Performance Evaluation 75–76 (2014) 36–49

Contents lists available at ScienceDirect

Performance Evaluation

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

Large deviations of an infinite-server system with a linearly scaled background process

K.E.E.S. De Turck[b,*], M.R.H. Mandjes[a,1]

[a] Korteweg–de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

[b] TELIN, Ghent University, St.-Pietersnieuwstraat 41, B9000 Gent, Belgium

ARTICLE INFO

Article history:
Received 16 April 2013
Received in revised form 30 September 2013
Accepted 13 January 2014
Available online 31 January 2014

Keywords:
Queues
Infinite-server systems
Markov modulation
Large deviations

ABSTRACT

This paper studies an infinite-server queue in a Markov environment, that is, an infinite-server queue with arrival rates and service times depending on the state of a Markovian background process. We focus on the probability that the number of jobs in the system attains an unusually high value. Scaling the arrival rates λ_i by a factor N and the transition rates ν_{ij} of the background process as well, a large-deviations based approach is used to examine such tail probabilities (where N tends to ∞). The paper also presents qualitative properties of the system's behavior conditional on the rare event under consideration happening.

1. Introduction

Queues with infinitely many servers have found widespread use in various application domains, often as an approximation for models with many servers. In these systems jobs arrive, are served in parallel, to leave when their service is completed. While rooted in communication networks, where the so-called Erlang model describes the dynamics of the number of calls in progress, applications in various other domains have been explored, such as road traffic [1] and biology [2,3].

In the standard infinite-server model, referred to as $M/G/\infty$, jobs arrive according to a Poisson process with rate λ, where their service times form a sequence of independent and identically distributed (i.i.d.) random variables (distributed as a random variable B with finite first moment), independent of the call arrival process. In such $M/G/\infty$ systems, a key result states that the stationary number of jobs in the system obeys a Poisson distribution with mean $\lambda\,\mathbb{E}B$ (irrespective of the precise distribution of the service times). This basic infinite-server system may be considered somewhat restrictive, though: in many practical situations the assumptions of a constant arrival rate and the jobs stemming from a single distribution are not realistic. A model that allows the input process to exhibit some sort of 'burstiness' is the so-called *Markov-modulated* infinite-server queue. In this model, a finite-state irreducible continuous-time Markov process (often referred to as the *background process*, or *modulating process*) modulates the input process: if the background process is in state i, the arrival process is a Poisson process with rate, say, λ_i, while the service times are distributed as a random variable, say, B_i (while the obvious independence conditions are imposed).

[*] Corresponding author. Tel.: +32 92643411.
E-mail addresses: kdeturck@telin.ugent.be (K.E.E.S. De Turck), M.R.H.Mandjes@uva.nl (M.R.H. Mandjes).

[1] M. Mandjes is also with Eurandom, Eindhoven University of Technology, Eindhoven, The Netherlands, and IBIS, Faculty of Economics and Business, University of Amsterdam, Amsterdam, The Netherlands.

http://dx.doi.org/10.1016/j.peva.2014.01.001
0166-5316/© 2014 Published by Elsevier B.V.

The Markov-modulated infinite-server queue has attracted some attention in recent years (but substantially less than the corresponding Markov-modulated single-server queue). The main focus in the literature so far has been on characterizing (through the derivation of moments, or even the full probability generating function) the steady-state number of jobs in the system. The most striking feature is that the number of jobs in the system still has a Poisson distribution, but now with a *random* parameter; a few key references are [4–7]. Interestingly, under an appropriate time-scaling [8,9] in which the transitions of the background process occur at a faster rate than the Poisson arrivals, we retrieve the Poisson distribution (with a *deterministic* parameter, that is) for the steady-state number of jobs in the system. Recently, transient results have been obtained as well, under specific scalings of the arrival rates and transition times of the modulating Markov chain [8,10].

Contribution. In this paper we focus on Markov-modulated infinite-server queues in a large-deviations setting. More precisely, we study the probability that the number of jobs present in the system at some time t attains some unusually high value. In the past in two short papers we have identified the corresponding tail asymptotics in two specific regimes: (i) one in which the transitions of the background process occur at a considerably slower rate than the job arrivals [11], and (ii) one in which the transitions of the background process occur at a considerably faster rate than the job arrivals [12]. In both cases the large deviations are those of a Poisson random variable; in the former case the (non-trivial) parameter value corresponds to the background process' 'worst-case behavior' (constructed so as to build up as many jobs as possible), whereas in the latter case the system essentially behaves as a standard $M/G/\infty$ queue with appropriately chosen arrival rate and service times (e.g., this arrival rate is a weighted sum of the λ_i, where the weights follow from the equilibrium distribution of the background process). These papers, however, do *not* cover the (technically challenging) case in which the timescale of the jumps of the background process and the timescale of the arrival process grow in a proportional manner, and it is a large deviations analysis of this linear regime that we present in this paper.

More formally, in our analysis we replace the arrival rates λ_i by $N\lambda_i$, whereas the transition rates of the background process ν_{ij} are replaced by $N\nu_{ij}$; the service time distributions are left unchanged. With $M^{(N)}(t)$ denoting the number of jobs in the system (starting empty) at time 0, the decay rate of $\mathbb{P}(M^{(N)}(t) \geq Na)$ is identified, for $a > \mathbb{E}M^{(N)}(t)/N$, in the regime that $N \to \infty$. As it turns out, this decay rate can be expressed in terms of the solution to a variational problem. In the paper we specialize to the case that the dimension d of the background process equals 2; it is indicated, though, how the analysis should be adapted for $d \in \{3, 4, \ldots\}$. It is noted that for the *non-modulated* system (i.e., the rates λ_i and μ_i independent of the state i) a large-deviations analysis was performed in [13, Chapter XII].

Organization. The organization of the rest of this paper is as follows. In Section 2, we provide a detailed model description and introduce some notation. Section 3 states and proves the main result of this paper, viz. an expression for the decay rate under study as the solution to a variational problem. Next, in Section 4, we discuss techniques for numerically solving this variational problem. Next, Section 5, contains some discussion of the results as well as a number of concluding remarks. Finally, numerical results are provided in Section 6.

2. Model description

As mentioned above, this paper studies an infinite-server queue with Markov-modulated Poisson arrivals and general service times. In full detail, the model can be described as follows.

Consider an irreducible continuous-time Markov process $(J(t))_{t\in\mathbb{R}}$ on a finite state space $\{1, \ldots, d\}$, with $d \in \mathbb{N}$. Its rate matrix is given by $\left(\nu_{ij}\right)_{i,j=1}^{d}$. Let π_i the stationary probability that the background process is in state i, for $i = 1, \ldots, d$. The time spent in state i (often referred to as the *transition time*) has an exponential distribution with mean $1/\nu_i$, where $\nu_i := -\nu_{ii}$.

While the process $(J(t))_{t\in\mathbb{R}}$, also called the *background process* or *modulating process*, is in state i, jobs arrive according to a Poisson process with rate $\lambda_i \geq 0$. The queueing model is an *infinite-server queue*: jobs are served in parallel—in other words: the sojourn time of a job equals its service time. The service times are assumed to be i.i.d. samples distributed as a random variable B_i if the job was generated when the background process was in state i. The usual independence assumptions apply. It is noted that we exclude the case that all λ_i as well as the distributions of the B_i coincide (as otherwise the queue is just an ordinary $M/G/\infty$).

In the sequel, we specialize to the case of a two-state background process ($d = 2$), and the random variable B_i corresponding to an exponential distribution with mean μ_i^{-1}. In the discussion section, we indicate how these assumptions can be relaxed.

3. Main result

In this paper, we consider the scaling $\nu_i \mapsto N\nu_i$, for $i = 1, 2$. We call the resulting background process $(J^{(N)}(s))_{s\in\mathbb{R}}$; in this scaling the background process jumps N times as fast. In addition, the arrival rates are scaled by N as well: $\lambda_i \mapsto N\lambda_i$. The objective of the section is to identify the tail asymptotics of the number of jobs present in our Markov-modulated infinite server at time t under this scaling. We let $M^{(N)}(t)$ denote the number of jobs in the system at time t, in the N-scaled model, where it is assumed that the system starts empty at time 0.