СИСТЕМА ОБСЛУЖИВАНИЯ С ГРУППОВЫМ ПОСТУПЛЕНИЕМ ЗАПРОСОВ В СОСТАВЕ СЕССИЙ

С.А. Дудин

Белорусский государственный университет, кафедра ТВиМС

пр. Независимости, д. 4, Минск, Беларусь

телефон: + (37529) 3134834; e-mail: dudin85@mail.ru

Рассматривается однолинейная система массового обслуживания с групповым поступлением запросов в составе сессий. Прибытие сессий в систему описывается марковским входным процессом. Поступление групп запросов в сессии и время обслуживания запросов описывается потоком фазового типа. Число сессий, которые могут быть приняты в систему одновременно, является управляемым параметром. Анализируются совместное распределение числа сессий и числа заявок в системе.

Ключевые слова – марковский входной поток (*MAP*), процесс фазового типа (*PH*), сессионное поступление запросов.

1 ВВЕДЕНИЕ

Математические модели систем массового обслуживания широко используются при исследовании процессов в различных телекоммуникационных сетях. Пользователь сети может генерировать не один запрос, а целое множество, поэтому при анализе систем массового обслуживания часто предполагают групповой приход заявок. Обычно считают, что в момент прихода группы все заявки из нее прибывают в систему одновременно. Однако, типичной особенностью многих современных телекоммуникационных сетей является то, что запросы, принадлежащие одной группе, поступают в систему не одновременно, а в течение некоторого случайного времени. Первый запрос из группы прибывает в систему в момент ее прихода, в то время как остальные проступают по одному через случайные интервалы. Данная ситуация характерна для многих телекоммуникационных сетей, где, например, запросом является пакет видео- или аудиоинформации, а группой является сеанс, фильм, разговор, сессия, поток и т.д. В данной статье группы запросов, приход которых распределен во времени, будем называть *сессиями*.

Системы с сессионным поступлением запросов могут успешно использоваться при описании работы многих телекоммуникационных сетей, например сетей с протоколом HTTP/1.1, в частности, сети Интернет, где предусматривается создание виртуальных каналов, по которым передаются пакеты информации для каждого пользователя, установившего соединение.

Сценарий с поступлением запросов в сессиях описывался и анализировался с помощью компьютерного моделирования в [1]. В статье [2], а затем и в [3], словесно описанная в [1] модель была сформулирована и проанализирована в терминах теории массового обслуживания. В [2] и [3] предполагалось, что число запросов в сессии имеет геометрическое распределение и что запросы из сессии прибывают по одному через экспоненциально распределенные временные интервалы. Более общей и сложной моделью поступления запросов в сессии является рассмотренная в данной статье модель, в которой приход групп запросов в сессии управляется *PH* (Phase Type) процессом.

2 МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Структура рассматриваемой системы обслуживания приведена на Рис. 1.

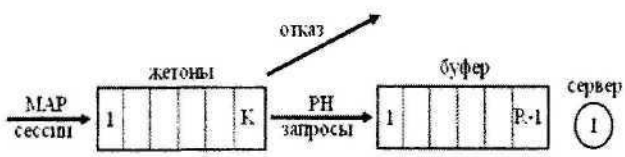

Рис. 1. Структура системы.

Система емкости $R, 1 \le R < \infty$, состоит из одного сервера и буфера размера $R-1 \ge 0$. Время обслуживания сервером имеет *PH* распределение. Это означает, что время обслуживания запроса сервером определяется как время, за которое цепь Маркова с непрерывным временем $\eta_t, t \ge 0$, имеющая несущественные состояния $\{1,\dots,M\}$ и поглощающее состояние $M+1$, достигнет поглощающего состояния. Начальное состояние цепи Маркова $\eta_t, t \ge 0$, в момент начала обслуживания запроса определяется вероятностным вектором $\boldsymbol{\beta} = (\beta_1,\dots,\beta_M)$. Переходы цепи $\eta_t, t \ge 0$, которые не приводят к окончанию обслуживания, задаются субгенератором S размера $M \times M$. Интенсивности переходов в поглощающее состояние описываются вектором

$\boldsymbol{S}_0 = -\boldsymbol{S}\boldsymbol{e}$. Здесь $\boldsymbol{e}$ – вектор-столбец, состоящий из единиц.

Запросы поступают в систему в сессиях. Сессии прибывают в соответствии с *MAP* (Markovian Arrival Process)-потоком. Обозначим управляющий процесс *MAP*-потока как $\nu_t, t \geq 0$. Этот процесс является неприводимой цепью Маркова с непрерывным временем, имеющей пространство состояний $\{0,1,\ldots,W\}$. Время пребывания цепи в состоянии ν экспоненциально распределено с положительным параметром λ_ν. Когда время пребывания в состоянии ν истекло, с вероятностью $p_{\nu,\nu'}^{(k)}$ процесс ν_t переходит в состояние ν', при этом генерируются k сессий, $k = 0,1, \nu, \nu' = \overline{0,W}$. Поведение данного *MAP*-потока сессий полностью характеризуется матрицами $D_k, k = 0,1,$ которые определяются следующим образом: $(D_k)_{\nu,\nu'} = \lambda_\nu p_{\nu,\nu'}^{(k)}, k = 1$ и $k = 0, \nu \neq \nu'; (D_0)_{\nu,\nu} = -\lambda_\nu, \nu', \nu = \overline{0,W}$.

Матрица $D(1) = D_0 + D_1$ является инфинитезимальным генератором цепи $\nu_t, t \geq 0$. Средняя интенсивность поступления сессий λ имеет вид $\lambda = \boldsymbol{\chi} D_1 \boldsymbol{e}$, где $\boldsymbol{\chi}$ – вектор стационарного распределения цепи Маркова $\nu_t, t \geq 0$. Вектор $\boldsymbol{\chi}$ является единственным решением системы линейных алгебраических уравнений $\boldsymbol{\chi} D(1) = \boldsymbol{0}, \boldsymbol{\chi}\boldsymbol{e} = 1$. Здесь $\boldsymbol{0}$ – вектор-строка, состоящий из нулей.

Предполагаем, что прием сессий в систему ограничивается с помощью жетонов. Общее число жетонов $K, K \geq 1$. Если в момент прихода сессии нет свободных жетонов или буфер полон, то сессия получает отказ и покидает систему. Иначе, сессия принимается на обслуживание и число свободных жетонов уменьшается на единицу.

Мы предполагаем, что первый запрос сессии прибывает в момент поступления сессии. Процесс поступления остальных запросов в сессии описывается *PH* входным потоком. То есть приход запросов из принятой сессии (кроме первого запроса) управляется процессом $j_t, t \geq 0$, который является цепью Маркова с непрерывным временем и пространством состояний $\{1,\ldots,J\}$. Начальное состояние цепи $j_t, t \geq 0$, в момент прихода сессии определяется вектором $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_J)$. Переход цепи Маркова j_t из состояния j в состояние $j', j \neq j', j, j' = \overline{1,J}$, с вероятностью $p_n, n = \overline{1,N}$, сопровождается приходом группы из n запросов, принадлежащих данной сессии. Переход из состояния j в состояние j не допускается. Интенсивности переходов процесса $j_t, t \geq 0$, определяются субгенератором Λ размера $J \times J$. Интенсивности переходов, которые приводят к окончанию прибытия сессии, задаются с помощью вектора $\boldsymbol{\Delta}_0 = -\Delta\boldsymbol{e}$.

Если в момент прихода группы запросов из принятой сессии число свободных мест в системе превышает размер группы, вся группа принимается в систему. В противном случае, запросы, которым не хватило места в системе, покидают систему навсегда. Если прибытие сессии окончено, жетон, который был у этой сессии, возвращается в пул свободных жетонов. Все запросы из оконченной сессии, находившиеся в буфере в момент возврата жетона, будут обслужены системой.

3 СОВМЕСТНОЕ РАСПРЕДЕЛЕНИЕ ЧИСЛА СЕССИЙ И ЗАПРОСОВ В СИСТЕМЕ

Пусть $k_t = \overline{0,K}$, – число сессий в системе, $i_t = \overline{0,R}$, – число запросов в системе, $\nu_t = \overline{0,W}$, – состояние управляющего процесса поступления сессий, $j_t^{(l)} = \overline{0,J}$, – состояние управляющего процесса поступления запросов в l-й сессии, $l = \overline{1,k_t}$, $\eta_t = \overline{0,M}$, – состояние управляющего процесса обслуживания прибором в момент времени $t, t \geq 0$. Тогда очевидно, что процесс $\xi_t = \{k_t, i_t, \nu_t, j_t^{(1)},\ldots, j_t^{(k_t)}, \eta_t\}, t \geq 0,$ является неприводимой регулярной цепью Маркова с непрерывным временем.

Текущие сессии предполагаются перенумерованными в порядке их открытия: номер 1 имеет наиболее долготекущая к данному моменту времени сессия. При завершении прихода сессии соответствующий ей номер вычеркивается из записи и производится перенумерация.

Когда число i заявок в системе равно 0, состояние процесса η_t не определено. Чтобы избежать отдельного рассмотрения такого случая, мы будем предполагать, что после момента, когда система станет пустой, значение компоненты η_t установится случайно в соответствии с вероятностным вектором $\boldsymbol{\beta}$ и сохранится до начала следующего обслуживания прибором.

Перенумеруем состояния цепи Маркова ξ_t в лексикографическом порядке и множество состояний, имеющих значение (k,i) двух первых компонент цепи, будем называть макросостоянием.

Пусть Q – генератор цепи Маркова $\xi_t, t \geq 0$, состоящий из блоков $Q_{k,j}$, состоящих из матриц $(Q_{k,j})_{i,i'}$ интенсивностей переходов этой цепи из макросостояния (k,i) в макросостояние $(j,i'), i,i' = \overline{0,R}$.

Введем следующие обозначения:

- $\tilde{\Delta}$ – это диагональная матрица, диагональные элементы которой совпадают с диагональными элементами матрицы Δ;
- $\hat{\Delta} = \Delta - \tilde{\Delta}$;

- O – нулевая матрица;
- I – единичная матрица;
- $\otimes$ и $\oplus$ – символы кронекеровых произведения и суммы матриц;
- $A^{\oplus i} = \underbrace{A \oplus A \oplus \ldots \oplus A}_{i}, i > 0, A^{\oplus 0} = O_{1\times 1}$;
- $\bar{W} = W + 1$.

Лемма 1. Генератор Q имеет следующую блочно-трехдиагональную структуру:

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & \ldots & O & O \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & \ldots & O & O \\ O & Q_{2,1} & Q_{2,2} & \ldots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \ldots & Q_{K-1,K-1} & Q_{K-1,K} \\ O & O & O & \ldots & Q_{K,K-1} & Q_{K,K} \end{pmatrix},$$

где ненулевые блоки $Q_{k,j}$ вычисляются как

$$Q_{k,k} =$$

$$\begin{pmatrix} A(k,0) & C(k,0,1) & C(k,0,2) & \ldots & C(k,0,R) \\ B(k) & A(k,1) & C(k,1,1) & \ldots & C(k,1,R-1) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ O & \ldots & B(k) & A(k,R-1) & C(k,R-1,1) \\ O & \ldots & O & B(k) & \hat{A}(k) \end{pmatrix},$$

$$k = \overline{0,K},$$

$$Q_{k,k+1} = \begin{pmatrix} O & E(k,0) & O & \ldots & O \\ O & O & E(k,1) & \ldots & O \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ O & O & \ldots & O & E(k,R-1) \\ O & O & \ldots & O & O \end{pmatrix}, k = \overline{0,K-1},$$

$$Q_{k,k-1} = \operatorname{diag}\{I_{\bar{W}} \otimes \Delta_0^{\oplus k} \otimes I_{M^{\min\{i,1\}}}, i = \overline{0,R}\}, k = \overline{1,K},$$

$$A(k,i) = \begin{cases} D_0 \oplus \tilde{\Delta}^{\oplus k} \oplus S^{\oplus \min\{i,1\}}, k < K, \\ D(1) \oplus \tilde{\Delta}^{\oplus k} \oplus S^{\oplus \min\{i,1\}}, k = K, \end{cases},$$

$$\hat{A}(k) = D(1) \oplus \Delta^{\oplus k} \oplus S,$$

$$C(k,i,n) = p_n^{(i)} I_{\bar{W}} \otimes \hat{\Delta}^{\oplus k} \otimes I_{M^{\min\{i,1\}}},$$

$$B(k) = I_{\bar{W}} \otimes I_{J^k} \otimes (\boldsymbol{S}_0 \boldsymbol{\beta}),$$

$$p_s^{(i)} = \begin{cases} p_s, & s + i < R, s \le N, \\ \sum_{l=s}^{N} p_l, & s + i = R, s \le N, \\ 0, & s > N, \end{cases}$$

$$E(k,i) = D_1 \otimes (I_{J^k} \otimes \boldsymbol{\delta}) \otimes I_{M^{\min\{i,1\}}}.$$

Доказательство Леммы 1 состоит в анализе переходов цепи Маркова $\xi_t, t \ge 0$, за бесконечно малый интервал времени и последующей группировке интенсивностей соответствующих переходов в блочные матрицы.

Так как цепь Маркова $\xi_t = \{k_t, i_t, \nu_t, j_t^{(1)}, \ldots, j_t^{(k_t)}, \eta_t\}, t \ge 0$, неприводимая, регулярная и имеет конечное пространство состояний, то существует единственное стационарное распределение вероятностей состояний этой цепи, совпадающее с эргодическим (предельным, финальным):

$$\pi(k,i,\nu,j^{(1)},\ldots,j^{(k)},\eta) =$$

$$= \lim_{t\to\infty} P\{i_t = i, k_t = k, \nu_t = \nu, j_t^{(1)} = j^{(1)}, \ldots, j_t^{(k)} = j^{(k)}, \eta_t = \eta\},$$

$$i = \overline{0,R}, k = \overline{0,K}, \nu = \overline{0,W}, j^{(l)} = \overline{0,J}, l = \overline{1,k}, \eta_t = \overline{0,M}.$$

Перенумеруем стационарные вероятности цепи Маркова ξ_t *в лексикографическом порядке и сформируем* из них векторы $\boldsymbol{\pi}(k,i)$, состоящие из вероятностей $\pi(k,i,\nu,j^{(1)},\ldots,j^{(k)},\eta)$, записанных в лексикографическом порядке по компонентам $(\nu, j^{(1)},\ldots,j^{(k)},\eta)$, и векторы

$$\boldsymbol{\pi}_k = (\boldsymbol{\pi}(k,0),\boldsymbol{\pi}(k,1),\ldots,\boldsymbol{\pi}(k,R)), k = \overline{0,K}.$$

Известно, что вектор $(\boldsymbol{\pi}_0,\ldots,\boldsymbol{\pi}_K)$ является единственным решением системы линейных алгебраических уравнений:

$$(\boldsymbol{\pi}_0,\ldots,\boldsymbol{\pi}_K)Q = \mathbf{0},\ (\boldsymbol{\pi}_0,\ldots,\boldsymbol{\pi}_K)\boldsymbol{e} = 1.$$

Для решения данной системы может быть использован устойчивый алгоритм, являющийся модификацией алгоритма, разработанного в [4], для случая генератора блочно-трехдиагональной структуры. Этот алгоритм изложен в следующей теореме.

Теорема 1. Векторы стационарных вероятностей $\boldsymbol{\pi}_k, k = \overline{0,K}$, вычисляются как

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_0 F_k, k = \overline{1,K},$$

где матрицы F_k вычисляются рекуррентно:

$$F_0 = I,\ F_k = -F_{k-1}Q_{k-1,k}(Q_{k,k} + Q_{k,k+1}G_k)^{-1}, k = \overline{1,K},$$

матрицы $G_k, k = \overline{0,K-1}$, вычисляются с помощью обратной рекурсии:

$$G_k = -(Q_{k+1,k+1} + Q_{k+1,k+2}G_{k+1})^{-1}Q_{k-1,k},$$

$$k = K-2, K-3, \ldots, 0,$$

при начальном условии

$$G_{K-1} = -(Q_{K,K})^{-1}Q_{K,K-1},$$

вектор $\boldsymbol{\pi}_0$ вычисляется как единственное решение следующей системы алгебраических уравнений:

$$\boldsymbol{\pi}_0(Q_{0,0}+Q_{0,1}G_0)=\mathbf{0},\quad \boldsymbol{\pi}_0\sum_{k=0}^{K}F_k\boldsymbol{e}=1.$$

Заметим, что все обратные матрицы в утверждении теоремы существуют, так как субгенераторы обратимы.

Алгоритм вычисления векторов стационарных вероятностей $\boldsymbol{\pi}_k,\ k=\overline{0,K},$ приведенный в Теореме 1, позволяет легко решить систему на компьютере практически при любых размерностях.

Найдя векторы стационарных вероятностей $\boldsymbol{\pi}_k,\ k=\overline{0,K},$ можно вычислить различные характеристики производительности системы:

Среднее число запросов в системе $L=\sum_{k=0}^{K}\sum_{i=1}^{R}i\boldsymbol{\pi}(k,i)\boldsymbol{e}.$

Среднее число сессий в системе $B=\sum_{k=1}^{K}k\boldsymbol{\pi}_k\boldsymbol{e}.$

Среднее число запросов, обслуженных в единицу времени, (пропускная способность)

$$T=\sum_{i=1}^{R}\sum_{k=0}^{K}\boldsymbol{\pi}(k,i)(\boldsymbol{e}_{\bar{W}}\otimes\boldsymbol{e}_{J^k}\otimes\boldsymbol{S}_0).$$

Вероятность отказа сессии на входе

$$P_s^{(loss)}=(\sum_{i=0}^{R-1}\boldsymbol{\pi}(K,i)(D_1\otimes I_{J^K M^{\min\{i,1\}}})+$$

$$+\sum_{k=0}^{K}\boldsymbol{\pi}(k,R)(D_1\otimes I_{J^k M}))\frac{1}{\lambda}\boldsymbol{e}.$$

Вероятность потери группы запросов из принятой сессии целиком

$$P_g^{(loss)}=\frac{\sum_{k=1}^{K}\boldsymbol{\pi}(k,R)\boldsymbol{e}_{\bar{W}}\otimes\hat{\Delta}^{\oplus k}\boldsymbol{e}_{J^k}\otimes\boldsymbol{e}_M}{\sum_{k=1}^{K}\sum_{i=0}^{R}\boldsymbol{\pi}(k,i)\boldsymbol{e}_{\bar{W}}\otimes\hat{\Delta}^{\oplus k}\boldsymbol{e}_{J^k}\otimes\boldsymbol{e}_{M^{\min\{i,1\}}}}.$$

Вероятность потери произвольного запросов из принятой сессии

$$P_c^{(loss)}=$$

$$\sum_{k=1}^{K}\sum_{m=0}^{R}\left(\boldsymbol{\pi}(k,R-m)\boldsymbol{e}_{\bar{W}}\otimes\hat{\Delta}^{\oplus k}\boldsymbol{e}_{J^k}\otimes\boldsymbol{e}_M\sum_{n=m+1}^{N}p_n(n-m)\right)\times$$

$$\times\left(\sum_{n=1}^{N}np_n\left(\sum_{k=1}^{K}\sum_{i=0}^{R}\boldsymbol{\pi}(k,i)\boldsymbol{e}_{\bar{W}}\otimes\hat{\Delta}^{\oplus k}\boldsymbol{e}_{J^k}\otimes\boldsymbol{e}_{M^{\min\{i,1\}}}\right)\right)^{-1}.$$

ЛИТЕРАТУРА

[1] A simple IP flow blocking model / A.A. Kist, B. Lloyd-Smith, R.J. Harris //Performance Challenges Efficient Next Generation Networks. Proceedings of 19 International Telegraphic Congress, 29 August – 2 September 2005, Beijing. – 2005. – P. 355-364.

[2] Queueing Model with Time-Phased Batch Arrivals / M.H. Lee, S. Dudin, V. Klimenok // Lecture Notes in Computer Science. – 2007. – V. 4516. – P. 716-730.

[3] The MAP/PH/1/N queue with time phased arrivals as model for traffic control in telecommunication networks / C.S. Kim, S.A. Dudin, V.I. Klimenok // Performance Evaluation. – 2009. – V. 66. – P. 564-579.

[4] Lack of invariant property of Erlang loss model in case of the MAP input / V. Klimenok, C.S. Kim, D. Orlovsky, A. Dudin // Queueing Systems. – 2005. – V. 49. – P. 187-213.