5-30-2020

A Retrial Queueing Model With Thresholds and Phase Type Retrial Times

Srinivas R. Chakravarthy
Kettering University, schakrav@kettering.edu

J. Appl. Math. & Informatics Vol. **38***(2020), No. 3 - 4, pp. 351 - 373*
https://doi.org/10.14317/jami.2020.351

A RETRIAL QUEUEING MODEL WITH THRESHOLDS AND PHASE TYPE RETRIAL TIMES

SRINIVAS R. CHAKRAVARTHY

Abstract. There is an extensive literature on retrial queueing models. While a majority of the literature on retrial queueing models focuses on the retrial times to be exponentially distributed (so as to keep the state space to be of a reasonable size), a few papers deal with nonexponential retrial times but with some additional restrictions such as constant retrial rate, only the customer at the head of the retrial queue will attempt to capture a free server, 2-state phase type distribution, and finite retrial orbit. Generally, the retrial queueing models are analyzed as level-dependent queues and hence one has to use some type of a truncation method in performing the analysis of the model. In this paper we study a retrial queueing model with threshold-type policy for orbiting customers in the context of nonexponential retrial times. Using matrix-analytic methods we analyze the model and compare with the classical retrial queueing model through a few illustrative numerical examples. We also compare numerically our threshold retrial queueing model with a previously published retrial queueing model that uses a truncation method.

AMS Mathematics Subject Classification : 60K20, 60J28.
Key words and phrases : Retrial, queueing, phase type distribution, thresholds.

1. Introduction

Retrial queueing models are queueing models in which an arriving customer not able to get into service enters into a retrial orbit and try to capture a free server by competing with other customers, if any, present in the orbit. These models have been identified to be very useful in many applications notably in telecommunications and service industries. Retrial queues have been studied extensively in the literature (see e.g., [1, 3, 6, 8, 21, 26]). Much of the literature deals with exponential retrials with the exception of a few studies(see e.g., [5, 15, 18, 23, 36, 37, 38, 40]). As pointed out in [15], the few

Received August 1, 2019. Revised December 21, 2019. Accepted February 11, 2020.

papers dealing with nonexponential retrials, such as phase type distributions ($PH-$distribution) for retrials, propose a variety of approximations in their studies. Such an approach to the study of retrial queueing models with nonexponential retrials is not by choice but rather due to the inherent complexity created by the exponential growth of the state space. Realizing the lack of results in multi-server retrial queueing models with nonexponential retrial times, Chakravarthy [15] proposed a different approach via simulation and reported some interesting results. Specifically, it was shown in [15] that assuming exponential retrial times in place of nonexponential ones could lead to under or over estimating the system performance measures.

In classical queues, a threshold approach is employed to manage any possible congestions in the system. For example, in a classical multi-server queueing system, one way to minimize the congestion (in terms of average number of customers present) in the system is to have threshold-type policies to determine the allocation of the jobs/customers among the heterogeneous servers (see, e.g., [19, 28, 29, 32, 35]). Further, in classical multi-server queues (with or without homogeneity assumption on the servers), threshold-type policies have been identified to minimize the number of servers (see, e.g., [11, 13, 14, 24]) in the system.

Efrosinin and Breuer [20] established that a threshold-type policy is also optimal for retrial queues. It should be pointed out the threshold they employ is only at the arrival points of new or retrial customers accessing a free server from a finite retrial buffer. That is, the threshold becomes active only when a new customer arrives (at which time a decision is made either to route to an idle server or to send the customer to the retrial orbit provided there is a space) or a retrial customer arrives (at which time a decision is made either to route to an idle server or send the customer back to the retrial orbit). However, to the best of our knowledge there is no literature that employs a threshold-type policy for the waiting customers in the retrial orbit. Thus, in this paper, we take a different approach to the study of the retrial queues by introducing the concepts of threshold based retrial times.

The main motivation for the study of a threshold-type retrial queueing model in this paper arose out of a need for including nonexponential retrial times and at the same time not to significantly increase the complexity of the retrial model. For example, Shin [37] studied an $M/M/c$-type retrial queueing model with phase type (restricted to only two phases due to the size of the underlying state space) retrial times and apply the level-dependent $QBD-$process approach to study the model. Relaxing the assumption of a $2-$state $PH-$distribution for the retrials, Shin and Moon [38] proposed an approximation for the distribution of the number of busy servers as well as the mean number of customers in retrial orbit. These two papers, for example, further illustrate the complexity involved in relaxing the exponential assumption for the retrial times. Our approach using threshold-type retrial times has an advantage in that one can use this model as

another approximation (through appropriate choice of the threshold parameters) to retrial queues with phase type retrial times.

For use in the sequel, we set up the following notation. By (a) $\boldsymbol{e}$, a column vector (of appropriate dimension) of 1's; (b) $\boldsymbol{e}_i$, a unit column vector (of appropriate dimension) with 1 in the i^{th} position and 0 elsewhere; (c) I is an identity matrix (of appropriate dimension); (d) $\Delta(T_1, \cdots, T_r)$, a diagonal matrix with diagonal entries given by $T_i, i = 1, \cdots, r$. Note that these entries can be scalars, vectors, or matrices, and will be clear from the context. The dimension of the vectors and matrices should be clear in the context of usage. We will use $\boldsymbol{e}(mn)$ to show that the column vector is of dimension mn when more clarity is needed. We also need to use Kronecker product (denoted by $\otimes$) and Kronecker sum (denoted by $\oplus$) of matrices. For details on these, we refer to [22, 30, 39].

The paper is organized as follows. In Section 2, we describe the threshold retrial queueing model under study in this paper. The QBD-process needed to study this model and its steady-state analysis are presented in Section 3. The comparison of our threshold model to the one discussed in [38] is carried out in Section 4. Illustrative numerical examples are presented in Section 5 and some concluding remarks including future research work are presented in Section 6.

2. Model description

We consider a multi-server retrial queueing model in which the customers arrive according to a Poisson process with rate λ. An arriving customer, finding all c servers busy, will enter into a retrial buffer of infinite capacity. The service times are assumed to be exponentially distributed with parameter μ. The customers who are in retrial orbit will attempt to capture a free server at random times. In classical retrial queueing model, it is generally assumed that the retrial rates are proportional to the number of customers waiting in the retrial orbit and that the random times to attempt to capture a free server are exponentially distributed. There are variations such as constant retrial policy (see, e.g., [2, 9, 10, 16]), finite retrial orbit (see, e.g., [4, 7]), and attempts to capture a free server are made only by the customer at the head of the orbit (see, e.g., [36]). Such variations lead to level-independent $QBD-$process to analyze the model. But these restrictions may not be valid or suitable in practice.

We introduce two threshold parameters, say, $N, 1 \leq N < \infty$, and $K, 1 \leq K < \infty$, such that when the number of retrial customers is between $(k-1)N+1$ to kN, the retrial rate will be $\theta_k, 1 \leq k \leq K$, and once the number in retrial exceeds KN, the retrial rate is taken to be θ.

In this paper we will assume that the underlying random variable governing the retrial times is of phase type with rates dictated by the parameters, N and K. Suppose that the representation $(\boldsymbol{\beta}, S)$ of order n with $\boldsymbol{\beta}(-S)^{-1}\boldsymbol{e} = 1$. Then the retrial times are modeled using this $PH-$distribution and the threshold parameters. That is, when the number of retrial customers is in the interval $[(k-1)N+1, kN], 1 \leq k \leq K$, the retrial times are of phase type with

representation $(\boldsymbol{\beta}, \theta_k S)$ of order n. Once the number in the retrial buffer exceeds KN, the retrial times will follow a $PH-$distribution with representation $(\boldsymbol{\beta}, \theta S)$ of order n. Note that by taking $N = 1$, we get the classical retrial queueing model in which the retrial times are of phase type with all customers' $PH-$distributions having a common set of states but possibly with different rates. The value of K may be chosen similar to how a level dependent $QBD-$process is truncated for numerical purposes in the literature (see, e.g., [12]). As can be seen, in this paper we assume that we have the same number of states for the $PH-$distribution for all customers. However, we will elaborate more on how this can be relaxed and the impact of this relaxation later on.

Suppose that we define $J_1(t)$ to be the number of customers in the retrial orbit; $J_2(t)$ to be the number of busy servers; and $J_3(t)$ to be the phase of the retrial times (when $J_1(t) > 0$) at time t. It is easy to verify that $\{(J_1(t), J_2(t), J_3(t)) : t \geq 0\}$ is a quasi-birth-and-death process (QBD) with state space given by

$$\Omega = \{(0, j) : 0 \leq j \leq c\} \bigcup \{(i, j, k) : i \geq 1, 0 \leq j \leq c,\ 1 \leq k \leq n\}.$$

Let $\mathbf{0} = \{(0, j),\ 0 \leq j \leq c\}$ denote the set of states corresponding to the system in which the retrial buffer is empty and that j servers are busy. Note that in this case there is no need to keep track of the phase of underlying $PH-$distribution; let $\boldsymbol{i} = \{(i, j, k),\ 0 \leq j \leq c,\ 1 \leq k \leq n\}$, denote the set of states wherein exactly i customers are in the retrial orbit with j servers being busy, and the underlying $PH-$distribution is in phase k. For the sake of not complicating the notation, we group the set of levels 1 through KN into K groups of N levels when displaying the generator Q below. Thus, the matrices $Q_i i, 1 \leq i \leq K$ are of dimension $N(c+1)n$, whereas the matrices A_0, A_1, and A_2 are of dimension $(c+1)m$.

The generator, Q, of the QBD process under consideration is of the form

$$Q = \begin{bmatrix}
\tilde{B}_{0,1} & \hat{B}_{0,0} & & & & & & & & & \\
\hat{B}_{1,2} & Q_{1,1} & \tilde{A}_0 & & & & & & & & \\
 & Q_{2,1} & Q_{2,2} & \tilde{A}_0 & & & & & & & \\
 & & Q_{3,2} & Q_{3,3} & \tilde{A}_0 & & & & & & \\
 & & \ddots & \ddots & \ddots & & & & & & \\
 & & & & Q_{K,K-1} & Q_{K,K} & \hat{A}_0 & & & & \\
 & & & & & \hat{A}_2 & A_1 & A_0 & & & \\
 & & & & & & A_2 & A_1 & A_0 & & \\
 & & & & & & & A_2 & A_1 & A_0 & \\
 & & & & & & & & \ddots & \ddots & \ddots
\end{bmatrix}, \quad (1)$$

where the (block) matrices appearing in Q are as follows.

$$\tilde{B}_{0,1} = \begin{bmatrix} -\lambda & \lambda & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & \\ & 2\mu & -(\lambda+2\mu) & \lambda & & \\ & \ddots & \ddots & \ddots & & \\ & & (c-1)\mu & -(\lambda+(c-1)\mu) & \lambda & \\ & & & c\mu & -(\lambda+c\mu) \end{bmatrix}, \quad (2)$$

$$\hat{B}_{0,0} = \boldsymbol{e}_1'(N) \otimes \tilde{B}_{0,0},\ \hat{B}_{1,2} = \boldsymbol{e}_1(N) \otimes \tilde{B}_{1,2}, \quad (3)$$

$$Q_{r,r} = \begin{bmatrix} B_{r,1} & A_0 & & & \\ B_{r,2} & B_{r,1} & A_0 & & \\ & B_{r,2} & B_{r,1} & A_0 & \\ & & \ddots & \ddots & \\ & & B_{r,2} & B_{r,1} & A_0 \\ & & & B_{r,2} & B_{r,1} \end{bmatrix},\ 1 \le r \le K, \quad (4)$$

$$Q_{r,r-1} = \boldsymbol{e}_1(N) \otimes \boldsymbol{e}_N'(N) \otimes B_{r,2},\ 2 \le r \le K,$$

$$\tilde{A}_0 = \boldsymbol{e}_N(N) \otimes \boldsymbol{e}_1'(N) \otimes A_0, \quad (5)$$

$$\hat{A}_2 = \boldsymbol{e}_N'(N) \otimes A_2,\ \hat{A}_0 = \boldsymbol{e}_N(N) \otimes A_0,$$

$$\tilde{B}_{0,0} = \lambda \boldsymbol{e}_{c+1}(c+1) \otimes \boldsymbol{e}_{c+1}'(c+1) \otimes \boldsymbol{\beta}, \quad (6)$$

$$\tilde{B}_{1,2} = \theta_1 \begin{bmatrix} \mathbf{0} & \boldsymbol{S^0} & & & \\ & & \boldsymbol{S^0} & & \\ & \ddots & \ddots & \ddots & \\ & & & \boldsymbol{S^0} & \\ & & & & \mathbf{0} \end{bmatrix}, \quad (7)$$

$$A_0 = \lambda \boldsymbol{e}_{c+1}(c+1) \otimes \boldsymbol{e}_{c+1}'(c+1) \otimes I,$$

$$B_{r,2} = \theta_r \begin{bmatrix} 0 & \boldsymbol{S^0\beta} & & & \\ & & \boldsymbol{S^0\beta} & & \\ & \ddots & \ddots & \ddots & \\ & & & \boldsymbol{S^0\beta} & \\ & & & & 0 \end{bmatrix}, \quad (8)$$

$$B_{r,1} = \Delta(\theta_r S, \cdots, \theta_r S, \theta_r(S + \boldsymbol{S^0\beta})) + \tilde{B}_{0,1} \otimes I,$$
$$2 \le r \le K,$$

$$A_2 = \theta \begin{bmatrix} 0 & \boldsymbol{S^0\beta} & & & \\ & & \boldsymbol{S^0\beta} & & \\ & \ddots & \ddots & \ddots & \\ & & & \boldsymbol{S^0\beta} & \\ & & & 0 & \end{bmatrix}, \tag{9}$$

$$A_1 = \Delta(\theta S, \cdots, \theta S, \theta(S + \boldsymbol{S^0\beta})) + \tilde{B}_{0,1} \otimes I.$$

3. Steady-state analysis

In this section the model described in Section 2 will be studied in steady-state. First, we recall that the stability condition of the classical retrial queueing model is same as that of the classical queueing model [8]. For example, a retrial queueing model of the type $M/M/c$ in which all the customers waiting in the retrial orbit (of infinite size) independently attempt to capture a free server is stable if and only if $\lambda < c\mu$. However, the moment one puts a restriction in the way the retrials are modeled, the stability condition needs to be modified. This is also the case in our model studied in this paper.

3.1. Stability condition. Suppose that $A = A_0 + A_1 + A_2$ (note that A is irreducible) and that $\boldsymbol{\pi}$ is its steady-state probability vector. That is,

$$\boldsymbol{\pi} A = \mathbf{0}, \ \boldsymbol{\pi e} = 1. \tag{10}$$

On noting that the model described in Section 2 is a $QBD-$process, one can apply the classical result due to Neuts (see, e.g., [31]) to arrive at the stability condition of the retrial queueing model under study. This is given in the following theorem.

Theorem 1: The threshold retrial queueing model under study with the generator given in (1) is stable if and only if

$$\lambda < \mu \sum_{j=1}^{c} j \boldsymbol{\pi}_j \boldsymbol{e}. \tag{11}$$

Proof: First note that the equations given in (10) can be rewritten as

$$\begin{gathered} \boldsymbol{\pi}_0(\theta S - \lambda I) + \mu \boldsymbol{\pi}_1 = \mathbf{0}, \\ \theta \boldsymbol{\pi}_{j-1} \boldsymbol{S^0\beta} + \lambda \boldsymbol{\pi}_{j-1} + \boldsymbol{\pi}_j(\theta S - (\lambda + j\mu I) + (j+1)\mu \boldsymbol{\pi}_{j+1} = \mathbf{0}, \ 1 \le j \le c-1, \\ \theta \boldsymbol{\pi}_{c-1} \boldsymbol{S^0\beta} + \lambda \boldsymbol{\pi}_{c-1} + \theta \boldsymbol{\pi}_c(S + \boldsymbol{S^0\beta}) - c\mu \boldsymbol{\pi}_c = \mathbf{0}, \\ \textstyle\sum_{j=0}^{c} \boldsymbol{\pi_j e} = 1, \end{gathered} \tag{12}$$

from which it is easy to verify that

$$\theta \boldsymbol{\pi}_j \boldsymbol{S^0} + \lambda \boldsymbol{\pi}_j \boldsymbol{e} = (j+1)\mu \boldsymbol{\pi}_{j+1} \boldsymbol{e}, \ 0 \le j \le c-1. \tag{13}$$

The proof follows immediately on noting (a) $\boldsymbol{\pi} A_2 \boldsymbol{e} = \theta \sum_{j=0}^{c-1} \boldsymbol{\pi}_j \boldsymbol{S^0}$; (b) $\boldsymbol{\pi} A_0 \boldsymbol{e} = \lambda \boldsymbol{\pi}_c \boldsymbol{e}$; (c) $\theta \sum_{j=0}^{c-1} \boldsymbol{\pi}_j \boldsymbol{S^0} + \lambda(1 - \boldsymbol{\pi}_c \boldsymbol{e}) = \mu \sum_{j=0}^{c} j \boldsymbol{\pi}_j \boldsymbol{e}$ and (d) applying the stability condition for the $QBD-$ process (see, e.g., [31]), which is $\boldsymbol{\pi} A_0 \boldsymbol{e} < \boldsymbol{\pi} A_2 \boldsymbol{e}$.

Note: (1) It is worth pointing out that the vector $\boldsymbol{\pi}$ is independent of $\theta_k, 1 \leq k \leq K$. This is as is to be expected as the matrices A_0, A_1 and A_2 govern transitions away from the boundary states.
(2) As $\theta \to \infty, \boldsymbol{\pi}_j \to \boldsymbol{0}, 0 \leq j \leq c-1, \boldsymbol{\pi}_c \to \boldsymbol{\beta}(-S)^{-1}$, and hence the condition $\frac{\lambda}{c\mu} < 1$ becomes a necessary and sufficient condition for the stability. This is also intuitive as the retrial queueing model approaches the corresponding classical queueing model when $\theta \to \infty$. It is worth mentioning that in the classical retrial queue, as the number of customers in the orbit grows without bound (by fixing the retrial rate), the total (linear) retrial rate will approach ∞; thus, the retrial queue approaches the corresponding classical queue.

3.2. Steady-state probability vector. Suppose that $\boldsymbol{x} = (\boldsymbol{x}(0), \boldsymbol{x}(1), \cdots)$ denote the steady-state probability vector of the generator Q given in 1. That is, $\boldsymbol{x}$ satisfies

$$\boldsymbol{x} Q = \boldsymbol{0}, \ \boldsymbol{x} \boldsymbol{e} = 1. \tag{14}$$

We partition the vectors, $\boldsymbol{x}(i)$, for $i \geq 0$, as $\boldsymbol{x}(0) = (x_0(0), \cdots, x_c(0))$ and $\boldsymbol{x}(i) = (\boldsymbol{x}_0(i), \cdots, \boldsymbol{x}_c(i))$. First, note that $\boldsymbol{x}(0)$ is of dimension $c+1$, while $\boldsymbol{x}(i), \ i \geq 1$ is of dimension $(c+1)n$. Secondly, the j^{th} component, for $0 \leq j \leq c$, gives the steady-state probability vector that the retrial orbit is empty with exactly j servers busy serving customers; the k^{th} component of the vector $\boldsymbol{x}_j(i), \ 0 \leq j \leq c$ gives the steady-state probability vector that the retrial orbit has i customers with exactly j servers busy serving customers, and that the underlying $PH-$distribution is in phase $k, 1 \leq k \leq n$. It is worth pointing out that there can be at least one free server with at least one customer in the retrial orbit.

Under the stability condition given in (11) the steady state probability vector $\boldsymbol{x}$ is obtained (see, e.g., [31]) as shown in the following theorem.

Theorem 2: Assume that the stability condition as given in (11) holds good. Then the steady-state vector, $\boldsymbol{x}$, of the threshold retrial queueing model under study with the generator given in (1) is obtained by solving the following system

of linear equations:

$$\begin{gathered}
\boldsymbol{x}(0)\tilde{B}_{0,1} + \boldsymbol{x}(1)\tilde{B}_{1,2} = \mathbf{0}, \\
\boldsymbol{x}(0)\tilde{B}_{0,0} + \boldsymbol{x}(1)B_{1,1} + \boldsymbol{x}(2)B_{1,2} = \mathbf{0}, \\
\boldsymbol{x}(i-1)A_0 + \boldsymbol{x}(i)B_{1,1} + \boldsymbol{x}(i+1)B_{1,2} = \mathbf{0},\ 2 \le i \le N-1, \\
\boldsymbol{x}(N-1)A_0 + \boldsymbol{x}(N)B_{1,1} + \boldsymbol{x}(N+1)B_{2,2} = \mathbf{0}, \\
\begin{gathered}\boldsymbol{x}(i-1)A_0 + \boldsymbol{x}(i)B_{j+1,1} + \boldsymbol{x}(i+1)B_{j+1,2} = \mathbf{0}, \\ jN+1 \le i \le (j+1)N-1,\ 1 \le j \le K-1,\end{gathered} \\
\boldsymbol{x}(i-1)A_0 + \boldsymbol{x}(i)B_{j,1} + \boldsymbol{x}(i+1)B_{j+1,2} = \mathbf{0},\ i = jN,\ 2 \le j \le K-1, \\
\boldsymbol{x}(KN-1)A_0 + \boldsymbol{x}(KN)[B_{K,1} + RA_2] = \mathbf{0}, \\
\textstyle\sum_{i=0}^{KN-1} \boldsymbol{x}(i)\boldsymbol{e} + \boldsymbol{x}(KN)(I-R)^{-1}\boldsymbol{e} = 1,
\end{gathered} \tag{15}$$

where the matrix R is the minimal nonnegative solution to the matrix-quadratic equation:

$$R^2 A_2 + RA_1 + A_0 = 0. \tag{16}$$

Proof: First note that the $QBD-$structure of the generator given in (1), under the stability condition, yields (see, e.g., [31]) a modified matrix-geometric solution. Thus, the non-boundary states, namely, for $i \ge KN$, are given by

$$\boldsymbol{x}(i+KN) = \boldsymbol{x}(KN)R^i,\ i \ge 0, \tag{17}$$

where R satisfies the matrix-quadratic equation given in (16). With regard to the non-boundary states, the stated equations follow once we rewrite the equations given in (14).

The following lemma, which is intuitively obvious, will serve as another accuracy check in numerical implementation of the steady-state probability vector.

Lemma 1: We have

$$\Big[\sum_{j=1}^{K}\theta_j \sum_{i=(j-1)N+1}^{jN} \boldsymbol{x}(i) + \theta\boldsymbol{x}(KN)R(I-R)^{-1}\Big](\boldsymbol{e}\otimes I) = d\boldsymbol{\beta}(-S)^{-1}, \tag{18}$$

where d is the normalizing constant and is given by

$$d = \Big[\sum_{j=1}^{K}\theta_j \sum_{i=(j-1)N+1}^{jN} \boldsymbol{x}(i)\boldsymbol{e} + \theta\boldsymbol{x}(KN)R(I-R)^{-1}\boldsymbol{e}\Big]. \tag{19}$$

Proof: Post-multiplying the first equation in (15) by $\boldsymbol{e}$ we get

$$-\lambda \boldsymbol{x}(0)\boldsymbol{e}_{c+1}(c+1) + \theta_1 \boldsymbol{x}(1)((\boldsymbol{e} - \boldsymbol{e}_{c+1}(c+1)) \otimes \boldsymbol{S^0}) = 0. \tag{20}$$

Post-multiplying all but the first and last equations in (15) by $\boldsymbol{e} \otimes I$ (note that we replace the last but one equation by $\boldsymbol{x}(KN-1)A_2 + \boldsymbol{x}(KN)B_{K,1} + \boldsymbol{x}(KN+1)A_2 = \boldsymbol{0}$ and $\boldsymbol{x}(i-1)A_0 + \boldsymbol{x}(i)A_1 + \boldsymbol{x}(i+1)A_2 = \boldsymbol{0},\ i \ \geq KN+1$), and adding the resulting equations we get

$$\begin{aligned} -\lambda(\boldsymbol{x}(0)\boldsymbol{e}_{c+1}(c+1) \otimes \boldsymbol{\beta}) + \Big[\theta_1 \boldsymbol{x}(1)[(\boldsymbol{e} \otimes S) + (\boldsymbol{e}_{c+1}(c+1) \otimes \boldsymbol{S^0}\boldsymbol{\beta})] \\ \textstyle\sum_{j=1}^{K-1} \theta_{j+1} \sum_{i=jN+1}^{(j+1)N} \boldsymbol{x}(i) + \theta \sum_{i=KN+1}^{\infty} \boldsymbol{x}(i)\Big](\boldsymbol{e} \otimes I) = \boldsymbol{0}. \end{aligned} \tag{21}$$

Now post-multiplying equation (20) by $\boldsymbol{\beta}$ and adding the resulting equation with (21), we get

$$\Big[\sum_{j=1}^{K} \theta_j \sum_{i=(j-1)N+1}^{jN} \boldsymbol{x}(i) + \theta \sum_{i=KN+1}^{\infty} \boldsymbol{x}(i)\Big]\Big[\boldsymbol{e} \otimes (S + \boldsymbol{S^0}\boldsymbol{\beta})\Big] = \boldsymbol{0}, \tag{22}$$

from which using the uniqueness of the stationary vector of the generator $S + \boldsymbol{S^0}\boldsymbol{\beta}$ the stated result follows.

The following lemma, which again can be used as another accuracy check in the numerical computations, is similar to the well-known result in the classical queue. That is, the average number of busy servers is given by $\frac{\lambda}{\mu}$. Towards this end, we define

$$\tilde{\boldsymbol{y}} = \textstyle\sum_{i=0}^{\infty} \boldsymbol{y}_i,\ \texttt{where}\ \boldsymbol{y}_i = (y_{i,0}, \cdots, y_{i,c}),\ i \geq 0,$$

$$\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{x}}_0, \cdots, \tilde{\boldsymbol{x}}_c), \tag{23}$$

$$\texttt{where}\ \tilde{\boldsymbol{x}}_j = \textstyle\sum_{k=1}^{K} \theta_k \sum_{r=1}^{N} \boldsymbol{x}_{(k-1)N+r,j} + \theta \sum_{i=KN+1}^{\infty} \boldsymbol{x}_{k,j},\ 0 \leq j \leq c.$$

Note that $\boldsymbol{y}_0 = \boldsymbol{x}(0)$ and $\boldsymbol{y}_i = (\boldsymbol{x}_0(i)\boldsymbol{e}, \cdots, \boldsymbol{x}_c(i)\boldsymbol{e})$.

Lemma 2: The fraction, $\frac{\lambda}{\mu}$, gives the mean number of busy servers in the system. That is,

$$\sum_{j=1}^{c} j\tilde{y}_j = \frac{\lambda}{\mu}. \tag{24}$$

Proof: First, from the steady-state equations given in (15), verify that

$$\tilde{\boldsymbol{y}}\begin{bmatrix} -\lambda & \lambda & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & \\ & 2\mu & -(\lambda+2\mu) & \lambda & & \\ & \ddots & \ddots & \ddots & & \\ & & (c-1)\mu & -(\lambda+(c-1)\mu) & \lambda & \\ & & & c\mu & -c\mu \end{bmatrix}$$

$$+\,\tilde{\boldsymbol{x}}\begin{bmatrix} -\boldsymbol{S^0} & \boldsymbol{S^0} & & & \\ & -\boldsymbol{S^0} & \boldsymbol{S^0} & & \\ & & -\boldsymbol{S^0} & \boldsymbol{S^0} & \\ & & & \ddots & \ddots \\ & & & -\boldsymbol{S^0} & \boldsymbol{S^0} \\ & & & & \boldsymbol{0} \end{bmatrix} = \boldsymbol{0}. \tag{25}$$

From equation (25) it is easy to verify

$$\tilde{\boldsymbol{x}}_j \boldsymbol{S^0} + \lambda \tilde{y}_j = (j+1)\mu \tilde{y}_{j+1},\ 0 \le j \le c-1, \tag{26}$$

from which we get

$$\sum_{j=0}^{c-1} \tilde{\boldsymbol{x}}_j \boldsymbol{S^0} + \lambda \sum_{j=0}^{c-1} \tilde{y}_j = \sum_{j=1}^{c} j\mu \tilde{y}_j. \tag{27}$$

The stated result follows from the above equation on noting that $\sum_{j=0}^{c} \tilde{y}_j = 1$ and $\lambda \tilde{y}_c = \sum_{j=0}^{c-1} \tilde{\boldsymbol{x}}_j \boldsymbol{S^0}$. Note that the latter equation is due to the fact that in steady-state the rate at which a customer enters a retrial orbit should be equal to the rate at which a retrial customer captures a free server.

3.3. Computational procedures. In this section we will briefly outline the computational procedures involved in obtaining the rate matrix R (of dimension $(c+1)n$) and the steady-state probability vector, $\tilde{\boldsymbol{x}}$ which are key ingredients to system performance measures to qualitatively study the model. First, we look at the R matrix. One can use a number of well-known methods such as logarithmic reduction [27] to compute R, especially when the dimension of R is of reasonable size. If the dimension is prohibitively large, say, when c or n or both large, then one should use (block) Gauss-Siedel iteration by exploiting the special structure of the coefficient matrices A_0, A_1, and A_2.

Due to structure of the coefficient matrices in (16), it is easy to verify that R is of the form

$$R = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \\ R_0 & R_1 & \cdots & R_c \end{bmatrix}. \tag{28}$$

The above form of R along with the structure of the coefficient matrices can be exploited to rewrite the matrix-quadratic equation given in (16) as follows.

$$R_0 = \mu R_1(\lambda I - \theta S)^{-1},$$

$$R_k = [\lambda R_{k-1} + (k+1)\mu R_{k+1} + \theta R_c R_{k-1}\boldsymbol{S^0\beta}][(\lambda + k\mu)I - \theta S]^{-1},\ 1 \le k \le c-1,$$

$$R_c = [\lambda R_{c-1} + \theta R_c R_{c-1}\boldsymbol{S^0\beta} + \lambda I][(\lambda + k\mu)I - \theta(S + \boldsymbol{S^0\beta})]^{-1}. \tag{29}$$

Similar to exploiting the structure in computing R, one can do the same for computing the steady-state probability vector, $\boldsymbol{x}$. We will briefly display two sets of such equations.

$$x_0(0) = \tfrac{\mu}{\lambda} x_1(0),$$

$$x_k(0) = \tfrac{1}{\lambda + k\mu}\Big[\lambda x_{k-1}(0) + (k+1)\mu x_{k+1}(0) + \theta_1 \boldsymbol{x}_{k-1}(1)\boldsymbol{S^0}\Big],\ 1 \le k \le c-1,$$

$$x_c(0) = \tfrac{1}{\lambda + k\mu}\Big[\lambda x_{c-1}(0) + \theta_1 \boldsymbol{x}_{c-1}(1)\boldsymbol{S^0}\Big],$$

$$\boldsymbol{x}_0(1) = \mu \boldsymbol{x}_1(1)(\lambda I - \theta_1 S)^{-1},$$

$$\boldsymbol{x}_k(1) = [\lambda \boldsymbol{x}_{k-1}(1) + (k+1)\mu \boldsymbol{x}_{k+1}(1) + \theta_1 \boldsymbol{x}_{k-1}(2)\boldsymbol{S^0\beta}][(\lambda + k\mu)I - \theta_1 S]^{-1},\ 1 \le k \le c-1,$$

$$\boldsymbol{x}_c(1) = [\lambda x_c(0)\boldsymbol{\beta} + \lambda \boldsymbol{x}_{c-1}(1) + \theta_1 \boldsymbol{x}_{c-1}(2)\boldsymbol{S^0\beta}][(\lambda + c\mu)I - \theta_1(S + \boldsymbol{S^0\beta})]^{-1}.$$

3.4. System performance measures. In order to analyze the model qualitatively, we need to look into some key system performance measures. Here, we will list a few performance measures along with the expressions for their computations.

(1) *Probability that the system is idle.* The probability, P_{idle}, that the system is idle (i.e. all servers are idle and the retrial buffer is empty) at an arbitrary time is given by

$$P_{idle} = x_{0,0}.$$

(2) *Probability that the orbit is empty.* The probability, $P_{O-empty}$, that the retrial buffer is empty at an arbitrary time is given by

$$P_{O-empty} = \boldsymbol{x}(0)\boldsymbol{e}.$$

(3) *Probability that the system is in various mode.* The probability, PSM_r, that the system is operating in mode $r, 0 \le r \le K+1$, is of interest. Note that the system is said to be in (a) mode 0 when there is no one in the retrial orbit; (b) in mode $i, 1 \le i \le K$ when the rate of retrial is

given by θ_i; and (c) in mode $K+1$ when the rate of retrials is given by θ.

$$PSM_r = \begin{cases} \boldsymbol{x}(0)\boldsymbol{e}, \ r = 0, \\ \sum_{i=1}^{N} \boldsymbol{x}((r-1)N+i)\boldsymbol{e}, \ 1 \le r \le K, \\ \boldsymbol{x}(KN)R(I-R)^{-1}\boldsymbol{e}, \ r = K+1. \end{cases}$$

(4) *Probability mass function of the number of busy servers.* The probability mass function of the number of busy servers is given by $\{\tilde{y}_j, \ 0 \le j \le c\}$ (see equation (23)). From this probability mass function, we can get the mean, μ_0, number of busy servers and the standard deviation, σ_0 of the number of busy servers.

(5) *Mean number of customers in retrial orbit.* The mean, μ_{RO}, is calculated as

$$\begin{aligned} \mu_{RO} &= \textstyle\sum_{i=1}^{KN-1} i\boldsymbol{x}(i)\boldsymbol{e} + KN \sum_{k=0}^{c-1} \boldsymbol{x}_k(KN)\boldsymbol{e} \\ &+KN\boldsymbol{x}_c(KN)(I-R_c)^{-1} \textstyle\sum_{k=0}^{c-1} R_k\boldsymbol{e} + KN\boldsymbol{x}_c(KN)(I-R_c)^{-1}\boldsymbol{e} \\ &+\boldsymbol{x}_c(KN)(I-R_c)^{-2}\Big(\boldsymbol{e} + \textstyle\sum_{k=0}^{c-1} R_k\boldsymbol{e}\Big) + (KN-1)\boldsymbol{x}_c(KN)(I-R_c)^{-1}\boldsymbol{e}. \end{aligned}$$

(6) *Probability of blocking* The probability, P_{block}, that an arriving customer finds all servers busy is obtained as

$$P_{block} = x_0(c) + \sum_{i=1}^{\infty} \boldsymbol{x}_c(i)\boldsymbol{e}.$$

(7) *Rate of successful capture of a free server by a customer from retrial orbit.* The rate, $\xi_r, 1 \le r \le K+1$, of successful capture of a free server from the retrial orbit is calculated as

$$\xi_r = \begin{cases} \theta_r \sum_{j=1}^{N} \sum_{k=0}^{c-1} \boldsymbol{x}_k((r-1)N+j)\boldsymbol{S^0}, \ 1 \le r \le K, \\ \theta\boldsymbol{x}_c(KN)R(I-R)^{-1}[(\boldsymbol{e}(c) - \boldsymbol{e}_c(c)) \otimes \boldsymbol{S^0}], \ r = K+1. \end{cases}$$

4. Comparison with Shin-Moon model [38]

In this section, we compare our threshold model with the retrial model studied in ([38]) through a few numerical examples. In [38], the authors report a number of numerical examples dealing with Erlang, hyperexponential, mixtures of Erlang, and compositions of mixtures of two Erlangs. In getting these numerical results, the authors in [38] recourse to truncation method (to find a cut-off point for the level-dependent $QBD-$process that is used to study the model) proposed in [37] and an approximation method for reporting a few system performance measures.

For the sake of completeness and for continuity of discussion, we will use the same notation for the distributions considered for the numerical examples as in [38]. In the following m_1 denote the mean of retrial times of each customer.

That is m_1 denotes the mean of the $PH-$distribution governing the retrials.
1. Erlang ($E_k(\nu)$): This is Erlang of order k with parameter ν.
2. Hyperexponential (H_2): This is hyperexponential with two states with probability density function, $f(t)$, given by $f(t) = p\nu_1 e^{-\nu_1 t} + (1-p)\nu_2 e^{-\nu_2 t},\ t \geq 0$, where $p = 0.5(1 + \sqrt{(\frac{2}{3})}), \nu_1 = \frac{2p}{m_1}, \nu_2 = \frac{2(1-p)}{m_1}$.
3. Mixture of two Erlangs ($MER_k(p; \nu_1, \nu_2)$): This is a mixture of two Erlangs of order k and is obtained as $MER_k(p; \nu_1, \nu_2) = pE_k(\nu_1) + (1-p)E_k(\nu_2)$.
4. Composition of the mixture of Erlangs ($CE_{k,j}(p; \nu_1, \nu_2)$): This is a composition of the mixture of two Erlangs of orders k and j, whose Laplace transform is given by $f^*(s) = p\Big(\frac{\nu_1}{\nu_1+s}\Big)^k \Big(\frac{\nu_2}{\nu_2+s}\Big)^j + (1-p)\Big(\frac{\nu_2}{\nu_2+s}\Big)^j,\ Re(s) \geq 0$.

The parameter values chosen for comparison purposes in this section are (see [38]) as follows. $\lambda = c\mu, c = 5$, and $\mu = 1$. The traffic intensity, ρ, and the mean retrial times, m_1 are varied and the values are displayed in the comparison tables below. In the following, the parameters for the mixture of two Erlangs, denoted by MER_3, are $MER_3\Big(0.0740741; \frac{4}{3m_1}, \frac{10}{3m_1}\Big)$; the two distributions corresponding to the composition of mixtures of Erlangs, $CE_{3,1}^{(1)}$ and $CE_{3,1}^{(2)}$ have parameters, respectively, given by $CE_{3,1}^{(1)} = CE_{3,1}\Big(0.007773; \frac{0.146991}{m_1}, \frac{1.188568}{m_1}\Big)$ and $CE_{3,1}^{(2)} = CE_{3,1}\Big(0.185487; \frac{0.61971}{m_1}, \frac{9.79811}{m_1}\Big)$.

While we used the same set of values for the common parameters such as λ, μ, c, and ρ, some other additional parameters that are in our model are taken as follows. We fixed $K = 100, N = 1, \theta_1 = \frac{1}{m_1}$, and $\theta_k = k\theta_1$. The value of θ is obtained as the minimum value for which $\frac{\lambda}{\mu \sum_{j=1}^{c} j\boldsymbol{\pi}_j \boldsymbol{e}}$ is close to a given value of ρ. That is, θ is such that

$$\Big|\frac{\lambda}{\mu \sum_{j=1}^{c} j\boldsymbol{\pi}_j \boldsymbol{e}} - \rho\Big| < 10^{-3}. \tag{30}$$

A few comments about the choice of K and N. One can think of K as the truncation point similar to the one chosen in [37] to truncate the level-dependent $QBD-$process. The idea of fixing $N = 1$ is to sort of resemble the classical retrial queue with phase type retrial times. Unlike the models in [37], wherein the number of phases is fixed at 2 in order to minimize the dimension of the problem on hand, and in [38], where no such restriction is placed but approximation methods are proposed due to dimensionality issues, we use a common $PH-$distribution whose mean depends on the threshold parameters, K and N.

In Tables 1 through 5 we display the difference percentages, which are calculated as $100|\frac{Threshold - ShinMoon}{ShinMoon}|\%$. The results due to Shin and Moon [38] correspond to the ones displayed in Tables 1 through 5 of their paper. Specifically, we use their exact results as presented in their Tables 1 through 4, the approximation denoted by $\hat{L}$ in their Table 5.

Table 1: Difference percentages for $P(block)$

	$\rho = 0.3$				$\rho = 0.5$				$\rho = 0.8$				
θ_1	10	1	0.2	0.1	10	1	0.2	0.1	10	1	0.2	0.1	0.05
E_2	0.37	0.18	0.07	0.40	0.01	0.05	0.10	0.12	0.21	0.39	0.40	0.85	1.06
H_2	0.11	0.36	0.32	0.13	0.28	0.30	0.33	0.55	1.17	0.34	3.53	4.65	6.30

Table 2: Difference percentages for σ_0

	$\rho = 0.3$				$\rho = 0.5$				$\rho = 0.8$				
θ_1	10	1	0.2	0.1	10	1	0.2	0.1	10	1	0.2	0.1	0.05
E_2	0.02	0.01	0.02	0.03	0.03	0.04	0.02	0.14	0.14	0.21	0.33	0.77	0.98
H_2	0.00	0.03	0.00	0.02	0.03	0.02	0.06	0.13	0.66	0.14	3.02	4.13	5.46

Table 3: Difference percentages for μ_{RO}

θ_1	10^6	10	1	0.2	0.1	0.05
E_2	0.02	0.28	1.23	1.24	0.35	0.56
H_2	0.02	0.66	0.61	0.81	1.40	1.83

Table 4: Difference percentages for μ_{RO}

	$\rho = 0.3$				$\rho = 0.5$				$\rho = 0.8$				
θ_1	10	1	0.2	0.1	10	1	0.2	0.1	10	1	0.2	0.1	0.05
E_2	0.37	1.30	1.02	0.66	0.28	1.23	1.16	0.35	0.95	3.14	6.71	8.18	8.81
H_2	0.88	1.13	0.37	0.22	0.66	0.61	0.81	1.40	7.05	18.91	30.19	33.88	31.14

Table 5: Difference percentages for μ_{RO}

	$\rho = 0.5$					$\rho = 0.8$				
θ_1	10	1	0.2	0.1	0.05	10	1	0.2	0.1	0.05
E_4	4.66	2.27	0.26	0.95	1.80	5.28	4.37	10.68	12.77	13.72
MER_3	4.79	2.32	0.10	0.45	1.27	10.10	2.48	7.20	9.21	10.14
$CE_{3,1}^{(1)}$	3.52	2.21	0.37	0.06	0.34	0.72	11.30	12.76	12.61	11.51
$CE_{3,1}^{(2)}$	9.44	11.88	0.19	7.22	29.64	5.98	11.18	63.43	86.73	74.52

The above tables show the difference percentages for all but a few cases to be reasonably acceptable. Those cases for which the percentages exceed 15%, we recalculated the values by increasing the value of θ_1. The modified difference percentages are displayed in Tables 4b and 5b, related, respectively, to Tables 4 and 5.

Table 4b: Modified difference percentages for μ_{RO} related to Table 4

	$\rho = 0.3$				$\rho = 0.5$				$\rho = 0.8$				
θ_1	10	1	0.2	0.1	10	1	0.2	0.1	10	1	0.2	0.1	0.05
E_2	0.37	1.30	1.02	0.66	0.28	1.23	1.16	0.35	0.95	3.14	6.71	8.18	8.81
H_2	0.88	1.13	0.37	0.22	0.66	0.61	0.81	1.40	7.05	7.13^*	$4.24^\dagger$	$5.23^\ddagger$	$1.60^\S$

$^*\theta_1 = 1.2;^\dagger\ \theta_1 = 0.26;^\ddagger\ \theta_1 = 0.13;^\S\ \theta_1 = 0.07.$

Table 5b: Modified difference percentages for μ_{RO} related to Table 5

	$\rho = 0.5$					$\rho = 0.8$				
θ_1	10	1	0.2	0.1	0.05	10	1	0.2	0.1	0.05
E_4	4.66	2.27	0.26	0.95	1.80	5.28	4.37	10.68	12.77	13.72
MER_3	4.79	2.32	0.10	0.45	1.27	10.10	2.48	7.20	9.21	10.14
$CE_{3,1}^{(1)}$	3.52	2.21	0.37	0.06	0.34	0.72	11.30	12.76	12.61	11.51
$CE_{3,1}^{(2)}$	9.44	11.88	0.19	7.22	10.71^*	5.98	11.18	$12.10^\dagger$	$3.24^\ddagger$	$14.15^\S$

$^*\theta_1 = 0.075;^\dagger\ \theta_1 = 0.4;^\ddagger\ \theta_1 = 0.2;^\S\ \theta_1 = 0.09.$

It is clear from the above tables that our threshold model with a special way of modeling the retrial times with $PH-$distribution appears to perform well. In the next section, we will discuss additional examples by varying N.

5. Illustrative numerical examples

In this section we will discuss a few illustrative examples to bring out the qualitative nature of our threshold model. Towards this end, we consider three distributions for the retrial times. It should be pointed out these are special types of $PH-$distributions.

A. E_{10}: This is an Erlang of order 10. The parameter will be chosen so that the rate will be θ_k or θ, depending on the context. Thus, if the rate is fixed at θ, then we have an Erlang of order 10 with parameter 10θ so that the mean of this Erlang will be $\frac{1}{\theta}$.
B. E_1: This is an exponential distribution whose parameter will be θ_k or θ depending on the context.
C. H_5: This is hyperexponential with 5 states. The mixing probabilities and the rates in each of those states are, respectively, $(0.60, 0.24, 0.10, 0.05, 0.01)$ and $a(1000, 100, 10, 1, 0.1)$, where $a = \theta_k$ or $a = \theta$ depending on the context.

Note that the above three distributions are qualitatively different in that they have different structure. While we will keep the means of these three distributions to be the same (so as to compare properly), they have different variance structure. The ratio of the standard deviations of E_1 and H_5 with respect to E_{10} are, respectively, 3.1623 and 27.9492.

One can choose the values of $\theta_k, 1 \leq k \leq K$, in a variety of ways. In our examples below, we take $\theta_k = k\theta_1, 2 \leq k \leq K$, where θ_1 will be is specified along with θ.

EXAMPLE 1: In this example, we investigate the effect of the traffic intensity, c, and the type of retrial distribution on the minimum value of θ for which the retrial queue traffic load will be close to a given value of $\rho = \frac{\lambda}{c\mu}$ (see (30)). Towards this end, we fix $\lambda = 1, \mu = \frac{1}{c\rho}$, and vary c and ρ. It is worth pointing out that θ does not depend on N, K, and $\theta_k, 1 \leq k \leq K$. The graph of the plot of $ln(\theta)$ as a function of c and ρ under the three distributions for the retrial times is given in Figure 1 below.

A quick look at Figure 1 reveals the following observations.

(1) Generally speaking, it appears that as c is increased (for fixed ρ), $ln(\theta)$ (and hence θ) approaches a constant value for all ρ up to 0.8. However, for $\rho > 0.8$, we see a decreasing trend in $ln(\theta)$ as c is increased. This latter phenomenon can be explained intuitively as follows. When ρ is large, it is more likely that an arriving customer will enter into the retrial orbit and hence even a moderate value of $ln(\theta)$ will be sufficient for a retrial customer to capture a free server.

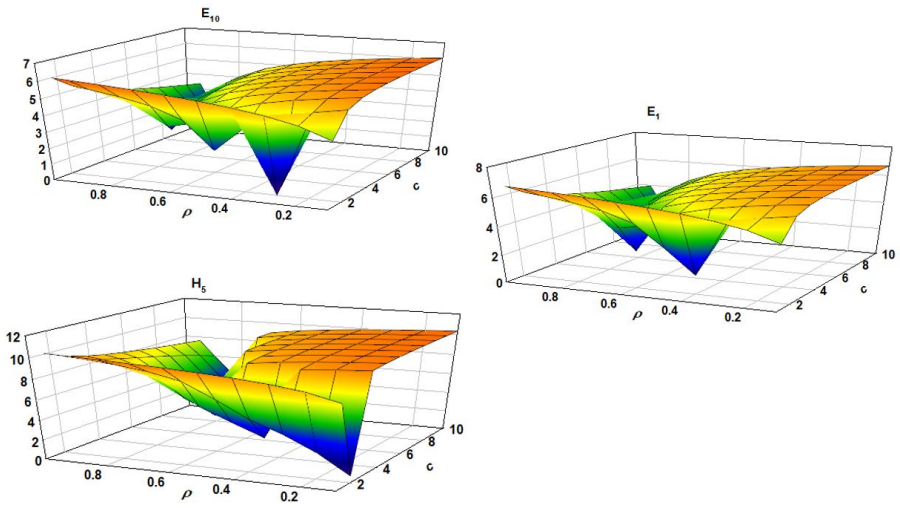

FIGURE 1. $Ln(\theta)$ under various scenarios

(2) Again, generally speaking, H_5 retrial time with a high variability appears to need a large $ln(\theta)$ compared to E_{10} and E_1 retrial times, especially for large values of ρ.
(3) For fixed c, we see an interesting behavior in $ln(\theta)$ as a function of ρ in that this measure appears to decrease up to a certain point (depending on the type of retrial and the value c) and then increases. The "dips", if any, appear to occur for low to moderate values of ρ for $c \geq 2$, while for $c = 1$ the measure appears to increase as ρ is increased.

EXAMPLE 2: The purpose of this example is to compare the two measures, P_{block} and μ_{RO}, when N is varied. Here we fix $K = 100, \lambda = 1, \theta_1 = 1.0, \mu = \frac{\lambda}{c\rho}$, and vary $c = 1, 2, 5$, and $\rho = 0.5, 0.95$. The value of θ is obtained so as to have the traffic intensity of the threshold retrial queue to be close to the given value of ρ (see (30)). By considering the three distributions, E_{10}, E_1, and H_5, and looking at $N = 1, 2, 3$, we display the two measures in Figure 2 below. It is clear from this figure that while the measure P_{block} appears to be not sensitive to the three values of N considered when $\rho = 0.5$, we notice sensitivity to N when $\rho = 0.95$ only in the case of H_5 retrial times. This indicates the role of variability in the retrial times. With respect to the measure, μ_{RO}, we see the sensitivity to N for all scenarios considered. As is to be expected this measure decreases as c is increased. Also, as seen in the classical queue, here also we see the mean number in retrial orbit appears to increase with increasing variability (in the retrial times).

A natural question that arises here is how should the retrial rate be increased so that the mean number in retrial orbit for the threshold model under study

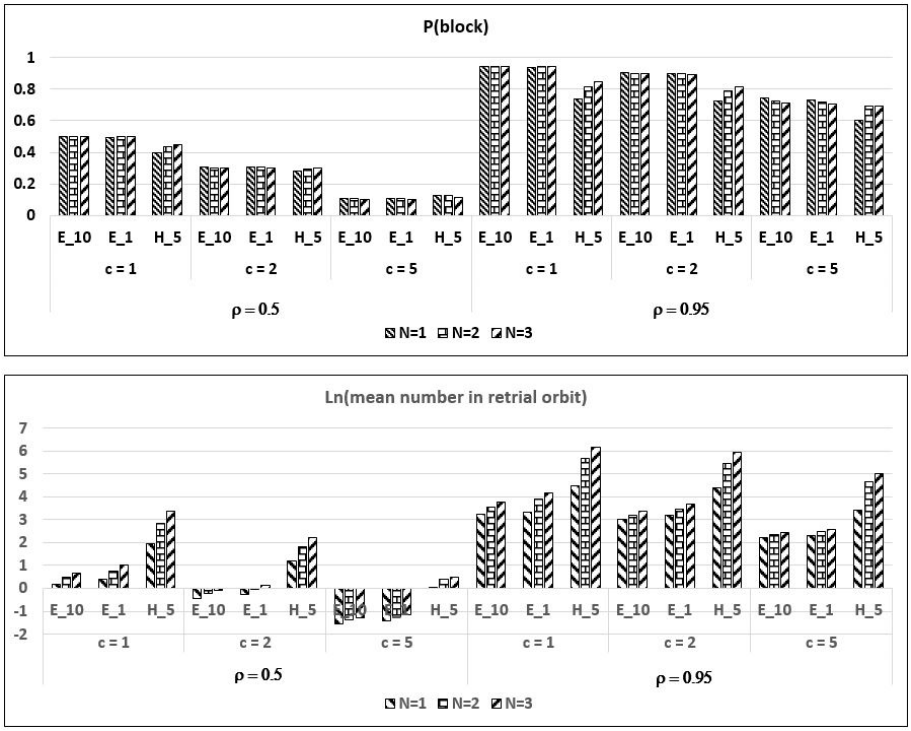

FIGURE 2. P_{block} and $Ln(\mu_{RO})$ under various scenarios

here will be close to the corresponding classical retrial queue. The next example discusses this interesting question. Since there are no analytical results available for multi-server retrial queue with phase type retrial times without placing any restriction on the retrial model, we use the simulated model developed in [15] to get μ_{RO}.

EXAMPLE 3: In this example we explore how the value of θ_1 varies under different scenarios so that the measure, μ_{RO}, is within 5% of the corresponding value in the classical retrial queue. Due to lack of theoretical results for these cases, we use the simulated values based on the model in [15]. Towards this end we fix $\lambda = 1, \mu = \frac{\lambda}{c\rho}$, and vary $c = 1, 2, 5$, and $\rho = 0.5, 0.95$. First, we obtain the value of θ (which does not depend on θ_1) so as to have the traffic intensity of the threshold retrial queue to be close to the given value of ρ (see (30)). By considering the three distributions, E_{10}, E_1, and H_5, and looking at $N = 1, 2, 3, 4, 5$ and $K = 50, 60, \cdots, 100$, we display the value of θ_1 in Figures 3 and 4 below.

An examination of the above two figures reveal that (a) as N is increased, θ_1 appears to increase; (b) the higher the variation in retrial times, the larger the

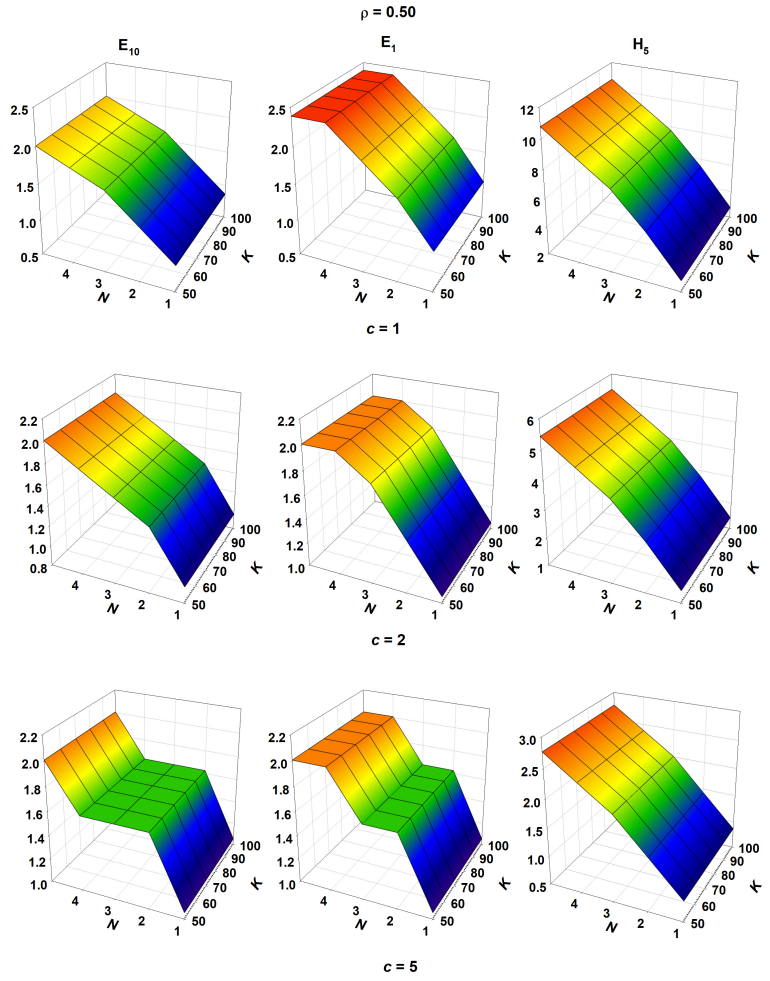

FIGURE 3. θ_1 under various scenarios when $\rho = 0.5$

value of θ_1 is required; (c) While the values of θ_1 appear to increase when going from $\rho = 0.5$ to $\rho = 0.95$, the increase is several fold for H_5 retrials as compared to the other two retrial times.

Note: The classical retrial models simulated in ([15]) are such that each orbiting customer attempts to capture a free server using his/her own (identical) $PH-$distribution. Thus, this example illustrates the worthiness of our threshold model (with reasonable size state space) to approximate a classical retrial queueing model with $PH-$distribution with proper choice of the retrial rates.

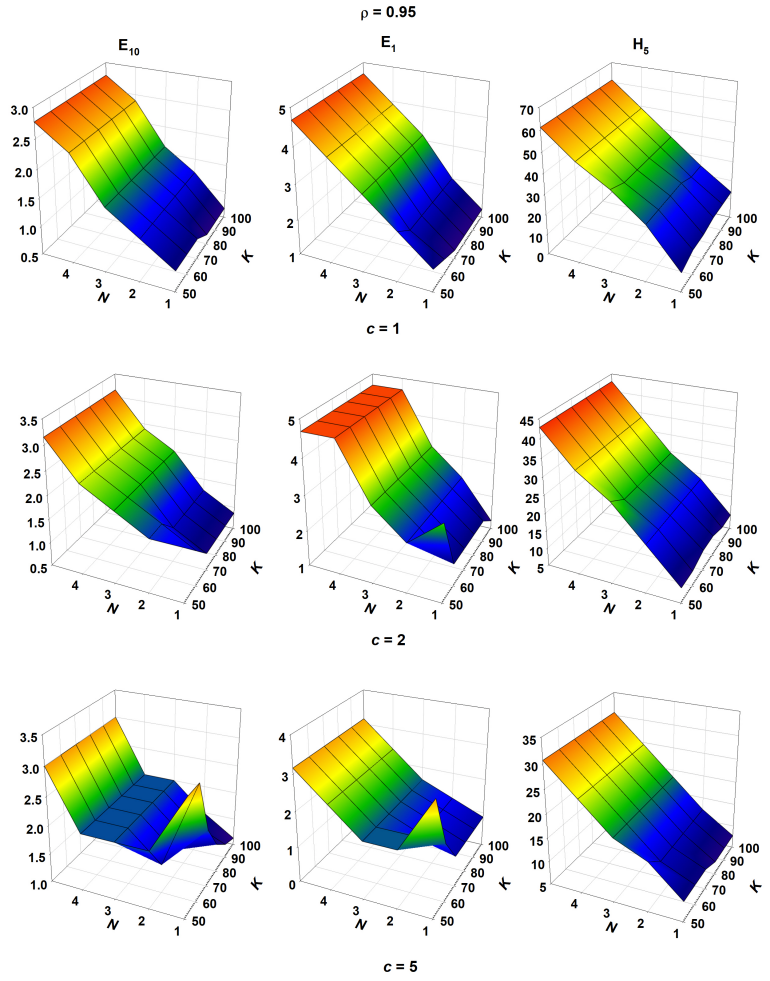

FIGURE 4. θ_1 under various scenarios when $\rho = 0.95$

In some applications it is of interest to see where the queueing system is spending most of the times as well as the state from where the customers have a high probability of capturing a free server. In the next example we look at the two measures, $PSM_r, 0 \leq r \leq K+1$ and $\xi_r, 1 \leq r \leq K+1$.

EXAMPLE 4: The purpose of this example is to identify the values of K, N, and r for which the probability that the system will be operating in mode r will be the largest among $K+1$ modes. Similarly, we identify the values for which

the customer has the highest probability of capturing a free server. That is, we look for $(K_1^*, r_1^*, N_1^*, \xi^*)$ and $(K_2^*, r_2^*, N_2^*, PSM^*)$, respectively, for the two measures, ξ and PSM. These values are displayed in Tables 6 and 7, respectively, under various scenarios. It should be pointed that we searched for these by fixing $\lambda = 1, \theta_1 = 1, \theta = KN\theta_1$; vary N from 1 to 10 and K from 1 to 50. The value of μ is chosen so that the traffic intensity of the threshold model is close to a given value of ρ. That is, choose μ so that $\left|\mu - \frac{\lambda}{\rho \sum_{j=1}^{c} \boldsymbol{\pi e}}\right| < 10^{-3}$.

Table 6: Mode for ξ

ρ	ToR	$c = 1$	$c = 2$	$c = 5$
0.5	E_{10}	(1, 1, 10, 0.447)	(2, 1, 10, 0.294)	(1, 1, 6, 0.117)
	E_1	(1, 1, 10, 0.408)	(3, 1, 10, 0.292)	(1, 1, 5, 0.121)
	H_5	(1, 1, 10, 0.191)	(1, 1, 10, 0.163)	(4, 1, 10, 0.107)
0.95	E_{10}	(1, 2, 9, 0.819)	(1, 2, 7, 0.730)	(1, 2, 3, 0.626)
	E_1	(1, 2, 10, 0.805)	(1, 2, 8, 0.724)	(1, 2, 4, 0.620)
	H_5	(50, 51, 10, 0.753)	(25, 26, 10, 0.676)	(11, 12, 10, 0.546)

Table 7: Mode for PSM

ρ	ToR	$c = 1$	$c = 2$	$c = 5$
0.5	E_{10}	(28, 1, 10, 0.724)	(1, 0, 1, 0.691)	(14, 0, 1, 0.882)
	E_1	(1, 1, 10, 0.779)	(1, 0, 1, 0.673)	(24, 0, 1, 0.868)
	H_5	(1, 1, 10, 0.698)	(1, 1, 10, 0.653)	(5, 0, 1, 0.687)
0.95	E_{10}	(1, 2, 1, 0.870)	(1, 2, 1, 0.852)	(1, 2, 1, 0.808)
	E_1	(1, 2, 1, 0.875)	(1, 2, 1, 0.850)	(1, 2, 1, 0.792)
	H_5	(1, 2, 1, 0.925)	(1, 2, 1, 0.915)	(1, 2, 1, 0.905)

An examination of the above tables reveals the following.

- With respect to the probability of capturing a free server, we notice that in the case of $\rho = 0.5$, both E_{10} and E_1 retrial times have the largest probability for a customer to capture a server. But for all the three retrial times and for $c = 1, 2$, and 5, we notice that their respective largest probability of capturing a free server occurs for customers from the orbit. This is somewhat counter intuitive at first as the traffic load is only 0.5 and so one would expect an arriving (new) customer to capture a free server, especially, in multiple server case. However, the retrial rate, θ, chosen for this example might be large enough that the retrial customers capture a server more often than a new arriving customer.
- When $\rho = 0.95$, we notice a similar behavior for the probability of capturing a free server; however, the values are significantly larger compared to $\rho = 0.5$. Another interesting observation is that for all the three retrial times we see the largest probability of capturing a free server occurs when the customers are seeking a free server when the system is operating in its last mode (i.e., namely when $r = K + 1$).
- The measure, PSM, indicates that the system seems to operate most of the times in extreme modes.

6. Concluding remarks

In this paper we introduced a new type threshold model in the context of the classical retrial queue. Assuming the retrial times to be of phase type but with the rates driven by the threshold parameters, we analyzed the model in steady-state. Some illustrative numerical examples to bring out the qualitative nature of the model were presented. We assumed that all the $PH-$distributions governing the orbiting customers' retrial attempts to have the same number of phases but with different rates that depend on the threshold parameters. However, this assumption can be relaxed to have different orders (essentially different $PH-$distributions). The needed modifications are in the matrices governing the transitions from one threshold interval to the next (adjacent) one and all other structure will still be preserved and one can analyze the model along the lines described here. We also showed how our threshold model can be used to approximate a classical retrial queueing model with phase type retrials. The model considered in this paper can be extended to include more versatile arrival process as well as more robust service times. These will be topics for future research.

Acknowledgments: The author is grateful to the editor and the anonymous referees for their constructive suggestions that improved the presentation of the paper.

References

1. J.R. Artalejo, *Accessible Bibliography on Retrial Queues*, Mathematical and Computer Modelling, **30** (1999), 1-6.
2. J.R. Artalejo, M.J. Lopez-Herrero, *On the Busy Period of the M/G/1 Retrial Queue*, Naval Research Logistics **47** (2000), 115-127.
3. J.R. Artalejo, A. Gomez-Corral, M.F. Neuts, *Analysis of multiserver queues with constant retrial rate*, Euro. J. Oper. Res. **135** (2002), 569-581.
4. J.R. Artalejo, S.R. Chakravarthy, *Algorithmic Analysis of a $MAP/PH/1$ Retrial Queue*, TOP **14** (2006), 293-332.
5. J.R. Artalejo, A. Economou, M.J. Lopez-Herrero, *Algorithmic approximations for the busy period distribution of the $M/M/c$ retrial queue*, Euro. J. Oper. Res. **176** (2007), 1687-1702.
6. J.R. Artalejo, A. Gomez-Corral, *Modelling communication systems with phase type service and retrial times*, IEEE Communications Letters **11** (2007), 955-957.
7. J.R. Artalejo, S.R. Chakravarthy, M.J. Lopez-Herrero, *The busy period and the waiting time analysis of a $MAP/M/c$ queue with finite retrial group*, Stochastic Analysis and Applications **25** (2007), 445-469.
8. J.R. Artalejo, A. Gomez-Corral, *Retrial Queueing Systems: A Computational Approach*, Springer-Verlag, Berlin, Heidelberg, 2008.
9. K. Avrachenko, U. Yechiali, *Retrail networks with finite buffers and their application to internet data traffic*, Probability in the Engineering and Informational Sciences **22** (2008), 519-536.
10. K. Avrachenko, U. Yechiali, *On tandem blocking queues with a common retrial queue*, Computers and Operations Research **37** (2010), 1174-1180.
11. N. Baer, R.J. Boucherie, J-K. Ommeren, *The $PH/PH/1$ multi-threshold queues*, In: B. Sericola, M. Telek, and G. Horvath (Eds.): ASMTA 2014, LNCS 8499, 95-109. Springer International Publishing Switzerland, 2014.

12. S.R. Chakravarthy, A. Krishnamoorthy and V.C. Joshua, *Analysis of a Multi-server Retrial Queue with Search of Customers from the Orbit*, Performance Evaluation **63** (2006), 776-798.
13. S.R. Chakravarthy, *A Multi-server Queueing Model with Markovian Arrivals and Multiple Thresholds*, Asia-Pacific Journal of Operational Research **24** (2007), 223-243.
14. S.R. Chakravarthy, *A multi-server synchronous vacation model with thresholds and a probabilistic decision rule*, European Journal of Operational Research **182** (2007), 305-320.
15. S.R. Chakravarthy, *Analysis of $MAP/PH/c$ retrial queue with phase type retrials - Simulation approach*, Communications in Computer and Information Science **356** (2013), 37-49.
16. B.D. Choi, Y.W. Shin, W.C. Ahn, *Retrial queues with collision arising from unslotted CSMA/CD protocol*, Queueing Systems **11** (1992), 335-356.
17. D.I. Choi, T.S. Kim, S. Lee, *Analysis of an MMPP/G/1/K Queue with Queue Length Dependent Arrival Rates, and its Application to Preventive Congestion Control in Telecommunication Networks*, European Journal of Operational Research **187** (2008), 652-659.
18. J.E. Diamond, A.S. Alfa, *Approximation method for M/PH/1 retrial queues with phase type inter-retrial times*, European Journal of Operational Research **113** (1999), 620-631.
19. D.V. Efrosinin, *Controlled Queueing Systems with Heterogeneous Servers*, Ph.D. Dissertation, Trier University, Germany, 2004.
20. D. Efrosinin, L. Breuer, *Threshold policies for controlled retrial queues with heterogeneous servers*, Ann Oper Res **141** (2006), 139-162.
21. G.I. Falin, J.G.C. Templeton, *Retrial Queues, Chapman and Hall*, London, 1997.
22. A. Graham, *Kronecker Products and Matrix Calculus with Applications*, Ellis Horwood, Chichester, UK, 1981.
23. Q.M. He, H. Li, Y.Q. Zhao, *Ergodicity of the $BMAP/PH/s/s+K$ retrial queue with PH-retrial times*, Queueing Systems **35** (2000), 323-347.
24. O.C. Ibe, J. Keilson, *Multi-server threshold queues with hysteresis*, Performance Evaluation **21** (1995), 185-2135.
25. S.B. Khodadadi, F. Jolai, *A fuzzy based threshold policy for a single server retrial queue with vacations*, Central European Journal of Operations Research **20** (2012), 281-297.
26. J. Kim, B. Kim, *A survey of retrial queueing systems*, Ann. Oper. Res. **247** (2016), 3-36.
27. G. Latouche, V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*, SIAM, 1999.
28. W. Lin, P.R. Kumar, *Optimal Control of a Queueing System with two Heterogeneous Servers* IEEE Trans. on Autom. Control **29** (1984), 696-703.
29. H. Luh, I. Viniotis, *Threshold control policies for heterogeneous server systems*, Mathematical Methods of OR **55** (2002), 121-142.
30. M. Marcus, H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, MA, 1964.
31. M.F. Neuts, *Matrix-geometric solutions in stochastic models: An algorithmic approach*, The Johns Hopkins University Press, Baltimore, MD. [1994 version is Dover Edition], 1981.
32. R. Nobel, H.C. Tijms, *Optimal Control of a Queueing System with Heterogeneous Servers*, IEEE Transactions on Autom. Control **45** (2002), 780-784.
33. V. Ponomarov, E. Lebedev, *Finite Source Retrial Queues with State-Dependent Service Rate*, Communications in Computer and Information Science **356** (2013), 140146.
34. V. Ponomarov, E. Lebedev, *Optimal Control of Retrial Queues with Finite Population and State-Dependent Service Rate*, Advances in Intelligent Systems and Computing **754** (2018), 359-369.
35. V.V. Rykov, D.V. Efrosinin, *Numerical Analysis of Optimal Control Polices for Queueing Systems with Heterogeneous Servers*, Information Processes **2** (2002), 252-256.
36. M. Senthilkumar, K. Sohraby, K. Kim, *On a multiserver retrial queue with phase type retrial time*, Mathematical and Computational Models. Eds: R. Nadarajan et al., 65-78, 2012, Narosa Publishing House, New Delhi, India.

37. Y.W. Shin, *Algorithmic solutions for $M/M/c$ retrial queue with PH2 retrial time*, Journal of Applied Mathematics and Informatics **29** (2011), 803-811.
38. Y.W. Shin, D.H. Moon, *Approximation of $M/M/c$ retrial queue with PH-retrial times*, European Journal of Operational Research **213** (2011), 205-209.
39. W-H. Steeb and Y. Hardy, *Matrix Calculus and Kronecker Product*, World Scientific Publishing, Singapore, 2011.
40. T. Yang, M.J.M. Posner, J.G.C. Templeton, H. Li, *An approximation method for the $M/G/1$ retrial queues with general retrial times*, European Journal of Operational Research **76** (1994), 552-562.

Srinivas R. Chakravarthy received B.Sc. in Mathematics, from the University of Madras, India; M.Sc. in Statistics from the University of Madras, India; and Ph.D. in Operations Research, from the University of Delaware, Newark, US. Since 1983, he has been in the Departments of Industrial and Manufacturing Engineering & Mathematics, Kettering University , Flint, Michigan, USA. His research interests include queueing, inventory, reliability, matrix-analytical methods, algorithmic probability, and simulation.

Departments of Industrial and Manufacturing Engineering & Mathematics, Kettering University , Flint, Michigan, USA.
e-mail: schakrav@kettering.edu