Available online at www.sciencedirect.com

ScienceDirect

Computers & Operations Research 34 (2007) 983–996

computers & operations research

www.elsevier.com/locate/cor

Applications of maximum queue lengths to call center management

J.R. Artalejo[a,*], A. Economou[b], A. Gómez-Corral[a]

[a]Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Madrid 28040, Spain
[b]Department of Mathematics, University of Athens, Panepistemiopolis, Athens 15784, Greece

Available online 29 June 2005

Abstract

This paper deals with the distribution of the maximum queue length in two-dimensional Markov models. In this framework, two typical assumptions are: (1) the stationary regime, and (2) the system homogeneity (i.e., homogeneity of the underlying infinitesimal generator). In the absence of these assumptions, the computation of the stationary queue length distribution becomes extremely intricate or, even, intractable. The use of maximum queue lengths provides an alternative queueing measure overcoming these problems. We apply our results to some problems arising from call center management.

Keywords: Call center; Maximum queue length; Level dependent quasi-birth-and-death processes; Customer behavior; Routing rules

1. Introduction

In recent years, there has been a rapidly growing interest on call centers making emphasis on design and management problems. A comprehensive review of the existing literature can be found in the survey papers by Gans et al. [1], and Koole and Mandelbaum [2].

* Corresponding author. Fax: +34 913944606.
E-mail addresses: jesus_artalejo@mat.ucm.es (J.R. Artalejo), aeconom@math.uoa.gr (A. Economou), antonio_gomez@mat.ucm.es (A. Gómez-Corral).

0305-0548/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cor.2005.05.020

A large number of studies in the call center literature deal with the problem of finding optimal measures of interest. The optimal staffing of call centers addresses the key problem of dimensioning parameters (number of agents, trunks) in order to guarantee maximal profit and a desired grade-of-service (GoS) measured in terms of acceptable waiting and blocking, see Brandt et al. [3] and Srinivasan et al. [4]. In addition, many modern call centers assume the existence of several types of customers which share common resources under certain conditions (allocation rules, routing to shared services, existence of specialized and flexible servers). In this general framework, the quality of service (QoS) is usually measured in terms of classical queueing performance descriptors: expected queue length, waiting time, throughput and blocking probabilities.

The computation of these queueing characteristics typically requires the stochastic model to be in steady state regime. For instance, many call centers are modelled using birth-and-death processes or their matrix generalizations, i.e., quasi-birth-and-death (QBD) processes. In this context, one needs to study the positive recurrence of the underlying Markov chain. Assuming a stationary regime is needed to guarantee the existence, and subsequent computation, of most classical queueing measures. The possibility of overcoming the burden of dealing with a system operating in stationary regime provides an initial motivation to our study.

Our main goal in this paper is to present the distribution of the maximum queue length in a busy period as a performance descriptor of practical relevance in call center management. This distribution can be even computed in non-stationary regime. Replacing the stationary queue length distribution by the maximum queue length distribution provides extra advantages when the underlying queueing system is space-heterogeneous. This is the case, for instance, for call centers allowing retrials, see Aguir et al. [5] and Whitt [6]. These models operate under the presence of a non-homogeneous flow of repeated attempts which is superimposed on the stream of primary calls. We also refer to Falin and Templeton [7] and Artalejo [8] for accounts of the literature on retrial queues. Besides these advantages, the maximum queue length has an intrinsic interest as a measure of system congestion. Small values of the maximum queue length show that the call center is operating without significant fluctuations. In contrast, a large maximum queue length gives support to the adoption of drastic decisions such as an increase of the number of agents or rescheduling of common resources.

Extreme values of the queue length can be investigated following different approaches. Serfozo [9] proposes an asymptotic analysis whereas Neuts [10] concentrates on the distribution of the maximum queue length during a busy period. We will adopt the latter approach to develop an algorithmic analysis of the maximum level visited in a non-homogeneous QBD process. In fact, the existing literature provides recent examples where call centers are modelled as QBD processes, see for instance Aguir et al. [5], Masi et al. [11], Shumsky [12] and references therein. Focusing on QBD processes is not restrictive, since our goals can also be extended to other stochastic models including Markov chains of $M/G/1$ and $GI/M/1$-type.

As related work, we mention Section 8.1 of Latouche and Ramaswami [13]. They established the relationship between the computation of the matrix $\mathbf{G}$ (which records the probabilities, starting from an initial level, of visiting the previous one in a finite time) and the maximum queue length in a homogeneous QBD process. Our analysis in this paper covers the study of the maximum queue length in the more general class of level dependent QBD processes. We also refer to a recent paper by Artalejo et al. [14] where the specific form both of the generator and the blocks of the $M/M/c$ retrial queue is exploited to derive an efficient algorithm for computing the distribution of interest.