УДК 519.71

МОДЕЛЬ ПРОЦЕССА ОБРАБОТКИ ЗАПРОСОВ СИСТЕМОЙ УПРАВЛЕНИЯ WEB-САЙТАМИ

М. Ю. Быков[1]

Системы управления сайтами, получившие широкое распространение при построении сложных Web-систем, предоставляют разнообразные функциональные возможности, однако, являясь дополнительной надстройкой над Web-сервером, требуют и дополнительных затрат процессорного времени. Для того чтобы спроектировать программно-аппаратную конфигурацию для конкретного сервера либо сформулировать требования для системы управления сайтами, необходима математическая модель процесса обработки запросов такой системой. В статье рассматривается математическая модель данного типа, основанная на принципах теории массового обслуживания.

Ключевые слова: модель, система управления сайтами, система массового обслуживания, Web-сервер, теория массового обслуживания.

1. Введение. Применение систем управления сайтами (СУС) позволяет значительно сэкономить усилия при разработке Web-систем, повысить повторное использование кода, сократить количество ошибок в коде. Однако применение СУС накладывает ограничения на производительность, поскольку использует дополнительное процессорное время сервера. Для оценки эффективности СУС необходимо рассмотреть математическую модель системы, выявить факторы, влияющие на производительность, и определить основные параметры системы.

2. Формализация процесса обработки запросов системой управления сайтами. При высокой нагрузке системы физические ресурсы компьютера могут быть исчерпаны и не все запросы к СУС будут обработаны. Такая ситуация является крайне нежелательной и должна быть проанализирована с целью выявления возможностей снижения вероятности ее появления. Система обработки запросов с математической точки зрения является системой массового обслуживания (СМО) [1, 2]. Модели СМО широко применяются для расчета производительности компьютерных систем обработки данных [3]. На рис. 1 представлена модель обработки заявок Web-сайтом.

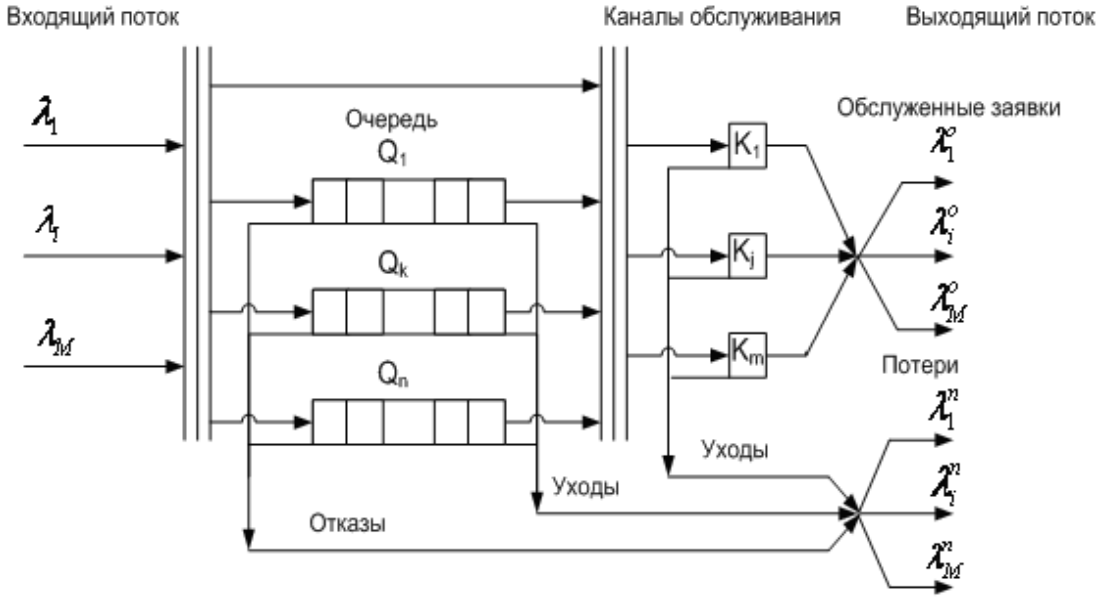

Рис. 1. Модель обслуживания заявок Web-сайтом

[1] Московский государственный институт электронной техники, г. Зеленоград, проезд 4806, д. 5, 124498, Москва; e-mail: mikhail.bykov@gmail.com

На вход подаются заявки различных типов — запросы на статические ресурсы, статические страницы, динамические страницы. Эти запросы обрабатываются свободными каналами обслуживания (потоками сервера). При отсутствии свободных каналов заявка помещается в очередь. При отсутствии мест в очереди происходит отказ в принятии заявки. У заявки есть максимально допустимое время пребывания в системе, при исчерпании которого она уходит либо из очереди, либо из канала обслуживания. Все ушедшие заявки и заявки, в обработке которых было отказано, считаются потерями. Задача настройки конфигурации Web-сервера состоит в минимизации потерь. Для каждого сервера эта задача имеет свои особенности, поскольку любая Web-система имеет свою специфику, поэтому в общем виде задача минимизации потерь сервера очень сложна. Однако система управления сайтами представляет собой более простую систему массового обслуживания. Модель такой системы представлена на рис. 2. Основное отличие системы управления сайтами состоит в том, что на вход системы подаются однотипные заявки, т.е. заявки на динамические страницы, поскольку остальные запросы Web-сервер обрабатывает самостоятельно. Таким образом, модель, а вместе с ней и задача минимизации потерь и повышения производительности, упрощаются.

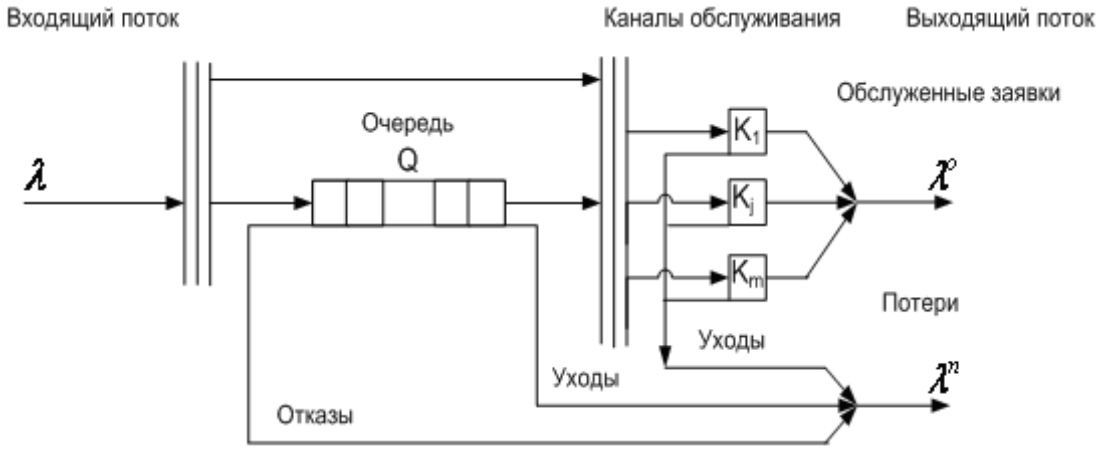

Рис. 2. Модель обработки запросов системой управления сайтами

Для определения того, какая модель СМО подходит для описания обработки запросов системой управления сайтами, рассмотрим физические свойства системы.

На вход системы приходят запросы пользователей в случайной последовательности. Поток запросов при рассмотрении в течение достаточно короткого периода времени (на практике это время может быть от одного часа до любого периода в случае интернациональных сайтов) является так называемым простейшим потоком, т.е. с характеристиками, не зависящими от времени, с событиями, происходящими поодиночке, и с интервалом времени от начала отсчета до наступления первого события, представляющим собой непрерывную случайную величину, распределенную по экспоненциальному закону. Систему можно рассматривать как систему, имеющую m однотипных каналов обслуживания (максимальное количество потоков обработки запросов) и характеризующуюся экспоненциальным распределением времени обслуживания со средним значением $\overline{\tau}_{\rm ob}$ или, что эквивалентно, являющуюся простейшим потоком обслуживаний с интенсивностью $\mu = 1/\overline{\tau}_{\rm ob}$ независимо от типа обслуживаемой заявки. Такое рассмотрение возможно, поскольку, как было указано выше, к системе приходят однотипные запросы страниц.

При полностью загруженных каналах обслуживания заявки могут ждать обслуживания в общей очереди, число мест в которой равно n. Дисциплина ожидания — FIFO (First In, First Out), т.е. первыми из очереди подаются заявки, первыми попавшие в очередь, приоритета заявок нет, они становятся в очередь в порядке поступления, при переполнении очереди вновь поступившая заявка получает отказ. Дисциплина обслуживания также FIFO, выбор заявки из очереди при освобождении какого-либо из каналов обслуживания делается из начала очереди. Поскольку заявки однотипны, рассматривается единственный входящий поток, который будет простейшим с интенсивностью λ.

Заявки в данной системе являются нетерпеливыми, т.е. имеющими право находиться в СМО не более $\tau_{\rm dop}$ единиц времени. Это время является максимальным временем обслуживания запроса и устанавливается на Web-сервере. Если время пребывания заявки в системе t_c превышает $\tau_{\rm dop}$, заявка покидает систему и считается потерянной для системы. Величина $\tau_{\rm dop}$ является постоянной и фиксированной.

Удобным для дальнейшего рассмотрения является представление о простейшем потоке уходов из СМО с интенсивностью $\eta = 1/\tau_{\rm dop}$. Уходы заявки возможны либо из очереди, если $t_{\rm ozh} > \tau_{\rm dop}$, либо из

канала обслуживания, если $t_{\text{ozh}} \leqslant \tau_{\text{dop}} \leqslant t_c$. Методически удобно рассматривать два потока уходов с интенсивностями, соответственно, η_{ozh} и η_{ob}: $\eta_{\text{ozh}} = \eta_{\text{ob}} = \eta = 1/\tau_{\text{dop}}$.

Данная модель обработки запросов является одним из классических вариантов СМО, называемой "разомкнутой СМО с ожиданием". Основные показатели эффективности такой СМО определяются теорией непрерывных Марковских цепей и приводятся ниже.

При приведении интенсивности всех потоков к интенсивности потока обслуживаний имеем

$\rho = \dfrac{\lambda}{\mu}$ — приведенная интенсивность входящего потока, представляющая собой среднее число заявок, поступающих на вход СМО за среднее время обслуживания одной заявки;

$\alpha_{\text{ozh}} = \dfrac{\eta_{\text{ozh}}}{\mu}$ — приведенная интенсивность потока уходов из очереди;

$\alpha_{\text{ob}} = \dfrac{\eta_{\text{ob}}}{\mu}$ — приведенная интенсивность потока уходов из канала обслуживания.

Вероятность P_0 (нет заявок в системе) равна

$$P_0 = \left[1 + \sum_{i=1}^{m} \frac{\rho^i}{i!(1+\alpha_{\text{ob}})^i} + \frac{\rho^m}{m!(1+\alpha_{\text{ob}})^m} \sum_{i=1}^{n} \prod_{j=1}^{l} \frac{\rho}{m(1+\alpha_{\text{ob}}) + j\alpha_{\text{ozh}}}\right]^{-1}.$$

Вероятность нахождения системы в одном из состояний P_i, $i = \overline{1,m}$, т.е. при нуле заявок в очереди и частичной загрузке каналов, равна $P_i = \dfrac{\rho^i}{i!(1+\alpha_{\text{ob}})^i} P_0$.

При полной загрузке каналов и наличии заявок в очереди ожидания

$$P_{m+1} = \frac{\rho^m}{m!(1+\alpha_{\text{ob}})^m} \prod_{j=1}^{l} \frac{\rho}{m(1+\alpha_{\text{ob}}) + j\alpha_{\text{ozh}}} P_0, \quad l = \overline{1,n}.$$

Одним из важных показателей эффективности является среднее число каналов, занятых обслуживанием: $\overline{K} = \sum\limits_{i=1}^{m} iP_i + m\Big(1 - \sum\limits_{i=0}^{m} P_i\Big)$. При этом средняя длина очереди $\overline{l}$ равна $\overline{l} = \sum\limits_{l=1}^{n} lP_{m+l}$. Среднее число заявок в СМО равно $\overline{Z} = \overline{K} + \overline{l}$.

В рассматриваемой СМО потери заявок возможны либо в форме отказа вследствие переполнения системы, либо в форме ухода нетерпеливых заявок из системы. Справедливы следующие формулы:

$P_{\text{otk}} = \dfrac{\rho^{m+n}}{m!(1+\alpha_{\text{ob}})^m \prod\limits_{j=1}^{n} \big(m(1+\alpha_{\text{ob}}) + j\alpha_{\text{ozh}}\big)} P_0$ — вероятность отказа,

$P_{\text{y}}^{\text{ozh}} = \dfrac{\overline{l}\,\eta_{\text{ozh}}}{\lambda}$ — вероятность ухода заявки во время ожидания,

$P_{\text{y}}^{\text{ob}} = \dfrac{\overline{K}\eta_{\text{ob}}}{\lambda}$ — вероятность ухода заявки во время обслуживания.

Общая вероятность ухода заявки из системы вследствие окончания времени ожидания или отказа системы принять заявку равна $P_{\text{p}} = P_{\text{otk}} + P_{\text{y}}^{\text{ozh}} + P_{\text{y}}^{\text{ob}}$.

3. Расчет параметров эффективности. Приведенные выше выражения позволяют рассчитать параметры эффективности системы при наличии данных о системе (потоке, пропускной способности, среднем времени обслуживания). Из всех параметров, от которых зависит эффективность системы, при разработке системы управления сайтами можно влиять только на среднее время обслуживания $\overline{\tau}_{\text{ob}}$. Остальные параметры либо являются внешними, например входной поток заявок, либо зависят от сервера системы и его настроек (максимальное количество каналов, длина очереди, максимальное время ожидания τ_{dop}). Настройка параметров сервера является индивидуальной для Web-системы и представляет собой сложную техническую задачу. Возможные значения настроек для различных систем и приемы настройки сервера обсуждаются в [4, 5]. Рассмотрим наиболее важный параметр — вероятность ухода заявки из системы, т.е. сбоя в зависимости от $\overline{\tau}_{\text{ob}}$, используя значения констант настройки Web-сервера IIS (Internet Information Server) по умолчанию. Значения параметров при этом следующие (данные компании Microsoft [4]):

m — количество каналов обслуживания; по умолчанию в сервере IIS эквивалент этого значения — максимальное количество потоков обслуживания — равен 10;

n — максимальная длина очереди; по умолчанию в сервере IIS эквивалент этого значения — максимальная длина очереди соединений — равен 15;

τ_{dop} — максимально допустимое время пребывания заявки в системе; по умолчанию в сервере IIS эквивалент этого значения — максимальное время выполнения сценария — равен 90 секундам;

λ — в качестве интенсивности входящего потока возьмем обычное пиковое значение количества запросов на поисковой машине Rambler, равное 100 запросам в секунду; такое значение эквивалентно системе с высокой нагрузкой.

При определенных выше значениях параметров мы можем проследить зависимость P_{p} от $\overline{\tau}_{\text{ob}}$. График зависимости приведен на рис. 3.

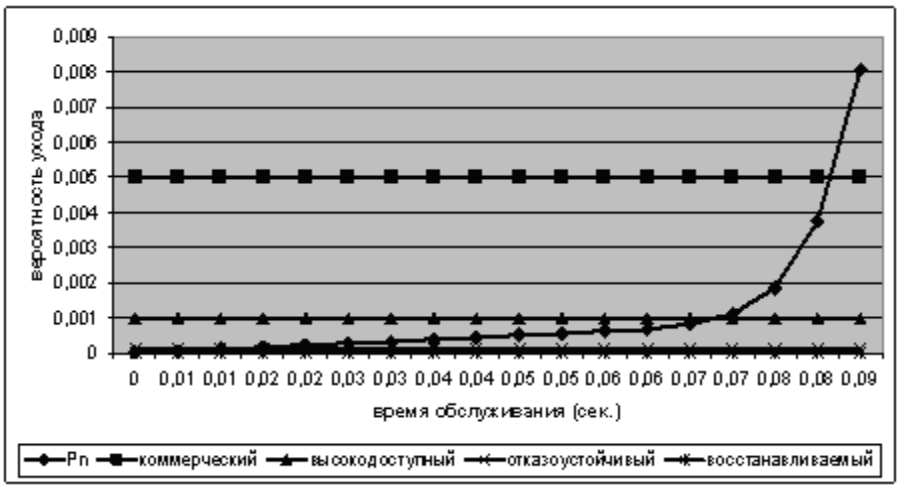

Рис. 3. Зависимость вероятности ухода заявки от времени обслуживания

Для определения надежности системы необходимо сравнить вероятность отказа с принятыми в индустрии критериями надежности. Возьмем для сравнения критерии надежности Web-систем, приводимые компанией DELL [6]. Значения этих критериев даны в таблице и отмечены линиями с маркерами разных типов на рис. 3 для расчетной системы. Кроме того, в таблице указано максимальное время обработки запроса для описанной конфигурации, соответствующее каждому уровню надежности. Таким образом, мы получили время, которое может быть оценено после реализации системы и тестирования для оценки ее надежности.

Уровень доступности	Доступность	Время обработки, с
Коммерческий	99,5 %	0,08187
Высокодоступный	99,9 %	0,0687
Отказоустойчивый	99,99 %	0,009
Восстанавливаемый	99,999 %	0,0009
Постоянный	100 %	—

Полученные формулы и практические результаты могут быть использованы для широкого ряда Web-систем и СУС вне зависимости от применяемой технологии реализации или архитектуры.

СПИСОК ЛИТЕРАТУРЫ

1. *Ивницкий В.А.* Теория сетей массового обслуживания. М.: Физматлит, 2004.
2. *Клейнрок Л.* Теория массового обслуживания. М.: Машиностроение, 1979.
3. *Nelson Randolph.* Probability, stochastic processes, and queuing theory: the mathematics of computer performance modeling. Berlin: Springer-Verlag, 1995.
4. *Curry B., Kaldestad H., Reilly G.* The art and science of Web server tuning with Internet Information Services 5.0. Microsoft Corporation (http://www.microsoft.com/TechNet/prodtechnol/iis/iis5/maintain/optimize/ iis5tune.asp).
5. *Moore M.* Tuning Internet Information Server performance. Microsoft Corporation, 2003 (http://www.microsoft.com/serviceproviders/whitepapers/tuningiis.asp).
6. *Graham J.* Maximizing Web server availability. Dell Inc., 2002 (http://www1.us.dell.com/content/topics/global.aspx/power/en/ps1q02_graham).

Поступила в редакцию
11.05.2005