

УДК 519.872

РАСЧЕТ СИСТЕМ СО СЛУЧАЙНЫМ ВЫБОРОМ НА ОБСЛУЖИВАНИЕ

Ю. И. Рыжиков,

доктор техн. наук, профессор

Военно-космическая академия им. А. Ф. Можайского

Проверка на имитационной модели известных частных методов расчета систем с очередями при случайном выборе на обслуживание выявила заметные ошибки в высших моментах распределения времени ожидания. Предлагается решать задачу построением вложенной цепи Маркова, описывающей изменение числа заявок в очереди на моменты перед завершением обслуживания в многоканальной системе — до выбора меченой заявки. Конечный результат получен в виде преобразования Лапласа распределения времени ее ожидания. Алгоритм тестирован на модели с простейшим входящим потоком и обобщен на рекуррентный поток.

The verification of known special approaches to queuing systems with random choice by imitation has detected noticeable errors in the high order moments of waiting time distribution. It is proposed to construct embedded Markov chain which describe the changing of queue length in multi-channel system before service completing — until the marked demand is selected. The final result is the Laplace transform of waiting time distribution. An algorithm is tested for Poissonian input flow and generalized for recurrent flow.

Постановка задачи

подавляющее большинство работ по теории очередей посвящено системам с выборкой из очереди по принципу FCFS (первый пришел — первый обслужен). Однако значительный интерес вызывает и другой принцип — случайный выбор из очереди (RANDOM). Он особенно типичен для связанных приложений и ситуаций с разъездными бригадами обслуживания (ремонтными, аварийными, скорой помощи, групп быстрого реагирования и т. п.), где очередность обработки заявок не определяется моментами их поступления. Различные подходы к этой задаче предлагаются или воспроизводятся в работах [1–7]. В работах [2, 3, 6] обсуждаются чисто марковские системы $M/M/n$. В частности, в них выводятся формулы для коэффициентов увеличения высших моментов времени ожидания в сравнении с дисциплиной FCFS:

$$R_2 = (1 - \rho/2)^{-1};$$

$$R_3 = (4 + 2\rho)/(2 - \rho)^2.$$

В статье Кингмана [5] выполнен анализ системы $M/G/1$, к сожалению, не доведенный до расчетных зависимостей. Дисперсия времени ожидания для этой же системы приводится в справочнике Дж. Мартина [1]. В работе Розенлунда [7] для

модели $GI/M/n$ получены преобразования Лапласа—Стилтьеса (ПЛС) распределения времени ожидания и рекуррентный алгоритм вычисления моментов. Фрагмент его итоговой таблицы приведен в табл. 1.

Наконец, в статье Картера и Купера [4] рассматриваются системы $GI/M/n$ и $M/G/n$. Здесь подтверждены известные результаты для систем $M/M/n$ и $M/D/1$, рассчитана система $M/E_k/1$.

Все перечисленные результаты выводятся весьма громоздкими методами, к тому же различными для разных классов моделей, с трудно прослеживаемой аргументацией и с привлечением эвристических приемов. Это делает весьма желательной их проверку на простой и надежной имитационной модели. В табл. 2 приведены результаты имитационного моделирования системы $M/M/3$ с дисциплинами FCFS и RANDOM при 500 тыс. испытаний.

Средний интервал между заявками предполагался единичным, интенсивность обслуживания выбиралась через заданный коэффициент загрузки ρ . Аналитические результаты хорошо согласуются с имитационной моделью FCFS с учетом общеизвестного нарастания погрешностей при увеличении порядка моментов и коэффициента загрузки системы. Переделка модели под случайный выбор из очереди затронула лишь четыре операто-

■ Таблица 1. Поправки к FIFO-моментам

Порядок моментов	D/M/n			E ₄ /M/n			M/M/n		
	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
2	1.2550	1.5005	1.8125	1.2884	1.5164	1.8148	1.3333	1.5385	1.8182
3	1.9782	3.0602	4.7698	2.0776	3.1149	4.7793	2.2222	3.1953	4.7934
4	3.6889	7.6951	16.238	3.9570	7.8927	16.286	4.3704	8.1896	16.356

■ Таблица 2. Расчет и имитация марковской системы

ρ	Моменты	FCFS (расчет)	FCFS (имитация)	RANDOM (имитация)	Отношение RANDOM/FCFS
0.5	w ₁	0.2368	0.2346	0.2346	1.0000
	w ₂	0.4737	0.4628	0.5362	1.1585
	w ₃	1.4211	1.3572	2.2109	1.6290
0.7	w ₁	1.1488	1.1494	1.1494	1.0000
	w ₂	5.3611	5.2108	7.1150	1.3654
	w ₃	37.528	34.162	86.600	2.5350
0.9	w ₁	7.3535	7.6507	7.3628	0.9624
	w ₂	132.36	140.17	226.70	1.6173
	w ₃	3573.9	3687.5	14980	4.0624

ра программы, и в ней было невозможно ошибиться. Правильность RANDOM-модели подтверждается и практическим совпадением средних времен ожидания. Однако отношения высших моментов заметно отличаются от приведенных в табл. 1, что вынуждает поставить вопрос о новых методах расчета систем обслуживания с RANDOM-дисциплиной.

Вложенная цепь Маркова для простейшего входящего потока

Рассмотрим режим полной занятости системы GI/G/n/R и обозначим:

J = R - n — максимальную длину очереди;

B_n(t) — распределение интервалов между последовательными завершениями обслуживания;

B_n^{*}(t) — распределение интервала от прибытия «меченой» заявки до ближайшего завершения обслуживания;

λ — интенсивность входящего потока.

Вычислим вероятности прибытия ровно j заявок за время до ближайшего обслуживания:

$$q_j^*(t) = \int_0^{\infty} \frac{(\lambda t)^j}{j!} e^{-\lambda t} dB_n^*(t), \quad j = 0, 1, \dots$$

и с заменой B_n^{*}(t) на B_n(t) — аналогичный набор {q_j} вероятностей прибытия ровно j заявок за время между смежными обслуживаниями. Введем также вектор P = {p₀, p₁, ..., p_J} стационарного распределения вероятностей длины очереди в момент прибытия меченой заявки. Построим вложенную цепь Маркова изменения числа заявок в оче-

реди непосредственно перед очередным выбором на обслуживание. Элементы матрицы R* переходов за время ожидания первого обслуживания должны вычисляться согласно

$$r_{i,j}^* = q_{j-i-1}^*(j-1)/j, \quad i = \overline{0, J}, j = \overline{i+1, J}$$

(было i заявок, пришли меченая и еще j - i - 1 ≥ 0). Дробный множитель учитывает вероятность невыбора меченой и, следовательно, продолжения процесса. Процесс завершается выбором меченой заявки с вероятностью 1/j. Вектор-столбец T* определяет вероятности перехода в поглощающее состояние. Его компоненты

$$t_i^* = \sum_{j=0}^J \frac{q_j^*}{i+j+1}.$$

Продолжение процесса (переходы на интервалах между смежными обслуживаниями) возможно при начальных состояниях i = 2, J. Соответственно переход i → j получается после прибытия j - i + 1 заявок (одна заявка уходит сразу после начала отсчета интервала). Итак, элементы новой матрицы R

$$r_{i,j} = q_{j-i+1}(j-1)/j, \quad i, j = \overline{2, J},$$

а элементы нового вектора вероятностей поглощения

$$t_i = \sum_{j=0}^J \frac{q_j}{i+j-1}.$$

Введем ПЛС временных распределений:
 $\beta^*(s)$ — от прибытия меченой заявки до ближайшего обслуживания;
 $\beta(s)$ — между смежными завершениями обслуживания;
 $\omega_k(s)$ — ожидания меченой заявки, завершаемого через k шагов марковской цепи, $k = 1, 2, \dots$
 Нетрудно видеть, что

$$\begin{aligned} \omega_1(s) &= \beta^*(s) \mathbf{P} \mathbf{T}^*, \\ \omega_2(s) &= \beta^*(s) \mathbf{P} \mathbf{R}^* \beta(s) \mathbf{T}, \\ \omega_3(s) &= \beta^*(s) \mathbf{P} \mathbf{R}^* \beta(s) [\mathbf{R} \beta(s)] \mathbf{T}, \\ &\dots \dots \dots \\ \omega_k(s) &= \beta^*(s) \mathbf{P} \mathbf{R}^* \beta(s) [\mathbf{R} \beta(s)]^{k-2} \mathbf{T}. \end{aligned}$$

Суммируя эти выражения по всем k , находим

$$\begin{aligned} \omega(s) &= \beta^*(s) \mathbf{P} \left(\mathbf{T}^* + \mathbf{R}^* \beta(s) \sum_{k=2}^{\infty} [\mathbf{R} \beta(s)]^{k-2} \mathbf{T} \right) = \\ &= \beta^*(s) \mathbf{P} \left(\mathbf{T}^* + \mathbf{R}^* \beta(s) [\mathbf{I} - \mathbf{R} \beta(s)]^{-1} \mathbf{T} \right). \end{aligned}$$

Искомые моменты распределения могут быть получены численным дифференцированием $w(s)$ (точнее, построенного в окрестности $s = 0$ интерполяционного многочлена с последующей сменой знака у нечетных производных).

Тестирование базовой схемы на модели $M/M/n$

Описанную расчетную схему целесообразно тестировать применительно к простейшей ситуации — с показательными распределениями. При этом нет проблем с потоком заявок: все остаточные распределения остаются показательными с исходным параметром λ , а связанные с процессом обслуживания распределения суть $B_n(t) = B_n^*(t) = 1 - e^{-n\mu t}$ и соответственно $\beta(s) = \beta^*(s) = n\mu / (n\mu + s)$. При $\lambda = 1$, коэффициенте загрузки $\rho = 0,7$ и шаге $h = 10^{-3}\mu$ для построения таблицы $\omega(s)$ при интерполировании по Стирлингу были получены значения моментов времени ожидания $w_1 = 1,1479$, $w_2 = 7,3481$, $w_3 = 91,338$. Соотнесенные с аналогичными результатами при дисциплине FCFS, они дают коэффициенты роста $k_1 = 1,0000$, $k_2 = 1,3706$, $k_3 = 2,4339$. Хорошее их согласие с полученными на имитационной модели позволяет считать основной алгоритм правильным и применить его к более сложным случаям: с немарковским распределением обслуживания и произвольным рекуррентным потоком заявок.

Немарковское распределение длительности обслуживания

В данном случае получение распределения интервалов до очередного обслуживания становится самостоятельной проблемой. Дополнительная

функция распределения (ДФР) интервала от прибытия меченой заявки до ближайшего завершения обслуживания может быть подсчитана согласно

$$\overline{B}_n^*(t) = \left[\overline{B}^*(t) \right]^n. \tag{1}$$

Для работы с этой формулой вычислим моменты $B_n^*(t)$ через моменты $B^*(t)$, аппроксимируя ДФР последнего распределением Вейбулла:

$$\overline{B}^*(t) = \exp(-t^k / W) \tag{2}$$

с моментами $b_m^* = W^{m/k} \Gamma(1 + m/k)$, $m = 1, 2, \dots$ Подставляя (2) в формулу (1), убеждаемся, что интересующее нас распределение $B_n(t)$ вновь описывается формулой (2) с заменой W на W/n . Соответственно вычисляются и его моменты.

Вторым удобным вариантом аппроксимации ДФР является гиперэкспоненциальная H_2 . В этом случае параметры ДФР $\overline{B}(t) = \sum_{i=1}^2 y_i e^{-\mu_i t}$ определяются по трем моментам $\{b_i\}$ исходного распределения, а остаточного $\overline{B}^*(t) = \sum_{i=1}^2 z_i e^{-\gamma_i t}$ — по трем модифицированным $b_i^* = b_{i+1} / [(i+1)b_1]$. Соответственно:

$$\overline{B}_n^*(t) = \left(\sum_{i=1}^2 z_i e^{-\gamma_i t} \right)^n = \sum_{j=0}^n \binom{n}{j} z_1^j z_2^{n-j} e^{-[j\gamma_1 + (n-j)\gamma_2]t}.$$

Моменты этого распределения

$$b_{m,n}^* = \int_0^{\infty} t^{m-1} \overline{B}_n^*(t) dt = m! \sum_{j=0}^n \binom{n}{j} \frac{z_1^j z_2^{n-j}}{[j\gamma_1 + (n-j)\gamma_2]^m},$$

$$m = 1, 2, \dots$$

Для основного режима марковской цепи придется строить распределение интервалов между смежными завершениями обслуживания в n -канальной системе. Здесь (см. [3]) $\overline{B}_n(t) = \left[\overline{B}^*(t) \right]^{n-1} \overline{B}(t)$. В данной ситуации аппроксимация распределениями Вейбулла не имеет смысла, а гиперэкспоненциальная аппроксимация приводит к

$$\begin{aligned} \overline{B}_n(t) &= \left(\sum_{i=1}^2 z_i e^{-\gamma_i t} \right)^{n-1} \left(\sum_{i=1}^2 y_i e^{-\mu_i t} \right) = \\ &= \sum_{j=0}^{n-1} \binom{n-1}{j} z_1^j z_2^{n-1-j} e^{-[j\gamma_1 + (n-1-j)\gamma_2]t} \left(\sum_{i=1}^2 y_i e^{-\mu_i t} \right) = \\ &= \sum_{i=1}^2 y_i \sum_{j=0}^{n-1} \binom{n-1}{j} z_1^j z_2^{n-1-j} e^{-[j\gamma_1 + (n-1-j)\gamma_2 + \mu_i]t}. \end{aligned}$$

Ее моменты можно считать по формулам

$$b_{m,n} = m! \sum_{i=1}^2 y_i \sum_{j=0}^{n-1} \binom{n-1}{j} \frac{z_1^j z_2^{n-1-j}}{[j\gamma_1 + (n-1-j)\gamma_2 + \mu_i]^m},$$

$$m = 1, 2, \dots$$

Далее по найденным моментам подбирается аппроксимация, обеспечивающая удобное вычисление $\{q_j\}$. В частности, для гамма-плотности $b(t) = m(mt)^{\alpha-1} e^{-mt} / \Gamma(\alpha)$ параметры выражаются через среднее b_1 и дисперсию D согласно $\alpha = b_1^2 / D$, $m = \alpha / b_1$. Теперь

$$q_0 = \left(\frac{m}{\lambda + m} \right)^\alpha;$$

$$q_j = q_{j-1} \frac{\lambda}{\lambda + m} \frac{\alpha + j - 1}{j}, \quad j = 1, 2, \dots$$

Применение гамма-плотности с поправочным множителем, который строится на основе теории обобщенных многочленов Лагерра, позволяет выравнять произвольное число моментов. При этом сохраняется возможность рекуррентного вычисления $\{q_j\}$, но по более сложному алгоритму.

Рекуррентный поток

При рекуррентном входящем потоке общего вида приходится учитывать следующие дополнительные обстоятельства.

1. Для начального шага марковской цепи коэффициенты $\{q_j\}$ считаются как вероятности прибытия ровно j заявок рекуррентного потока за *модифицированный* интервал между смежными обслуживаниями (отсчет времени начинается с прибытия меченой заявки).

2. На последующих шагах расчет начинается с завершения очередного обслуживания, которое приходится на случайную точку интервала между смежными заявками. Соответственно, поток заявок представляется как рекуррентный с *запаздыванием*, а интервалы между обслуживаниями не модифицируются.

Рассмотрим способ расчета распределения числа событий рекуррентного потока на фиксированном интервале времени. Пусть $A_k(t)$ есть функция распределения суммы k независимых случайных величин, подчиняющихся распределению

$A_1(t) \equiv A(t)$. Очевидно, вероятность неприбытия заявок $q_0(t) = 1 - A(t) = \bar{A}(t)$, а вероятность появления ровно $k \geq 1$ требований потока $q_k(t) = A_k(t) - A_{k+1}(t)$.

Обозначим через $\alpha(s)$ ПЛС от плотности распределения $A(t)$. Тогда производящая функция ПЛС интересующих нас вероятностей

$$Q(s) = \frac{1}{s} \sum_{k=0}^{\infty} z^k [\alpha^k(s) - \alpha^{k+1}(s)] = a \frac{1 - \alpha(s)}{as} \sum_{k=0}^{\infty} z^k \alpha^k(s).$$

Здесь a есть средний интервал между смежными заявками, а дробь перед суммой — ПЛС от *остаточного* распределения интервалов между ними. Таким образом, для рекуррентного потока $q_k(t) = aL^{-1} [\alpha^*(s) \alpha^k(s)]$, $k = 0, 1, \dots$ (через L^{-1} обозначен оператор обратного ПЛС). Контрольное суммирование ПЛС дает $1/s$, откуда следует равенство суммы $\{q_k(t)\}$ единице. Произведение сверток в последней формуле имеет прозрачную вероятностную интерпретацию: чтобы за время t пришло ровно k заявок, в него должны уложиться k полных интервалов между заявками и один остаточный (последний). Напомним, что отсчет времени здесь начинается сразу после прибытия очередной заявки.

В случае рекуррентного потока с *запаздыванием* интервал времени до прибытия первой заявки является случайной модификацией типового. Это позволяет по аналогии с предыдущим вариантом сразу записать итоговые формулы в виде

$$q_0(t) = \bar{A}^*(t);$$

$$q_k(t) = aL^{-1} [(\alpha^*(s))^2 \alpha^{k-1}(s)], \quad k = 1, 2, \dots$$

Условие нормировки $\{q_j\}$ здесь проверяется (и выполняется) аналогично. Эти выражения должны интегрироваться с учетом вышеуказанных распределений длительности шагов марковской цепи.

Заключение

Результаты статьи позволяют с единых позиций подойти к расчету ПЛС распределения времени ожидания для многоканальных систем со случайным выбором заявок из очереди и произвольными распределениями длительности обслуживания и интервалов между заявками.

Литература

1. Мартин Дж. Системный анализ передачи данных: Пер. с англ. М.: Мир, 1975. Т. 2. 431 с.
2. Риордан Дж. Вероятностные системы обслуживания: Пер. с англ. М.: Связь, 1966. 164 с.
3. Саати Т. Л. Элементы теории массового обслуживания и ее приложения: Пер. с англ. М.: Сов. радио, 1965. 510 с.

4. Carter G. M., Cooper R. B. Queues with service in random order // Opns res. 1972. Vol. 20. N 2. P. 389–407.
5. Kingman J. F. C. On queues in which customers are served in random order // Proc. Camb. Phil. Soc. 1962. Vol. 58. N 1. P. 79–91.
6. Morse P. M. Queues, inventories, and maintenance. N.Y.: Wiley, 1958. 202 p.
7. Rosenlund S. I. The random order G/M/m queue // Naval Res. Logistics. 1980. Vol. 27. N 2. P. 207–215.