11-2022

A Retrieval Queueing Model With Feedback

Agassi Melikov

Srinivas Chakravarthy

Sevinc Aliyeva

Queueing Models and
Service Management
Vol. 6, No. 1, page 63-95, 2023

QMSM
©AU 2023

A Retrial Queueing Model with Feedback

Agassi Melikov[1,*], Srinivas R. Chakravarthy[2] and Sevinc Aliyeva[3]

[1]Department of Teletraffic Theory
Institute of Control System, Baku AZ1141, Azerbaijan
[2]Departments of Industrial and Manufacturing Engineering & Mathematics
Kettering University, Flint, MI-48504, USA
[3] Departments of Information Technologies and Programming
Baku State University, Baku AZ1148, Azerbaijan

(*Received May 2022; accepted November 2022*)

Abstract: A multi-server retrial queuing model with feedback is considered in this paper. Input flow of calls is modeled using a Markovian Arrival Process (MAP) and the service time is assumed to follow an exponential distribution. An arriving call enters into service should there be a free server. Otherwise, in accordance to Bernoulli trials, the call will enter into an infinite orbit (referred to as a *retrial* orbit) to retry along with other calls to get into service or will leave the system forever. After obtaining a service each call, independent of the others, will either enter into a finite orbit (referred to as a *feedback* orbit) for another service or leave the system forever. The decision to enter into the feedback orbit or not is done according to another Bernoulli trial. Calls from these two buffers will compete with the main source of calls based on signals received from two independent Poisson processes. The rates of these processes depend on the phase of the MAP. The steady-state analysis of the model is carried out and illustrative numerical examples including economical aspects are presented.

Keywords: queueing, retrial queue, delayed feedback, signals generation of orbital calls, MAP flow

1. Introduction and literature review

In classical retrial queueing models, the calls (or customers) will leave the system upon receiving a service. However, in practice we encounter retrial queues in which the calls need to be served again through a feedback mechanism. For example, in multiple access telecommunication systems, the calls that are erroneously transmitted (or served) need to be retransmitted. Such calls when finding all servers busy at the time of their arrivals usually enter into a retrial orbit and compete for a free server. Hence, using retrial queues with

Corresponding author
Email: agassi.melikov@gmail.com

feedback ($RQwFB$) will be more appropriate. Similar situations occur in widely used client-server systems as well as in queuing-inventory systems. In the latter case after ordering the inventory, customers may have to return to the system to receive the inventory.

In dealing with queuing systems with feedback ($QSwFB$) it is necessary to distinguish between two options to provide (re)services: instantaneous feedback (IFB) and delayed feedback (DFB). In the case of IFB, any call requiring another service (with a certain probability) will get attention immediately, whereas in DFB the call will enter into a buffer (usually a finite buffer) with a certain probability if it needs to be served again. In both the mechanisms with the corresponding complementary probability the calls will leave the system. Such kinds of feedbacks are referred to Bernoulli feedback in the literature. Here, we consider only models with Bernoulli feedback.

In the last three decades, $QSwFB$ of both feedback types have been extensively studied in the literature. First, we look at the works dealing with $QSwIFB$. In the pioneer work of Takacs (1963), the feedback model $M/G/1$ was studied using probability generating functions (PGF). The same model $M/G/1$ with vacations of server was investigated in Wortman, Disney and Kiessler (1991). In D'Avignon and Disney (1977), the model $M/G/1$ with two independent Poisson traffics and with non-preemptive priorities is examined. Berg and Boxma (1991) studied model $M/M/1$ with general feedback mechanism and it is shown that under some conditions the given model approaches the $M/G/1$ with processor sharing queue discipline. In Hunter (1989) the Laplace-Stieltjes transforms (LST) of the sojourn time distribution for the model $M/G/1/N$ with state-dependent feedback is derived and the difficulty to invert these LST except for the cases when $N = 1$ and $N = 2$ is pointed out. In the context of $MAP/PH/1/N$ Dudin et al (2005), incorporated state-dependent departure and feedback probabilities in random environment. They applied Neuts' (Neuts (1981)) matrix-analytic method (MAM) to study the model in steady state. A similar model is considered in Krieger et. al. (2005). Assuming that there are two Poisson flows to a single server system in which there are two separate (infinite) buffers to hold each type of p-calls and only one type calls can make feedback requests, Krishnamoorthy and Manjunath (2018) analyze the model in steady-state using MAM. It is assumed that feedback calls (f-calls) are sent to the queue of low priority calls. In Bouchentouf, Cherfaoui and Boualem (2019) the feedback model M/M/1/N with server vacation, balking, reneging and retention of reneged calls is analyzed. The steady-state probabilities for the number of calls in the system when the server is in busy and vacation periods are derived through a recursive method. Similar but a multi-server model is investigated in Bouchentouf, Cherfaoui and Boualem (2020). A single server Markovian system with group arrivals, multiple vacation of server, impatient calls and retention of reneged calls is examined in Bouchentouf and Guendouzi (2021).

Note that, in all of the above mentioned works it is assumed that primary calls (p-calls) and f-calls have the same channel holding times. Further, in these works it is assumed that the departure and feedback probabilities are governed by Bernoulli trials with constant parameters. Melikov, Ponomarenko and Kuliyeva (2015a) consider an un-buffered multi-channel system with different channel occupancy times for p-calls and f-calls along with state-dependent Bernoulli probabilities for calls' departures and feedback probabilities. The model of single-channel and buffered system with $MMPP$ flow of p-calls is investigated in Melikov and Aliyeva (2019). The model of single-server $QSwIFB$ with an exponential server switching time (to serve f-calls) are investigated in Melikov, Aliyeva and Shahmaliyev (2020). In Melikov, Aliyeva and Sztrik (2020) models of $QSwIFB$ with two heterogeneous servers are investigated. Note that in the works co-authored by Melikov, the space merging method was used to find the approximate values of steady-state probabilities (see Ponomarenko, Kim and Melikov (2010) and Melikov, Ponomarenko and Sztrik (2016)). In a recent paper, Ayyapan and Thilagavathy (2021) look at $MAP/PH/1$ with unreliable main server and a standby server who is serving at a lower rate under IFB scenario.

Now we look at the works involving $QSwDFB$ models. Takacs (1977) also did the first work in this direction. Using PGF approach, the author obtains expressions for the mean queue length, mean waiting time in the queue, and the mean sojourn time in the system. Pekoz and Joglekar (2002) examine a multi-server queueing model in which the arrivals occur according to a stationary ergodic flow process, general services, a finite-capacity for p-calls and an infinite orbit for f-calls, where the number of feedbacks for each call is assumed to be random. In Lee and Seo (1997), the model $M/G/1$ with finite orbit is considered where it is assumed that, when the orbit becomes full, all calls from the orbit will instantaneously fed back to the buffer of the queuing system. A similar model is examined in Lee and Ahn (2000) but with the assumption that the feedback calls require a random time to move from the orbit to the service area. Foley and Disney (1983) study an $M/G/1$ queue with infinite buffer with the assumption that the feedback calls spend an exponential amount of time in the orbit. The model of the type $M/M/n$ with finite buffer and exponential sojourn time in the orbit by the calls is investigated in Melikov, Ponomarenko and Kuliyeva (2015b). Here, the feedback probabilities depend on the number of busy channels in the system at the call departure epochs. The theory and applications of RQ can be found in the books of Artalejo and Gomez-Corral (2008) and Falin and Templeton (1997), and in the review papers by Kim and Kim (2016) and Phung-Duc (2019). In the literature, there are a few works devoted to retrial queues with delayed feedback (RQwDFB). In this paper, we focus on one such RQwDFB model. In such systems, orbit(s) are not only for p-calls, but also for the f-calls. It is important to note that often retrial calls (r-calls) generated by p-calls differ from calls generated by f-calls in some way, for example, in importance, the cost of losing them, waiting in orbit, and so on (see Section 5 as well). Therefore, in this paper, we consider separate orbits for r-calls and f-calls, especially since one of the orbits has a finite size, and the other has an infinite size.

To the best of our knowledge, the first work on $RQwDFB$ was by Choi, Kim and Lee (1998). In this paper, the authors consider $M/M/c/c$ with infinite orbit. The joint PGF of the number of busy servers and the number of calls in the retrial group is obtained for the cases $c = 1, 2$. Krishna Kumar, Rukmani and Thangaraj (2009) considered an $M/M/c/N + c$ queue with constant retrial rate. The authors employ Neuts' matrix-analytic method (MAM) to study the model. The same model was investigated by Do (2010) and applied the spectral expansion method of Mitrani I., Chakka R. (1995) to calculate the steady-state probabilities. Ayyapan, Subramanian and Sekar considered an $M/M/1$ queue with loss under non-preemptive (2010a) and preemptive (2010b) priority service disciplines. Two types of calls arrive and retrial, loss and feedback mechanisms are allowed only for low priority calls only. In both the papers, the authors use MAM to analyze the models.

Mokaddis, Metwall and Zaki (2007) considered $RQwDFB$ of the type $M/G/1/1$ with constant retrial rate and single vacation where the server is subject to failures and repairs. In Lee (2005), a single server system in which two types of calls arrive according to independent Poisson flows is considered. An arriving priority call finding a free server immediately starts getting a service; otherwise, the call will join the queue. An arriving non-priority call finding a free server will get into service immediately; otherwise, the call will enter into a retrial orbit.

In Melikov, Aliyeva and Sztrik (2019) $MMPP/M/K/K$ with DFB model is considered. $M/G/1$-type $RQwDFB$ queueing models with negative calls under working vacation and working breakdowns are studied in Rajadurai, Sundararaman and Narasimhan (2020). The PGF of the numbers of calls in the orbit under different server status are derived.

Recently Dimitriou and Phung-Duc (2018) considered the Markovian single-server system with two separated queues for retrial and feedback calls. They obtain the stability conditions and uses the generating function technique to calculate the joint queue length distribution.

Note that the works indicated above are devoted to single-server $RQwDFB$ models with the exceptions of Krishna Kumar, Rukmani and Thangaraj (2009), Do (1998) and Melikov, Aliyeva and Sztrik (2019). Further, in these works with the exception of Melikov, Aliyeva and Sztrik (2019)), it is assumed that the retrial and feedback probabilities are constant. These assumptions significantly limit the applications of such models in practice since in practice $RQwDFB$ models will have many servers and further the decisions to enter the orbit or leave the system without getting additional services often depend on the current state of the system.

In all of the above mentioned works, queueing models with instantaneous and delayed feedbacks are investigated separately. Models with both instantaneous and delayed feedbacks are investigated in Melikov, Ponomarenko and Rustamov (2015) and Dudin and Dudina (2019). In Dudin and Dudina (2019) the model of RQ with both instantaneous and delayed feedback of the type $MAP/PH/1$ with unreliable transmission of calls is considered. Here it is assumed that feedback phenomena can occur not after the completion of the service of the call, but as a result of a failure in the period of transmission of the call. Orbit is common for retrial and feedback calls.

It is worth pointing out that in all the works devoted to $RQwDFB$, r-calls that are generated by p-calls and r-calls that are generated by f-calls are not distinguished. However, the calls that have not received any service should be different from the calls that have received at least some service before leaving the system. The related to our study papers have been summarized according to their main assumptions and presented in Table 1. This table clearly illustrate the contribution of our study. In addition to this table note that the main differences between the $RQwDFB$ model studied here and the models considered in literature are as follows:

1. We consider a multi-server $RQwDFB$ with a versatile point process, namely, Markovian arrival process (MAP) to model the arrivals of the primary calls.
2. We consider separate orbits for retrial and feedback calls so as to distinguish between these types of calls.
3. The retrial rates from the orbit as well as from the feedback group depend on the phase of the MAP.

In the following table, we use the following abbreviations: AP - Arrival Process; SP - Service Process; NoS - Number of Servers; SR - Server Reliability (R - Reliable, U - Unreliable); SV - Server vacation; BS - Buffer Size (F - Finite, I - Infinite); $S\&SO$ - Size and Structure of Orbits for r- and f-calls (C = Common orbit; S = Separate); NCT - Number of p-Calls Types; RP - Retrial process (LRR - Linear Retrial Rate; CRR - Constant Retrial Rate.

Table 1. Comparisons between the published $RQwDFB$ models

Reference	AP	SP	NoS	SR	SV	BS	$S\&SO$	NCT	RP
[2]	Poisson	Exp	1	R	No	F	I, C	2	Exp, LRR
[3]	Poisson	Exp	1	R	No	F	I, C	2	Exp, LRR
[11]	Poisson	Exp	1 & 2	R	No	0	I, C	1	Exp, LRR
[13]	Poisson	Exp	multiple	R	No	F	I, C	1	Exp, CRR
[22]	Poisson	Exp	multiple	R	No	F	I, C	1	Exp, CRR
[24]	Poisson	G	1	R	Yes	I	I, C	2	Exp, LRR
[33]	$MMPP$	Exp	multiple	R	No	0	I, C	1	Exp, LRR
[39]	Poisson	G	1	U	Yes	0	I, C	1	G
[46]	Poisson	G	1	U	Yes	I	I, C	1†	G, CRR
Here	MAP	Exp	multiple	R	No	0	I, S, F	1	Signal

†: with negative calls

It is clear that, taking into account the above listed assumptions will increase the utility of the model in practice.

For use in sequel, we register a number of notation. By $\boldsymbol{e}$, we denote a column vector of appropriate dimension (which will be clear in the context) consisting of 1's. By $\Delta(\boldsymbol{a})$ we denote a diagonal matrix whose diagonal elements are given by the elements of the vector $\boldsymbol{a}$.

By $\hat{I}_r$ we denote a square matrix of order r such that the only non-zero number is 1 and it occurs in the last (namely, r^{th}) diagonal position.

By $\tilde{I}_r$ we denote the following square matrix of order r.

$$\tilde{I}_r = \begin{pmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & 0 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & 0 & 1 \\ & & & & & 0 \end{pmatrix}. \tag{1}$$

The Kronecker product of A and B, denoted by $A \otimes B$, is a matrix of dimension $mp \times nq$ and is given by

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{12}B & \cdots & a_{1n}B \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}B & a_{n2}B & \cdots & a_{mn}B \end{pmatrix}.$$

The Kronecker sum, denoted by, $A \oplus B$, is defined as

$$A \oplus B = A \otimes I_n + I_m \otimes B,$$

where I_r is an identity matrix of dimension r.

The paper is organized as follows. In Section 2, the model under study is described and the steady-state analysis of the model is performed in Section 3. Illustrative numerical are presented in Section 4. Some concluding remarks are given in Section 5.

2. Model assumptions

We assume that calls arrive according to a MAP with parameter matrices (D_0, D_1) of order m. These calls are referred to as *primary calls* or simply $p-$calls. It is well-known that MAP, a versatile Markovian point process introduced by Neuts (1979), enables one to model even correlated arrivals into the arrival process and generalizes some of the classical ones such as Poisson process, Markov modulated Poisson process, $PH-$ renewal process among others.

A (continuous-time) MAP is characterized by parameter matrices (D_0, D_1) such that $D = D_0 + D_1$ is an irreducible generator of the underlying continuous-time Markov chain governing the MAP. While, transitions within D_0 correspond to no arrivals to the system, the transitions governed by D_1 are those of the arrivals to the system. Suppose that $\boldsymbol{\delta}$ is the invariant vector of D. That is,

$$\boldsymbol{\delta} D = \mathbf{0}, \ \ \boldsymbol{\delta}\, \boldsymbol{e} = 1. \tag{2}$$

The arrival rate, λ, is given by $\lambda = \boldsymbol{\delta} D_1 \boldsymbol{e}$.

We refer the reader to, for example, Chakravarthy (2020), He (2014), Lucantoni, et.al. (1990), Lucantoni (1991), and Neuts (1989), for more details on Markovian arrival processes and their applications.

The $p-$calls are processed by one of c servers in the system. The service times of the calls are exponential and are independent of the type of calls served. Let the service rate be denoted by μ.

An arriving $p-$call finding all servers busy will be lost with probability α; with probability $1 - \alpha$ will enter into an orbit, denoted as O_r, of infinite size, and try to capture a free server through a retrial mechanism. Calls coming out of orbit O_r are referred as $r-$calls.

A processed $p-$call leaves the system with probability β; or with probability $1-\beta$ needs to be served again. Such calls are referred to as *feedback calls* or simply $f-$call. These calls enter into an orbit, say, O_f, of finite size, say, N, and try to capture a free server through a retrial mechanism. If this buffer is full when a feedback call needs to get processed, it will be considered lost.

Calls in orbits O_r and O_f enter into service only through the signals arriving at their respective orbits. That is, at the time the signals reach O_r (or O_f) and if there is at least one call waiting in the orbit with at least one free server, the signals will be considered successful and one of the waiting calls will enter into service. The signals are independent of the numbers in the orbits. Since we are dealing with continuous time, only one call (either from O_r or O_f) will be successful and the signals are transferrable in the sense that if there is no customer in one orbit at the time of receiving a signal, a call from another orbit cannot enter into service. Thus, each orbit receives their signals to let one of the waiting calls to enter into service. If there are no calls waiting or no free server, then the signals will be unsuccessful.

The signals are generated according to two Poisson processes, one for $r-$calls and one for $f-$ calls. The rates of these two processes depend on the phase of the MAP. Thus, the signals to O_r occur according to a Poisson process with rate $\theta_j^{(r)}, 1 \leq j \leq m,$ and the signals to O_f occur according to a Poisson process with rate $\theta_j^{(f)}, 1 \leq j \leq m.$ This idea of using signals was recently introduced by Chakravarthy (2021) as a way to provide a reasonable fair treatment to all orbiting customers seeking service. Further, this allows one to study retrial queueing models as level-independent queues.

A pictorial description of the model under study is displayed in the Figure 1 below.

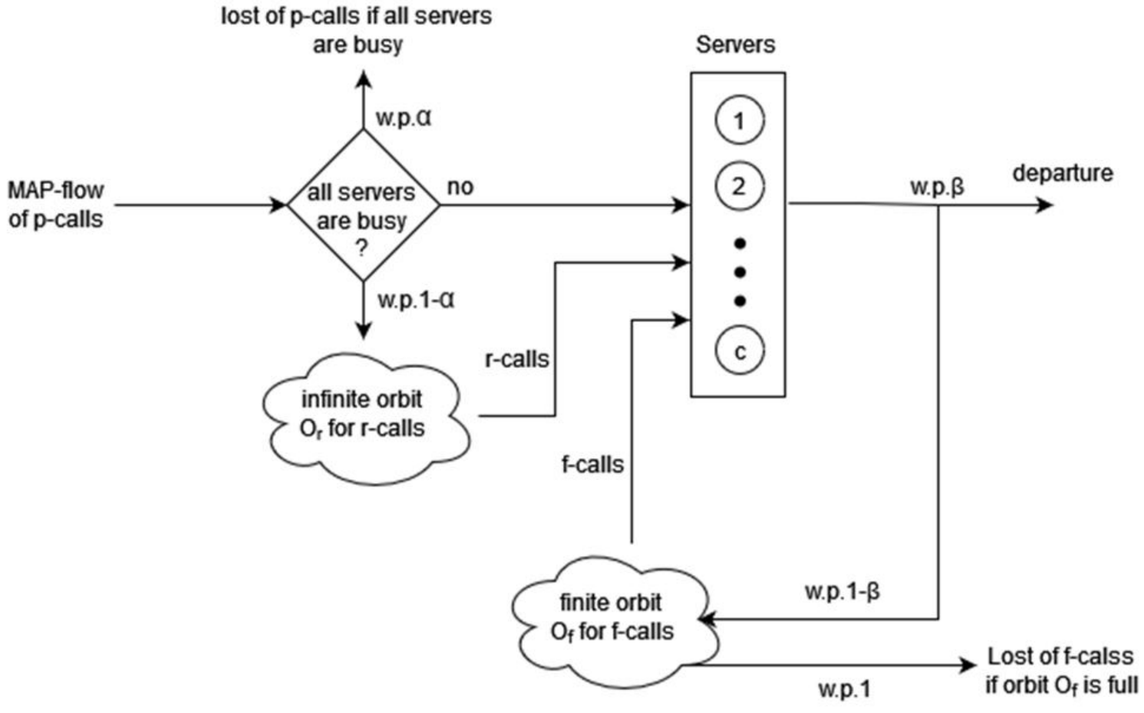
lost of p-calls if all servers are busy
w.p.α
MAP-flow of p-calls
all servers are busy ?
no
w.p.1-α
infinite orbit O_r for r-calls
r-calls
Servers
1
2
c
w.p.β
departure
f-calls
finite orbit O_f for f-calls
w.p.1-β
w.p.1
Lost of f-calss if orbit O_f is full

Figure 1. Pictorial description of the model under study

By keeping track of

1. $N_1(t)$ = Number of calls in O_r
2. $N_2(t)$ = Number of calls in O_f
3. $J(t)$ = Number of busy servers
4. $K(t)$ = Phase of the MAP

we can study the model as a continuous-time Markov process.

The process $\{(N_1(t), N_2(t), J(t), K(t)) : t \geq 0\}$ can be verified to be a Markov process with state space given by

$$\Omega = \{(i_1, i_2, j, k) : i_1 \geq 0, 0 \leq i_2 \leq N, 0 \leq j \leq c,\ 1 \leq k \leq m\}.$$

Let $\boldsymbol{i}_1$ denote the level consisting of the states $\{(i_1, i_2, j, k) : 0 \leq i_2 \leq N, 0 \leq j \leq c,\ 1 \leq k \leq m\}$, for $i_1 \geq 0$. Note that this level consists of $(N+1)(c+1)m$ states.

3. Steady-state analysis

The model described in Section 2 can be studied as a $QBD-$ process with infinitesimal generator Q given by

$$Q = \begin{pmatrix} B_0 & A_0 & & & \\ A_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & A_2 & A_1 & A_0 \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{3}$$

where the matrices B_0, A_0, A_1, and A_2 are as given below in (block) partitioned form.

$$B_0 = \begin{pmatrix} \tilde{B}_1 & \tilde{B}_0 & & \\ \tilde{B}_2 & \tilde{B}_1 & \tilde{B}_0 & \\ & \tilde{B}_2 & \tilde{B}_1 & \tilde{B}_0 \\ & & \ddots & \ddots & \ddots \\ & & & \tilde{B}_2 & \tilde{B}_1 + \tilde{B}_0 \end{pmatrix}, \tag{4}$$

with

$$\tilde{B}_1 = \begin{pmatrix} D_0 & D_1 & & & \\ \beta\mu I & D_0 - \mu I & D_1 & & \\ & 2\beta\mu I & D_0 - 2\mu I & D_1 & \\ & & 3\beta\mu I & D_0 - 3\mu I & D_1 \\ & & \ddots & \ddots & \ddots \\ & & & c\beta\mu I & D_0 + \alpha D_1 - c\mu I \end{pmatrix}, \tag{5}$$

$$\tilde{B}_0 = \begin{pmatrix} 0 & & & & & \\ (1-\beta)\mu I & 0 & & & & \\ & 2(1-\beta)\mu I & 0 & & & \\ & & 3(1-\beta)\mu I & 0 & & \\ & & & \ddots & & \\ & & & & c(1-\beta)\mu I & 0 \end{pmatrix}, \tag{6}$$

$$\tilde{B}_2 = \begin{pmatrix} 0 & \Delta(\boldsymbol{\theta}^{(F)}) & & & \\ & 0 & \Delta(\boldsymbol{\theta}^{(F)}) & & \\ & & \ddots & & \\ & & & 0 & \Delta(\boldsymbol{\theta}^{(F)}) \\ & & & & 0 \end{pmatrix}, \tag{7}$$

$$\begin{aligned} A_0 &= (1-\alpha)\big(I_{N+1} \otimes \hat{I}_{c+1} \otimes D_1\big), \\ A_1 &= B_0 - \big(I \otimes [(I_{c+1} - \hat{I}_c(c+1))] \otimes \Delta(\boldsymbol{\theta}^{(R)})\big), \\ A_2 &= I_{N+1} \otimes \tilde{I}_{c+1} \otimes \Delta(\boldsymbol{\theta}^{(R)}), \end{aligned} \tag{8}$$

$$\boldsymbol{\theta}^{(R)} = (\theta_1^{(r)}, \cdots, \theta_m^{(r)}), \;\; \boldsymbol{\theta}^{(F)} = (\theta_1^{(f)}, \cdots, \theta_m^{(f)}).$$

Suppose that $A = A_0 + A_1 + A_2$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \cdots, \boldsymbol{\pi}_N)$ is the invariant vector of A. That is,

$$\boldsymbol{\pi} A = \mathbf{0} \;\texttt{and}\; \boldsymbol{\pi}\boldsymbol{e} = 1. \tag{9}$$

First, verify that the matrix A is of the form

$$A = \begin{pmatrix} \tilde{C}_1 & \tilde{B}_0 & & & & \\ \tilde{B}_2 & C_1 & \tilde{B}_0 & & & \\ & \tilde{B}_2 & C_1 & \tilde{B}_0 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \tilde{B}_2 & C_1 & \tilde{B}_0 \\ & & & & \tilde{B}_2 & C_1 + \tilde{B}_0 \end{pmatrix}, \;\texttt{with}\; C_1 = \tilde{C}_1 - (I - \hat{I}_{c+1}) \otimes \Delta(\boldsymbol{\theta}^{(F)}), \tag{10}$$

and

$$\tilde{C}_1 = \begin{pmatrix} D_0 - \Delta(\boldsymbol{\theta}^{(R)}) & D_1 + \Delta(\boldsymbol{\theta}^{(R)}) & & & & \\ \beta\mu I & D_0 - \Delta(\boldsymbol{\theta}^{(R)}) - \mu I & D_1 + \Delta(\boldsymbol{\theta}^{(R)}) & & & \\ & 2\beta\mu I & D_0 - \Delta(\boldsymbol{\theta}^{(R)}) - 2\mu I & D_1 + \Delta(\boldsymbol{\theta}^{(R)}) & & \\ & & \ddots & \ddots & \ddots & \\ & & & (c-1)\beta\mu I & D_0 - \Delta(\boldsymbol{\theta}^{(R)}) - (c-1)\mu I & D_1 + \Delta(\boldsymbol{\theta}^{(R)}) \\ & & & & c\beta\mu I & D_0 + D_1 - c\mu I \end{pmatrix}. \tag{11}$$

From the steady-state equations related to the vector $\boldsymbol{\pi}$ of A, which in partitioned form $\boldsymbol{\pi} = \boldsymbol{\pi}_{0,0}, \boldsymbol{\pi}_{0,1}, \cdots, \boldsymbol{\pi}_{N,c-1}, \boldsymbol{\pi}_{N,c}$, we get the following results.

Result 1. We have

$$\sum_{i=0}^{N}\sum_{j=0}^{c} \boldsymbol{\pi}_{i,j} = \boldsymbol{\delta}, \tag{12}$$

where $\boldsymbol{\delta}$ is as given in Equation (2).

Result 2. We have

$$\sum_{i=1}^{N}\sum_{j=0}^{c-1} \boldsymbol{\pi}_{i,j}\Delta(\boldsymbol{\theta}^{(F)})\boldsymbol{e} = (1-\beta)\mu\sum_{i=0}^{N-1}\sum_{j=1}^{c} j\,\boldsymbol{\pi}_{i,j}\boldsymbol{e}. \tag{13}$$

Using the classical result on $QBD-$ process (see, Neuts(1981)), the stability condition, namely, $\boldsymbol{\pi}A_0\boldsymbol{e} < \boldsymbol{\pi}A_2\boldsymbol{e}$, for our model reduces to the one given in the following result.

Result 3. The stability condition for our model under study is

$$(1-\alpha)\sum_{i=0}^{N} \boldsymbol{\pi}_{i,c}D_1\boldsymbol{e} < \sum_{i=0}^{N}\sum_{j=0}^{c-1} \boldsymbol{\pi}_{i,j}\Delta(\boldsymbol{\theta}^{(R)})\boldsymbol{e}. \tag{14}$$

Note that in equation (14) the left-hand side corresponds to $\boldsymbol{\pi}A_0\boldsymbol{e}$ and that the right-hand side is $\boldsymbol{\pi}A_2\boldsymbol{e}$. Also, the stability condition can be intuitively explained as follows noting that orbit O_r is of infinite size and the orbit O_f is of finite size. The left-hand side of equation (14) corresponds to the rate of customers getting into orbit O_r while the right-hand side of the same equation gives the rate of O_r customers getting out of O_r. Obviously, input rate should be less than the output rate. For use in sequel, we define the traffic intensity, ρ, as

$$\rho = \frac{\boldsymbol{\pi}A_0\boldsymbol{e}}{\boldsymbol{\pi}A_2\boldsymbol{e}}.$$

3.1. Steady-state vector

Since we are dealing with $QBD-$ process, the steady-state vector, $\boldsymbol{x}$ is of matrix-geometric type (see, Neuts(1981)) under the stability condition that $\rho < 1$ (or equivalently the condition given in Equation (14) and is given by

$$\boldsymbol{x}_i = \boldsymbol{x}_0R^i,\ i \geq 0, \tag{15}$$

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

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

and the vector $\boldsymbol{x}_0$ is obtained by solving the following system of equations:

$$\boldsymbol{x}_0[B_0 + RA_2] = \mathbf{0} \ \texttt{and}\ \boldsymbol{x}_0(I-R)^{-1}\boldsymbol{e} = 1. \tag{17}$$

While one can use the logarithmic-reduction algorithm of Latouche and Ramaswami(1999) for computing R when the dimension of R is manageable, it is strongly advised to exploit the structure of the sparsity of the coefficient matrices. First note that the sparsity of A_0 makes the structure of R to be of the form:

$$R = \begin{pmatrix} R_{0,0} & R_{0,1} & \cdots & R_{0,N} \\ R_{1,0} & R_{1,1} & \cdots & R_{1,N} \\ \vdots & \vdots & \cdots & \vdots \\ R_{N,0} & R_{N,1} & \cdots & R_{N,N} \end{pmatrix}, \quad R_{i,j} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ R_0^{(i,j)} & R_1^{(i,j)} & \cdots & R_c^{(i,j)} \end{pmatrix}, \tag{18}$$

where $R_r^{(i,j)},\ 0 \le i, j \le N,\ 0 \le r \le c$, are matrices of dimension m. We need to exploit it to solve Equation (16). The details are as follows.

Simplifications: The following simplifications can easily be verified.

1. The matrix, $RA_2 = \{R_{i,j}(\tilde{I}_{c+1} \otimes \Delta(\boldsymbol{\theta}^{(R)}))\}$, is such that the non-zero blocks (which are the last row blocks in $(c+1)^{st}$ block) in $R_{i,j}(\tilde{I}_{c+1} \otimes \Delta(\boldsymbol{\theta}^{(R)}))$ are given by

$$R_{i,j}(\tilde{I}_{c+1} \otimes \Delta(\boldsymbol{\theta}^{(R)})) = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & R_0^{(i,j)}\Delta(\boldsymbol{\theta}^{(R)}) & R_1^{(i,j)}\Delta(\boldsymbol{\theta}^{(R)}) & \cdots & R_{c-1}^{(i,j)}\Delta(\boldsymbol{\theta}^{(R)}) \end{pmatrix}.$$

2. The matrix $R^2A_2 = \{S_{i,j}\}$ is such that the non-zero blocks (which are the last row blocks in $(c+1)^{st}$ block) in $S_{i,j}$ are given by

$$S_{i,j} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & \sum\limits_{k=0}^{N} R_c^{(i,k)}R_0^{(k,j)}\Delta(\boldsymbol{\theta}^{(R)}) & \sum\limits_{k=0}^{N} R_c^{(i,k)}R_1^{(k,j)}\Delta(\boldsymbol{\theta}^{(R)}) & \cdots & \sum\limits_{k=0}^{N} R_c^{(i,k)}R_{c-1}^{(k,j)}\Delta(\boldsymbol{\theta}^{(R)}) \end{pmatrix}.$$

3. Suppose that $T = RA_1 = \{T_{i,j}\}$. Then using the structure of R and A_1, we first note that $T_{i,j}$ is of the form by

$$T_{i,j} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ T_0^{(i,j)} & T_1^{(i,j)} & \cdots & T_c^{(i,j)} \end{pmatrix}.$$

The non-zero blocks of T are as follows.

$$T_0^{(i,0)} = R_0^{(i,0)}(D_0 - \Delta(\boldsymbol{\theta}^{(R)})) + \beta\mu R_1^{(i,0)},$$

For $1 \le i \le N,\ 1 \le r \le c-1$,

$$T_r^{(i,0)} = R_{r-1}^{(i,0)}D_1 + R_r^{(i,0)}(D_0 - r\mu I - \Delta(\boldsymbol{\theta}^{(R)})) + (r+1)\beta\mu R_{r+1}^{(i,0)} + R_{r-1}^{(i,1)}\Delta(\boldsymbol{\theta}^{(F)}),$$

$$T_c^{(i,0)} = R_{c-1}^{(i,0)} D_1 + R_c^{(i,0)}(D_0 - c\mu I + \alpha D_1) + R_{c-1}^{(i,1)} \Delta(\boldsymbol{\theta}^{(F)}),$$

For $1 \le i \le N,\ 1 \le j \le N-1,$

$$T_0^{(i,j)} = (1-\beta)\mu R_1^{(i,j-1)} + R_0^{(i,j)}(D_0 - \Delta(\boldsymbol{\theta}^{(R)}) - \Delta(\boldsymbol{\theta}^{(F)})) + \beta\mu R_1^{(i,j)},$$

For $1 \le i \le N,\ 1 \le j \le N-1,\ 1 \le r \le c-1,$

$$T_r^{(i,j)} = (r+1)(1-\beta)\mu R_{r+1}^{(i,j-1)} + R_{r-1}^{(i,j)} D_1 + R_r^{(i,j)}(D_0 - r\mu I - \Delta(\boldsymbol{\theta}^{(R)}) - \Delta(\boldsymbol{\theta}^{(F)}))$$

$$+(r+1)\beta\mu R_{r+1}^{(i,j)} + R_{r-1}^{(i,j+1)} \Delta(\boldsymbol{\theta}^{(F)}),$$

$$T_c^{(i,j)} = R_{c-1}^{(i,j)} D_1 + R_c^{(i,j)}(D_0 - c\mu I + \alpha D_1) + R_{c-1}^{(i,j+1)} \Delta(\boldsymbol{\theta}^{(F)}),$$

For $1 \le i \le N,$

$$T_0^{(i,j)} = (1-\beta)\mu R_1^{(i,j-1)} + R_0^{(i,j)}(D_0 - \Delta(\boldsymbol{\theta}^{(R)}) - \Delta(\boldsymbol{\theta}^{(F)})) + \beta\mu R_1^{(i,j)},$$

For $1 \le i \le N,\ 1 \le j \le N-1,\ 1 \le r \le c-1,$

$$T_r^{(i,j)} = r(1-\beta)\mu R_{r+1}^{(i,j-1)} + R_{r-1}^{(i,j)} D_1 + R_r^{(i,j)}(D_0 - r\mu I - \Delta(\boldsymbol{\theta}^{(R)}) - \Delta(\boldsymbol{\theta}^{(F)}))$$

$$+(r+1)\beta\mu R_{r+1}^{(i,j)} + R_{r-1}^{(i,j+1)} \Delta(\boldsymbol{\theta}^{(F)}),$$

$$T_c^{(i,j)} = R_{c-1}^{(i,j)} D_1 + R_c^{(i,j)}(D_0 - c\mu I + \alpha D_1) + R_{c-1}^{(i,j+1)} \Delta(\boldsymbol{\theta}^{(F)}),$$

$$T_0^{(i,N)} = (1-\beta)\mu R_1^{(i,N-1)} + R_0^{(i,N)}(D_0 - \Delta(\boldsymbol{\theta}^{(R)}) - \Delta(\boldsymbol{\theta}^{(F)})) + \mu R_1^{(i,N)},$$

For $1 \le i \le N,\ 1 \le r \le c-1,$

$$T_r^{(i,N)} = (r+1)(1-\beta)\mu R_{r+1}^{(i,N-1)} + R_{r-1}^{(i,N)} D_1 + R_r^{(i,N)}(D_0 - r\mu I - \Delta(\boldsymbol{\theta}^{(R)}) - \Delta(\boldsymbol{\theta}^{(F)}))$$
$$+(r+1)\beta\mu R_{r+1}^{(i,j)},$$

$$T_0^{(i,0)} = R_{c-1}^{(i,N)} D_1 + R_c^{(i,N)}(D_0 - c\mu I + \alpha D_1).$$

4. In terms of matrices of dimension m, the matrix R is computed in blocks and the needed equations are as follows.

$$R_0^{(i,0)} = \beta\mu R_1^{(i,0)} [\Delta(\boldsymbol{\theta}^{(R)}) - D_0]^{-1},\ 0 \le i \le N, \tag{19}$$

$$\begin{aligned} R_r^{(i,0)} = \Big[\sum_{k=0}^{N} R_c^{(i,k)} R_{c-1}^{(k,0)} \Delta(\boldsymbol{\theta}^{(R)}) + R_{r-1}^{(i,0)} D_1 + +(r+1)\beta\mu R_{r+1}^{(i,0)} \\ +R_{r-1}^{(i,1)} \Delta(\boldsymbol{\theta}^{(F)}) \Big] [r\mu I + \Delta(\boldsymbol{\theta}^{(R)}) - D_0]^{-1},\ 0 \le i \le N,\ 1 \le r \le c-1, \end{aligned} \tag{20}$$

$$\begin{aligned} R_c^{(i,0)} = \Big[\sum_{k=0}^{N} R_c^{(i,k)} R_{c-1}^{(k,0)} \Delta(\boldsymbol{\theta}^{(R)}) + R_{c-1}^{(i,0)} D_1 + R_{c-1}^{(i,1)} \Delta(\boldsymbol{\theta}^{(F)}) \\ +(1-\alpha) D_1 \Big] [c\mu I - D_0 - \alpha D_1]^{-1},,\ 0 \le i \le N, \end{aligned} \tag{21}$$

$$R_0^{(i,j)} = \Big[\beta\mu R_1^{(i,j)} + (1-\beta)\mu R_1^{(i,j-1)}\Big][\Delta(\boldsymbol{\theta}^{(R)}) + \Delta(\boldsymbol{\theta}^{(F)}) - D_0]^{-1}, \\ 0 \le i \le N,\ 1 \le j \le N-1, \tag{22}$$

$$R_r^{(i,j)} = \Big[\sum_{k=0}^{N} R_c^{(i,k)} R_{c-1}^{(k,j)}\Delta(\boldsymbol{\theta}^{(R)}) + (r+1)(1-\beta)\mu R_{r+1}^{(i,j-1)} + R_{r-1}^{(i,j)} D_1 \\ +(r+1)\beta\mu R_{r+1}^{(i,j)} + R_{r-1}^{(i,j+1)}\Delta(\boldsymbol{\theta}^{(F)})\Big][r\mu I + \Delta(\boldsymbol{\theta}^{(R)}) + \Delta(\boldsymbol{\theta}^{(F)}) - D_0]^{-1}, \\ 0 \le i \le N,\ 1 \le j \le N-1,\ 1 \le r \le c-1, \tag{23}$$

$$R_c^{(i,j)} = \Big[\sum_{k=0}^{N} R_c^{(i,c)} R_{c-1}^{(k,j)}\Delta(\boldsymbol{\theta}^{(R)}) + R_{c-1}^{(i,j)} D_1 + R_{c-1}^{(i,j+1)}\Delta(\boldsymbol{\theta}^{(F)})\Big][c\mu I - D_0 - \alpha D_1]^{-1}, \\ 0 \le i \le N,\ 1 \le j \le N-1, \tag{24}$$

$$R_0^{(i,N)} = \Big[\mu R_1^{(i,N)} + (1-\beta)\mu R_1^{(i,N-1)}\Big][\Delta(\boldsymbol{\theta}^{(R)}) + \Delta(\boldsymbol{\theta}^{(F)}) - D_0]^{-1}, \tag{25}$$

$$R_r^{(i,N)} = \Big[\sum_{k=0}^{N} R_c^{(i,k)} R_{c-1}^{(k,N)}\Delta(\boldsymbol{\theta}^{(R)}) + (r+1)\beta\mu R_{r+1}^{(i,j)} + R_{r-1}^{(i,N)} D_1 \\ +(r+1)(1-\beta)\mu R_{r+1}^{(i,N-1)}\Big][r\mu I + \Delta(\boldsymbol{\theta}^{(R)}) + \Delta(\boldsymbol{\theta}^{(F)}) - D_0]^{-1}, \\ 0 \le i \le N,\ 1 \le r \le c-1, \tag{26}$$

$$R_c^{(i,N)} = \Big[\sum_{k=0}^{N} R_c^{(i,k)} R_{c-1}^{(k,N)}\Delta(\boldsymbol{\theta}^{(R)}) + R_{c-1}^{(i,N)} D_1 \\ +(1-\alpha)D_1\Big][c\mu I - D_0 - \alpha D_1]^{-1},\ 0 \le i \le N. \tag{27}$$

Remarks: 1. Note that the recursive equations given in (19) through (27) are computed in that order until two successive iterates (element-by-element) are close enough (say to a pre-specified tolerance level, ϵ. Typically this level is chosen as $\epsilon = 10^{-9}$ or so.

2. It is worth pointing out that the computational complexity when using equation (16) versus equations (19) through (27) to find the rate matrix R. In the former case (without exploiting the structure of the rate matrix) the complexity is of the order $O((N+1)^3(c+1)^3m^3)$ per iterate, whereas when exploiting the structure of the coefficient matrices, this complexity reduces to $O(m^3)$ per iterate. [Note here we used the facts that (i) $O(km) = O(m)$; (ii) $O(m^2+m) = O(m^2)$; (iii) $O(km^3) = O(m^3)$, where k is some positive constant.] It is well-known that using the logarithmic-quadratic procedure reduces the number of iterates and hence the cut-off points (for N and c) to determine when to use one over the other depends on the values of N and c. Specifically, this depends on the value of $[(N+1)(c+1)]^3$. Obviously, exploiting the structure of R results in significant savings in the computational time and operations when N and c are large.

3.2. System performance measures

In order to qualitatively compare several models (or scenarios) we need to develop a few system performance measures. Here, we will give a few and others can be similarly developed. We first partition the vector $\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots)$ and $\boldsymbol{x}_0 = (\boldsymbol{x}_{0,0}, \cdots, \boldsymbol{x}_{0,c}, \cdots, \boldsymbol{x}_{N,0}, \cdots, \boldsymbol{x}_{N,c})$.

We define

$$\boldsymbol{a} = (\boldsymbol{a}_{0,0}, \cdots, \boldsymbol{a}_{0,c}, \cdots, \boldsymbol{a}_{N,0}, \cdots, \boldsymbol{a}_{N,c})\boldsymbol{a}_0(I - R)^{-1}.$$

1. The probability that the system is idle is $\boldsymbol{x}_{0,0}\boldsymbol{e}$.
2. The probability that all servers are idle is $\boldsymbol{a}_{0,0}\boldsymbol{e}$.
3. The probability that exactly i servers are busy is $\sum\limits_{k=0}^{N} \boldsymbol{a}_{k,i}\boldsymbol{e},\ 0 \le i \le c$.
4. The mean number of servers busy is $\sum\limits_{i=0}^{c}\sum\limits_{k=0}^{N} \boldsymbol{a}_{k,i}\boldsymbol{e}$.
5. The probability that exactly i are waiting in the feedback buffer is
$$\sum_{j=0}^{c} \boldsymbol{a}_{i,j}\boldsymbol{e},\ 0 \le i \le N.$$
6. The probability that exactly i r-calls in the orbit is $\boldsymbol{x}_i\boldsymbol{e} = \boldsymbol{x}_0R^i\boldsymbol{e},\ i \ge 0$.
7. The mean number of $r-$calls in the orbit is $\sum\limits_{i=1}^{\infty} i\boldsymbol{x}_i\boldsymbol{e} = \boldsymbol{x}_0(I - R)^{-2}\boldsymbol{e} - 1$.
8. The probability that an arrival is lost due to all server busy is $\frac{\alpha}{\lambda}\sum\limits_{k=0}^{N} \boldsymbol{a}_{k,c}D_1\boldsymbol{e}$.
9. The probability that an served $f-$ call (needing another service) is lost due to lack of buffer is $(1 - \beta)\sum\limits_{k=1}^{N} k\boldsymbol{a}_{N,k}\boldsymbol{e}$.

4. Illustrative examples

In this section we illustrate a few numerical examples out of many we generated. Further, we will discuss an optimization numerically. These examples are generated using a Fortran code. We used a number of internal accuracy checks to verify the written code.

First we look at the behavior of selected performance measures as α and β are varied. Towards this end, we consider five MAP arrivals: 1) the inter-arrival times are modeled using an Erlang of order 2 with parameter 2; 2) the inter-arrival times are exponentially distributed with parameter 1; 3) the inter-arrival times are modeled using hyperexponential distribution with parameters 1.9 and 0.19 with mixing probabilities, respectively, 0.9 and 0.1; 4) the inter-arrival times are modeled using a MAP that has a 1-lag correlation coefficient (CC) value of - 0.3267; 5) the inter-arrival times are modeled using a MAP that has a 1-lag CC value of 0.3267. The representation matrices (D_0, D_1) for these five cases are as follows.

MAP_1 : MAP for Erlang arrivals

$$D_0 = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix},\ D_1 = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}.$$

MAP_2 : MAP for Poisson arrivals

$$D_0 = (-1),\ \ D_1 = (1)$$

MAP_3 : MAP for arrival with hyperexponential distribution

$$D_0 = \begin{pmatrix} -1.9 & 0.19 \\ 0 & -0.19 \end{pmatrix}, \ D_1 = \begin{pmatrix} 1.71 & 0.19 \\ 0.171 & 0.019 \end{pmatrix}.$$

MAP_4 : MAP with negative CC

$$D_0 = \begin{pmatrix} -1.25 & 1.25 & 0 \\ 0 & -1.25 & 0 \\ 0 & 0 & -2.5 \end{pmatrix}, \ D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1.2375 & 0 & 0.0125 \\ 0.0250 & 0 & 2.4750 \end{pmatrix}.$$

MAP_5 : MAP with positive CC

$$D_0 = \begin{pmatrix} -1.25 & 1.25 & 0 \\ 0 & -1.25 & 0 \\ 0 & 0 & -2.5 \end{pmatrix}, \ D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.0125 & 0 & 1.2375 \\ 2.4750 & 0 & 0.0250 \end{pmatrix}.$$

The five $MAPs$ are qualitatively different as seen above. Further, the standard deviation (SDA) of the five $MAPs$ are, respectively, 0.7071, 1, 2.2447, 1.0392, and 1.0392. We fix $\lambda = 1, \mu = 2, c = 5, N = 10$, and vary α and β from 0 to 0.9. Note that the rate matrix R is of dimension $11 \times 6 \times m = 66m$.

In Figure 2 we plot the probability that the system is idle. A quick look at this shows that as β is increased this probability also increases for all five $MAPs$ considered. This is probably due to the fact that an increase in β creates a reduction in the load due to f-calls needing additional services. Further, this measure appears to be insensitive to the changes in α when β is increased.

The plot of the probability that all servers idle (this one is different from the system idle probability as it is possible for $r-$ and $f-$ calls to be waiting in their respective buffers) is displayed in Figure 3. The trend for this measure when varying β is similar to the ones seen in Figure 2.

In Figure 4 the mean number of busy servers is plotted as a function of α and β under various scenarios. As is to be expected (due to a reduction in the load to the system from f-calls), we see a decreasing trend in all the cases as β is increased. However, this measure is not significantly affected by a change in α. This fact (of insensitivity to α) is not surprising as it is known from classical queues as well as classical retrial queues that the mean number of busy servers depend only on the arrival and the service rates.

The mean number of $r-$ calls waiting in the retrial orbit is plotted under various scenarios in Figure 5. While this measure appears to decrease as β is increased (for a fixed α) under all scenarios, the rate of decrease depends not only on the type of MAP but also on the value of α. In the case of $MAPs$ with a large variability MAP_3 or with a positive correlation (MAP_5, it appears that the rate of change is small as compared to the other $MAPs$.

In Figures 6 and 7, respectively, we plot the mean number of $f-$calls in the feedback buffer and the loss probability of an $f-$call due to the feedback buffer being full at the time of such a requirement. A decreasing trend under all scenarios as β is increased (when α is fixed) in seen, which are as expected. While varying of α has almost no effect on the mean number of f-calls in orbit O_f, we do notice the sensitivity in varying α on the probability of f-call getting lost for β up to a certain point. Beyond this point we notice the insensitivity to α on this measure too.

Finally, in Figure 8, we plot the loss probability of an $r-$call at an arrival epoch due to all servers being busy. Obviously, one would expect this measure to exhibit sensitivity to α and is clearly seen in the plots under all scenarios. For all the $MAPs$ we notice a value of β, say, β^* (note that β^* depends on α and the type of MAP), such that an upward trend in this measure is seen as β is increased (for a fixed α) up to $\beta = \beta^*$ and after this point the measure decreases as β is increased (for a fixed α). This is seen for all values of α. Further, we see an interesting fact for positively correlated arrivals, namely, for MAP_5. There appears to be more than one peak as β is increased when α is fixed. This phenomenon of an upward trend followed by a downward trend is due to a similar trend in the traffic intensity. The intricate dependence of the traffic intensity ρ (see Equation (14)) on the parameters α and β for fixed values of the other input parameters, is plotted under various scenarios in Figure 9.

These figures allow us to study the sensitivity of performance measures to the type of MAP, i.e. flow with zero, positive and negative CC. From figures 1-7 we conclude that the impact of CC on performance measures is negligible.

4.1. Investigation of economic measures

We now investigate the effects of the parameters α and β of the Bernoulli trials on some selected economic measures of the system in the case when $N = 10$ and $c = 5$. Towards this end we define a number of costs and the economic measures along with the needed notations. These are displayed in Table 2 below.

Table 2. Costs, economic measures, and notations

$Notation$	$Definition$
c_p^b	Cost per unit time when a p-call balks
c_f^b	Cost per unit time when a f-call balks
c_r^w	Cost per unit time when a r-call waits in orbit
c_f^w	Cost per unit time when a f-call waits in orbit
c_{ser}	Cost per unit time when one server is idle
r_{ser}	Revenue earned by providing service to a call
P_p	Probability that p-call is lost
P_f	Probability that f-call is lost
L_r	Mean number of r-calls in O_r
L_f	Mean number of f-calls in O_f
BC_{av}	Mean number of busy servers
TC	Total expected cost per unit time of the system
TR	Total expected revenue per unit time of the system
TP	Total expected profit per unit time of the system

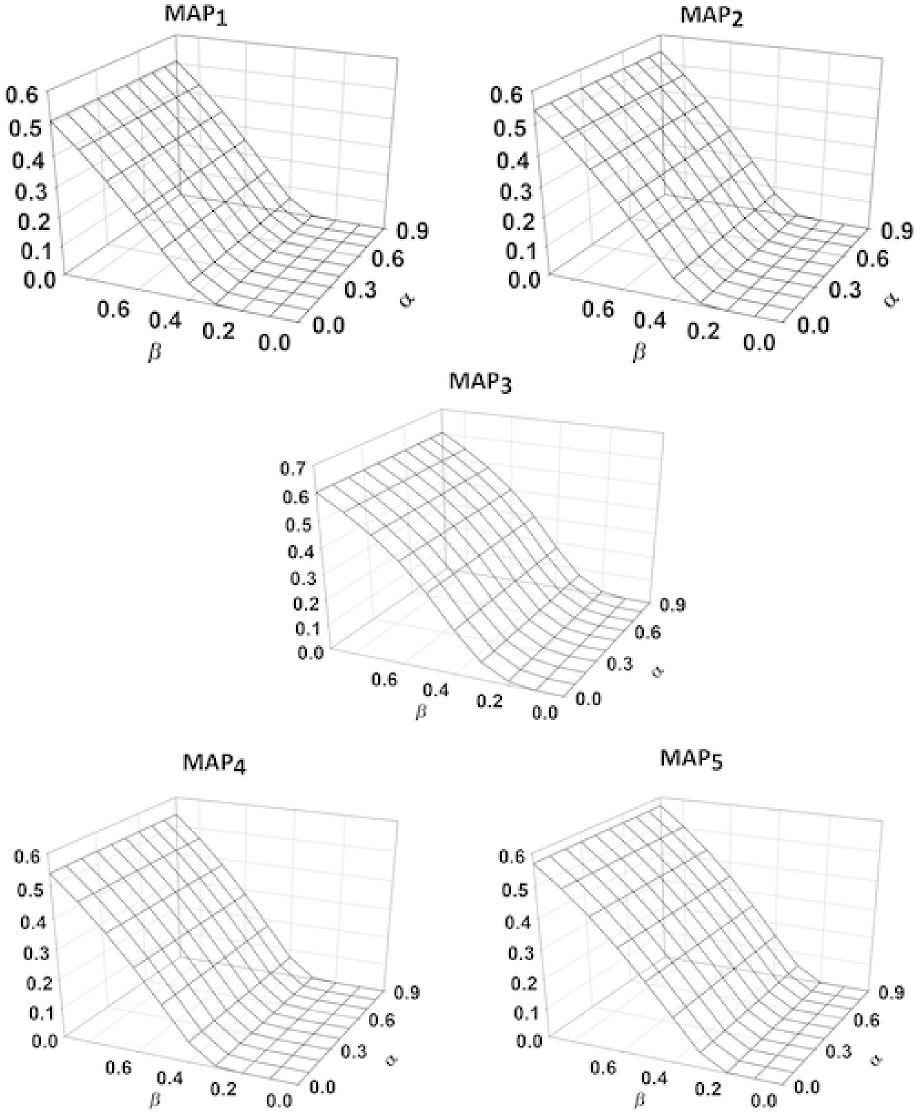

Figure 2. P(system is idle) under various scenarios

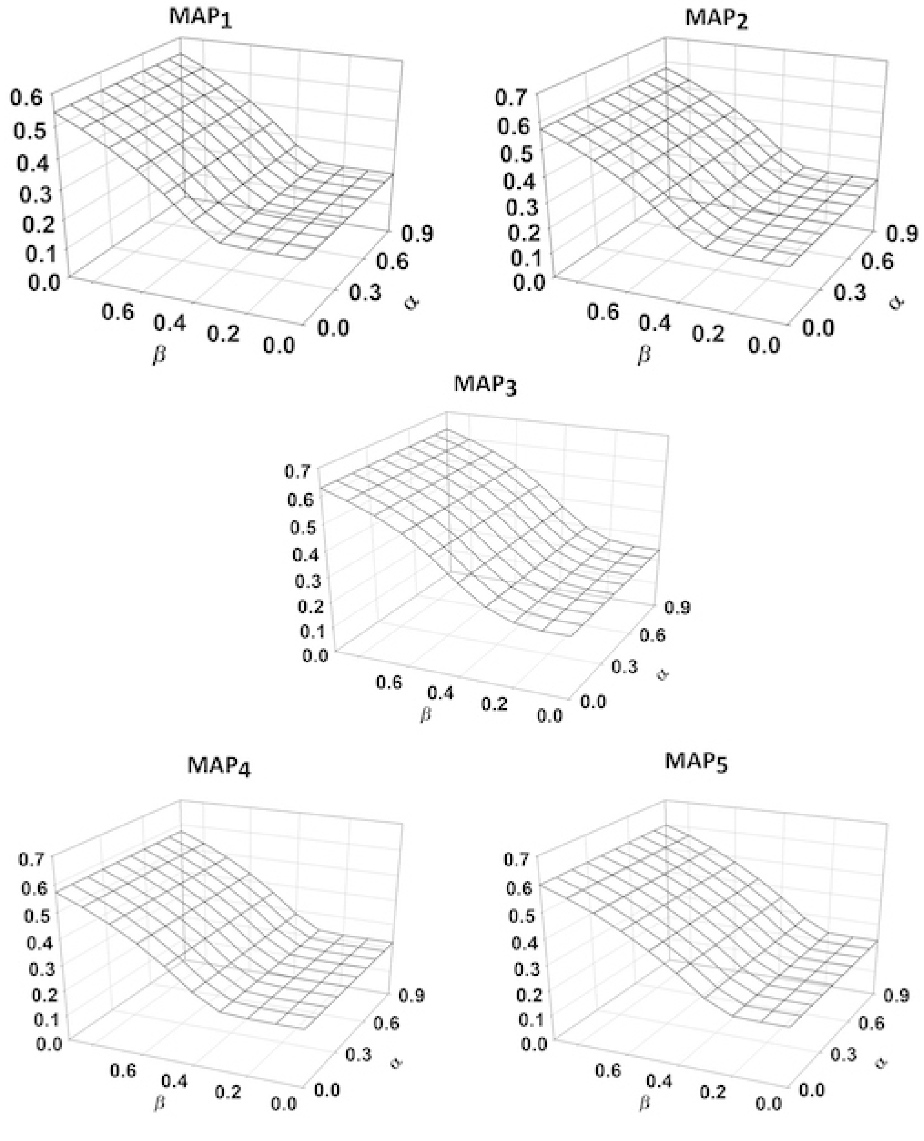

Figure 3. P(all servers are idle) under various scenarios

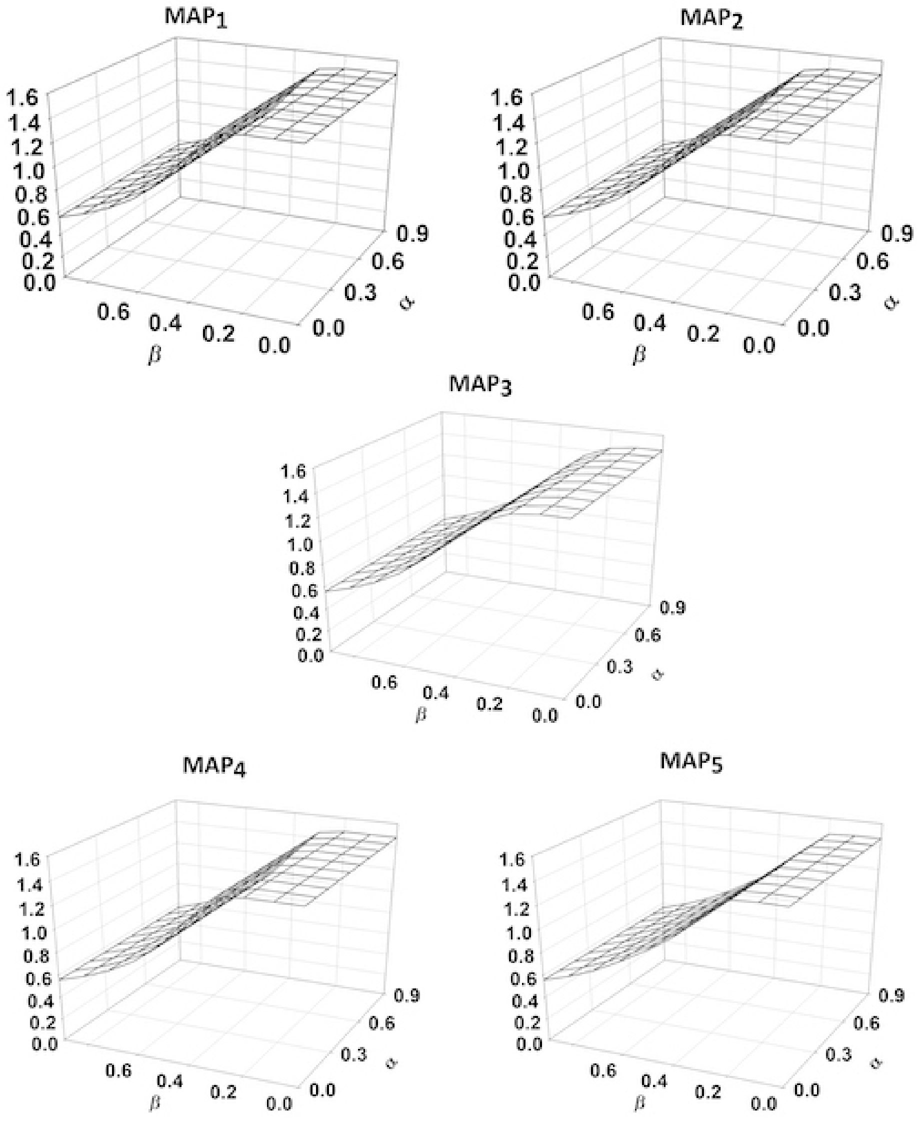

Figure 4. Mean number of servers busy under various scenarios

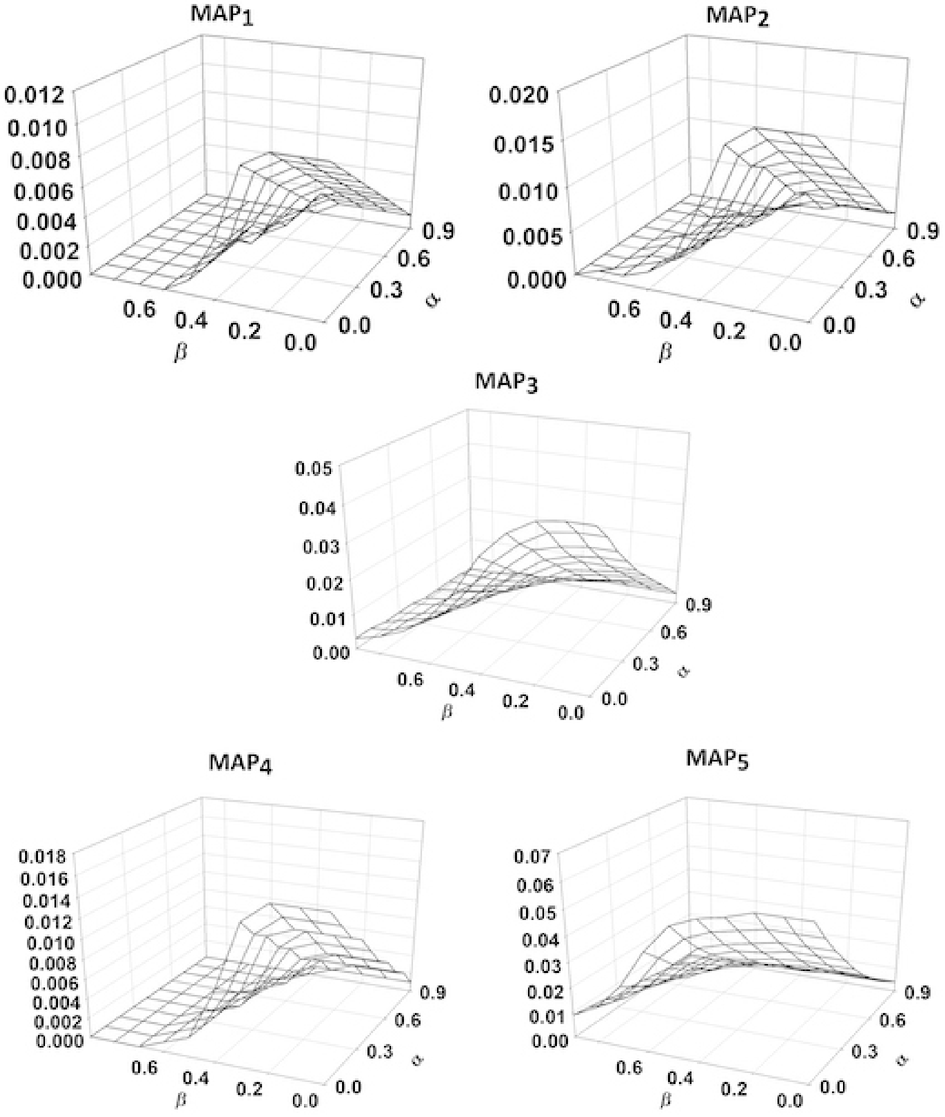

Figure 5. Mean number in retrial orbit under various scenarios

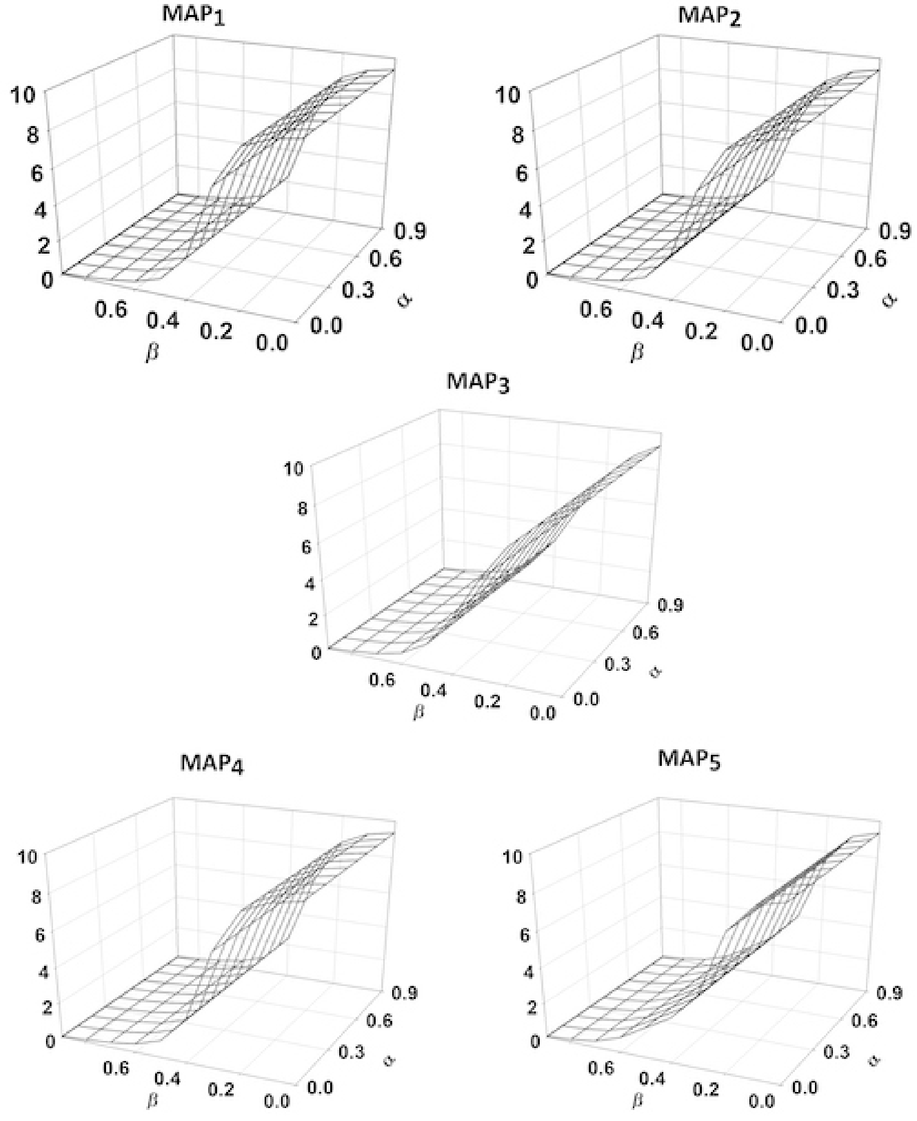

Figure 6. Mean number in feedback orbit under various scenarios

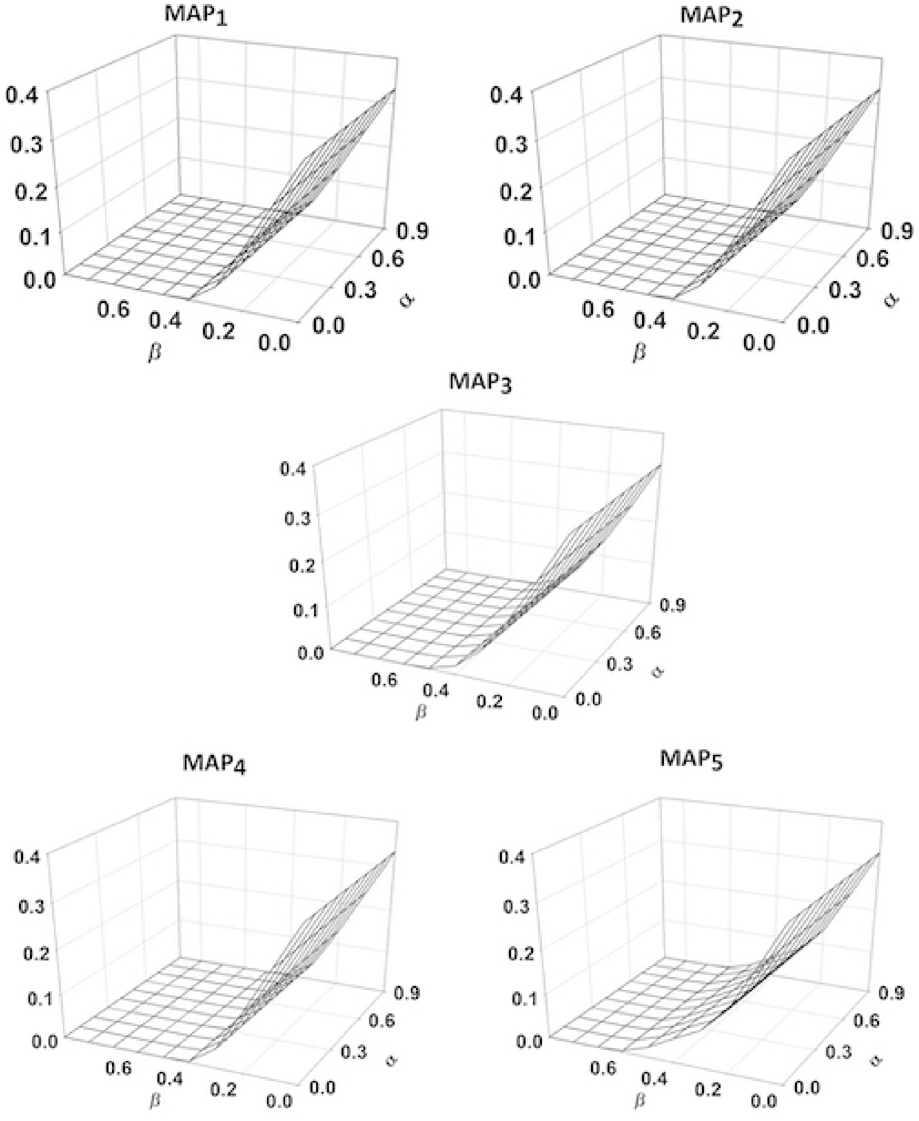

Figure 7. P(an $f-$call is lost) under various scenarios

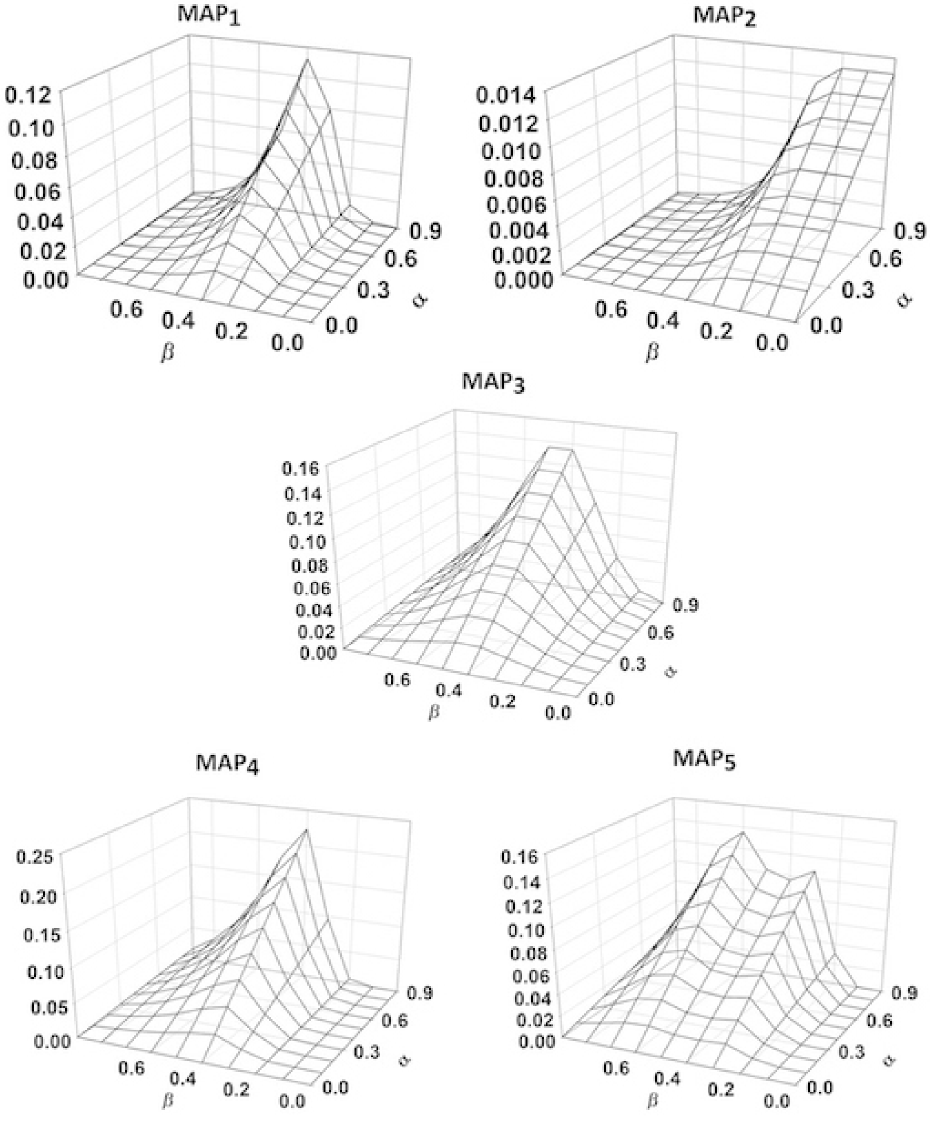

Figure 8. P(an $r-$call is lost) under various scenarios

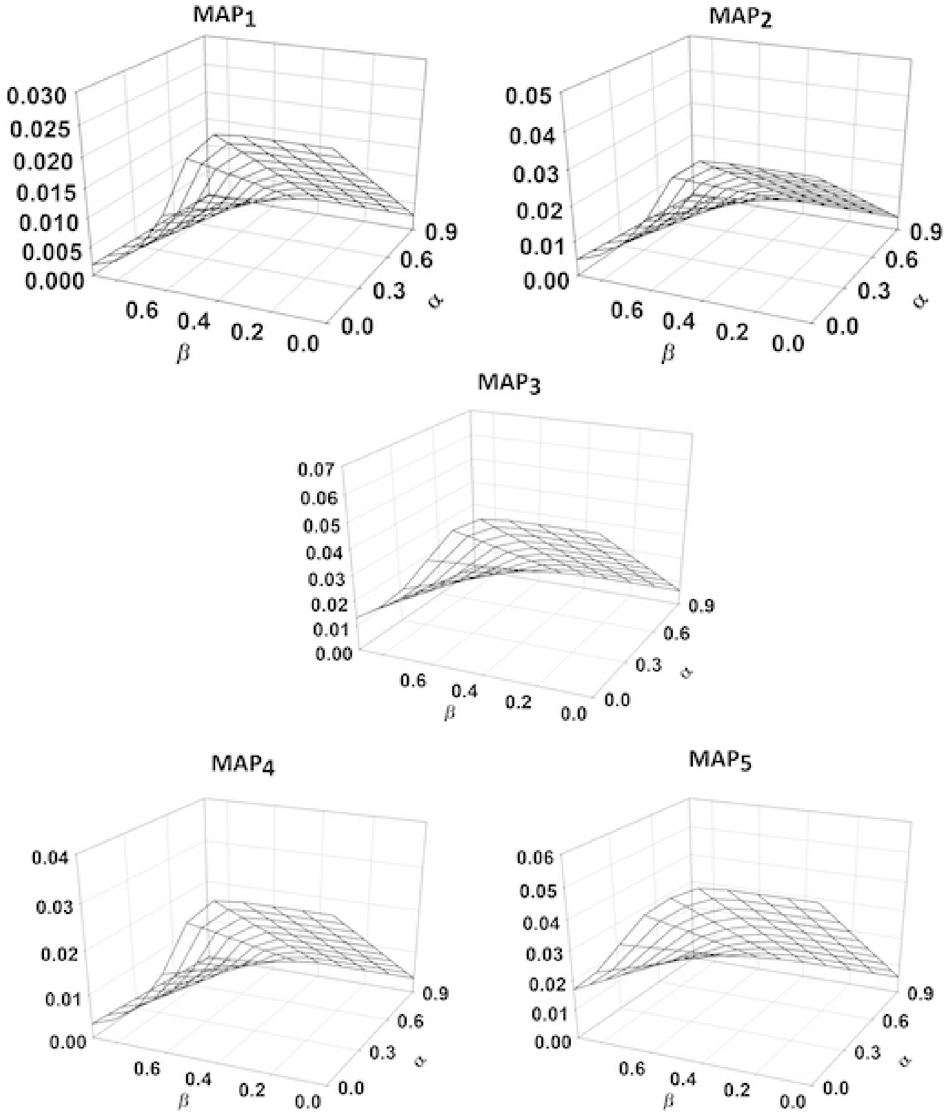

Figure 9. Traffic intensity (ρ) under various scenarios

The three economic measures, $TC, TR,$ and TP are computed as follows.

$$TC = c_p^b P_p + c_f^b P_f + c_r^w L_r + c_f^w L_f \ + c_{ser}\big(5 - BC_{av}\big),$$

$$TR = r_{ser}\mu BC_{av},$$

$$TP = TR - TC.$$

We look at different costs: $c_p^b = 5,\ c_f^b = 2.5,\ c_r^w = 1,\ c_f^w = 0.5,\ c_{ser} = 0.6,\ r_{ser} = 25,$ and compute the economic measures as functions of α and β under different scenarios for the arrival process.

In Figures 10-12, we display the graphs of the three economic measures under different scenarios. The optimum values, namely, the minimum for TC and the maximum for TR and TP, along with their corresponding α and β values under various scenarios are listed in Table 3 below. Note that values of (α, β) are shown within square parentheses.

The maximum for TR occurring at (0, 0) is intuitive since BC_{av} is a decreasing function of both α and β. Thus, when all balked calls go to their orbit with certainty (which occurs when $\alpha = 0$), and when all processed calls enter into feedback buffer with certainty (which occurs when $\beta = 0$), the total revenue has to go up. A similar argument justifies the minimum of TC occurring at $(0, 0.9)$. This is the case for all five arrival processes considered. The maximum for TP occurs when β is closer to 0 than 0.9, and depends on the type of the arrival process.

While the optimum values of $TR, TP,$ and TC appear to be close to each other for the $MAPs$ considered, the graphs of these measures show the sensitivity to the variability in the inter-arrival times as well as to the 1-lag correlation coefficient, especially when β is varied.

Table 3. Optimum values of the three economic measures

$Arrival$	$TC^*(min)$	$TR^*(max)$	$TP^*(max)$
MAP_1	0.0581 [(0,0.9)]	74.370 [(0,0)]	66.48305 [(0,0.3)]
MAP_2	0.0588 [(0,0.9)]	74.270 [(0,0)]	66.08205 [(0,0.3)]
MAP_3	0.0625 [(0,0.9)]	74.070 [(0,0)]	64.82940 [(0,0.2)]
MAP_4	0.0586 [(0,0.9)]	74.295 [(0,0)]	66.2831 [(0,0.3)]
MAP_5	0.0656 [(0,0.9)]	74.105 [(0,0)]	65.6587 [(0,0.2)]

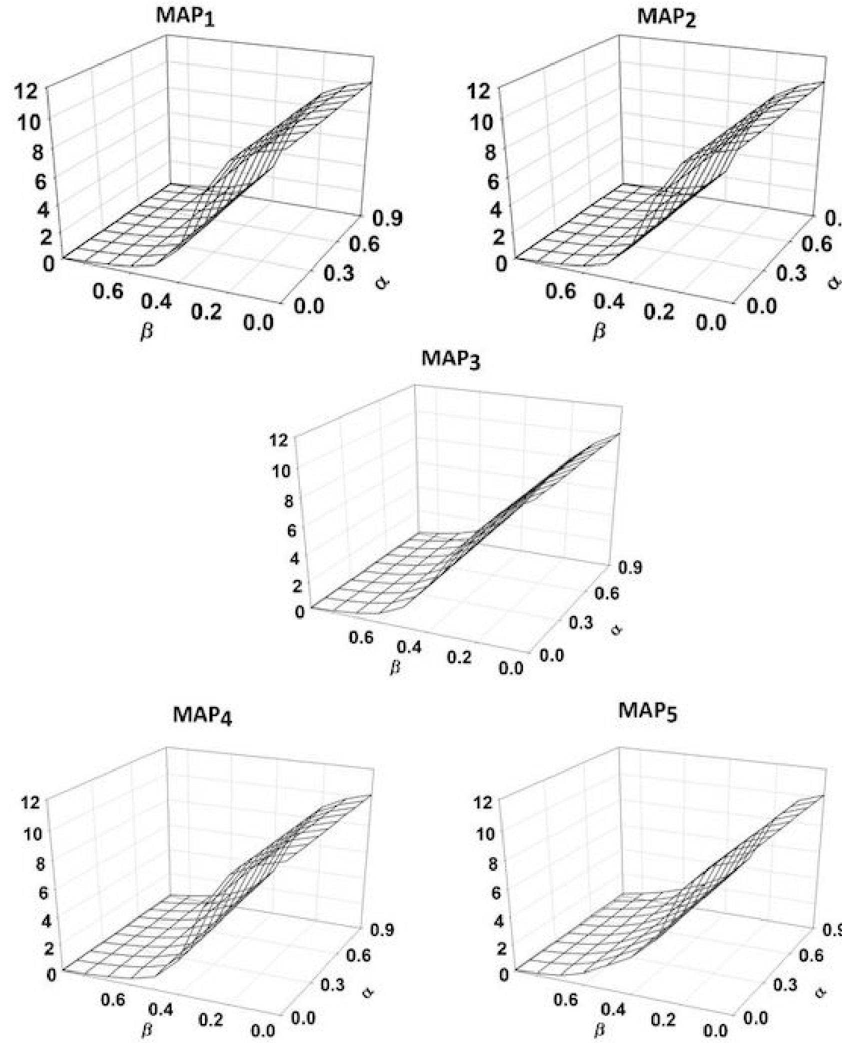

Figure 10. Economic measure, TC, as a function of α and β under various scenarios

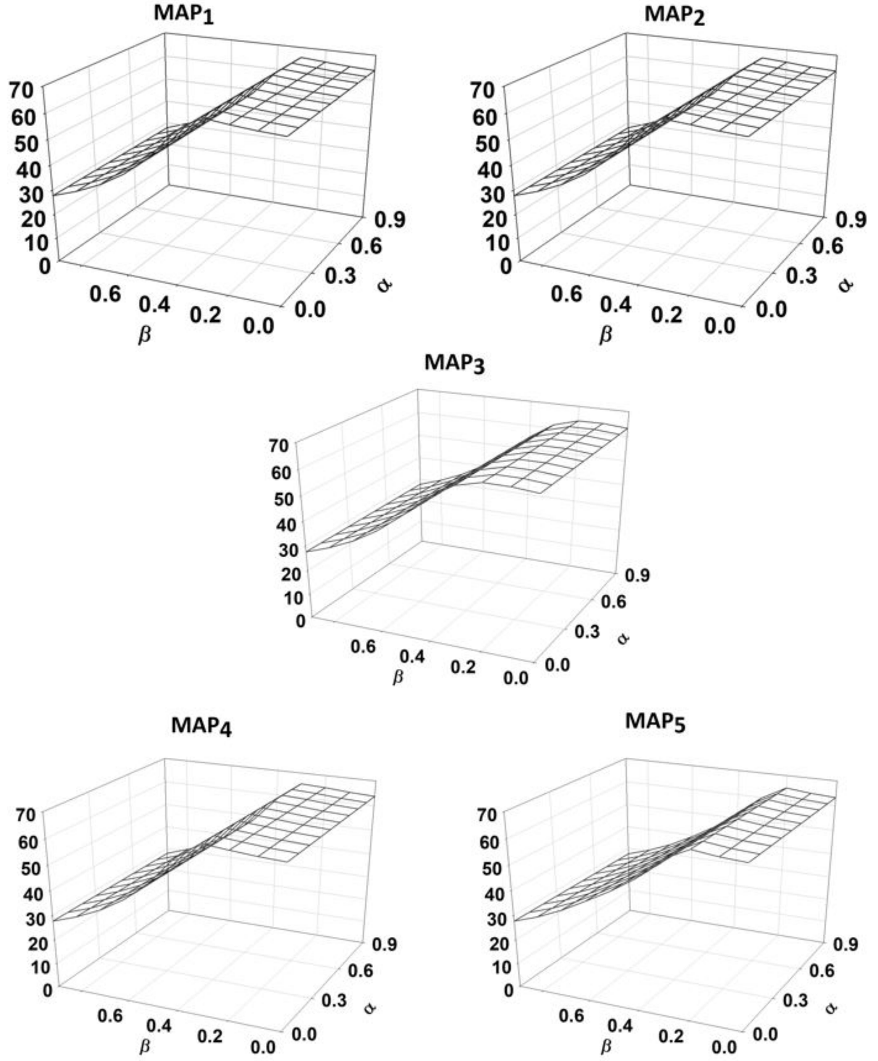

Figure 11. Economic measure, TR, as a function of α and β under various scenarios

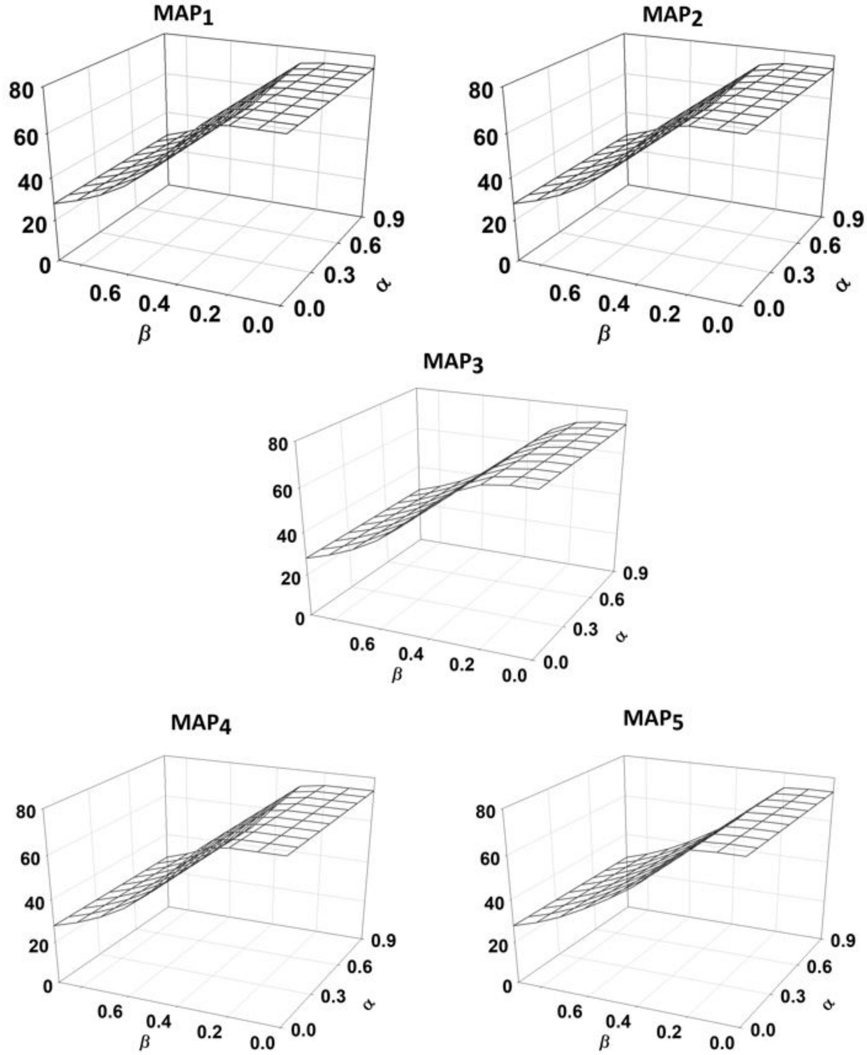

Figure 12. Economic measure, TP, as a function of α and β under various scenarios

5. Concluding remarks

We analyzed a multi-server $RQwDFB$ in which the calls arrive according to a versatile Markovian point process. There are two buffers, one infinite to hold the primary calls when all servers are busy (modeled using a Bernoulli trial), and the other a finite buffer to hold feedback calls (again using possibly a different Bernoulli trial). All feedback calls finding the buffer full will be lost. Calls from the two orbits try to capture a free served based on getting their respective signals that are generated by two independent Poisson processes with (possibly) different parameters which depend on the phase of the arrival process. The steady-state analysis of the model is carried out using the matrix-analytic method. Behavior of selected performance measures versus retrial and feedback probabilities for different MAP arrivals is investigated. Moreover, the cost-profit analysis of the model is performed and the impact of various scenarios for the arrival process as well as retrial and feedback probabilities on the total expected cost-profit of the system are presented in tabular form. The analysis carried out in this paper is very useful to choose the best service scheme that is used to maximize the profit of the system. The model can further be optimized for other parameters controllable. The model studied in this paper can be generalized in a number of ways. First, we can replace the Poisson signals with non-Poisson signals such as phase type or MAP. Secondly, exponential service times can be replaced with phase type distributions. Finally, the servers can be made as heterogeneous. In these cases, the dimensions of the problem will be increased and a careful planning the numerical implementation is needed. These issues are subject further investigations.

Acknowledgement

The authors appreciate the comments and suggestions of the reviewers that improved the presentation of this paper.

References

[1] Artalejo, J. R., & Gomez-Corral, A. (2008). Retrial Queueing Systems: A Computational Approach. London: Springer. doi: 10.1007/978-3-540-78725-9.

[2] Ayyapan, G., Subramanian, A. M. G., & Sekar, G. (2010a). M/M/1 Retrial Queuing System with Loss and Feedback Under Pre-emptive Priority Service. *International Journal of Computer Applications*, 2, 27–34. doi:10.5120/672-945.

[3] Ayyapan, G., Subramanian, A. M. G., & Sekar, G. (2010b). M/M/1 Retrial Queueing System with Loss and Feedback Under Non-pre-emptive Priority Service by Matrix Geometric Method. *Applied Mathematical Sciences*, 4, 2379–2389.

[4] Ayyapan, G., & Thilagavathy, K. (2021). Analysis of MAP/PH/1 queueing model with immediate feedback, starting failures, single vacation, standby server, delayed repair, breakdown and impatient customers. *International Journal of Mathematics in Operational Research*, 19(3), 269–301.

[5] Berg, J. L., & Boxma, O. J. (1991). The M/G/1 Queue with Processor Sharing and its Relation to Feedback Queue. *Queueing Systems*, 9(4), 365–402.

[6] Bouchentouf, A. A., Cherfaoui, M., & Boualem M. (2019). Performance and Economic Analysis of a Single Server Feedback Queueing Model with Vacation and Impatient Customers. *OPSEARCH*, 56, 300–323.

[7] Bouchentouf, A. A., Cherfaoui, M., & Boualem, M. (2020). Analysis and Performance Evaluation of Markovian Feedback Multi-Server Queueing Model with Vacation and Impatience. *American Journal of Mathematical and Management Sciences*. doi: 10.1080/01966324.2020.1842271

[8] Bouchentouf, A. A., & Guendouzi, A. (2021). Single Server Batch Arrival Bernoulli Feedback Queueing System with Waiting Server, Variant Vacations and Impatient Customers. *Operations Research Forum*, 2(14). doi.org/10.1007/s43069-021-00057-0.

[9] Chakravarthy, S. R. (2010). Markovian arrival process. Wiley Encyclopedia of Operation Research and Management Science. OI: 10.1002/9780470400531.eorms0499

[10] Chakravarthy, S. R. (2021). A multi-server retrial queueing model with Poisson signals. *Journal of Applied Mathematics and Informatics*, 39(5-6), 601 – 616.

[11] Choi, B. D., Kim, Y., & Lee, Y. (1998). The M/M/c Retrial Queue with Geometric Loss and Feedback. *Computers & Mathematics with Applications*, 36, 41–52. doi: 10.1016/S0898-1221 (98) 00160-6.

[12] D'Avignon, G. R., & Disney, R. L. (1977). Queues with Instantaneous Feedback. *Management Science*, 24(2), 168–180.

[13] Dimitriou, I and Phung-Duc, T. (2018). A Riemann-Hilbert boundary value problem for single-server systems with two queues for blocked and feedback customers. *AIP Conference Proceedings*, 1978(1), 190004. https://doi.org/10.1063/1.5043831

[14] Do, T. V. (2010). An Efficient Computation Algorithm for a Multi-Server Feedback Retrial Queue with a Large Queuing Capacity. *Applied Mathematical Modeling*, 34, 2272–2278.

[15] Dudin, A. N., Kazimirsky, A. V., Klimenok, V. I., Breuer, L., & Krieger, U. (2005). The Queueing Model MAP/PH/1/N with Feedback Operating in a Markovian Random Environment. *Austrian Journal of Statistics*, 34(2), 101–110.

[16] Dudin, S., & Dudina, O. (2019). Retrial multi-server queuing system with PHF service time distribution as a model of a channel with unreliable transmission of information. *Applied Mathematical Modeling*, 65, 676–695.

[17] Falin, G., & Templeton, J. G. C. (1997). Retrial Queues. London: Chapman & Hall. doi: 10.1007/BF02564732.

[18] Foley R. D., & Disney R. L. (1983). Queues with Delayed Feedback. *Advances in Applied Probability*, 15(1), 162–182.

[19] He, Q. M. (2014). Fundamentals of Matrix-Analytic Methods, Springer-Verlag, New York.

[20] Hunter, J. J. (1989). Sojourn Time Problems in Feedback Queue. *Queueing Systems*, 5(1-3), 55-76.

[21] Kim, J., & Kim, B. (2016). A Survey of Retrial Queueing Systems. *Annals of Operations Research*, 247, 3–36. doi: 10.1007/s10479-015-2038-7.

[22] Krieger, U., Klimenok, V. I., Kazimirsky, A. V., Breuer, L., & Dudin, A. N. (2005). A BMAP/PH/1 Queue with Feedback Operating in a Random Environment. *Mathematical and Computer Modelling*, 41, 867–882.

[23] Krishna Kumar, B., Rukmani, R., & Thangaraj, V. (2009). On Multi-Server Feedback Retrial Queue with Finite Buffer. *Applied Mathematical Modeling*, 33, 2062–2083.

[24] Krishnamoorthy, A., & Manjunath, A. S. (2018). On Queues with Priority Determined by Feedback. *Calcutta Statistical Association Bulletin*, 70, 33–56.

[25] Lee, Y. W. (2005). The M/G/1 Feedback Retrial Queue with Two Types of Customers. *Bulletin of the Korean Mathematical Society*, 42(4), 875–887.

[26] Latouche, G., & Ramaswami, V. (1999). Introduction to matrix analytic methods in stochastic modeling. ASA-SIAM series on statistics and applied probability. Philadelphia.

[27] Lee, H. W., & Ahn, B. Y. (2000). Analysis of a Production System with Feedback Buffer and General Dispatching Time. *Mathematical Problems in Engineering*, 5, 421–439.

[28] Lee, H. W., & Seo, D. W. (1997). Design of a Production System with Feedback Buffer. *Queueing Systems*, 26(1), 187–202. doi.org/10.1023/A:1019129107476.

[29] Lucantoni, D. (1991). New results on the single server queue with a batch Markovian arrival process, Communications in Statistics. *Stochastic Models*, 7, 1–46.

[30] Lucantoni, D. M., Meier-Hellstern, K. S., & Neuts, M. F. (1990). A single-server queue with server vacations and a class of nonrenewal arrival processes. *Advances in Applied Probability*, 22(3), 676–705.

[31] Mitrani, I., & Chakka, R. (1995). Spectral Expansion Solution for a Class of Markov Models: Application and Comparison with the Matrix-Geometric Method. *Performance Evaluation*, 23, 241–260. doi: 10.1016/0166-5316 (94)00025-F.

[32] Melikov, A. Z., & Aliyeva, S. H. (2019). Mathematical Models of the Queuing Systems with MMPP Flow and Instantaneous Feedback. *Communications in Computer and Information Sciences*, 1141, 288–301.

[33] Melikov, A. Z., Aliyeva, S. H., & Shahmaliyev, M. O. (2020). Methods to Calculate the System with Instantaneous Feedback and Varying Arrival Rate. *Automation and Remote Control*, 81(9), 1647–1658.

[34] Melikov, A. Z., Aliyeva, S. H., & Sztrik, J. (2019). Analysis of Queueing System MMPP/M/K/K with Delayed Feedback. *Mathematics*, 7(11), Article 1128. doi: 10.3390/math7111128.

[35] Melikov, A. Z., Aliyeva, S. H., & Sztrik, J. (2020). Analysis of Instantaneous Feedback Queue with Heterogeneous Servers. *Mathematics*, 89(12), Article 2186. doi: 10.3390/math8122186.

[36] Melikov, A. Z., Ponomarenko, L. A., & Kuliyeva, K. N. (2015a). Calculation of the Characteristics of Multi-Channel Queuing System with Pure Losses and Feedback. *Journal of Automation and Information Sciences*, 47(5), 19–29.

[37] Melikov, A. Z., Ponomarenko, L. A., & Kuliyeva, K. N. (2015b). Numerical Analysis of the Queuing System with Feedback. *Cybernetics and Systems Analysis*, 51(4), 566–570.

[38] Melikov, A. Z., Ponomarenko, L. A., & Rustamov, A. M. (2015). Methods for Analysis of Queuing Models with Instantaneous and Delayed Feedbacks. *Communications in Computer and Information sciences*, 564, 185–199.

[39] Melikov, A. Z., Ponomarenko, L. A., & Sztrik, J. (2016). Hierarchical Space Merging Algorithm for Analysis of Two Stage Queuing Network with Feedback. *Communications in Computer and Information Science*, 638, 238–249.

[40] Mokaddis, G. S., Metwally, S. A., & Zaki, B. M. (2007). A Feedback Retrial Queuing System with Starting Failures and Single Vacation. *Tamkang Journal of Science and Engineering*, 10, 183–192.

[41] Neuts, M. F. (1979). A versatile Markovian point process. *Journal of Applied Probability*, 16(4), 764–79.

[42] Neuts M. F. (1981). Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Baltimore: Johns Hopkins University.

[43] Neuts, M. F. (1989). Structured Stochastic Matrices of $M/G/1$ Type and Their Applications. Marcel Dekker, Inc., New York.

[44] Pekoz E. A., & Joglekar N. (2002). Poisson Traffic Flow in a General Feedback. *Journal of Applied Probability*, 39(3), 630–636.

[45] Phung-Duc, T. (2019). Retrial Queueing Models: A Survey on Theory and Applications. arXiv:1906.09560. Vol. 1.

[46] Ponomarenko, L., Kim, C. S., & Melikov A. (2010). Performance Analysis and Optimization of Multi-traffic on Communication Networks. London: Springer.

[47] Rajadurai, P., Sundararaman, M., & Narasimhan, D. (2020). Performance Analysis of an M/G/1 Retrial Queue with Feedback Under Working Breakdown Services. *Songklanakarin Journal of Science & Technology*, 42, 236–247.

[48] Takacs, L. (1963). A Single-Server Queue with Feedback. *Bell System Technical Journal*, 42, 505–519.

[49] Takacs, L. (1977). A Queuing Model with Feedback. *Operations Research*, 11, 345–354.

[50] Wortman, M. A., Disney, R. L., & Kiessler, P. C. (1991). The M/GI/1 Bernoulli Feedback Queue with Vacations. *Queueing Systems*, 9(4), 353–363.