^*(u) = p_{i,j}(u), \quad i = \overline{0, H-2}, \quad j = \overline{0, H-3},$$

$$p^*_{i,H-2}(u) = p_{i,H-2}(u) + \frac{p_{i,H-1}(u)p_{H-1,H-2}(u)}{1 - p_{H-1,H-1}(u)}, \quad i = \overline{0, H-2},$$

а координаты вектора

$$\vec{Q}^*(u)^T = (q^*_0(u), q^*_1(u), q^*_2(u), \ldots, q^*_{H-2}(u))$$

размерности $(H-1)$ — выражением

$$q^*_i(u) = q_i(u) + \frac{p_{i,H-1}(u)q_{H-1}(u)}{1 - p_{H-1,H-1}(u)}, \quad i = \overline{0, H-2}.$$

Таким образом, приходим к точно такой же системе (3) линейных алгебраических уравнений, как и исходная система (1), но размерности на единицу меньше. Продолжая описанную процедуру, в конечном счете получаем уравнение

$$V^0_0(u) = \tilde{p}_{0,0}(u)V^0_0(u) + \tilde{q}_0(u),$$

из которого находится $V^0_0(u)$ по формуле

$$V^0_0(u) = \frac{\tilde{q}_0(u)}{1 - \tilde{p}_{0,0}(u)}.$$

Остальные функции $V^0_n(u)$, $n = \overline{1, H-1}$ вычисляются последовательно по n от $n = 1$ до $n = H-1$ из аналогов формулы (2).

3.2. Вычисление ПЛС $V^1_n(u)$ и $V^2_n(u)$

Прежде, чем перейти к нахождению ПЛС $V^1_n(u)$ и $V^2_n(u)$ решим сначала вспомогательную задачу. Предположим, что в системе находится $n \geq H+1$ заявок и начинается обслуживание заявки. Вычислим время (в терминах ПЛС) до того момента, когда в системе впервые окажется H заявок.

Пусть в начальный момент в системе находится n, $n = \overline{H+1, R-1}$, заявок и система начинает обслуживать заявку в режиме 1. Обозначим через $w^1_n(u)$, $n = \overline{H+1, R-1}$, ПЛС времени до того момента, когда в системе впервые окажется H заявок, и вероятность того, что до этого момента система будет работать только в режиме 1, а через $W^1_n(u)$, $n = \overline{H+1, R-1}$, ПЛС времени до того момента, когда система перейдет в режим 2, и вероятность того, что до этого момента в системе не будет менее $(H+1)$ заявок.

Введем векторы

$$\vec{w}(u)^T = (w^1_{H+1}(u), w^1_{H+2}(u), w^1_{H+3}(u), \ldots, w^1_{R-1}(u)),$$

$$\vec{W}(u)^T = (W^1_{H+1}(u), W^1_{H+2}(u), W^1_{H+3}(u), \cdots, W^1_{R-1}(u))$$

размерности $(R-H-1)$. Анализируя возможные изменения числа заявок в системе за один шаг (длительности, равной времени обслуживания заявки), а затем воспользовавшись формулой полной вероятности и свойствами ПЛС, получаем, что для $\vec{w}(u)$ и $\vec{W}(u)$ справедливы уравнения

$$\vec{w}(u) = P(u)\vec{w}(u) + \vec{q}(u), \tag{4}$$

$$\vec{W}(u) = P(u)\vec{W}(u) + \vec{Q}(u), \tag{5}$$

где квадратная матрица

$$P(u) = \begin{pmatrix} p_{H+1,H+1}(u)\, p_{H+1,H+2}(u)\, p_{H+1,H+3}(u) \ldots p_{H+1,R-1}(u) \\ p_{H+2,H+1}(u)\, p_{H+2,H+2}(u)\, p_{H+2,H+3}(u) \ldots p_{H+2,R-1}(u) \\ p_{H+3,H+1}(u)\, p_{H+3,H+2}(u)\, p_{H+3,H+3}(u) \ldots p_{H+3,R-1}(u) \\ \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \ddots \qquad\quad \vdots \\ p_{R-1,H+1}(u)\, p_{R-1,H+2}(u)\, p_{R-1,H+3}(u) \ldots p_{R-1,R-1}(u) \end{pmatrix}$$

порядка $(R-H-1)$ определяется выражением

$$P(u) = \begin{pmatrix} \alpha_1^1(u)\, \alpha_2^1(u)\, \alpha_3^1(u) \ldots \alpha_{R-H-1}^1(u) \\ \alpha_0^1(u)\, \alpha_1^1(u)\, \alpha_2^1(u) \ldots \alpha_{R-H-2}^1(u) \\ 0 \quad \alpha_0^1(u)\, \alpha_1^1(u) \ldots \alpha_{R-H-3}^1(u) \\ \vdots \quad\ \vdots \quad\ \vdots \quad \ddots \quad\ \vdots \\ 0 \quad\ 0 \quad\ 0 \quad \ldots \quad \alpha_1^1(u) \end{pmatrix},$$

а векторы $\vec{q}(u)$ и $\vec{Q}(u)$ размерности $(R-H-1)$ имеют вид

$$\vec{q}(u)^T = (q_{H+1}(u) q_{H+2}(u) \ldots q_{R-1}(u)) = (\alpha_0^1(u) 0 \ldots 0),$$

$$\vec{Q}(u)^T = (Q_{H+1}(u), Q_{H+2}(u), \ldots, Q_{R-1}(u)) = (A_{R-H}^1(u), A_{R-H-1}^1(u), \ldots, A_2^1(u)).$$

Системы (4) и (5) решаются по тому же алгоритму, что и система (1).

Снова предположим, что в начальный момент система начала обслуживать заявку. Обозначим через $v_n^1(u)$, $n = \overline{H+1, R-1}$, ПЛС времени до того момента, когда в системе впервые останется H заявок, при условии, что в начальный момент в системе находилось n, $n = \overline{H+1, R-1}$, заявок и система работала в режиме 1, а через $v_n^2(u)$, $n = \overline{H+1, R}$, — ПЛС времени до того момента, когда в системе впервые останется H заявок (система перейдет в режим 1), при условии, что в начальный момент в системе находилось n, $n = \overline{H+1, R}$, заявок и система работала в режиме 2. Тогда

$$v_n^2(u) = \beta^{n-H}(u), \quad n = \overline{H+1, R}, \tag{6}$$

$$v_n^1(u) = w_n^1(u) + W_n^1(u) v_R^2(u), \quad n = \overline{H+1, R-1}. \tag{7}$$

Теперь все готово для нахождения ПЛС $V_n^1(u)$, $n = \overline{L, H}$, времени до того момента, когда система впервые перейдет в режим 0, при условии, что в начальный момент в системе было n заявок, система работала в режиме 1 и началось обслуживание заявки на приборе.

Введем вектор

$$\vec{V}(u)^T = (V_L^1(u), V_{L+1}^1(u), V_{L+2}^1(u), \ldots, V_H^1(u))$$

размерности $(H-L+1)$. Тогда, воспользовавшись снова формулой полной вероятности и свойствами ПЛС, получаем, что для $\vec{V}(u)$ справедливо уравнение

$$\vec{V}(u) = P(u)\vec{V}(u) + \vec{Q}(u), \tag{8}$$

где квадратная матрица

$$P(u) = \begin{pmatrix} p_{L,L}(u) \quad p_{L,L+1}(u) \quad p_{L,L+2}(u) \quad \ldots \quad p_{L,H}(u) \\ p_{L+1,L}(u)\, p_{L+1,L+1}(u)\, p_{L+1,L+2}(u) \ldots p_{L+1,H}(u) \\ p_{L+2,L}(u)\, p_{L+2,L+1}(u)\, p_{L+2,L+2}(u) \ldots p_{L+2,H}(u) \\ \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \ddots \qquad\quad \vdots \\ p_{H,L}(u) \quad p_{H,L+1}(u) \quad p_{H,L+2}(u) \quad \ldots \quad p_{H,H}(u) \end{pmatrix}$$

порядка $(H-L+1)$ определяется выражением

$$P(u)=\begin{pmatrix}\alpha_1^1(u)\,\alpha_2^1(u)\,\alpha_3^1(u)\ldots & \alpha_{H-L}^1(u) & \tilde\alpha_0(u)\\ \alpha_0^1(u)\,\alpha_1^1(u)\,\alpha_2^1(u)\ldots & \alpha_{H-L-1}^1(u) & \tilde\alpha_1(u)\\ 0\quad \alpha_0^1(u)\,\alpha_1^1(u)\ldots & \alpha_{H-L-2}^1(u) & \tilde\alpha_2(u)\\ \vdots\quad\ \vdots\quad\ \vdots\quad \ddots & \vdots & \vdots\\ 0\quad\ 0\quad\ 0\quad\ldots & \alpha_0^1(u) & \tilde\alpha_{H-L}(u)\end{pmatrix},$$

функции $\tilde\alpha_i(u)$, $i=\overline{0,H-L}$, определяются по формуле

$$\tilde\alpha_i(u)=\alpha_{H-L+1-i}^1(u)+\sum_{j=1}^{R-H-1}\alpha_{H-L+1+j-i}^1(u)v_{H+j}^1(u)+A_{R-L+1-i}^1(u)v_R^2(u), \tag{9}$$

а вектор $\vec Q(u)$ размерности $(H-L+1)$ имеет вид

$$\vec Q(u)^T=(Q_L(u),Q_{L+1}(u),\ldots,Q_{H-L}(u))=(\alpha_0^1(u),0,\ldots,0). \tag{10}$$

Система (8) решается точно таким же образом, что и система (1).

Наконец заметим, что если в системе находится $n\ge H+1$ заявок, то время до того момента, когда система впервые перейдет в режим 0 (т.е. в системе впервые окажется $L-1$ заявок) равно сумме двух времен. Первое – это время, необходимое для того, чтобы число заявок в системе стало равным H (при условии, что было n), а второе – это время, необходимое для того, чтобы число заявок в системе стало равным $(L-1)$ (при условии, что было H). Поэтому

$$V_n^1(u)=v_n^1(u)V_H^1(u),\quad n=H+1,R-1, \tag{11}$$

$$V_n^2(u)=v_n^2(u)V_H^1(u),\quad n=H+1,R. \tag{12}$$

4. РАСПРЕДЕЛЕНИЕ ВРЕМЕН ВОЗВРАЩЕНИЯ

С точки зрения показателей качества обслуживания SIP-сервера интерес представляют времена перехода случайного процесса, описывающего функционирование системы, из множества состояний нормальной нагрузки в множество состояний перегрузки и сброса нагрузки и, наоборот, из множества состояний перегрузки и сброса нагрузки в множество состояний нормальной нагрузки. За первую из этих характеристик в терминах ПЛС можно принять, например $V_{L-1}^0(u)$, но, вообще говоря, то или иное значение $V_n^0(u)$ должно выбираться из практических соображений.

В этом разделе займемся вычислением второй характеристики – времени перехода из множества состояний перегрузки и сброса нагрузки в множество состояний нормальной нагрузки (или, сокращенно, времени возврата), ПЛС которого обозначим через $V^*(u)$.

Основной задачей при вычислении времени возврата является определение вероятностей π_n, $n=\overline{H,R}$, того, что при переходе системы из множества состояний режима 0 в множество состояний режима 1 или в множество состояний режима 2 в системе окажется n заявок (очевидно, что $n=\overline{H,R-1}$ при переходе в множество состояний режима 1 и $n=R$ — в множество состояний режима 2). Найдем эти вероятности. Введем матрицу

$$\pi=\begin{pmatrix}\pi_{0,H} & \pi_{0,H+1} & \pi_{0,H+2} & \cdots & \pi_{0,R}\\ \pi_{1,H} & \pi_{1,H+1} & \pi_{1,H+2} & \cdots & \pi_{1,R}\\ \pi_{2,H} & \pi_{2,H+1} & \pi_{2,H+2} & \cdots & \pi_{2,R}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \pi_{H-1,H} & \pi_{H-1,H+1} & \pi_{H-1,H+2} & \cdots & \pi_{H-1,R}\end{pmatrix} \tag{13}$$

размера $H\times(R-H+1)$. Здесь $\pi_{i,j},\ i=\overline{0,H-1},\ j=\overline{H,R}$, — вероятность того, что при выходе из множества состояний режима 0 в системе окажется j заявок, при условии что в начальный момент было i заявок. Тогда для π справедливо уравнение

$$\pi = P\pi + Q, \tag{14}$$

где квадратная матрица

$$P = \begin{pmatrix} p_{0,0} & p_{0,1} & p_{0,2} & \cdots & p_{0,H-1} \\ p_{1,0} & p_{1,1} & p_{1,2} & \cdots & p_{1,H-1} \\ p_{2,0} & p_{2,1} & p_{2,2} & \cdots & p_{2,H-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{H-1,0} & p_{H-1,1} & p_{H-1,2} & \cdots & p_{H-1,H-1} \end{pmatrix}$$

порядка H, в которой $p_{0,j},\ j=\overline{0,H-1}$, — вероятность того, что в свободную систему поступит группа из j заявок, а $p_{i,j},\ i,j=\overline{0,H-1}$, — вероятность того, что сразу же после окончания обслуживания заявки на приборе в системе окажется j заявок, при условии что в начальный момент было i заявок, определяется выражением

$$P = \begin{pmatrix} 0 & \omega_1 & \omega_2 & \ldots & \omega_{H-1} \\ \alpha_0 & \alpha_1 & \alpha_2 & \ldots & \alpha_{H-1} \\ 0 & \alpha_0 & \alpha_1 & \ldots & \alpha_{H-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \alpha_1 \end{pmatrix}.$$

Наконец, матрица Q размера $H\times(R-H+1)$ имеет вид

$$Q = \begin{pmatrix} q_{0,H} & q_{0,H+1} & q_{0,H+2} & \cdots & q_{0,R} \\ q_{1,H} & q_{1,H+1} & q_{1,H+2} & \cdots & q_{1,R} \\ q_{2,H} & q_{2,H+1} & q_{2,H+2} & \cdots & q_{2,R} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{H-1,H} & q_{H-1,H+1} & q_{H-1,H+2} & \cdots & q_{H-1,R} \end{pmatrix} = \begin{pmatrix} \omega_H & \omega_{H+1} & \omega_{H+2} & \ldots & \Omega_R \\ \alpha_H & \alpha_{H+1} & \alpha_{H+2} & \ldots & A_R \\ \alpha_{H-1} & \alpha_H & \alpha_{H+1} & \ldots & A_{R-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_2 & \alpha_3 & \alpha_4 & \ldots & A_{R-H+2} \end{pmatrix},$$

где также $q_{0,j},\ j=\overline{H,R}$, есть вероятность того, что после поступления группы в свободную систему в системе окажется j заявок, а $q_{i,j},\ i,j=\overline{0,H-1}$, — вероятность того, что сразу же после окончания обслуживания заявки на приборе в системе окажется j заявок, при условии что в начальный момент было i заявок.

Алгоритм решения системы уравнений (14) в идейном плане ничем не отличается от алгоритмов решения систем (1), (4) и (5). Из последних уравнений системы (14) находим

$$\pi_{H-1,i} = \frac{p_{H-1,H-2}\pi_{H-2,i} + q_{H-1,i}}{1 - p_{H-1,H-1}}, \quad i=\overline{H,R}. \tag{15}$$

Теперь, вводя новую матрицу

$$\pi^* = \begin{pmatrix} \pi_{0,H} & \pi_{0,H+1} & \pi_{0,H+2} & \cdots & \pi_{0,R} \\ \pi_{1,H} & \pi_{1,H+1} & \pi_{1,H+2} & \cdots & \pi_{1,R} \\ \pi_{2,H} & \pi_{2,H+1} & \pi_{2,H+2} & \cdots & \pi_{2,R} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \pi_{H-2,H} & \pi_{H-2,H+1} & \pi_{H-2,H+2} & \cdots & \pi_{H-2,R} \end{pmatrix}$$

размера $(H-1)\times(R-H+1)$, после подстановки в (14) выражения для $\pi_{H-1,i}$ по формуле (15), получаем новое уравнение

$$\pi^* = P^*\pi^* + Q^*, \tag{16}$$

где коэффициенты матрицы

$$P^* = \begin{pmatrix} p^*_{0,0} & p^*_{0,1} & p^*_{0,2} & \cdots & p^*_{0,H-2} \\ p^*_{1,0} & p^*_{1,1} & p^*_{1,2} & \cdots & p^*_{1,H-2} \\ p^*_{2,0} & p^*_{2,1} & p^*_{2,2} & \cdots & p^*_{2,H-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p^*_{H-2,0} & p^*_{H-2,1} & p^*_{H-2,2} & \cdots & p^*_{H-2,H-2} \end{pmatrix}$$

порядка $(H-1)$ определяются выражениями

$$p^*_{i,j} = p_{i,j}, \quad i = \overline{0,H-2}, \quad j = \overline{0,H-3}, \tag{17}$$

$$p^*_{i,H-2} = p_{i,H-2} + \frac{p_{i,H-1}p_{H-1,H-2}}{1-p_{H-1,H-1}}, \quad i = \overline{0,H-2}, \tag{18}$$

а коэффициенты матрицы

$$Q^* = \begin{pmatrix} q^*_{0,H} & q^*_{0,H+1} & q^*_{0,H+2} & \cdots & q^*_{0,R} \\ q^*_{1,H} & q^*_{1,H+1} & q^*_{1,H+2} & \cdots & q^*_{1,R} \\ q^*_{2,H} & q^*_{2,H+1} & q^*_{2,H+2} & \cdots & q^*_{2,R} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q^*_{H-2,H} & q^*_{H-2,H+1} & q^*_{H-2,H+2} & \cdots & q^*_{H-2,R} \end{pmatrix} \tag{19}$$

размера $(H-1)\times(R-H+1)$ — выражением

$$q^*_{i,j} = q_{i,j} + \frac{p_{i,H-1}q_{H-1,j}}{1-p_{H-1,H-1}}, \quad i = \overline{0,H-2}, \quad j = \overline{H,R}. \tag{20}$$

Продолжая эту процедуру, приходим к матричному уравнению

$$\tilde\pi = \tilde P\tilde\pi + \tilde Q, \tag{21}$$

где $\tilde\pi = \begin{pmatrix}\pi_{0,H}\ \pi_{0,H+1}\ \pi_{0,H+2}\ldots\pi_{0,R}\end{pmatrix}$ и $\tilde Q = \begin{pmatrix}\tilde q_{0,H}\ \tilde q_{0,H+1}\ \tilde q_{0,H+2}\ldots\tilde q_{0,R}\end{pmatrix}$ — матрицы размера $1\times(R-H+1)$, а $\tilde P = \begin{pmatrix}\tilde p_{0,0}\end{pmatrix}$ — матрица порядка 1, т. е. число. Матричное уравнение (21) фактически является $(R-H+1)$ отдельными уравнениями, которые имеют вид

$$\pi_{0,i} = \tilde p_{0,0}\pi_{0,i} + \tilde q_{0,i}, \quad i = \overline{H,R} \tag{22}$$

из которых находим

$$\pi_{0,i} = \frac{\tilde q_{0,i}}{1-\tilde p_{0,0}}, \quad i = \overline{H,R}. \tag{23}$$

Остальные вероятности $\pi_{n,i}$, $n = \overline{1,H-1}$, $i = \overline{H,R}$, вычисляются последовательно от $n=1$ до $n=H-1$ по аналогам формулы (15).

Теперь можно привести формулу для ПЛС $V^*(u)$ времени возврата. Предположим, что исходное состояние системы, из которого считается время возврата, соответствует тому, что в системе находится n, $n=\overline{0,H-1}$, заявок (и система находится в режиме 0). Тогда

$$V^*_n(u) = \sum_{i=H}^{R-1}\pi_{n,i}V^1_i(u) + \pi_{n,R}V^2_R(u).$$

5. НЕКОТОРЫЕ ЧИСЛЕННЫЕ РЕЗУЛЬТАТЫ

В качестве примера приведем некоторые результаты расчетов моментов распределения времени возврата с ПЛС $V_n^*(u)$ при $n = L-1$. В качестве значений исходных параметров системы выберем те же значения, что и в работе [13], где для рассмотренной системы рассчитывались только вероятностные характеристики очереди. Пусть $L = 12$, $H = 18$, $R = 30$ и $\omega_n = 0.25$ при $n = 1, 2, 3, 4$. В качестве распределения времени обслуживания рассмотрим распределение Эрланга с функцией распределения $B_1(x) = 1-e^{-2x}-2xe^{-2x}$, $x > 0$, и гипер-экспоненциальное распределение с функцией распределения $B_2(x) = 1 - \frac{2}{3}e^{-2x} - \frac{1}{3}e^{-\frac{x}{2}}$, $x > 0$. Таким образом, в обоих случаях среднее время обслуживания равно 1, а соответствующие коэффициенты вариации c_1 и c_2 удовлетворяют неравенству $c_1 = 1/\sqrt{2} < c_2 = \sqrt{3/2}$. На рис. 1 приведены результаты расчёта значения среднего времени возврата (дисперсии времени возврата) при различных интенсивностях потока заявок 1-го типа λ_1.

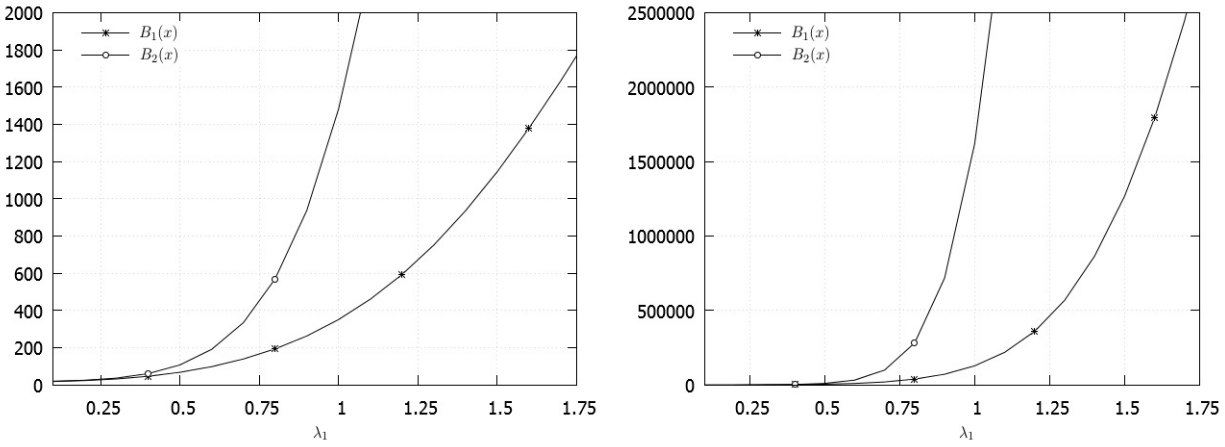

Рис. 1. Зависимость значения среднего времени возврата (слева) и дисперсии времени возврата (справа) от интенсивности потока заявок 1-го типа

Как видно из рисунка, при значениях загрузки в режиме перегрузки не превосходящей 0.5 средние времена возврата при различных распределениях отличаются незначительно, тогда как при росте нагрузки разница, как и ожидалось, становится все более значительно. Численные эксперименты показывают, что при изменении распределения времени обслуживания (при сохранении среднего значения) на другое, близкое к детерминированному (например, при распределения Эрланга с 20-ю фазами и средним временем обслуживания 0.05 на каждой фазе), значения среднего и дисперсии времени возврата в указанных диапазонах изменения интенсивности λ_1 потока заявок 1-го типа заметно увеличиваются (за исключением случаев очень малых значений λ_1).

6. ЗАКЛЮЧЕНИЕ

Заметим, что при численной реализации предложенных методов достаточно каждый раз хранить только одну матрицу из соответствующих элементов, что позволяет производить расчеты для больших значений исходных параметров системы и гистерезисной стратегии (L, H и R). Стоит отметить, что полученные формулы позволяют не только рассчитывать моменты соответствующих случайных величин, но также и находить (численно) их распределения с помощью известных алгоритмов (Gaver-Stehfest, Talbot и др.). Кроме того, используемый в данной работе метод, допускает простое обобщение на семейство гистерезисных петель, рас-

положенных по тому же принципу, что описан в разделе 2. Учитывая, что полученные результаты справедливы в том случае, когда смена режима работы системы происходит в момент окончания обслуживания заявки на приборе, то несомненный интерес представляет также аналогичная задача, но когда переключение режима функционирования системы осуществляется в моменты изменения числа заявок в системе.

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. P. Abaev, Y. Gaidamaka, K. Samouylov, A. Pechinkin, R. Razumchik and S. Shorgin, "Hysteretic control technique for overload problem solution in network of sip servers" in *Computing and Informatics*, 2014, vol. 33, no. 1, pp. 1–18.
3. Dshalalow J.H. Queueing systems with state dependent parameters. In: *Frontiers in Queueing: Models and Applications in Science and Engineering*, 1997, pp. 61–116.
4. Chydzinski A. The oscillating queue with finite buffer. *Performance Evaluation*, 2004, vol. 57, no. 3, pp. 341–355.
5. 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.
6. 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.
7. Жерновый Ю.В, Жерновый К.Ю. Вероятностные характеристики системы $M_2^\theta/G/1/m$ с двухпетельным гистерезисным управлением длительностью обслуживания и интенсивностью входящего потока *Информационные процессы*, 2014. т. 14, № 2, стр. 137–150.
8. Сегхайер А., Цитович И.И. Об интервальной модели для процесса рождения и гибели с гистерезисом. *Информационные процессы*, 2012, т. 12, № 1, стр. 117–126.
9. 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.
10. Милованова Т.А., Печинкин А.В. Стационарные характеристики системы обслуживания с инверсионным порядком обслуживания, вероятностным приоритетом и гистерезисной политикой. *Информатика и ее применения*, 2013, т. 7, вып. 1, стр. 26–38.
11. 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.
12. 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.
13. Y. Gaidamaka, A. Pechinkin, R. Razumchik, K. Samouylov, E. Sopin, "Analysis of an $M|G|1|R$ queue with batch arrivals and two hysteretic overload control policies" in *Applied Mathematics and Computer Science*, 2014. vol. 24, no. 3, pp. 519–534.

First passage times between modes in the queueing system with batch Poisson arrivals, general service and hysteresis policy

Pechinkin A.V., Razumchik R.V.

Consideration is given to the model of SIP-server as a single-line queueing system of finite capacity R with two batch Poisson flows of customers, general service time distribution and bi-level hysteretic load control policy, defined by two natural numbers L and H $(0 < L < H < R)$. Hysteretic control policy implies that at each time instant the system can be in one of the three modes: normal, overload, discard. When in normal mode the system accepts customers from both flow. When overloaded customer from one of the flows are accepted. In discard mode the system rejects all incoming customers. It is assumed that the switching of the mode occurs only on service completion epochs. The analytical method for the calculation of distributions (in terms of Laplace-Stieltjes transform) of the first passage times between modes as well as return times is proposed. Numerical example is presented.

KEYWORDS: overload, batch arrivals, queueing system, hysteretic control.