Computers & Operations Research 61 (2015) 89–99

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Single server retrial queue with group admission of customers

A.N. Dudin [a,*], R. Manzo [b], R. Piscopo [b]

[a] Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus
[b] Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, Via Giovanni Paolo II, 132, 84084, Fisciano, SA, Italy

ARTICLE INFO

Available online 20 March 2015

Keywords:
Retrial queueing system
Group admission
Stationary distribution
Optimization

ABSTRACT

We consider a retrial queueing system with a single server and novel customer's admission discipline. The input flow is described by a Markov Arrival Process. If an arriving customer meets the server providing the service, it goes to the orbit and repeats attempts to get service in random time intervals whose duration has exponential distribution with parameter dependent on the customers number in orbit. Server operates as follows. After a service completion epoch, customers admission interval starts. Duration of this interval has phase type distribution. During this interval, primary customers and customers from the orbit are accepted to the pool of customers which will get service after the admission interval. Capacity of this pool is limited and after the moment when the pool becomes full before completion of admission interval all arriving customers move to the orbit. After completion of an admission interval, all customers in the pool are served simultaneously by the server during the time having phase type distribution depending on the customers number in the pool. Using results known for Asymptotically Quasi-Toeplitz Markov Chains, we derive stability condition of the system, compute the stationary distribution of the system states, derive formulas for the main performance measures and numerically show advantages of the considered customer's admission discipline (higher throughput, smaller average number of customers in the system, higher probability to get a service without visiting the orbit) in case of proper choice of the capacity of the pool and the admission period duration.

1. Introduction

Retrial queueing models play an important role in performance evaluation and capacity planning of many telecommunication networks including wireless communication networks and so they are intensively studied in the literature, for references see, e.g., the books [4,12] and survey paper [13].

In retrial queueing models, an arriving primary customer seeing idle servers immediately starts the service. The arriving primary customer seeing all servers busy moves to some virtual place called as orbit from which it will try to get the service later on.

The overwhelming majority of results obtained for retrial queues concern systems with a stationary Poisson arrival process. However, such an arrival process is a poor descriptor of information flows in modern telecommunication networks. The Batch Markovian Arrival Process (*BMAP*) is recommended in the literature (what is also supported by our own experience of applied research) for description of such flows. First papers where single server retrial queues with the *BMAP* and arbitrary distribution of service time were analyzed are [10] and [11] where the $BMAP/G/1$ and the $BMAP/SM/1$ retrial queues, respectively, were under study. Multi-server retrial queues with the *BMAP* and phase (*PH*) type distribution of service time were studied in [7,8,16–19,22,23].

In contrast to the queues with a buffer, in which the server is permanently busy while the queue is not empty, in retrial queues the server stays idle during a random time interval after each service completion moment until a primary customer arrives or a customer from the orbit makes an attempt to get a service. So, some time is wasted after the service of each customer. In this paper we consider customers admission discipline which allows one to reduce server wasted time (and increase system throughput) as well as to provide better quality of service for customers. These can be achieved by means of providing the service to customers not individually, but in groups. Operation time of the server alternates between two states: service providing period and customers admission period. After a service completion moment, customers admission interval starts. During this interval, primary customers and customers from the orbit are accepted to a pool of customers which will get service after this admission interval. Capacity of this pool is limited and after the moment when the pool becomes full before completion of

* Corresponding author. Tel.: +375 296482181.
E-mail addresses: dudin@bsu.by (A.N. Dudin), rmanzo@unisa.it (R. Manzo), rpiscopo@unisa.it (R. Piscopo).

http://dx.doi.org/10.1016/j.cor.2015.03.008
0305-0548/

admission interval all arriving customers move to the orbit. After completion of an admission interval, all customers presenting in the pool are served by the server simultaneously during a random time having distribution depending on the number of customers in the pool.

It is worth to note that the admission discipline we propose in this paper was not considered before in queueing literature. The main reason likely is the fact that this discipline assumes simultaneous, but not sequential, service of many customers by one server what was not motivated by possible application in important technical systems. Such an admission discipline recently became realistic in some wireless networks, e.g., in multi-rate IEEE802.11 WLAN if the length of the service period has distribution of the maximum of several independent random variables. Each of these random variables has the distribution function coinciding with the distribution function of service time of individual customer while the number of these random variables is defined by the number of customers which get the service simultaneously. Because expectation of the maximum of a fixed number of independent random variables is less than the sum of expectations of these random variables, the average time devoted to the service to an arbitrary customer under the proposed service discipline may be much less than this time under the classical service discipline. So, throughput of the system under the proposed service discipline is higher. Disadvantage of the proposed service discipline consists of the fact that a customer has no chance to start service immediately after its arrival or making a retrial. The customer has to wait some time in the pool even if the server is idle. Thus, to justify the proposed service discipline and its usefulness, it is necessary to show that, under the proper choice of the admission period duration and the pool capacity, the main performance measures of the system benefit from the proposed discipline. In particular, the latter provides higher throughput, smaller average number of customers in the system, higher probability to get a service without visiting the orbit if the parameters of the strategy are chosen in an optimal way. This motivates the analysis described in this paper.

A similar admission discipline in context of queues with discrete time was recently discussed in the paper [27]. All customers arriving to the system or making the repeated attempt during a time slot next after the service completion are admitted to the system and further be served uninterruptedly in some random order. The total service time of these customers is the sum of individual service times. Because we deal with a model in continuous time and do not have a mandatory slot during which server does not start the service, we specially introduce an admission period having a length with arbitrary distribution. Two other essential differences of our model with respect to [27] are that only a finite number of customers can be accepted by the pool and distribution of duration of the simultaneous service of customers from the pool may be arbitrary, not necessarily equal to convolution of the distributions of individual service times.

The rest of the paper is organized as follows. In Section 2, the mathematical model is described. The process of the system states as a continuous time five-dimensional Asymptotically Quasi-Toeplitz Markov chain is introduced in Section 3 and its generator in block-matrix structured form is described. In Section 4, ergodicity condition is derived in a simple intuitively tractable form. In Section 5, an algorithm for computing the stationary distribution of the system states is described in brief and the expressions for computation of the key performance measures of the system are presented. Results of numerical experiments are given in Section 6. They show advantages of the considered customer's admission discipline in case of proper choice of the pool capacity and admission period duration. Section 7 concludes the paper.

2. The mathematical model

We consider a retrial queueing system with a single server. The input flow is described by a *MAP* (Markov Arrival Process). Customer's arrival in the *MAP* is directed by an underlying irreducible continuous time Markov chain $\nu_t,\ t\geq 0$, with the finite state space $\{0,\dots,W\}$. Sojourn time of the Markov chain $\nu_t, t\geq 0$, in the state ν has exponential distribution with parameter $\lambda_\nu,\ \nu=\overline{0,W}$. Here and in the sequel notation of type $\nu=\overline{0,W}$ means that ν takes values from the set $\{0,\dots,W\}$. After this sojourn time expires, with probability $p_k(\nu,\nu')$, the process ν_t transits to the state ν', and k customers, $k=0,1$, arrive into the system. The intensities of jumps of underlying Markov chain from one state into another, which are accompanied by an arrival of k cust omers, are combined into the matrices $D_k,\ k=0,1$, of size $(W+1)\times(W+1)$. The matrix generating function of these matrices is $D(z)=D_0+D_1z,\ |z|\leq 1$. The matrix $D(1)$ is the infinitesimal generator of the process $\nu_t, t\geq 0$. The stationary distribution vector $\boldsymbol{\theta}$ of this process is the unique solution to the equations $\boldsymbol{\theta}D(1)=\mathbf{0},\ \boldsymbol{\theta}\mathbf{e}=1$. Here and in the sequel $\mathbf{0}$ is the zero row vector and $\mathbf{e}$ is the column vector of appropriate size consisting of ones. In case the dimensionality of the vector is not clear from context, it is indicated as a lower index, e.g. $\mathbf{e}_{\overline{W}}$ denotes the unit column vector of dimensionality $\overline{W}=W+1$.

The average intensity λ (fundamental rate) of the *MAP* is defined as

$$\lambda=\boldsymbol{\theta}D_1\mathbf{e}.$$

The variance v of intervals between customer arrivals is calculated as

$$v=2\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e}-\lambda^{-2},$$

the squared coefficient c_{var} of variation is equal to

$$c_{var}=2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e}-1,$$

while the correlation coefficient c_{cor} of intervals between successive arrivals is given by

$$c_{cor}=(\lambda^{-1}\boldsymbol{\theta}(-D_0)^{-1}D_1(-D_0)^{-1}\mathbf{e}-\lambda^{-2})/v.$$

For more information about the *MAP*, its special cases, properties and related research see [24] and the survey paper [9]. Usefulness of the *MAP* in modeling customers flows in telecommunication systems is mentioned in [15,20]. Among the papers devoted to the queues with the *MAP*, we can mention [1–3,5,6].

If an arriving customer meets the server providing the service, it goes to the orbit and repeats attempts to get service in random time intervals whose duration has exponential distribution. Parameter of this distribution is α_i when $i,\ i\geq 0$, customers stay in the orbit. We admit any dependence of the intensity α_i on i such as α_i is monotonically increasing when i becomes large and tends to infinity when i approaches infinity. Popular in the literature cases $\alpha_i=i\alpha$ and $\alpha_i=i\alpha+\gamma,\ \alpha>0,\ \gamma\geq 0$, satisfy the mentioned conditions.

Server operates as follows. After a service completion instant, customer's admission interval starts. Duration of this interval has *PH* type distribution with irreducible representation $(\boldsymbol{\tau},T)$. It means the following. Duration of customer's admission interval is governed by the underlying process $\eta_t^{(a)},\ t\geq 0$, which is a continuous time Markov chain with state space $\{1,\dots,M^{(a)}\}$. The initial state of the process $\eta_t^{(a)},\ t\geq 0$, at the epoch of starting the admission interval is determined by the probabilistic row-vector $\boldsymbol{\tau}=(\tau_1,\dots,\tau_{M^{(a)}})$. The transitions of the process $\eta_t^{(a)},\ t\geq 0$, that do not lead to admission interval completion, are defined by the irreducible matrix T of size $M^{(a)}\times M^{(a)}$. The intensities of transitions, which lead to admission interval completion, are defined by the vector $\mathbf{T}_0=-T\mathbf{e}$. The admission interval time distribution function has the form $T(x)=1-\boldsymbol{\tau}e^{Tx}\mathbf{e}$. Laplace–Stieltjes transform