Performance Analysis and Statistical Modeling of the Single-Server Non-reliable Retrial Queueing System with a Threshold-Based Recovery

Dmitry Efrosinin[1](✉) and Janos Sztrik[2]

[1] Johannes Kepler University, Altenbergerstrasse 69, 4040, Linz, Austria
dmitry.efrosinin@jku.at
http://www.jku.at, http://www.unideb.hu
[2] Debrecen University, Egyetem Ter. 1, 4032 Debrecen, Hungary
sztrik.janos@inf.unideb.hu

Abstract. In this paper we study a single-server Markovian retrial queueing system with non-reliable server and threshold-based recovery policy. The arrived customer finding a free server either gets service immediately or joins a retrial queue. The customer at the head of the retrial queue is allowed to retry for service. When the server is busy, it is subject to breakdowns. In a failed state the server can be repaired with respect to the threshold policy: the repair starts when the number of customers in the system reaches a fixed threshold level. Using a matrix-analytic approach we perform a stationary analysis of the system. The optimization problem with respect to the average cost criterion is studied. We derive expressions for the Laplace transforms of the waiting time. The problem of estimation and confidence interval construction for the fully observable system is studied as well.

Keywords: Quasi-birth-and-death process · Retrial queues · Performance analysis · Confidence intervals

1 Introduction

Different types of single server retrial queueing systems have found applications in local area networks and communication protocols. In a retrial queue a customer who finds the server busy is assigned to a queue of retrial customers. It is assumed that the arrived customer finding a free server with probability p gets service immediately or joins a retrial queue with probability $1 - p$. Many

D. Efrosinin—This work was funded by the COMET K2 Center "Austrian Center of Competence in Mechatronics (ACCM)", funded by the Austrian federal government, the federal state Upper Austria, and the scientific partners of the ACCM.
J. Sztrik— The research is supported by the Austro-Hungarian Cooperation Grant No. 90u6, OMAA 2014, Stiftung Aktion Österreich-Ungarn.

A. Dudin et al. (Eds.): ITMM 2015, CCIS 564, pp. 330–343, 2015.
DOI: 10.1007/978-3-319-25861-4_28

papers study the case $p = 1$, where customers have a direct access to the server, or $p = 0$, when a customer upon arrival goes always to the retrial queue. The bibliography for these two particular cases as well as a description of a general model for arbitrary value p can be found in [2].

In our model the customer at the head of the retrial queue is allowed to retry for service, i.e. the system has a retrial queue with a constant retrial policy or FCFS retrial queue. The constant retrial policy was introduced by [8] and it was used in many applications to local area networks and communication protocols, e.g. in [3,4,6,10]. The system with constant retrial policy is simpler to analyze that one with the classical retrial policy assuming the state-dependent retrial intensity, since in the latter case the QBD process with three diagonal block infinitesimal matrix can be constructed. Moreover the constant retrial policy can be used in a truncation model of classical policy exhibiting spatial homogeneity from some orbit level upwards.

The systems with an unreliable server have been studied extensively. But the systems which combines server breakdowns with a retrial effect are still not exhaustively examined. We refer the interested readers to the papers of [1,9] and bibliographies therein. The system under study is assumed to be controllable in the sense that the repair process in a failed state starts according to the threshold-based recovery policy. This policy prescribes to switch on/off the repair facility if the number of customers in the system is higher or lower than a fixed threshold level $q_r \geq 1$. The threshold-based recovery was first introduced by [5] in case of the system with an ordinary queue. Then the obtained results were generalized by [7] to case of the retrial queue with a constant retrial rate.

Whenever a queueing system is fully observable with respect to their random time periods such as inter-arrival time, service time, time to failure, repair time, inter-retrial time and so on, standard parametric estimation methods of mathematical statistics seems to be quite appropriate. But the most papers include only the results about transient and stationary solutions and very few consider the associated statistical problems. [14] have evaluated confidence intervals for the mean waiting time of the single server and tandem queues with blocking. Maximum likelihood estimates of multi-server system with heterogeneous servers were obtained by [13]. In [12] have studied the estimation of arrival and service rates for queues based only on queue length data.

The analysis of the presented retrial queue with constant retrial rate, non-reliable server and threshold-based recovery includes the following contributions:

(a) We model the system as a quasi-birth-and-death (QBD) process with threshold dependent block-tridiagonal infinitesimal matrix and apply a general theory of matrix-analytic solutions to derive the stationary distribution of the system states and stability condition.
(b) We formulate optimization problem to calculate a threshold level which minimizes the long-run average cost per unit of time for the given cost structure.
(c) We derive the main performance characteristics of the system for the given threshold policy.

(d) We obtain the Laplace transforms of the waiting time distribution.
(e) We perform a parameter estimation and construct confidence intervals for the performance measures.

In further sections we will use the notation I for the identity matrix, O and $\mathbf{0}$ – respectively for the square matrix and row vector with zero entries. Furthermore $\mathbf{e}$ will denote a column vector of ones and $\mathbf{e}_j$ – a column vector with 1 in the j-th position and 0 elsewhere. Vectors and matrices are assumed to have an appropriate size. The symbol ∇ will stand for the gradient.

2 Mathematical Model and Stability

We consider a $M/M/1$ queueing system illustrated in Fig. 1. Customers arrive to the system according to a Poisson stream with intensity $\lambda > 0$. The server servers the customers according to an exponentially distributed time with parameter $\mu > 0$. If the server is idle at the time of an external arrival, the customer proceeds to the server with probability p or to the orbit with probability $1 - p$. In particular case $p = 1$ we get a system with a direct access to the server which was already studied in [7]. The servicing customer leaves the system after service completion. If the server is found to be blocked, i.e. busy or failed, the customer has to enter the infinite capacity retrial queue. We assume a constant retrial policy, i.e. FCFS discipline for the retrial queue, when the customer at the head of the queue repeats its requests for service in exponentially distributed retrial times with intensity $\tau > 0$. The server is assumed to be unreliable. During a service process it may fail in exponentially distributed time with intensity $\alpha > 0$. The repair time is again exponentially distributed with intensity $\beta > 0$. The system under study is regulated by a controller who switches the repair facility on only when the number of customers in the system reaches a fixed threshold level $q_r \geq 1$ and switch it off if the number of customers decreases below this level. The inter-arrival times, intervals of successive retrials, service, breakdown and repair times are assumed to be mutually independent.

The system states at time t are described by random vector $\{N(t), D(t)\}_{t \geq 0}$, where $N(t)$ – the number of customers in the queueing system and $D(t)$ – the

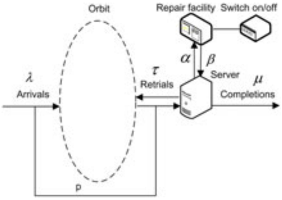

Fig. 1. Scheme of the queueing system

server state, where

$$D(t) = \begin{cases} 0 & \text{the server is idle,} \\ 1 & \text{the server is busy,} \\ 2 & \text{the server is failed.} \end{cases}$$

Note that if $D(t) = 0$, the component $N(t) \in \mathbb{N}_0$, otherwise $N(t) \in \mathbb{N}$. The random process

$$\{X(t)\}_{t \geq 0} = \{N(t), D(t)\}_{t \geq 0} \tag{1}$$

is an irreducible continuous-time Markov chain with a state space

$$E = \{x = (n, d); n \geq 0, d = 0 \vee n \geq 1, d \in \{1, 2\}\} \tag{2}$$

and transition intensities $\lambda_{xy}(q_r)$ from state $x = (n, d) \in E$ to state $y = (n', d')$ $\in E$,

$$\lambda_{xy}(q_r) = \begin{cases} \lambda p, & n' = n + 1,\, d' = 1,\, d = 0,\, n \geq 0, \\ \lambda(1 - p), & n' = n + 1,\, d' = d = 0,\, n \geq 0, \\ \lambda, & n' = n + 1,\, d' = d \in \{1, 2\},\, n \geq 0, \\ \mu, & n' = n - 1,\, d' = 0,\, d = 1,\, n > 0, \\ \tau, & n' = n,\, d' = 1,\, d = 0,\, n > 0, \\ \alpha, & n' = n,\, d' = 2,\, d = 1,\, n > 0, \\ \beta, & n' = n,\, d' = 1,\, d = 2,\, n \geq q_r. \end{cases} \tag{3}$$

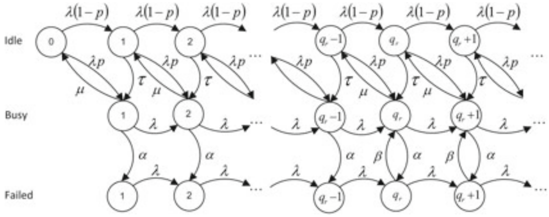

Fig. 2. The state-transition-intensity diagram for the given threshold q_r

Figure 2 illustrates the state transition rates. Now we define a macro-state $\mathbf{n}$ consisting of three states,

$$\mathbf{n} = \{(n, 0), (n + 1, 1), (n + 1, 2)\},\ n \geq 0.$$

Define by $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$ a row vector of stationary state probabilities with subvectors $\boldsymbol{\pi}_n = (\pi_{(n,0)}, \pi_{(n+1,1)}, \pi_{(n+1,2)})$ for the macro-state $\mathbf{n}$, where

$$\pi_{(n,d)} = \lim_{t \to \infty} \mathbb{P}[N(t) = n, D(t) = d].$$

Theorem 1. *For the fixed threshold q_r the Markov chain (1) belongs to a class of the QBD processes with boundary states and block tri-diagonal infinitesimal matrix $\Lambda = [\lambda_{xy}(q_r)]$,*

$$\Lambda = \left(\begin{array}{cccccccc} Q_{1,0} & Q_0 & O & O & O & O & O & \dots \\ Q_2 & Q_{1,1} & Q_0 & O & O & O & O & \dots \\ O & Q_2 & Q_{1,1} & Q_0 & O & O & O & \dots \\ \ddots & & \ddots & \ddots & \ddots & & & \dots \\ \hline O & O & O & Q_2 & Q_{1,2} & Q_0 & O & \dots \\ O & O & O & O & Q_2 & Q_{1,2} & Q_0 & \dots \\ \ddots & & & & & \ddots & \ddots & \ddots \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\1\\1\\1\end{array}}\right\} q_r - 1 \\ \\ \\ \\ \end{array},$$

where

$$Q_{1,j} = \begin{pmatrix} -(\lambda + \tau 1_{\{j \neq 0\}}) & \lambda p & 0 \\ \mu & -(\alpha + \lambda + \mu) & \alpha \\ 0 & \beta 1_{\{j=2 \vee q_r=1\}} & -(\lambda + \beta 1_{\{j=2 \vee q_r=1\}}) \end{pmatrix},$$

$$Q_0 = \begin{pmatrix} \lambda(1-p) & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}, \ Q_2 = \begin{pmatrix} 0 & \tau & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and row vector $\boldsymbol{\pi}$ satisfies the matrix system

$$\boldsymbol{\pi}\Lambda = \mathbf{0}, \ \boldsymbol{\pi}\mathbf{e} = 1.$$

Proof. The result follows by arranging the balance equations according to the macro states $\mathbf{n}$ and collecting them in matrix form.

The next statement reveals the condition that is necessary and sufficient to ensure system stability.

Theorem 2. *The necessary and sufficient stability condition for the process $\{X(t)\}_{t \geq 0}$ is given by*

$$\rho = \frac{\lambda}{\mu}\Big(1 + \frac{\alpha}{\beta}\Big) + \frac{\lambda}{\lambda p + \tau} < 1. \tag{4}$$

Proof. Consider the matrix $A = Q_0 + Q_{1,2} + Q_2$, composed of matrices defined above, and let $\boldsymbol{p}$ be its stationary distribution. Since A is irreducible, the vector $\boldsymbol{p}$ exists such that $\boldsymbol{p}A = \mathbf{0}$ and $\boldsymbol{p}\,\mathbf{e} = 1$. It is given by

$$\boldsymbol{p} = \frac{1}{\alpha(p\lambda + \tau) + \beta(p\lambda + \mu + \tau)}(\beta\mu, \beta(p\lambda + \tau), \alpha(p\lambda + \tau)). \tag{5}$$

According to the mean drift result of [11] the stability condition is given by the inequality $\boldsymbol{p}\,Q_2\boldsymbol{e} > \boldsymbol{p}\,Q_0\boldsymbol{e}$ which leads to the proposed inequality.

3 Evaluation of Performance Measures

Using the general theory of the QBD processes (see e.g. [11]) we get the following result.

Theorem 3. *Subvectors $\boldsymbol{\pi}_n, n \geq 0$, of stationary state probabilities are calculated by*

$$\begin{aligned} \boldsymbol{\pi}_n &= \boldsymbol{\pi}_{q_r} \prod_{i=1}^{q_r-n} M_{q_r-i}, \quad 0 \leq n \leq q_r - 1, \\ \boldsymbol{\pi}_n &= \boldsymbol{\pi}_{q_r} R^{n-q_r}, \quad n \geq q_r, \end{aligned} \tag{6}$$

where the matrices M_i are given by

$$\begin{aligned} M_0 &= -Q_2 Q_{1,0}^{-1}, \\ M_i &= -Q_2(M_{i-1}Q_0 + Q_{1,1})^{-1}, \quad 1 \leq i \leq q_r - 2, \\ M_{q_r-1} &= -Q_2(M_{q_r-2}Q_0 + Q_{1,2})^{-1}. \end{aligned} \tag{7}$$

The vector $\boldsymbol{\pi}_{q_r}$ is a unique solution of the system of equations

$$\boldsymbol{\pi}_{q_r}(M_{q_r-1}Q_0 + Q_{1,2} + RQ_2) = \mathbf{0}, \tag{8}$$

$$\boldsymbol{\pi}_{q_r}\Big(\sum_{n=0}^{q_r-1} \prod_{i=1}^{q_r-n} M_{q_r-i} + (I-R)^{-1}\Big)\mathbf{e} = 1. \tag{9}$$

Matrix R is the minimal non-negative solution to matrix equation $R^2Q_2 + RQ_{1,2} + Q_0 = \mathrm{O}$ and has the following explicit representation,

$$R = \begin{pmatrix} \frac{\lambda(1-p)}{\tau} & \frac{\lambda^2(1-p)}{\mu\tau} & \frac{\alpha\lambda^2(1-p)}{\mu\tau(\beta+\lambda)} \\ \frac{\lambda}{\tau} & \frac{\lambda(\lambda+\theta)}{\mu\tau} & \frac{\lambda(\lambda+\tau)\alpha}{\mu\tau(\beta+\lambda)} \\ \frac{\lambda}{\tau} & \frac{\lambda(\lambda+\tau)}{\mu\tau} & \frac{\lambda(\lambda+\tau)\alpha+\lambda\mu\tau}{\mu\tau(\beta+\lambda)} \end{pmatrix}. \tag{10}$$

Proof. Due to the general theory of the QBD processes (see [11, Chapter3,pp.82–83.]), subvectors $\boldsymbol{\pi}_n$ which correspond to the macro-states $\mathbf{n}$ with homogeneous blocks in the matrix Λ, have geometric structure,

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_{q_r} R^{n-q_r}, \quad n > q_r.$$

For the probabilities $\boldsymbol{\pi}_n$, $0 \leq n \leq q_r$, of the boundary states the system of balance equations can be transformed in the form

$$\begin{aligned} &\boldsymbol{\pi}_0 Q_{1,0} + \boldsymbol{\pi}_1 Q_2 = \mathbf{0}, \\ &\boldsymbol{\pi}_{n-1}Q_0 + \boldsymbol{\pi}_n Q_{1,1} + \boldsymbol{\pi}_{n+1}Q_2 = \mathbf{0},\ 1 \leq n \leq q_r - 2, \\ &\boldsymbol{\pi}_{q_r-2}Q_0 + \boldsymbol{\pi}_{q_r-1}Q_{1,2} + \boldsymbol{\pi}_{q_r}Q_2 = \mathbf{0}. \end{aligned}$$

The last system implies the recurrent relation

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n+1} M_n, \quad 0 \le n \le q_r - 1,$$

where matrices M_n can be evaluated recursively using (7). For the boundary state $n = q_r$ we get

$$\boldsymbol{\pi}_{q_r-1} Q_0 + \boldsymbol{\pi}_{q_r} Q_{1,2} + \boldsymbol{\pi}_{q_r+1} Q_2 = \mathbf{0}.$$

Subsequent substitution of $\boldsymbol{\pi}_{q_r-1} = \boldsymbol{\pi}_{q_r} M_{q_r-1}$ and $\boldsymbol{\pi}_{q_r+1} = \boldsymbol{\pi}_{q_r} R$ to the last equality leads to (8). This equation can be solved with respect to the last unknown vector $\boldsymbol{\pi}_{q_r}$ together with the normalizing condition which follows from the relation

$$\sum_{n=0}^{\infty} \boldsymbol{\pi}_n \mathbf{e} = \sum_{n=0}^{q_r-1} \boldsymbol{\pi}_n \mathbf{e} + \boldsymbol{\pi}_{q_r} \sum_{n=q_r}^{\infty} R^{n-q_r} \mathbf{e} = 1.$$

Finally, the structure of matrices $Q_0, Q_{1,2}, Q_2$ together with a relation $RQ_2\mathbf{e} = Q_0\mathbf{e}$ implies the form (10).

Corollary 1. *Using the probabilities $\boldsymbol{\pi}_n$, $n \ge 0$, we can evaluate different performance measures:*

Utilization of the system

$$U = 1 - \boldsymbol{\pi}_0 \boldsymbol{e}_{\mathbf{1}}. \tag{11}$$

Probability of a server being blocked

$$P_{blocking} = \Big(\sum_{n=0}^{q_r-1} \boldsymbol{\pi}_n + \boldsymbol{\pi}_{q_r} (I - R)^{-1} \Big) (\mathbf{e}_2 + \mathbf{e}_3). \tag{12}$$

Mean number of customers in the queue

$$\bar{Q} = \Big(\sum_{n=0}^{q_r-1} n \boldsymbol{\pi}_n + \boldsymbol{\pi}_{q_r} (q_r (I - R) + R)(I - R)^{-2} \Big) \mathbf{e}. \tag{13}$$

Mean number of customers in the system

$$\bar{N} = \bar{Q} + P_{blocking}. \tag{14}$$

Mean waiting and sojourn times

$$\bar{W} = \frac{\bar{N}}{\lambda}, \quad \bar{T} = \frac{\bar{Q}}{\lambda}. \tag{15}$$

A natural question that may arise in practice is a calculation of an optimal threshold policy which leads to the minimum of the system operating costs per unit of time. To find the optimal threshold q^* the following cost structure is introduced: c_0 – holding cost per unit time for each customer in the system, $c_{0,0}$, $c_{1,0}$ and $c_{2,0}$ – usage costs per unit time if the server is idle, busy or failed for $N(t) < q_r$. For $N(t) \ge q_r$ the costs $c_{0,1}$, $c_{1,1}$ $c_{2,1}$ – usage costs together with the operational costs of the repair facility.

Corollary 2. *The average cost function $g(q_r)$ is of the form*

$$g(q_r) = c_0\bar{N} + \sum_{n=0}^{q_r-2} \boldsymbol{\pi}_n(c_{0,0}, c_{1,0}, c_{2,0})' + \boldsymbol{\pi}_{q_r-1}(c_{0,0}, c_{1,1}, c_{2,1})' \qquad (16)$$
$$+ \boldsymbol{\pi}_{q_r}(I - R)^{-1}(c_{0,1}, c_{1,1}, c_{2,1})',$$

where $(c_{0,i}, c_{1,i}, c_{2,i})'$ is a column-vector of the costs per unit of time, $i = 0, 1$.

In many cases a simple exhaustion method is quite appropriate to calculate the optimal value q_r^*. Setting $\frac{d}{dq_r}g(q_r) = 0$ the optimal threshold level q_r^* can also be numerically evaluated.

4 The Waiting Time Distribution

Here we want to calculate the distribution function of the waiting time W of the customer in the retrial queue. In comparison to the classical queue, where the conditional waiting time of the tagged customer is Erlang distributed, the conditional waiting time in a present system will depend on the future arrivals. It happens due to the presence of the threshold-based policy for the recovering of the server and due to the fact that with probability p a new arrival is served according to the LCFS (last come first served) discipline. Therefore it is required to observe the state of the tagged customer up to the time where its service begins. The further calculation is performed by analyzing of the auxiliary Markov chain just after an arrival of the tagged customer at time t^+,

$$\{\hat{X}(t)\}_{t \geq t^+} = \{N(t), D(t), M(t)\}_{t \geq t^+}.$$

The state space of this process is

$$\hat{E} = \{\hat{x} = (n, d, m) | n \geq 0, d = 0 \vee n \geq 1, d \in \{1, 2\}, \, 0 \leq m \leq n\}$$

with an absorption states with $m = 0$ when the tagged customer receives the service. The component $M(t)$ of the process denotes the position of the tagged customer in the list of waiting customers at time t. This component can only decrease at retrial time when the server is idle. If the server is busy or failed then we obviously have $M(t^*) = N(t^*) - 1$. The process is absorbed when the component $M(t)$ becomes equal to zero.

The waiting time distribution of the tagged customer is obtained as follows. First we calculate the Laplace transform of the conditional waiting time distribution given the system state and the position of the tagged customer after the arrival. Using the law of total probability and the state distribution just after the arrival of the tagged customer, the conditioning is removed. Numerical inversion of the Laplace transform completes the calculation.

Denote by $w_{(n,d,m)}(t)$ the probability density function of the conditional waiting time given state $\hat{x} = (n, d, m) \in \hat{E}$ and $\tilde{w}_{n,d,m}(s) = \int_0^\infty e^{-st} w_{(n,d,m)}(t)dt$ the corresponding Laplace transform (LT). Then due to the PASTA property and the law of total probability the unconditional LT is of the form

$$\tilde{w}(s) = \sum_{n=0}^{\infty} \pi_{(n,0)} p + \sum_{n=0}^{\infty} \pi_{(n,0)}(1-p)\tilde{w}_{(n+1,0,n+1)}(s) \qquad (17)$$
$$+ \sum_{n=1}^{\infty} \pi_{(n,1)} \tilde{w}_{(n+1,1,n)}(s) + \sum_{n=1}^{\infty} \pi_{(n,2)} \tilde{w}_{(n+1,2,n)}(s),$$

where the first summand represents the stationary probability that a tagged customer does not have to wait for service, i.e. $W = 0$; the last three terms represent the LT of the waiting time given $W > 0$.

Now we partition the conditional LT $\tilde{w}_{(n,d,m)}(s), (n,d,m) \in \hat{E}$, according to the number of customers in the system and define the column-vectors

$$\tilde{\mathbf{w}}_{n,m}(s) = (\tilde{w}_{(n,0,m)}(s), \tilde{w}_{(n+1,1,m)}(s), \tilde{w}_{(n+1,2,m)})', \; m \le n \le q_r + m - 1.$$

For the calculation of the conditional waiting time we make use of the Laplace transform of conditional service time for $n \ge q_r$: Let $h_1(t)$ and $h_2(t)$ denote the probability density functions from the start in an busy or failed state to the next departure. Obviously we have

$$h_1(t) = \frac{\mu}{\mu+\alpha}(\mu+\alpha)e^{-(\mu+\alpha)t} + \frac{\alpha}{\mu+\alpha}\int_0^t (\mu+\alpha)e^{-(\mu+\alpha)x h_f(t-x)dx},$$
$$h_2(t) = \int_0^t \beta e^{-\beta x} h_1(t-x)dx.$$

Denote by $\tilde{h}_1(s)$ and $\tilde{h}_2(s)$ the corresponding LT. For these functions we get

$$\tilde{h}_1(s) = \frac{\mu(\beta+s)}{(\mu+\alpha+s)(\beta+s)-\alpha\beta}, \quad \tilde{h}_2(s) = \frac{\beta}{\beta+s}\tilde{h}_1(s).$$

Theorem 4. *The vector of conditional Laplace transforms $\tilde{\mathbf{w}}_{n,m}(s), m \le n \le q + m - 2$ under stability condition satisfy the following recurrent relations*

$$\begin{aligned}
\tilde{\mathbf{w}}_{n,m}(s) &= M_1^{q-n-1}(s) M_2^m(s) \tilde{\mathbf{w}}_{q+m-1,m}(s) \qquad (18)\\
&+ M_1^{q-n-1}(s) \sum_{r=0}^{m-1} M_2^r(s) L_2(s) \tilde{\mathbf{w}}_{q+r-1,m-1}(s)\\
&+ \sum_{r=0}^{q-n-2} M_1^r(s) L_1(s) \tilde{\mathbf{w}}_{n+r-1,m-1}(s), \; n \le q-2,\\
\tilde{\mathbf{w}}_{n,m}(s) &= M_2^{q-n+m-1}(s) \tilde{\mathbf{w}}_{q+m-1,m}(s)\\
&+ \sum_{r=0}^{q-n+m-2} M_2^r(s) L_2(s) \tilde{\mathbf{w}}_{n+r-1,m-1}(s), \; n \ge q-1,\\
\tilde{\mathbf{w}}_{q+m-1,m}(s) &= \tilde{\vartheta}_1^{m-1}(s)(\tilde{\vartheta}_0(s), \tilde{\vartheta}_1(s), \tilde{\vartheta}_2(s))', \; \tilde{\mathbf{w}}_{n,0}(s) = (0,1,1)',
\end{aligned}$$

where

$$M_j(s) = -(Q_{1,j} - sI)^{-1}Q_0,\, L_j(s) = -(Q_{1,j} - sI)^{-1}Q_2,\, i = 1, 2, \tag{19}$$

$$\tilde{\vartheta}_0(s) = \frac{\tau}{\tau + \lambda p(1 - \tilde{h}_1(s)) + s},\, \tilde{\vartheta}_1(s) = \tilde{h}_1(s)\tilde{\vartheta}_0(s),\, \tilde{\vartheta}_2(s) = \tilde{h}_2(s)\tilde{\vartheta}_0(s). \tag{20}$$

Proof. The Markov property of the process $\{\hat{X}(t)\}$ implies the following system

$$(\lambda + \theta + s)\tilde{w}_{(n,0,m)}(s) = \lambda p\tilde{w}_{(n+1,1,m)}(s) + \lambda(1-p)\tilde{w}_{(n+1,0,m)} + \theta\tilde{w}_{(n,1,m-1)}(s),$$
$$(\alpha + \lambda + \mu + s)\tilde{w}_{(n+1,1,m)}(s) = \alpha\tilde{w}_{(n+1,2,m)}(s) + \lambda\tilde{w}_{(n+2,1,m)}(s) + \mu\,\tilde{w}_{(n,0,m)}(s),$$
$$(\lambda + \beta I_{\{n\geq q-1\}} + s)\tilde{w}_{(n+1,2,m)}(s) = \lambda\tilde{w}_{(n+2,2,m)}(s) + \beta\tilde{w}_{(n+1,1,m)}(s)I_{\{n\geq q-1\}},$$

where $\tilde{w}_{(n,1,0)}(s) = \tilde{w}_{(n,2,0)}(s) = 1$ and $\tilde{w}_{(n,0,0)}(s) = 0$. After routing block identification taking into account the difference of the transition rates for the states below and above threshold level, this system can be expressed in matrix form

$$(Q_{1,1}I_{\{n\leq q_r-2\}} + Q_{1,2}I_{\{n\geq q_r-1\}} - sI)\tilde{\mathbf{w}}_{n,m}(s) + Q_0\tilde{\mathbf{w}}_{n+1,m}(s) + Q_2\tilde{\mathbf{w}}_{n-1,m-1}(s) = 0.$$

The recursive forward substitution applied $q_r + m - 1 - n$ times using the notations (19) leads to the expressions (18). Note that the Laplace transforms $\tilde{\mathbf{w}}_{q_r+m-1,m}(s)$ do not depend on future arrivals to the queue, since the number of customers in the retrial queue always exceeds the given threshold level q_r during the waiting time of the tagged customer. To calculate the components of this vector we derive the Laplace transforms $\tilde{\vartheta}_d(s)$ of the waiting time for the customer at the head of the retrial queue given the initial server state $d \in \{0, 1, 2\}$. Obviously these LTs satisfies (20), since the random time to absorption is equal to the sum of the service time given states 1 or 2 of the server plus the time to absorption given state 0.

If we take into account the sequence of epochs at which the queue size decreases in one unit, we easily find the expression (18).

Corollary 3. *For the unconditional LT of the waiting time distribution we have*

$$\tilde{W}(s) = \frac{1}{s}(1 - \boldsymbol{\pi}_W\mathbf{e} + \boldsymbol{\pi}_W\tilde{\mathbf{w}}(s)),$$

where the contribution $1-\boldsymbol{\pi}_W\mathbf{e}$ *is equal to the first summand of (17) and* $\boldsymbol{\pi}_W\tilde{\mathbf{w}}(s)$ *stands for the last three terms defined in (17).*

5 Confidence Intervals for Performance Measures

Consider a real life system which runs without control, i.e. $q_r = 1$, and system parameters are unknown. Using random samples drawn from observed data we derive simple parameter estimators. An estimator $\hat{q}_r^*$ for the optimal threshold q_r^* is calculated from them next. Given a system which runs under threshold $\hat{q}_r^*$ we provide consistent and asymptotically normal estimators and corresponding confidence intervals for its performance measures. Numerical examples illustrate the performance improvement due to introduced control at the end of the section.

5.1 System Parameter Estimators

Let $(X_1, X_2, \dots, X_n), (Y_1, Y_2, \dots, Y_n), (Z_1, Z_2, \dots, Z_n), (U_1, U_2, \dots, U_n)$ and $(H_1, H_2, \dots, H_n)$ each be random samples of size n, which, respectively, are drawn from different exponentially distributed inter-arrival time populations with parameter λ, exponentially distributed service time populations with parameter μ, exponentially distributed time to failure populations with parameter α, exponentially distributed repair time populations with parameter β and exponentially distributed inter-retrial time populations with parameter τ. It follows that $\mathbb{E}[\bar{X}] = \frac{1}{\lambda}$, $\mathbb{E}[\bar{Y}] = \frac{1}{\mu}$, $\mathbb{E}[\bar{Z}] = \frac{1}{\alpha}$, $\mathbb{E}[\bar{U}] = \frac{1}{\beta}$ and $\mathbb{E}[\bar{H}] = \frac{1}{\tau}$, where $\bar{X}, \bar{Y}, \bar{Z}$, $\bar{U}$ and $\bar{H}$ are the sample means of inter-arrival time, service time, time to failure, repair time and inter-retrial time. It is obvious that $\bar{X}, \bar{Y}, \bar{Z}, \bar{U}$ and $\bar{H}$ are the maximum likelihood estimators of $\frac{1}{\lambda}$, $\frac{1}{\mu}$, $\frac{1}{\alpha}$, $\frac{1}{\beta}$ and $\frac{1}{\tau}$. Let $(J_1, J_2, \dots, J_n)$ be the random sample of size n with

$$J_i = \begin{cases} 1 & \text{if the i-th arrived customer finding server idle proceeds to the server,} \\ 0 & \text{if the i-th arrived customer finding server idle proceeds to the orbit.} \end{cases}$$

Obviously,

$$\bar{J} \cdot n \xrightarrow{d} \mathcal{B}(n, p)$$

is binomially distributed with parameters n and p. It follows that $\mathbb{E}[\bar{J}] = p$ and $\mathbb{V}[\bar{J}] = \frac{p(1-p)}{n}$. Thus, this relative frequency $\bar{J}$ serves as an unbiased estimator for probability p.

5.2 Optimal Threshold Estimator

We use the average cost function $g(q_r)$ from (16) to derive an estimator $\hat{q}_r^*$ for the optimal threshold q_r^*. For this reason we transform the optimization problem

$$\min_{q_r \in \mathbb{N}} g(q_r) = \min_{q_r \in \mathbb{N}} g(q_r, \lambda, \mu, \alpha, \beta, \tau, p) = g(q_r^*) \tag{21}$$

into

$$\min_{q_r \in \mathbb{N}} g(q_r, \bar{X}^{-1}, \bar{Y}^{-1}, \bar{Z}^{-1}, \bar{U}^{-1}, \bar{H}^{-1}, \bar{J}) = g(\hat{q}_r^*) \tag{22}$$

and numerically evaluate $\hat{q}_r^*$.

5.3 The Consistent and Asymptotically Normal Estimator

Let $\phi(\lambda, \mu, \alpha, \beta, \tau, p)$ denote any function from corollaries 3.1 and 3.2, which characterizes the performance of the system which runs under threshold $\hat{q}_r^*$. For example it can be the cost function $g(\hat{q}_r^*)$ or the mean number of customers in the system $\bar{N}(\hat{q}_r^*)$ with $q_r = \hat{q}_r^*$. In order to derive an estimator for ϕ we linearize

$$\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J}) = \phi(\bar{X}^{-1}, \bar{Y}^{-1}, \bar{Z}^{-1}, \bar{U}^{-1}, \bar{H}^{-1}, \bar{J})$$

around the point $\boldsymbol{\mu} = (\lambda^{-1}, \mu^{-1}, \alpha^{-1}, \beta^{-1}, \tau^{-1}, p)$ and get the approximation

$$\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J}) \approx \hat{\phi}(\boldsymbol{\mu}) - \boldsymbol{\mu}\ \nabla\hat{\phi}(\boldsymbol{\mu}) + (\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J})\ \nabla\hat{\phi}(\boldsymbol{\mu}). \tag{23}$$

The random vector

$$(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J}) \xrightarrow{d} \mathcal{MN}(\boldsymbol{\mu}, \Sigma)$$

is asymptotically multi-normal distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma = diag(\frac{1}{\lambda^2 n}, \frac{1}{\mu^2 n}, \frac{1}{\alpha^2 n}, \frac{1}{\beta^2 n}, \frac{1}{\tau^2 n}, \frac{p(1-p)}{n})$ due to the multivariate central limit theorem. We employ the theorem of the affine transformation on the above approximation and get

$$\begin{aligned}\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J}) &\xrightarrow{d} \mathcal{N}\left(\hat{\phi}(\boldsymbol{\mu}), \nabla\hat{\phi}^t(\boldsymbol{\mu})\ \Sigma\ \nabla\hat{\phi}(\boldsymbol{\mu})\right) \\ &= \mathcal{N}\left(\phi(\lambda, \mu, \alpha, \beta, \tau, p), \nabla\hat{\phi}^t(\boldsymbol{\mu})\ \Sigma\ \nabla\hat{\phi}(\boldsymbol{\mu})\right).\end{aligned} \tag{24}$$

Hence, it is a consistent and asymptotically normal estimator of any performance measure $\phi(\lambda, \mu, \alpha, \beta, \tau, p)$.

5.4 Confidence Intervals for Performance Measures

Using Slutsky's theorem we get from (24)

$$\frac{\left(\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J}) - \phi(\lambda, \mu, \alpha, \beta, \tau, p)\right)}{\sqrt{\nabla\hat{\phi}^t(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J})\ \bar{\Sigma}\ \nabla\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J})}} \xrightarrow{d} \mathcal{N}(0,1)$$

with $\bar{\Sigma} = \frac{1}{n} \cdot diag(\bar{X}^2, \bar{Y}^2, \bar{Z}^2, \bar{U}^2, \bar{H}^2, \bar{J}(1-\bar{J}))$. In other words we have

$$\mathbb{P}\left[n_{\frac{\alpha}{2}} < \frac{\left(\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J}) - \phi(\lambda, \mu, \alpha, \beta, \tau, p)\right)}{\sqrt{\nabla\hat{\phi}^t(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J})\ \bar{\Sigma}\ \nabla\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J})}} < -n_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

where $n_{\frac{\alpha}{2}}$ is obtained from Normal tables. This implies the following $100(1-\alpha)\%$ asymptotic confidence interval:

$$\begin{aligned}&\phi(\lambda, \mu, \alpha, \beta, \tau, p) \in \\ &\Big[\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J}) + n_{\frac{\alpha}{2}}\ \sqrt{\nabla\hat{\phi}^t(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J})\ \bar{\Sigma}\ \nabla\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J})}, \\ &\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J}) - n_{\frac{\alpha}{2}}\ \sqrt{\nabla\hat{\phi}^t(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J})\ \bar{\Sigma}\ \nabla\hat{\phi}(\bar{X}, \bar{Y}, \bar{Z}, \bar{U}, \bar{H}, \bar{J})}\Big].\end{aligned}$$

5.5 Numerical Examples

The following numerical examples show the performance improvement due to introduced optimal threshold $\hat{q}_r^*$. We use the general representation of the confidence interval derived above in order to calculate confidence intervals for different performance measures, e.g. the mean number of customers in the system $\bar{N}$ or the average cost function g.

In order to get some numerical results we simulate our system with parameters fixed in the following way:

$$\lambda = 1.0,\ \mu = 2.0,\ \alpha = 0.2,\ \beta = 5.0,\ \tau = 10.0,\ p = 0.5.$$

To perform a simulation we use the probability of an initial state of system which is calculated by (6). The samples $(X_1, X_2, \ldots, X_n)$, $(Y_1, Y_2, \ldots, Y_n)$, $(Z_1, Z_2, \ldots, Z_n)$,$(U_1, U_2, \ldots, U_n)$ and $(H_1, H_2, \ldots, H_n)$ are drawn as described in Subsect. 5.1. The costs are: $c_0 = 0.1$, $c_{0,0} = 0.5$, $c_{1,0} = 0.5$, $c_{2,0} = 0.5$, $c_{0,1} = 2.0$, $c_{1,1} = 2.0$, $c_{2,1} = 2.0$.

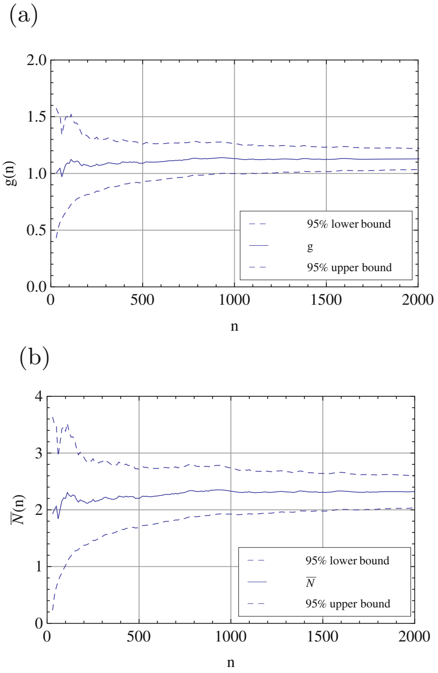

Fig. 3. Confidence intervals for g (a) and $\bar{N}$ (b)

Figure 3(a) illustrates the confidence intervals for the long-run average cost $g(\hat{q}_r^*)$ versus sample size n. Take note that exact value q_r^* as well as the estimation $\hat{q}_r^*$ equal to 3 and the exact value of the cost function $g(3) = 1.107$. For the repair threshold $q_r = 1$ we have $g(1) = 2.14$. Thus, an optimized threshold considerably reduces the system costs. The costs estimation becomes stable for $n \geq 500$.

On the other hand, the confidence intervals for the mean number of customers in the system $\bar{N}(\hat{q}^*)$ versus n are illustrated in Fig. 3(b). The exact value $\bar{N}(q_r^*) = 2.25$ for $q_r^* = 3$. For the threshold level $q_r = 1$ we have $\bar{N}(1) = 1.42$. Thus, the average number of customers in the system increases with q_r. The estimation becomes stable for $n \geq 500$.

References

1. Aissani, A., Artalejo, J.: On the single server retrial queue subject to breakdowns. Queueing Syst. **30**(3–4), 309–321 (1998)
2. Artalejo, J., Gomez-Corral, A., Neuts, M.F.: Analysis of multiserver queues with constant retrial rate. Eur. J. Oper. Res. **135**, 569–581 (2001)
3. Choi, B.D., Rhee, K.H., Park, K.K.: The $M/G/1$ retrial queue with retrial rate control policy. Probab. Eng. Inf. Sci. **7**, 29–46 (1993)
4. Choi, B.D., Shin, Y.W., Ahn, W.C.: Retrial queues with collision arising from unslotted CSMA/CD protocol. Queueing Syst. **11**, 335–356 (1955)
5. Efrosinin, D., Semenova, O.: An M/M/1 system with an unreliable device and threshold recovery policy. J. Commun. Technol. Electron. **55**(12), 1526–1531 (2010)
6. Efrosinin, D., Sztrik, J.: Performance analysis of a two server heterogeneous retrial queue with threshold policy. Qual. Technol. Quant. Manag. **8**, 211–236 (2011)
7. Efrosinin, D., Winkler, A.: Queueing system with a constant retrial rate, non-reliable server and threshold-based recovery. Eur. J. Oper. Res. **210**, 594–605 (2011)
8. Fayolle, G.: A simple telephone exchange with delayed feedbacks. In: Cohen, J.W., Tijms, M.C. (eds.) Teletraffic Analysis and Computer Performance Evaluation OJ Boxma, pp. 245–253. North-Holland, Amsterdam (1986)
9. Li, W., Zhao, Y.Q.: A retrial queue with a constant retrial rate, server downs and impatient customers. Stoch. Models **21**, 531–550 (2005)
10. Martin, M., Artalejo, J.: Analysis of an $M/G/1$ queue with two types of impatient units. Adv. Appl. Probab. **27**, 840–861 (1995)
11. Neuts, M.F.: Matrix-geometric Solutions in Stochastic Models. The John Hopkins University Press, Baltimore (1981)
12. Ross, J.V., Taimre, T., Pollett, P.K.: Estimation for queues from queue length data. Queueing Syst. **55**, 131–138 (2007)
13. Wang, T.Y., Ke, J.C., Wang, K.H., Ho, S.C.: Maximum likelihood estimates and confidence intervals of an $M/M/R$ queue with heterogeneous servers. Math. Methods Oper. Res. **63**, 371–384 (2006)
14. Yadavalli, V.S.S., Adendorff, K., Erasmus, G., Chandrasekhar, P., Deepa, S.P.: Confidence limits for expected waiting time of two queueing models. Oper. Res. Soc. S. Afr. **20**(1), 1–6 (2004)