

# РАСПРЕДЕЛЕНИЕ ВРЕМЕНИ ВЫХОДА ИЗ МНОЖЕСТВА СОСТОЯНИЙ ПЕРЕГРУЗКИ В СИСТЕМЕ $M|M|1|\langle L, H \rangle|\langle H, R \rangle$ С ГИСТЕРЕЗИСНЫМ УПРАВЛЕНИЕМ НАГРУЗКОЙ\*

Ю. В. Гайдамака<sup>1</sup>, А. В. Печинкин<sup>2</sup>, Р. В. Разумчик<sup>3</sup>, А. К. Самуйлов<sup>4</sup>, К. Е. Самуйлов<sup>5</sup>, И. А. Соколов<sup>6</sup>, Э. С. Сопин<sup>7</sup>, С. Я. Шоргин<sup>8</sup>

**Аннотация:** Одним из наиболее простых в реализации и эффективных решений проблемы перегрузок, обеспечивающим наименьшее число переключений режимов функционирования системы, является гистерезисное управление нагрузкой. В статье предложен аналитический метод исследования параметров гистерезисного управления. В качестве математической модели рассмотрена система массового обслуживания (СМО)  $M|M|1|\langle L, H \rangle|\langle H, R \rangle$  с двумя петлями гистерезисного управления, где  $H$  — порог обнаружения перегрузки;  $L$  — порог снижения перегрузки;  $R$  — порог сброса нагрузки. Получены два метода вычисления преобразования Лапласа–Стилтьеса (ПЛС) времени возврата системы из множества состояний перегрузки в множество состояний нормальной нагрузки: первый — путем решения системы уравнений с ПЛС неизвестных времен возврата для каждого состояния перегрузки; второй — с помощью рекуррентного представления ПЛС времен возврата в виде дробно-рациональных функций. Оба метода позволяют при вычислениях эффективно применять инструментальные программные средства общего назначения, что показано на численном примере.

**Ключевые слова:** перегрузка сервера; система массового обслуживания (СМО); гистерезисное управление нагрузкой; время возврата в множество состояний нормальной нагрузки; преобразование Лапласа–Стилтьеса (ПЛС); функция распределения

DOI: 10.14357/19922264130403

## 1 Введение

Проблема защиты жизненно важных узлов телекоммуникационной сети от перегрузок вновь стала критичной в сетях связи последующих поколений (NGN, Next Generation Network). В сетях второго поколения (2G) с коммутацией каналов перегрузки были вызваны, главным образом, поведением пользователей, порождающим взрывной рост трафика в часы наивысшей нагрузки. В современных сетях 3G и 4G основной причиной непредсказуемого по объему трафика, с обработкой которого не справляется самое современное оборудование, стал в первую очередь стремительный рост числа телекоммуникационных услуг, характеризующихся высокими требованиями к производительности сетевых узлов и серверов различного назначения. Примером проявления проблемы являются пере-

грузки серверов протокола SIP (Session Initiation Protocol), порождаемые лавинообразным потоком запросов пользователей на предоставление широкополосных услуг [1]. Так, при установлении соединения по протоколу SIP даже в простейшем случае установления речевого соединения один запрос пользователя требует передачи и обработки несколькими серверами до семи сообщений протокола SIP.

Различные сценарии перегрузок SIP-серверов описаны в документах Рабочей группы по инженерным проблемам сети Интернет (IETF, Internet Engineering Task Force), играющих роль международных стандартов [1–5] для сетей на базе протокола IP. Типичным проявлением перегрузки служит лавинный перезапуск, который происходит, когда слишком много пользователей одновременно пытаются зарегистрироваться на SIP-серверах. Примером служит

\* Работа выполнена при поддержке РФФИ (проекты №№ 11-07-00112 и 12-07-00108).

<sup>1</sup> Российский университет дружбы народов, ygaidamaka@sci.pfu.edu.ru

<sup>2</sup> Институт проблем информатики Российской академии наук, apchinkin@ipiran.ru

<sup>3</sup> Институт проблем информатики Российской академии наук, rrazumchik@ieee.org

<sup>4</sup> Российский университет дружбы народов, asam1988@gmail.com

<sup>5</sup> Российский университет дружбы народов, ksam@sci.pfu.edu.ru

<sup>6</sup> Институт проблем информатики Российской академии наук, isokolov@ipiran.ru

<sup>7</sup> Российский университет дружбы народов, sopin-eduard@yandex.ru

<sup>8</sup> Институт проблем информатики Российской академии наук, sshorgin@ipiran.ru

сценарий так называемого «манхэттенского перезапуска» (англ. Manhattan reboots scenario), когда в результате аварии произошло отключение электричества в этом крупнейшем районе города и после восстановления электроснабжения все SIP-терминалы одновременно пытались зарегистрироваться на серверах, создав тем самым большой поток сообщений REGISTER.

Для успешного управления перегрузками, по сути, требуется ответить на два вопроса: как определить начало перегрузки и как ее устранить. Наиболее естественным решением является введение порогового управления очередью сообщений (заявок) на обработку сервером подобно тому, как это было сделано в рекомендациях Международного союза электросвязи (ITU, International Telecommunications Union) для протоколов сетевого и канального уровней общеканальной системы сигнализации № 7 (ОКС7) [6].

В [7–9] был сделан краткий обзор и анализ механизма гистерезисного управления нагрузкой, применяемого в ОКС7, а также разработана математическая модель локального управления перегрузками в сети SIP-серверов. Управление перегрузками осуществляется путем введения трех порогов в очереди на обработку сигнальных сообщений сервером — порога  $H$  обнаружения перегрузки, порога  $L$  снижения перегрузки и порога  $R$  сброса нагрузки. Пока общее число сообщений в очереди не превышает  $(H - 1)$ , сервер функционирует в режиме нормальной нагрузки. Если длина очереди стала равной  $H$ , система переходит в режим снижения нагрузки и остается в этом режиме до тех пор, пока длина очереди не достигнет значения  $(L - 1)$  или  $(R - 1)$ . При уменьшении длины очереди до значения  $(L - 1)$  система возвращается в режим нормальной нагрузки. Если очередь увеличилась до значения  $R$ , включается режим сброса нагрузки, в котором система находится до тех пор, пока длина очереди не станет равной  $H$ , после чего система возвращается в режим снижения нагрузки. Значения порогов выбираются так, что  $0 < L < H < R$ , и поэтому между парами порогов  $\langle L, H \rangle$  и  $\langle H, R \rangle$  возникает так называемый эффект гистерезиса [10, 11] в виде двух петель — по одной на каждую пару порогов. В ОКС7 гистерезисное управление было введено для сокращения числа переключений системы управления из режима перегрузки в режим нормальной нагрузки [12, 13], при этом задача решалась путем выбора значений порогов с целью минимизации среднего времени возврата системы из режима перегрузки в режим нормальной нагрузки.

Исследованию СМО с гистерезисным управлением посвящено достаточно много работ, при-

чем чаще всего встречаются работы по системам с гистерезисным обслуживанием [14–16] и реже встречаются работы по системам с гистерезисным управлением входящим потоком заявок [17, 18]. Объемный обзор результатов по гистерезисному управлению содержится в работах [19, 20], а наиболее близки к исследованиям настоящей статьи модель и методы исследования, разработанные в [21], где также можно найти обширный список источников по проблеме анализа СМО с гистерезисным управлением интенсивностью входящего потока заявок (далее для краткости — с гистерезисным управлением нагрузкой). В работах [7–9] с участием части авторов данной статьи проведен обзор работ, посвященных математическому и имитационному моделированию систем с гистерезисным управлением нагрузкой. В этих же статьях, а также в [22–32] были построены и исследованы марковские модели в виде СМО с пуассоновским входящим потоком и экспоненциально распределенной длительностью обслуживания заявок.

С точки зрения показателей качества обслуживания SIP-сервера интерес представляет время перехода случайного процесса, описывающего функционирование системы, из множества состояний перегрузки и сброса нагрузки в множество состояний нормальной нагрузки. Эту случайную величину принято называть временем возврата в режим нормальной нагрузки, или для краткости — временем возврата, а ее характеристики, такие как математическое ожидание или 95%-ная квантиль, подлежат минимизации при заданных ограничениях на нагрузочные и структурные параметры системы. В [32] для марковского случая получен алгоритм расчета среднего времени возврата и численно решена задача его минимизации для близких к реальным исходных данных. В данной статье, в отличие от известных результатов, предложен метод расчета ПЛС времени возврата, что позволяет вычислять функцию распределения (ФР) и находить квантили этой величины, не прибегая к средствам имитационного моделирования. Кроме того, результаты промежуточных вычислений позволяют эффективно проводить анализ и других важных вероятностно-временных характеристик исследуемой СМО.

Для решения задачи были разработаны два метода. Первый метод основан на решении системы уравнений, где неизвестными являются ПЛС времени возврата из каждого состояния множеств перегрузки и сброса нагрузки. Второй метод основан на рекуррентном представлении ПЛС времени возврата через дробно-рациональные функции. В обоих случаях получены алгоритмы вычисления ПЛС времени возврата в таком виде, что они могут быть достаточно просто запрограммированы, например, с

помощью систем символьных вычислений, встроенных в известные инструментальные программные средства, такие как MATLAB [33], Mathematica [34], Maple [35] и т. п.

Статья организована следующим образом. В разд. 2 детально описана исследуемая модель СМО, введены все необходимые понятия и обозначения. В разд. 3 предложены два метода вычисления ПЛС времени возврата. В разд. 4 представлен пример численного анализа, а в заключении подведены итоги статьи и обсуждаются полученные результаты.

## 2 Постановка задачи

Приведем детальное описание СМО, которая будет рассматриваться далее. Система представляет собой однолинейную СМО с двумя типами заявок. Входящие потоки заявок являются независимыми пуассоновскими с интенсивностью  $\lambda_k, k = 1, 2$ , поступления заявок  $k$ -го типа. Через  $\lambda = \lambda_1 + \lambda_2$  обозначим суммарную интенсивность входящего потока. Время обслуживания заявки любого типа распределено по экспоненциальному закону с параметром  $\mu$ .

Процесс обслуживания происходит следующим образом. Пусть в начальный момент в системе отсутствуют заявки. Тогда до того момента, когда в системе впервые окажется  $H$  заявок, к обслуживанию принимаются заявки обоих типов. Но как только число заявок становится равным  $H$ , прекращается прием заявок второго типа и принимаются только заявки первого типа. Так продолжается до того момента, когда число заявок в системе не становится равным  $L - 1$  или  $R$ . В первом случае снова начинается прием заявок второго типа, а во втором — прекращается прием всех заявок (в том числе и первого типа), причем прием заявок первого типа

возобновляется в тот момент, когда число заявок в системе становится равным  $H$ . Схематическое изображение функционирования рассматриваемой СМО приведено на рис. 1.

Введем обозначения:

$\tilde{V}_n(s), n = \overline{0, H-1}$ , — ПЛС времени до того момента, когда в системе впервые окажется  $H$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки любого типа;

$V_n^*(s), n = \overline{H+1, R}$ , — ПЛС времени до того момента, когда в системе впервые останется  $(L-1)$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и не принимались заявки (любого типа);

$V_n(s), n = \overline{L, R-1}$ , — ПЛС времени до того момента, когда в системе впервые останется  $(L-1)$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа.

Заметим, что  $V_H(s)$  представляет собой не что иное, как выраженное в терминах ПЛС распределение времени возврата (в режим нормальной нагрузки) — одного из основных используемых на практике показателей качества обслуживания SIP-сервера, о чем уже говорилось во введении. Далее,  $\tilde{V}_{L-1}(s)$  — ПЛС времени от момента попадания системы в состояние с  $(L-1)$  заявками до момента первого после этого попадания в состояние с  $H$  заявками, или времени пребывания функционирующего в стационарном режиме SIP-сервера в множестве состояний нормальной нагрузки. И, наконец,  $T(s) = \tilde{V}_{L-1}(s)V_H(s)$  — ПЛС времени между соседними попаданиями системы в состояние с  $(L-1)$  заявками с обязательным попаданием в состояние с  $H$  заявками, или цикла функционирования SIP-сервера, т. е. суммарное время пребывания

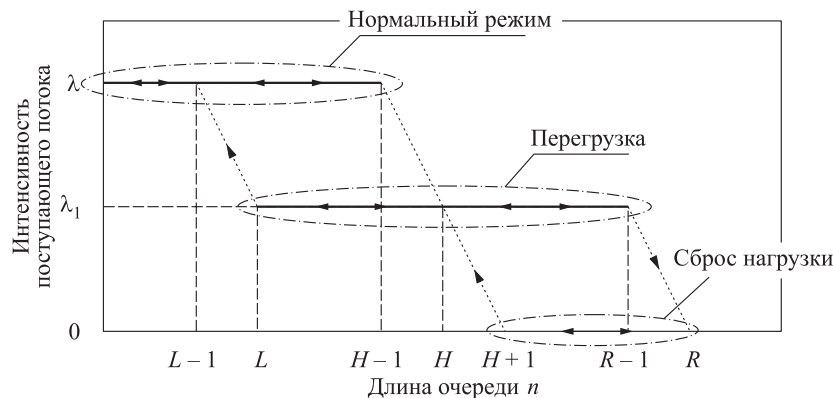


Рис. 1 Схематическое изображение функционирования системы

SIP-сервера в множествах состояний нормальной нагрузки и перегрузки.

Основным результатом настоящей статьи является решение следующих двух задач:

- (1) найти распределение времени до того момента, когда в системе впервые останется  $H$  заявок, при условии, что в начальный момент в системе было  $n$ ,  $n = \overline{0, H-1}$ , заявок и принимались заявки любого типа, т. е. в терминах ПЛС найти  $\tilde{V}_n(s)$ ,  $n = \overline{0, H-1}$ ;
- (2) найти распределение времени до того момента, когда в системе впервые останется  $(L-1)$  заявок, при условии, что в начальный момент в системе было  $n$ ,  $n = \overline{H+1, R}$ , заявок и не принимались заявки (любого типа), или при условии, что было  $n$  заявок и принимались заявки только первого типа, т. е. в терминах ПЛС найти  $V_n^*(s)$ ,  $n = \overline{H+1, R}$ , и  $V_n(s)$ ,  $n = \overline{L, R-1}$ .

Очевидно, дифференцируя  $\tilde{V}_n(s)$ ,  $V_n^*(s)$  и  $V_n(s)$  нужное число раз в точке  $s = 0$ , можно получить моменты любого порядка.

Найденное в терминах ПЛС распределение времени возврата используется далее для численных расчетов, причем предлагается два варианта формул. Первый вариант предпочтительно употреблять в том случае, когда обращение ПЛС производится по точкам. Второй вариант, дающий ответ в виде дробно-рациональной функции, удобен тогда, когда имеются хорошие программы разложения дробно-рациональной функции на сумму простейших дробей.

Отметим, что используемый в работе метод, более подробное изложение которого можно найти в [36], элементарно переносится на решение задачи нахождения ПЛС ФР времени перехода из любого состояния в любое другое. Кроме того, заметим, что  $\tilde{V}_n(s)$  представляет собой не что иное, как ПЛС времени до первой потери заявки в системе  $M|M|1|(H-2)$ , исходя из состояния  $n$ , и хорошо известно (см., например, результат для более общего случая в [37]). Здесь выводы этого ПЛС приводятся для лучшего понимания дальнейших выкладок.

### 3 Вычисление преобразования Лапласа–Стилтьеса времени выхода

#### 3.1 Вариант 1

Вычислим сначала ПЛС  $\tilde{V}_n(s)$ ,  $n = \overline{1, H-1}$ . Эти ПЛС удовлетворяют уравнению:

$$\begin{aligned} \tilde{V}_n(s) &= \\ &= \frac{\mu + \lambda}{s + \mu + \lambda} \left( \frac{\mu}{\mu + \lambda} \tilde{V}_{n-1}(s) + \frac{\lambda}{\mu + \lambda} \tilde{V}_{n+1}(s) \right) = \\ &= \frac{1}{s + \mu + \lambda} \left[ \mu \tilde{V}_{n-1}(s) + \lambda \tilde{V}_{n+1}(s) \right], \\ & \quad n = \overline{1, H-1}. \end{aligned} \quad (1)$$

Граничные условия для системы уравнений (1) определяются формулами:

$$\tilde{V}_0(s) = \frac{\lambda}{s + \lambda} \tilde{V}_1(s); \quad (2)$$

$$\tilde{V}_H(s) = 1. \quad (3)$$

Решение системы уравнений (1) имеет вид:

$$\tilde{V}_n(s) = \tilde{c}_1 \tilde{z}_1^{H-n}(s) + \tilde{c}_2 \tilde{z}_2^{H-n}(s), \quad n = \overline{0, H},$$

где  $\tilde{z}_1 = \tilde{z}_1(s)$  и  $\tilde{z}_2 = \tilde{z}_2(s)$  — решения уравнения

$$\lambda - (s + \mu + \lambda)z + \mu z^2 = 0,$$

т. е.

$$z_{1,2} = \frac{s + \mu + \lambda \pm \sqrt{(s + \mu + \lambda)^2 - 4\lambda\mu}}{2\mu},$$

а  $\tilde{c}_1 = \tilde{c}_1(s)$  и  $\tilde{c}_2 = \tilde{c}_2(s)$  определяются из граничных условий (2) и (3). Из (3) находим

$$\tilde{c}_2 = 1 - \tilde{c}_1$$

и, значит,

$$\begin{aligned} \tilde{V}_n(s) &= \tilde{c}_1 \tilde{z}_1^{H-n}(s) - (1 - \tilde{c}_1) \tilde{z}_2^{H-n}(s), \\ & \quad n = \overline{0, H-1}. \end{aligned} \quad (4)$$

Подставляя (4) в (2), получаем после элементарных преобразований

$$\begin{aligned} \tilde{c}_1 &= \left\{ [(s + \lambda)\tilde{z}_2(s) - \lambda]\tilde{z}_2^{H-1}(s) \right\} / \left\{ [(s + \lambda)\tilde{z}_1(s) - \right. \\ & \quad \left. - \lambda]\tilde{z}_1^{H-1}(s) + [(s + \lambda)\tilde{z}_2(s) - \lambda]\tilde{z}_2^{H-1}(s) \right\}. \end{aligned}$$

Для нахождения  $V_n^*(s)$ ,  $n = \overline{H+1, R}$ , и  $V_n(s)$ ,  $n = \overline{L, R-1}$ , введем сначала следующие вспомогательные функции:

$W_n^*(s)$ ,  $n = \overline{H+1, R}$ , — ПЛС времени до того момента, когда в системе впервые останется  $H$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и не принимались заявки (любого типа);

$w_n(s)$ ,  $n = \overline{H+1, R-1}$ , — ПЛС времени до того момента, когда в системе впервые останется  $H$  заявок, и вероятность того, что до этого момента в системе не было  $R$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа;

$\bar{w}_n(s)$ ,  $n = \overline{H+1, R-1}$ , — ПЛС времени до того момента, когда в системе впервые окажется  $R$  заявок, и вероятность того, что до этого момента в системе не было  $H$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа.

Очевидно,

$$W_n^*(s) = \left(\frac{\mu}{s+\mu}\right)^{n-H}, \quad n = \overline{H+1, R}.$$

Функции  $w_n(s)$ ,  $n = \overline{H+1, R-1}$ , удовлетворяют уравнению

$$w_n(s) = \frac{\mu + \lambda_1}{s + \mu + \lambda_1} \left( \frac{\mu}{\mu + \lambda_1} w_{n-1}(s) + \frac{\lambda_1}{\mu + \lambda_1} w_{n+1}(s) \right) = \frac{1}{s + \mu + \lambda_1} [\mu w_{n-1}(s) + \lambda_1 w_{n+1}(s)], \quad n = \overline{H+1, R-1}, \quad (5)$$

с граничными условиями

$$w_H(s) = 1; \quad (6)$$

$$w_R(s) = 0. \quad (7)$$

Решение системы уравнений (5) имеет вид:

$$w_n(s) = c_1 z_1^{R-n}(s) + c_2 z_2^{R-n}(s), \quad n = \overline{H, R},$$

где  $z_1, z_2$  — решения уравнения

$$\lambda_1 - (s + \mu + \lambda_1)z + \mu z^2 = 0,$$

т. е.

$$z_{1,2} = \frac{s + \mu + \lambda_1 \pm \sqrt{(s + \mu + \lambda_1)^2 - 4\lambda_1\mu}}{2\mu}.$$

Из граничного условия (7) имеем

$$c_1 = -c_2 = c$$

и, следовательно,

$$w_n(s) = c [z_1^{R-n}(s) - z_2^{R-n}(s)], \quad n = \overline{H, R}.$$

Наконец, из граничного условия (6) получаем

$$c [z_1^{R-H}(s) - z_2^{R-H}(s)] = 1,$$

т. е.

$$c = \frac{1}{z_1^{R-H}(s) - z_2^{R-H}(s)}.$$

Далее, для ПЛС  $\bar{w}_n(s)$ ,  $n = \overline{H+1, R-1}$ , справедливо уравнение

$$\begin{aligned} \bar{w}_n(s) &= \frac{\mu + \lambda_1}{s + \mu + \lambda_1} \left( \frac{\mu}{\mu + \lambda_1} \bar{w}_{n-1}(s) + \frac{\lambda_1}{\mu + \lambda_1} \bar{w}_{n+1}(s) \right) = \\ &= \frac{1}{s + \mu + \lambda_1} [\mu \bar{w}_{n-1}(s) + \lambda_1 \bar{w}_{n+1}(s)], \quad n = \overline{H+1, R-1}, \quad (8) \end{aligned}$$

с граничными условиями

$$\bar{w}_H(s) = 0; \quad (9)$$

$$\bar{w}_R(s) = 1. \quad (10)$$

Решение системы уравнений (8) имеет вид:

$$\bar{w}_n(s) = \bar{c}_1 \bar{z}_1^{n-H}(s) + \bar{c}_2 \bar{z}_2^{n-H}(s), \quad n = \overline{H, R},$$

где  $\bar{z}_1, \bar{z}_2$  — решения уравнения

$$\mu - (s + \mu + \lambda_1)\bar{z} + \lambda_1\bar{z}^2 = 0,$$

т. е.

$$\begin{aligned} \bar{z}_{1,2} &= \frac{s + \mu + \lambda_1 \pm \sqrt{(s + \mu + \lambda_1)^2 - 4\lambda_1\mu}}{2\lambda_1} = \\ &= \frac{1}{z_{2,1}}. \quad (11) \end{aligned}$$

Из граничных условий (9) и (10) имеем

$$\bar{c}_1 = -\bar{c}_2 = \bar{c}$$

и

$$\bar{c} = \frac{1}{z_1^{R-H}(s) - z_2^{R-H}(s)}.$$

Введем теперь  $W_n(s)$ ,  $n = \overline{H+1, R-1}$ , — ПЛС времени до того момента, когда в системе впервые останется  $H$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа. Тогда

$$W_n(s) = w_n(s) + \bar{w}_n(s)W_R^*(s), \quad n = \overline{H+1, R-1}.$$

Вычислим  $V_n(s)$  при  $n = \overline{L, H}$ . Напомним, что  $V_n(s)$ ,  $n = \overline{L, H}$ , — ПЛС времени до того момента, когда в системе впервые останется  $(L-1)$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа. Преобразование Лапласа–Стилтьеса  $V_n(s)$  при  $n = \overline{L, H-1}$  удовлетворяет уравнению

$$\begin{aligned} V_n(s) &= \frac{\mu + \lambda_1}{s + \mu + \lambda_1} \left( \frac{\mu}{\mu + \lambda_1} V_{n-1}(s) + \frac{\lambda_1}{\mu + \lambda_1} V_{n+1}(s) \right) = \frac{1}{s + \mu + \lambda_1} [\mu V_{n-1}(s) + \\ &+ \lambda_1 V_{n+1}(s)], \quad n = \overline{L, H-1}, \quad (12) \end{aligned}$$

с граничными условиями

$$V_{L-1}(s) = 1; \quad (13)$$

$$V_H(s) = \frac{\mu + \lambda_1}{\mu + \lambda_1} \cdot \left( \frac{\mu}{\mu + \lambda_1} V_{H-1}(s) + \frac{\lambda_1}{\mu + \lambda_1} W_{H+1}(s) V_H(s) \right),$$

второе из которых перепишем в виде

$$[s + \mu + \lambda_1 - \lambda_1 W_{H+1}(s)] V_H(s) = \mu V_{H-1}(s). \quad (14)$$

Решение системы уравнений (12) имеет вид:

$$V_n(s) = \bar{c}_1 \bar{z}_1^{n-L+1}(s) + \bar{c}_2 \bar{z}_2^{n-L+1}(s), \quad n = \overline{L-1, H},$$

где  $\bar{z}_1, \bar{z}_2$  определяются формулой (11). Из граничных условий (13) и (14) имеем

$$\bar{c}_2 = 1 - \bar{c}_1,$$

$$\bar{c}_1 = \left( \mu \bar{z}_2^{H-L}(s) - [s + \mu + \lambda_1 - \lambda_1 W_{H+1}(s)] \bar{z}_2^{H-L+1}(s) \right) / \left( [s + \mu + \lambda_1 - \lambda_1 W_{H+1}(s)] [\bar{z}_1^{H-L+1}(s) - \bar{z}_2^{H-L+1}(s)] - \mu [\bar{z}_1^{H-L}(s) - \bar{z}_2^{H-L}(s)] \right).$$

Теперь можно привести окончательные выражения для ПЛС  $V_n^*(s)$  при  $n = \overline{H+1, R}$  и ПЛС  $V_n(s)$  при  $n = \overline{H+1, R-1}$ :

$$V_n^*(s) = W_n^*(s) V_H(s), \quad n = \overline{H+1, R};$$

$$V_n(s) = W_n(s) V_H(s), \quad n = \overline{H+1, R-1}.$$

### 3.2 Вариант 2

В этом подразделе будут найдены выражения для тех же самых ПЛС  $\tilde{V}_n(s)$ ,  $n = \overline{0, H-1}$ , ПЛС  $V_n^*(s)$ ,  $n = \overline{H+1, R}$ , и ПЛС  $V_n(s)$ ,  $n = \overline{L, R-1}$ , что и в подразд. 3.1, но только в виде дробно-рациональных функций.

Как и прежде, начнем с  $\tilde{V}_n(s)$ .

Обозначим через  $\tilde{w}_n(s)$ ,  $n = \overline{0, H-1}$ , ПЛС времени до того момента, когда в системе впервые окажется  $(n+1)$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались все заявки. Тогда для  $\tilde{w}_n(s)$ ,  $n = \overline{0, H-1}$ , справедливы соотношения:

$$\tilde{w}_0(s) = \frac{\lambda}{s + \lambda};$$

$$\tilde{w}_n(s) = \frac{\mu + \lambda}{s + \mu + \lambda} \left[ \frac{\lambda}{\mu + \lambda} + \frac{\mu}{\mu + \lambda} \tilde{w}_{n-1}(s) \tilde{w}_n(s) \right] = \frac{1}{s + \mu + \lambda} [\lambda + \mu \tilde{w}_{n-1}(s) \tilde{w}_n(s)],$$

$$n = \overline{1, H-1}. \quad (15)$$

Из уравнения (15) получаем

$$\tilde{w}_n(s) = \frac{\lambda}{s + \mu + \lambda - \mu \tilde{w}_{n-1}(s)}, \quad n = \overline{1, H-1}.$$

Запишем ПЛС  $\tilde{w}_n(s)$ ,  $n = \overline{0, H-1}$ , в виде дробно-рациональной функции

$$\tilde{w}_n(s) = \frac{\tilde{q}_n(s)}{\tilde{r}_n(s)}, \quad n = \overline{0, H-1}, \quad (16)$$

где  $\tilde{q}_n(s)$ ,  $n = \overline{0, H-1}$ , — полином  $n$ -й степени, а  $\tilde{r}_n(s)$ ,  $n = \overline{0, H-1}$ , — полином  $(n+1)$ -й степени. Полиномы  $\tilde{q}_n(s)$  и  $\tilde{r}_n(s)$  можно последовательно вычислить из следующих соотношений:

$$\tilde{q}_0(s) = \lambda; \quad \tilde{r}_0(s) = s + \lambda;$$

$$\tilde{q}_n(s) = \lambda \tilde{r}_{n-1}(s), \quad n = \overline{1, H-1}; \quad (17)$$

$$\tilde{r}_n(s) = (s + \mu + \lambda) \tilde{r}_{n-1}(s) - \mu \tilde{q}_{n-1}(s), \quad n = \overline{1, H-1}.$$

Теперь можно привести формулу для  $\tilde{V}_n(s)$ ,  $n = \overline{0, H-1}$ :

$$\tilde{V}_n(s) = \prod_{i=n}^{H-1} \tilde{w}_i(s), \quad n = \overline{0, H-1}. \quad (18)$$

Подставляя в (18) вместо ПЛС  $\tilde{w}_i(s)$  его значение по формуле (16), а вместо ПЛС  $\tilde{q}_i(s)$  — его значение по формуле (17), окончательно получаем

$$\tilde{V}_n(s) = \frac{\lambda^{H-n-1} \tilde{q}_n(s)}{\tilde{r}_{H-1}(s)}, \quad n = \overline{0, H-1}.$$

Перейдем к нахождению  $V_n^*(s)$ ,  $n = \overline{H+1, R}$ , и  $V_n(s)$ ,  $n = \overline{L, R-1}$ . Введем следующие вспомогательные функции:

$w_n(s)$ ,  $n = \overline{H+1, R-1}$ , — ПЛС времени до того момента, когда в системе впервые останется  $(n-1)$  заявок, и вероятность того, что до этого момента в системе не было  $R$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа;

$\bar{w}_n(s)$ ,  $n = \overline{H+1, R-1}$ , — ПЛС времени до того момента, когда в системе впервые окажется  $(n+1)$  заявок, и вероятность того, что до этого момента в системе не было  $H$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа;

$W_n^*(s)$ ,  $n = \overline{H+1, R}$ , — ПЛС времени до того момента, когда в системе впервые останется  $H$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и не принимались заявки (любого типа);

$\tilde{W}_n(s)$ ,  $n = \overline{H+1, R-1}$ , — ПЛС времени до того момента, когда в системе впервые останется  $H$  заявок, и вероятность того, что до этого момента в системе не было  $R$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа;

$\overline{W}_n(s)$ ,  $n = \overline{H+1, R-1}$ , — ПЛС времени до того момента, когда в системе впервые окажется  $R$  заявок, и вероятность того, что до этого момента в системе не было  $H$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа;

$W_n(s)$ ,  $n = \overline{H+1, R-1}$ , — ПЛС времени до того момента, когда в системе впервые останется  $H$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа.

Преобразование Лапласа–Стилтьеса  $w_n(s)$ ,  $n = \overline{H+1, R-1}$ , удовлетворяет соотношению

$$w_n(s) = \frac{\mu + \lambda_1}{s + \mu + \lambda_1} \left( \frac{\mu}{\mu + \lambda_1} + \frac{\lambda_1}{\mu + \lambda_1} w_{n+1}(s) w_n(s) \right), \quad n = \overline{H+1, R-1},$$

из которого имеем

$$w_n(s) = \frac{\mu}{s + \mu + \lambda_1 - \lambda_1 w_{n+1}(s)}, \quad n = \overline{H+1, R-1},$$

где положено

$$w_R(s) = 0.$$

Отсюда получаем следующие соотношения для последовательного вычисления полиномов  $q_n(s)$  и  $r_n(s)$  в представлении  $w_n(s) = q_n(s)/r_n(s)$ :

$$\begin{aligned} q_{R-1}(s) &= \mu; \\ r_{R-1}(s) &= s + \mu + \lambda_1; \\ q_n(s) &= \mu r_{n+1}(s), \quad n = \overline{H+1, R-2}, \\ r_n(s) &= (s + \mu + \lambda_1) r_{n+1}(s) - \lambda_1 q_{n+1}(s), \\ &\quad n = \overline{H+1, R-2}. \end{aligned}$$

Преобразование Лапласа–Стилтьеса  $\overline{w}_n(s)$ ,  $n = \overline{H+1, R-1}$ , удовлетворяет соотношению

$$\overline{w}_n(s) = \frac{\mu + \lambda_1}{s + \mu + \lambda_1} \left( \frac{\lambda_1}{\mu + \lambda_1} + \frac{\mu}{\mu + \lambda_1} \overline{w}_{n-1}(s) \overline{w}_n(s) \right), \quad n = \overline{H+1, R-1},$$

что дает равенство

$$\overline{w}_n(s) = \frac{\lambda_1}{s + \mu + \lambda_1 - \mu \overline{w}_{n-1}(s)}, \quad n = \overline{H+1, R-1},$$

где положено

$$\overline{w}_H(s) = 0.$$

Это приводит к следующим выражениям для полиномов  $\overline{q}_n(s)$  и  $\overline{r}_n(s)$  в представлении  $\overline{w}_n(s) = \overline{q}_n(s)/\overline{r}_n(s)$ :

$$\begin{aligned} \overline{q}_{H+1}(s) &= \lambda_1; \\ \overline{r}_{H+1}(s) &= s + \mu + \lambda_1; \\ \overline{q}_n(s) &= \lambda_1 \overline{r}_{n-1}(s), \quad n = \overline{H+2, R-1}; \\ \overline{r}_n(s) &= (s + \mu + \lambda_1) \overline{r}_{n-1}(s) - \mu \overline{q}_{n-1}(s), \\ &\quad n = \overline{H+2, R-1}. \end{aligned}$$

Далее,

$$W_n^*(s) = \left( \frac{\mu}{s + \mu} \right)^{n-H}, \quad n = \overline{H+1, R};$$

ПЛС  $\tilde{W}_n(s)$ ,  $n = \overline{H+1, R-1}$ , и  $\overline{W}_n(s)$ ,  $n = \overline{H+1, R-1}$ , определяются выражениями:

$$\tilde{W}_n(s) = \prod_{i=H+1}^n w_i(s) = \frac{\mu^{n-H-1} q_{H+1}(s)}{r_n(s)}, \quad n = \overline{H+1, R-1};$$

$$\overline{W}_n(s) = \prod_{i=n}^{R-1} \overline{w}_i(s) = \frac{\lambda_1^{R-n-1} \overline{q}_{R-1}(s)}{\overline{r}_n(s)}, \quad n = \overline{H+1, R-1},$$

а ПЛС  $W_n(s)$ ,  $n = \overline{H+1, R-1}$ , имеет вид:

$$\begin{aligned} W_n(s) &= \tilde{W}_n(s) + \overline{W}_n(s) W_R^*(s) = \\ &= \frac{\mu^{n-H-1} q_{H+1}(s)}{r_n(s)} + \frac{\lambda_1^{R-n-1} \mu^{R-H} \overline{q}_{R-1}(s)}{(s + \mu)^{R-H} \overline{r}_n(s)} = \\ &= \frac{Q_n(s)}{R_n(s)}, \quad n = \overline{H+1, R-1}, \end{aligned}$$

где  $Q_n(s)$  и  $R_n(s)$  находятся по формулам

$$Q_n(s) = (s + \mu)^{R-H} \mu^{n-H-1} q_{H+1}(s) \overline{r}_n(s) + \lambda_1^{R-n-1} \mu^{R-H} \overline{q}_{R-1}(s) r_n(s), \quad n = \overline{H+1, R-1};$$

$$R_n(s) = (s + \mu)^{R-H} r_n(s) \overline{r}_n(s), \quad n = \overline{H+1, R-1}.$$

Обозначим теперь через  $W_n(s)$ ,  $n = \overline{L, H}$ , ПЛС времени до того момента, когда в системе впервые останется  $(n - 1)$  заявок, при условии, что в начальный момент в системе было  $n$  заявок и принимались заявки только первого типа. Тогда справедливо уравнение:

$$W_n(s) = \frac{\mu + \lambda_1}{s + \mu + \lambda_1} \left( \frac{\lambda_1}{\mu + \lambda_1} W_{n+1}(s)W_n(s) + \frac{\mu}{\mu + \lambda_1} \right), \quad n = \overline{L, H},$$

из которого находим

$$W_n(s) = \frac{\mu}{s + \mu + \lambda_1 - \lambda_1 W_{n+1}(s)}, \quad n = \overline{L, H}.$$

Таким образом, имеем следующие выражения для полиномов  $Q_n(s)$  и  $R_n(s)$  в представлении  $W_n(s) = Q_n(s)/R_n(s)$ ,  $n = \overline{L, H}$ :

$$\begin{aligned} Q_n(s) &= \mu R_{n+1}(s), \quad n = \overline{L, H}; \\ R_n(s) &= (s + \mu + \lambda_1)R_{n+1}(s) - \lambda_1 Q_{n+1}(s), \\ & \quad n = \overline{L, H}. \end{aligned}$$

Окончательно получаем:

$$\begin{aligned} V_n(s) &= \prod_{i=L}^n W_i(s) = \frac{\mu^{n-L} Q_n(s)}{R_L(s)}, \quad n = \overline{L, H}; \\ V_n^*(s) &= W_n^*(s) V_H(s), \quad n = \overline{H+1, R}; \\ V_n(s) &= W_n(s) V_H(s), \quad n = \overline{H+1, R-1}. \end{aligned}$$

## 4 Вычислительный алгоритм

Полученные в предыдущем разделе математические соотношения были использованы для численных расчетов, причем в качестве рассчитываемой характеристики была взята наиболее интересная с точки зрения показателей качества обслуживания SIP-сервера ФР времени возврата, в том числе математическое ожидание и 95%-ная квантиль времени возврата.

В этом разделе приведем алгоритм вычисления ПЛС  $V_H(s)$  времени возврата только для варианта 2 (см. подразд. 3.2), поскольку проведенные расчеты показали, что и точность, и трудоемкость обоих вариантов для всех предложенных исходных данных практически совпали. Кроме того, так как расчеты проводились в среде MATLAB с использованием символьных вычислений, то действия над многочленами производились автоматически. Поэтому приводимый ниже алгоритм не содержит формул для определения многочленов дробно-рациональных представлений ПЛС  $V_H(s)$  и промежуточных

рассчитываемых ПЛС. Далее, поскольку алгоритм предназначен для вычисления только характеристик, связанных с временем возврата, здесь не представлены формулы для нахождения других параметров, хотя эти формулы практически не отличаются от приведенных ниже. Наконец, в рамках шагов предложенного алгоритма удобно показать также, как вычисляется математическое ожидание  $T = -V_H'(0)$  времени возврата.

Алгоритм состоит из следующих шагов.

**Шаг 1.** Последовательно по  $n$  от  $n = R - 1$  до  $n = H + 1$  вычисляются ПЛС  $w_n(s)$  и производные  $w_n'(0)$  по формулам:

$$\begin{aligned} w_{R-1}(s) &= \frac{\mu}{s + \mu + \lambda_1}; \\ w_n(s) &= \frac{\mu}{s + \mu + \lambda_1 - \lambda_1 w_{n+1}(s)}, \quad n = \overline{H+1, R-2}; \\ w_{R-1}'(0) &= -\frac{\mu}{(\mu + \lambda_1)^2}; \\ w_n'(0) &= -\frac{\mu[1 - \lambda_1 w_{n+1}'(0)]}{[\mu + \lambda_1 - \lambda_1 w_{n+1}(0)]^2}, \quad n = \overline{H+1, R-2}. \end{aligned}$$

Далее понадобятся только  $w_{H+1}(s)$  и  $w_{H+1}'(0)$ .

**Шаг 2.** Последовательно по  $n$  от  $n = H + 1$  до  $n = R - 1$  вычисляются ПЛС  $\bar{w}_n(s)$  и производные  $\bar{w}_n'(0)$  по формулам:

$$\begin{aligned} \bar{w}_{H+1}(s) &= \frac{\lambda_1}{s + \mu + \lambda_1}; \\ \bar{w}_n(s) &= \frac{\lambda_1}{s + \mu + \lambda_1 - \mu \bar{w}_{n-1}(s)}, \quad n = \overline{H+2, R-1}; \\ \bar{w}_{H+1}'(0) &= -\frac{\lambda_1}{(\mu + \lambda_1)^2}; \\ \bar{w}_n'(0) &= -\frac{\lambda_1[1 - \mu \bar{w}_{n-1}'(0)]}{[\mu + \lambda_1 - \mu \bar{w}_{n-1}(0)]^2}, \quad n = \overline{H+2, R-1}. \end{aligned}$$

**Шаг 3.** Вычисляются ПЛС  $W_R^*(s)$  и производная  $W_R^{*'}(0)$  по формулам:

$$\begin{aligned} W_R^*(s) &= \left( \frac{\mu}{s + \mu} \right)^{R-H}; \\ W_R^{*'}(0) &= -\frac{R-H}{\mu}. \end{aligned}$$

**Шаг 4.** Вычисляются ПЛС  $\bar{W}_{H+1}(s)$  и производная  $\bar{W}_{H+1}'(0)$  по формулам:

$$\begin{aligned} \bar{W}_{H+1}(s) &= \prod_{i=H+1}^{R-1} \bar{w}_i(s); \\ \bar{W}_{H+1}'(0) &= \bar{W}_{H+1}(0) \sum_{i=H+1}^{R-1} \frac{\bar{w}_i'(0)}{\bar{w}_i(0)}. \end{aligned}$$

**Шаг 5.** Вычисляются ПЛС  $W_{H+1}(s)$  и производная  $W'_{H+1}(0)$  по формулам:

$$W_{H+1}(s) = w_{H+1}(s) + \overline{W}_{H+1}(s)W_R^*(s);$$

$$W'_{H+1}(0) = w'_{H+1}(0) + \overline{W}'_{H+1}(0)W_R^*(0) + \overline{W}_{H+1}(0)W_R^{*'}(0).$$

**Шаг 6.** Последовательно по  $n$  от  $n = H$  до  $n = L$  вычисляются ПЛС  $W_n(s)$  и производные  $W'_n(0)$  по формулам:

$$W_n(s) = \frac{\mu}{s + \mu + \lambda_1 - \lambda_1 W_{n+1}(s)}, \quad n = \overline{L, H};$$

$$W'_n(0) = -\frac{\mu[1 - \lambda_1 W'_{n+1}(0)]}{[\mu + \lambda_1 - \lambda_1 W_{n+1}(0)]^2}, \quad n = \overline{L, H}.$$

**Шаг 7.** Вычисляются ПЛС  $V_H(s)$  и производная  $V'_H(0)$  по формулам:

$$V_H(s) = \prod_{i=L}^H W_i(s),$$

$$T = -V'_H(0) = -\sum_{i=L}^H W'_i(0).$$

Дальнейшие шаги, связанные с вычислением плотности распределения времени возврата с помощью обратного преобразования Лапласа, а затем ФР и 95%-ной квантили времени возврата зависят от используемого программного обеспечения. Приведенный выше алгоритм удобно реализовать в среде MATLAB, используя встроенные возможности символьных вычислений, интегрирования и др.

## 5 Пример расчетов

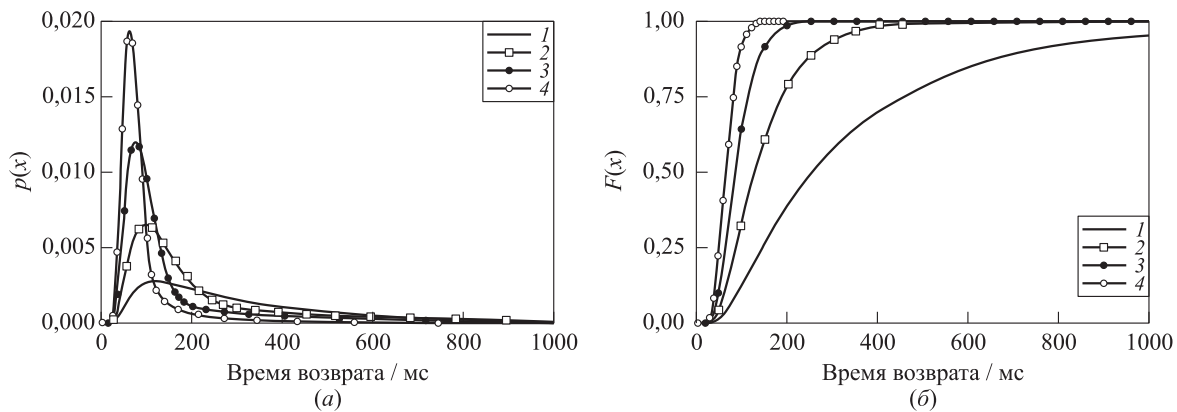
Приведем пример численного анализа характеристик времени возврата для одного из возможных

наборов исходных данных, разработанного авторами данной статьи в рамках решения прикладной задачи по анализу показателей качества обслуживания протокола SIP [8]. Рассматривается случай, когда  $L = 74$ ,  $H = 85$ ,  $R = 100$ , суммарная интенсивность входящего потока  $\lambda = 240$  заявок/с, причем интенсивность  $\lambda_2$  потока заявок второго типа связана с  $\lambda$  следующим образом:  $\lambda_2 = q\lambda$ . Здесь  $q$  — вероятность сброса заявки, т. е. вероятность поступления в систему заявки второго типа. Среднее время обслуживания заявки любого типа  $\mu^{-1} = 5$  мс.

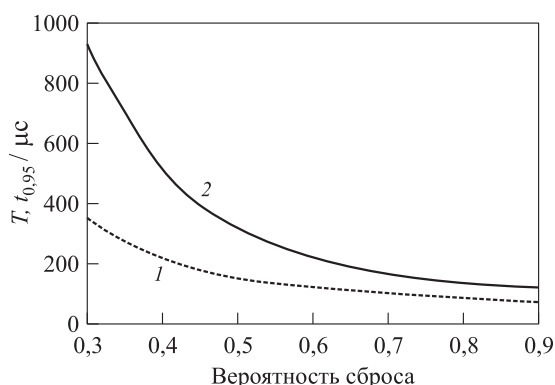
На рис. 2, а приведен график плотности распределения  $p(x)$  времени возврата, построенный для значений  $q = 0,3; 0,5; 0,7; 0,9$ . График ФР  $F(x)$  времени возврата построен на рис. 2, б для тех же самых значений  $q$ . Наконец, на рис. 3 показан график 95%-ной квантили  $t_{0,95}$  и среднего значения  $T$  времени возврата в зависимости от величины вероятности сброса  $q \in [0,3, 0,9]$ .

Из графиков видно, что с ростом вероятности  $q$  сброса заявки время возврата уменьшается, причем его среднее значение  $T$  всегда меньше значения  $t_{0,95}$  95%-ной квантили. В [8] численно решалась задача нахождения значений порогов  $L$  и  $H$  таких, что при фиксированных значениях всех остальных параметров среднее время возврата является минимальным. Полученные в данной статье алгоритмы позволяют решать более сложные задачи оптимизации параметров гистерезисного управления нагрузкой SIP-сервера, в том числе задачу минимизации 95%-ной квантили времени возврата.

При расчетах было отмечено, что наиболее трудоемко вычисление значений 95%-ной квантили, оно может занять до 30 мин машинного времени на компьютере с процессором Intel Core i5 @ 3,30 ГГц и 8 ГБ оперативной памяти. Численный анализ и результаты имитационного моделирования пока-



**Рис. 2** Плотность распределения  $p(x)$  (а) и функция распределения  $F(x)$  (б) времени возврата: 1 —  $q = 0,3$ ; 2 —  $0,5$ ; 3 —  $0,7$ ; 4 —  $q = 0,9$



**Рис. 3** Среднее значение  $T$  времени возврата (1) и 95%-ная квантиль  $t_{0,95}$  (2)

зали достоверность обоих вариантов метода вычислений ПЛС ФР времени возврата для исследуемой СМО.

## 6 Заключение

В статье предложены новые аналитические методы исследования СМО с гистерезисным управлением входящей нагрузкой. По сравнению с известными ранее результатами получен эффективный метод, предназначенный для анализа и расчета не только математического ожидания времени возврата системы из состояний перегрузки в состоянии нормальной нагрузки, но и других его характеристик, в том числе ФР и квантилей. Отметим, что 95%-ную квантиль рекомендовано использовать как показатель качества телекоммуникационных систем международными стандартизирующими организациями, такими как ИТУ и IETF. Разработка программных средств в среде MATLAB и численный анализ на близких к реальным исходных данных показали эффективность разработанных в статье методов, предназначенных для вычисления ПЛС ФР времени возврата, причем время вычисления обоими способами оказалось практически одинаковым, а сложность вычислений такова, что для проведения вычислительного эксперимента достаточно использования персонального компьютера с процессором Intel Core i5 @ 3,30 ГГц и 8 ГБ оперативной памяти. Дальнейшие исследования будут направлены на решение задач нахождения оптимальных пороговых значений для управления перегрузками SIP-серверов.

## Литература

- Hilt V., Noel E., Shen C., Abdelal A. Design considerations for Session Initiation Protocol (SIP) overload control // Internet Engineering Task Force RFC-6357, 2011. [Электронный ресурс] Режим доступа: <http://tools.ietf.org/html/rfc6357>, свободный (дата обращения 01.10.2013).
- Hilt V., Rosenberg J., Schulzrinne H., et al. SIP: Session Initiation Protocol // Internet Engineering Task Force RFC-3261, 2002. [Электронный ресурс] Режим доступа: <http://tools.ietf.org/html/rfc3261>, свободный (дата обращения 01.10.2013).
- Rosenberg J. Requirements for management of overload in the Session Initiation Protocol // Internet Engineering Task Force RFC-5390, 2008. [Электронный ресурс] Режим доступа: <http://tools.ietf.org/html/rfc5390>, свободный (дата обращения 01.10.2013).
- Gurbani V., Hilt V., Schulzrinne H. Session Initiation Protocol overload control // Internet Engineering Task Force Draft, 2013. [Электронный ресурс] Режим доступа: <http://tools.ietf.org/pdf/draft-ietf-soc-overload-control-13.pdf>, свободный (дата обращения 01.10.2013).
- Noel E., Williams P.M. Session Initiation Protocol rate control // Internet Engineering Task Force Draft, 2013. [Электронный ресурс] Режим доступа: <http://tools.ietf.org/pdf/draft-ietf-soc-overload-rate-control-05.pdf>, свободный (дата обращения 01.10.2013).
- ITU-T Recommendation Q.704. 1996. Signalling System No.7 — Message Transfer Part, Signalling network functions and messages. [Электронный ресурс] Режим доступа: <http://www.itu.int/rec/T-REC-Q.704-199607-1/en>, свободный (дата обращения 01.10.2013).
- Абаев П. О., Гайдамака Ю. В., Самуйлов К. Е. Гистерезисное управление сигнальной нагрузкой в сети SIP-серверов // Вестник РУДН. Серия Математика. Информатика. Физика, 2011. № 4. С. 55–73.
- Абаев П. О., Гайдамака Ю. В., Печинкин А. В., Разумчик Р. В., Шоргин С. Я. Simulation of overload control in SIP server networks // 26th Conference (European) on Modelling and Simulation ECMS Proceedings. — Koblenz, 2012. P. 533–539.
- Абаев П., Гайдамака Ю., Самуйлов К. Queuing model for loss-based overload control in a SIP server using a hysteretic technique // Internet of things, smart spaces, and next generation networking / Eds. S. Andreev, S. Balandin, Ye. Koucheryavy. — Lecture notes in computer science ser. — Heidelberg: Springer-Verlag, 2012. Vol. 7469. P. 371–378.
- Gebhart R. F. A queuing process with bilevel hysteretic service-rate control // Nav. Res. Logist. Q., 1967. Vol. 14. P. 55–68.
- Красносельский М. А., Покровский А. В. Системы с гистерезисом. — М.: Наука, 1983. 272 с.
- Yum T., Yen H. Design algorithm for a hysteresis buffer congestion control strategy // IEEE Conference (International) on Communications Proceedings, 1983. P. 499–503.
- Brown P., Chemouil P., Delosme B. A congestion control policy for signalling networks // 7th IeCC Proceedings, 1984. P. 717–724.

14. Golubchik L., Lui J. C. S. Bounding of performance measures for a threshold-based queueing system with hysteresis // *Newsl. ACM SIGMETRICS Performance Evaluation Rev.*, 1997. Vol. 25. No. 1. P. 147–157.
15. Sindal R., Tokekar S. Modeling and analysis of voice/data call admission control scheme in CDMA cellular network for variation in soft handoff threshold parameters // 16th IEEE Conference (International) on Networks (ICON 2008) Proceedings. P. 1–6.
16. Жерновский К. Ю., Жерновский Ю. В. Система  $M^0/G/1$  с гистерезисным переключением интенсивности обслуживания // *Информационные процессы*, 2012. Т. 12. № 3. С. 176–190.
17. Takagi H. Analysis of a finite-capacity  $M/G/1$  queue with a resume level // *Perform. Evaluation*, 1985. Vol. 5. P. 197–203.
18. Benaboud H., Mikou N. Analysis by queueing model of multi-threshold mechanism in ATM switches // 5th IEEE Conference (International) on High Speed Networks and Multimedia Communications (HSNMC 2002) Proceedings. P. 147–151.
19. Dshalalow J. H. Queues with state dependent parameters // *Frontiers in queueing: Models and applications in science and engineering* / Ed. J. H. Dshalalow. — Probability and stochastic ser. — CRC Press, 1997. P. 61–116.
20. Bekker R. Queues with Levy input and hysteretic control // *Queueing Syst.*, 2009. Vol. 63. No. 1. P. 281–299.
21. Roughan M., Pearce C. A martingale analysis of hysteretic overload control // *Adv. Perform. Anal. J. Teletraffic Theory Perform. Anal. Communication Syst. Networks*, 2000. Vol. 3. No. 1. P. 1–30.
22. Abaev P., Gaidamaka Yu., Samouylov K. Modeling of hysteretic signalling load control in next generation networks // *Internet of things, smart spaces, and next generation networking* / Eds. S. Andreev, S. Balandin, Ye. Koucheryavy. — Lecture notes in computer science ser. — Heidelberg: Springer-Verlag, 2012. Vol. 7469. P. 440–452.
23. Абаев П. О., Разумчик Р. В. Моделирование работы SIP-сервера с помощью системы массового обслуживания с гистерезисом и прогулками в дискретном времени // *T-Comm — Телекоммуникации и транспорт*, 2012. № 7. С. 5–8.
24. Гайдамака Ю. В., Самуйлов К. Е., Сопин Э. С. Модель одной системы массового обслуживания типа  $M/G/1$  с гистерезисным управлением входящим потоком // *T-Comm — Телекоммуникации и Транспорт*, 2012. № 7. С. 60–62.
25. Abaev P., Pechinkin A., Razumchik R. On analytical model for optimal SIP server hop-by-hop overload control // 4th Congress (International) on Ultra Modern Telecommunications and Control Systems ICUMT-2012 Proceedings. — IEEE, 2012. P. 299–304. doi:10.1109/ICUMT.2012.6459680.
26. Abaev P., Pechinkin A., Razumchik R. Analysis of queueing system with constant service time for SIP server hop-by-hop overload control // *Modern Probab. Meth. Anal. Telecommunication Networks Communications Computer Information Sci.*, 2013. Vol. 356. P. 1–10. doi:10.1007/978-3-642-35980-4\_1.
27. Abaev P., Pechinkin A., Razumchik R. On mean return time in queueing system with constant service time and bi-level hysteric policy // *Modern Probab. Meth. Anal. Telecommunication Networks Communications Computer Information Sci.*, 2013. Vol. 356. P. 11–19. doi:10.1007/978-3-642-35980-4\_2.
28. Abaev P., Gaidamaka Yu., Samouylov K., Shorgin S. Design and software architecture of SIP server for overload control simulation // 27th Conference (European) on Modelling and Simulation (ECMS 2013) Proceedings. Aalesund, Norway, 2013. P. 533–539. doi:10.7148/2013-0580.
29. Pechinkin A. V., Razumchik R. V. Approach for analysis of finite  $M_2/M_2/1/R$  with hysteric policy for SIP server hop-by-hop overload control // 27th Conference (European) on Modelling and Simulation (ECMS 2013) Proceedings. Aalesund, Norway, 2013. P. 573–579. doi:10.7148/2013.
30. Shorgin S., Samouylov K., Gaidamaka Yu., Etezoov Sh. Polling system with threshold control for modeling of SIP server under overload // 18th Conference (International) on Systems Science (ICSS 2013) Proceedings. Advances in intelligent systems and computing ser., 2014. Vol. 240. P. 97–107. doi:10.1007/978-3-319-01857-7\_10.
31. Гайдамака Ю. В. Модель с пороговым управлением нагрузкой для анализа серверов протокола SIP в режиме перегрузок // *Автоматика и вычислительная техника*, 2013. № 4. С. 65–75.
32. Abaev P., Gaidamaka Yu., Samouylov K., Pechinkin A., Razumchik R., Shorgin S. Hysteretic control technique for overload problem solution in network of SIP servers // *Comput. Inform.*, 2014 (in press). Vol. 33. No. 1.
33. MathWorks — MATLAB and Simulink for Technical Computing. [Электронный ресурс] Режим доступа <http://www.mathworks.com>, свободный (дата обращения 01.10.2013).
34. Wolfram Mathematica: Программное обеспечение для технических вычислений. [Электронный ресурс] Режим доступа: <http://www.wolfram.com/mathematica>, свободный (дата обращения 01.10.2013).
35. Maple — Technical Computing Software for Engineers, Mathematicians, Scientists, Instructors, and Students. [Электронный ресурс] Режим доступа: <http://www.maplesoft.com/products/maple>, свободный (дата обращения 01.10.2013).
36. Bocharov P. P., D'Apice C., Pechinkin A. V., Salerno S. Queueing theory. — Utrecht, Boston: VSP, 2004. 446 p.
37. Kempa W. On time to buffer saturation in a  $GI/M/1/N$ -type queue // *Int. J. Adv. Telecommunications Electrotechnics Signals Syst.*, 2012. Vol. 1. No. 2–3. С. 60–66.

Поступила в редакцию 26.09.13

# THE DISTRIBUTION OF THE RETURN TIME FROM THE SET OF OVERLOAD STATES TO THE SET OF NORMAL LOAD STATES IN A SYSTEM $M|M|1|\langle L, H \rangle|\langle H, R \rangle$ WITH HYSTERETIC LOAD CONTROL

Yu. V. Gaidamaka<sup>1</sup>, A. V. Pechinkin<sup>2</sup>, R. V. Razumchik<sup>2</sup>, A. K. Samuylov<sup>1</sup>, K. E. Samouylov<sup>1</sup>, I. A. Sokolov<sup>2</sup>, E. S. Sopin<sup>1</sup>, and S. Ya. Shorgin<sup>2</sup>

<sup>1</sup>Peoples' Friendship University of Russia, Moscow 117198, Russian Federation

<sup>2</sup>Institute of Informatics Problems, Russian Academy of Sciences, Moscow 119333, Russian Federation

**Abstract:** An analytical method for studying the parameters of the hysteretic control, which is implemented as one of the effective solutions to the overload problem in the network of SIP-servers, is suggested. As a mathematical model, the queuing system  $M|M|1|\langle L, H \rangle|\langle H, R \rangle$  with two loops hysteretic control was developed, where  $H$  is the overload onset threshold,  $L$  is the overload abatement threshold, and  $R$  is the discard threshold. Two methods of calculating the Laplace–Stieltjes transform of the distribution function of the return time from the set of overload system states to the set of normal load system states were obtained. The first method consists in solving a system of equations with return times for each state of the set of overload system states as unknowns, the second deals with the recurrence for the Laplace–Stieltjes transform of the distribution function of the return time for each state of the set of overload system states as rational fractional expressions. Both methods allow the effective calculations with standard software tools, as shown in the numerical example.

**Keywords:** SIP-server overload; queueing system; hysteretic load control; return time to normal load states; Laplace–Stieltjes transform; distribution function

**DOI:** 10.14357/19922264130403

## Acknowledgments

This work was supported by the Russian Foundation for Basic Research (projects Nos. 11-07-00112 and 12-07-00108).

## References

1. Hilt, V., E. Noel, C. Shen, and A. Abdelal. Design considerations for Session Initiation Protocol (SIP) overload control. 2011. Internet Engineering Task Force RFC-6357. Available at: <http://tools.ietf.org/html/rfc6357> (accessed October 1, 2013).
2. Hilt, V., J. Rosenberg, H. Schulzrinne, *et al.* SIP: Session Initiation Protocol. 2002. Internet Engineering Task Force RFC-3261. Available at: <http://tools.ietf.org/html/rfc3261> (accessed October 1, 2013).
3. Rosenberg, J. Requirements for management of overload in the Session Initiation Protocol. 2008. Internet Engineering Task Force RFC-5390. Available at: <http://tools.ietf.org/html/rfc5390> (accessed October 1, 2013).
4. Gurbani, V., V. Hilt, and H. Schulzrinne. Session Initiation Protocol overload control. 2013. Internet Engineering Task Force Draft. Available at: <http://tools.ietf.org/pdf/draft-ietf-soc-overload-control-13.pdf> (accessed October 1, 2013).
5. Noel, E., and P.M. Williams. Session Initiation Protocol rate control. 2013. Internet Engineering Task Force Draft. Available at: <http://tools.ietf.org/pdf/draft-ietf-soc-overload-rate-control-05.pdf> (accessed October 1, 2013).
6. ITU-T Recommendation Q.704: Signalling system No. 7 — Message Transfer Part, Signalling network functions and messages. 1996. Available at: <http://www.itu.int/rec/T-REC-Q.704-199607-1/en> (accessed October 1, 2013).
7. Abaev, P., Yu. Gaidamaka, and K. Samouylov. 2011. Gisterezisnoe upravlenie signal'noy nagruzkoy v seti SIP-serverov [Hysteretic overboard control in a SIP signaling network]. *Vestnik RUDN. Seriya Matematika, Informatika, Fizika [Bulletin of Peoples' Friendship University of Russia. Mathematics. Informatics. Physics ser.]* 4:55–73.
8. Abaev, P., Yu. Gaidamaka, A. Pechinkin, R. Razumchik, and S. Shorgin. 2012. Simulation of overload control in SIP server networks. *26th Conference (European) on Modelling and Simulation ECMS Proceedings*. Koblenz.

- 533–539.
9. Abaev, P., Yu. Gaidamaka, and K. Samouylov. 2012. Queuing model for loss-based overload control in a SIP server using a hysteretic technique. *Internet of things, smart spaces, and next generation networking*. Eds. S. Andreev, S. Balandin, and Ye. Koucheryavy. Lecture notes in computer science ser. Heidelberg: Springer-Verlag. 7469:371–378.
  10. Gebhart, R. F. 1967. A queuing process with bilevel hysteretic service-rate control. *Nav. Res. Logist. Q.* 14:55–68.
  11. Krasnoselskii, M., and A. Pokrovskii. 1989. *Systems with hysteresis*. Berlin – Heidelberg – New York – London – Paris – Tokio: Springer-Verlag. 410 p.
  12. Yum, T., and H. Yen. 1983. Design algorithm for a hysteresis buffer congestion control strategy. *IEEE Conference (International) on Communications Proceedings*. 499–503.
  13. Brown, P., P. Chemouil, and B. Delosme. 1984. A congestion control policy for signalling networks. *7th IeCC Proceedings*. 717–724.
  14. Golubchik, L., and J. Lui. 1997. Bounding of performance measures for a threshold-based queuing system with hysteresis. *Newsl. ACM SIGMETRICS Performance Evaluation Rev.* 25(1):147–157.
  15. Sindal, R., and S. Tokekar. 2008. Modeling and analysis of voice/data call admission control scheme in CDMA cellular network for variation in soft handoff threshold parameters. *16th IEEE Conference (International) on Networks (ICON 2008) Proceedings*. 1–6.
  16. Zhernovyi, K., and Yu. Zhernovyi. 2012. Sistema  $M^0/G/1$  s gisterizatsionnym pereklyucheniem intensivnosti obsluzhivaniya [The system  $M^0/G/1$  with a hysteretic switching intensity of service]. *Informatsionnye Protssesy – Information Processes* 12(3):176–190.
  17. Takagi, H. 1985. Analysis of a finite-capacity  $M/G/1$  queue with a resume level. *Perform. Evaluation* 5:197–203.
  18. Benaboud, H., and N. Mikou. 2002. Analysis by queuing model of multi-threshold mechanism in ATM switches. *5th IEEE Conference (International) on High Speed Networks and Multimedia Communications (HSNMC 2002) Proceedings*. 147–151.
  19. Dshalalow, J. H. 1997. Queuing systems with state dependent parameters. *Frontiers in queuing: Models and applications in science and engineering*. Ed. J. H. Dshalalow. Probability and stochastic ser. CRC Press. 61–116.
  20. Bekker, R. 2009. Queues with Levy input and hysteretic control. *Queueing Syst.* 63(1):281–299.
  21. Roughan, M., and C. Pearce. 2000. A martingale analysis of hysteretic overload control. *Adv. Perform. Anal. J. Teletraffic Theory Performance Anal. Communication Syst. Networks.* 3(1):1–30.
  22. Abaev, P., Yu. Gaidamaka, and K. Samouylov. 2012. Modeling of hysteretic signalling load control in next generation networks. *Internet of things, smart spaces, and next generation networking*. Eds. S. Andreev, S. Balandin, and Ye. Koucheryavy. Lecture notes in computer science ser. Heidelberg: Springer-Verlag. 7469:440–452.
  23. Abaev, P., and R. Razumchik. 2012. Modelirovanie raboty SIP servera s pomoshch'yu sistemy massovogo obsluzhivaniya s gisterizatsionnoy i progulkami v diskretnom vremeni [Modeling of SIP-server with hysteric overload control as discrete time queueing system]. *T-Comm — Telekommunikacii i transport [T-Comm — Telecommunications and Transport]* 7:5–8.
  24. Gaidamaka, Yu., K. Samouylov, and E. Sopin. 2012. Model' odnoy sistemy massovogo obsluzhivaniya tipa  $M/G/1$  s gisterizatsionnym upravleniem vkhodyashchim potokom [On queuing system of  $M|G|1$  type with hysteretic input flow control]. *T-Comm — Telekommunikatsii i transport [T-Comm — Telecommunications and Transport]* 7:60–62.
  25. Abaev, P., A. Pechinkin, and R. Razumchik. 2012. On analytical model for optimal SIP server hop-by-hop overload control. *4th Congress (International) on Ultra Modern Telecommunications and Control Systems Proceedings*. IEEE. 299–304. doi:10.1109/ICUMT.2012.6459680.
  26. Abaev, P., A. Pechinkin, and R. Razumchik. 2013. Analysis of queuing system with constant service time for SIP server hop-by-hop overload control. *Modern Probab. Meth. Anal. Telecommunication Networks Communications Computer Information Sci.* 356: 1–10. doi:10.1007/978-3-642-35980-4\_1.
  27. Abaev, P., A. Pechinkin, and R. Razumchik. 2013. On mean return time in queuing system with constant service time and bi-level hysteric policy. *Modern Probab. Meth. Anal. Telecommunication Networks Communications Computer Information Sci.* 356:11–19. doi:10.1007/978-3-642-35980-4\_2.
  28. Abaev, P., Yu. Gaidamaka, K. Samouylov, and S. Shorgin. 2013. Design and software architecture of SIP server for overload control simulation. *27th Conference (European) on Modelling and Simulation Proceedings*. Aalesund, Norway. 533–539. doi:10.7148/2013-0580.
  29. Pechinkin, A., and R. Razumchik. 2013. Approach for analysis of finite  $M_2/M_2/1/R$  with hysteric policy for SIP server hop-by-hop overload control. *27th Conference (European) on Modelling and Simulation Proceedings*. Aalesund, Norway. 573–579. doi:10.7148/2013.
  30. Shorgin, S., K. Samouylov, Yu. Gaidamaka, and Sh. Ete-zov. 2013. Polling system with threshold control for modeling of SIP server under overload. *18th Conference (International) on Systems Science Proceedings*. Wroclaw: Springer International Pubs. 240:97–107. doi:10.1007/978-3-319-01857-7\_10.
  31. Gaidamaka, Yu. 2013. Model' s porogovym upravleniem nagruzkoy dlya analiza serverov protokola SIP v rezhime peregruzok [On model with threshold load control for SIP server overload analysis]. *Avtomatika i Vychislitel'naya Tekhnika [Automatic Control and Computer Sciences]* 4:65–75.
  32. Abaev, P., Yu. Gaidamaka, K. Samouylov, A. Pechinkin, R. Razumchik, and S. Shorgin. 2014 (in press). Hysteretic control technique for overload problem solution in network of SIP servers. *Comput. Inform.* 33(1).
  33. MathWorks — MATLAB and Simulink for Technical Computing. Available at: <http://www.mathworks.com> (accessed October 1, 2013).

34. Wolfram Mathematica: Programmnoe obespechenie dlya tekhnicheskikh vychisleniy [Software for numerical computations]. Available at: <http://www.wolfram.com/mathematica> (accessed October 1, 2013).
35. Maple — Technical Computing Software for Engineers, Mathematicians, Scientists, Instructors and Students. Available at: <http://www.maplesoft.com/products/maple> (accessed October 1, 2013).
36. Bocharov, P. P., C. D'Apice, A. V. Pechinkin, and S. Salerno. 2004. *Queueing theory*. Utrecht, Boston: VSP. 446 p.
37. Kempa, W. 2012. On time to buffer saturation in a  $GI/M/1/N$ -type queue. *Int. J. Adv. Telecommunications Electrotechnics Signals Syst.* 1(2-3):60–66.

Received September 26, 2013

## Contributors

**Gaidamaka Yuliya V.** (b. 1971) — Candidate of Science (PhD) in physics and mathematics, associate professor, Peoples' Friendship University of Russia, Moscow 117198, Russian Federation; [ygaidamaka@sci.pfu.edu.ru](mailto:ygaidamaka@sci.pfu.edu.ru)

**Pechinkin Alexander V.** (b. 1946) — Doctor of Science in physics and mathematics; principal scientist, Institute of Informatics Problems, Russian Academy of Sciences, Moscow 119333, Russian Federation; professor, Peoples' Friendship University of Russia, Moscow 117198, Russian Federation; [apechinkin@ipiran.ru](mailto:apechinkin@ipiran.ru)

**Razumchik Rostislav V.** (b. 1984) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Russian Academy of Sciences, Moscow 119333, Russian Federation; [rrazumchik@ieee.org](mailto:rrazumchik@ieee.org)

**Samuylov Andrey K.** (b. 1988) — PhD student, Peoples' Friendship University of Russia, Moscow 117198, Russian Federation; [asam1988@gmail.com](mailto:asam1988@gmail.com)

**Samouylov Konstantin E.** (b. 1955) — Doctor of Science in technology, professor, Head of Department, Peoples' Friendship University of Russia, Moscow 117198, Russian Federation; [ksam@sci.pfu.edu.ru](mailto:ksam@sci.pfu.edu.ru)

**Sokolov Igor A.** (b. 1954) — Academician of the Russian Academy of Sciences, Doctor of Science in technology, Director, Institute of Informatics Problems, Russian Academy of Sciences, Moscow 119333, Russian Federation; [isokolov@ipiran.ru](mailto:isokolov@ipiran.ru)

**Sopin Eduard S.** (b. 1987) — PhD student, Peoples' Friendship University of Russia, Moscow 117198, Russian Federation; [sopin-eduard@yandex.ru](mailto:sopin-eduard@yandex.ru)

**Shorgin Sergey Ya.** (b. 1952) — Doctor of Science in physics and mathematics, professor, Deputy Director, Institute of Informatics Problems, Russian Academy of Sciences, Moscow 119333, Russian Federation; [sshorgin@ipiran.ru](mailto:sshorgin@ipiran.ru)