Многоканальные системы обслуживания с марковским нетерпением[1]

Ю.И. Рыжиков

Институт информатики и автоматизации РАН,
Военно-космическая академия им. А.Ф. Можайского,
г. Санкт-Петербург, Россия

Среди многочисленных применений теории очередей заметную роль играют ситуации с нетерпеливыми заявками, имеющими случайные ограничения на время пребывания заявки в системе. Чисто марковская система указанного вида была рассмотрена в [1, 2]. В данной статье рассматривается гораздо более интересная для практики задача с гиперэкспоненциальным распределением обслуживания и марковским нетерпением. Предлагаемый метод позволяет обобщить задачу применительно к сети обслуживания.

Итерационный метод для модели $M/H_2/n$

Работа системы $M/H_2/n$ может быть интерпретирована как процесс обслуживания потока заявок двух типов. Каждая заявка имеет случайное время терпения, подчиненное экспоненциальному закону с параметром γ.

На рис. 1 представлен фрагмент диаграммы переходов между микросостояниями системы $M/H_2/3$-M по уходу заявок для 2-го, 3-го и 4-го ярусов. Кодовые комбинации вида (21) указывают типы обслуживаемых заявок, проставленные на концах стрелок дополнительные множители $\{y_i\}$ – вероятности выбора заявок соответствующего типа из очереди. На ярусах $j>n$ отличие будет состоять лишь в интенсивности вертикальных переходов вверх: γ, 2γ, 3γ … Заметим, что скорость продвижения в очереди новой (меченой) заявки определяется только судьбой ранее пришедших.

В соответствии с диаграммой переходов для выбранной модели построим матрицы интенсивностей инфинитезимальных переходов: A_j по прибытию заявки, B_j – по завершению обслуживания, D_j – ухода из состояний яруса j.

[1] Работа выполнялась по госзаданию 0073 2018-0003.

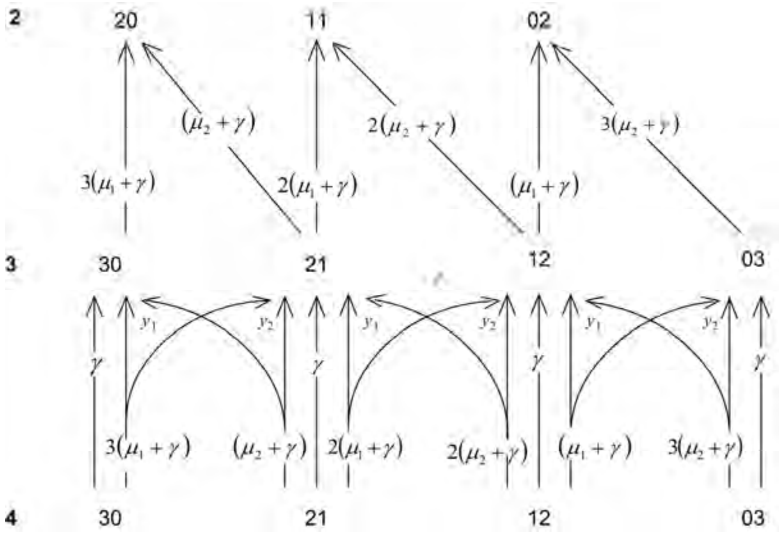

Рис. 1. Фрагмент диаграммы переходов

Введем векторы-строки g_j={$g_{j,1}$, $g_{j,2}$,…, $g_{j,mj}$ } нахождения системы в состоянии (j,i) , j=0,1,… . Теперь можно записать векторно-матричные уравнения баланса переходов между состояниями

$$\begin{aligned} g_0 D_0 &= g_1 B_1, \\ g_j D_j &= g_{j-1} A_{j-1} + g_{j+1} B_{j+1}, \quad j = 1, 2, \ldots . \end{aligned} \tag{1}$$

Положим $t_j = g_j / p_j$, где p_j – суммарная вероятность наличия в системе ровно j заявок, и обозначим

$$x_j = p_{j+1}/p_j, \qquad z_j = p_{j-1}/p_j. \tag{2}$$

Тогда систему (1) можно переписать относительно векторов условных вероятностей $\{t_j\}$, нормированных к единице в пределах яруса:

$$\begin{aligned} t_0 D_0 &= x_0 t_1 B_1, \\ t_j D_j &= z_j t_{j-1} A_{j-1} + x_j t_{j+1} B_{j+1}, \quad j = 1, 2, \ldots . \end{aligned} \tag{3}$$

Алгоритм [3, 4] расчета набора векторов $\{t_j\}$ и чисел $\{x_j\}$ и $\{z_j\}$, удовлетворяющих соотношениям (1) – (3) , в случае разомкнутой системы опирается на существование предельного вектора условных вероятностей $t_\infty = \lim_{j\to\infty} t_j$, которое является следствием стабилизации переходных матриц при $j > n$. Для граничного яруса N в случае прогона сверху вниз последняя из формул (3) записывается в виде

$$t_N D_N = z_N t_{N-1} A_{N-1} + x_N t_{N-1} B_N \qquad (4)$$

(отсутствующий нижележащий вектор условных вероятностей микросостояний заменяется на только что вычисленный вышележащий). В случае с нетерпеливыми заявками $\{B_j\}$ и $\{D_j\}$ стабилизируются лишь при $j \to \infty$, но вследствие нарастающей при этом интенсивности уходов по нетерпению суммарные вероятности состояний $\{p_j\}$ при прочих равных условиях будут убывать значительно быстрее, и ошибки от упомянутого допущения будут играть все меньшую роль. Это оправдывает применение для граничного яруса N условия (4).

После прекращения итераций можно переходить к нахождению кумулянтных вероятностей. Полагая $p_0 = 1$, последовательно вычисляем

$$p_{j+1} = p_j x_j, \qquad j = \overline{0, N-1}, \qquad (5)$$

а затем нормируем их к единице. Имитационный эксперимент показал хорошее согласие расчетных и статистических вероятностей порядка 10^{-4} и больше, для которых число наблюдений превышало 100.

Распределение времени ожидания

Наибольшие трудности расчета систем с нетерпением связаны с определением их временных характеристик. Здесь общая формула Брюмелля для моментов распределения длительности ожидания

$$w_j = q_{[j]}/\lambda^j, \qquad j = 1, 2, \ldots, \qquad (6)$$

в которой $\{q_{[j]}\}$ – факториальные моменты распределения длины очереди, как показали имитационные эксперименты, дает неудовлетворительную точность.

Вычислим

- стационарные векторы-строки $g_k = p_k * t_k$ вероятностей микросостояний, $k = 0, 1, \ldots,$
- диагональные матрицы суммарных интенсивностей переходов вверх с элементами $\{\sigma_{k,i}\}$,

• диагональные матрицы $\{U_k(s)\}$ преобразований Лапласа – Стилтьеса (ПЛС) распределений длительности переходов вверх по соответствующим суммарным интенсивностям $\{\sigma_k\}$ с элементами $\{\sigma_{k,i}/(\sigma_{k,i}+s)\}$,

• произведение $\tilde{U}_n(s)$ матрицы $U_n(s)$ на единичный вектор-столбец,

• матрицы $\{T_k\}$ с элементами $\{b_{k,i,j}/\sigma_{k,i}\}$ вероятностей переходов на вышележащий ярус.

Кроме того, заменим диагональную матрицу $U_n(s)$ на одноименный вектор-столбец. Полагая $F_k(s)=U_k(s)T_k$, можно сразу записать матрицы $\{F_k\}$ как совокупности элементов вида $\{b_{k,i,j}/(\sigma_{k,i}+s)\}$. Суммируя результаты по всем возможным стартовым ярусам, получаем окончательную формулу для ПЛС распределения времени ожидания:

$$\omega(s)=[\sum_{k=0}^{\infty}g_{n+k}\coprod_{i=0}^{k}F_{n+i}(s)]\tilde{U}_n(s). \qquad (7)$$

В этой формуле перевернутый символ произведения используется для указания на обратный порядок следования сомножителей. Начальное значение $F_n(s)=I$.

Многократным численным дифференцированием таблицы ПЛС [4] можно получить моменты распределения $w(t)$ виртуального ожидания начала обслуживания

Осталось учесть возможность «нетерпения» самой меченой заявки. Поскольку для каждой заявки вероятность перетерпеть время u равна $e^{-\gamma u}$, функция распределения длительности успешного ожидания

$$W^+(t)=\int_0^t e^{-\gamma u}w(u)du.$$

Соответственно ПЛС этого распределения

$$\omega^+(s)=\int_0^\infty e^{-st}d[\int_0^t e^{-\gamma u}w(u)du]dt=\int_0^\infty e^{-st}e^{-\gamma t}w(t)dt=\omega(s+\gamma)..$$

Моменты успешного ожидания, полученные таким способом, должны быть разделены на вероятность успешного ожидания, включая нулевое, то есть на

$$\pi_w = \sum_{k=0}^{n-1} p_k + \omega^+(\gamma).$$

Сопоставим численные результаты, полученные по этой методике и посредством имитационного моделирования (2 млн обслуженных заявок). Для трехканальной системы с интенсивностью входящего простейшего потока $\lambda = 1.5$, среднем временем обслуживания $b_1 = 4.0$, коэффициентом вариации обслуживания $v_b = 2.0$ и интенсивностью нетерпения $\gamma = 0.2$ результаты сведены в таблицу.

Моменты распределения успешного ожидания

Способ	w_1^+	w_2^+	w_3^+
Имитация	0.483	1.048	3.319
Расчет	0.479	1.039	3.287

Распределение времени пребывания успешной заявки

Сделанное в данной статье предположение о марковском распределении допустимого терпения позволяет при поступлении заявки из очереди на обслуживание отсчитывать ее терпение заново. Нас в конечном счете интересует ожидание «успешных» заявок, дождавшихся завершения обслуживания. Примем распределение последнего двухфазовым гиперэкспоненциальным с параметрами $\{y_m, \mu_m\}$. Тогда j-й момент успешного обслуживания

$$b_j^+ = \int_0^\infty [\int_0^\theta t^j b(t)dt]\gamma e^{-\gamma\theta} d\theta = \int_0^\infty [\int_0^\theta t^j \sum_{m=1}^{2} y_m \mu_m e^{-\mu_m t} dt].$$

Можно показать, что

$$b_j^+ \quad = \quad j!\sum_{m=1}^{2} \frac{y_m}{\mu_m^j}[1 - \frac{\gamma}{\mu_m + \gamma} \sum_{i=0}^{j} (\frac{\mu_m}{\mu_m + \gamma})^i]. \qquad (8)$$

Расчет сети с нетерпеливыми заявками

Предположение о марковском распределении допустимого терпения позволяет перенести результаты предыдущего раздела на расчет сетей методом потокоэквивалентной декомпозиции. В нашем случае в общую схему расчета разомкнутой сети обслуживания [4] приходится внести следующие изменения:

- В сети не должно быть циклических маршрутов.
- Интенсивности выходящих потоков должны рассчитываться умножением интенсивностей входящих на вероятность успеха.
- Нумерация и, соответственно, очередность расчета узлов должны определяться на основе отношений предшествования – например, с помощью известного алгоритма Флойда.

Выделим из матрицы передач сети с M рабочими узлами:

- вектор-строку $P=\{r_{0,1},r_{0,2},\ldots,r_{0,M}\}$ вероятностей перехода из источника в конкретные рабочие узлы;
- вектор-столбец $T=\{r_{1,M+1},r_{2,M+1},\ldots,r_{M,M+1}\}^T$ вероятностей перехода из рабочих узлов в сток;
- матрицу $Q=\{r_{i,j}\},\quad i,j=\overline{1,M}$, вероятностей переходов между рабочими узлами.

Кроме того, определим диагональную матрицу $N(s)$ преобразований Лапласа $\{\nu_i(s)\}$ распределений времени пребывания в рабочих узлах сети и сформируем матрицу

$$\Phi(s)=N(s)Q.$$

ПЛС распределения длительности k-шаговых переходов в сток

$$\varphi_k(s)=P\Phi^{k-1}(s)N(s)T.$$

Полное ПЛС распределения времени пребывания заявки в сети получается суммированием переходов кратностей до N включительно, где N – максимальное количество посещаемых заявкой узлов:

$$\varphi(s)=\sum_{k=1}^{N}\varphi_k(s)=P[\sum_{k=0}^{N-1}\Phi^k(s)]N(s)T.$$

Моменты $\{g_i\}$ распределения времени пребывания в сети можно получить численным дифференцированием $\varphi(s)$ в нуле.

Заключение

Основные результаты данной работы сводятся к следующему:

1. С учетом экспоненциального нетерпения находящихся в системе заявок. скорректированы правила расчета матриц интенсивностей переходов $\{B_j\}$ и $\{D_j\}$.

2. Обоснована возможность применения метода Такахаси-Таками к рассматриваемой задаче с переменными матрицами интенсивностей переходов.

3. Разработан метод расчета распределения времени ожидания для «успешных» заяок.

4. Предложены формулы для расчета средних потерь на прерванное обслуживание и интенсивности потока обслуженных заявок.

5. Показано применение перечисленных результатов к расчету сетей обслуживания с нетерпеливыми заявками.

Перечисленные результаты применимы к широкому кругу практически важных ситуаций. В технике связи и военном деле это может быть «обслуживание» подвижных объектов с ограниченным временем пребывания в зоне досягаемости, в проблематике МЧС – спасение людей в аварийных ситуациях, в юстиции – затяжные процессы с ограниченным сроком давности, в медицине – «тяжелые» больные, состояние которых быстро ухудшается в ожидании оказания экстренной помощи, и т. п.

ЛИТЕРАТУРА

1. *Бочаров П.П., Печинкин А.В.* Теория массового обслуживания: Учебник. М.: Изд-во РУДН, 1995. 529 с.
2. *Takagi H. Waiting* Time in the M/M/m/(m+c) Queue with Impatient Customers // Int. J. of Pure and Applied Mathematics. 2014. V. 90. No. 4. P. 519–559.
3. *Рыжиков Ю.И.* Итеративный метод расчета многоканальных систем обслуживания – основы, модификации и предельные возможности // Труды 9-й Российской мультиконференции по проблемам управления. Информационные технологии в управлении. СПб.: ГНЦ РФ «Концерн Электроприбор», 2016. С. 224–233.
4. *Рыжиков Ю.И.* Алгоритмический подход к задачам массового обслуживания: Монография. СПб.: ВКА им. А.Ф. Можайского, 2013. 496 с.
5. *Вишневский В.М.* Теоретические основы проектирования компьютерных сетей. М.: Техносфера, 2003. 512 с.
6. *Takahashi Y., Takami Y. A* Numerical Method for the Steady-State Probabilities of a GI/G/c Queuing System in a General Class // J. Operat. Res. Soc. of Japan. 1976. V. 19. No. 2. P. 147–157.
7. *Малинковский Ю.В.* Сети Джонсона с однолинейными узлами и ограниченным временем пребывания и.или ожидания // Автоматика и Телемеханика. 2015. № 4. С. 67–79.

Рыжиков Юрий Иванович, д.т.н., профессор; ryzhbox@yandex.ru