У.Б. Амирсаидов

Ташкентский университет информационных технологий, г.Ташкент

УДК 621.39

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ УПРАВЛЕНИЯ ПЕРЕГРУЗКОЙ В СЕТИ IMS

Аннотация. В статье рассмотрены механизмы контроля и устранения перегрузок в серверах IMS, предложены метод и математическая модель управления перегрузкой в сети IMS.

Ключевые слова: управление потоками, перегрузка, производительность серверов, SIP-сервер, механизмы управления перегрузкой, облачная система, виртуальная машина.

Введение

Платформа IMS(Internet Multimedia Subsystem), является развитием архитектуры сети NGN (Next Generation Network), обеспечивающей расширение возможностей сети связи в реализации разнообразных сервисов на базе отделения функций уровня инфраструктуры от уровня сервиса. Введение новых элементов(серверов) сети (функция управления вызовами и сессиями CSCF - Call Session Control Function), которые с одной стороны, управляют соединением, а с другой- взаимодействуют с серверами предоставления услуг по протоколу SIP (Session Initiation Protocol), упрощает введение новых услуг и позволяет оператором связи повысить конкурентоспособность.

Производительность серверов IMS определяется количеством вызовов, которые способен обслужить сервер в единицу времени. Когда интенсивность вызовов превышает пропускную способность сервера, он переходит в режим перегрузки, время ожидания обслуживания на сервере увеличивается. Клиент ждет ответа на запрос и, не дождавшись, отправляет

повторный запрос на установление соединения. Это приводит к еще большему увеличению входной нагрузки на сервер и усугубляет перегрузку.

Перегрузка в сети IMS может быть вызвана различными причинами. Это может быть результатом активности пользователей, пытающихся установить соединения примерно в одно и то же время. Другая распространенная причина перегрузки – отказ одного из элементов в кластере SIP серверов, что снижает общую производительность и требует распределения входящей нагрузки между остальными серверами кластера.

Проблема перегрузок SIP-серверов возникает не только вследствие поведения пользователей в часы наивысшей нагрузки, но и в результате предоставления некоторых мультимедийных услуг, которые существенно меняют характер сигнального трафика. Еще одна причина перегрузки, когда оконечные терминалы после сбоя в сети пытаются одновременно зарегистрироваться.

Как показали многочисленные исследования [1-3], протоколе SIP имеются недоработки в механизме контроля перегрузок, в соответствии с которым в случае перегрузки прокси-сервера предусмотрена отправка сообщения 503 Service Unavailable. В частности, не решены следующие проблемы:

- проблема усугубления перегрузки, которая заключается в тенденции значительного увеличения нагрузки в период перегрузки;
- проблема неполного использования кластера серверов.

В условиях перегрузки серверов SIP необходимо применять механизмы управления нагрузкой: пороговое и приоритетное управление.

Различают локальное, межузловое и сквозное управление нагрузкой [2]. Одним из самых простых инструментов предотвращения перегрузок в SIP-серверах является механизм локального управления поступающей нагрузки на основе порогов длины очереди [1]. При локальном контроле перегрузок SIP-сервера, отклоненные сервером в режиме перегрузки сообщения сбрасываются, а при межузловом механизме контроля перегрузок

- для каждого отклоненного сервером в режиме перегрузки сообщения на вышележащий сервер отправляется уведомление о невозможности обслужить сообщение. Очевидно, что межузловой механизм контроля перегрузок позволяет более эффективно разгрузить сервер, поскольку при сбросе сервером сообщения без уведомления вышележащего сервера, на последнем применение различных таймеров приводит к множественным ретрансляциям сброшенных сообщений на нижележащий сервер, находящийся в режиме перегрузки, что усугубляет ситуацию.

Работа над новыми механизмами управления перегрузками была поручена рабочей группе SOC (SIP Overload Control) комитета IETF [3]. Текущая работа группы сконцентрирована на двух межузловых схемах контроля перегрузки:

- схема со сбросом сообщений на стороне отправителя (LBOC- Loss-Based Overload Control);

- схема с ограничением скорости потока сигнальных сообщений (RBOC-Rate-Based Overload Control).

Основная идея LBOC схемы заключается в том, что отправитель по запросу получателя уменьшает число отправляемых сообщений на указанный получателем процент от общего числа сообщений. Принцип работы RBOC схемы основывается на указании отправителю максимального числа сообщений, которые получатель хотел бы принять от отправителя в течение указанного им интервала времени. Схема LBOC призвана заменить существующий базовый механизм контроля перегрузки и со временем должна быть интегрирована в существующую версию протокола SIP.

Цель работы

В настоящее время все большую популярность приобретает облачная или виртуализированная архитектура IMS (vIMS). В «облачных» SIP-серверах появляется возможность управлять перегрузками не только механизмами LBOC и RBOC, но и путем изменения количества виртуальных машин. Исходя из этого, в данной работе разрабатывается математическая

модель управления перегрузками в «облачных» SIP-серверах на основе ограничения интенсивности входящего потока запросов (LBOC) и (или) увеличения количество виртуальных машин, обслуживающих заявки.

Математическая модель управления перегрузками

Процесс обслуживания запроса на установление соединения (INVITE) в подсистеме IMS является многоэтапным, в нем участвуют несколько серверов IMS. Каждый сервер в многоэтапном процессе обслуживания запроса для вышестоящего сервера является клиентом, т.е. $i-1$ сервер является клиентом i -го сервера, а i-ый сервер является клиентом $i+1$ сервера.

Рассмотрим функционирование i –го сервера как управляемую систему массового обслуживания. Управляемыми параметрами являются интенсивность поступления запросов (заявок, сообщений) λ и количество включенных виртуальных машин v, обслуживающих запросы с интенсивностью μ. Количество виртуальных машин v может изменяться от 1 до максимального значения V, в зависимости от числа запросов в системе $n(t)$ в момент времени $t>0$, $0 \le n(t) \le N$, $N = L + V$, где L – максимальная длина очереди.

В начальный момент сервер содержит одну виртуальную машину, остальные виртуальные машины могут быть включены при достижении числа запросов в системе порогового значения M_j, $j = \overline{1,V}$. Расстояние между порогами равно:

$$m = \left\lfloor \frac{N}{V} \right\rfloor, \quad (1)$$

Где: $\lfloor \quad \rfloor$- знак операции округления числа в меньшую сторону.

Если количество запросов в системе $n(t) \le M_1$, то параметры управления не изменяются:

$$\lambda_1 = \lambda, \ \mu_1 = \mu. \quad (2)$$

Если $n(t)$ увеличивается и переходит порог M_1, то повышается интенсивность обслуживания путем увеличения количества виртуальных машин на единицу, при условии отсутствия перегрузки $i+1$-го сервера:

$$\mu_2 = 2\mu(1-P) + P, \tag{3}$$

где: P - вероятность перегрузки $i+1-$сервера.

Если $i+1$- сервер перегружен, то интенсивность поступления запросов от $i-1$ - го сервера снижается:

$$\lambda_2 = P\lambda(1-q_1) + (1-P)\lambda, \tag{4}$$

где: q_1- доля снижения интенсивности входящего потока, когда $M_1 < n(t) \le M_2,\ 0 < q_1 \le 1$.

Таким образом, когда $n(t) > M_i$ $(i = \overline{2,V})$ интенсивности поступления и обслуживания запросов изменяются по формулам:

$$\mu_i = i\mu(1-P) + P\mu, \tag{5}$$

$$\lambda_i = P\lambda(1-q_i) + (1-P)\lambda, \tag{6}$$

где: $q_1 < q_2 < q_3 \ldots . < q_v$.

Система уравнений равновесия для рассматриваемого виртуального сервера имеет вид:

$$\begin{cases} -P_k(\lambda_i + \mu_i) + P_{k-1}\lambda_i + P_{k+1}\mu_i = 0, \quad k = \overline{M_{i-1}+1, M_i - 1}\ , i = \overline{1,V}, \\ -P_k(\lambda_{i+1} + \mu_i) + P_{k-1}\lambda_i + P_{k+1}\mu_{i+1} = 0, \quad k = M_i\ , i = \overline{1,V-1}\ , \\ P_{N-1}\lambda_v - P_N\mu_v = 0, \end{cases} \tag{7}$$

где: P_k- вероятность того, что система находится в -ом состоянии, т.е. в системе имеются k-запросов.

Решая систему уравнений, находим:

$$P_k = \begin{cases} P_0\left(\frac{\lambda_1}{\mu_1}\right)^k, \ k = \overline{1,m} \\ P_o\left(\frac{\lambda_{\lfloor k/m \rfloor + 1}}{\mu_{\lfloor k/m \rfloor + 1}}\right)^{mod(k,m)} \prod_{i=1}^{\lfloor k/m \rfloor}\left(\frac{\lambda_i}{\mu_i}\right)^m, \ k = \overline{(m+1),N}, \end{cases} \tag{8}$$

где: $mod(k,m)$- остаток от деления k на m.

Из условия нормировки $\sum_{k=0}^{N} P_k = 1$ определяем:

$$P_0 = \left[1 + \sum_{k=1}^{m} \left(\frac{\lambda_1}{\mu_1}\right)^k + \sum_{k=m+1}^{N} \left(\frac{\lambda_{\lfloor k/m \rfloor + 1}}{\mu_{\lfloor k/m \rfloor + 1}}\right)^{mod(k,m)} \prod_{i=1}^{\lfloor k/m \rfloor} \left(\frac{\lambda_i}{\mu_i}\right)^m \right]^{-1} \quad (9)$$

Зная стационарные вероятности состояний (8,9), можно оценить вероятностно-временные характеристики исследуемого виртуального SIP-сервера, формулы для расчета которых приведены ниже.

Среднее число запросов в SIP-сервере:

$$\bar{N} = \sum_{k=0}^{N} k P_k. \quad (10)$$

Вероятность потери запросов (блокировки системы):

$$P_{loss} = P_N. \quad (11)$$

Среднее время задержки (пребывания) запросов в сервере:

$$\bar{T} = \frac{\bar{N}}{\lambda(1-P_{loss})}. \quad (12)$$

Пример численного анализа

Проведен расчет вероятностно-временных характеристик виртуального сервера при следующих исходных данных:

$N = 50, m = 10, V = 5, \mu = 1$ мс$^{-1}$, $P = 0.5$, $q_1 = 0.2$, $q_2 = 0.4$, $q_3 = 0.6$, q_4=0.8. Графики зависимости среднего времени задержки от нагрузки $(\rho = \lambda/\mu)$ приведены на рисунке 1. Для сравнительного анализа на этом рисунке 1 пунктирной линией показана характеристика сервера с механизмом LBOC.

Из рисунка видно, что преимущество предложенного метода управления перегрузкой наблюдается в областях высокой нагрузки.

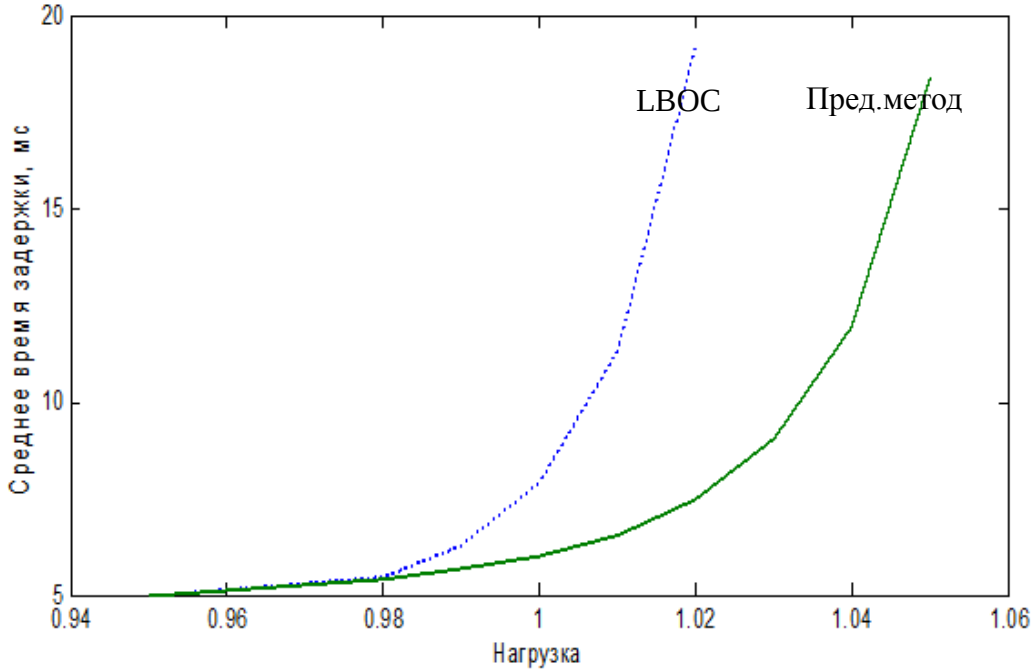

Рис.1 Графики зависимости среднего времени задержки запросов от нагрузки

Заключение

Как показано в работе, предложенный метод обеспечивает снижение среднего время задержки запросов от 1.2 раза до 5 раз в областях высокой нагрузки по сравнению с механизмом LBOC. Это обосновывает целесообразность проведения исследований в направлении совершенствования подходов и методов управления потоками в современных сетях.

СПИСОК ЛИТЕРАТУРЫ

1. Э.С. Сопин. Модели серверов подсистемы IMS с групповым поступлением заявок. XII Всероссийское совещание по проблемам управления, Москва, 2014, с.8735-8742.
2. Самуйлов К.Е., Зарипова З.Р. Модель локального механизма контроля перегрузок SIP-сервера. T-Comm, №7, 2012, с.185-187.
3. Самуйлов К.Е., Абаев П.О., Гайдамака Ю.В. и др. Аналитические и имитационные модели для оценки показателей функционирования SIP серверов в условиях перегрузок T!Comm №8, 2014, с.83-88.