УДК 621.39

Модель SIP-сервера с дисциплинами шлюзового и исчерпывающего обслуживания очередей

Ю. В. Гайдамака, Э. Р. Зарипова

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

В статье исследуются характеристики СМО как системы поллинга с циклическим порядком опроса очередей, исчерпывающей и шлюзовой дисциплинами обслуживания. Представлены формулы в явном виде, характеризующие среднее время ожидания в очереди. Проведена реализация численного эксперимента.

Ключевые слова: система поллинга, SIP-сервер, исчерпывающая дисциплина обслуживания, шлюзовая дисциплина обслуживания, математическая модель.

1. Введение

Протокол инициирования сеансов связи (Session Initiation Protocol, SIP) является неотъемлемой частью концепции сетей связи следующего поколения (Next Generation Network, NGN). Протокол применяется для установления, модификации и завершения сеансов связи и его характеризуют масштабируемость, расширяемость, взаимодействие с другими протоколами, интеграция в стек уже существующих протоколов TCP/IP. Разрабатывала протокол SIP Рабочая группа по проектированию Интернет-технологий (Internet Engineering Task Force, IETF), а основным стандартизующим документом является RFC 3261 [1]. В последние несколько лет ведутся интенсивные исследования по анализу перегрузок в сетях на базе протокола SIP. Как показало практическое применение, прописанный в RFC 3261 базовый алгоритм контроля перегрузок, использующий код ошибки 503 для определения недоступности услуги (Service Unavailable), не обладает необходимой эффективностью, о чем говорится, в том числе, и в документе IETF [2]. Если SIP-сервер не может обработать запрос из-за временной перегрузки, он отклоняет его с кодом ошибки 503. Однако с увеличением числа услуг, предоставляемых с использованием протокола SIP, значительно увеличивается и нагрузка на SIP-сервера, порождающая, соответственно, вероятность перегрузки сервера. Таким образом, возникает задача создания алгоритма, способного предотвратить возникновение перегрузок на SIP-серверах.

Данная задача решалась многими авторами [3–5], но до сих пор не исследован вопрос о выборе эффективной дисциплины обслуживания сервером сообщений протокола SIP, которая учитывает поступление потоков сообщений двух типов – запросов INVITE и всех остальных сообщений – ответов. Разделение сообщений на два потока целесообразно потому, что механизмы защиты от перегрузок предусматривают в первую очередь сброс сообщений-запросов (INVITE), а не сообщений-ответов (non-INVITE).

В статье исследуется простейшая модель функционирования SIP-сервера, учитывающая обслуживание двух очередей – очередь сообщений INVITE (1-заявки) и очередь сообщений non-INVITE (2-заявки). Предлагается провести сравнительный анализ двух известных моделей поллинга – со шлюзовой и с исчерпывающей дисциплиной обслуживания очередей [6]. Для этого в статье в явном виде получены формулы для расчёта среднего времени пребывания сообщений обоих типов в очереди и проведён численный анализ, который для типичного набора исходных данных показал преимущество шлюзовой дисциплины обслуживания.

Статья поступила в редакцию 26 октября 2012 г.

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

Авторы благодарят проф. К.Е. Самуйлова за полезные замечания и внимание к работе.

2. Описание модели

Рассмотрим поллинговую модель SIP-сервера, схематично изображённую на рис. 1. Здесь и далее будем, в основном, придерживаться понятий и обозначений, принятых в монографии [6]. Исследуются две дисциплины обслуживания очередей – исчерпывающая и шлюзовая. При исчерпывающей дисциплине обслуживания очередей прибор обслуживает заявки до тех пор, пока очередь не опустеет. При шлюзовой дисциплине обслуживания прибор обслуживает лишь те заявки, которые находились в очереди на момент опроса (момент завершения подключения прибора к очереди), а заявки, поступившие в очередь после момента опроса, обслуживаются в следующем цикле.

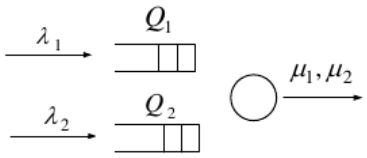

Рис. 1. Схема модели SIP-сервера

Предполагается, что времена обслуживания i-заявок в i-очереди (Q_i) – независимые одинаково распределённые случайные величины (СВ), с функцией распределения (ФР) $B_i(t)$. Среднее время обслуживания в очереди Q_i определяется, как $b_i = \int\limits_0^\infty t\, dB_i(t)$, $b_i = \frac{1}{\mu_i}$, а второй начальный момент, как $b_i^{(2)} = \int\limits_0^\infty t^2\, dB_i(t)$. Предполагается, что времена переключения между очередями являются независимыми СВ с ФР $S_i(t)$ с первыми двумя начальными моментами s_i и $s_i^{(2)}$ соответственно. Для рассматриваемой системы справедливы следующие соотношения: $s = s_1 + s_2$ и $s^{(2)} = s^2 + \sum\limits_{i=1}^{2} (s_i^{(2)} - s_i^2)$. Нагрузку на очередь Q_i обозначим $\rho_i = \frac{\lambda_i}{\mu_i}$, а суммарную нагрузку на систему $\rho = \rho_1 + \rho_2$. Здесь λ_i - интенсивность поступающего в очередь Q_i пуассоновского потока i-заявок, и пусть $\lambda = \lambda_1 + \lambda_2$. Обозначим $C = \frac{s}{1-\rho}$ среднюю длину цикла – интервала времени, за который прибор посетит обе очереди системы ровно один раз. Отметим, что в системе поллинга стационарный режим существует в том случае, если длины всех очередей в моменты опроса имеют стационарное распределение, а длина цикла имеет конечное математическое ожидание, т.е. $\rho < 1$.

Обозначим $M[W_i]$ – среднее время ожидания обслуживания i-заявками в очереди Q_i, $i = 1, 2$. Именно эти вероятностные характеристики являются предметом анализа в данной статье, в следующем разделе которой в явном виде получены формулы для их вычисления, а в заключительном разделе с их помощью проводится сравнение исчерпывающей и шлюзовой дисциплин обслуживания.

3. Анализ времени ожидания сообщений в очереди

Рассмотрим модель поллинга со шлюзовой и с исчерпывающей дисциплинами обслуживания очередей. Сделаем допущение, что стационарный режим существует. Используя предложенные в [6] выражения для производящих функций получены следующие формулы для средних времён ожидания заявок в очереди.

Лемма 1. *Среднее время ожидания в очереди для модели СМО типа $M_2|G_2|1$ с поллингом и со шлюзовой дисциплиной обслуживания очередей определяется формулами:*

$$M[W_1] = \frac{(1+\rho_1)}{2\lambda_1^2 C} \frac{(1-\rho_1)^2 (1-\rho_2)^2}{(1-\rho)(1-\rho_1+\lambda_2 b_1)(1-\rho_2+\lambda_1 b_2)} \times$$

$$\times\lambda_1\left(\lambda_1\left(s_1^{(2)}+\lambda_1^2 s_1^2\left(\rho_1\frac{s}{1-\rho}+s_2\right)^2+\lambda_2\frac{s}{1-\rho}\left(\frac{2b_2s_1}{1-\rho_2}+\frac{b_2^{(2)}}{(1-\rho_2)^3}\right)\right)+\right.$$
$$+\frac{b_2}{1-\rho_2}\left(\lambda_2^2\left(s_2^{(2)}+\lambda_2^2 s_2^2\left(\rho_2\frac{s}{1-\rho}+s_1\right)^2+\lambda_1\frac{s}{1-\rho}\left(\frac{2b_1s_2}{1-\rho_1}+\frac{b_1^{(2)}}{(1-\rho_1)^3}\right)\right)+\right.$$
$$\frac{\rho_1\lambda_2^2}{1-\rho_1}\left(s_1^{(2)}+s_1\left(\frac{\rho_1 s}{1-\rho}+s_2\right)\left(1+\frac{\lambda_1 b_2}{1-\rho_2}\right)\right)+\lambda_2^2 s_1^{(2)}+$$
$$\left.+\lambda_1\lambda_2\left(s_2^{(2)}+s_2\left(\frac{\rho_2 s}{1-\rho}+s_1\right)\left(1+\frac{\lambda_2 b_1}{1-\rho_1}\right)\right)\right)+\lambda_1 s_2^{(2)}+\lambda_1\left(\frac{\rho_2}{1-\rho_2}\right)^2\times$$
$$\times\left(s_2^{(2)}++s_2^2\lambda_2^2\left(\rho_2\frac{s}{1-\rho}+s_1\right)^2+\frac{\lambda_1 s}{1-\rho}\left(\frac{2b_1s_2}{1-\rho_1}+\frac{b_1^{(2)}}{(1-\rho_1)^3}\right)+\right.$$
$$\left.+\frac{\rho_1}{1-\rho_1}\left(s_1^{(2)}+s_1\left(\frac{\rho_1 s}{1-\rho}+s_2\right)\left(1+\frac{\lambda_1 b_2}{1-\rho_2}\right)\right)+s_1^{(2)}\right),$$

$$M[W_2]=\frac{(1+\rho_2)}{2\lambda_2^2 C}\frac{(1-\rho_1)^2(1-\rho_2)^2}{(1-\rho)(1-\rho_1+\lambda_2 b_1)(1-\rho_2+\lambda_1 b_2)}\times$$
$$\times\lambda_2\left(\lambda_2\left(s_2^{(2)}+\lambda_2^2 s_2^2\left(\rho_2\frac{s}{1-\rho}+s_1\right)^2+\lambda_1\frac{s}{1-\rho}\left(\frac{2b_1s_2}{1-\rho_1}+\frac{b_1^{(2)}}{(1-\rho_1)^3}\right)\right)+\right.$$
$$+\frac{b_1}{1-\rho_1}\left(\lambda_1^2\left(s_1^{(2)}+\lambda_1^2 s_1^2\left(\rho_1\frac{s}{1-\rho}+s_2\right)^2+\right.\right.$$
$$\left.\left.+\lambda_2\frac{s}{1-\rho}\left(\frac{2b_2s_1}{1-\rho_2}+\frac{b_2^{(2)}}{(1-\rho_2)^3}\right)\right)+\right.$$
$$+\frac{\rho_2\lambda_1^2}{1-\rho_2}\left(s_2^{(2)}+s_2\left(\frac{\rho_2 s}{1-\rho}+s_1\right)\left(1+\frac{\lambda_2 b_1}{1-\rho_1}\right)\right)+\lambda_1^2 s_2^{(2)}+$$
$$\left.+\lambda_1\lambda_2\left(s_1^{(2)}+s_1\left(\frac{\rho_1 s}{1-\rho}+s_2\right)\left(1+\frac{\lambda_1 b_2}{1-\rho_2}\right)\right)\right)+\lambda_2 s_1^{(2)}+\lambda_2\left(\frac{\rho_1}{1-\rho_1}\right)^2\times$$
$$\times\left(s_1^{(2)}+s_1^2\lambda_1^2\left(\rho_1\frac{s}{1-\rho}+s_2\right)^2+\frac{\lambda_2 s}{1-\rho}\left(\frac{2b_2s_1}{1-\rho_2}+\frac{b_2^{(2)}}{(1-\rho_2)^3}\right)+\right.$$
$$\left.+\frac{\rho_2}{1-\rho_2}\left(s_2^{(2)}+s_2\left(\frac{\rho_2 s}{1-\rho}+s_1\right)\left(1+\frac{\lambda_2 b_1}{1-\rho_1}\right)\right)+s_2^{(2)}\right).$$

Лемма 2. *Среднее время ожидания в очереди для модели СМО типа $M_2|G_2|1$ с поллингом и с исчерпывающей дисциплиной обслуживания очередей определяется формулами:*

$$M[W_1]=\frac{\lambda_1^2 b_1^{(2)}}{2(1-\rho_1)}+\frac{1}{2\lambda_1(1-\rho_1)C}\left(\lambda_1 s_1^{(2)}+\lambda_1^2\left(2s_1s_2+s_1\rho_2 C(2-\rho_2)\right)\right)+$$
$$+\frac{1}{2\lambda_1(1-\rho_1)C}\left(\lambda_1^2\left(\frac{2\rho_2}{(1-\rho)(1-\rho_2)}\left(\rho_1(1-\rho_2)\left(\rho_2 s_1 C-s_2^{(2)}-s^2\right)\right)\right)\right)+$$
$$+\frac{1}{2\lambda_1(1-\rho_1)C}\left(s_2\rho_1\left(\frac{s_2}{1-\rho_1}+C\right)+s_2^{(2)}\right)+$$
$$+\frac{1}{2\lambda_1(1-\rho_1)C}\left(\lambda_1^2\left(\frac{2\rho_2}{(1-\rho)(1-\rho_2)}\left(\rho_1(1-\rho_2)\left(s_2C(1-\rho_1)\right)\right)\right)\right)+$$

$$+\frac{1}{2\lambda_1(1-\rho_1)C}\left(\lambda_1^2\left(\frac{2\rho_2}{(1-\rho)(1-\rho_2)}\left((\rho-\rho_1\rho_2-1)\left(s_1^{(2)}+s_1^2\right)\right)\right)\right),$$

$$M[W_2]=\frac{\lambda_2^2 b_2^{(2)}}{2(1-\rho_2)}+\frac{1}{2\lambda_2(1-\rho_2)C}\left(\lambda_2 s_2^{(2)}+\lambda_2^2\left(2s_1s_2+s_2\rho_1C(2-\rho_1)\right)\right)+$$
$$+\frac{1}{2\lambda_2(1-\rho_2)C}\left(\lambda_2^2\left(\frac{2\rho_1}{(1-\rho)(1-\rho_1)}\left(\rho_2(1-\rho_1)\left(\rho_1\, s_2C-s_1^{(2)}-s^2\right)\right)\right)\right)+$$
$$+\frac{1}{2\lambda_2(1-\rho_2)C}\left(s_1\rho_2\left(\frac{s_1}{1-\rho_2}+C\right)+s_1^{(2)}\right)+$$
$$+\frac{1}{2\lambda_2(1-\rho_2)C}\left(\lambda_2^2\left(\frac{2\rho_1}{(1-\rho)(1-\rho_1)}\left(\rho_2(1-\rho_1)\left(s_1C(1-\rho_2)\right)\right)\right)\right)+$$
$$+\frac{1}{2\lambda_1(1-\rho_1)C}\left(\lambda_1^2\left(\frac{2\rho_2}{(1-\rho)(1-\rho_2)}\left((\rho-\rho_1\rho_2-1)\left(s_2^{(2)}+s_2^2\right)\right)\right)\right).$$

Перейдём теперь к численному анализу шлюзовой и исчерпывающей дисциплин обслуживания.

Для сравнения моделей SIP сервера со шлюзовой и с исчерпывающей дисциплинами обслуживания очередей проведён численный анализ. В обеих моделях время обслуживания заявок выбрано постоянным [7] $b_1 = 0,01c$ и $b_1 = 0,004c$. Время переключения сервера к каждой из очередей является также постоянным, причём $s_1 = s_2 = 0,002c$. На рис. 2 показаны графики среднего времени ожидания обслуживания i-заявок в очереди в зависимости от величины предложенной нагрузки $\rho = \rho_1 + \rho_2$ для моделей с шлюзовой и с исчерпывающей дисциплинами обслуживания очередей.

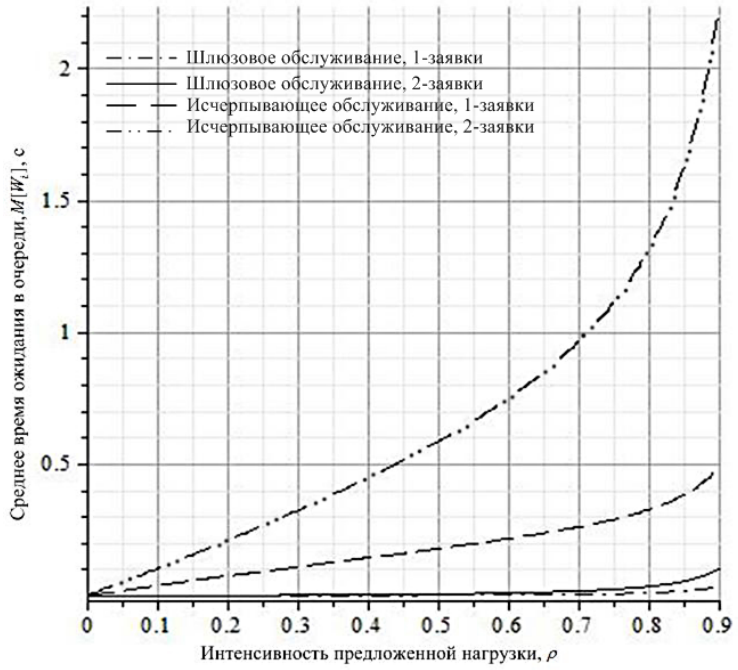

Рис. 2. Среднее время ожидания в очереди в СМО $M_2|D_2|1$ со шлюзовой и с исчерпывающей дисциплинами обслуживания очередей

Принято, что исследуется типовая процедура установления соединения по протоколу SIP, поэтому, как и в [7], принято, что $\rho_2 = 6\rho_1$. Как видно из графиков,

для указанных исходных данных шлюзовая дисциплина обслуживания гарантирует меньшее время пребывания заявок в очереди по сравнению с исчерпывающей дисциплиной. К тому же при увеличении нагрузки среднее время ожидания в очереди для модели со шлюзовым обслуживанием увеличивается медленнее, чем для модели с исчерпывающим обслуживанием.

4. Заключение

Итак, численный анализ с использованием полученных в статье аналитических формул показал, что модель поллинга со шлюзовой дисциплиной обслуживания очередей по критерию времени ожидания в очереди является предпочтительной по сравнению с исчерпывающей дисциплиной. Однако вывод о том, что следует отказаться от применения исчерпывающей дисциплины обслуживания в SIP-сервере, является преждевременным. В дальнейшем необходимо рассмотреть поллинговые модели обслуживания очередей с возможностью применения порогового управления входящей нагрузкой так, как это было предложено в работах [8–11].

Литература

1. *Rosenberg J., Schulzrinne H., Camarillo G.* RFC 3261. SIP: Session Initiation Protocol. — 2002.
2. *Rosenberg J.* RFC 5390. Requirements for Management of Overload in the Session Initiation Protocol. — 2008.
3. *Shen C., Schulzrinne H., Nahum E.* Session Initiation Protocol (SIP) Server Overload Control: Design and Evaluation // Lecture Notes in Computer Science. — Berlin: Springer-Verlag, 2008. — Vol. 5310. — Pp. 149–173.
4. Queueing Strategies for Local Overload Control in SIP Server / R. G. Garroppo, S. Giordano, S. Spagna, S. Niccolini // Global Telecommunications Conference. — Dept. of Inf. Eng., Univ. of Pisa, Pisa, Italy: IEEE, 2009. — 1-6 p.
5. *Самуйлов К. Е., Зарипова Э. Р.* Модель локального механизма контроля перегрузок SIP-сервера // T-COMM. — 2012. — С. 185–187. [Samouylov K. E., Zaripova E. R. Modelling Local Overload Control Mechanism of SIP Server // T-COMM. № 7. — 2012. — P. 185–187]
6. *Вишневский В. М., Семенова О. В.* Системы поллинга: теория и применение в широкополосных беспроводных сетях. — М.: Техносфера, 2007. — 312 с. [Vishnevskiy V.M., Semenova O. V. Polling systems: Theory and Applications in Broadband Wireless Networks. — M.: Tekhnosfera, 2007. — 312 p.]
7. Simulation of Overload Control in SIP Server Networks / P. O. Abaev, Y. V. Gaidamaka, A. V. Pechinkin et al. // Proc. of the 26th European Conference on Modelling and Simulation ECMS. — Koblenz: 2012. — Pp. 533–539.
8. *Abaev P., Gaidamaka Y., Samouylov K.* Modeling of Hysteretic Signalling Load Control in Next Generation Networks // Proceedings 12th International Conference, NEW2AN 2012, and 5th Conference, ruSMART 2012. — St. Petersburg: 2012. — Pp. 440–452.
9. *Abaev P., Gaidamaka Y., Samouylov K.* Queuing Model for Loss-Based Overload Control in a SIP Server Using a Hysteretic Technique // Proceedings 12th International Conference, NEW2AN 2012, and 5th Conference, ruSMART 2012. — St. Petersburg: 2012. — Pp. 371–378.
10. *Gaidamaka Y., Samouylov K., Sopin E.* Analysis of M|G|1 queue with hysteretic load control // XXX International Seminar on Stability Problems for Stochastic Models, the Autumn Session of the VI International Seminar on Applied Problems of Probability Theory and Mathematical Statistics related to Modeling of Information Systems. — Svetlogorsk: 2012. — Pp. 87–89.

11. *Basharin G., Gaidamaka Y., Samouylov K.* Actual Tasks of Mathematical Theory of Teletraffic // International Seminar "Stochastic Models and Statistical Inference" dedicated to the 100th Anniversary of Boris V. Gnedenko 17–20 October 2012. — Riga: 2012. — Pp. 104–105.

UDC 621.39

Polling Model of a SIP Server with Exhaustive and Gated Service Disciplines

Y. V. Gaidamaka, E. R. Zaripova

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

This paper investigates characteristics of the polling cycling models with exhaustive and gated service disciplines. We have derived mean queuing delays for suggested models and evaluated results.

Key words and phrases: polling system, SIP-server, exhaustive service discipline, gated service discipline, mathematical model.