Math-Net.Ru

Общероссийский математический портал

Л. Г. Афанасьева, Т. Н. Белорусов, Предельные теоремы для систем с нетерпеливыми клиентами в условиях высокой загрузки, *Теория вероятн. и ее примен.*, 2011, том 56, выпуск 4, 788–796

DOI: <https://doi.org/10.4213/tvp4423>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением
<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 44.207.124.84

26 апреля 2023 г., 18:15:14

КРАТКИЕ СООБЩЕНИЯ

© 2011 г. АФАНАСЬЕВА Л. Г.*, БЕЛОРУСОВ Т. Н.**

ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ ДЛЯ СИСТЕМ С НЕТЕРПЕЛИВЫМИ КЛИЕНТАМИ В УСЛОВИЯХ ВЫСОКОЙ ЗАГРУЗКИ

В работе изучаются системы обслуживания с регенерирующим входным потоком и нетерпеливыми клиентами. На основе результатов, полученных для систем без ограничений, и построении мажорирующих систем доказываются предельные теоремы о сходимости стационарных распределений процессов виртуального времени ожидания и количества требований в системе к экспоненциальному. Приводятся примеры.

Ключевые слова и фразы: система обслуживания, нетерпеливые клиенты, высокая нагрузка, предельная теорема, регенерирующий поток.

1. Введение. Системы массового обслуживания с различными ограничениями (ограниченное время ожидания или пребывания, число мест для ожидания и тому подобное) привлекли внимание исследователей еще в середине прошлого века [1], [2]. Это связано не только с потенциально широкой областью приложений, но и возникающими здесь моделями, представляющими безусловный интерес для математиков. Можно выделить два основных направления развития исследований. Во-первых, изучение систем с достаточно общими входными потоками, такими как дважды стохастический пуассоновский (ДСП) [3], марковски модулированный (ММ) [4], полумарковский (ПМ) [5]. Во-вторых, асимптотический анализ операционных характеристик в условиях большой и малой загрузки [6].

В работе рассматриваются системы с нетерпеливыми клиентами, которые с вероятностью, зависящей от числа требований в очереди, навсегда покидают систему в момент своего поступления. Прикладная интерпретация может быть следующей. Каждое требование обладает неким лимитом времени. При обнаружении в системе в момент своего прихода некоторого числа других требований оно оценивает шансы попасть на обслуживание в имеющееся в его распоряжении время. Это и приводит к модели, в которой при наличии j требований в системе поступившее требование остается в ней с вероятностью f_j , а с вероятностью $1 - f_j$ уходит. Новизна предпринятого исследования в том, что входной поток требований предполагается регенерирующим. Заметим, что большинство потоков, обычно используемых в теории очередей, являются таковыми. К этому классу относится упомянутый выше ДСП поток со случайной интенсивностью, представляющий собой регенерирующий процесс, ММ поток,

* Московский государственный университет им. М. В. Ломоносова, механико-математический факультет, кафедра теории вероятностей, Ленинские горы, 119991 Москва, Россия; e-mail: afanas@mech.math.msu.su

** Московский государственный университет им. М. В. Ломоносова, механико-математический факультет, кафедра теории вероятностей, Ленинские горы, 119991 Москва, Россия; e-mail: timophey.belorusev@gmail.com

ПМ поток и некоторые другие. Описание этих потоков можно найти в [7]. Еще одно свойство регенерирующих потоков состоит в том, что при довольно общих условиях они остаются регенерирующими при прохождении через систему обслуживания. Это позволяет, например, находить условия эргодичности последовательно соединенных систем обслуживания, а также иерархических сетей, опираясь на результаты, касающиеся отдельных узлов [8].

В работе для указанного класса систем с ограничениями при регенерирующем входном потоке изучается поведение операционных характеристик (виртуальное время ожидания, число требований в системе) в ситуации высокой загрузки. Находятся условия сходимости стационарных распределений этих процессов к экспоненциальному. Доказательства опираются на построение мажорирующих процессов и результаты для систем без ограничений, полученные в [9]. Этот же подход позволяет установить при определенной нормировке C -сходимость на конечном интервале указанных процессов к диффузионному. Приводятся примеры, в которых при нарушении условий доказанных теорем имеет место сходимость к другим распределениям.

2. Описание модели. Рассматривается одноканальная система обслуживания с регенерирующим входным потоком $A(t)$.

О п р е д е л е н и е 1. Пусть $\{A(t), t \geq 0\}$ — случайный процесс с неубывающими непрерывными слева траекториями и $A(0) = 0$ (такой процесс называется потоком). Скажем, что случайный поток является регенерирующим, если существует возрастающая последовательность неотрицательных случайных величин $\{\theta_i\}_{i=1}^{\infty}$, $\theta_0 = 0$, такая, что

$$\chi_i = \{\theta_i - \theta_{i-1}, A(t) - A(\theta_{i-1}), t \in [\theta_{i-1}, \theta_i)\}, \quad i \geq 1,$$

образуют последовательность независимых одинаково распределенных случайных элементов.

Обозначим $\tau_i = \theta_i - \theta_{i-1}$, $a_i(t) = A(t) - A(\theta_{i-1})$, $\xi_i = a_i(\theta_i)$, $i \geq 1$. Моменты $\{\theta_i\}$, $i \geq 1$, являются точками регенерации потока $A(t)$, а интервалы $\{\tau_i\}$, $i \geq 1$, — периодами регенерации. Далее предполагается, что у распределений τ_1 и ξ_1 существуют по крайней мере два момента. Положим

$$a = \mathbf{E} \xi_1, \quad \mu = \mathbf{E} \tau_1, \quad \sigma_\xi^2 = \mathbf{D} \xi_1, \quad \sigma_\tau^2 = \mathbf{D} \tau_1, \quad r_{\xi, \tau} = \text{cov}(\xi_1, \tau_1).$$

Поток $A(t)$ обладает интенсивностью λ , которая определяется следующим образом $\lambda = \lim_{t \rightarrow \infty} A(t)/t$ п.н. и $\lambda = a/\mu$.

Пусть задана последовательность чисел $\{f_j\}_{j=0}^{\infty}$, $f_j \in [0, 1]$. Требование, заставшее в системе j требований, независимо от других с вероятностью f_j остается в системе до завершения обслуживания, а с вероятностью $1 - f_j$ навсегда покидает ее. Если $f_j = 1$ при $j \leq m$ для некоторого m и $f_j = 0$ при $j > m$, мы получаем систему с m местами для ожидания, при $m = 0$ систему с отказами, а при $f_j \equiv 1$ классическую модель без ограничений.

Времена обслуживания $\{\eta_j\}_{j=1}^{\infty}$ являются независимыми одинаково распределенными случайными величинами с функцией распределения $B(x)$ и конечными моментами $b = \mathbf{E} \eta_j$, $b_2 = \mathbf{E} \eta_j^2$, $\sigma_\eta^2 = b_2 - b^2$. Эта последовательность не зависит от входного потока. Введем также последовательность $\{\gamma_n\}_{n=1}^{\infty}$, где γ_n — суммарное время обслуживания требований, поступивших на n -м периоде регенерации потока $A(t)$.

Изучаются случайные процессы $Q(t)$, представляющие собой число требований в системе в момент t , и $W(t)$ — остаточное суммарное время обслуживания этих требований.

О п р е д е л е н и е 2. Случайный процесс $W(t)$ называется эргодичным, если при любом $W(0)$ существует

$$\lim_{t \rightarrow \infty} \mathbf{P}\{W(t) < x\} = \Phi(x), \quad (1)$$

где $\Phi(x)$ является функцией распределения, не зависящей от $W(0)$.

Очевидно, предел (1) и

$$\lim_{t \rightarrow \infty} \mathbf{P}\{Q(t) < x\} = F(x), \quad (2)$$

существуют одновременно, а $\Phi(x)$ и $F(x)$ либо обе являются функциями распределения, либо нет. Если $W(t)$ и $Q(t)$ эргодичны, то систему обслуживания назовем эргодичной. Положим $\rho = \lambda b$.

Для системы с нетерпеливыми клиентами в [10] показано, что когда

$$\rho \limsup_{j \rightarrow \infty} f_j < 1$$

система эргодична, а когда

$$\rho \liminf_{j \rightarrow \infty} f_j > 1,$$

процессы $W(t) \rightarrow \infty$, $Q(t) \rightarrow \infty$ при $t \rightarrow \infty$ п.н. При существовании предела $\lim_{j \rightarrow \infty} f_j = f$ система эргодична, если $\rho f < 1$, и неэргодична, если $\rho f > 1$. В случае $\rho f = 1$ система может оказаться как эргодичной, так и неэргодичной. В [10] построены соответствующие примеры. В случае, когда последовательность $\{f_j\}$ убывает, начиная с некоторого j , условие $\rho f < 1$ необходимо и достаточно для эргодичности.

Поскольку входной поток $A(t)$ регенерирующий, основную роль при доказательстве эргодических теорем играет известная теорема Смита [11]. Чтобы избежать чисто технических трудностей при ее применении, в дальнейшем будем предполагать выполненными следующие условия:

А. $\mathbf{P}\{\gamma_1 < \tau_1\} > 0$;

В. Распределение τ_1 имеет абсолютно непрерывную компоненту.

Процессы $W(t)$ и $Q(t)$ регенерирующие, и точками регенерации являются моменты регенерации входного потока, такие, что $W(\theta_i - 0) = Q(\theta_i - 0) = 0$. Условия А и В позволяют применить теорему Смита [11], из которой следует существование пределов (1) и (2), не зависящих от $W(0)$ и $Q(0)$.

3. Предельная теорема для систем с неограниченной очередью в условиях высокой загрузки. Пусть S — система без ограничений, т.е. $f_j \equiv 1$, с регенерирующим входным потоком, удовлетворяющим условиям А и В. Как показано в [12], необходимое и достаточное условие эргодичности состоит в том, что $\rho < 1$. Ситуация высокой загрузки означает, что коэффициент загрузки ρ близок к единице. Возникают два класса задач. Если $\rho < 1$ и $\rho \uparrow 1$, исследуют поведение стационарных распределений процессов $W(t)$ и $Q(t)$ (см. [9]). При произвольном ρ изучают нестационарные процессы $W(t)$ и $Q(t)$ на конечном промежутке времени. В [6] доказана весьма общая предельная теорема, устанавливающая при определенной нормировке C -сходимость этих процессов к диффузионному. На основании этой теоремы в [13] получена предельная теорема для систем с ДСП потоком. Мы начнем с задач, касающихся стационарных распределений.

Предположим, что задано семейство $\{S_\varepsilon\}$ систем обслуживания с входным потоком $A_\varepsilon(t)$, определяемым соотношением

$$A_\varepsilon(t) = A(\alpha_\varepsilon t), \quad (3)$$

где $\alpha_\varepsilon = (1 - \varepsilon)/\rho$. Тогда коэффициент загрузки S_ε равен $\rho_\varepsilon = 1 - \varepsilon$ и $\rho_\varepsilon \uparrow 1$, $\varepsilon \rightarrow 0$. Пусть $W_\varepsilon(t)$ и $Q_\varepsilon(t)$ — процессы $W(t)$ и $Q(t)$, а $\Phi_\varepsilon(x)$ и $F_\varepsilon(x)$ — функции $\Phi(x)$ и $F(x)$ для S_ε .

Теорема 1. Пусть существует $\delta > 0$, такое, что

$$\mathbf{E} \tau_1^{2+\delta} < \infty, \quad \mathbf{E} \xi_1^{2+\delta} < \infty, \quad \mathbf{E} \eta_1^{2+\delta} < \infty. \quad (4)$$

Тогда в асимптотике сжатия времени (3)

$$\Phi_\varepsilon\left(\frac{x}{\varepsilon}\right) \rightarrow 1 - \exp\left\{-\frac{2ab}{\sigma^2}x\right\}, \quad \varepsilon \rightarrow 0, \quad (5)$$

$$F_\varepsilon\left(\frac{x}{\varepsilon}\right) \rightarrow 1 - \exp\left\{-\frac{2ab^2}{\sigma^2}x\right\}, \quad \varepsilon \rightarrow 0, \quad (6)$$

где

$$\sigma^2 = a\sigma_\eta^2 + b^2\sigma_\xi^2 + \frac{a^2b^2}{\mu^2}\sigma_\tau^2 - \frac{2ab^2}{\mu}r_{\xi,\tau}. \quad (7)$$

Доказательство. Поскольку $\Phi_\varepsilon(bx/\varepsilon)$ асимптотически эквивалентна $F(x/\varepsilon)$ (см., например, [9]), достаточно доказать (5).

Введем вложенный процесс $W_n(\varepsilon) = W_\varepsilon(\theta_n - 0)$. Справедливы неравенства

$$W_n(\varepsilon) - \tau_n^\varepsilon \leq W_\varepsilon(t) \leq W_n(\varepsilon) + \gamma_n, \quad t \in [\theta_{n-1}, \theta_n) \quad \text{п.н.}$$

Из этих неравенств следует, что теорему достаточно доказать для вложенного процесса $W_n(\varepsilon)$. Далее индекс ε опускается, если это не приводит к неясности. Определим W_n^- при помощи рекуррентных соотношений

$$W_0^- = W_0, \quad W_n^- = [W_{n-1}^- + \gamma_n - \tau_n]^+,$$

так что

$$W_n^- \leq W_n, \quad n \geq 1 \quad \text{п.н.}$$

Пусть $k(n) = \max\{j < n: W(t) = 0 \text{ для некоторого } t \in [\theta_{j-1}, \theta_j)\}$, $n \geq 1$, т.е. $k(n)$ — номер последнего периода регенерации перед n -м, на котором система была свободна. Как показано в [13], с вероятностью единицы выполняется неравенство

$$W_n(\varepsilon) \leq W_{k(n)}^-(\varepsilon) + \tau_{k(n)}^\varepsilon \quad (8)$$

и

$$\varepsilon \tau_{k(n)}^\varepsilon \xrightarrow{\mathbf{P}} 0, \quad \varepsilon \rightarrow 0. \quad (9)$$

Из (8) и (9) следует, что теорему достаточно доказать для функции $\Phi_\varepsilon^-(x) = \lim_{n \rightarrow \infty} \mathbf{P}\{W_n^-(\varepsilon) < x\}$. Воспользуемся теоремой из [9, §24], которую приведем здесь в несколько измененной форме. Пусть $\{\zeta_n^\varepsilon, n \geq 1\}$ — последовательность независимых одинаково распределенных случайных величин, зависящая от параметра ε . Положим $Y_n^\varepsilon = \sum_{j=1}^n \zeta_j^\varepsilon$, $Y_0^\varepsilon = 0$, и $a_\varepsilon = \mathbf{E}\zeta_n^\varepsilon < 0$, для всех достаточно малых $\varepsilon > 0$. Тогда с вероятностью единицы случайная величина $\bar{Y}^\varepsilon = \sup_{n \geq 0} Y_n^\varepsilon$ конечна.

Теорема 2 (А. А. Боровков). *Если для некоторого $\delta > 0$ существует $\mathbf{E}|\zeta_n^\varepsilon|^{2+\delta}$, $a_\varepsilon < 0$ и $\lim_{\varepsilon \rightarrow 0} a_\varepsilon = 0$, $\lim_{\varepsilon \rightarrow 0} \mathbf{D}\zeta_n^\varepsilon = \sigma^2$, то*

$$\mathbf{P}\left\{\bar{Y}^\varepsilon > -\frac{x}{a_\varepsilon}\right\} \rightarrow \exp\left\{-\frac{2x}{\sigma^2}\right\}, \quad \varepsilon \rightarrow 0.$$

Для рассматриваемой системы положим $\zeta_n^\varepsilon = \gamma_n - \tau_n^\varepsilon$, $n \geq 1$. Тогда $\Phi_\varepsilon^-(y) = \mathbf{P}\{\bar{Y}^\varepsilon < y\}$. Для завершения доказательства теоремы остается заметить, что $\mathbf{E}\zeta_n^\varepsilon = -ba\varepsilon/(1-\varepsilon)$, а $\mathbf{D}\zeta_n^\varepsilon \rightarrow \sigma^2$, $\varepsilon \rightarrow 0$, где σ^2 задается соотношением (7).

4. Предельные теоремы для систем с нетерпеливыми клиентами в условиях высокой загрузки. Поскольку для произвольной сходящейся последовательности $\{f_j\}$ условие $\rho f < 1$ не является, вообще говоря, необходимым для эргодичности, будут рассматриваться системы с монотонно убывающей последовательностью $\{f_j\}$. Формализация условия высокой загрузки такая же, как в разд. 3, т.е. входной поток в S_ε задается формулой (3), где

$$\alpha_\varepsilon = \frac{1-\varepsilon}{\rho f}. \quad (10)$$

Процессы $W_\varepsilon(t)$ и $Q_\varepsilon(t)$ имеют тот же смысл, что и выше, а $\Phi_\varepsilon(x)$ и $F_\varepsilon(x)$ — их одномерные предельные функции распределения при $t \rightarrow \infty$ соответственно.

Теорема 3. Пусть $f_j \downarrow f$, $j \rightarrow \infty$, выполнены условия (4) теоремы 1 и

$$\sum_{j=1}^{\infty} (f_j - f) < \infty. \quad (11)$$

Тогда для системы с нетерпеливыми требованиями в асимптотике (3), (10) справедливы соотношения (5) и (6) с заменой σ^2 на σ_f^2 , где

$$\sigma_f^2 = a\sigma_\eta^2 + a(1-f)b^2 + fb^2\sigma_\xi^2 + \frac{a^2b^2f}{\mu^2}\sigma_\tau^2 - \frac{2b^2af}{\mu}r_{\xi,\tau}. \quad (12)$$

Доказательство. Рассмотрим вспомогательную систему S^+ , в которой поступающие требования разделяются на два типа. Пусть в момент j -го скачка входного потока $A(t)$ в системе S^+ находятся k требований. Новое требование присоединяется к системе и считается требованием первого типа с вероятностью f . Оно присоединяется к системе, но считается требованием второго типа с вероятностью $f_k - f$, уходит из системы с вероятностью $1 - f_k$. Предположим, что в S^+ обслуживаются лишь заявки первого типа, заявки второго типа не обслуживаются и остаются в системе навсегда. Обозначим $Q_i^+(t)$ количество заявок i -го типа в момент t в системе S^+ ($i = 1, 2$), а $Q(t)$ — количество требований в системе S с нетерпеливыми требованиями и вероятностями присоединения $\{f_j\}$.

Лемма 1. Если $Q(0) = Q_1^+(0) = Q_2^+(0) = 0$, то выполняется стохастическое неравенство

$$Q(t) \leq Q_1^+(t) + Q_2^+(t), \quad t \geq 0. \quad (13)$$

Доказательство. Используем метод одного вероятностного пространства, т.е. организуем функционирование S и S^+ таким образом, что (13) будет выполняться с вероятностью единица, а распределения всех участвующих процессов останутся прежними. Предположим, что заданы последовательные моменты скачков $\{t_n\}_{n=1}^{\infty}$ входного потока $A(t)$, а также независимые последовательности $\{e_n\}_{n=1}^{\infty}$ и $\{\eta_n\}_{n=1}^{\infty}$ независимых одинаково распределенных случайных величин. При этом e_n равномерно распределена в $[0, 1]$, а η_n имеет функцию распределения $B(x)$. Функционирование S и S^+ происходит по следующим правилам.

1. В момент t_n в обе системы приходит требование. Пусть в S (S^+) в этот момент имеется j (k) требований. Если $e_n \leq f_j$, то требование остается в S , в противном случае уходит. Если $e_n \leq f$, то требование остается в S^+ и считается требованием первого типа. При $f < e_n \leq f_k$ оно остается в S^+ и считается требованием второго типа, а при $e_n > f_k$ покидает S^+ .

2. Требованию, поступающему n -м на обслуживание в S , назначается время обслуживания η_n . Если в S и S^+ требования начинают обслуживаться одновременно, то им назначается одно и то же время обслуживания.

3. Пусть в момент поступления требования на прибор в S^+ в системе S уже происходит обслуживание некоторого требования, например, j -го. Тогда требованию в S^+ назначается время обслуживания η_j .

Рассмотрим случайный процесс

$$Z(t) = \{Q(t), \zeta(t); Q_1^+(t), Q_2^+(t), \zeta^+(t)\},$$

где $\zeta(t)$ ($\zeta^+(t)$) — остаточное время обслуживания требования на приборе в S (S^+). Как обычно, $\zeta(t)$ ($\zeta^+(t)$) не определяются, если $Q(t) = 0$ ($Q^+(t) = 0$). Состояния процесса $Z(t)$ описываются набором чисел

$$\{(j, x; k_1, k_2, y), j > 0, k_1 > 0; (0; k_1, k_2, y), k_1 > 0; \\ (j, x; 0, k_2), j > 0; (0; 0, k_2)\}.$$

Пусть $\mathcal{K}$ — множество состояний, удовлетворяющих одному из условий:

- i. $j = 0$;
- ii. $0 < j < k_1 + k_2$, $x \leq y$, $k_1 > 0$;
- iii. $j = k_1 + k_2$, $k_1 > 0$, $x = y$;
- iv. $j \leq k_2$, $k_1 = 0$.

Тогда на множестве $\mathcal{K}$ неравенство (13) выполняется. Далее индукцией по моментам поступления t_n так же, как в [10], убеждаемся, что при соблюдении перечисленных правил, процесс $Z(t)$, находясь в начальный момент в $\mathcal{K}$, навсегда останется в этом множестве, что и доказывает лемму.

Отметим также очевидное неравенство

$$Q(t) \geq Q_1^+(t). \quad (14)$$

Лемма 2. Если $f_j \downarrow f$ и выполнено (11), процесс $Q_2^+(t)$ при $t \rightarrow \infty$ стохастически ограничен.

Доказательство. Поскольку $Q_2^+(t)$ монотонно возрастает по t , достаточно доказать, что случайная величина $Q_2^+(\infty)$ конечна с вероятностью единица. Необходимо оценить вероятность события

$$\{Q_2^+(\infty) = \infty\} = \bigcap_{k=1}^{\infty} \{Q_2^+(\infty) > k\} = \bigcap_{k=1}^{\infty} \bigcup_{j>k} \{Q_2^+(\infty) = j\}.$$

Пусть D_n — событие, состоящее в том, что n -е требование останется в S^+ и будет требованием второго типа. Поскольку

$$\{Q_2^+(\infty) = j\} \subseteq \bigcup_{n=1}^{\infty} \{Q_2^+(t_n - 0) = j - 1\} \cap D_n,$$

получаем

$$\{Q_2^+(\infty) = \infty\} \subseteq \bigcap_{k=1}^{\infty} \bigcup_{j>k} \bigcup_{n=1}^{\infty} \{Q_2^+(t_n - 0) = j - 1\} \cap D_n.$$

Обозначим $\Lambda_j = \bigcup_{n=1}^{\infty} \{Q_2^+(t_n - 0) = j - 1\} \cap D_n$. Тогда

$$\{Q_2^+(\infty) = \infty\} \subseteq \limsup_{j \rightarrow \infty} \Lambda_j. \quad (15)$$

В силу того, что $\{f_j\}$ убывает, $\mathbf{P}\{\Lambda_j\} \leq f_j - f$. Из условия (11) и леммы Бореля–Кантелли следует, что $\mathbf{P}\{\limsup_{j \rightarrow \infty} \Lambda_j\} = 0$. Это вместе с (15) доказывает лемму.

Более того $\mathbf{P}\{Q_2^+(\infty) \geq k\} \leq \sum_{j=k}^{\infty} (f_j - f)$ и эта оценка верна для системы S_ε при любом ε , так что $\varepsilon Q_2^+(t, \varepsilon) \xrightarrow{\mathbf{P}} 0$, $\varepsilon \rightarrow 0$. Теперь из леммы 1 и соотношения (14) вытекает, что теорему 3 достаточно доказать для предельного (при $t \rightarrow \infty$) распределения процесса $Q_1^+(t, \varepsilon)$.

Заметим, что $Q_1^+(t)$ — это число требований в системе без ограничений с входным потоком $A_f(t)$, получающимся из процесса $A(t)$ операцией разрежения. Каждый скачок $A(t)$ независимо от других с вероятностью f является скачком $A_f(t)$, а с вероятностью $1 - f$ уничтожается. Процесс $A_f(t)$ является регенерирующим, более того, при $f < 1$ условие А для него выполняется автоматически. Периоды регенерации $\{\tau_j\}$ не зависят от f , а ξ_j^f по распределению совпадает с суммой $\sum_{i=1}^{\xi_j} \delta_i$, где $\mathbf{P}\{\delta_i = 1\} = f$, $\mathbf{P}\{\delta_i = 0\} = 1 - f$. Высокая загрузка в асимптотике сжатия времени определяется соотношениями (3), (10). Поскольку все условия теоремы 3 выполнены, остается найти σ_f^2 . Несложные вычисления приводят к (12).

В качестве примера рассмотрим систему $GI|GI|1$ с нетерпеливыми клиентами. Здесь $\{\tau_j\}_{j=1}^{\infty}$ — интервалы между поступлениями требований, а $\mathbf{P}\{\xi_j = 1\} = 1$, так что $a = 1$, $\sigma_\xi^2 = 0$, $r_{\xi, \tau} = 0$. Предположим, что τ_1 содержит абсолютно непрерывную компоненту. Справедливо следующее утверждение.

Следствие 1. Пусть для некоторого $\delta > 0$

$$\mathbf{E} \tau_1^{2+\delta} < \infty, \quad \mathbf{E} \eta_1^{2+\delta} < \infty,$$

последовательность $\{f_j\}$ убывает и выполнено (11). Тогда справедливы предельные соотношения (5) и (6) с заменой σ^2 на σ_f^2 , где

$$\sigma_f^2 = b_2 - fb^2 \left(1 - \frac{\sigma_\tau^2}{\mu^2} \right). \quad (16)$$

Для системы $M|M|1$ из (16) находим $\sigma_f^2 = 2b^2$ и

$$\Phi_\varepsilon \left(\frac{x}{\varepsilon} \right) \rightarrow 1 - e^{-\frac{x}{b}}, \quad F_\varepsilon \left(\frac{x}{\varepsilon} \right) \rightarrow 1 - e^{-x}, \quad \varepsilon \rightarrow 0. \quad (17)$$

Пример 1 (Модифицированная система с ограниченным временем ожидания). Рассмотрим одноканальную систему с показательным распределенным временем обслуживания и регенерирующим входным потоком требований двух типов. Требования первого типа присоединяются к очереди и ждут обслуживания сколь угодно долго. Требования второго типа имеют в распоряжении конечное время ожидания и при поступлении, обнаружив в системе некоторое число других требований, оценивают шансы уложиться в это время. В зависимости от полученной оценки они либо становятся в очередь и ждут обслуживания, либо навсегда покидают систему. Формальная постановка может быть следующей. Требование с вероятностью f имеет первый тип, а с вероятностью $1 - f$ второй. Для требований второго типа задана последовательность $\{\varkappa_j\}_{j=1}^\infty$ независимых одинаково распределенных случайных величин с функцией распределения $C(y)$. Если в момент поступления j -го требования в системе имеется k требований, требование второго типа присоединяется к очереди с вероятностью

$$c_k = \mathbf{P}\{\eta_1 + \dots + \eta_k < \varkappa_j\} = \int_0^\infty [1 - C(y)] \nu \frac{(\nu y)^{k-1}}{(k-1)!} e^{-\nu y} dy,$$

где $\nu = b^{-1}$ — параметр распределения η_1 . Таким образом, мы получаем описанную выше систему с нетерпеливыми клиентами и $f_k = f + (1 - f)c_k$. Последовательность f_k убывает, $f_k \rightarrow f$, $k \rightarrow \infty$, а условие (11) эквивалентно существованию y распределения $C(y)$ первого момента.

В [14] установлена диффузионная аппроксимация для процесса виртуального времени ожидания в системе без ограничений с регенерирующим входным потоком. Для систем с нетерпеливыми клиентами на основании лемм 1 и 2 из теоремы 5 в [14] вытекает следующее утверждение.

Теорема 4. Пусть выполнены условия теоремы 3 и $W_\varepsilon(0) = 0$. Тогда на каждом конечном интервале $[0, u]$ для процесса $\widehat{W}_\varepsilon(t) = \varepsilon W_\varepsilon(t/\varepsilon^2)$ имеет место C -сходимость к диффузионному процессу с отражением от нулевой границы и коэффициентами $(-1, \sigma_f^2 a/b)$, где σ_f^2 определяется (12).

5. Система $M|M|1$ с нетерпеливыми клиентами. Несколько примеров.

Цель этого пункта — проанализировать условия теоремы 3. Для этого рассматривается система с пуассоновским входным потоком интенсивности λ и экспоненциально распределенным временем обслуживания с параметром ν . Поскольку распределение числа требований в системе в стационарном режиме находится в явном виде, в некоторых случаях удастся выяснить, что происходит с этим распределением в ситуации высокой загрузки, когда какие-то условия теоремы 3 нарушаются. В сделанных предположениях $Q(t)$ представляет собой процесс размножения и гибели. Необходимое и достаточное условие эргодичности состоит в сходимости ряда

$$C = \sum_{j=1}^{\infty} \rho^j \prod_{i=0}^{j-1} f_i < \infty,$$

где $\rho = \lambda/\nu$ (см., например, [15]). Стационарное распределение имеет вид

$$P_0 = (C + 1)^{-1}, \quad P_j = (C + 1)^{-1} \rho^j \prod_{i=0}^{j-1} f_i. \quad (18)$$

Пример 2. Пусть $f_j = f(1 + 1/j)$ для $j \geq r$, где $r = [f/(1 - f)] + 1$, $f_j = 1$ при $j < r$, а $[y]$ обозначает целую часть y . Тогда $f_j \downarrow f$, $j \rightarrow \infty$, но условие (11) нарушается. Предположим, что $\rho f < 1$, так что система эргодична. Для производящей функции стационарного распределения процесса $Q(t)$ из (18) получаем выражение

$$P(z) = \mathbf{E} z^{Q(t)} = P_0 \left[\frac{(\rho z)^{r+1} - 1}{\rho z - 1} + \frac{f(\rho z)^{r+1}}{r(1 - \rho f z)^2} + \frac{f(\rho z)^{r+1}}{1 - \rho f z} \right]. \quad (19)$$

В условиях высокой загрузки $\rho_\varepsilon f = 1 - \varepsilon$ и $\varepsilon \rightarrow 0$. Из (18) и (19) несложно найти, что

$$P_0^\varepsilon = \mathbf{P}\{Q_\varepsilon(t) = 0\} = r f^r \varepsilon^2 (1 + o(1)), \quad \varepsilon \rightarrow 0. \quad (20)$$

Преобразование Лапласа

$$\mathbf{E} e^{-s \varepsilon Q_\varepsilon(t)} = P_\varepsilon(e^{-\varepsilon s}), \quad \operatorname{Re} s \geq 0,$$

также определяется из (19) и с учетом (20)

$$\lim_{\varepsilon \rightarrow 0} P(e^{-s \varepsilon}) = (1 + s)^{-2}.$$

Это означает, что распределение $\varepsilon Q_\varepsilon(t)$ сходится при $\varepsilon \rightarrow 0$ к распределению Эрланга второго порядка.

Если $f_j = f(1 + k/j)$ при $j \geq k[f/(1 - f)] + 1$ и $f_j = 1$ в остальных случаях, из (18) и (20) нетрудно установить сходимость распределения $\varepsilon Q_\varepsilon(t)$ при $\varepsilon \rightarrow 0$ к распределению Эрланга порядка $k + 1$. Это обстоятельство можно объяснить следующим образом. Если выполнено (11), то число требований второго типа в S^+ стохастически ограничено, так что длина очереди в S определяется только требованиями, которые остаются в системе независимо от длины очереди (т.е. требованиями первого типа S^+). Если ряд $\sum_{j=1}^\infty (f_j - f)$ расходится, в образовании очереди существенную роль начинают играть требования второго типа.

Пример 3. Этот пример демонстрирует, что последовательность $\{f_j\}$ может не быть убывающей, а сходимость к показательному распределению имеет место. Нетрудно установить справедливость теоремы 3 в случае, когда $\{f_j\}$ убывает, начиная с некоторого номера j_0 . Это можно сделать, используя лемму о мажорировании из [10]. Здесь мы рассмотрим периодическую последовательность $\{f_j\}$ с периодом m . Положим $\beta = f_0 \cdots f_{m-1}$. Из (18) получаем, что условие эргодичности состоит в выполнении неравенства $\beta \rho^m < 1$. Ситуация высокой загрузки предполагает, что $\rho_\varepsilon^m = (1 - \varepsilon)\beta^{-1}$, $\varepsilon \rightarrow 0$. С помощью (18) находим производящую функцию для числа требований в системе в стационарном режиме

$$P(z) = P_0 \left(1 + \frac{\gamma(z)}{1 - \beta(\rho z)^m} \right),$$

где

$$\gamma(z) = \sum_{j=1}^m (z\rho)^j \prod_{i=0}^{j-1} f_i$$

и $P_0 = (1 - \beta \rho^m)/(1 - \beta \rho^m + \gamma(1))$. Несложные выкладки приводят к асимптотическим соотношениям

$$F_\varepsilon\left(\frac{mx}{\varepsilon}\right) \rightarrow 1 - e^{-x}, \quad P_0^\varepsilon = \varepsilon/\gamma(1)(1 + o(1)), \quad \varepsilon \rightarrow 0. \quad (21)$$

Мы получаем экспоненциальное распределение в качестве предельного. Сравнение с формулой (17) для случая, когда $f_j \downarrow f$, $j \rightarrow \infty$, показывает, что в (17) нормирующий коэффициент $\varepsilon \sim 1 - \rho f$, $\varepsilon \rightarrow 0$, а в (21) имеем $\varepsilon \sim 1 - \rho^m \beta$, $\varepsilon \rightarrow 0$.

6. Заключение. Предпринятые исследования являются первым шагом в асимптотическом анализе систем обслуживания с нетерпеливыми требованиями в ситуации высокой загрузки. Получены условия, при которых предельное распределение определяется лишь «терпеливыми» требованиями, т.е. присоединяющимися к очереди независимо от ее длины. При высокой загрузке такая система в некотором смысле эквивалентна системе без ограничений. Как известно, исследование таких систем сводится к изучению верхней грани случайного блуждания с отражением в нуле. Пример 2 показывает, что предельное распределение может существовать и в том случае, когда действительно нетерпеливые требования также участвуют в образовании очереди при большой загрузке. Асимптотический анализ поведения очереди для таких систем, описание класса возможных предельных распределений приведет к постановкам новых задач, требующих для своего решения разработки новых подходов и методов.

Исследование произведено при поддержке гранта РФФИ 10-01-00266а.

СПИСОК ЛИТЕРАТУРЫ

1. Гнеденко Б. В., Коваленко И. В. Введение в теорию массового обслуживания. М.: КомКнига, 2005, 400 с.
2. *Astussen S.* Applied Probability and Queues, 2nd ed. Berlin: Springer-Verlag, 2003, 438 p.
3. *Grandell J.* Doubly Stochastic Poisson Processes. Berlin: Springer-Verlag, 1976, 234 p.
4. *Astussen S.* Ladder heights and the Markov-modulated $M|G|1$ queue. — *Stochastic Process. Appl.*, 1991, v. 37, № 2, p. 313–326.
5. *Astussen S.* Semi-Markov queues with heavy tails. — *Semi-Markov Models and Applications*, Kluwer Academic Publishers, 1999, p. 269–284.
6. Боровков А. А. Асимптотические методы в теории массового обслуживания. М.: Наука, 1980, 381 с.
7. Афанасьева Л. Г., Булинская Е. В. Надежность систем с регенерирующим потоком отказов элементов. — *Автомат. телемех.*, 2010, т. 7, с. 15–28.
8. Афанасьева Л. Г. Об эргодичности открытой сети обслуживания. — *Теория вероятн. и ее примен.*, 1987, т. 32, в. 4, с. 777–781.
9. Боровков А. А. Вероятностные процессы в теории массового обслуживания. М.: Наука, 1972, 369 с.
10. Белорусов Т. Н. Эргодичность многоканальной системы обслуживания с возможностью неприсоединения к очереди. — *Теория вероятн. и ее примен.*, 2011, т. 56, в. 1, с. 145–152.
11. *Smith W. L.* Regenerative stochastic processes. — *Proc. Roy. Soc., London. Ser. A*, 1955, v. 232, p. 6–31.
12. Афанасьева Л. Г. Системы массового обслуживания с циклическими управляющими процессами. — *Кибернетика и системный анализ*, 2005, т. 41, в. 1, с. 54–68.
13. Афанасьева Л. Г., Баиштова Е. Е. Предельные теоремы для систем обслуживания с дважды стохастическим пуассоновским потоком (условия высокой загрузки). — *Пробл. передачи информ.*, 2008, т. 44, в. 4, с. 72–91.
14. Афанасьева Л. Г., Баиштова Е. Е. Предельные теоремы для систем массового обслуживания в условиях высокой загрузки. — *Современные проблемы математики и механики*, 2009, т. 4, в. 3, с. 40–54.
15. Саати Т. Л. Элементы теории массового обслуживания и ее приложения. М.: Либроком, 2010, 520 с.

Поступила в редакцию
11.III.2011