СИСТЕМНЫЙ
АНАЛИЗ

И.Н. КОВАЛЕНКО, Е.В. КОБА

УДК 519.872

К КЛАССИФИКАЦИИ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ С ПОВТОРЕНИЕМ ВЫЗОВОВ[1]

Ключевые слова: *система массового обслуживания с повторением вызовов, орбита, цикл орбиты, система типа Лакатоша, чистая система с повторением, система с повторением и приоритетом*.

С 80-х годов XX столетия начала стремительно развиваться теория систем массового обслуживания (СМО) с повторением вызовов (возвращением заявок).

В классической теории массового обслуживания, в создание которой основополагающий вклад сделала, в частности, школа Б.В. Гнеденко, рассматривают системы без блокирования заявок; таким образом, при наличии свободного канала заявка, которая находится в системе, направляется в него немедленно. Очевидно, такие модели представляют собой идеализированную картину реальных процессов. В большинстве моделей с повторением, особенно тех, которые описывают работу современных компьютерных систем и сетей, заявка блокируется от обслуживания до момента, когда создаются условия для ее обслуживания, даже в случае свободного канала. Одним из важных типов систем с блокированием являются системы с возвращением заявок. Так, при обычной телефонной связи вызов, заставший все m приборов занятыми, получает отказ, и поэтому очередь не образуется; однако часто вызов, получивший отказ, повторяет попытку соединиться с абонентом через случайное время, т.е. в системе возникает поток повторных вызовов. После каждой неудачной попытки дозвониться к абоненту вызов (как вновь поступивший, так и повторный) может уйти из системы с определенной вероятностью.

Следует отметить, что если моделируется работа телефонных станций, то, как правило, предполагают системы с повторными вызовами, в других случаях с подобной организацией принятия на обслуживание (например, компьютерные системы и сети, аэропорты) говорят о системах с возвращением заявок.

В качестве примера системы с возвращением заявок можно привести модель системы посадки воздушных судов. Во время посадки воздушного судна в случае занятости взлетно-посадочной полосы (или по каким-либо другим причинам) оно отправляется в зону ожидания, чтобы вернуться к попытке посадки через время, кратное некоторому постоянному (в достаточном приближении) T.

Неотъемлемым атрибутом моделей функционирования компьютерных сетей и систем также является возвращающаяся заявка. В частности, функционирование сети связи случайного множественного доступа в общем случае можно описать следующим образом. Имеется некоторое количество абонентских станций (АС), которые пытаются передавать сообщения между собой, используя разделяемую среду передачи данных. Сообщение от АС, поступившее в среду и заставшее ее свободной, начинает немедленно передаваться, если же сообщение пришло в момент передачи другого, считается, что оно попало в конфликт, и оба сообщения — вновь пришедшее и передаваемое — снимаются с обслуживания и требуют повтор-

[1]Работа выполнена при поддержке МОН Украины (дог. № М/202-2009).

ной передачи. В среде передачи распространяется сигнал оповещения о конфликте. Таким образом, станции, которые отправили сообщения, но не получили подтверждения о доставке, уведомляются о том, что информацию нужно передавать заново. Сообщения, пришедшие от абонентских станций на интервале оповещения о конфликте, также требуют повторной передачи.

Отметим наиболее известные монографии [1, 2] и обзоры [3, 4] по СМО с повторением вызовов. В последнее время в общей литературе по теории массового обслуживания (ТМО) также уделяется внимание СМО с возвращением заявок (например, [5, 6]).

Ввиду широкого использования таких систем имеются определенные подходы к их классификации. Так, Янг и Темплтон в обзоре [3] дают следующее описание и кодировку СМО с возвращением заявок.

Имеется система обслуживания с входным потоком Пуассона с параметром λ и s $(s \geq 1)$ идентичными каналами обслуживания; время обслуживания для всех каналов случайное с функцией распределения $B(x)$. В системе существует $m-s$ $(m \geq s)$ мест ожидания. При наличии свободных каналов обслуживания поступившая заявка начинает обслуживаться немедленно, в противном случае при наличии свободных мест ожидания она немедленно отправляется туда. В то же время, если при поступлении заявки в системе все каналы обслуживания и все места ожидания заняты, то заявка с вероятностью $1-H_0$ навсегда покидает систему или с вероятностью H_0 покидает ее на случайный период времени, чтобы снова сделать попытку обслужиться.

О заявках, которые возвращаются в систему и делают попытку обслуживания, говорят, что они находились на орбите. Емкость орбиты O может быть конечна или бесконечна. В случае, когда O конечна и заполнена, заявка, которая пыталась попасть на орбиту, навсегда покидает систему. Допускается, что заявка, которая находится на орбите, с вероятностью $\theta\Delta t + o(\Delta t)$ будет делать попытку снова попасть в систему в интервале времени $(t, t+\Delta t)$, формируя независимый поток заявок с параметром θ. При возвращении в систему любая заявка обрабатывается системой как впервые поступившая: если есть свободные каналы или места ожидания, она соответственно обслуживается немедленно или присоединяется к очереди заявок; если все каналы обслуживания и все места ожидания заняты, она оставляет систему навсегда с вероятностью $1-H_k$ (если это k-е независимое возвращение) или отправится на орбиту (если она не заполнена) с вероятностью H_k.

В классическом обозначении систем по Кендаллу (например, [3]) СМО с возвращением заявок описываются как $A/B/s/m/O/H$, где A и B — распределения интервалов времени между поступлением заявок в систему и времени обслуживания заявок соответственно, s — число каналов обслуживания, m — количество мест в очереди плюс число каналов, O — емкость орбиты (максимальное количество заявок, которое может находиться на орбите), H означает, что модель с потерями, которые могут быть описаны рядом $(H_0, H_1, H_2, \dots)$. Когда $H_k = 1$ для $k \geq 0$, система становится системой без потерь (при выполнении некоторых условий любая заявка в конце концов будет обслужена, если только емкость орбиты O бесконечна). В этом случае H в обозначениях Кендалла записывается как *NL* (no loss — без потерь). Когда $H_k = \alpha < 1$ для $k \geq 0$, система называется системой с геометрическими потерями и H записывается как *GL* (geometric loss).

Если в обозначениях Кендалла модели СМО m, O, H отсутствуют, то полагается $m = s, O = \infty, H = NL$.

Исходя из анализа литературы по СМО с повторениями вызовов, можно сделать следующие выводы. СМО с повторением достаточно разнообразны и широко распространены. Однако практически все системы исследованы при условии показательного распределения времени пребывания на орбите, что часто не соответствует реальным системам (аэродромные системы, компьютерные системы и т.д.). В обозначениях СМО с повторением по Кендаллу никак не описывается время пребывания на орбите (его функция распределения). Таким образом, круг СМО с повторением вызовов, который охватывает обозначения из [3], недостаточно широк.

Исследуя системы с общим распределением времени пребывания на орбите, выделяем некоторый ограниченный класс систем, который удовлетворяет следующим условиям:

— данный класс достаточно широк, чтобы охватить наиболее типовые ситуации возвращения заявок;

— каждую систему из этого класса можно описать с помощью трех распределений (интервалы времени между поступлением заявок, времени обслуживания и времени их возвращения), а также алгоритма выбора требования из очереди и орбиты;

— системы, принадлежащие данному классу, достаточно сложны для анализа; они не сводятся к схеме цепей Маркова с конечным или бесконечным множеством состояний.

Перейдем к общему описанию систем. Математическая модель системы с повторением включает такие элементы: входящий поток, орбита, ограничения на ожидания, каналы обслуживания, время обслуживания, дисциплина обслуживания, выходящий поток.

Входящий поток заявок в рассматриваемых моделях — рекуррентный, независимый от процесса обслуживания. Таким образом, заявки поступают в систему в моменты $t_n, n \geq 0$, где $0 = t_0 < t_1 < \ldots$ Интервалы $\xi_n = t_n - t_{n-1}$, $n \geq 1$, — независимые случайные величины с общей функцией распределения $A(x) = \mathrm{P}\{\xi_n \leq x\}$. Среднее время между поступлением заявок $a = \mathrm{E}\{\xi_n\} < \infty$. В частности, поток может быть простейшим (стационарным пуассоновским); параметр потока обозначим $\lambda = 1/a$. Однако в отдельных моделях поток зависит от состояния системы.

Число каналов обслуживания обозначим m. Будем различать одноканальные системы $(m = 1)$ и многоканальные $(1 \leq m < \infty)$.

Орбита — виртуальная среда, в которой сосредоточиваются заявки, не получившие доступа на каналы обслуживания в момент поступления в систему. Если на орбите может находиться произвольное число заявок, то емкость орбиты не ограничена, если на орбите не могут пребывать более чем K заявок, то емкость орбиты равна K. В рамках ТМО единственной характеристикой орбиты является время возвращения с нее заявок, т.е. с формальной точки зрения орбита — элемент задержки на определенное или случайное время.

Циклом орбиты назовем период времени между попытками заявки обслужиться.

Время обслуживания n-й заявки обозначим Y_n, $B(x) = \mathrm{P}\{Y_n \leq x\}, n \geq 0$; среднее время обслуживания $\tau = \mathrm{E}\{Y_n\} < \infty$.

Введем в рассмотрение практически важный класс систем, которые назовем системами с T-возвращением и системами с $(\geq T)$-возвращением. Для первых n-я заявка может быть принята на обслуживание только в некоторый момент из множества $\{t_n + kT, k = 0, 1, 2, \ldots\}$; какой именно момент реализуется, определяется дисциплиной обслуживания. Для систем с $(\geq T)$-возвращением, если n-я заявка не принимается на обслуживание в момент t_n, то она отправляется на орбиту — в конец очереди задержанных заявок, причем эта заявка может быть принята на обслуживание не раньше момента $t_n + T$.

Обобщая понятие систем с T-возвращением, рассматриваем также системы с γ-возвращением, для которых n-я заявка может быть принята на обслуживание только в момент из множества $\{t_n, t_n + \gamma_{n1}, t_n + \gamma_{n1} + \gamma_{n2}, \ldots\}$, где время возвращения с орбиты γ_{nk} — случайная величина с функцией распределения $D(x) = \mathrm{P}\{\gamma_{nk} \leq x\}$.

Допускаем, что все случайные величины ξ_n, Y_n, γ_{nk} независимы в совокупности.

Рассмотрим дисциплину обслуживания. Систему с γ-возвращением назовем чистой RQ (RQ — retrial queue, СМО с возвращением заявок), если:

— заявка, поступившая в систему извне или с орбиты при наличии хотя бы одного канала обслуживания, принимается на обслуживание немедленно;

— заявка при занятых каналах обслуживания направляется (возможно, снова) на орбиту.

Таким образом, для чистой RQ исключается какая-либо очередность в обслуживании.

Чистая RQ с ограниченным числом циклов — система с той же дисциплиной обслуживания, что и выше, но в ней разрешается только r возвращений с орбиты. Если n-я заявка не принята на обслуживание до момента $(t_n + \gamma_{n1} + \gamma_{n2}, \ldots + \gamma_{nr})$ включительно, то эта заявка теряется в указанный момент времени.

Система типа Лакатоша (система типа L) — это система с γ-возвращением (в частности, с T-возвращением), в которой заявка, возвратившаяся с орбиты либо вновь поступившая, опять возвращается на орбиту, если еще не обслужена хотя бы одна заявка, которая поступила в систему раньше данной.

Таким образом, система типа Лакатоша объединяет в себе два принципа: повторение заявок и обслуживание в порядке очереди.

Отметим, что эти системы названы по имени венгерского математика Ласло Лакатоша, который первым начал их изучать в связи с построением модели посадки воздушных судов в аэропорту.

Профессор Лакатош рассмотрел модель посадки воздушного судна как систему обслуживания $M/M/1$ с повторением и детерминированным временем пребывания на орбите [7]. В работе [8] удалось обобщить задачу типа Лакотоша $GI/G/1$ с общим распределением времени пребывания на орбите.

Система $PRIORB$ с размещением заявок (с приоритетом задержанных заявок) — это система с γ-возвращением (в частности, с T-возвращением), в которой происходит диспетчеризация заявок, поступающих в систему. В момент t_n поступления n-й заявки становится известно, когда она будет обслужена. Это определяется на основе информации об интервалах времени, занятых обслуживанием предыдущих заявок, при условии, что данная заявка не нарушает обслуживания предыдущих.

Формализуем описание данной системы для случая $m = 1$. Пусть $t_n + W_n$ — момент поступления n-й заявки в канал обслуживания. Тогда $(t_n + W_n, t_n + W_n + Y_n)$ — интервал времени, на протяжении которого система занята обслуживанием n-й заявки. Определим k_n как число циклов n-й заявки на орбите:

$$W_n = \gamma_{n1} + \gamma_{n2} + \ldots + \gamma_{nk_n}.$$

В данных обозначениях k_n, а вместе с ним и W_n, определяется через $(W_k, k < n)$ следующим образом:

$$k_n = \min\{k \geq 0 : (t_n + \gamma_{n1} + \ldots + \gamma_{nk}, t_n + \gamma_{n1} + \ldots + \gamma_{nk} + Y_n) \cap$$
$$\cap (t_k + W_k, t_k + W_k + Y_k) = \varnothing, \ 0 \leq k < n\}.$$

Проиллюстрируем для сравнения систему $PRIORB$ с размещением заявок и систему типа Лакатоша (рис. 1) в случае $m = 1$.

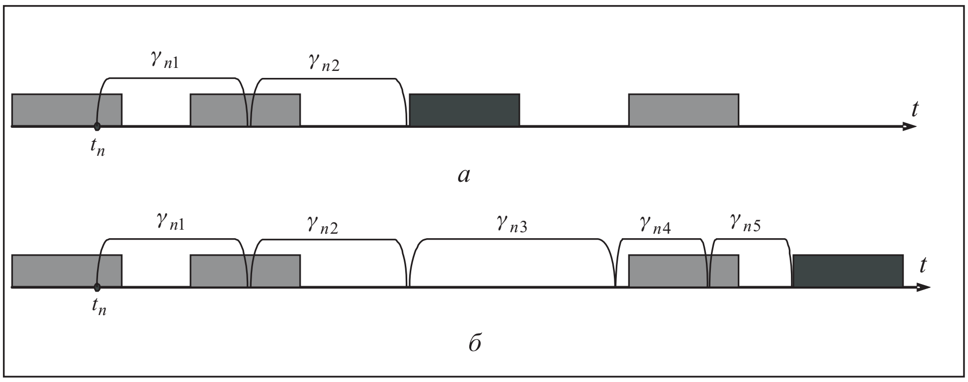

Рис. 1. Система *PRIORB* (*а*) и система типа *L* (*б*)

В обоих случаях еще необслуженные предыдущие заявки, которые размещены в интервалах времени, показаны на рис. 1 серыми прямоугольниками, а n-я заявка размещается в интервале времени, показанном черным прямоугольником.

По индукции легко показать, что если $(\xi_n, Y_n, \gamma_{nk}, n \ge 1, k \ge 1)$ в системе с размещением и системе типа Лакатоша те же самые, то для любого n имеем $W_{n(PRIORB)} \le W_{n(L)}$, откуда $\mathrm{P}\{W_{n(PRIORB)} \le x\} \ge \mathrm{P}\{W_{n(L)} \le x\}, x \in \Re$.

Заметим, что одноканальная система с $(\ge T)$-возвращением также может рассматриваться как система с размещением заявок. Для нее время ожидания W_n n-й заявки определяется следующим образом:

$$W_n = 0, \text{ если } (t_n, t_n + Y_n) \cap (t_k, t_k + Y_k) = \varnothing, k < n;$$

$$W_n = \min\{t_n + T, \max_{k<n}(t_k - t_n + W_k + Y_k)\} \text{ в противном случае.}$$

Кодируя системы с повторением по схеме Кендалла, а также используя обозначения Янга и Темплетона [3], введем еще одну позицию для обозначения функции распределения времени цикла орбиты. Так, если рассматривается система $M / M / 1 / 3 / 8 / 0{,}5 / E_2$, то она представляет собой одноканальную систему обслуживания с пуассоновским входящим потоком, экспоненциальным временем обслуживания, двумя местами в классической очереди, емкостью орбиты — 8 мест, с вероятностью 0,5 обслужиться при каждой попытке и распределением Эрланга второго порядка времени цикла орбиты. При необходимости описание систем дополняется словесно.

Отметим, что в работе [9] сведены результаты исследований эргодичности и устойчивости систем с повторением вызовов рассмотренных выше классов в различных модификациях.

В целях сравнения систем обслуживания с повторением применялся также метод статистического моделирования.

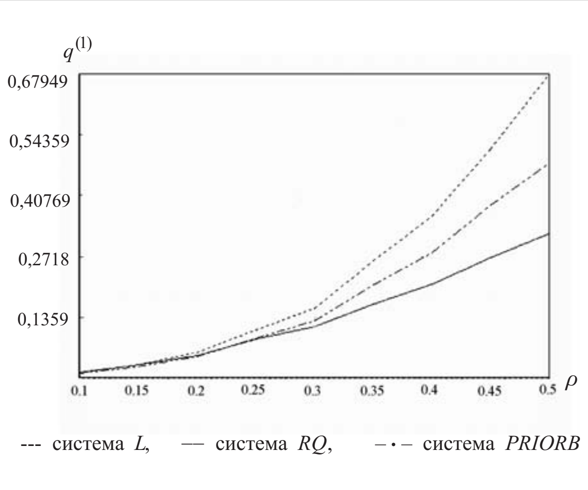

Рис. 2. Зависимость $q^{(1)}$ от ρ для систем L, RQ, $PRIORB$ ($\alpha = 1$, $0{,}1 \le \rho \le 0{,}5$)

Разработан алгоритм статистического моделирования типовых систем с повторением в пределах периода занятости, который базируется на рекуррентном алгоритме локальных изменений случайного процесса (регенеративный метод моделирования). От оценок, полученных для периода занятости для систем L, RQ, $PRIORB$ типа $GI / D / 1$ с T-возвращением, осуществляется переход к глобальному параметру: вероятности $q^{(1)}$ того, что заявке требуется по меньшей мере дважды побывать на орбите. Оценка глобального параметра дается через усреднение соответствующих величин, полученных по $n = 20000$ независимым реализациям периода занятости. Результаты моделирования приведены на рис. 2 с временем обслуживания, равным единице.

В условиях малой загрузки системы ρ для сравнения тех же систем (L, RQ, $PRIORB$ типа $GI/D/1$ с T-возвращением) был применен метод расслоенной выборки, который значительно уменьшает дисперсию оценки параметра $q^{(1)}$. Результаты моделирования приведены на рис. 3.

Заметим, что при реализации алгоритмов, разработанных как на основе регенеративного метода, так и метода расслоенной выборки, интервал времени между поступлениями заявок подчинялся Г-распределению с параметрами α и λ.

Также для систем L, RQ, $PRIORB$ типа $M/D/1$ с T-возвращением методом прямого моделирования установлено граничное значение c загрузки системы ρ такое, что при $\rho < c$ система эргодична, а при $\rho > c$ — не эргодична. В частности, эмпирично подтверждена гипотеза Л.Г. Афанасьевой, что для системы RQ с T-возвращением типа $M/D/1$ условие $\rho < 1$ достаточное для эргодичности соответствующего марковского процесса. Графики возможных зависимостей количества требований в системе от времени для систем RQ и $PRIORB$ типа $M/D/1$ с T-возвращением приведен на рис. 4–7.

При реализации алгоритмов этим методом требования поступают по закону Пуассона с параметром λ. Время их обслуживания τ постоянно $(T \geq \tau)$. Представленная выше классификация систем использовалась в [10–12].

Продолжая классификацию систем с повторением, выделим систему с синхронизацией потока заявок, которая характеризуется тем, что все заявки, поступающие в нее в полуинтервале $[(k-1)h, kh)$, задерживаются до момента $kh + lT$, $l \in N$, где T кратное h. Такая система изучалась в [13].

Выделим также систему с расщеплением потока заявок. Ее можно описать в следующей, «автобусной», ин-

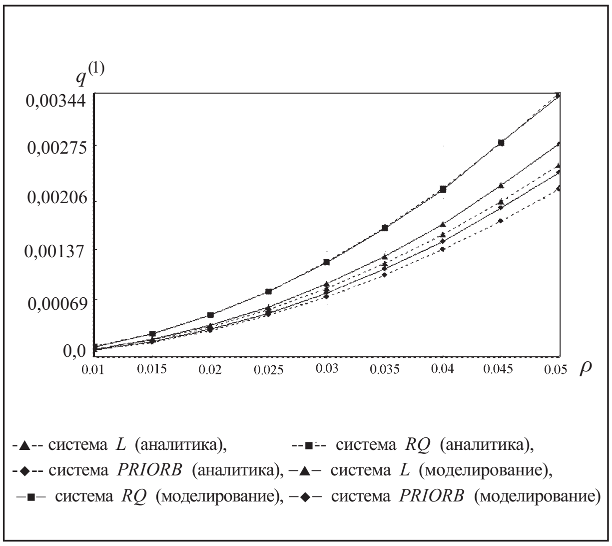

Рис. 3. Зависимость $q^{(1)}$ от ρ для систем $L, RQ, PRIORB$ ($\alpha = 1$, $0{,}01 \leq \rho \leq 0{,}05$)

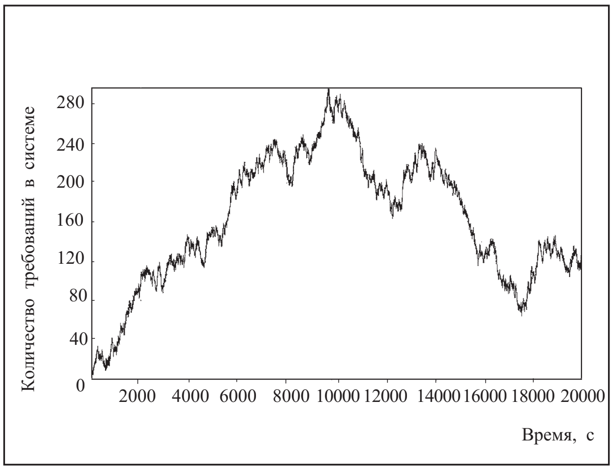

Рис. 4. Зависимость количества требований в системе RQ типа $M/D/1$ с T-возвращением от времени ($\rho = 0{,}99$; $\lambda = 1$; $T = 1{,}5$)

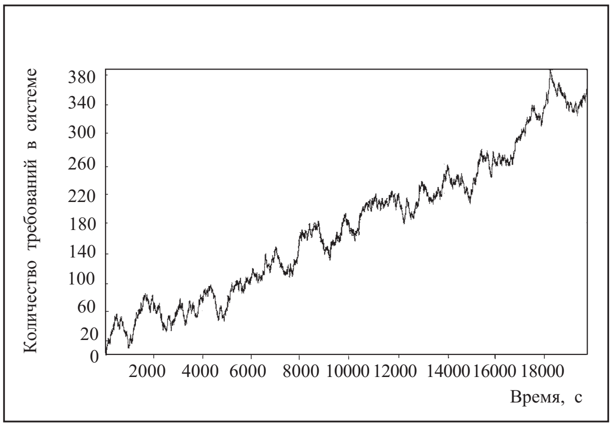

Рис. 5. Зависимость количества требований в системе RQ типа $M/D/1$ с T-возвращением от времени ($\rho = 1$; $\lambda = 1$; $T = 1{,}5$)

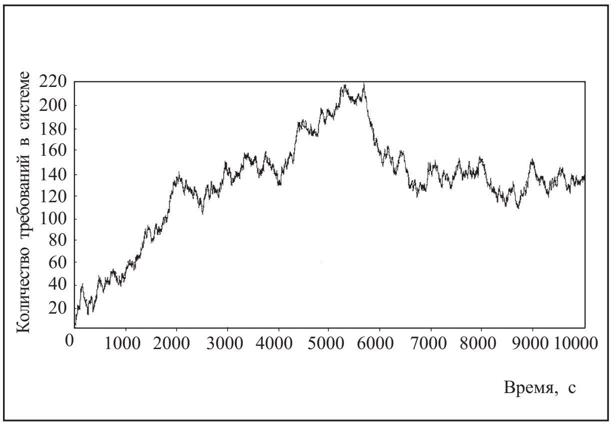

Рис. 6. Зависимость количества требований в системе *PRIORB* типа $M/D/1$ с *T*-возвращением от времени (ρ = 0,774; λ = 1; T = 1,5)

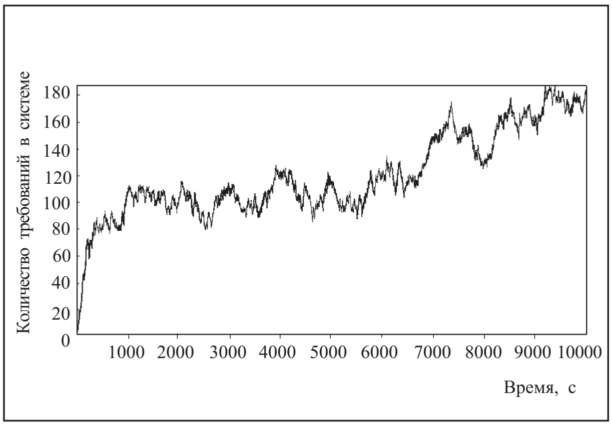

Рис. 7. Зависимость количества требований в системе *PRIORB* типа $M/D/1$ с *T*-возвращением от времени (ρ = 0,775; λ = 1; T = 1,5)

терпретации. По кольцевому маршруту, разделенному на N равных участков, равномерно движется автобус. Потоки пассажиров на каждом участке — пуассоновские, с параметром λ / N. Забрав с i-го участка одного пассажира (если он есть), водитель высаживает его через m участков и забирает пассажира с участка $(i+m)$ по модулю N. Если на i-м участке очереди нет, то автобус едет пустой по тому же алгоритму, как если бы он взял пассажира. Числа N и m подобраны так, что они взаимно просты: $(N, m)=1$; время прохождения автобусом кольцевого маршрута равно T.

Таким образом, на каждом участке высадка-посадка происходит периодически, т.е. через одинаковые интервалы времени. Такая система эквивалентна одной из разновидностей системы с расщеплением времени [10]. Она расщепляется на N независимых одноканальных СМО.

Введем понятия медленного и быстрого возвращения заявки с орбиты. Рассмотрим отношение T/τ, где τ — среднее время обслуживания заявок, T — среднее время возвращения, т.е. цикла орбиты. Для системного анализа важны все три случая, которые назовем:

— быстрое возвращение заявки: $T << \tau$, что соответствует математической модели $T/\tau \to 0$;

— умеренное возвращение заявки: τ и T — постоянные положительные величины;

— медленное возвращение заявки: $\tau << T$, что соответствует математической модели $T/\tau \to \infty$.

Необходимо подчеркнуть, что введенные понятия суть интуитивные; строгого критерия, по которому время T можно считать, например, быстрым возвращением, не существует.

Модель медленного типа интуитивно соответствует процессу посадки воздушного судна на взлетно-посадочную полосу, модель быстрого возвращения — телефонной связи, когда разговор продолжается несколько минут, а повторение происходит через несколько секунд. Однако есть и примеры умеренного времени возвращения заявки: повторения вызовов во время международной телефонной связи, где частое повторение практически невозможно.

На первый взгляд системы с быстрым возвращением заявок не имеют особого значения, поскольку при $T/\tau \to 0$ система приближается к обычной СМО с ожиданием. Однако это не так, поскольку при различных дисциплинах обслуживания граничной системой может быть или система с обслуживанием в порядке очереди, или система со случайным выбором из очереди.

Можно также привести неожиданный результат преимущества использования системы типа Лакатоша. В работе [14] показано, что при квадратичной функции ущерба, например $\sigma(x)=x^2$, при достаточно быстром возвращении заявок с орбиты средний ущерб на одну заявку для системы типа Лакатоша меньше, чем для системы типа RQ.

Следует подчеркнуть, что в применении асимптотических методов к анализу явлений природы каждое конкретное значение параметра или отношение двух параметров нельзя считать малым или большим. При этом понятие малости понимают как асимптотическую формулу, которую потом применяют для приближения конкретной величины. Так, существует теория малых колебаний математического маятника, которую широко используют в расчетах; однако каждое, конкретно взятое значение амплитуды колебаний можно считать «малым» только приближенно, в зависимости от точности расчетов. Именно в этом смысле говорим о быстром или медленном возвращении заявки с орбиты.

СПИСОК ЛИТЕРАТУРЫ

1. Falin G.I., Templeton J.G.C. Retrial queues. — London: Chapmen & Hall, 1997. — 395 p.
2. Artolejo J.A., Gomez-Corral A. Retrial queueing systems: A computational approach. — Berlin; Heidelberg: Springer-Verlag, 2008. — 318 p.
3. Yang T., Templeton J.G.C. A survey on retrial queues // Queueing Systems. — 1987. — N 3. — P. 201–233.
4. Falin G. A survey of retrial queues // Ibid. — 1990. — N 7. — P. 127–167.
5. Гнеденко Б.В., Коваленко И.Н. Введение в теорию массового обслуживания. — Изд. 3-е, испр. и доп. — М.: Комкнига, 2005. — 400 с.
6. Бочаров П.П., Печинкин А.В. Теория массового обслуживания. — М.: РУДН, 1995. — 529 с.
7. Lakatos L. On a cycling-waiting queueing problem // Theory Stoch. Proc. — 1996. — N 2 (19). — P. 176–180.
8. Koba E.V. On a $GI/G/1$ retrial queueing system with a FIFO queueing discipline // Ibid. — 2002. — N 8(24). — P. 201–207.
9. Коба Е.В. Условие устойчивости некоторых типовых систем обслуживания с возвращением заявок // Кибернетика и системный анализ. — 2005. — № 1. — С. 124–127.
10. Коба Е.В. Исследование эргодичности систем обслуживания с возвращением заявок методом статистического моделирования // Проблемы управления и информатики. — 2003. — № 2. — С. 106–115.
11. Mykhalevych K. Sensitivity of typical queueing systems: PhD thesis of London Metropolitan Univ. — London, 2005. — 224 p.
12. Пустова С.В. Аналітичні та статистичні моделі оцінювання показників ефективності функціонування call-центрів: Автореф. дис. ... канд. техн. наук. — Київ, 2009. — 19 с.
13. Коба Е.В. Система обслуживания $M/D/1$ с заявками, повторяющимися через постоянное время, при частичной синхронизации входящего потока // Кибернетика и системный анализ. — 2000. — № 6. — С. 177–180.
14. Коба О.В., Михалевич К.В. Порівняння систем типу $M/D/1$ з швидким поверненням заявок при різних дисциплінах обслуговування // Системні дослідж. та інформ. технології. — 2003. — № 2. — С. 59–68.

Поступила 26.10.2009