Applied Mathematics and Computation 235 (2014) 124–136

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Maximum queue lengths during a fixed time interval in the $M/M/c$ retrial queue ☆

A. Gómez-Corral [a,*], M. López García [b]

[a] Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, 28040 Madrid, Spain
[b] Department of Applied Mathematics, School of Mathematics, University of Leeds, Leeds LS2 9JT, United Kingdom

ARTICLE INFO

Keywords:
Absorbing Markov chain
Eigenvalues/eigenvectors
Maximum queue length
Retrial queue
Splitting method

ABSTRACT

We are concerned with the problem of characterizing the distribution of the maximum number $Z(t_0)$ of customers during a fixed time interval $[0, t_0]$ in the $M/M/c$ retrial queue, which is shown to have a matrix exponential form. We present a simple condition on the service and retrial rates for the matrix exponential solution to be explicit or algorithmically tractable. Our methodology is based on splitting methods and the use of eigenvalues and eigenvectors. A particularly appealing feature of our solution is that it allows us to obtain global error control. Specifically, we derive an approximating solution $p(x;t_0) \equiv p(x;t_0;\varepsilon)$ verifying $|P(Z(t_0) \leqslant x|X(0) = (i,j)) - p(x;t_0)| < \varepsilon$ uniformly in $x \geqslant i+j$, for any $\varepsilon > 0$ and initial numbers i of busy servers and j of customers in orbit.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

In this paper we consider maximum queue lengths in the $M/M/c$ retrial queue. The $M/M/c$ retrial queue is the main multiserver retrial queue (see [11, Chapter 2]) in which primary customers arrive according to a Poisson process of rate λ, the service facility consists of c identical servers, and service times are exponentially distributed with parameter ν. If a primary customer finds some server free, he instantly occupies one server and leaves the system after service. Any customer who finds all servers busy upon arrival is obliged to leave the service area, but he repeats his demand after an exponential time with parameter μ; i.e., inter-retrial times of each customer are assumed to be independent and exponentially distributed with intensity μ. We assume that inter-arrival times, service times and inter-retrial times are mutually independent.

The system state at time t can be described by means of a bivariate process $\mathcal{X} = \{X(t) = (C(t), N(t)) : t \geqslant 0\}$, where $C(t)$ is the number of busy servers and $N(t)$ is the number of customers in orbit, that is, customers repeating their demand. Under the above distributional assumptions, the process $\mathcal{X}$ is a regular continuous-time Markov chain (CTMC) with the lattice semistrip $\mathcal{S} = \{0, 1, \ldots, c\} \times \mathbb{N}_0$ as the state space. Its non-null infinitesimal transition rates $q_{(i,j),(i',j')}$ are specified as follows:

☆ Financial support for this work was provided by the Government of Spain (Ministry of Economy and Competitiveness) and the European Commission through the project MTM-2011-23864, and the Grant BES-2009-018747.

* Corresponding author.
E-mail addresses: antonio_gomez@mat.ucm.es (A. Gómez-Corral), M.LopezGarcia@leeds.ac.uk (M. López García).
URLs: http://www.mat.ucm.es/~mcqt/qmg/gomezc.html (A. Gómez-Corral), https://www1.maths.leeds.ac.uk/~lopezgarcia/index.html (M. López García).

http://dx.doi.org/10.1016/j.amc.2014.02.074
0096-3003/© 2014 Elsevier Inc. All rights reserved.

(a) For $0 \leqslant i \leqslant c-1$,

$$q_{(i,j),(i',j')} = \begin{cases} \lambda, & \text{if } (i',j') = (i+1,j), \\ i\nu, & \text{if } (i',j') = (i-1,j), \\ j\mu, & \text{if } (i',j') = (i+1,j-1), \end{cases} \tag{1}$$

and $q_{(i,j)} = -q_{(i,j),(i,j)} = \lambda + i\nu + j\mu$.

(b) For $i = c$,

$$q_{(c,j),(i',j')} = \begin{cases} \lambda, & \text{if } (i',j') = (c,j+1), \\ c\nu, & \text{if } (i',j') = (c-1,j), \end{cases} \tag{2}$$

and $q_{(c,j)} = -q_{(c,j),(c,j)} = \lambda + c\nu$.

If we express $\mathcal{S}$ in terms of *levels* $\mathcal{S} = \cup_{j=0}^{\infty} l(j)$ with $l(j) = \{(i,j-i) : 0 \leqslant i \leqslant \min\{j,c\}\}$ for $j \geqslant 0$, then the infinitesimal generator $\mathbf{Q} = (q_{(i,j),(i',j')})$ of the process $\mathcal{X}$ has the structured form

$$\mathbf{Q} = \begin{pmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,1} & & & \\ \mathbf{A}_{1,0} & \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & & \\ & \mathbf{A}_{2,1} & \mathbf{A}_{2,2} & \mathbf{A}_{2,3} & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{3}$$

where $\mathbf{A}_{j,j'}$ contains transition rates related to jumps of $\mathcal{X}$ from states of $l(j)$ to states of the level $l(j')$, for $j' \in \{j-1,j,j+1\}$, and diagonal elements of $\mathbf{A}_{j,j}$ are given by $-q_{(i,j-i)}$, for $0 \leqslant i \leqslant \min\{j,c\}$. Specifications for $\mathbf{A}_{j,j'}$ are readily derived from (1) and (2); see Appendix A.

The main analytical difficulties and the most interesting properties of the $M/M/c$ retrial queue are connected with the level dependence exhibited by the infinitesimal generator $\mathbf{Q}$ in (3). To show the nature of these difficulties in more detail, we may consider the simplest problem, that is, the calculation of the stationary distribution $\{P_{i,j} : (i,j) \in \mathcal{S}\}$ of the process $\mathcal{X}$ under the assumption that the traffic load $\rho = (c\nu)^{-1}\lambda$ is less than one. It is worth mentioning that, for $c \leqslant 2$, the partial sequences $\{P_{i,j} : j \in \mathbb{N}_0\}$ with $0 \leqslant i \leqslant c$ satisfy sets of equations of *birth-and-death* type and, consequently, explicit expressions for the stationary probabilities $P_{i,j}$ are recursively derived. The consideration of more than two servers complicates the transitions among states, which implies that the underlying structure of birth-and-death type is not preserved. The particular case $c = 3$ is treated in [13], where the problem is reduced to finding the probabilities $P_{0,0}$ and $P_{0,1}$, which can be recursively computed in terms of a limit condition. The papers by Phung-Duc et al. [23,24] overcome this technical condition and investigate in more detail the stationary probabilities $P_{i,j}$. Based on the Kolmogorov equations some theoretical approaches provide solutions in terms of contour integrals [8] or as limit of extended continued fractions [22]; see [26] for a matrix application of the continued fraction approach to multiserver retrial queues. However, from a practical point of view, the stationary probabilities $P_{i,j}$ cannot be expressed in a tractable form and do not lead to a direct recursive computation when $c > 3$. This drawback motivates the implementation of numerically tractable approximations, such as approximations based on truncated models [11,28,31] and generalized truncated models [2,10,21], the RTA (*retrials see time averages*) approximation due to Greenberg and Wolff [15,32], the Fredericks and Reisner approximation [12], and approximations by interpolation [5,27]. The book by Artalejo and Gómez-Corral [5] presents a comparative review of these approximations for the $M/M/c$ retrial queue, as well as of other performance measures. In the recent monograph by Dayar [9] and the paper by Bright and Taylor [7] the reader can find details on computational issues for an appropriate numerical evaluation of stationary distributions in the more general setting of Markov chains and level-dependent QBD processes. A recent development of algorithmic tools in level-dependent QBD processes can be found in [25].

In this paper, we aim to study the maximum queue length $Z(t_0)$ during a fixed time interval $[0, t_0]$ in the $M/M/c$ retrial queue defined by (1) and (2). One way of analyzing the maximum size distribution is to record maximum values during a busy period, instead of a predetermined time interval. Let us recall that, in the $M/M/c$ retrial queue, a busy period is defined as the period $[0, T]$ that, starting with the arrival of a customer who finds the system (i.e., servers and orbit) empty, ends at the first service completion epoch at which the system becomes empty again. The distribution of the maximum queue length during a busy period in the $M/M/c$ retrial queue can be numerically evaluated from Algorithm 1 of [4]. More concretely, Artalejo et al. [4] reduce the problem to computation of certain absorption probabilities that constitute the unique solution of a block-tridiagonal linear system obtained by employing first principles based on first-step analysis; as a related work, see [3] where the focus is on level-dependent QBD processes. The maximum queue length appears to be a performance descriptor of practical relevance in retrial queues since it allows us to deal with queueing systems that do not necessarily operate in stationary regime; see e.g. [3, Sections 3 and 4]. Artalejo et al. [3] apply the *busy-period* version $Z(T)$ of the maximum queue length to two problems arising from call center management. To be concrete, they consider the retrial queueing system proposed by Aguir et al. [1] to model a call center operating under the simultaneously presence of customer balking and retrials due to impatience, as well as the external-rule system investigated by Masi et al. [19] to formulate a routing rule for a resource-sharing call center.

Download English Version:

https://daneshyari.com/en/article/4627820

Download Persian Version:

https://daneshyari.com/article/4627820

Daneshyari.com