УДК 519.152

Г. И. Фалин

Фазовые переходы в системах обслуживания, связанные с виртуальным временем ожидания

Фазовые переходы в физических системах, т. е. переходы системы из одного состояния в другое, качественно отличное от него, изучены относительно подробно. Подобные явления встречаются и в системах массового обслуживания. Самое известное и исследованное из них — это переход от эргодичности к неэргодичности при росте поступающей нагрузки. В настоящей работе описан еще один вид фазовых переходов в системах обслуживания. Этот вид переходов будет связан с виртуальным временем ожидания. Напомним, что виртуальное время ожидания $W(t)$ в момент t в некоторой системе обслуживания — это время, в течение которого пришлось бы ожидать начала обслуживания требованию, если бы оно поступило в систему в момент t. Процесс виртуального времени ожидания является одним из наиболее важных процессов, связанных с системой обслуживания (правда, в ряде случаев естественнее работать с виртуальным временем пребывания).

В отличие от классического подхода, когда исследуют условия эргодичности (неэргодичности), расчитывают одномерные стационарные распределения $W(t)$ и т. п., основной вопрос относительно процесса $W(t)$ заключается в нахождении условий, при которых для фиксированного t $W(t)$ является собственной случайной величиной (и имеет конечное среднее).

Простейшей системой, для которой ответ на этот вопрос не является тривиальным, будет классическая система $M/G/1/\infty$ со случайным выбором на обслуживание (если в момент освобождения прибора в очереди — $n \geqslant 1$ требований, то с равной вероятностью $1/n$ любое из них может быть направлено в канал). Будем изучать эту систему в обычных предположениях — входящий поток — пуассоновский с некоторой интенсивностью λ, времена обслуживания не зависят от входящего потока, независимы в совокупности и одинаково распределены по закону $B(x)$ с конечным средним β_1. Пусть, кроме того, $\beta(s) = \int_0^\infty e^{-sx} dB(x)$ — преобразование Лапласа времени обслуживания, $\rho = \lambda\beta_1$ — загрузка системы.

Теорема 1. *Если t фиксировано, то независимо от загрузки системы виртуальное время ожидания $W(t)$ конечно почти наверное. Однако $MW(t) < \infty$ тогда и только тогда, когда $\rho < 2$.*

Доказательство. Пусть в момент ухода некоторого вызова в системе осталась очередь длины $n \geqslant 1$. Пометим один из этих вызовов и обозначим через T_n его время ожидания. Ясно, что конечность величин $W(t)$ (и их средних) при фиксированных t равносильна конечности величин T_n (и их средних) при фиксированных n. Для изучения T_n введем цепь Маркова ξ_k с пространством состояний Z_+ и следующими вероятностями перехода за один шаг: $p_{i0} = 1/i$, $p_{ii+n} = \frac{i-1}{i} k_{n+1}$, $n \geqslant -1$, $i \geqslant 1$, где $k_n = \int_0^\infty \frac{(\lambda x)^n}{n!} e^{-\lambda x} dB(x)$; переходы из состояния 0 можно задать произвольным образом.

Состояние $i \geqslant 1$ интерпретируем как наличие в системе после окончания обслуживания некоторого требования i вызовов, включая выделенное, а переход в состояние 0—как поступление выделенного вызова на обслуживание. Теперь задача сводится к изучению условий конечности числа шагов до перехода цепи ξ_k из состояния n в состояние 0. Ясно, что конечность T_k равносильна возвратности цепи ξ_k, а конечность MT_n — эргодичности цепи ξ_k. Для изучения этих вопросов удобно использовать критерии, основанные на среднем сносе.

1. Рассмотрим пробную функцию $\varphi(i) = \ln(i+1)$. Ее средний снос из состояния $i \geqslant 1$ имеет вид

$$x_i = \frac{1}{i}(\ln 1 - \ln(i+1)) + \sum_{n=0}^{\infty} \frac{i-1}{i} k_n \ln \frac{i+n}{i+1} = -\frac{\ln(i+1)}{i} +$$

$$+ \frac{i-1}{i} \sum_{n=0}^{\infty} k_n \ln\left(1 + \frac{n-1}{i+1}\right) < -\frac{\ln(i+1)}{i} + \frac{i-1}{i} \sum_{n=0}^{\infty} k_n \frac{n-1}{i+1} =$$

$$= [-\ln(i+1) + \frac{i-1}{i}(\rho - 1)]/i.$$

Поэтому финально $x_i < 0$. Используя известные критерии возвратности по среднему сносу (см., например, [1], теорема 2), можно гарантировать возвратность ξ_k при любой загрузке ρ.

2. Рассмотрим пробную функцию $\varphi(i) = i$. Ее средний снос из состояния i есть $x_i = \frac{1}{i}(0 - i) + \sum_{n=0}^{\infty} \frac{i-1}{i} k_n (i + n - 1 - i) = -1 +$

$+ \frac{i-1}{i} \sum_{n=0}^{\infty} k_n (n-1) = -1 + \frac{i-1}{i}(\rho - 1) = \rho - 2 - \frac{1}{i}(\rho - 1)$. При $i \to 2$ $x_i \to \rho - 2$. Поэтому, если $\rho < 2$, то финально $x_i \leqslant -\varepsilon < 0$ и в силу критерия Мустафы — Фостера (см. [1], теорема 2), цепь ξ_k эргодична.

Если $\rho > 2$, то финально $x_i \geqslant \varepsilon > 0$ и в силу критерия Каплана [2] цепь ξ_k неэргодична; для проверки справедливости условия Каплана удобно использовать теорему 3 работы [3]: величины $\delta_i = \sum_{j \leqslant i} \rho_{ij} (j - i) =$

$= \frac{1}{i}(-i) + \frac{i-1}{i}(k_0(-1))$ очевидно ограничены снизу.

Неэргодичность цепи ξ_k в пограничном случае $\rho = 2$ тоже можно было бы установить с помощью функций Ляпунова, но это потребовало бы гораздо более тонких рассуждений, чем приведенные выше. Поэтому применим другой метод. Он интересен и сам по себе, так как позволит очень просто получить в явном виде распределение случайной величины T_k и (в случае $\rho < 2$) MT_k (хотя распределение T_k известно давно [4], однако оно получено несравненно более громоздким методом, чем наш). При этом будем использовать метод «катастроф» (см. [5], гл. 1, § 1).

Пусть фиксировано некоторое $s > 0$ и независимо от функционирования системы наступают «катастрофы», поток которых пуассоновский с параметром s. Предположим, что в момент ухода некоторого вызова (который считаем нулевым) в системе осталась очередь длины $n \geqslant 1$. Пометим один из этих n вызовов и обозначим через $p_m^{(i)}(s)$ вероятность того, что в момент i-го ухода в системе $m \geqslant 1$ вызовов, включая выделенный, и до этого момента не наступала «катастрофа». Нетрудно видеть, что справедливо следующее основное уравнение:

$$p_m^{(i+1)}(s) = \sum_{l=2}^{m+1} p_l^{(i)}(s) \frac{l-1}{l} k_{m-l+1}(s), \quad i \geqslant 0, \quad m \geqslant 1,$$

где $k_n(s) = \int_0^{\infty} \frac{(\lambda x)^n}{n!} e^{-\lambda x} e^{-sx} dB(x)$ — вероятность того, что за время обслуживания некоторого требования поступило ровно n новых вызовов и не произошла «катастрофа». Начальное условие имеет вид $p_m^{(0)}(s) = \delta_{m,n}$. Для производящей функции $p(s, z, y) = \sum_{i=0}^{\infty} y^i \sum_{m=1}^{\infty} \frac{1}{m} z^{m-1} p_m^{(i)}(s)$ эти уравнения примут вид

$$p_z'(s, z, y)[y\beta(s + \lambda - \lambda z) - z] = p(s, z, y) - z^{n-1}. \tag{1}$$

Интересующее нас решение уравнения (1) должно быть ограничено при $|z| < 1$. Техника решения задач такого рода разработана в [6]. С ее помощью легко показать, что

$$p(s, z, y) = \begin{cases} \int_z^{\pi(s,y)} \frac{u^{n-1}}{y\beta(s + \lambda - \lambda u) - u} \exp\left\{\int_u^z \frac{1}{y\beta(s + \lambda - \lambda v) - v} dv\right\} du, \\ \qquad\qquad \text{если } z \neq \pi(s, y); \\ [\pi(s, y)]^{n-1} \qquad \text{если } z = \pi(s, y). \end{cases}$$

Здесь $\pi(s, y) = Me^{-sL}y^I$, где L — длина периода занятости, а I — число вызовов, обслуженных за период занятости.

Теперь можно найти преобразование Лапласа времени ожидания выделенного требования: $Me^{-sT_n} = P$ (за время T_n не произойдет «катастрофы») $= \sum_{i=0}^{\infty} P$ (за время T_n не произойдет «катастрофы» и перед выделенным вызовом будет обслужено ровно i требований) $= \sum_{i=0}^{\infty}\sum_{m=1}^{\infty} p_m^{(i)}(s)\frac{1}{m} =$

$$= p(s, 1, 1) = \int_1^{\pi(s)} \frac{u^{n-1}}{\beta(s+\lambda-\lambda u)-u} \exp\left\{\int_u^1 \frac{1}{\beta(s+\lambda-\lambda v)-v}\,dv\right\} du =$$

$$= \int_{\pi(s)}^{1} u^{n-1} d_u \exp\left\{\int_u^1 \frac{dv}{\beta(s+\lambda-\lambda v)-v}\right\}. \qquad (2)$$

Здесь $\pi(s) = \pi(s, 1) = Me^{-sL}$ — преобразование Лапласа длины периода занятости.

Дифференцируя (2) по s, получаем $MT_n = (n-1)\beta_1/(2-\rho)$, и поэтому при $\rho = 2MT_n = \infty$.

Теорема 1 показывает, что с точки зрения ответа на поставленный в начале работы вопрос при росте загрузки системы возникает своеобразный фазовый переход, причем его точка $\rho = 2$ не совпадает с точкой $\rho = 1$ перехода эргодичность — неэргодичность для процесса $W(t)$.

Рассуждения, проведенные при доказательстве теоремы 1, могут служить основой дальнейшего изучения времени ожидания. В качестве примера приведем следующее утверждение.

Теорема 2. *При $n \to \infty$ распределение случайной величины T_n/n слабо сходится к распределению с плотностью*

$$f(t) = \begin{cases} \beta_1^{-1}(1-(1-\rho)t/\beta_1)^{\rho/(1-\rho)}, & 0 \leqslant t \leqslant \beta_1/(1-\rho), \text{ если } \rho < 1; \\ \beta_1^{-1}e^{-t/\beta_1}, & 0 \leqslant t < +\infty, \text{ если } \rho = 1; \\ \beta_1^{-1}(1+(\rho-1)t/\beta_1)^{-\rho/(\rho-1)}, & 0 \leqslant t < +\infty, \text{ если } \rho > 1. \end{cases}$$

Доказательство. Пусть $N(t)$ — число требований в системе в момент t, $I(t)$ — число вызовов, обслуженных к моменту t. Тогда $p(s, z, y) = Me^{-sT_n}z^{N(T_n)}y^{I(T_n)}$. Для совместного преобразования Лапласа $\varphi(s, t, u) = Me^{-sT_n/n - tN(T_n)/n - uI(T_n)/n}$ нормированных величин T_n/n, $N(T_n)/n$, $I(T_n)/n$ справедливо уравнение

$$-\left[e^{-u/n}\beta\left(\frac{s}{n}+\lambda-\lambda e^{-t/n}\right)-e^{-t/n}\right]\cdot n\cdot e^{t/n}\cdot\varphi_t'(s, t, u) =$$

$$= \varphi(s, t, u) - e^{-t(n-1)/n}.$$

Этому уравнению соответствует следующее уравнение для $\psi(s, t, u) = \lim_{n\to\infty}\varphi(s, t, u)$:

$$[u + \beta_1 s + (\rho-1)t]\,\psi_t'(s, t, u) = \psi(s, t, u) - e^{-t}. \qquad (3)$$

Интересующее нас решение этого уравнения должно быть ограничено при s, t, $u \geqslant 0$. Уравнение (3) легко интегрируется (при этом, конечно, нужно различать три случая в соответствии с тем, положителен, отрицателен

или равен нулю коэффициент $\rho - 1$ при t в левой части (3)):

$$\psi(s, t, u) = \begin{cases} [u + \beta_1 s + (\rho - 1)\, t]^{1/(\rho-1)} \int\limits_t^{+\infty} e^{-x}\, [u + \beta_1 s + (\rho - 1)\, x]^{-\rho/(\rho-1)} dx, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{если } \rho > 1; \\ \dfrac{1}{1 + u + \beta_1 s}\, e^{-t}, \text{ если } \rho = 1; \\ [u + \beta_1 s - (1 - \rho)\, t]^{-1/(1-\rho)} \int\limits_t^{(u+\beta_1 s)/(1-\rho)} e^{-x} [u + \beta_1 s - \\ - (1 - \rho)\, x]^{\rho/(1-\rho)} dx, \text{ если } \rho < 1,\ 0 \leqslant t < \dfrac{u + \beta_1 s}{1 - \rho}; \\ [(1 - \rho)\, t - (u + \beta_1 s)]^{-1/(1-\rho)} \int\limits_{\frac{u+\beta_1 s}{1-\rho}}^{t} e^{-x} [(1 - \rho)\, x - \\ - (u + \beta_1 s)]^{\rho/(1-\rho)} dx, \text{ если } \rho < 1,\ \dfrac{u + \beta_1 s}{1 - \rho} < t < +\infty; \\ e^{-(u+\beta_1 s)/(1-\rho)}, \text{ если } \rho < 1,\ t = (u + \beta_1 s)/(1 - \rho). \end{cases}$$

При $t = u = 0$ отсюда получим преобразование Лапласа предельного распределения случайной величины T_n/n. Оно легко обращается, откуда и следует утверждение теоремы.

Вопрос о конечности виртуального времени ожидания $W(t)$ (и его моментов) при фиксированном t было бы интересно изучить и для других систем обслуживания. Это позволит лучше понять особенности процесса ожидания в перегруженных (с классической точки зрения) системах. Методы, примененные при доказательстве теоремы 1, могут быть использованы и для изучения более сложных систем. Рассмотрим, например, одноканальную систему с повторными вызовами типа $M/G/1/\infty$. От классической системы $M/G/1/\infty$ с ожиданием, рассмотренной выше, она отличается тем, что блокированные вызовы не выстраиваются в очередь, а образуют так называемые источники повторных вызовов, которые через экспоненциально распределенные интервалы (со средним $1/\mu$) независимо друг от друга возобновляют попытки получения обслуживания (подробнее см., например, [6]). Как и раньше, дело сводится к исследованию возвратности и эргодичности некоторой вспомогательной цепи Маркова, а именно, цепи ξ_k с фазовым пространством Z_+ и вероятностями перехода за один шаг вида $p_{i0} = \dfrac{\mu}{\lambda + i\mu}$, $p_{ii+n} = \dfrac{\lambda}{\lambda + i\mu} k_n + \dfrac{(i-1)\,\mu}{\lambda + i\mu} k_{n+1}$, $n \geqslant -1$, $i \geqslant 1$; переходы из состояния 0 можно задать произвольным образом.

Состояние $i \geqslant 1$ интерпретируем как наличие в системе после окончания обслуживания некоторого требования i источников повторных вызовов, включая некоторый выделенный, а переход в состояние 0 — как поступление выделенного источника на обслуживание.

Для пробной функции $\varphi(i) = \ln(i + 1)$ средний снос за один шаг есть $x_i = -\dfrac{\mu}{\lambda + i\mu} \ln(i + 1) + \sum\limits_{n=-1}^{\infty} \left(\dfrac{\lambda}{\lambda + i\mu} k_n + \dfrac{(i-1)\,\mu}{\lambda + i\mu} k_{n+1} \right) \ln \dfrac{i+n+1}{i+1} \leqslant$

$\leqslant \dfrac{1}{\lambda + i\mu} \left[-\mu \ln(i + 1) + \dfrac{\lambda}{i + 1} \rho + \dfrac{(i-1)\,\mu}{i + 1} (\rho - 1) \right] \leqslant 0$ при достаточно больших i. Поэтому цепь ξ_k всегда возвратна.

Для пробной функции $\varphi(i) = i$ средний снос за один шаг есть

$$x_i = -\frac{\mu}{\lambda + i\mu}\, i + \sum_{n=-1}^{\infty} \left(\frac{\lambda}{\lambda + i\mu} k_n + \frac{(i-1)\,\mu}{\lambda + i\mu} k_{n+1} \right) n = -\frac{\mu i}{\lambda + i\mu} +$$

$$+\frac{\lambda}{\lambda+i\mu}\rho+\frac{(i-1)\mu}{\lambda+i\mu}(\rho-1)\to\rho-2$$

при $i\to\infty$. Поэтому при $\rho<2$ цепь ξ_h эргодична, а при $\rho>2$ — неэргодична $\left(\delta_i=-\frac{\mu i}{\lambda+i\mu}-\frac{(i-1)\mu}{\lambda+i\mu}k_0 \text{ ограничено снизу}\right)$.

Так же, как и выше, могут быть получены явные формулы для Me^{-sT_n} и MT_n:

$$Me^{-sT_n}=\int_1^{\pi(s)}\frac{u^{n-1}}{\beta(s+\lambda-\lambda u)-u}\exp\left\{\int_u^1\frac{\mu+s+\lambda-\lambda\beta(s+\lambda-\lambda v)}{\mu[\beta(s+\lambda-\lambda v)-v]}\,dv\right\}du,$$

$$MT_n=\frac{(2+\rho)/\mu+(n-1)\beta_1}{2-\rho}\,.$$

В случае экспоненциального обслуживания последняя формула была получена в [7] (см. также [8], где изучалось число повторных попыток, которое необходимо сделать для поступления на обслуживание).

Предельное (при $n\to\infty$) распределение нормированного времени ожидания T_n/n от μ не зависит и совпадает с соответствующим распределением для системы со случайным выбором на обслуживание.

1. *Pakes A. G.* Some conditions for ergodicity and recurrence of Markov chains // Oper. Res.— 1969.— **17**.— P. 1058—1061.
2. *Kaplan M.* A sufficient condition for nonergodicity of a Markov chain // IEEE Trans. Inform. Theory.— 1979.— IT-25, N 4.— P. 470—471.
3. *Sennott L. I., Humblet P. A., Tweedie R. L.* Mean drifts and the nonergodicity of Markov, chains // Oper. Res.— 1983.— **31**, N 4.— P. 783—789.
4. *Kingman J. F. C.* On queues in which customers are served in random order // Proc. Cambridge Phil. Soc. (Math. and Phys. Scie).— 1962.— **52**, pt 1.— P. 79—91.
5. *Гнеденко Б. В. и др.* Приоритетные системы обслуживания.— М. : Изд-во Моск. ун-та, 1973.— 320 с.
6. *Фалин Г. И.* Однолинейная система с повторными вызовами // Изв. АН СССР. Техн. кибернетика.— 1979.— № 2.— С. 107—114.
7. *Falin G. I.* Single-line repeated orders queueing systems // Math. Operationsforsch. und Statist. Optim.— 1986.— № 5.— P. 649—667.
8. *Falin G. I.* On the waiting time process in a single-line queue with repeated calls // J. Appl. Probab.— 1986.— **23**, N 1.— P. 185—192.

Моск. ун-т

Получено 05.08.86