ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆԻ ԳԻՏԱԿԱՆ ՏԵՂԵԿԱԳԻՐ
УЧЕНЫЕ ЗАПИСКИ ЕРЕВАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

Բնական գիտություններ 1, 2007 Естественные науки

Математика

УДК 519.21

СУФИАН МОХАММЕД ДВЕЙДАРИ

ИССЛЕДОВАНИЕ ХАРАКТЕРИСТИК СЕТЕВОГО СЕРВЕРА С ДВУХЭТАПНЫМ ОБСЛУЖИВАНИЕМ ЗАПРОСОВ

Рассматривается очередь MAP$|G|1$ с двухэтапным обслуживанием. Исследованы длительность периода занятости модели, распределение очереди и вероятность свободного состояния очереди в нестационарном режиме. Анализ модели проводится методом введения дополнительной переменной.

Введение. В ряде сетей (интернет-сервисы, сетевые серверы, системы электронной коммерции, дистанционного обучения, поисковые и др.) обслуживание запросов пользователей производится поэтапно [1, 2]. Выделяют три функционально значимых этапа обработки запросов. Первый этап включает прием, регистрацию и идентификацию поступающих запросов. В результате запрос пользователя либо отклоняется, либо принимается на обслуживание. Характеристики этого этапа зависят от параметров поступающего на вход системы потока запросов, принципов их буферизации, дисциплин обслуживания. Второй этап включает обработку запросов. Запросы проходят по подсистемам, обрабатываются и преобразуются. Характеристики этапа зависят от типа и содержания запроса, специфики системы, состава и структуры подсистем, принципов организации вычислительного процесса. Результатом выполнения данного этапа является генерация ответного сообщения и/или выходного файла. Третий этап включает передачу-доставку по сетевым каналам ответного сообщения или файла пользователю. Характеристики этапа зависят от типа и параметров каналов передачи данных, методов разделения ресурсов, принципов организации буферизации выходных сообщений.

Как показывают измерения, характеристики различных сетей и сервисов зависят от характера и параметров входящего потока. С другой стороны, анализ циркулирующего в сетях трафика показал, что он отличается от пуассоновского потока, характеризуется сильной корреляцией между прибытиями пакетов, имеет модулированный, самоподобный, пульсирующий характер [2–5]. Учет характера реального сетевого трафика при расчете параметров и характеристик сетевых сервисов и систем потребовал разработки новых моделей, которые адекватно описывают реальные потоки пакетов в сетях и определяют соответствующие измерениям оценки характеристик. Для опи-

сания самоподобного потока пакетов применение находят модели MAP (Markovian Arrival Process), предложенные Ньютсом в [6]. MAP моделируют широкий спектр потоков, в том числе многие, используемые для моделирования компьютерных сетей, модели трафика.

Для анализа компьютерных сетей используются модели и методы теории очередей [6,7]. Разработан матрично-аналитический метод, который позволяет с помощью матричной экспоненты исследовать очереди с MAP-потоками [6, 8]. Модель MAP$|G|1$ с дисциплиной FIFO (first in–first out), с конечной и/или бесконечной очередью методом вложенных цепей Маркова (ВЦМ) исследована в [6, 8]. В преобразованиях Лапласа–Стилтьеса (ПЛС) получены соотношения для функций распределений (ФР) периода занятости (ПЗ) модели, времени ожидания и пребывания в модели, а также производящей функции (ПФ) длины очереди. Получены матричные обобщения формул Поллячека–Хинчина для времени ожидания и длины очереди. В [9] исследован переходный режим в очереди MAP$|G|1$, получены двумерные ПЛС для распределения длины очереди, а также времени ожидания при дисциплине FIFO.

В настоящей работе рассматривается очередь MAP$|G|1$ с дисциплиной FIFO и двухэтапным обслуживанием запросов.

Описание входящего потока. MAP задается марковским процессом (МП) $\xi(t)=(N(t),J(t))$ с множеством состояний $\{(i,j);\ i\ge 0, 1\le j\le m\}$, где $N(t)$ – количество поступивших за время t запросов, а $J(t)$ – управляющая процессом поступления цепь Маркова (ЦМ) с множеством состояний $(1,2,...,m)$. Инфинитезимальный генератор МП $\xi(t)$ Q имеет структуру

$$Q=\begin{pmatrix} C & D & 0 & \dots \\ 0 & C & D & \dots \\ 0 & 0 & C & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Диагональные и верхне-диагональные элементы Q равны соответственно матрицам $C=\|C_{ij}\|$ и $D=\|D_{ij}\|$ размерности $m\times m$. Матрица C имеет отрицательные диагональные элементы $C_{ii}<0$ и неотрицательные недиагональные элементы $C_{ij}\ge 0,\ i\ne j$. Матрица D имеет неотрицательные элементы $D_{ij}\ge 0$. Здесь $C_{ij},\ i\ne j$, – интенсивность перехода из состояния i в состояние j ЦМ $J(t)$ без генерации (поступления) запроса, а D_{ij} – интенсивность перехода ЦМ $J(t)$ из состояния i в состояние j с генерацией (поступлением) запроса. Компоненты матрицы Q удовлетворяют условию $(C+D)\boldsymbol{e}=0$, где $\boldsymbol{e}$ – вектор-столбец, все элементы которого равны 1.

Кроме того, $C+D$ является инфинитезимальным генератором ЦМ $J(t)$. Стационарное распределение ЦМ $J(t)$ определяется из уравнений $\pi(C+D)=0,\ \pi\boldsymbol{e}=1$. Для интенсивности трафика имеем $\lambda=\pi D\boldsymbol{e}$.

Пусть $P(n,t)$, $n\ge 0$, – матрица вероятностей переходов МП $\xi(t)$, элемент которой равен вероятности перехода МП из начального состояния

$(0,i)$ в состояние (n,j) за время t:

$$P_{ij}(n,t) = P\{N(t) = n, \quad J(t) = j / N(0) = 0, J(0) = i\}.$$

Справедливы следующие уравнения Колмогорова:

$$\frac{d}{dt} P(n,t) = P(n,t)C + P(n-1,t)D\,, \quad n \geq 1\,, \quad t \geq 0\,, \quad P(0,0) = 1\,, \qquad (1)$$

решение которых представимо через матричные экспоненциальные функции.

Пусть $\tilde{P}(z,t)$ есть ПФ МП $\xi(t)$: $\tilde{P}(z,t) = \sum_{n \geq 0} z^n P(n,t)\,, \; |z| \leq 1\,.$

Тогда из (1) для ПФ $\tilde{P}(z,t)$ получаем

$$\frac{d}{dt} \tilde{P}(z,t) = \tilde{P}(z,t)[C + zD]\,, \; |z| < 1, \quad t \geq 0\,.$$

Отсюда для $\tilde{P}(z,t)$ находим: $\tilde{P}(z,t) = e^{(C+zD)t}$, где $e^Q = \sum_{k=0}^{\infty} \frac{Q^k}{k!}$ – матричная экспонента [6, 7].

Процесс обслуживания. Запросы обслуживаются двухэтапно. Вначале они поступают на первый этап обслуживания, после которого с вероятностью q запрос уходит из модели, а с вероятностью $1-q$ поступает на второй этап. После второго этапа запросы уходят из модели. Длительности первого и второго этапов – независимые, одинаково распределенные (НОР) случайные величины v_1 и v_2 с ФР $B_1(t), B_2(t)$, плотностями $b_1(t), b_2(t)$ и средними значениями b_1, b_2. ФР полного времени обслуживания запроса $H(t)$ равна:

$$H(t) = qB_1(t) + (1-q)\int_0^t B_1(t-x)dB_2(x)\,.$$

Отсюда для ПЛС ФР $\tilde{H}(s)$ и среднего значения h_1 полного времени обслуживания получаем

$$\tilde{H}(s) = q\tilde{B}_1(s) + (1-q)\tilde{B}_1(s)\tilde{B}_2(s)\,, \; h_1 = qb_1 + (1-q)(b_1 + b_2) = b_1 + (1-q)b_2\,.$$

Период занятости. Пусть $G(t)$ – матрица размерности $m \times m$, элемент которой $G_{ij}(t)$ есть ФР ПЗ при условии, что в момент начала ПЗ входящий процесс находился в состоянии i, а в момент окончания – в состоянии j.

Для модели MAP|G|1 с дисциплиной FIFO известно [10], что при $\rho < 1$, где $\rho = \lambda h_1$ – загрузка модели, ПЛС ФР ПЗ модели $\tilde{G}(s)$ удовлетворяет функциональному уравнению

$$\tilde{G}(s) = \tilde{h}(sI - D[\tilde{G}(s)]) = \int_0^{\infty} e^{-sx} e^{D[G(s)]x} dH(x)\,, \qquad (2)$$

где $D[\tilde{G}(s)] = C + D \cdot \tilde{G}(s)$. Подставляя значение ПЛС ФР $\tilde{H}(s)$ в (2), для $\tilde{G}(s)$ получаем

$$\tilde{G}(s) = \beta_1(sI - D[\tilde{G}(s)])[q + (1-q)\beta_2(sI - D[\tilde{G}(s)])]\,.$$

$G = \tilde{G}(0)$ – стохастическая матрица, удовлетворяющая функциональному

уравнению $G = \int_0^\infty e^{D(G)x} dH(x) = \tilde{h}[-D(G)]$. Здесь $D(G) = C + D \cdot G$.

Обозначим через $\boldsymbol{g} = (g_1, ..., g_m)$ вектор стационарного распределения матрицы G, i-ый элемент которой g_i равен стационарной вероятности того, что в свободной модели входящий процесс находится в i-ом состоянии [11]. $\boldsymbol{g}$ определяем из уравнений Колмогорова

$$\boldsymbol{g} = \boldsymbol{g}G, \ \boldsymbol{g}\boldsymbol{e} = 1.$$

Длина очереди. Воспользуемся методом введения дополнительной переменной. Рассмотрим четырехкомпонентный случайный процесс $(L(t), J(t), I(t), x(t))$, где $L(t)$ – количество запросов в модели в момент t; $J(t)$ – состояние входящего процесса; $I(t)$ – индикатор состояния обслуживания, $I(t)$=1 – запрос находится на первом этапе, $I(t)$=2 – на втором этапе; $x(t) = 0$, если в момент t система свободна, а в противном случае $x(t)$ равен времени, которое прошло до момента t с начала обслуживания запроса, находящегося в момент t на первом или втором этапе. Пусть

$$\pi(n, i, j, x, t) = \frac{\partial}{\partial x} P(\xi(t) = n, J(t) = i, I(t) = j, x(t) < x),$$

$$\pi_i(n, j, t) = \int_0^\infty \pi(n, i, j, x, t) dx,$$

$$\pi(n, j, x, t) = (\pi_1(n, j, x, t), \pi_2(n, j, x, t), ..., \pi_m(n, j, x, t)),$$

$$\pi(n, x, t) = (\pi(n, 1, x, t), \pi(n, 2, x, t)), \ \pi(n, t) = (\pi_1(n, t), \pi_2(n, t), ..., \pi_m(n, t)).$$

Для $\pi(n,t)$ и $\pi(n,x,t)$ выполнены условия

$$\pi(n, t) = \int_0^\infty \pi(n, x, t) dx, \quad \sum_{j=1}^{2} \sum_{n=0}^{\infty} \pi(n, j, t) \boldsymbol{e} = 1, \quad 0 \le t.$$

Теорема. ПФ $\tilde{\pi}(z, j, 0, s)$, $\tilde{\pi}(z, j, x, s)$, $\tilde{\pi}(0, s)$ и $\tilde{\pi}(z, j, s)$ распределения количества запросов в модели определяются по формулам

$$\tilde{\pi}(z, j, x, s) = \tilde{\pi}(z, j, 0, s) e^{-[sI - D(z)]x} (1 - B_j(x)), \ \tilde{\pi}(0, s) = \vartheta_0 (sI - D(\tilde{G}(s)))^{-1}.$$

В частности,

$$\tilde{\pi}(z, 1, 0, s) = \vartheta_0 [(sI - D(\tilde{G}(s)))^{-1} (D(z) - sI) + I] \times$$
$$\times \left[I - \frac{1}{z} \tilde{B}_1 [sI - D(z)] \{q - (1 - q) \tilde{B}_2 [sI - D(z)]\} \right]^{-1}. \qquad (3)$$

Доказательство. Для векторов $\pi(n,x,t)$ составляем уравнения Колмогорова

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x} \right) \pi(n, j, x) = \pi(n, j, x, t)(C - \eta_j(x) I) + \pi(n - 1, j, x, t) D,$$

$$\frac{d}{dt} \pi(0, t) = \pi(0, t) C + \int_0^\infty \pi(1, 2, x, t) \eta_2(x) dx + q \int_0^\infty \pi(1, 1, x, t) \eta_1(x) dx,$$

где I – столбец из единиц. Запишем граничные условия

$$\pi(n,1,0,t) = q\int_0^\infty \pi(n+1,1,x,t)\eta_1(x)dx + \int_0^\infty \pi(n+1,2,x,t)\eta_2(x)dx\,,$$

$$\pi(n,2,0,t) = (1-q)\int_0^\infty \pi(n,1,x,t)\eta_1(x)dx\,,$$

(4)

$$\pi(1,1,0,t) = q\int_0^\infty \pi(2,1,x,t)\eta_1(x)dx + \int_0^\infty \pi(2,2,x,t)\eta_2(x)dx + \pi(0,t)D\,;$$

а также начальные условия

$$\pi(0,0) = \vartheta_0,\ \pi(n,i,x,0) = 0,\ 0 \le n,\ i = 1,2,\ 0 \le x\,.$$

Введем матричные ПФ и ПЛС:

$$\pi(z,i,x,t) = \sum_{n=1}^\infty z^n \pi(n,i,x,t),\quad \pi(z,i,0,t) = \sum_{n=1}^\infty z^n \pi(n,i,0,t),$$

$$\tilde\pi(z,i,x,s) = \int_0^\infty e^{-st}\pi(z,i,x,t)dt,\quad \tilde\pi(z,i,0,s) = \int_0^\infty e^{-st}\pi(z,i,0,t)dt.$$

Тогда из (4) для $\tilde\pi(z,x,s)$ и $\tilde\pi(z,0,s)$ получим

$$\frac{\partial}{\partial x}\tilde\pi(z,i,x,s) = \tilde\pi(z,i,x,s)[C + zD - (s - \eta_i(x))I],$$

$$\tilde\pi(0,s)[sI - C] = \vartheta_0 + \int_0^\infty \pi(1,2,x,s)\eta_2(x)dx + q\int_0^\infty \pi(1,1,x,s)\eta_1(x)dx\,,$$

$$\tilde\pi(z,1,0,s) = \frac{1}{z}q\int_0^\infty \tilde\pi(z,1,x,s)\eta_1(x)dx + \frac{1}{z}\int_0^\infty \tilde\pi(z,2,x,s)\eta_2(x)dx + \tilde\pi(0,s)(C + Dz - sI) + \vartheta_0.$$

Отсюда для $\tilde\pi(z,x,s)$ и $\tilde\pi(z,0,s)$ получаем

$$\begin{aligned}
&\tilde\pi(z,i,x,s) = \tilde\pi(z,i,0,s)e^{-[sI - D(z)]x}(1 - B_i(x)), \quad i = 1,2,\\
&\tilde\pi(z,1,0,s) = [\tilde\pi(0,s)(D(z) - sI) + \vartheta_0]\times\\
&\qquad\qquad \times\left[I - \frac{1}{z}\tilde B_1[sI - D(z)]\{q - (1-q)\tilde B_2[sI - D(z)]\}\right]^{-1},\\
&\tilde\pi(z,2,0,s) = (1-q)\tilde\pi(z,1,0,s)\tilde B_1[sI - D(z)],
\end{aligned} \tag{5}$$

где $D(z) = C + Dz$, $\tilde B(sI - D(z)) = \int_0^\infty e^{-[sI - D(z)]x}dB(x)$.

Для определения $\tilde\pi(0,s)$ заметим, что $\tilde G(s)$ – единственное решение функционального уравнения $z = \tilde B_1[sI - D(z)]\{q - (1-q)\tilde B_2[sI - D(z)]\}$, $|z| < 1$.

Из условия ограниченности ПФ $\tilde\pi(z,1,0,s)$ определяем $\tilde\pi(0,s)$:

$$\tilde\pi(0,s) = \vartheta_0(sI - D(\tilde G(s)))^{-1}. \tag{6}$$

Подставляя (6) в (5), получаем (3).

В стационарном режиме $\pi(0)$ определяем из уравнения $\pi(0)Q = 0$, где матрица Q удовлетворяет матричному функциональному уравнению

$$Q = C + D\int_0^\infty e^{Qx}dH(x) = C + D\tilde H(-Q) = C + D\tilde B_1[-Q]\{q - (1-q)\tilde B_2[-Q]\}\,.$$

Окончательно для $\pi(0)$ получаем: $\pi(0)\boldsymbol{e} = 1 - \lambda[b_1 + (1-q)b_2] = 1 - \rho$.

ПФ числа запросов в системе определяется по формуле

$$\tilde{\pi}(z,s) = \tilde{\pi}(0,s) + \int_0^\infty \tilde{\pi}(z,1,x,s)dx + \int_0^\infty \tilde{\pi}(z,2,x,s)dx, \quad \tilde{\pi}(1,s)\boldsymbol{e} = 1.$$

С помощью вектора $\tilde{\pi}(z,x,s)$ определяются характеристики модели MAP|G|1.

В заключение отметим, что разработанная модель позволяет моделировать и оценить характеристики многих сетевых сервисов и систем с учетом характера реального сетевого трафика и структуры организации обслуживания запросов. Для оценки объема буферной памяти сетевых систем в разработанной модели необходимо учесть также распределение размера запросов и его влияние на время обслуживания.

Кафедра теории вероятностей и математической статистики

Поступила 29.11.2006

ЛИТЕРАТУРА

1. **Григорян Т.А.** Разработка и исследование моделей проектирования интернет-сервиса для доступа к базам данных: Автореф. дисс. на соискание уч. степ. канд. техн. наук. Ер., 2004.
2. **Crovella M. and Bestavros A.** – IEEE/ACM Trans. on Networking, 1997, v. 5, № 6, p. 835–846.
3. **Leland W.E., Taqqu M.S., Willinger W., Wilson V.** – IEEE/ACM Trans. on Networking, 1994, v. 2, № 1, p. 1–15.
4. **Park K., Kim G. and Crovella M.** On the Effect of Traffic Self-Similarity on Network Performance. Proc. SPIE Int' Conf. Perf. & Control of Network Sys., 1997, p. 296–310.
5. **Paxson V. and Floyd S.** – IEEE/ACM Trans. on Networking, 1995, № 3, p. 226–244.
6. **Neuts M.F.** Matrix-analytic methods in the theory of queues, in Advances in Queueing: Theory, Methods and Open Problems, 1995, p. 265–292.
7. **Chao X., Miyazawa M. and Pinedo M.** Queueing Networks: Customers, Signals, and Product Form Solutions, Wiley, Chichester, 1999.
8. **Latouche G. and Ramaswami V.** Introduction to Matrix Analytic Methods in Stochastic Modeling, SIAM, Philadelphia, 1999.
9. **Lucantoni M., Choudhury G.L., Whitt W.** – Commun. Stat.-Stoch. Models, 1994, v. 10, № 1, p. 145–182.
10. **Lucantoni D.L. and Neuts M.F.** – J. Appl. Prob., 1994, v. 31, p. 235–243.
11. **Choi B.D., Hwang G.U., Han D.H.** – J. Austral. Math. Soc., Ser. B, 1998, v. 40, p. 86–96.

ՍՈՒՖԻԱՆ ՄՈՀԱՄՄԵԴ ԴՎԵՅԴԱՐԻ

ՊԱՀԱՆՋՆԵՐԻ ԵՐԿՓՈՒԼ ՍՊԱՍԱՐԿՄԱՄԲ ՑԱՆՑԱՅԻՆ ՍԵՐՎԵՐԻ ԲՆՈՒԹԱԳՐԻՉՆԵՐԻ ՈՒՍՈՒՄՆԱՍԻՐՈՒԹՅՈՒՆԸ

Ամփոփում

Աշխատանքում դիտարկվում է MAP հոսքով և պահանջների երկփուլ սպասարկումով MAP|G|1 տիպի հերթ: Ուսումնասիրված են մոդելի զբաղ-

վածության տարբերության տևողությունը, հերթի բաշխումը և ոչ ստացիոնար ռեժիմում ազատ վիճակի հավանականությունը: Անալիզը կատարված է լրացուցիչ փոփոխականի ներմուծման եղանակով:

SUFIAN MOHAMMED DVEIDARY

ANALYSIS OF NETWORK SERVER WITH CUSTOMERS TWO-STAGE SERVICE

Summary

In the present work MAP|G|1 type queue with MAP stream of customers and two-stage service is considered. The busy period duration, queue length and the probability of empty queue state in non-stationary situation are obtained. The analysis is done with the help of additional variable method.