Computers & Industrial Engineering 62 (2012) 890–897

Contents lists available at SciVerse ScienceDirect

Computers & Industrial Engineering

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

A queueing system with batch arrival of customers in sessions

Chesoong Kim [a], Alexander Dudin [b], Sergey Dudin [b,*], Valentina Klimenok [b]

[a] Sangji University, Wonju, Kangwon 220-702, Republic of Korea
[b] Belarusian State University, 4, Nezavisimosti Ave., Minsk 220030, Belarus

ARTICLE INFO

Article history:
Received 31 March 2011
Received in revised form 8 December 2011
Accepted 9 December 2011
Available online 17 December 2011

Keywords:
Queueing
Markovian arrival process
Phase type service
Session admission control
Performance modeling

ABSTRACT

This paper describes and analyzes a single-server queueing model with a finite buffer and session arrivals. Generation of the sessions is described by the Markov Arrival Process (*MAP*). Arrival of the groups of the requests within any admitted session is described by the Terminating Batch Markov Arrival Process (*TBMAP*). Service time of the request has Phase (*PH*) type distribution. The number of the sessions that can be simultaneously admitted to the system is under control.

Analysis of the joint distribution of the number of sessions and requests in the system is implemented by means of the matrix technique. Analysis of the sojourn time of an arbitrary and admitted session is performed by means of the extension of the method of catastrophes. Effect of control on the main performance measures of the system is numerically demonstrated.

1. Introduction

Queueing systems are successfully applied for performance evaluation, capacity planning, and optimization of many real world telecommunication systems. Typically, a user of a network generates not a single request but a whole bunch of requests. This is why the batch arrivals are often assumed in the analysis of queueing systems. It is usually assumed that, at a batch arrival epoch, all requests of the batch arrive to the system simultaneously.

However, nowadays, it is a typical feature of many communication networks, *IP* networks in particular, that requests arrive in batches, but the arrival of requests of a batch is not instantaneous. To distinguish the standard batches from the batches considered in this paper the latter ones are named sessions.

Session arrival is typical for multiple access telecommunication system where resources are shared by a set of users. A user establishes a session (sends the first request) when it enters the system. If this request of the user is admitted to the system, the session is considered as established. Once the user has established the session, it can generate a sequence of requests.

This scenario was described and analyzed by means of computer simulation in Kist, Lloyd-Smith, and Harris (2005) with purpose to analyze performance measures of so called *SAPOR* (*Alternative Packet Overflow Routing*) scheme of routing in *IP* networks. In this scheme, the session is called as *flow* and represents a set of requests (packets) that should be sequentially routed in the same channel. When a packet arrives, it is determined (e.g., by means of *IP* address) if the packet is part of a flow, already tracked. If the packet belongs to an existing flow, the packet is marked for transmission. If the flow is not yet tracked and the buffer and channel capacity is still available, the packet is admitted into the system and flow count is increased. Otherwise the flow is routed on the overflow link (or is dropped at all) and the packet is rejected in the considered channel. Tracked flows are cleared after they are finished. Clearing of inactive flow is done if no more packets belonging to this flow are received within a certain time interval. Tracking and clearing of flows is performed by a token mechanism. Physically, the token can be interpreted, e.g., as some timer which is switched-on at a flow admission epoch and is restarted at epochs of other requests from this flow arrival and is switched-off if some fixed timeout expires but new request from this flow does not arrive. The number of tokens (timers), which defines the maximal number of flows that can be admitted into the system simultaneously, is very important control parameter. If this number is small, the channel may be under utilised. If this number is too large, the channel may become congested. Many packets from admitted flows may be lost and level of service decreases. Simultaneously, delay and jitter of flows may increase essentially. So, the problem of defining the optimal number of tokens is practically important and non-trivial.

In the paper of Lee, Dudin, and Klimenok (2007), the model verbally described in Kist et al. (2005) was formulated and investigated in terms of the novel finite capacity queueing model of $M/M/N/R$ type with request arrivals in sessions. In paper Kim, Dudin, and Klimenok (2009), the $MAP/PH/1/N$ queueing system with session arrivals was investigated. It was assumed in Kist et al.

* Corresponding author.
E-mail addresses: dowoo@sangji.ac.kr (C. Kim), dudin@bsu.by (A. Dudin), dudin@madrid.com (S. Dudin), vklimenok@yandex.ru (V. Klimenok).

0360-8352/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cie.2011.12.010

(2005) and, then, in Lee et al. (2007) and Kim et al. (2009), that the sessions arrival is regulated by means of so called tokens. The pool of the tokens consists of K tokens and a new session is admitted to the system only if there is an available token and buffer is not full at a session arrival epoch. Otherwise, the session leaves the system permanently.

In this paper, we significantly generalize the mechanism of requests arrival within a session comparing to the model considered in Kim et al. (2009) in five directions. It was assumed in Kim et al. (2009) that: (i) the user established a session can generate further requests one by one, (ii) the intervals between requests arrival within a session are independent identically exponentially distributed random variables, and (iii) the number of requests in a session has the geometrical distribution. Here we significantly weaken all these three assumptions by suggesting that arrival of the groups of the requests from the admitted session is directed by the *TBMAP* arrival process.

Due to these generalizations, the model considered here is more complicated for analysis from the mathematical point of view, especially the analysis of the session sojourn time. Although behavior of the system is described by a finite state Markov chain, its analysis is very far from trivial because this Markov chain is multidimensional one with complicated internal structure of the generator. However, as the main contribution of the paper we consider not the presented mathematical analysis itself, but the outcome of this analysis: the adequate mathematical model of a complicated mechanism of session processing of user's requests that is typical for many modern telecommunication networks, Internet and IP networks in particular. This queueing model is well suited for description and optimization of many other real world systems, e.g., maritime terminal, logistic center, beauty salon, etc. where the service to customers (cargo ships, tracks, railway carriages, clients, etc.) is provided on basis of service contracts, contract agreements, season tickets, etc. These contracts, agreements, tickets correspond to tokens in our model and the optimal choice of the number of simultaneously running contracts allows to get maximal profit from providing a service under the predefined in a contract requirements to the quality of service, e.g., average waiting time.

The rest of the paper is organized as follows. In Section 2, the model is described. The steady-state joint distribution of the number of sessions and requests in the system is given in Section 3. In Section 4, an analysis of a session sojourn time distribution is presented. Numerical illustrations are given in Section 5.

2. Mathematical model

We consider a single server queueing system with a finite buffer of capacity $R - 1, 1 \leqslant R < \infty$.

Service time of a request is assumed having *PH* distribution. It means the following. Request service time is governed by the directing process $\eta_t,\ t \geqslant 0$, which is a continuous time Markov chain with the state space $\{1, \ldots, M\}$. The initial state of the process $\eta_t, t \geqslant 0$, at the epoch of starting the service is determined by the probabilistic row-vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$. The transitions of the process $\eta_t, t \geqslant 0$, that do not lead to service completion, are defined by the irreducible matrix S of size $M \times M$. The intensities of transitions, which lead to service completion, are defined by the vector $\mathbf{S}_0 = -S\mathbf{e}$. Here and in the sequel $\mathbf{e}$ is the column vector of appropriate size consisting of ones. The more detailed description of the *PH*-type service time distribution can be found, e.g., in Neuts (1981). Usefulness of *PH* distribution in description of service process in telecommunication networks is stated, e.g., in Pattavina and Parini (2005) and Riska, Diev, and Smirni (2002).

Requests arrive at the system in sessions. Sessions arrive at the system according to the *MAP*. Sessions arrival in the *MAP* is directed by an irreducible continuous time Markov chain $\nu_t, t \geqslant 0$, with the finite state space $\{0, \ldots, W\}$. The sojourn time of the Markov chain $\nu_t, t \geqslant 0$, in the state ν has an exponential distribution with the parameter $\lambda_\nu,\ \nu = \overline{0, W}$. Here, notation such as $\nu = \overline{0, W}$ means that ν assumes values from the set $\{0, \ldots, W\}$. After this sojourn time expires, with probability $p_k(\nu, \nu')$ the process ν_t transits to the state ν' and k sessions, $k = 0, 1$, arrive at the system. The intensities of transitions from one state to another, which are accompanied by an arrival of k sessions, are combined to the matrices $D_k, k = 0, 1$, of size $(W + 1) \times (W + 1)$. The matrix generating function of these matrices is $D(z) = D_0 + D_1 z,\ |z| \leqslant 1$. The matrix $D(1)$ is an infinitesimal generator of the process $\nu_t, t \geqslant 0$. The stationary distribution vector $\boldsymbol{\theta}$ of this process satisfies the system of equations $\boldsymbol{\theta} D(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1$. Here and in the sequel $\mathbf{0}$ is a zero row vector. In case if the dimensionality of a vector is not clear from the 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 by $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$. The variance v of session inter-arrival time is calculated by

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

the squared coefficient c_{var} of the variation is calculated by

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

while the coefficient of correlation 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 and properties and related research see Fisher and Meier-Hellstern (1993), Lucantoni (1991) and the survey paper Chakravarthy et al. (2001). Usefulness of the *MAP* in modeling information flows in modern telecommunication systems is mentioned in Heyman and Lucantoni (2003), Klemm, Lindermann, and Lohmann (2003).

By analogy with Kist et al. (2005), we assume that admission of the sessions is limited via *tokens*. The total number of available tokens is assumed to be K, $K \geqslant 1$. K can be considered as a control parameter, and various optimization problems can be solved. If no token is available or the buffer is full at a session arrival epoch, the session leaves the system permanently. Otherwise it is admitted to the system.

We assume that the first request of a session arrives at the session arrival epoch. If the session is admitted to the system, other requests from this session arrive according to the *terminating BMAP* arrival process. It means the following. The arrival of requests is directed by the continuous time Markov chain $j_t, t \geqslant 0$. This Markov chain has transient states $\{1, \ldots, J\}$ and a single absorbing state. The *TBMAP* is defined by the probabilistic row-vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_J)$ and the set of square matrices $\Delta_n, n = \overline{0, N}$, of dimension J. The initial state of the chain j_t at the epoch of the session arrival is randomly chosen according to the probabilistic distribution $(\delta_1, \ldots, \delta_J)$. Further arrivals of the requests from the session can occur at the epochs of transition of the Markov chain j_t within the set $\{1, \ldots, J\}$. Intensity of a transition from the state j to the state $j', j,\ j' = \overline{1, J}$, with generation of a group consisting of n requests is defined by the entry $(\Delta_n)_{j,j'}$ by the matrix $\Delta_n, n = \overline{1, N}$. Intensities of transition of the Markov chain $j_t$without generation of requests are defined by the non-diagonal entries of the matrix Δ_0. The diagonal entries of the matrix Δ_0 are negative and define, up to the sign, intensity of leaving the corresponding state of the Markov chain j_t. The vector $\boldsymbol{\delta}^0$ given by $\boldsymbol{\delta}^0 = -\sum_{n=0}^{N}\Delta_n\mathbf{e}$ defines the intensities of transitions to absorbing state. Transition to this state corresponds to the termination of the session.