Operations Research Perspectives 5 (2018) 245–255

Contents lists available at ScienceDirect

Operations Research Perspectives

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

Analysis of queueing model with processor sharing discipline and customers impatience

A.N. Dudin[*,a,b], S.A. Dudin[a,b], O.S. Dudina[a,b], K.E. Samouylov[b]

[a] Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., Minsk 220030, Belarus
[b] Department of Applied Probability and Informatics, RUDN University, 6, Miklukho-Maklaya st., Moscow 117198, Russia

ARTICLE INFO

Keywords:
Processor sharing
Admission control
Markovian arrival process
Impatience

ABSTRACT

Queueing systems with processor sharing represent the adequate models for sharing the resources, e.g., components of a computer or a bandwidth of communication systems. In this paper, we consider a queueing system with processor sharing discipline under quite general assumptions about the arrival and service processes. Arrivals are defined by the Markovian arrival process. The service time has a phase type distribution. Possible impatience of customers is taken into account. The number of customers, which can simultaneously obtain service, is limited. We compare two approaches for monitoring service of customers, namely, the approach counting the number of customers at each phase of service and the approach counting the phase of service of each customer and show the significant advantage of the former approach. We obtain the joint distribution of the number of customers in the system and the states of the underlying arrival and service processes as well as the loss probabilities. It is shown that the sojourn time in the system of an arbitrary customer has phase type distribution and an irreducible representation of this distribution is obtained. Numerical examples are presented. A possibility of optimal choice of the server capacity (e.g., multi-programming level) is numerically illustrated. An opportunity of increasing the speed of computations via the use of the graphics processing unit is discussed.

1. Introduction

Processor sharing discipline is very popular in computers, communication systems and networks. For references and examples of real-world applications see, e.g. [16], the surveys [32,33] as well as the recent papers [19,31]. In particular, this discipline is very popular for tasks scheduling in multi-programming computer systems. The model considered in our paper significantly extends possibility of adequate modelling of such systems. We do not impose restrictive assumptions like an exponential distribution of all times characterizing the behavior of the system and a flow of tasks as well as on the number of tasks that share the computer resources. The presented results allow to consider a task processing in a computer or communication system as a whole sequence of various operations, e.g., using CPU, GPU, RAM, I/O devices, etc, not just a single operation duration of which has an exponential distribution.

In the classical settings, a processor can be shared by the unlimited number of users and the majority of the existing literature is devoted to the analysis of queueing systems under this assumption. However, in many applications of this discipline in computer systems and communication networks this assumption is not fulfilled because a certain minimal share of the bandwidth of the computer or channel has to be guaranteed to provide acceptable quality of service to a customer. Therefore, the *limited* processor sharing or processor sharing with a finite capacity is often considered. This kind of processor sharing suggests that the maximal number, say N, $N < \infty$, of users who may obtain service simultaneously is fixed. Customers arriving when the capacity of the server is not exhausted immediately start service with the rate which is, in general, inversely proportional to the number of customers in service. The majority of the existing research is addressed to analysis of the simple $M/M/1$ type queues where it is assumed that the arrivals are described by the stationary Poisson process and the service time distribution is exponential. However, both these assumptions look quite artificial in many real-world systems. In particular, it is already well recognized that the stationary Poisson arrival process is not a good descriptor of the real-world information flows and the Markovian arrival process (*MAP*) suits much better for the description of such flows, see, e.g. [6,17,30]. An exponential distribution is a very particular case of the phase type (*PH*) distribution successfully used for approximation of an arbitrary distribution, see, e.g. [1]. In our paper, to provide the advanced model, we assume that the arrival process is defined by the *MAP* and the service time distribution is of phase type. A short list of related papers, in which at least one of the unrealistic assumptions that

* Corresponding author at: Belarusian state university, 4, Independence Ave., Minsk 220030, Belarus.
E-mail addresses: dudin@bsu.by (A.N. Dudin), dudin85@mail.ru (S.A. Dudin), dudina@bsu.by (O.S. Dudina), ksam@sci.pfu.edu.ru (K.E. Samouylov).

https://doi.org/10.1016/j.orp.2018.08.003
Received 8 June 2018; Received in revised form 19 August 2018; Accepted 20 August 2018
Available online 22 August 2018
2214-7160/ © 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

the arrivals are defined by the stationary Poisson process and the service time distribution is exponential is omitted, is as follows. The model with the infinite capacity of the server and the *MAP* is considered, e.g., in [10,18,20]. The model with the finite capacity and the *MAP* is considered, e.g., in [7,26]. It is worth to note that as a rule the problem of computation of the stationary distribution of the number of customers in the system under processor sharing discipline has a known solution which coincides with the solution for the corresponding system with first-in-first-out service discipline. The problem of computation of the sojourn time distribution is more complicated. This problem for the *M/M/*1 and *MAP/M/*1 systems with an infinite capacity was addressed in [20,34], correspondingly. The moments of the sojourn time distribution for the unreliable *MAP/M/*1 system with a finite capacity are computed in [26]. In all cited above papers, it was assumed that the service time has an exponential distribution. This assumption is more or less suitable for modelling the systems with the coefficient of variation of the service time equal to 1. However, in some real-world systems, including cellular wireless communication networks, the distribution of the service time may have higher variation, see, e.g. [23] and the hyper-exponential distribution describes the duration of holding times in such networks better. The hyper-exponential distribution as well as the Erlangian distribution is very particular case of the *PH* distribution. The model of *M/PH/*1 type with unlimited processor sharing was considered in the paper [27]. The mathematical technique exploited for analysis there can be hardly used in the case of the *MAP* arrival process.

In this paper, we consider the *MAP/PH/*1 queue with limited processor sharing. The very recent paper [28] is devoted to detailed consideration of an analogous system along with a survey of the related research. However, there are three essential differences between our paper and [28]. (i) We assume that a customer arriving when the capacity of the server is exhausted is lost while in [28] it is assumed that such a customer joins the buffer of an infinite capacity to obtain service later. It seems that the model with customer loss better suits, e.g., for modelling bandwidth sharing in wireless communication networks. (ii) In real-world systems, customers may be impatient and leave the system before service completion due to long processing. When the processor is shared by many customers, service of each customer becomes slower and importance of account of an impatience phenomenon increases. In our model, we account possible impatience of customers. (iii) We use another description of the system states by the multi-dimensional Markov chain. This description allows to compute characteristics of the system faster and for much larger capacity N of the server. E.g., even in the case when the state spaces of the underlying Markov processes of the *MAP* arrival process and the *PH* distribution consists of only two states, it is more or less realistic to compute characteristics of the system based on the classical description of the system states only for N up to 12. The effective description applied in our paper allows to make computations even for N equal to 1000.

The rest of the paper is organized as follows. In Section 2, the mathematical model of the system under study is described. The stationary distribution of the number of customers in the system is analysed in Section 3. The dynamics of the system is described by the multi-dimensional Markov chain, the generator of which is derived and equilibrium equations are written down. Formulas for the throughput of the system and the customer loss probabilities (due to the server capacity exhausting and due to impatience) are presented. In Section 4, it is shown that the sojourn time of an arbitrary customer has a phase type distribution. Section 5 contains the numerical results illustrating the dependence of the key performance measures of the system on its capacity, correlation in the arrival process and variance of the service times. An optimization problem is considered in brief. An advisability of using for computations the graphics processing unit (GPU) is discussed. Section 6 concludes the paper.

2. Description of the model

We consider a single-server queueing system without a buffer. The arrival process is the *MAP*. Arrivals are controlled by the underlying irreducible continuous-time Markov chain $\nu_t, t \geq 0$, with a finite state space $\{0, 1, ..., W\}$. The *MAP* is defined by the square matrices D_k, $k = 0, 1$, of size $W + 1$ consisting of the intensities of transitions of the Markov chain ν_t accompanied by the arrival of k customers. The matrix $D_0 + D_1$ is an infinitesimal generator of the process ν_t. The stationary distribution vector $\boldsymbol{\theta}$ of this process is the unique solution of the system $\boldsymbol{\theta}(D_0 + D_1) = \boldsymbol{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$ where $\mathbf{e}$ is a column vector consisting of1's', and $\boldsymbol{0}$ is a zero row vector. The average intensity λ (fundamental rate) of the *MAP* is given by $\lambda = \boldsymbol{\theta}D_1\mathbf{e}$. We assume that $\lambda < \infty$. For more detailed and exact definition of the *MAP* and motivation of its importance for description of the correlated bursty arrival flows in modern communication networks see [6,17,30].

The service time of an individual customer (service in absence of other customers) has a *PH* distribution with an irreducible representation $(\boldsymbol{\beta}, S)$. This service time can be interpreted as the time until the underlying Markov process $\eta_t, t \geq 0$, with a finite state space $\{1, \cdots, M, M + 1\}$ reaches the single absorbing state $M + 1$, conditioned on the fact that the initial state of this process is selected among the transient states $\{1, \cdots, M\}$ with probabilities defined by the entries of the probabilistic row vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)$. The transition rates of the process η_t within the set $\{1, \cdots, M\}$ are defined by the sub-generator S and the transition rates into the absorbing state (which leads to service completion) are given by the entries of the column vector $\boldsymbol{S}_0 = -S\mathbf{e}$. The Laplace-Stieltjes transform of the distribution having an irreducible representation $(\boldsymbol{\beta}, S)$ is defined as $\boldsymbol{\beta}(sI - S)^{-1}\boldsymbol{S}_0$, $Re\, s > 0$. For more detailed information about the *PH* distribution see [21].

The problem of constructing the matrices D_0, D_1, S and the vector $\boldsymbol{\beta}$ based on traces of real arrival and service processes is extensively addressed in the literature and may be more or less easily solved based on the results from, e.g. [4,5,22].

We assume that up to N customers can be served simultaneously. The number N is called the capacity of the server. If during an arbitrary customer arrival epoch the number of customers in service is less than N, the customer is admitted and immediately starts obtaining service. If the number of customers in service is equal to N, the arriving customer leaves the system permanently (is lost). The most well-known results relating to the systems with processor sharing assume the exponential distribution of individual customer service time. Let us denote the parameter of this distribution (rate) by μ. It is assumed that when i customers simultaneously receive service each customer is served with the rate $\frac{\mu}{i}$. Because here we assume the *PH* distribution of service time, it is necessary at first to specify the interaction of simultaneous services. It is reasonable to do this in the following way. It follows from the description of the *PH* distribution given above that the service time of a customer can be interpreted as the walking time of a customer in the open network consisting of M nodes. The customer starts walking from the node m with the probability β_m, $m = \overline{1, M}$. Here, denotation like $m = \overline{1, M}$ means that the variable m takes the values from the set $\{1, \cdots, M\}$. Then, the customer makes the transitions within this network. The intensities of the transitions are given by the entries of the matrix S. Then, the customer leaves the network with the intensities given by the entries of the column vector $\boldsymbol{S}_0$. From this interpretation, it is clear that the starting phase of the service of any customer should be chosen independently of other customers receiving service. The individual intensities of the transitions within the network during the periods when i customers present in the system are defined by the components of the sub-generator $\frac{1}{i}S$, $i = \overline{1, N}$. The intensities of transitions leading to service completion are defined by the components of the vector $\boldsymbol{S}_{0,i} = -\frac{1}{i}S\mathbf{e}$, $i = \overline{1, N}$.

It is worth to note that the presented below results can be easily extended to the case of more general, than the supposed above, inversely proportional dependence of the intensities of the transitions between the phases on the number i of customers presenting in the system.

As it was mentioned in Introduction, account of customers impatience is very important in analysis of the processor sharing discipline

Download English Version:

https://daneshyari.com/en/article/11002640

Download Persian Version:

https://daneshyari.com/article/11002640

Daneshyari.com