Performance Evaluation 91 (2015) 56–79

Contents lists available at ScienceDirect

Performance Evaluation

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

Interpolation approximations for the steady-state distribution in multi-class resource-sharing systems

A. Izagirre [b,e,*], U. Ayesta [b,c,d,e], I.M. Verloop [a,e]

[a] CNRS, IRIT, 2 rue C. Camichel, 31071 Toulouse, France
[b] CNRS, LAAS, 7 avenue du colonel Roche, 31400 Toulouse, France
[c] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
[d] UPV/EHU, University of the Basque Country, 20018 Donostia, Spain
[e] Univ. de Toulouse, INP, INSA, LAAS, 31400 Toulouse, France

ARTICLE INFO

Article history:
Available online 6 July 2015

Keywords:
Light traffic
Interpolation approximation
Discriminatory processor sharing
Random order of service

ABSTRACT

We consider a single-server multi-class queue that implements relative priorities among customers of the various classes. The discipline might serve one customer at a time in a non-preemptive way, or serve all customers simultaneously. The analysis of the steady-state distribution of the queue-length and the waiting time in such systems is complex and closed-form results are available only in particular cases. We therefore set out to develop approximations for the steady-state distribution of these performance metrics. We first analyze the performance in light traffic. Using known results in the heavy-traffic regime, we then show how to develop an interpolation-based approximation that is valid for any load in the system. An advantage of the approach taken is that it is not model dependent and hence could potentially be applied to other complex queueing models. We numerically assess the accuracy of the interpolation approximation through the first and second moments.

1. Introduction

In this paper we are interested in analyzing the steady-state performance of two multi-class single-server models: discriminatory processor sharing (DPS) and relative-priorities (RP). The behavior of both systems is determined by a vector of class-dependent weights, which we will denote by $(g_1, \ldots, g_K)$ for DPS and by $(p_1, \ldots, p_K)$ for RP. DPS is a time-sharing discipline in which all customers in the system get served simultaneously, being $\frac{g_k}{\sum_j n_j g_j}$, the fraction of the service that is allocated to a class-k customer, with n_j the number of class-j customers in the system. On the other hand RP operates in a non-preemptive manner, and the probability that the next customer to be served is from class k is given by $\frac{n_k p_k}{\sum_j n_j p_j}$. The intra-class scheduling discipline under RP can be any non-anticipating policy, e.g. First Come First Served (FCFS), Last Come First Served (LCFS), or Random Order of Service (ROS).

Both DPS and RP are versatile queueing models providing a natural framework to model service differentiation in systems. DPS is a multi-class extension of the well-studied egalitarian Processor Sharing (PS) policy, where the various classes are assigned positive weight factors. The DPS queue has received lot of attention due to its application to model the performance

* Corresponding author at: CNRS, LAAS, 7 avenue du colonel Roche, 31400 Toulouse, France.
E-mail address: izagirreane0@gmail.com (A. Izagirre).

http://dx.doi.org/10.1016/j.peva.2015.06.005
0166-5316/

of bandwidth sharing policies in communication networks, see for example [1–4]. The RP model can have applications in various domains, in particular in ATM networks [5], telecommunication networks [6], or genetic networks, where molecules are analogous to customers, the enzyme is analogous to the server and protein species correspond to classes, see [7].

The exact analysis of both DPS and RP is difficult, and closed-form results are scarce and exist only under limiting assumptions. For DPS with *exponential service time distributions*, in [8] the authors established that the generating function of the queue length vector satisfies a differential equation. From this equation, the authors further show that the moments can be determined numerically as the solution of a system of equations. RP is more amenable to analyze because it is non-preemptive. In [9] the authors established for *general service requirements* a set of equations for the generating function of the queue length vector *and* the Laplace–Stieltjes Transform (LST) of the waiting time. For both DPS and RP, a closed-form expression for the mean queue length is available only for the case of two classes, see [10] for DPS (with exponential service times) and [11] for RP. The heavy-traffic limits for DPS and RP have been studied in [12–14]. For both models it has been shown that a so-called "state-space collapse" appears, which describes that the queue lengths of the various classes become proportional in the heavy-traffic regime.

Motivated by the difficulty in analyzing both systems in exact form, in this paper we derive closed-form approximations for the steady-state distribution of the queue length vector and waiting time. We have chosen these metrics since they are among the most frequently considered measures in the performance evaluation literature. More precisely, we will first investigate the performance of both systems in light traffic, that is, when the arrival rate tends to 0. This approach was pioneered in a series of papers by Reiman & Simon, see for example [15], where the objective was the mean number of customers or mean sojourn time, and extended to the distribution of the sojourn time for Markovian queues in [16,17]. In one of our main contributions, we will derive the distribution of performance metrics under DPS in a light-traffic regime for general service times. We emphasize that in that case no analytical characterizations are available for DPS. In the case of RP, we will show that the light-traffic approximation can be obtained directly from the differential equations obtained in [9]. We will then combine our light-traffic approximations with the heavy-traffic characterization in order to develop an interpolation approximation that aims at capturing the performance for any load. We investigate the accuracy of our approximations for several service time distributions to illustrate the applicability of the approach.

We note that this paper is a generalization of [18] where we developed closed-form approximations for the mean conditional and unconditional sojourn times for the DPS policy. The main result in [18, Proposition IV.1] is a particular case of Proposition 6.6, as described in Section 6.3.

The remainder of the paper is organized as follows. In Section 2 we provide a short overview of the related literature. In Section 3 we present the main modeling assumptions and notation used in this paper. In Section 4 we provide a detailed explanation of how to obtain the light-traffic derivatives and how to build the interpolation approximation. Sections 5 and 6 focus on the RP model and the DPS model, respectively. We first introduce the known results from the literature (including known heavy-traffic results), and then explain how to derive the light-traffic approximation and the interpolation approximation. In Section 7 we numerically illustrate the accuracy of our approximations.

2. Related work

In this section we present a brief overview of the main results available on the models DPS and RP, and on light-traffic approximations.

The DPS model was introduced by Kleinrock in [19]. Despite the simplicity of the model description and the fact that the properties of the egalitarian Processor-Sharing queue (equal weights) are quite thoroughly understood, the analysis of DPS has proven to be extremely difficult. In a seminal paper Fayolle et al. [10] studied the mean conditional (on the service requirement) and unconditional sojourn time. For general service time distributions, the authors obtained the mean conditional sojourn time as the solution of a system of integro-differential equations. Asymptotics of the sojourn time have received considerable attention for example in [20,21]. Time-scale separations have been studied in [22,23]. The performance of DPS in overload and its application to model TCP flows is considered in [24]. The application of DPS to analyze the performance of TCP is also considered in [1] and for more applications of DPS in communication networks see [2–4]. DPS under a heavy-traffic regime (when the traffic load approaches the available capacity) was analyzed in Grishechkin [13] assuming finite second moments of the service requirement distributions. Subsequently, assuming exponential service requirement distributions, a direct approach to establish a heavy-traffic limit for the joint queue length distribution was described by Rege & Sengupta [8] and extended to *phase-type* distributions in [14]. For an overview of the literature on DPS we refer to the survey [25].

A special case of RP is when the intra-class scheduling discipline is uniformly random, that is, within a class a customer is selected randomly. This model was proposed in [26] and it is referred to as discriminatory-random-order-of-service (DROS). In recent years several interesting studies have been published on DROS [11,27,9,12]. Expressions for the mean waiting time of a customer given its class have been obtained in [11]. In [27,9] the authors derive differential equations that the transform of the joint queue lengths and the waiting time in steady-state must satisfy, respectively, and this allows the authors to find the moments of the queue lengths as a solution of linear equations. In [12] the authors obtain that the scaled waiting time of a customer of a given class in heavy traffic is distributed as the product of two exponentially distributed random variables, see Section 5.1 for more details.