УДК 621.39

Анализ некоторых характеристик СМО $M|G|1|r$ с гистерезисным управлением для исследования перегрузок SIP-сервера

Ю. В. Гайдамака, Р. И. Закирова

Кафедра систем телекоммуникаций
Российский университет дружбы народов
ул. Миклухо-Маклая, д. 6, Москва, Россия, 117198

В современных телекоммуникационных сетях существует ряд задач, среди которых выделяют задачу поиска наиболее эффективного механизма управления перегрузками на SIP-серверах. В общем случае перегрузки связаны с тем, что интенсивность поступления вызовов на SIP-сервер превышает возможности по их обработке. Проблемы такого рода могут привести к снижению производительности SIP-сервера, а также могут быть причиной его полного отказа.

В стандартах комитета IETF в зависимости от типа перегрузок выделяют ряд решений проблемы, среди которых: увеличение числа SIP-серверов, механизм 503, метод просеивания потока, метод снижения скорости. Однако оптимального решения для управления перегрузок на SIP-сервере не найдено.

В работе предлагается упрощённый механизм контроля перегрузок, который позволяет осуществить управление интенсивностью поступления вызовов на SIP-сервер путём ввода порога снижения нагрузки. Разработана упрощённая математическая модель в виде системы массового обслуживания типа $M|G|1|r$ с пороговым управлением нагрузкой. Получено стационарное распределение вероятностей состояний системы методом вложенных цепей Маркова. Описан алгоритм для расчёта вероятностно-временных характеристик, таких как вероятность потери заявки, средняя длина очереди модели, время возврата из режима перегрузки в режим нормальной нагрузки. Численно решена оптимизационная задача, которая заключается в минимизации данной характеристики, проведён эксперимент, а также численный анализ полученных результатов.

Ключевые слова: SIP-сервер, пороговое управление, полумарковский процесс, время возврата из режима перегрузки.

1. Введение

Вследствие высокой популярности услуг в телекоммуникационной сети, архитектура которой поддерживает протокол SIP, могут возникать перегрузки, связанные с тем, что интенсивность поступления вызовов на SIP-сервер превышает возможности по их обработке. Перегрузки можно разделить на две категории: сервер – клиент и сервер – сервер. Проблема перегрузок типа клиент – сервер может решиться увеличением числа SIP-серверов, для контроля перегрузок типа сервер – сервер существует несколько механизмов. Так, спецификация RFC 3261 [1] описывает механизм 503, который в случае, когда сервер не может обработать запрос, предусматривает передачу уведомляющего об этом сообщения 503 Service Unavailable серверу-отправителю. Кроме того, выделяют три типа управления:

- механизм сквозного управления, который предусматривает сбор информации по маршруту о состоянии всех узлов и ограничение поступающей по маршруту нагрузки как можно ближе к источнику;
- механизм межузлового управления, который осуществляет контроль перегрузок только между соседними узлами [2–4];

Статья поступила в редакцию 25 ноября 2013 г.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №12-07-00108.

Авторы выражают благодарность профессору Самуйлову К. Е. за внимание к работе и ряд важных замечаний, а также ассистенту кафедры систем телекоммуникаций Сопину Э. С. за помощь в проведении исследований.

– механизм локального контроля перегрузок, который предполагает, что сервер на основе мониторинга текущего уровня использования своих ресурсов принимает решение о сбросе части нагрузки.

В статье построен пороговый механизм межузлового управления нагрузкой. Статья организована следующим образом. В разделе 2 построена упрощённая математическая модель процесса обработки сообщений SIP-сервером в терминах теории массового обслуживания. В разделе 3 получено стационарное распределение вероятностей для рассматриваемой модели. Раздел 4 содержит описание вероятностно-временных характеристик, таких как вероятность потери заявки, средняя длина очереди, время возврата из режима перегрузки в режим нормальной нагрузки. Агоритм для расчёта вероятностно-временных характеристик приведен в разделе 5 статьи. В разделе 6 рассматривается пример численного анализа.

2. Описание математической модели

Рассматривается однолинейная система массового обслуживания (СМО), состоящая из одного обслуживающего прибора, буферного накопителя ёмкости r с порогом снижения перегрузки L (рис. 1). Поток заявок, поступающих на прибор, является пуассоновским с интенсивностью λ. Если поступающая заявка застаёт прибор свободным, она немедленно начинает обслуживаться, в противном случае заявка занимает место в накопителе. В случае, если мест в накопителе нет, заявка теряется. Длительность обслуживания является случайной величиной (СВ) с произвольной функцией распределения (ФР) $B(x)$ и средним $b^{(1)} < \infty$.

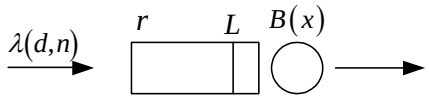

Рис. 1. СМО с пороговым управлением нагрузкой

Порог снижения перегрузки L служит для предотвращения частых переключений между режимами функционирования системы — режимом нормальной нагрузки и режимом перегрузки. В случае, когда длина очереди n в буфере достигает значения r, система переходит в режим перегрузки. Возврат в нормальный режим функционирования происходит не сразу, а при снижении длины очереди до значения L. Описанный механизм относится к механизмам гистерезисного (порогового) управления перегрузкой и является частным случаем механизма из [5], где исследована СМО $M|M|1|r$ с несколькими группами порогов.

На рис. 2 показана зависимость входящего потока заявок от длины очереди, что отражает описанный выше механизм гистерезисного управления.

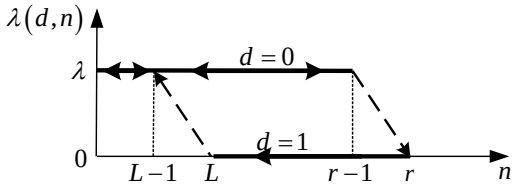

Рис. 2. Гистерезисное управление нагрузкой

Множество состояний СМО (рис. 1) с интенсивностью входящего потока $\lambda(d,n)$ (рис. 2) представляет собой объединение двух множеств: $X = X_0 \cup X_1$, где $X_0 = \{(d,n) : d = 0,\ 0 \leqslant n \leqslant r-1\}$ — множество состояний нормальной нагрузки, $X_1 = \{(d,n) : d = 1,\ L \leqslant n \leqslant r\}$ — множество состояний перегрузки.

Введём случайный процесс (СП) $\xi(t)$ — число заявок в СМО в момент времени t. Пусть t_k, $k \geqslant 0$, — моменты ухода заявок из системы [6]. Обозначим $\xi_k = \xi(t_k+0)$, тогда последовательность $\{\xi_k,\, k \geqslant 0\}$ образует вложенную цепь Маркова (ЦМ).

Введём стационарное распределение СП $\xi(t)$ и ЦМ $\{\xi_k,\, k \geqslant 0\}$:

$$P_{d,j} = \lim_{t\to\infty} P\{\xi(t) = j\}, \quad j = 0, \ldots, r+1, \quad d \in \{0,1\},$$

$$q_{d,j} = \lim_{t\to\infty} P\{\xi(t_k+0) = j\}, \quad j = 0, \ldots, r, \quad d \in \{0,1\}.$$

Получение аналитических формул для стационарных вероятностей состояний СП $\xi(t)$ позволит находить основные вероятностно-временные характеристики системы, такие как вероятность потери заявки, средняя длина очереди, среднее время нахождения системы в множествах X_0 и X_1 [7].

3. Стационарное распределение вероятностей

Построим граф переходных вероятностей для ЦМ $\{\xi_k,\, k \geqslant 0\}$ (рис. 3), и для нахождения стационарных вероятностей СП $\xi(t)$ воспользуемся методом, предложенным в [8].

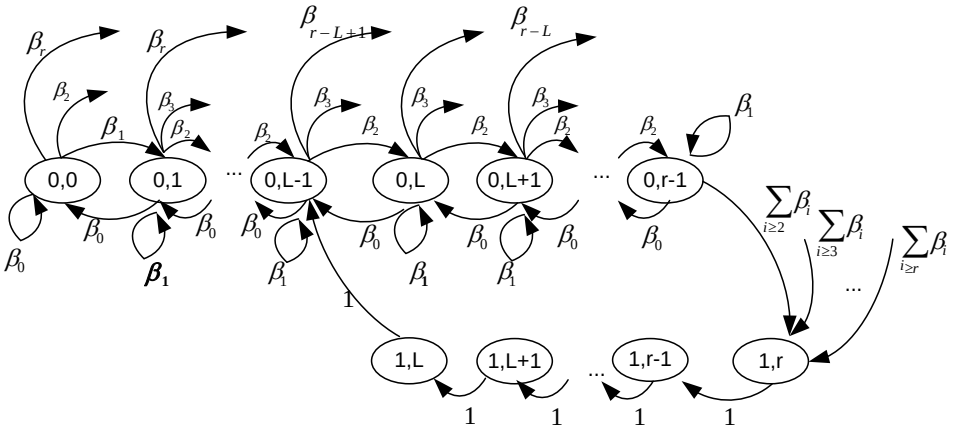

Рис. 3. Граф переходных вероятностей ЦМ $\{\xi_k,\, k \geqslant 0\}$

Нетрудно убедиться, что стационарное распределение $\{q_{d,j}\}_{j=0,\ldots,r}^{d\in\{0,1\}}$ имеет вид

$$\begin{cases} q_{0,j} = q_{0,0}k_j, \quad j = 0, \ldots, L-1, \\ q_{0,j} = q_{0,0}(k_j - a_1 q_{0,0}), \quad j = L, \ldots, r-1, \\ q_{1,r} = a_1 q_{0,0}, \end{cases} \tag{1}$$

где k_j и h_j вычисляются по рекуррентным формулам

$$\begin{gathered} k_0 = 1, \quad k_1 = \frac{1-\beta_0}{\beta_0}, \\ k_j = \frac{1}{\beta_0}\left(k_{j-1} - \beta_{j-i} - \sum_{i=1}^{j-1} k_i \beta_{j-1}\right), \quad j = 2, \ldots, L-1, \end{gathered} \tag{2}$$

$$h_0 = \frac{1}{\beta_0}, \quad h_j = \frac{1}{\beta_0}\left(h_{j-1} - \sum_{i=1}^{j} h_{i-1}\beta_{j-i+1}\right), \quad j = 1, \ldots, r-L-1, \tag{3}$$

$$a_1 = \frac{\alpha_r + \sum\limits_{i=1}^{r-1} k_i \alpha_{r-i+1}}{1 + \sum\limits_{i=L}^{r-1} h_{i-L}\alpha_{r-i+1}}, \tag{4}$$

$$q_{0,0} = \frac{1}{\sum\limits_{j=0}^{r-1} k_j + a_1 \sum\limits_{j=L}^{r-1} h_{j-L} + (r-L+1)a_1}. \tag{5}$$

Отметим, что α_l — вероятность поступления в СМО не менее, чем l заявок за случайное время наблюдения, распределённое в соответствии с функцией распределения $B(x)$,

$$\alpha_l = \int_0^\infty (1 - B(x)) \exp^{-\lambda x} \frac{\lambda^l x^{l-1}}{(l-1)!} \mathrm{d}x, \quad l \geqslant 1, \quad \sum_{l=1}^{\infty} \alpha_l = \lambda b^{(1)}.$$

При этом β_l — вероятность поступления в СМО ровно l заявок за случайное время наблюдения, распределённое в соответствии с функцией распределения $B(x)$,

$$\beta_l = \int_0^\infty \exp^{-\lambda x} \frac{(\lambda x)^l}{l!} \mathrm{d}x, \quad l \geqslant 1, \quad \sum_{l=0}^{\infty} \beta_l = 1,$$

а соотношение величин α_l и β_l определяется в виде $\alpha_{l+1} = 1 - \sum\limits_{j=0}^{l} \beta_j, \quad l \geqslant 0.$

Однако для практических целей интерес представляют стационарные вероятности по времени $P_{d,j} = \lim\limits_{t\to\infty} P\{\xi(t) = j\}$, $d \in \{0, 1\}$, поскольку именно эти вероятности определяют распределение очереди в момент поступления заявки в систему.

Введём величину C — цикл восстановления, который определяется по формуле [9]:

$$C = \left(\frac{1}{\lambda} + b^{(1)}\right) q_{0,0} + (1 - q_{0,0})\, b^{(1)} = b^{(1)} + \frac{1}{\lambda} q_{0,0}. \tag{6}$$

,

В результате получаем стационарное распределение вероятностей СП $\xi(t)$ через стационарное распределение вероятностей ЦМ $\{\xi_k,\, k \geqslant 0\}$:

$$\begin{cases} P_{0,0} = \dfrac{C^{-1}}{\lambda} q_{0,0}, \\ P_{0,j} = \dfrac{C^{-1}}{\lambda}\left[q_{0,0}\alpha_j + \sum\limits_{i=1}^{\min(j,r-1)} q_{0,i}\alpha_{j-i+1}\right], \quad j = 1, \ldots, r, \\ P_{1,r+1} = \dfrac{C^{-1}}{\lambda}\left[q_{0,0}\left(\lambda b^{(1)} - \sum\limits_{l=1}^{r} \alpha_l\right) + \sum\limits_{i=1}^{r-1} q_{0,i}\left(\lambda b^{(1)} - \sum\limits_{l=1}^{r-i+1} \alpha_l\right)\right], \\ P_{1,j} = C^{(-1)} b^{(1)} q_{1,r}, \quad j = L, \ldots, r. \end{cases} \tag{7}$$

4. Вероятностно-временные характеристики

Зная стационарное распределение вероятностей $P_{d,j}$, можно рассчитать такие вероятностно-временные характеристики (ВВХ) системы, как средняя длина очереди в режиме нормальной нагрузки и в режиме перегрузки, соответственно, Q_0 и Q_1, и вероятность потери заявки π:

$$Q_0 = \sum_{j=1}^{r-1} jP_{0,j}, \quad Q_1 = \sum_{j=L}^{r} jP_{1,j},$$

$$\pi = P(X_1) = \sum_{j=L}^{r+1} P_{1,j} = P_{1,r+1} + (r - L + 1)C^{-1}b^{(1)}q_{1,r}.$$

Ещё одной важной характеристикой является время возврата. Данная характеристика определяется как среднее время возврата из множества состояний перегрузки (множество X_1) в множество состояний нормальной нагрузки (множество X_0) [5], и представляет собой интервал от момента первого попадания в состояние $(1, r)$ множества перегрузки до момента первого выхода из множества состояний перегрузки в состояние $(0, L - 1)$ нормальной нагрузки.

Среднее время возврата τ вычисляется по формуле

$$\tau = -\varphi'(s)|_{s=0}, \tag{8}$$

где

$$\varphi(s) = \beta^{r-L+1}(s). \tag{9}$$

Здесь $\varphi(s)$ — преобразование Лапласа–Стилтьеса (ПЛС) ФР времени возврата, а $\beta(s)$ — ПЛС ФР $B(x)$.

С учётом (9) формула (8) для расчёта среднего времени возврата будет иметь вид:

$$\tau = (r - L + 1)b^{(1)}.$$

Дисперсия среднего времени возврата рассчитывается по формуле

$$D\tau = (r - L + 1)(b^{(2)} - (b^{(1)})^2).$$

5. Алгоритм расчёта вероятностно-временных характеристик

На основании полученных выше формул алгоритм вычисления стационарного распределения сводится к следующим шагам.

Шаг 1. Вычисление α_j, β_j, $j = 0, \ldots, r$. Алгоритм вычисления данных величин зависит от вида соответствующих функций распределения. Величины β_j, $j \geqslant 0$ связаны с ПЛС $\beta(s)$ следующим соотношением:

$$\beta_j = \frac{(-1)^j}{j!}\lambda^j\beta^{(j)}(\lambda), \quad j \geqslant 0,$$

которое при явном задании функции $B(x)$ даёт возможность получить расчётные формулы.

Шаг 2. Вычисление величин k_j, h_j, $j \geqslant 0$ по формулам (2), (3).

Шаг 3. Вычисление величины a_1 по формуле (4).

Шаг 4. Вычисление стационарных вероятностей $\{q_{d,j}\}_{j=0,\ldots,r}^{d\in\{0,1\}}$ по формулам (5) и (1).

Шаг 5. Вычисление величины C по формуле (6).

Шаг 6. Вычисление стационарных вероятностей $\{P_{d,j}\}_{j=1,\ldots,r+1}^{d\in\{0,1\}}$ по формулам (7).

6. Пример численного анализа

Рассмотрим пример, где время обслуживания распределенно по экспоненциальному закону $B(x) = 1 - exp^{-\mu x}$, ПЛС $\beta(s) = \dfrac{\mu}{\mu + s}$, а время возврата для модели типа $M|M|1|r$ с пороговым управлением нагрузкой имеет вид $\tau = (r - L + 1)\mu^{-1}$.

Перейдём к численному анализу времени возврата. В качестве исходных данных использовались значение поступающей нагрузки $\rho = 1,2$, среднее время обслуживания сообщений $\mu^{-1} = 5$ мс и объём буферного накопителя $r = 150$.

Решается оптимизационная задача, которая заключается в нахождении таких значений порога L снижения перегрузки, для которых время возврата τ было бы минимальным, с учётом двух ограничений.

Первое ограничение накладывается на вероятность π попадания в множество состояний перегрузки. Как видно из рис. 4, при условии $\pi < 0,2$ порог L принимает значение от 1 до 150, что соответствует тому, что при выбранном диапазоне значений порога L система будет находиться в режиме перегрузки менее 20% от среднего времени цикла.

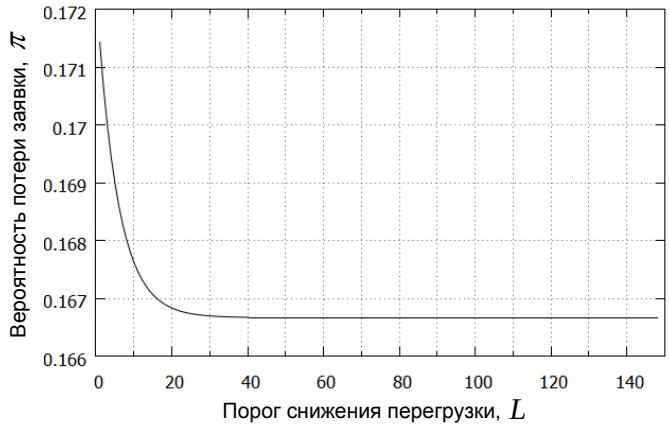

Рис. 4. Зависимость вероятности π от порога L

Второе ограничение определяется как условие, накладываемое на среднее время цикла управления $\tau_{\text{общ}}$, которое представляет собой средний интервал от момента попадания процесса в множество X_0 состояний нормальной нагрузки из множества X_1 состояний перегрузки до момента следующего попадания процесса в множество X_0 из множества X_1. Для избежания осцилляций системы между состояниями нормальной нагрузки и состояниями перегрузки время $\tau_{\text{общ}}$ следует максимизировать.

Времена $\tau_{\text{общ}}$ и τ связаны следующим соотношением:

$$\tau_{\text{общ}} = \tau + \tau\frac{P(X_0)}{P(X_1)} = \frac{\tau}{P(X_1)}.$$

В рассмотренном примере второе ограничение имеет вид $\tau_{\text{общ}} \geqslant 0,45$ с. Заметим, что значение $0,45$ с в данном случае выбрано для иллюстрации метода, а при расчёте реальных систем должно удовлетворять требованиям международных стандартов.

Как видно из рис. 5, накладываемому ограничению соответствует диапазон значений порога $L \in (1; 136)$.

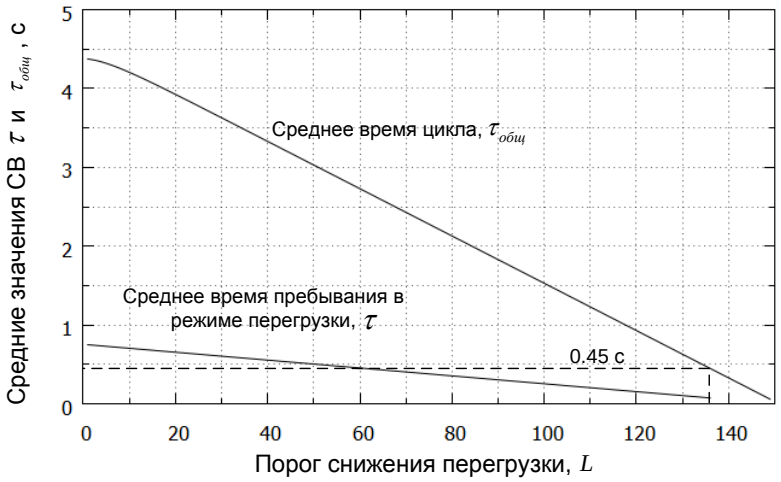

Рис. 5. Зависимость среднего времени возврата $\tau_{\text{общ}}$ от порога L

Построив график зависимости среднего времени возврата τ относительно выбора порога $L \in (1; 136)$, получили решение оптимизационной задачи: $\tau(136) = 0,075$ с. Таким образом, минимальное время $0,075$ с пребывания системы в множестве состояний перегрузки достигается при значении порога $L = 136$.

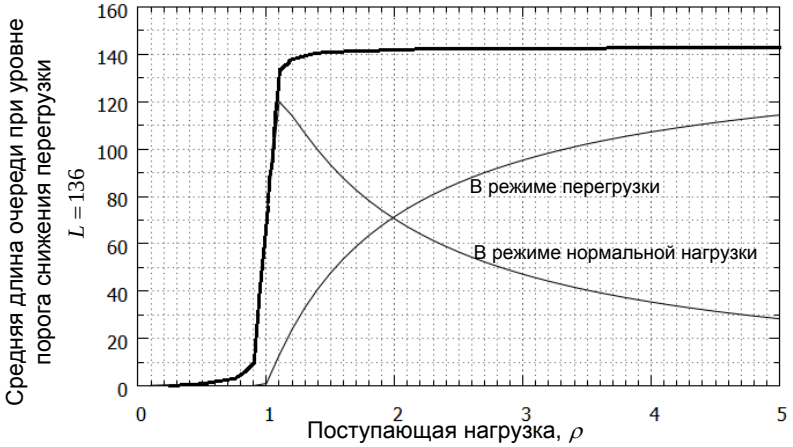

Рис. 6. Зависимость средней длины очереди при пороге снижения нагрузки $L = 136$ от поступающей нагрузки

Далее для порога снижения перегрузки $L = 136$ построен график зависимости средней длины очереди Q от поступающей нагрузки ρ (рис. 6). Перегиб на графике при $\rho = 1,1$ объясняется тем, что в этот момент система переходит в режим перегрузки. При увеличении значений ρ средняя длина очереди в режиме нормальной нагрузки убывает, поскольку вероятности пребывания системы в состояниях множества нормальной нагрузки падают. Аналогично средняя длина очереди в режиме перегрузки растёт. Таким образом, средняя длина очереди при заданных значениях $r = 150$, $L = 136$, $\rho = 1,2$ принимает значение $Q = 138$.

Задачей дальнейших исследований может стать построение и анализ модели с несколькими группами порогов, а также исследование сквозного механизма управления перегрузками SIP-серверов.

Литература

1. *Rosenberg J., Schulzrinne H., Camarillo G. et al.* RFC 3261 — SIP: Session Initiation Protocol. — 2002. — `http://www.ietf.org/rfc/rfc3261.txt`.
2. *Hilt V., Noel E., Shen C., Abdelal A.* Design Considerations for Session Initiation Protocol (SIP) Overload Control. — 2011. — `http://tools.ietf.org/html/rfc6357`.
3. *Hilt V.* Session Initiation Protocol (SIP) Overload Control. — 2012. — `http://tools.ietf.org/html/draft-ietf-soc-overload-control-11`.
4. *Williams P. M.* Session Initiation Protocol (SIP) Rate Control // IETF. — 2012. — `http://tools.ietf.org/html/draft-ietf-sipcore-event-rate-control-09`.
5. *Абаев П. О., Гайдамака Ю. В., Самуйлов К. Е.* Гистерезисное управление нагрузкой в сети SIP-серверов. // Вестник РУДН. Серия «Математика. Информатика. Физика». — 2011. [Abaev P. O., Gaidamaka Yu. V., Samouylov K. E. Load Control technique with Hysteresis in SIP Signaling Server. — Moscow: PFUR, 2011. — (in russian).]
6. *Бочаров П. П., Печинкин А. В.* Теория массового обслуживания. — М.: РУДН, 1995. [Bocharov P. P., Pechinkin A. V. Queueing Theory. — M.: PFUR, 1995. — (in russian).]
7. *Вентцель Е. С., Овчаров Л. А.* Теория случайных процессов и ее инженерные приложения. — М.: Наука, 1990. [Ventsel' E. S., Ovcharov L. A. The Theory of Stochastic Processes and its Engineering Applications. — Moscow: Nauka, 1990. — (in russian).]
8. *Takagi H.* Analysis of a Finite-Capacity M/G/1 Queue with a Resume Level // Perform. Eval. — 1985. — Vol. 5, No 3. — Pp. 197–203.
9. *Ивченко Г. И., Каштанов В. А., Коваленко И. Н.* Теория массового обслуживания. — М.: Высшая школа, 1982. [Ivchenko G. I., Kashtanov V. A., Kovalenko I. N. Queueing Theory. — Moscow: Vysshaya Shkola, 1982. — (in russian).]

UDC 621.39

Analysis of a Finite-Capacity $M|G|1|r$ Queue with Threshold Overload Control

Y. V. Gaidamaka, R. I. Zakirova

Telecommunication Systems Department
Peoples' Friendship University of Russia
6, Miklukho-Maklaya str., Moscow, Russia, 117198

One of the main challenges faced by telecommunications industry today is an issue of searching for the most effective overload control mechanisms on SIP servers. Generally, overload occurs in SIP networks when SIP servers have insufficient resources to handle all SIP messages they receive to handle all incoming SIP traffic. Such problems can decrease performance of SIP server or even cause its crash.

The IETF offers several solutions depending on types of overloads: to increase the number of SIP servers, through 503 (Service Unavailable) response code (IETF RFC 3261), rate-based overload control, loss-based overload control. However, SIP servers are still vulnerable to overload.

In this paper we have built and analyzed the $M|G|1|r$ queue with one level hysteretic input load control. Stationary distribution has been achieved based on the Embedded Markov chain method. Approach that allows computation of probability of loss and an average length of queue is developed. Another important parameter, the return time from overloading states to normal state is also considered. A numerical example illustrating the control mechanism that minimizes this characteristic is given to demonstrate some optimization issues.

Key words and phrases: SIP-server, threshold control, finite-capacity queue, semi-Markov process, return time.