An $M/G/1$ Retrial Queue With Unreliable Server for Streaming Multimedia Applications

Nathan P. Sherman[1]
Department of Operational Sciences
Air Force Institute of Technology
2950 Hobson Way (AFIT/ENS)
Wright Patterson AFB, OH 45433-7765 USA
Email: Nathan.Sherman@pentagon.af.mil

Jeffrey P. Kharoufeh[2]
Department of Industrial Engineering
University of Pittsburgh
1048 Benedum Hall
Pittsburgh, PA 15261 USA
Email: jkharouf@pitt.edu

Mark A. Abramson[3]
Department of Mathematics and Statistics
Air Force Institute of Technology
2950 Hobson Way (AFIT/ENC)
Wright Patterson AFB, OH 45433-7765 USA
Email: Mark.A.Abramson@boeing.com

Final version appears in
Probability in the Engineering and Informational Sciences, 23 (2009), 281-304.

Abstract

As a model for streaming multimedia applications, we study an unreliable retrial queue with infinite-capacity orbit and normal queue for which the retrial rate and the server repair rate are controllable. Customers join the retrial orbit if and only if their service is interrupted by a server failure. Interrupted customers do not rejoin the normal queue but repeatedly attempt to access the server at i.i.d. intervals until it is found functioning and idle. We provide stability conditions, queue length distributions, stochastic decomposition results, and performance measures. The joint optimization of the retrial and server repair rates is also studied.

[1]Now at Directorate of Force Management Policy; U.S. Air Force Headquarters, Manpower and Personnel.
[2]Corresponding author. Phone: (412) 624-9832.
[3]Now at The Boeing Company, Seattle, Washington.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
An M/G/1 Retrial Queue With Unreliable Server for Streaming Multimedia Applications	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Institute of Technology (AFIT/ENS),Department of Operational Sciences,2950 Hobson Way,Wright Patterson AFB,OH,45433-7765**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

As a model for streaming multimedia applications, we study an unreliable retrial queue with in?nite-capacity orbit and normal queue for which the retrial rate and the server repair rate are controllable. Customers join the retrial orbit if and only if their service is interrupted by a server failure. Interrupted customers do not rejoin the normal queue but repeatedly attempt to access the server at i.i.d. intervals until it is found functioning and idle. We provide stability conditions queue length distributions, stochastic decomposition results, and performance measures. The joint optimization of the retrial and server repair rates is also studied.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**25**	

1 Introduction

In this article, we analyze an M/G/1 retrial queue with an unreliable server whose orbit and normal queue have infinite storage capacity and whose retrial and server repair rates are controllable. Customers in service join the retrial orbit if and only if they are interrupted by a server breakdown and do not rejoin the normal queue, but rather attempt to access the server directly at random intervals independently of arrivals or other retrial customers. However, these interrupted customers can regain access to the server only when it is operational and idle and repeat service until they have been completed. Arriving customers who find a failed server join the normal queue. We allow for both *active* breakdowns which occur during a service cycle, and *idle* breakdowns which occur while the server is not failed but idle. The server may not breakdown while under repair. The times between customer arrivals, breakdowns, retrials, and repairs are assumed to be exponentially distributed while the service times are general.

Over the past two decades, advances in telecommunications and computer networking technologies have reinvigorated the study of queueing systems in general and retrial queueing systems in particular. The model we present here is well-suited to model computer network streaming multimedia applications. The primary (or normal) queue is similar to a 1-persistent carrier-sense multiple-access (CSMA) system. When the oldest packet in the normal queue detects that the transmission medium (or server) is free, transmission begins immediately. If the communication medium fails during transmission, the packet is sent to a retrial queue which is analogous to a non-persistent CSMA system. If the medium is unavailable (i.e., busy or failed), then the retrial packet waits a random amount of time before checking the status of the medium again. This process repeats until the retrial packet finds the transmission medium operational and idle. In this sense, the model is a priority queue wherein the packets that are not interrupted by a transmission failure have non-preemptive priority over those awaiting availability of the transmission medium. An important application is that of streaming voice or video wherein transmitted packets are used for playback upon reception and also stored for future use. The packets used for immediate playback are time-sensitive in that, if they are not received within a given time threshold, they are effectively

useless. These packets correspond to the priority customers. Packets that are interrupted during transmission can still be used for later playback from the stored copy of the stream, but their transmission time is no longer important. These packets constitute customers in the orbit. By developing analytical expressions for congestion and delay measures in stable systems, it is possible to simultaneously select a packet retrial rate and server repair rate that minimize the long-run average cost of holding customers in either queue.

In addition to its practical relevance, the model we present also exhibits extremely interesting mathematical properties that warrant investigation in their own right. The presence of an infinite waiting space for primary customers introduces an interaction between the two infinite queues. This dynamic does not exist in the vast majority of retrial queueing models which include only an infinite retrial orbit and do not consider an infinite waiting space for primary arrivals. We will show that the steady state orbit size and the overall system size both possess a stochastic decomposition property. Moreover, an interesting stability result emerges, namely that the normal queue may remain stable even if the condition for system stability is violated. Using the steady state distributions and corresponding queueing performance measures, we illustrate the means by which to simultaneously select the optimal retrial rate and repair capacity to minimize a long-run average cost criterion.

The literature addressing retrial queues with unreliable servers is relatively sparse but growing at a rapid pace. The seminal papers in this area are [1] and [15]. All models considering retrial queues with server breakdowns assume an M/G/1/1 loss system with the exception of [6] and [21]. Although [10] considered an M/G/1 retrial queue with infinite-capacity orbit and normal queue, the authors did not consider an unreliable server. For retrial models with no waiting room and server breakdowns, customers arriving to find the server unavailable (busy or failed) join the orbit. Some models (cf. [2], [3], [7], [16], [20], [23], [25], [27]) force these customers into the orbit while others ([4], [5], [6], [11], [15], [26]) provide the option of joining the orbit or departing the system. With the exception of two cases ([3] and [25]), these models also either force, or provide the option for, in-service customers interrupted by a server failure to join the orbit. Our model differs from

these in that arriving customers who find the server busy or failed join the normal queue whereas interrupted customers always join the orbit and attempt to re-access the server at random intervals. A variety of failure types are considered in the literature including starting failures ([16], [20], [27]), vacations ([7], [23]), active breakdowns ([6], [25], [26]), and like our model, both active and idle breakdowns ([2], [3], [4], [5], [11], [15]). Most orbits are assumed to behave as infinite-server queues with identical exponential service times; however some models (cf. [7], [16], and [26]) consider orbits as FCFS queues.

For retrial systems with no breakdowns and zero capacity in the normal queue, the most common optimal control strategies include the optimal routing of arriving customers ([9], [10], [22]) and selection of the optimal retrial rate ([8], [13], [14]). For general queueing systems (non-retrial queues), researchers such as [17], [19], and [24] have considered optimal N-policies wherein the server remains idle until exactly N ($N \geq 1$) customers are present in the queue. The current literature addressing the optimal design or control of unreliable retrial queues is very sparse. It appears that only [7] formally addressed these issues for a retrial queue with vacations. In that work, the author presents an optimal N-policy, an optimal T-policy and he computes the optimal retrial rate that minimizes costs using an N-policy. An informal, graphical approach to the optimal control and design of a retrial queue with vacations was presented in [20] wherein the authors examined the impact of the retrial rate, the number of input sources, the arrival rate, and the service rate on the mean waiting time and throughput.

As a model for streaming multimedia applications, this paper is concerned with the analysis and control of an M/G/1 retrial queue with an infinite-capacity orbit and normal queue. In particular, we consider the problem of simultaneously selecting an optimal retrial rate and optimal repair rate with the objective of minimizing the long-run average operating cost which is a function of the key queueing performance measures. To this end, using the method of supplementary variables and a classical generating function approach, we derive the steady state joint distribution of the orbit size and normal queue size when the server is *idle* (operational and not occupied), *failed* (non-operational and being repaired), or *busy* (operational and occupied). Using these results, we

obtain the joint generating function of the orbit size and normal queue size as well as the generating function for the overall system size (the total number of customers in orbit, normal queue and in service), independent of the server's status. We provide a necessary and sufficient condition for stability of the orbit and system as well as a (distinct) condition for stability of the normal queue. Moreover, we show that the steady state length of the retrial queue and the system size can be stochastically decomposed.

The remainder of the paper is organized as follows. Section 2 provides the model description and mathematical notation. In section 3 we establish stability conditions and, by means of generating functions, derive the queue length distributions, key queueing performance measures, as well as the steady state distribution of the server's status. In section 4 we present stochastic decomposability results for queue length distributions. Finally, section 5 presents and illustrates a nonlinear optimization problem for the optimal selection of the retrial and server repair rates.

2 Model Description

Customers arrive to the system according to a homogeneous Poisson process with rate $\lambda > 0$. Service times form an independent and identically distributed (i.i.d.) sequence of random variables with absolutely continuous distribution function (d.f.) B, probability density function (p.d.f.) b, and service completion rate

$$\mu(x) = \frac{b(x)}{1 - B(x)}, \qquad x \geq 0.$$

For $s \geq 0$, let

$$b^*(s) = \int_0^\infty e^{-sx} b(x) dx$$

denote the Laplace transform of b. Server failures occur according to a Poisson process with rate $\xi > 0$ when the server is not being repaired. The repair time is exponentially distributed with rate parameter $\alpha > 0$. An in-service customer interrupted by a server failure enters the orbit and spends an exponential amount of time there with rate $\theta > 0$, after which it either enters service (if possible) or remains in the orbit for an additional exponentially distributed time with rate θ. The arrival, service, failure, repair, and retrial processes are assumed to be mutually independent.

Denote by Q_t the number of customers in the normal queue at time t, excluding any customer that might be in service, and let R_t denote the number of customers in the orbit at time t. The random variable U_t is the occupation status of the server given by

$$U_t = \begin{cases} 1, & \text{if the server is occupied at time } t \\ 0, & \text{if the server is not occupied at time } t \end{cases}$$

while S_t describes the operational status of the server at time t defined by

$$S_t = \begin{cases} 1, & \text{if the server is not failed at time } t \\ 0, & \text{if the server is failed at time } t \end{cases} .$$

Let X_t denote the elapsed service time of the customer in service at time t so that the continuous-time stochastic process, $\{(Q_t, U_t, R_t, S_t, X_t) : t \geq 0\}$ describes the state of the system. Let N_t denote the total number of customers in the system at time t (i.e., in orbit, normal queue, and in service). Assume that as $t \to \infty$, $Q_t \Rightarrow Q$, $R_t \Rightarrow R$, $S_t \Rightarrow S$ and $N_t \Rightarrow N$ where "$\Rightarrow$" denotes weak convergence.

Now define

$$\begin{aligned} \pi_{0,0,j,1} &= \lim_{t\to\infty} P(Q_t = 0, U_t = 0, R_t = j, S_t = 1), \quad j \geq 0 \\ \pi_{k,0,j,0} &= \lim_{t\to\infty} P(Q_t = k, U_t = 0, R_t = j, S_t = 0), \quad j,k \geq 0 \\ \pi_{k,1,j,1}(x) &= \lim_{t\to\infty} P(Q_t = k, U_t = 1, R_t = j, S_t = 1, X_t < x), \quad j,k \geq 0 \end{aligned}$$

as the limiting probabilities that the system is in an idle, failed, or busy state, respectively. With the transform variables z_1 and z_2 corresponding to the orbit size and normal queue size, define

$$\begin{aligned} \phi_{0,0,1}(z_1) &= \sum_{j=0}^{\infty} z_1^j \pi_{0,0,j,1}, \\ \phi_{k,0,0}(z_1) &= \sum_{j=0}^{\infty} z_1^j \pi_{k,0,j,0}, \\ \phi_{k,1,1}(x, z_1) &= \sum_{j=0}^{\infty} z_1^j \pi_{k,1,j,1}(x). \end{aligned}$$

These are, respectively, the generating functions for $\pi_{0,0,j,1}$, $\pi_{k,0,j,0}$, and $\pi_{k,1,j,1}(x)$ with respect to

the orbit size. Further define, respectively,

$$\psi_{0,0}(z_1, z_2) = \sum_{k=0}^{\infty} z_2^k \phi_{k,0,0}(z_1),$$
$$\psi_{1,1}(x, z_1, z_2) = \sum_{k=0}^{\infty} z_2^k \phi_{k,1,1}(x, z_1),$$

the generating functions for $\phi_{k,0,0}(z_1)$ and $\phi_{k,1,1}(x, z_1)$ with respect to the normal queue size. The joint p.g.f. of the orbit and normal queue size when the server is not failed and busy, is given by

$$\psi_{1,1}(z_1, z_2) = \int_0^{\infty} \psi_{1,1}(x, z_1, z_2) dx.$$

Let p denote the joint probability mass function (p.m.f.) of R and Q while q denotes the p.m.f. of N. By summing over the three distinct and exhaustive server states, we denote by

$$G(z_1, z_2) = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} p(j, k) z_1^j z_2^k = \phi_{0,0,1}(z_1) + \psi_{0,0}(z_1, z_2) + \psi_{1,1}(z_1, z_2),$$

the joint generating function for the orbit and normal queue size. In a similar manner, we denote by

$$H(z) = \sum_{j=0}^{\infty} q(j) z^j = \phi_{0,0,1}(z) + \psi_{0,0}(z, z) + z\psi_{1,1}(z, z),$$

the generating function for the overall system size.

In the next section, we provide stability conditions and formally derive the generating functions defined in this section. Subsequently, we use these to characterize queue length distributions and performance measures.

3 Stability Analysis and Steady State Equations

In this section, we provide a necessary and sufficient condition for stability of the overall queueing system and derive the steady state joint distribution of the orbit and normal queue size when the server is idle, failed, or busy, respectively. Subsequently, we obtain the joint distribution of the orbit size and normal queue size, and the distribution of the system size, independent of the server's status. Additionally, we obtain standard queueing performance measures as well as the limiting distribution of the server's status.

Before proceeding to the main result, we first provide a lemma that is needed to characterize the stability conditions and steady state distributions. As in Aissani and Artalejo [4], define the fundamental server period as the time from which a service cycle begins until the next time at which the server is able to initiate a new service cycle. Denote this random duration by $\mathcal{T}$. Let N_r and N_q respectively denote the number of customers entering the orbit and normal queue during $(0, \mathcal{T}]$, and let $a(i,j) = P(N_r = i, N_q = j)$, $i, j \geq 0$. Define the generating function

$$Q(z_1, z_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a(i,j) z_1^i z_2^j, \qquad |z_1| \leq 1, |z_2| \leq 1.$$

Then one can verify (see [4]) that

$$Q(z_1, z_2) = b^*(\xi + \lambda(1 - z_2)) + \frac{\alpha z_1 \xi (1 - b^*(\xi + \lambda(1 - z_2))}{(\alpha + \lambda(1 - z_2))(\xi + \lambda(1 - z_2))}.$$

Now let us define the quantity

$$\rho_1 = -\frac{d}{d\varepsilon} Q(1, 1 - \varepsilon)\Big|_{\varepsilon = 0} = \frac{\lambda(1 - b^*(\xi))(\alpha + \xi)}{\alpha \xi},$$

where $b^*(\xi)$ is the Laplace transform of the service time p.d.f. evaluated at ξ. Using these definitions we have the following important result which is needed to obtain our main results.

Lemma 1 *For either $|z_1| < 1$ or $|z_1| \leq 1$ and $\rho_1 > 1$ the relation*

$$z_2 - Q(z_1, z_2)$$

has, as a function of z_2, one and only one zero, $g(z_1)$, inside the region $|z_2| < 1$. In case $z_1 = 1$, $g(1)$ is the smallest positive real zero with $g(1) < 1$ if $\rho_1 > 1$, and $g(1) = 1$ if $\rho_1 \leq 1$.

Proof. The proof is similar to that of Theorem 3 in [18, p. 351-352]. Now for the first part, on $|z_1| < 1$, applying Rouche's theorem to the function z_2 and the generating function $Q(z_1, z_2)$, we conclude that there is one and only one zero, $g(z_1)$, for each z_2 inside the unit disk $|z_2| < 1$. For the second part, consider the quantity ρ_1 and the case when $z_1 = 1$. The function $Q(z_1, z_2)$ is monotonically increasing in z_2 for $z_2 \in [0, 1]$ such that $0 < Q(1, 0) < 1$ and $Q(1, 1) = 1$. Thus, if $z_1 = 1$ and $\rho_1 > 1$, then $g(1)$ is the minimal, positive real zero with $g(1) < 1$. On the other hand, if $\rho_1 \leq 1$, $g(1) = 1$ is the unique zero, and this completes the proof. ∎

Using Lemma 1, we now characterize the stability condition for the overall system size (and orbit size), as well as the generating functions $\phi_{0,0,1}(z_1)$, $\psi_{0,0}(z_1, z_2)$, and $\psi_{1,1}(z_1, z_2)$. Theorem 1 states that the long-run proportion of time the server is available for serving customers must exceed the long-run proportion of time the server is busy if the system is to remain stable.

Theorem 1 *The queueing system is stable if and only if $\rho < 1$ where*

$$\rho = \frac{\lambda(1 - b^*(\xi))(\alpha + \xi)}{\alpha b^*(\xi)\xi}. \tag{1}$$

In such a case, the generating functions $\phi_{0,0,1}(z_1)$, $\psi_{0,0}(z_1, z_2)$, and $\psi_{1,1}(z_1, z_2)$ are, respectively, given by

$$\phi_{0,0,1}(z_1) = \frac{\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)}{\xi b^*(\xi)(\alpha + \xi)} \exp\left\{-\frac{1}{\theta}\int_{z_1}^{1} \frac{\lambda(1 - g(u)) + \xi(1 - \frac{\alpha}{\alpha + \lambda(1 - g(u))})}{g(u) - u} du\right\}, \tag{2}$$

$$\begin{aligned}\psi_{0,0}(z_1, z_2) = \Bigg\{ & \frac{(g(z_1) - z_1)[\alpha + \lambda(1 - g(z_1))][\xi + \lambda(1 - z_2)][z_2 - \hat{B}(z_2) - z_1(1 - \hat{B}(z_2))]}{(z_2 - \hat{B}(z_2))\,[\alpha + \lambda(1 - z_2)]\,[\xi + \lambda(1 - z_2)] - \alpha\xi(1 - \hat{B}(z_2))z_1} \\ & + \frac{\lambda z_1(1 - \hat{B}(z_2))(z_2 - z_1)(1 - g(z_1))[\alpha + \xi + \lambda(1 - g(z_1))]}{(z_2 - \hat{B}(z_2))\,[\alpha + \lambda(1 - z_2)]\,[\xi + \lambda(1 - z_2)] - \alpha\xi(1 - \hat{B}(z_2))z_1} \Bigg\} \frac{\xi\phi_{0,0,1}(z_1)}{(g(z_1) - z_1)[\alpha + \lambda(1 - g(z_1))]}\end{aligned} \tag{3}$$

and

$$\begin{aligned}\psi_{1,1}(z_1, z_2) = \Bigg\{ & \frac{(z_2 - z_1)(1 - g(z_1))[\alpha + \xi + \lambda(1 - g(z_1))][\alpha + \lambda(1 - z_2)]}{(z_2 - \hat{B}(z_2))\,[\alpha + \lambda(1 - z_2)]\,[\xi + \lambda(1 - z_2)] - \alpha\xi(1 - \hat{B}(z_2))z_1} \\ & - \frac{(1 - z_2)(g(z_1) - z_1)[\alpha + \lambda(1 - g(z_1))][\alpha + \xi + \lambda(1 - z_2)]}{(z_2 - \hat{B}(z_2))\,[\alpha + \lambda(1 - z_2)]\,[\xi + \lambda(1 - z_2)] - \alpha\xi(1 - \hat{B}(z_2))z_1} \Bigg\} \frac{\lambda(1 - \hat{B}(z_2))\phi_{0,0,1}(z_1)}{(g(z_1) - z_1)[\alpha + \lambda(1 - g(z_1))]}\end{aligned} \tag{4}$$

where

$$\hat{B}(z_2) = b^*(\xi + \lambda(1 - z_2))$$

and, for $z_1 \in [0, 1]$, $g(z_1)$ verifies

$$g(z_1) = b^*(\xi + \lambda(1 - g(z_1))) + \frac{\alpha\xi z_1[1 - b^*(\xi + \lambda(1 - g(z_1)))]}{[\alpha + \lambda(1 - g(z_1))][\xi + \lambda(1 - g(z_1))]}.$$

Proof. For $j \geq 0$ and $k \geq 1$ with $\pi_{k,i,-1,l}(x) = 0$, the balance equations are

$$(\alpha + \lambda)\pi_{0,0,j,0} = \xi\pi_{0,0,j,1} + \xi\int_0^\infty \pi_{0,1,j-1,1}(x)dx, \tag{5}$$

$$(\alpha + \lambda)\pi_{k,0,j,0} = \lambda\pi_{k-1,0,j,0} + \xi\int_0^\infty \pi_{k,1,j-1,1}(x)dx, \tag{6}$$

$$(\lambda + \xi + j\theta)\pi_{0,0,j,1} = \alpha\pi_{0,0,j,0} + \int_0^\infty \mu(x)\pi_{0,1,j,1}(x)dx, \tag{7}$$

$$\frac{d}{dx}\pi_{0,1,j,1}(x) = -[\mu(x) + \lambda + \xi]\pi_{0,1,j,1}(x), \tag{8}$$

$$\frac{d}{dx}\pi_{k,1,j,1}(x) = -[\mu(x) + \lambda + \xi]\pi_{k,1,j,1}(x) + \lambda\pi_{k-1,1,j,1}(x) \tag{9}$$

with boundary conditions

$$\pi_{0,1,j,1}(0) = \alpha\pi_{1,0,j,0} + \lambda\pi_{0,0,j,1} + \int_0^\infty \mu(x)\pi_{1,1,j,1}(x)dx + (j+1)\theta\pi_{0,0,j+1,1}, \quad (10)$$

$$\pi_{k,1,j,1}(0) = \alpha\pi_{k+1,0,j,0} + \int_0^\infty \mu(x)\pi_{k+1,1,j,1}(x)dx. \quad (11)$$

Multiplying both sides of Eqs. (5) through (11) by z_1^j and summing over all j, we obtain, respectively, the following equations:

$$(\alpha+\lambda)\phi_{0,0,0}(z_1) = \xi\phi_{0,0,1}(z_1) + \xi z_1 \int_0^\infty \phi_{0,1,1}(x,z_1)dx, \quad (12)$$

$$(\alpha+\lambda)\phi_{k,0,0}(z_1) = \lambda\phi_{k-1,0,0}(z_1) + \xi z_1 \int_0^\infty \phi_{k,1,1}(x,z_1)dx, \quad (13)$$

$$(\lambda+\xi)\phi_{0,0,1}(z_1) + \theta z_1 \frac{d}{dz_1}\phi_{0,0,1}(z_1) = \alpha\phi_{0,0,0}(z_1) + \int_0^\infty \mu(x)\phi_{0,1,1}(x,z_1)dx, \quad (14)$$

$$\frac{\partial}{\partial x}\phi_{0,1,1}(x,z_1) = -[\mu(x)+\lambda+\xi]\phi_{0,1,1}(x,z_1), \quad (15)$$

$$\frac{\partial}{\partial x}\phi_{k,1,1}(x,z_1) = -[\mu(x)+\lambda+\xi]\phi_{k,1,1}(x,z_1) + \lambda\phi_{k-1,1,1}(x,z_1), \quad (16)$$

with boundary conditions

$$\phi_{0,1,1}(0,z_1) = \alpha\phi_{1,0,0}(z_1) + \lambda\phi_{0,0,1}(z_1) + \int_0^\infty \mu(x)\phi_{1,1,1}(x,z_1)dx + \theta\frac{d}{dz_1}\phi_{0,0,1}(z_1), \quad (17)$$

$$\phi_{k,1,1}(0,z_1) = \alpha\phi_{k+1,0,0}(z_1) + \int_0^\infty \mu(x)\phi_{k+1,1,1}(x,z_1)dx. \quad (18)$$

Multiplying both sides of Eq. (12) by z_2^0 and Eq. (13) by z_2^k and summing over all $k \geq 0$, we obtain

$$[\alpha+\lambda(1-z_2)]\psi_{0,0}(z_1,z_2) = \xi\phi_{0,0,1}(z_1) + \xi z_1 \int_0^\infty \psi_{1,1}(x,z_1,z_2)dx. \quad (19)$$

Performing a similar operation on Eqs. (15) and (16) as well as (17) and (18), we obtain, respectively,

$$\frac{\partial}{\partial x}\psi_{1,1}(x,z_1,z_2) = -[\mu(x)+\xi+\lambda(1-z_2)]\psi_{1,1}(x,z_1,z_2) \quad (20)$$

and

$$\psi_{1,1}(0,z_1,z_2) = \lambda\phi_{0,0,1}(z_1) + \theta\frac{d}{dz_1}\phi_{0,0,1}(z_1) + \frac{\alpha}{z_2}[\psi_{0,0}(z_1,z_2) - \phi_{0,0,0}(z_1)] + \frac{1}{z_2}\int_0^\infty \mu(x)[\psi_{1,1}(x,z_1,z_2) - \phi_{0,1,1}(x,z_1)]dx. \quad (21)$$

Solving Eq. (20), we obtain

$$\psi_{1,1}(x, z_1, z_2) = \psi_{1,1}(0, z_1, z_2)e^{-[\xi+\lambda(1-z_2)]x}(1 - B(x)). \tag{22}$$

Using Eqs. (14), (19), and (22) in (21) we obtain

$$\psi_{1,1}(0, z_1, z_2) = \frac{\theta(z_2 - z_1)\frac{d}{dz_1}\phi_{0,0,1}(z_1) - [\lambda(1 - z_2) + \xi(1 - \frac{\alpha}{\alpha+\lambda(1-z_2)})]\phi_{0,0,1}(z_1)}{z_2 - \left[b^*(\xi + \lambda(1 - z_2)) + \frac{\alpha z_1 \xi(1-b^*(\xi+\lambda(1-z_2))}{(\alpha+\lambda(1-z_2))(\xi+\lambda(1-z_2))}\right]}. \tag{23}$$

Now by Lemma 1, the denominator of Eq. (23) has, for any z_1 in the unit disk, a zero in the region $|z_2| < 1$. This must also be a zero for the numerator; therefore, we have from (23)

$$\theta(z_1 - g(z_1))\frac{d}{dz_1}\phi_{0,0,1}(z_1) + \left[\ \lambda(1 - g(z_1)) + \xi\left(1 - \frac{\alpha}{\alpha + \lambda(1 - g(z_1))}\right)\right]\phi_{0,0,1}(z_1) = 0. \tag{24}$$

In order to solve the differential equation (24), we first examine the function $k(z_1) = z_1 - g(z_1) = z_1 - Q(z_1, g(z_1))$. Note that

$$\left.\frac{d}{dz_1}g(z_1)\right|_{z_1=1} = \left.\frac{d}{dz_1}Q(z_1, g(z_1))\right|_{z_1=1} = \frac{\alpha\xi(1 - b^*(\xi))}{\alpha\xi - \lambda(1 - b^*(\xi))(\alpha + \xi)} = \frac{\frac{1}{b^*(\xi)} - 1}{\frac{1}{b^*(\xi)} - \rho},$$

where ρ is defined in Eq. (1). We then observe that, for $\rho \leq 1$, the quantity $k(z_1)$ never becomes zero in $|z_1| < 1$, while for $\rho > 1$ this quantity has one and only one zero, call it β, such that $\beta \in (0, 1)$ (see for example [12] or [18]). Now rearranging Eq. (24), define the function

$$h(z_1) = \frac{\lambda(1 - g(z_1)) + \xi\left(1 - \frac{\alpha}{\alpha+\lambda(1-g(z_1))}\right)}{z_1 - g(z_1)},$$

and note that, for $\rho < 1$,

$$\lim_{z_1\to 1} h(z_1) = -\frac{\lambda\xi(1 - b^*(\xi))(\alpha + \xi)}{\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)} = -\frac{\xi\rho}{1 - \rho} < \infty.$$

Thus, we conclude that $h(z_1)$ is analytic on the open disk $|z_1| < 1$, and $h(z_1)$ can be defined at the point $z_1 = 1$. Therefore, on the closed disk $|z_1| \leq 1$, the differential equation,

$$\frac{d}{dz_1}\phi_{0,0,1}(z_1) + h(z_1)\phi_{0,0,1}(z_1) = 0,$$

is verified by the function

$$\phi_{0,0,1}(z_1) = K \exp\left\{-\frac{1}{\theta}\int_{z_1}^{1} \frac{\lambda(1 - g(u)) + \xi\left(1 - \frac{\alpha}{\alpha+\lambda(1-g(u))}\right)}{g(u) - u}du\right\} \tag{25}$$

where K is a constant of integration. For $\rho < 1$, $\phi_{0,0,1}(z_1)$ in Eq. (25) makes all generating functions analytic in $|z_1| \leq 1, |z_2| \leq 1$. In particular, (2), (3), and (4) are obtained up to the multiplicative constant K. To obtain this constant, we apply the normalization condition, $\phi_{0,0,1}(1) + \psi_{0,0}(1,1) + \psi_{1,1}(1,1) = 1$, which yields

$$K = \frac{\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)}{\xi b^*(\xi)(\alpha + \xi)} = \phi_{0,0,1}(1),$$

the steady state probability that the server is not failed and idle.

To see that $\rho < 1$ is also necessary for system stability, assume that $\rho > 1$ and the system is stable. By Lemma 1, when $\rho > 1$, the function $z_1 - Q(z_1, g(z_1))$ has one and only one zero, denoted by β, with $0 < \beta < 1$. Thus, the function $h(z_1)$ is not analytic in $|z_1| < 1$. Substituting $z_1 = \beta$ in Eq. (24) yields

$$\left[\lambda(1 - \beta) + \xi\left(1 - \frac{\alpha}{\alpha + \lambda(1 - \beta)}\right)\right]\phi_{0,0,1}(\beta) = 0,$$

which implies that $\phi_{0,0,1}(\beta) = 0$ since the coefficient of $\phi_{0,0,1}(\beta)$ does not equal zero. We must therefore conclude that

$$\phi_{0,0,1}(\beta) = \sum_{j=0}^{\infty} \beta^j \pi_{0,0,j,1} = 0.$$

However, it is not possible to find positive values $\pi_{0,0,j,1}$, $j \geq 0$, to satisfy the above relation, so the system is not stable; thus, a contradiction.

Finally, suppose that $\rho = 1$ and the system is stable. When $\rho = 1$, the function $z_1 - Q(z_1, g(z_1))$ never becomes zero in the unit disk $|z_1| < 1$; however, in this case

$$\lim_{z_1 \to 1} h(z_1) = -\frac{\xi\rho}{1 - \rho} = -\infty.$$

Differentiating Eq. (24) and evaluating at the point $z_1 = 1$, we obtain the relation

$$-\lambda\left(\frac{\alpha + \xi}{\alpha}\right)\frac{d}{dz_1}g(z_1)\Big|_{z_1=1}\phi_{0,0,1}(1) = 0,$$

implying that $\phi_{0,0,1}(1) = 0$ since $\frac{d}{dz_1}g(z_1)\big|_{z_1=1} = 1$ when $\rho = 1$. However, this contradicts the hypothesis that the system is stable. Therefore, we conclude that the system cannot be stable unless $\rho < 1$. ■

Using Theorem 1, we next obtain the joint distribution of the orbit and normal queue size, as well as the distribution of the overall system size, independent of server status.

Corollary 1 *For $\rho < 1$, the probability generating functions $G(z_1, z_2)$ and $H(z)$ are given by*

$$G(z_1,z_2) = \left\{ \frac{\lambda(1-\hat{B}(z_2))[\alpha+\xi z_1+\lambda(1-z_2)](z_2-z_1)(1-g(z_1))[\alpha+\xi+\lambda(1-g(z_1))]}{[g(z_1)-z_1][\alpha+\lambda(1-g(z_1))]\left\{(z_2-\hat{B}(z_2))[\alpha+\lambda(1-z_2)][\xi+\lambda(1-z_2)]-\alpha\xi z_1(1-\hat{B}(z_2))\right\}} \right.$$
$$\left. -\frac{\lambda(1-\hat{B}(z_2))[\alpha+\xi z_1+\lambda(1-z_2)](1-z_2)[\alpha+\xi+\lambda(1-z_2)]}{[\alpha+\lambda(1-z_2)]\left\{(z_2-\hat{B}(z_2))[\alpha+\lambda(1-z_2)][\xi+\lambda(1-z_2)]-\alpha\xi z_1(1-\hat{B}(z_2))\right\}} + \frac{[\alpha+\xi+\lambda(1-z_2)]}{[\alpha+\lambda(1-z_2)]} \right\}$$
$$\times \phi_{0,0,1}(z_1) \quad (26)$$

and

$$H(z) = \frac{[\alpha+\xi+\lambda(1-z)]\left\{\alpha\xi\hat{B}(z)+\lambda\hat{B}(z)(1-z)[\alpha+\xi+\lambda(1-z)]\right\}}{[\alpha+\lambda(1-z)]\left\{\alpha\xi\hat{B}(z)-\lambda(z-\hat{B}(z))[\alpha+\xi+\lambda(1-z)]\right\}}\phi_{0,0,1}(z). \quad (27)$$

Using standard methods, Eqs. (26) and (27) can be used to obtain the mth moment ($m \geq 1$) of R, Q, and N, respectively, as well as their probability distributions. The first moments are provided in the following corollary.

Corollary 2 *The steady state mean orbit size, mean normal queue size, and mean number in system are respectively given by*

$$\mathbb{E}(R) = \frac{\rho}{1-\rho}\left[\frac{\alpha}{\alpha+\xi}\cdot\frac{\xi b^*(\xi)[\xi-\lambda(1-b^*(\xi))]+(\alpha+\xi)[\lambda(1-b^*(\xi))-\xi\hat{B}']}{b^*(\xi)[\alpha\xi-\lambda(1-b^*(\xi))(\alpha+\xi)]}+\frac{\xi}{\theta}\right] \quad (28)$$

$$\mathbb{E}(Q) = \lambda\frac{\xi^3 b^*(\xi)-(\alpha+\xi)^2[\xi\hat{B}'-\lambda(1-b^*(\xi))]}{\xi b^*(\xi)(\alpha+\xi)\left[\alpha\xi-\lambda(1-b^*(\xi))(\alpha+\xi)\right]}, \quad (29)$$

and

$$\mathbb{E}(N) = \frac{\lambda b^*(\xi)\left\{\xi^3+(1-b^*(\xi))[\alpha\xi(\alpha+2\xi)+\lambda(\alpha+\xi)^2]\right\}-\lambda\xi(\alpha+\xi)^2\hat{B}'}{\xi b^*(\xi)(\alpha+\xi)[\alpha\xi b^*(\xi)-\lambda(1-b^*(\xi))(\alpha+\xi)]}+\frac{\xi\rho}{\theta(1-\rho)}, \quad (30)$$

where

$$\hat{B}' = \frac{d}{dz_2}\hat{B}(z_2)\bigg|_{z_2=1} = \lambda\int_0^\infty xe^{-\xi x}b(x)dx.$$

Let S denote the duration of an arbitrary service time with c.d.f. B. Then as $\xi \to 0$ in Eqs. (28) to (30), it can be shown that

$$\mathbb{E}(R) = 0,$$

$$\mathbb{E}(Q) = \frac{\lambda^2 \mathbb{E}(S^2)}{2(1 - \lambda \mathbb{E}(S))},$$

and

$$\mathbb{E}(N) = \lambda \mathbb{E}(S) + \frac{\lambda^2 \mathbb{E}(S^2)}{2(1 - \lambda \mathbb{E}(S))}.$$

These expressions are consistent with results for the standard M/G/1 queue with no failures and no retrials. Furthermore, denote by W_R, W_Q and W, the time spent in the orbit, normal queue and system by an arbitrary customer in the long-run, respectively. The expected values of these random variables are obtained by applying Little's Law to Eqs. (28) to (30), i.e., $\mathbb{E}(W_R) = \lambda^{-1}\mathbb{E}(R)$, $\mathbb{E}(W_Q) = \lambda^{-1}\mathbb{E}(Q)$, and $\mathbb{E}(W) = \lambda^{-1}\mathbb{E}(N)$.

Corollary 2 also confirms that $\rho < 1$ is necessary for the stability of R (and N), and by (29) we see that $\rho_1 < 1$ is necessary for the stability of Q. That is, the normal queue can be stable even if the orbit stability condition is violated. Owing to the nature of the orbit dynamics, retrial customers are subordinate to normal customers and may be served only when the server is idle and operational. Hence, normal queue customers experience a greater effective service rate than do retrial customers, and thus, it is possible that the orbit may continue to grow while the normal queue remains stable.

Finally, we characterize the steady state distribution of the server's status by directly applying the results of Theorem 1. Let p_I, p_F, and p_B respectively denote the limiting probability that the server is idle, failed, or busy.

Corollary 3 *For $\rho < 1$, the steady state distribution of the server's status is given by*

$$p_I = \lim_{z_1 \to 1} \phi_{0,0,1}(z_1) = \frac{\alpha}{\alpha + \xi} - \frac{\lambda(1 - b^*(\xi))}{\xi b^*(\xi)},$$

$$p_F = \lim_{\substack{z_1 \to 1 \\ z_2 \to 1}} \psi_{0,0}(z_1, z_2) = \frac{\xi}{\alpha + \xi},$$

and

$$p_B = \lim_{\substack{z_1 \to 1 \\ z_2 \to 1}} \psi_{1,1}(z_1, z_2) = \frac{\lambda(1 - b^*(\xi))}{\xi b^*(\xi)}.$$

In the following section we show that the orbit and system size can be stochastically decomposed before considering the optimal selection of the retrial and repair rates in section 5.

4 Stochastic Decomposition

In this section, we demonstrate that both the orbit and system size exhibit a stochastic decomposition property which has been observed for the system size distribution of many M/G/1 models including those with vacations, retrial queues, and breakdowns (cf. [4, 11, 16, 27]). Falin and Templeton [12] provide several stochastic decomposition results, including the decomposability of the vector of server status and the orbit size, in the standard M/G/1 retrial queue (i.e., one with no infinite waiting space and no server breakdowns).

Allowing $\theta \to \infty$ in our model yields a model in which retrial customers instantaneously attempt to re-access the server (i.e., an instantaneous feedback model). Let $\hat{R}$ denote the steady state orbit size in the instantaneous feedback model, and denote the generating function of $\hat{R}$ by $\mathbb{E}(z^{\hat{R}})$ for $|z| \leq 1$. Let $\hat{N}$ denote the steady state total number of customers in the system in the instantaneous feedback model and denote its generating function by $\mathbb{E}(z^{\hat{N}})$ for $|z| \leq 1$. Finally, let V be a random variable whose generating function is given by

$$\mathbb{E}(z^V) = \exp\left\{-\frac{1}{\theta}\int_{z_1}^{1} \frac{\lambda(1-g(u)) + \xi(1 - \frac{\alpha}{\alpha+\lambda(1-g(u))})}{g(u)-u} du\right\}, \quad |z| \leq 1. \tag{31}$$

The following two propositions describe the decomposability of the orbit and system size distributions.

Proposition 1 *The random variable R may be expressed as the sum of two independent random variables, one of which is the steady state orbit size in the instantaneous feedback model and the other is V, i.e.,*

$$R = \hat{R} + V. \tag{32}$$

Proof. Note that Eqs. (2), (3), (4), (26), and (27) depend on the retrial rate θ only through the generating function $\mathbb{E}(z_1^V)$. Therefore, we may write the generating function for $\hat{R}$ as

$$\mathbb{E}(z_1^{\hat{R}}) = \lim_{\theta\to\infty} G(z_1, 1) = A_G(z_1)$$

where, using L'Hospital's rule, it can be shown that

$$A_G(z_1) = (1-\rho)\left[1 + \frac{\lambda(\alpha+\xi z_1)(1-g(z_1))[\alpha+\xi+\lambda(1-g(z_1))]}{\xi(\alpha+\xi)(g(z_1)-z_1)[\alpha+\lambda(1-g(z_1))]}\right].$$

Now since $\mathbb{E}(z_1^R) \equiv G(z_1, 1)$, we may write

$$\begin{aligned}\mathbb{E}(z_1^R) &= A_G(z_1) \exp\left\{-\frac{1}{\theta}\int_{z_1}^1 \frac{\lambda(1-g(u)) + \xi(1-\frac{\alpha}{\alpha+\lambda(1-g(u))})}{g(u)-u}du\right\} \\ &= \mathbb{E}(z_1^{\hat{R}})\mathbb{E}(z_1^V) \\ &= \mathbb{E}(z_1^{\hat{R}+V}).\end{aligned}$$

■

Similar behavior may be observed for the steady state system size as noted in Proposition 2.

Proposition 2 *The random variable N may be expressed as the sum of two independent random variables, one of which is the steady state system size in the instantaneous feedback model and the other is V, i.e.,*

$$N = \hat{N} + V. \tag{33}$$

Proof. The proof is analogous to that of Proposition 1. Note that by setting $z_1 = z_2 = z$ in Eqs. (2), (3), and (4) we obtain

$$\mathbb{E}(z^N) \equiv H(z) = \phi_{0,0,1}(z) + \psi_{0,0}(z,z) + z\psi_{1,1}(z,z).$$

The generating function of $\hat{N}$ is given by

$$\mathbb{E}(z^{\hat{N}}) = \lim_{\theta\to\infty} H(z) = A_H(z)$$

where

$$A_H(z) = \frac{\alpha(1-\rho)[\alpha+\xi+\lambda(1-z)]\left\{\alpha\xi\hat{B}(z) + \lambda\hat{B}(z)(1-z)[\alpha+\xi+\lambda(1-z)]\right\}}{(\alpha+\xi)[\alpha+\lambda(1-z)]\left\{\alpha\xi\hat{B}(z) - \lambda(z-\hat{B}(z))[\alpha+\xi+\lambda(1-z)]\right\}}.$$

The generating function for N is the product of these two; that is,

$$\begin{aligned}\mathbb{E}(z^N) &= A_H(z) \exp\left\{-\frac{1}{\theta}\int_{z_1}^1 \frac{\lambda(1-g(u)) + \xi(1-\frac{\alpha}{\alpha+\lambda(1-g(u))})}{g(u)-u}du\right\} \\ &= \mathbb{E}(z^{\hat{N}})\mathbb{E}(z^V) \\ &= \mathbb{E}(z^{\hat{N}+V}).\end{aligned}$$

■

In the next section we formulate an optimization problem for the selection of the optimal retrial and repair rates that minimize the long-run average cost of operating the queueing system, subject to a budget constraint. We also provide two illustrative examples using distinct service time distributions.

5 Optimal Retrial and Repair Rates

We now consider the simultaneous optimal selection of the retrial and repair rates that minimize the long-run average operating costs. The cost function includes the cost of service, the cost of holding customers in the normal queue, and the cost of holding customers in the orbit. Required for the optimization are the queueing performance measures $\mathbb{E}(R)$, $\mathbb{E}(Q)$, $\mathbb{E}(W_R)$, and $\mathbb{E}(W_Q)$, as well as the expected number of customers in service, $\lambda(1-b^*(\xi))/\xi b^*(\xi)$, and the expected time to complete service, $(1-b^*(\xi))/\xi b^*(\xi)$. The cost per unit time per customer in service is c_S while the holding costs per unit time per customer in the orbit and normal queue are respectively denoted by c_R and c_Q. The coefficient c_θ is the cost of one "unit" of retrial rate while c_α is the cost of one "unit" of repair rate. Using Eqs. (28) and (29), we solve the optimization problem

$$\begin{aligned}
\text{Minimize} \quad & C(\theta,\alpha) = c_S \frac{\lambda(1-b^*(\xi))^2}{\xi^2 b^*(\xi)^2} + c_R \mathbb{E}(R)\mathbb{E}(W_R) + c_Q \mathbb{E}(Q)\mathbb{E}(W_Q) \\
\text{Subject to} \quad & \lambda(1-b^*(\xi))(\alpha+\xi) - \alpha\xi b^*(\xi) \;<\; 0 \qquad (34) \\
& c_\theta\theta + c_\alpha\alpha \;\leq\; D \qquad (35) \\
& \theta,\alpha \;>\; 0 \qquad (36)
\end{aligned}$$

where D is a fixed budget ($D < \infty$). Constraint (34) enforces the stability condition ($\rho < 1$) and (35) is a budget constraint that limits the attainable repair capacity and the rate at which interrupted customers may attempt to re-access the server.

5.1 Convexity Analysis

The uniqueness of a global solution to the above optimization problem can be established by showing it is a convex program. The existence of a solution will then be shown directly.

The feasible region, defined by

$$X = \{(\theta, \alpha) : \lambda(1 - b^*(\xi))(\alpha + \xi) - \alpha\xi b^*(\xi) < 0;\ c_\theta\theta + c_\alpha\alpha \leq D;\ \theta, \alpha > 0\},$$

is a convex set since it is defined by a finite set of linear constraints. Strict convexity of the objective function, which will now be proved, will complete the uniqueness proof. Expanding the terms of $C(\theta, \alpha)$ gives

$$\begin{aligned} C(\theta,\alpha) = c_S\frac{\lambda(1-b^*(\xi))^2}{\xi^2 b^*(\xi)^2} + c_Q\lambda\left(\frac{\xi^3 b^*(\xi) - (\alpha+\xi)^2[\xi\hat{B}' - \lambda(1-b^*(\xi))]}{\xi b^*(\xi)(\alpha+\xi)\left[\alpha\xi - \lambda(1-b^*(\xi))(\alpha+\xi)\right]}\right)^2 \\ + c_R\lambda\left(\alpha\lambda(1-b^*(\xi))\frac{\xi b^*(\xi)[\xi - \lambda(1-b^*(\xi))] + (\alpha+\xi)[\lambda(1-b^*(\xi)) - \xi\hat{B}']}{b^*(\xi)[\alpha\xi - \lambda(1-b^*(\xi))(\alpha+\xi)][\alpha\xi b^*(\xi) - \lambda(1-b^*(\xi))(\alpha+\xi)]}\right. \\ \left. + \frac{\lambda\xi(\alpha+\xi)(1-b^*(\xi))}{\theta[\alpha\xi b^*(\xi) - \lambda(1-b^*(\xi))(\alpha+\xi)]}\right)^2. \quad (37) \end{aligned}$$

The first term on the right-hand side (r.h.s.) of Eq. (37) depends on neither θ nor α, and hence, does not affect the convexity of C. To prove the convexity of the other two terms, the following functions are defined:

$$f_{R_1}(\alpha) = \alpha\frac{\xi b^*(\xi)[\xi - \lambda(1-b^*(\xi))] + (\alpha+\xi)[\lambda(1-b^*(\xi)) - \xi\hat{B}']}{[\alpha\xi - \lambda(1-b^*(\xi))(\alpha+\xi)][\alpha\xi b^*(\xi) - \lambda(1-b^*(\xi))(\alpha+\xi)]},$$

$$f_{R_2}(\theta,\alpha) = \frac{\lambda\xi(\alpha+\xi)(1-b^*(\xi))}{\theta[\alpha\xi b^*(\xi) - \lambda(1-b^*(\xi))(\alpha+\xi)]},$$

$$f_R(\theta,\alpha) = \frac{\lambda(1-b^*(\xi))}{b^*(\xi)}f_{R_1}(\alpha) + f_{R_2}(\theta,\alpha),$$

and

$$f_Q(\alpha) = \frac{\xi^3 b^*(\xi) - (\alpha+\xi)^2[\xi\hat{B}' - \lambda(1-b^*(\xi))]}{(\alpha+\xi)\left[\alpha\xi - \lambda(1-b^*(\xi))(\alpha+\xi)\right]}.$$

The next two lemmas are needed to prove the strict convexity of $C(\theta, \alpha)$ on X.

Lemma 2 *The function $f_R^2(\theta, \alpha)$ is strictly convex on X.*

Proof. We first establish the positivity of the quantity

$$\lambda(1 - b^*(\xi)) - \xi\hat{B}' = \lambda\left(1 - \int_0^\infty (\xi x + 1)b(x)e^{-\xi x}dx\right).$$

Since its derivative with respect to ξ,

$$\lambda\xi \int_0^\infty x^2 b(x) e^{-\xi x} dx,$$

is strictly positive, $\lambda(1 - b^*(\xi)) - \xi\hat{B}'$ is strictly increasing for $\xi \in [0, \infty)$ and thus attains its minimum value of zero at the left endpoint $\xi = 0$. Now the second derivative of f_{R_1} with respect to α is given by

$$\begin{aligned}
\frac{\partial^2 f_{R_1}(\alpha)}{\partial \alpha^2} = 2\xi \Bigg\{ & \frac{\alpha b^*(\xi)[\xi - \lambda(1 - b^*(\xi))]^2[\xi b^*(\xi) - \lambda(1 - b^*(\xi))]}{[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]^2[\alpha\xi - \lambda(1 - b^*(\xi))(\alpha + \xi)]^2} \\
& + \frac{[\lambda(1 - b^*(\xi)) - \xi\hat{B}'] \left\{\alpha[\xi - \lambda(1 - b^*(\xi))][\xi b^*(\xi) - \lambda(1 - b^*(\xi))] + \lambda^2\xi(1 - b^*(\xi))^2\right\}}{[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]^2[\alpha\xi - \lambda(1 - b^*(\xi))(\alpha + \xi)]^2} \\
& + \frac{\lambda(1 - b^*(\xi))(\alpha + \xi)[\xi - \lambda(1 - b^*(\xi))][\lambda(1 - b^*(\xi)) - \xi\hat{B}'][\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]^2}{[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]^3[\alpha\xi - \lambda(1 - b^*(\xi))(\alpha + \xi)]^3} \\
& + \frac{\lambda\xi b^*(\xi)(1 - b^*(\xi))[\xi - \lambda(1 - b^*(\xi))]^2[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]^2}{[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]^3[\alpha\xi - \lambda(1 - b^*(\xi))(\alpha + \xi)]^3} \\
& + \frac{\lambda(1 - b^*(\xi))(\alpha + \xi)[\xi b^*(\xi) - \lambda(1 - b^*(\xi))][\lambda(1 - b^*(\xi)) - \xi\hat{B}'][\alpha\xi - \lambda(1 - b^*(\xi))(\alpha + \xi)]^2}{[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]^3[\alpha\xi - \lambda(1 - b^*(\xi))(\alpha + \xi)]^3} \\
& + \frac{\lambda\xi b^*(\xi)(1 - b^*(\xi))[\xi - \lambda(1 - b^*(\xi))][\xi b^*(\xi) - \lambda(1 - b^*(\xi))][\alpha\xi - \lambda(1 - b^*(\xi))(\alpha + \xi)]^2}{[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]^3[\alpha\xi - \lambda(1 - b^*(\xi))(\alpha + \xi)]^3} \Bigg\}.
\end{aligned}$$

It follows from the stability condition that $[\alpha\xi - \lambda(1 - b^*(\xi))(\alpha + \xi)] > 0$, $[\xi b^*(\xi) - \lambda(1 - b^*(\xi))] > 0$, and $[\xi - \lambda(1 - b^*(\xi))] > 0$. Since $\lambda(1 - b^*(\xi)) - \xi\hat{B}'$ is nonnegative,

$$\frac{\partial^2 f_{R_1}(\alpha)}{\partial \alpha^2} > 0,$$

and thus, $f_{R_1}(\alpha)$ is strictly convex for all $\alpha > 0$. Taking second partial derivatives of f_{R_2} with respect to θ and α, respectively, and assuming stability, shows that

$$\frac{\partial^2 f_{R_2}(\theta, \alpha)}{\partial \theta^2} = \frac{2\lambda\xi(1 - b^*(\xi))(\alpha + \xi)}{\theta^3[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]} > 0$$

and

$$\frac{\partial^2 f_{R_2}(\theta, \alpha)}{\partial \alpha^2} = \frac{2\lambda\xi^3 b^*(\xi)(1 - b^*(\xi))[\xi b^*(\xi) - \lambda(1 - b^*(\xi))]}{\theta[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]^3} > 0.$$

Let $\mathbf{H}(\theta, \alpha)$ denote the Hessian matrix of $f_{R_2}(\theta, \alpha)$. Then it can be shown that

$$\det(\mathbf{H}(\theta, \alpha)) = \frac{\lambda^2\xi^4 b^*(\xi)(1 - b^*(\xi))^2 \left\{3\xi^2 b^*(\xi) + 4[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]\right\}}{\theta^4[\alpha\xi b^*(\xi) - \lambda(1 - b^*(\xi))(\alpha + \xi)]^4} > 0.$$

Hence, $f_{R_2}(\theta, \alpha)$ is strictly convex on X. Consequently, $f_R = f_{R_1} + f_{R_2}$ is strictly convex on X, and thus, f_R^2 is strictly convex on X. ■

Lemma 3 *The function $f_Q^2(\alpha)$ is strictly convex for all $\alpha > 0$.*

Proof. The second derivative of f_Q with respect to α is given by

$$\frac{d^2 f_Q(\alpha)}{d\alpha^2} = 2\xi^2 \left\{ \frac{(\alpha+\xi)[\xi - \lambda(1-b^*(\xi))]\{\xi^3 b^*(\xi) - (\alpha+\xi)^2[\xi\hat{B}' - \lambda(1-b^*(\xi))]\}}{(\alpha+\xi)^3[\alpha\xi - \lambda(1-b^*(\xi))(\alpha+\xi)]^3} \right.$$
$$\left. + \frac{\xi b^*(\xi)[\alpha\xi - \lambda(1-b^*(\xi))(\alpha+\xi)]\{2\xi^2 + 3[\alpha\xi - \lambda(1-b^*(\xi))(\alpha+\xi)]\}}{(\alpha+\xi)^3[\alpha\xi - \lambda(1-b^*(\xi))(\alpha+\xi)]^3} \right\}.$$

It follows from the stability condition that $[\alpha\xi - \lambda(1-b^*(\xi))(\alpha+\xi)] > 0$ and $[\xi - \lambda(1-b^*(\xi))] > 0$, and Lemma 2 ensures that $\lambda(1-b^*(\xi)) - \xi\hat{B}'$ is nonnegative. Hence,

$$\frac{d^2 f_Q(\alpha)}{d\alpha^2} > 0,$$

which establishes strict convexity of f_Q for $\alpha > 0$. The strict convexity of $f_Q^2(\alpha)$ follows directly. ■

The following theorem is the main result of this section.

Theorem 2 *The cost function $C(\theta, \alpha)$ is strictly convex on X.*

Proof. The proof follows directly from Lemmas 2 and 3. In particular, the strict convexity of $f_Q^2(\alpha)$ for all $\alpha > 0$ ensures that $f_R^2(\theta, \alpha) + f_Q^2(\alpha)$, and thus $C(\theta, \alpha)$, is strictly convex on X. ■

Theorem 2, along with the convexity of X, show that the optimization problem is a convex program (CP). A CP guarantees that any stationary point (Karush-Kuhn-Tucker point) is a global minimizer, but to ensure existence of a solution, the feasible region must be closed and bounded. Although X is bounded, it is not closed. To circumvent this complication, we note that, for all values of θ and α,

$$\frac{\partial C(\theta, \alpha)}{\partial \theta} = \frac{-2c_R\lambda^2\xi(\alpha+\xi)(1-b^*(\xi))}{\theta^2[\alpha\xi b^*(\xi) - \lambda(1-b^*(\xi))(\alpha+\xi)]} f_R(\theta, \alpha) < 0$$

and

$$\frac{\partial C(\theta,\alpha)}{\partial \alpha} =$$

$$-2c_R\lambda^2\xi(1-b^*(\xi))f_R(\theta,\alpha)\left\{\frac{\lambda(1-b^*(\xi))\left\{(\alpha+\xi)[\lambda(1-b^*(\xi))-\xi\hat{B}']+\xi b^*(\xi)[\xi-\lambda(1-b^*(\xi))]\right\}}{b^*(\xi)[\alpha\xi b^*(\xi)-\lambda(1-b^*(\xi))(\alpha+\xi)]^2[\alpha\xi-\lambda(1-b^*(\xi))(\alpha+\xi)]}\right.$$

$$\left.+\frac{\xi^2 b^*(\xi)}{\theta[\alpha\xi b^*(\xi)-\lambda(1-b^*(\xi))(\alpha+\xi)]}+\frac{\alpha\left\{\xi[\lambda(1-b^*(\xi))-\xi\hat{B}']+b^*(\xi)[\alpha\xi-\lambda(1-b^*(\xi))(\alpha+\xi)]^2\right\}}{b^*(\xi)[\alpha\xi b^*(\xi)-\lambda(1-b^*(\xi))(\alpha+\xi)][\alpha\xi-\lambda(1-b^*(\xi))(\alpha+\xi)]^2}\right\}$$

$$-2\lambda c_Q f_Q(\alpha)\left\{\frac{\xi^3 b^*(\xi)-(\alpha+\xi)^2[\xi\hat{B}'-\lambda(1-b^*(\xi))]+2\xi b^*(\xi)[\alpha\xi-\lambda(1-b^*(\xi))(\alpha+\xi)]}{b^*(\xi)^2(\alpha+\xi)^2[\alpha\xi-\lambda(1-b^*(\xi))(\alpha+\xi)]^2}\right\}<0.$$

Hence, the cost function $C(\theta,\alpha)$ is monotonically decreasing in both θ and α and is bounded below by

$$c_S\frac{\lambda(1-b^*(\xi))^2}{\xi^2 b^*(\xi)^2}+c_R\frac{\lambda^3(1-b^*(\xi))^2[\lambda(1-b^*(\xi))-\xi\hat{B}']^2}{b^*(\xi)^2[\xi b^*(\xi)-\lambda(1-b^*(\xi))]^2[\xi-\lambda(1-b^*(\xi))]^2}+c_Q\frac{\lambda[\xi\hat{B}'-(1-b^*(\xi))]^2}{\xi^2 b^*(\xi)^2[\xi-\lambda(1-b^*(\xi))]^2}.$$

Therefore, the budget constraint (35) is always binding, and we may substitute

$$\theta=\frac{D-c_\alpha\alpha}{c_\theta}$$

into $C(\theta,\alpha)$. Differentiating the objective function with respect to α and setting it equal to zero, the optimal repair rate, denoted by α^*, is the unique root that satisfies Eq. (34). Subsequently, the equation,

$$\theta^*=\frac{D-c_\alpha\alpha^*}{c_\theta},$$

may be solved to obtain the optimal retrial rate θ^*. Thus, (θ^*,α^*) is a stationary point of C on X, and hence, the global minimizer.

5.2 Numerical Examples

We now illustrate the solution procedure in two distinct scenarios. In the first case, we assume the service times are exponentially distributed with positive rate parameter μ. In this case, the Laplace transform of the service time distribution is given by

$$b^*(s)=\frac{\mu}{\mu+s},$$

which gives

$$b^*(\xi) = \frac{\mu}{(\mu + \xi)}$$

and

$$\hat{B}' = \frac{\lambda\mu}{(\mu + \xi)^2}.$$

In the second scenario, we assume the service times are uniformly distributed on the interval $(0, 2/\mu)$. In this case, the Laplace transform of the service time distribution is

$$b^*(s) = \frac{\mu(1 - e^{-2s/\mu})}{2s}$$

so that

$$b^*(\xi) = \frac{\mu(1 - e^{-2\xi/\mu})}{2\xi}$$

and

$$\hat{B}' = \frac{\lambda\mu}{2\xi^2}\left[1 - \exp\left(\frac{-2\xi}{\mu}\right)\left(1 + \frac{2\xi}{\mu}\right)\right].$$

The cost coefficients in both cases are: $c_\theta = c_\alpha = c_S = 1$, and $c_R = c_Q = 100$. The remaining parameters, as well as the optimal solutions (indicated by $*$), are specified in Table 1.

Table 1: Optimal repair and retrial rates for two numerical examples.

Service time distribution	λ	μ	ξ	D	θ^*	α^*	$C(\theta^*, \alpha^*)$
Exponential (μ)	5.0	10.0	1.5	30.0	11.23	18.77	5.85
Uniform on (0,2/μ)	5.0	10.0	1.5	30.0	11.69	18.31	5.18

As expected, Table 1 verifies that the budget constraint is binding at the optimal solution. In particular, with $c_\theta = c_\alpha = 1$, we have $\theta^* + \alpha^* = D = 30.0$ in both examples.

Acknowledgements. We wish to express our gratitude to an anonymous referee and Professor Mark Lewis for several useful comments. This research was sponsored by the Air Force Office of Scientific Research (F1ATA0634J001).

References

[1] Aissani, A. (1988). On the M/G/1/1 queueing system with repeated orders and unreliable server. *Journal of Technology,* **6**, 98-123 (in French).

[2] Aissani, A. (1993). Unreliable queuing with repeated orders. *Microelectronics and Reliability,* **33** (14), 2093-2106.

[3] Aissani, A. (1994). A retrial queue with redundancy and unreliable server. *Queueing Systems,* **17** (3-4), 431-449.

[4] Aissani, A. and J. R. Artalejo (1998). On the single server retrial queue subject to breakdowns. *Queueing Systems,* **30** (3-4), 309-321.

[5] Anisimov, V. V. and K. L. Atadzhanov (1994). Diffusion approximation of systems with repeated calls and an unreliable server. *Journal of Mathematical Sciences,* **72** (2), 3032-3034.

[6] Artalejo, J. R. (1994). New results in retrial queueing systems with breakdown of the servers. *Statistica Neerlandica,* **48** (1), 23-36.

[7] Artalejo, J. R. (1997). Analysis of an M/G/1 queue with constant repeated attempts and server vacations. *Computers and Operations Research,* **24** (6), 493-504.

[8] Artalejo, J. R., A. Gomez-Corral, and M. F. Neuts (2001). Analysis of multiserver queues with constant retrial rate. *European Journal of Operational Research,* **135** (3), 569-581.

[9] Atencia, I. and P. Moreno (2005). A single-server retrial queue with general retrial times and Bernoulli schedule. *Applied Mathematics and Computation,* **162** (2), 855-880.

[10] Choi, B. D. and K. K. Park (1990). The M/G/1 retrial queue with Bernoulli schedule. *Queueing Systems,* **7** (2), 219-228.

[11] Djellab, N. V. (2002). On the M/G/1 retrial queue subjected to breakdowns. *RAIRO Operations Research,* **36** (4), 299-310.

[12] Falin, G.I. and J.G.C. Templeton (1997). *Retrial Queues*. Chapman and Hall, London.

[13] Hassin, R. (1996). On optimal and equilibrium retrial rates in a queueing system. *Probability in the Engineering and Informational Sciences,* **10** (2), 223-227.

[14] Kulkarni, V. G. (1983). A game theoretic model for two types of customers competing for service. *Operations Research Letters,* **2** (3), 119-122.

[15] Kulkarni, V. G. and B. D. Choi (1990). Retrial queues with server subject to breakdowns and repairs. *Queueing Systems,* **7** (2), 191-208.

[16] Kumar, B. K., S. P. Madheswari, and A. Vijayakumar (2002). The M/G/1 retrial queue with feedback and starting failures. *Applied Mathematical Modelling,* **26** (11), 1057-1075.

[17] Lam, Y., Y. L. Zhang, and Q. Liu (2006). A geometric process model for M/M/1 queueing system with a repairable service station. *European Journal of Operational Research,* **168** (1), 100-121.

[18] Langaris, C. and E. Moutzoukis (1995). A retrial queue with structured batch arrivals, priorities and server vacations. *Queueing Systems,* **20** (3-4), 341-368.

[19] Lee, H. S. (1995). Optimal control of the M^X/G/1/K queue with multiple server vacations. *Computers and Operations Research,* **22** (5), 543-552.

[20] Li, H. and T. Yang (1995). A single-server retrial queue with server vacations and a finite number of input sources. *European Journal of Operational Research,* **85** (1), 149-160.

[21] Li, H. and Y. Q. Zhao (2005). A retrial queue with a constant retrial rate, server break downs and impatient customers. *Stochastic Models,* **21** (2-3), 531-550.

[22] Liang, H. M. and V. G. Kulkarni (1999). Optimal routing control in retrial queues. In *Applied Probability and Stochastic Processes: International Series in Operations Research and Management Science, Vol. 19*, (J.G. Shanthikumar and Ushio Sumita, eds.), 203-218.

[23] Moutzoukis, E. and C. Langaris (1996). Non-preemptive priorities and vacations in a multiclass retrial queueing system. *Communications in Statistics: Stochastic Models,* **12** (3), 455-472.

[24] Wang, K. H., H. T. Kao, and G. Chen (2004). Reliability analysis of the retrial queue with server breakdowns and repairs. *Quality Technology and Quantitative Management,* **1** (2), 325-339.

[25] Wang, J., J. Cao, and Q. Li (2001). Reliability analysis of the retrial queue with server breakdowns and repairs. *Queueing Systems,* **38** (4), 363-380.

[26] Wu, X., P. Brill, M. Hlynka, and J. Wang (2005). An M/G/1 retrial queue with balking and retrials during service. *International Journal of Operational Research,* **1** (1-2), 30-51.

[27] Yang, T. and H. Li (1994). The M/G/1 retrial queue with the server subject to starting failures. *Queueing Systems,* **16** (1-2), 83-96.