UDC 519.872 (045)

O. V. Koba

STABILITY OF RETRIAL QUEUEING SYSTEM M/D/1 WITH LOSSES

V. M. Glushkov Institute of Cybernetics, Kyiv, Ukraine.
E-mail: e-koba@yandex.ru

Abstract—A *stability condition is derived for a retrial queueing system with a Poisson input with parameter* λ *and constant service time* τ. *If the virtual waiting time is less than a constant a, then the call can be serviced; otherwise, it is repeated in exponentially distributed time or is lost with a probability q. The notion of system stability is defined. Two theorems are proved, defined the conditions for stability of system.*

Index Terms—Call; retrial queueing system; stability; queueing system; waiting time.

I. INTRODUCTION

Retrial (return call) queuing systems theory have rapidly developed since the 1980s.

The classical queuing theory (Gnedenko's school made a substantial contribution to its creation) consider queues without call blocking; thus, given an idle channel, a call being in the system is forwarded to this channel immediately. Such models are obviously an idealized pattern of real processes. In the majority of retrial models, especially those describing modern computer systems and networks, a call is blocked from service until conditions for its service are provided, even if there is an idle channel. An important type of blocking systems is retrial (return call) systems. For example, in ordinary telecommunication, a call that found all the *m* devices busy is rejected and, therefore, a queue does not form; however, the rejected call retries to connect to the subscriber in a random time, i.e., a flow of repeated calls forms in the system. After each failed retry, a call (either newly arrived or repeated) can leave the system with certain probability.

If a telephone station is modeled, retrial (repeated call) queues are implied; in other cases with a similar access to service (for example, computer systems and networks, airports) return call systems are assumed.

An example of a return call system may be a model of an aircraft landing system. If the runway is occupied while an airplane is landing (or for any other reasons), the aircraft is sent to a waiting zone to retry landing in a time multiple of some constant (in sufficient approximation) *T*.

A return call is also an inherent attribute of the models of computer networks and systems. For example, the operation of a random multiple-access communication network can generally be described as follows. Let there be several user stations (USs) that try to communicate with each other using a shared transmission medium. A message arrived at the medium from the US and found it free, starts to be transmitted immediately; if the message has arrived when another one is transmitted, it is assumed to face a conflict, and both messages (newly arrived and that being transmitted) cease to be served and need a repeated transmission. A conflict signal propagates in the medium. Thus, stations that sent messages without delivery confirmation are notified that the information needs to be transmitted again. Messages arrived from USs within the conflict notification interval also need retransmission.

Noteworthy are the best known monographs [1], [2] and reviews [3], [4] on retrial queues. The recent literature on queuing theory also pays attention to return call queuing systems (see, for example, [5], [6]). Retrial queueing systems are widely used for simulation of operation of telephone stations [7], computer networks, telecommunication and computer systems [8], [9], air traffic control systems [10], and many other real systems.

II. PROBLEM FORMULATION

Yang and Templeton [3] give the following description and coding of retrial queues.

Let there be a queuing system with Poisson's input flow with the parameter λ and s ($s \geq 1$) identical service channels (servers); the service time for all the channels is random with the distribution function $B(x)$. There are $m-s\,(m \geq s)$ waiting places in the system. If there are free servers, the call arrived is immediately served; otherwise, it is immediately forwarded to free waiting places (if any). At the same time, all the channels and waiting places are occupied when a calls arrives, the call leaves the system forever with probability $1-H_0$ or for a random period of time with probability H_0 and will make a new attempt to be served.

The calls that come back to the system and make an attempt to be served are said to be in the orbit. The orbit may have either finite or infinite capacity

O. If O is finite and occupied, the call that tried to get into the orbit leaves the system forever. It is assumed that a call in the orbit will retry, with probability $\theta\Delta t + o(\Delta t)$, to get into the system within the time interval $(t, t+\Delta t)$, by forming an independent call flow with the parameter θ. Any call returned to the system is processed as if it arrived for the first time: if there are free channels or waiting places, it is served immediately or is queued, respectively; if all the servers and waiting places are busy, it leaves the system forever with probability $1-H_k$ (if it is the kth independent return) or goes to the orbit (if it is free) with probability H_k.

In the classical Kendall notation (see, for example, [11]), retrial queues are described as $A/B/s/m/O/H$, where A and B are the distributions of interarrival times and service times, respectively, s is the number of servers, m is the number of places in the queue plus the number of servers, O is the orbit capacity (the maximum number of calls that may stay in the orbit), H means that it is a model with losses, which can be described by a series ($H_0, H_1, H_2 ...$). If $H_k = 1$ for $k \ge 0$, the system becomes loss-free In this case, H in Kendall's notation is written as *NL* (no loss). If $H_k = \alpha < 1$ for $k \ge 0$, the system is called a system with geometric loss and H is written as *GL* (geometric loss).

If *m, O,* and *H* are absent in Kendall's notation, it is assumed that $m= s$, $O = \infty$ and $H = NL$.

Let us consider a queueing system with one service channel and a limited demand buffer. The parameter of the input Poisson flow is λ , and the service time $\tau > 0$ is constant. Let us define the following random processes: $N(t)$ is the number of demands stored in the buffer, including that being serviced; if $N(t) > 0$, then $X(t)$ is the residual time of servicing the demand in the channel. It is obvious that a newly arriving demand should wait for service during the time $N(t)\tau + X(t)$. If this time is less than a constant a , the demand starts being serviced; otherwise, it is rejected. The rejected demand is lost with the probability q and comes back again with the probability $p = 1-q$ in a time exponentially distributed with the parameter ν . Thus, for $q > 0$ the number of attempts to obtain service is a geometrically distributed random variable γ ; $P\{\lambda = k\} = qp^{k-1}$, $k = 1,2,...$. If $q = 0$, then the number of retrials is unlimited. The times between retrials of one demand are independent. When $q = 1$, the system under consideration will have a limited (by a) waiting time. It can also be regarded as a limited-buffer system used in models of computing systems [8].

Our task is to establish the conditions of stability of the system. Note that the present paper is apparently the first to jointly allow for the boundedness of the waiting time and that of retrials.

III. Stability Conditions

Denote by $N_0(t)$ the number of available demands that have not been placed in the buffer, i.e., those that will return to the system. By the well-known terminology [l], these demands are on the orbit. Thus, there are $N_0(t) + N(t)$ demands in the system. According to the definition, by the stability of the given random process will be meant the uniform boundedness in probability

$$\forall\varepsilon > 0\ \exists N : \forall t > 0\ \{P\{N_0(t) + N(t) > N\} < \varepsilon\}.$$

For the sake of determinacy, we assume that the input flow starts at $t = 0$, i.e., $N_0(0) = N(0) = 1$.

Theorem 1. For $q > 0$, the systems is stable for any λ , τ and $a < \infty$.

Proof. Denote by t_i the arrival time. Let $t_i^{'}$.be the maximum time during which a demand exists, i.e., the instant of the last return of the demand, if it has not been placed in the buffer before. Let us define an impulse $U_i(t) = 1$ for $t_i < t < t_i^{'}$, outside this interval $U_i(t) = 0$.

Denote $\omega_i = t_i^{'} - t_i$ If the number γ_i of retrials of the given demand is equal to k , then ω_i is the sum of $k-1$exponentially distributed random variables with the parameter ν . Then the mathematical expectation of these quantities is determined by the equation

$$M\omega_i = \sum_{k=1}^{\infty} qp^{k-1}(k-1)/\nu = p/(\nu q) = A < \infty .$$

Let $\Phi(x) = P\{\omega_i > x\}$. Then

$$MN_0(0) \le \int_0^t \lambda\Phi(t-x)dx, \qquad (1)$$

where λdx is the probability of arrival of a demand in the interval dx; $\Phi(t-x)$ is the probability of existence of an impulse of the given demand at the instant t . From (1), we obtain

$$MN_0(t) \le \lambda \int_0^t \Phi(t)dt = \lambda A. \qquad (2)$$

Since $N(t)$ is limited $N(t) < c$, from (2) we have

$$P\{N_0(t) + N(t) > N\} \le (\lambda A + c)/N. \qquad (3)$$

The right-hand side of (3) tends to zero as $N \to \infty$ irrespective of t.

Theorem 2. If $q = 0$, $\rho = \lambda\tau < 1$, then for any a, $0 \le a \le \infty$ the system is stable.

Proof. For $a = \infty$ this result is well known [11], therefore, let us consider the case of finite a.

During operation of the system, alternation of unavailability intervals (when a demand cannot be placed since the virtual waiting time $W(t) > a$) and availability intervals ξ (when $W(t) < a$) takes place. Each unavailability interval lasts no more than τ. The availability interval depends on the number of rejected demands. Let at the beginning their number be $k \ge c$. Since the number of accepted demands is $N(t) \le c$, is stochastically less than the time of waiting, for arrival of c primary or secondary demands

$$\xi \le \xi_{\lambda+kv} + \xi_{\lambda+(k-1)\nu} + \ldots + \xi_{\lambda+(k-c+1)\nu},$$

where the terms in the right-hand side are independent exponentially distributed random variables with parameters designated by subscripts. From here $M\xi < c/(k-c+1)\nu$.

The probability that at least one new demand will be accepted in this interval is no greater than $\sigma = c\lambda/(k-c+1)\nu$.

These properties are also fulfilled for the rest of the availability interval after some instant *t* by virtue of the property of the exponential distribution.

Let us consider two instants of time: t and $t + m\tau$. Let, $N_0(t) = k > 2(m+c)$ and the previous history of the process to the instants t be fixed. No greater than $m+c$ demands can be accepted during the interval $(t,\ t+m\tau)$, and, therefore. The number of availability intervals is no greater than $m+c$. The probability that a repeated demand will be placed in queue in the *i*th availability interval is no less than

$$1 - \frac{c\lambda}{(k-c-i+2)\nu} \ge 1 - \frac{c\lambda}{m\nu}.$$

Let us estimate the mean value of the total length η of availability intervals

$$M\eta \le \frac{(m+c)c}{m\nu}.$$

Since service is continuous within the unavailability intervals, the mean operating time of the service channel is no less than $m\tau - \frac{(m+c)c}{m\nu}$. From here it follows that for a sufficiently large m the average number of demands accepted in the interval $(t, t+m\tau)$ is greater than $m(1-\varepsilon)$. Summing up, we obtain

$$M\{N_0(t+m\tau) - N_0(t) \mid N_0(t) = k\} \le \lambda m\tau - m(1-\varepsilon) + \frac{(m+c)c\lambda}{m\nu}, \qquad (4)$$

for $k > 2(m+c)$.

The first term on the right-hand side of (4) is the average number of new demands in the given interval, the second term is the estimate of the average number of accepted demands, and the third one (compensating for the second term) is the estimate of the average number of new demands that have been accepted. For a sufficiently large m, the right-hand side of (4) is less than $-\delta$, $(\delta > 0)$, taking into account the fact that $\lambda\tau < 1$. For any $k \ge 0$, we have

$$M\{N_0(t+m\tau) - N_0(t)\} \le \lambda m\tau. \qquad (5)$$

By virtue of the Mustafa criterion [12], stability of the process $N_0(t)$ follows from (4), (5).

IV. Conclusions

Queueing system M/D/1 with limited by constant *a* waiting time are considered. However, if the waiting time is less than some constant *a*, then the call is accepted for servicing, otherwise is rejected. A rejected call with probability *q* is lost, and with the probability $1-q$ of back through time, distributed by an exponential law. The notion of stability of the system are defined. Two theorems are proved, defined the conditions for stability of system. The results obtained can be used when designing the computer and telephone systems.

References

[1] G. I. Falin and J. G. C. Templeton, Retrial Queues, Chapmen & Hall, London, 1997.

[2] J. A. Artolejo and A. Gomez-Corral, Retrial Queueing Systems: A Computational Approach, Springer-Verlag, Berlin—Heidelberg, 2008.

[3] T. Yang and J. G. C. Templeton, "A survey on retrial queues," *Queueing Systems*, no. 3, 201–233, 1987.

[4] G. Falin, "A survey of retrial queues," *Queueing Systems*, no. 7, 127–167, 1990.

[5] J. Artalejo, "A classified bibliography of research in retrial queueing." *Progress in 1990-1999. Top.* no. 7, 1999. pp. 187–211.

[6] J. Artalejo, "A classified bibliography of research in retrial queueing." *Progress in 2000-2009. Mathematical and Computer Modeling*, vol 51, 2010, pp. 1071–1081.

[7] S. V. Pustova, "Investigation of call centers as retrial queueing systems", *Cybern. Syst. Analysis*, vol. 46, no. 3, pp. 494–499, 2010.

[8] S. F. Yashkov, Queue Analysis in a Computer, Radio i Svyaz', Moscow, 1989. [in Russian]

[9] D. Yu. Kuznetsov and A. A.Nazarov, Adaptive random access network, Deltaplan, Tomsk, 2002. [in Russian]

[10] L. Lacatos, "On a simple continuous cyclic-waiting problem." *Ann. Univ. Sci.. Budapest Sect. Comp.*, no. 14, pp. 105–113, 1994.

[11] B. V. Gnedenko and I. N. Kovalenko, An Introduction to Queuing Theory, Komkniga, Moscow, 2005. [in Russian]

[12] P.P. Bocharov and A. V. Pechenkin, Queueing Theory, Izd. RUDN. Moscow, 1995. [in Russian]

Received January 29, 2016

Koba Olena. DSc (Phys. and Math.). Associate Professor. Leading researcher.
V.M.Glushkov Institute of Cybernetics, Department of mathematical methods in reliability theory of complex systems, Kyiv, Ukraine.
Education: Kyiv State Shevchenko University (1973).
Research interests: queueing systems and networks, retrial queueing system.
Publications: 75.
E-mail: e-koba@yandex.ru

О. В. Коба. Стійкість системи масового обслуговування M/D/1 з повторенням і втратами

Виводяться умови стійкості системи масового обслуговування з повторенням заявок, вхідним потоком Пуассона λ і сталим часом обслуговування τ. Якщо віртуальний час очікування менше, ніж константа a, то виклик обслуговується, у протилежному випадку виклик повторюється через експоненціально розподілений час або губиться з імовірністю q. Наведено поняття стійкості системи. Доведено дві теореми, що визначають ці умови.

Ключові слова: виклик; система масового обслуговування з повторенням; стійкість, система масового обслуговування; час очікування.

Коба Олена Вікторівна. Доктор фіз.-мат. наук. Доцент. Провідний науковий співробітник.
Інститут кібернетики ім. В. М. Глушкова, відділ математичних методів теорії надійності складних систем, Київ, Україна.
Освіта: Київський державний університет ім. Т. Г. Шевченка (1973).
Напрям наукової діяльності: системи та мережі масового обслуговування, системи масового обслуговування з повторенням заявок.
Кількість публікацій: 75.
E-mail: e-koba@yandex.ru.

Е. В. Коба. Устойчивость системы массового обслуживания M/D/1 с повторением и потерями

Выводятся условия устойчивости системы массового обслуживания с повторением заявок, входящим потоком Пуассона λ и постоянным временем обслуживания τ. Если виртуальное время ожидания меньше, чем константа *a,* вызов обслуживается, в противном случае вызов повторяется через экспоненциально распределенное время или теряется с вероятностью *q*. Приведено понятие устойчивости системы. Доказаны две теоремы, которые определяют эти условия.

Ключевые слова: вызов, система массового обслуживания с повторением, устойчивость, система массового обслуживания, время ожидания.

Коба Елена Викторовна. Доктор физ.-мат. наук. Доцент. Ведущий научный сотрудник.
Институт кибернетики им. В. М. Глушкова, отдел математических методов теории надежности сложных систем, Киев, Украина.
Образование: Киевский государственный университет им. Т. Г. Шевченко (1973).
Направление научной деятельности: системы и сети массового обслуживания, системы массового обслуживания с повторением заявок.
Количество публикаций: 75.
E-mail: e-koba@yandex.ru.