European Journal of Operational Research 233 (2014) 625–639

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

A multi-server queueing model with server consultations

Srinivas R. Chakravarthy *

Department of Industrial and Manufacturing Engineering, Kettering University, Flint, MI 48504, USA

ARTICLE INFO

Article history:
Received 22 March 2013
Accepted 3 October 2013
Available online 18 October 2013

Keywords:
Markovian arrival process
Preemptive priority
Interruptions
Consultations
Algorithmic probability

ABSTRACT

We consider a multi-server queueing model in which the arrivals occur according to a Markovian arrival process. One of the servers, henceforth referred to as the main server, offers consultation to fellow servers (referred to as regular servers) apart from serving the customers. A regular server may request a consultation only when serving a customer and is offered consultation on a first-come-first-served basis by the main server. The main server gives a preemptive priority to regular servers (for consulting) over customers. Thus, the main server can undergo interruptions during his/her servicing the customers. Under the assumptions of exponential services and consultations, the model is analyzed in steady-state using the well-known matrix-analytic methods. Illustrative numerical examples to bring out the qualitative nature of the model under study are presented.

1. Introduction and model description

The stochastic model considered in this paper was motivated by a personal experience of this author. Recently, I visited a local branch office of the Secretary of State of Michigan for some paperwork. I was one of fifteen or so customers to enter into the office (after waiting for it to open) at 9am. After waiting for a few minutes, I was attended by a personnel. During the time I was in "service", the personnel was interrupted by her fellow colleagues for items like (a) clarification on some paperwork, (b) needing a change (for a twenty/fifty/hundred dollar bill), and (c) computer related questions. Only this personnel was interrupted and hence my service. None of the customers attended by other personnel were interrupted in the sense that their service requirements were the ones that made their respective servers to consult my server. One can think of this consulting activity as part of the services. After about four such interruptions my service was completed. Queues with interruptions have been studied extensively in the literature and we refer the reader to the recently published review article on this topic (Krishnamoorthy, Pramod, & Chakravarthy, 2012). The interruption process modeled in the literature can be broadly grouped into three categories: (a) here an independent process (mostly Poisson process) generates the interruptions and affects the busy servers and or the system; (b) here two or more priority customers arrive to the system and the services are of preemptive priority nature leading to service interruptions; and (c) in this case an independent process (again mostly Poisson process) generates the interruptions that affect the customer who is in the service facility (consisting of a single server) (Jacob, Chakravarthy, & Krishnamoorthy, 2012). Thus, the interruptions occur independent of the number of busy servers/customers in service. Thus, to our knowledge, the interruptions generated by the servers have not been addressed in the literature. Further, to this author's knowledge this model cannot be made as a special case of any of the published models with interruptions. The model studied here can be applied in many other areas in real-life applications similar to the one outlined here. For example, in call centers it is common to see a supervisor or manager being approached by call center attendants to seek help in matters such as billing adjustments, and addressing grievances. Supervisors may also choose to take the customers' calls. In fast food restaurants, banks, and grocery stores, the manager may be approached by the tellers for assistance.

In this paper, we assume that the customers arrive according to a Markovian arrival process (*MAP*) with parameter matrices D_0 and D_1 of order m. The *MAP* is a rich class of point processes that includes many well-known processes such as Poisson, phase type (PH) renewal processes, and Markov-modulated Poisson process. The matrix D_0 governs the transitions corresponding to no arrival and D_1 governs those corresponding to an arrival. By assuming D_0 to be a nonsingular matrix, the interarrival times will be finite with probability one and the arrival process does not terminate. Hence, we see that D_0 is a stable matrix. The generator Q^* is then given by $Q^* = D_0 + D_1$. It can be shown that *MAP* is equivalent to Neuts' versatile Markovian point process. One of the most significant features of the *MAP* is the underlying Markovian structure and fits ideally in the context of matrix-analytic solutions to stochastic models. Matrix-analytic methods were first introduced and studied by Neuts (1981). The point process described by the *MAP* is a special class of semi-Markov processes with kernel given by

* Tel.: +1 8107627906; fax: +1 8107629924.
E-mail address: schakrav@kettering.edu

0377-2217/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2013.10.008

$$\int_0^x e^{D_0 t} dt D_1 = [I - e^{D_0 x}](-D_0)^{-1} D_1, x \geqslant 0.$$

For more details on *MAP* processes and their usefulness in stochastic modelling, we refer to (Lucantoni, 1991; Neuts, 1989; Neuts, 1992), and for a review and recent work on *MAP* we refer the reader to (Artalejo, Gomez-Corral, & He, 2010; Chakravarthy, 2001, 2010).

The system has c servers offering services to the customers. Of these c servers there is one server, henceforth referred to as the *main* server, who offers consulting work to fellow servers, henceforth referred to as *regular* servers, in addition to providing services to the customers. An arriving customer finding at least one idle server will get into service immediately; otherwise joins the queue of infinite capacity. We assume that, whenever the main server is idle along with at least one regular server, the arriving customer will always be served by the main server. It is easy to modify this assumption to include various other options. Note that this modification will affect only the boundary states of the Markov chain describing the system under study.

The service times of the customers attended by any of the regular servers are exponentially distributed with parameter μ_1 and those of the main server are exponential with parameter μ_2. Whenever a service is initiated by a regular server, a clock of exponential duration with parameter θ is started. Whenever this clock, which is independent of the service time, expires before the service is completed, a consultation is needed for the server. Otherwise, the service will be completed without any consultation and the clock will have no bearing and will be restarted at the epoch of the next service. Note that if there are i, $1 \leqslant i \leqslant c-1$, regular servers busy, then the rate at which a consulting work will be required is $i\theta$. The server requiring consulting work will be attended by the main server as follows. If, at the time of such a request, the main server is idle or busy serving a customer the consulting request will be attended immediately. In the latter case, the main server (and hence that customer's service) is said to be interrupted. However, if the main server is busy offering consulting work to other regular server(s), the server needing the consulting work will be queued up. Thus, a regular server may request a consultation only when serving a customer and is offered consultation on a first-come-first-served basis by the main server. Note that at any given time there can be a maximum of $c-1$ servers needing consulting work. The interrupted customer's service will be resumed by the main server only after all consulting work is finished. We assume that the main server's consulting time of a regular server is exponentially distributed with parameter ξ.

In this paper, the services of the customers whose servers need consulting work during their services are not considered as interrupted ones. We assume that such consultations are part of their services. Thus, only the main server's customers face possible service interruptions before leaving the system.

The inter-arrival times of the customers, the service times of the customers, consulting clocks and the consulting times are all assumed to be independent of each other.

For use in sequel, let $\boldsymbol{e}(r)$, $\boldsymbol{e}_j(r)$ and I_r denote, respectively, the (column) vector of dimension r consisting of 1's, column vector of dimension r with 1 in the jth position and 0 elsewhere, and an identity matrix of dimension r. When there is no need to emphasize the dimension of these vectors we will suppress the suffix. Thus, $\boldsymbol{e}$ will denote a column vector of 1's of appropriate dimension. The notation "′" will stand for the transpose of a matrix and the symbol $\otimes$ denotes the Kronecker product of matrices. Thus, if A is a matrix of order $m \times n$ and if B is a matrix of order $p \times q$, then $A \otimes B$ will denote a matrix of order $mp \times nq$ whose (i,j)th block matrix is given by $a_{ij}B$. For more details on Kronecker products, we refer the reader to Marcus and Minc (1964). Finally, $\Delta(\boldsymbol{a})$ denotes the diagonal matrix whose diagonal elements are given by the components of the vector $\boldsymbol{a}$.

Let ζ denote the stationary probability vector of the Markov processes with generator Q^*. That is, ζ is the unique (positive) probability vectors satisfying

$$\zeta Q^* = \boldsymbol{0}, \quad \zeta \boldsymbol{e} = 1. \tag{1}$$

Thus, in this paper we consider a multi-server queueing model in which the main server not only offers services to the customers but also provides consulting work to regular servers. The service times, the request (for consulting) times, and the consulting times are exponentially distributed. The rest of the paper is organized as follows. In Section 2 the steady-state analysis of the system is presented and some illustrative numerical results are given in Section 3. Concluding remarks including some future work are given in Section 4.

2. The steady-state analysis

In this section we will analyze the model in steady state. Let $N_1(t)$, $N_2(t)$, $J_1(t)$, and $J_2(t)$ denote, respectively, the number of customers in the system, the number of servers in consulting mode, the status of the main server (idle – when not serving or consulting; busy consulting with no interruption; busy consulting with interruption; busy serving a customer), and the phase of the *MAP* arrival process at time t. Even though there are four possible states for the status of the main server we can limit the number of combinations by properly grouping the states as shown below. The process $\{(N_1(t),N_2(t),J_1(t),J_2(t)): t \geqslant 0\}$ is a continuous-time Markov chain with state space given by

$$\begin{aligned}\Omega = {} & \{(*,k) : 1 \leqslant k \leqslant m\} \\ & \bigcup \{(i,j,0,k) : 1 \leqslant i \leqslant c-1, 0 \leqslant j \leqslant i, 1 \leqslant k \leqslant m\} \\ & \bigcup \{(i,j,1,k) : 2 \leqslant i \leqslant c-1, 1 \leqslant j \leqslant i-1, 1 \leqslant k \leqslant m\} \\ & \bigcup \{(i,k) : 1 \leqslant i \leqslant c-1, 1 \leqslant k \leqslant m\} \\ & \bigcup \{(i,j,0,k) : i \geqslant c, 0 \leqslant j \leqslant c-1, 1 \leqslant k \leqslant m\} \\ & \bigcup \{(i,j,1,k) : i \geqslant c, 1 \leqslant j \leqslant c-1, 1 \leqslant k \leqslant m\}.\end{aligned}$$

A brief description of the state space along with defining the levels of the states are given below to better understand the generator of the Markov process.

- The set, $\underline{*}$, of states $\{(*,k): 1 \leqslant k \leqslant m\}$ denotes that the system is in idle state and the phase of the *MAP* is in k.
- The level $\underline{\boldsymbol{i}}^*$, $1 \leqslant i^* \leqslant c-1$, consisting of the states $\{(i,k): 1 \leqslant k \leqslant m\}$ corresponds to the case where the main server along with $i-1$ regular servers are busy serving customers, and the phase of the *MAP* is in k.
- The level $\underline{\tilde{\boldsymbol{i}}}$, $1 \leqslant \tilde{i} \leqslant c-1$, consisting of the states $\{(i,j,0,k): 1 \leqslant i \leqslant c-1, 0 \leqslant j \leqslant i, 1 \leqslant k \leqslant m\}$ corresponds to the case where there are i customers in the system with j servers in consulting mode (the other $i-j$ servers are busy serving) and the phase of the *MAP* is in k. Note that the main server can either be idle (when $j = 0$) or offering consulting work (when $j > 0$).
- The level $\underline{\hat{\boldsymbol{i}}}$, $2 \leqslant \hat{i} \leqslant c-1$, consisting of the states $\{(i,j,1,k): 2 \leqslant i \leqslant c-1, 1 \leqslant j \leqslant i-1, 1 \leqslant k \leqslant m\}$ corresponds to the case in which there are i customers in the system with j servers in consulting mode (and $i-1-j$ servers busy serving customers and the main server has been interrupted), and the phase of the *MAP* is in k.
- The set of the states $\{(i,j,0,k): i \geqslant c, 0 \leqslant j \leqslant c-1, 1 \leqslant k \leqslant m\}$ corresponds to the case where the system has i customers with the main server either busy serving the customers along with other $c-1$ servers or the main server is busy giving consulting to one