35th European Modeling & Simulation Symposium
20th International Multidisciplinary Modeling & Simulation Multiconference

2724-0029 © 2020 The Authors.
doi: 10.46354/i3m.2023.emss.039

A hysteresis type strategy for server reservation in a multi-server queueing system with priority and retrials

Ciro D'Apice[1], Alexander Dudin[2], Rosanna Manzo[3] and Luigi Rarità[1*]

[1]Dipartimento di Scienze Aziendali, Management & Innovation Systems, University of Salerno, Via Giovanni Paolo II, 132, Fisciano (SA), 84084, Italy
[2]Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus
[3]Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno, Via Giovanni Paolo II, 132, Fisciano (SA), 84084, Italy

[*]Corresponding author. Email address: lrarita@unisa.it

Abstract

This work deals with a possible model of a cell of cognitive radio architectures via a multi-server queueing system with two different types of requests and preemptive priority of one type of requests on the other. Service demands arrive according to Markovian Arrival Processes in order to capture traffic correlation. Cognitive engine tries to find a solution or optimize a performance goal based on the received inputs that define the radio's current internal state and operating environment. Hence, beside priority mechanisms, a possible deactivation of information to transmit is considered via a hysteresis type strategy for the access to the service, with the aim of smoothing the effects of service interruptions for low priority requests. Fixed values for the thresholds of the strategy are assumed. The system is statistically described by a level-dependent multi-dimensional Markov chain, which allows the computation of stationary probabilities and robust performance indices. Numerical results are presented to compare the system performances in the case of uncorrelated flows and flows with different correlation levels.

Keywords: multi-server queueing system; cognitive radio; priority; hysteresis mechanism.

1. Introduction

A retrial impatient queueing system consisting of N identical servers is analysed. Two types of requests are processed. The service times of a type-k request are independent identically exponentially distributed random variables. Type-1 requests have preemptive (absolute) priority over type-2 requests. Both types of requests arrive according to Markovian Arrival Processes (MAPs). The admission of requests of type-2 to the system follows a hysteresis strategy.

In order to avoid possible oscillations due to the frequent change of the operation mode of the system, a threshold strategy is presented as follows. Two thresholds are considered, M_1 and M, with $M_1 \leq M$. The admission of non-priority requests ends when the number of requests in the system exceeds M and is resumed when this number becomes less than M_1.
The service of a type-2 request may end at the arrival of a type-1 request and a type-2 request leaves the system permanently with a certain probability or moves to a virtual place, the orbit. The requests inside the orbit are impatient and depart from the system without service after a random time having an exponential distribution.

1.1. Possible applications

Different real life phenomena are described via multi-server queueing systems with various type of customers and/or requests. Indeed, many mathematical models are useful to simulate and optimize situations in some contexts, such as emergency departments in hospitals, supply chains as well as cells for cognitive radio systems. In such architectures, primary requests (i.e. they are considered as licensed users or patients with a severe injury) have preemptive priority over secondary ones (seen as cognitive customers or other types of patients). Secondary requests are dropped if, at their arrival moment, all servers give service to the primary requests. If all servers are busy, but some of them provide service to the secondary requests, the service of a secondary request ends and the primary request enters the server. When a service ends by interruption, there are various negative effects (i.e. the loss of the throughput). Hence, a suitable admission of the secondary requests is necessary, for instance by adopting some thresholds values to discriminate the possible interruption.

In normal contexts, radio cognitive cells have mechanisms that permit to discriminate among possible signals to transmit. Input impulses are accepted but not always immediately elaborated: in some cases, they are "served"; in others, they can be lost or put on hold. This justifies the possible adoption of a model that, considering multiple flows of incoming information, distinguishes priorities on requests as well as an acceptance mechanism for the elaboration of outputs. The management of the various elaborations are already assumed in some scientific works where, if there are N servers, the non-priority request is accepted to the system only when the number of busy servers is less than a threshold M such that $0 < M \leq N$. Such a modelling strategy has been improved by the introduction of: arrivals described by the Marked Markovian Arrival Process (MMAP), generalization of Markovian Arrival Process (MAP) for heterogeneous requests; an orbit, that represents a virtual place where non-priority requests, that are not immediately accepted to the system or interrupted during a service, have the option to be deactived instead of abandoning the system, and have then to retry for service after a random time.

The possibility of retrials is indeed typical in telecommunication systems. On the other hand, threshold and multi-threshold strategies of control are widely described in cases of flows of types MAP or MMAP. In particular, MAPs are able to describe correlated arrivals, and MMAPs allow to capture cross-correlations between the arrivals belonging to different classes as well.

1.2. Brief literature review

For an adequate presentation of examples of multi-server queueing systems with various type of customers and/or requests, see for instance (2), (12) and (27). Different mathematical ways to simulate and optimize situations such as emergency departments in hospitals, supply chains as well as cells for cognitive radio systems, are shown in (1), (6), (11), (17), (18), (29), (30), (32). As for the management of priorities, consider the work (33) where all non-priority requests enter the server if $0 < M \leq N$, where M is a suitable threshold and N represents the number of servers.

Notice that the model, seen in (33), has been improved by works like (31) where arrivals are described by flows MMAP and MAP, see (28). Eventual insights are also in (5), (13), (22) and (24).

The possibility of retrials, essential aspect for the described model, is widely considered in some works like (3), (4), (10), (18), (20), while (9) represents a suitable example of a hysteresis strategy of admission control. Finally, the paper (19) describes in a wide theoretical way the proposed model, whose performance indices are deduced by (16).

1.3. Contribution of the paper

The paper (19) presents a deep theoretical analysis for the queuing model that is here shortly described, but there is not an exhaustive description of some features, such as the possible differences in adopting various input flows. Hence, several numerical examples are presented in Section 5 to underline that general models of arrival processes with high correlations (like MAPs for instance) lead to quite different results from the ones obtained by arrivals of Poisson types. Such last flows, interested by a zero correlation, are nowadays still widely used but are not able to focus on the features of real systems and networks. Hence, the main contribution of this work is due to numerical studies, that clearly show the evident differences in using different and high correlations for input flows.

1.4. Organization of the work

The paper is organized as follows. Section 2 deals with the mathematical model of the system. Section 3 presents a level dependent multi-dimensional continuous-time Markov chain for the dynamics of the overall system, and suitable conditions for the ergodicity. Section 4 describes the main performance indices. Numerical results are shown in Section 5. Conclusions end the paper in Section 6.

2. Mathematical model

A queueing system, that consists of N identical servers with no buffer, is considered. For such a system, two types of requests are assumed. Type-1 requests have preemptive priority over type-2 ones. Service times of a type-k request are independent, identically exponentially distributed random variables with rate μ_k, $k = 1, 2$. The arrival of type-k, $k = 1, 2$, requests is described by a Markovian Arrival Process, defined by an irreducible continuous-time Markov chain $\nu_t^{(k)}$, $t \geq 0$ with $\bar{W}_k = W_k + 1$ states $\{0, ..., W_k\}$. The transition intensities of $\nu_t^{(k)}$ within the state space are defined by the matrices $D_0^{(k)}$ and $D_1^{(k)}$ of size

$\bar{W}_k$. The matrix $D_0^{(k)}$ has non-diagonal entries that define the intensities of transitions that are not accompanied by arrival of type-k request. The diagonal entries of the matrix $D_0^{(k)}$ indicate the rates of the process $\nu_t^{(k)}$ exit from its states. The entries of the matrix $D_1^{(k)}$ represent the intensities of transitions that are accompanied by arrival of a type-k request.

The infinitesimal generator of the Markov chain $\nu_t^{(k)}$ is represented by the matrix $D^{(k)}(1) = D_0^{(k)} + D_1^{(k)}$. The row vector $\theta^{(k)}$ indicates the stationary distribution of this Markov chain. Such vector represents the unique solution to the system $\theta^{(k)} D^{(k)}(1) = \mathbf{0}, \ \theta^{(k)}\mathbf{e} = 1$. Here and throughout this paper, $\mathbf{e}$ is a column vector of appropriate size consisting of 1's, and $\mathbf{0}$ is a row vector of appropriate size consisting of 0's.

The fundamental (average) arrival rate λ_k of type-k requests is defined by $\lambda_k = \theta^{(k)} D_1^{(k)}\mathbf{e}$.

Requests of type-1 have preemptive priority over requests of type-2. Hence, an arriving type-1 request is always admitted to the system except the case when, during the arrival moment, all servers are busy by providing service to type-1 requests. In such a context, a type-1 request leaves the system without service (it is lost). If all servers during type-1 request arrival epoch are busy, but at least one of them provides service to type-2 requests, the service of one of these requests ends and the type-1 request occupies the corresponding server.

The admission of requests of type-2 to the system follows the hysteresis strategy as follows. The decision to admit or to reject an arriving type-2 request depends on the current state of the managing stochastic process ξ_t that has two possible values: 0 and 1. Indeed, the value 0 represents an off-period during which arriving requests of type-2 are not admitted into the service; the value 1 indicates an on-period during which arriving requests of type-2 can be admitted for service.

The mechanism of switching the states of the managing process is defined by the current number of busy servers and two integer thresholds M_1 and M, $0 \leq M_1 \leq M \leq N$.

If the number of busy servers is less than M during the stay of the process ξ_t in the state 1, then any request, that tries to access, is admitted to the system and immediately starts service. If the number of busy servers during the stay of the process ξ_t in the state 1 equals M and a new request arrives from outside, then the process ξ_t transits to the state 0. The arriving request is accepted if it is of type-1 and $M < N$ and is rejected if it is of type-2. With probability $1 - q$, $0 \leq q \leq 1$, a rejected request leaves the system permanently (it is lost). With probability q, this request decides to retry to get the access in a second moment. In particular, this request moves to a virtual place called "orbit". A request inside the orbit repeats the attempts to get access, independently of other requests inside the orbit, after a random time interval that has exponential distribution with rate α, $\alpha > 0$. An attempt is successful if the managing process ξ_t is in state 1 and the number of busy servers is less than M. If the attempt is successful, the request immediately occupies a free server and starts service. If the attempt is not successful, with probability $1 - q$ the retrying request departs from the system. With probability q, the request comes back to the orbit.

When the process ξ_t is in the state 0, the number of busy servers equals M_1 and the service of a request ends. Then, the process ξ_t transits to the state 1 (the on-period begins). At all the other moments (when no arrival occurs during the on-line period in presence of M busy servers or no service completion occurs during the off-line period in presence of M_1 busy servers) no switches of the states of ξ_t occur.

The service of a type-2 request admitted for service may end at the arrival of a type-1 request. In this case, a type-2 request leaves the system permanently with probability $1 - p$, $0 \leq p \leq 1$, or moves to the orbit. The requests inside the orbit are impatient and depart from the system without service after a random time having an exponential distribution with parameter γ, $\gamma > 0$.

3. Markov Process for the system states and ergodicity

Consider the following quantities at the time t, $t \geq 0$:

- $i_t, i_t \geq 0$, are the number of requests in the orbit;
- n_t, $n_t = \overline{0, N}$, are the number of busy servers;
- l_t, $l_t = \overline{0, \min\{n_t, M\}}$, are the number of requests of type-2 inside the service;
- $\nu_t^{(k)}$, $\nu_t^{(k)} = \overline{0, W_k}$, are the state of the process MAP_k, $k = 1, 2$;
- ξ_t represents the state of the admission managing process: $\xi_t = 0$ during an off-period, while $\xi_t = 1$ during an on-period.

The six-dimensional process

$$\eta_t = \{i_t, n_t, l_t, \nu_t^{(1)}, \xi_t, \nu_t^{(2)}\}, \quad t \geq 0,$$

represents an irreducible continuous-time Markov chain.

The states of the chain ξ_t are enumerated in the direct lexicographic order of the components $(i, n, l, \nu^{(1)}, \xi, \nu^{(2)})$. The set of the states with value (i, n) of two first components is called *macro-state* (i, n) and the set of macro-states $((i, 0), \ldots, (i, N))$ as *level* i, $i \geq 0$.

Assume that Q is the generator of the Markov chain ξ_t, $t \geq 0$. The generator Q has the blocks $Q_{i,j}$, that, in turn, consist of the matrices $(Q_{i,j})_{n,n'}$ of the intensities of the transition of the chain ξ_t from the macro-state (i, n) to the macro-state (j, n'), n, $n' = \overline{0, N}$. The exact expression for the generator Q is not reported here, but further details are in (19).

For the ergodicity of the process η_t, we use the results of (26) and define the quantities:

$$Y_0 = \lim_{i \to \infty} R_i^{-1} Q_{i,i-1}, \ Y_1 = \lim_{i \to \infty} R_i^{-1} Q_{i,i} + I, \ Y_2 = \lim_{i \to \infty} R_i^{-1} Q_{i,i+1},$$

where R_i is a diagonal matrix with elements defined as the moduli of the corresponding diagonal entries of the matrix $Q_{i,i},\ i \geq 0$. From (26) we get the following sufficient condition for ergodicity:

$$\mathbf{y}Y_0\mathbf{e} > \mathbf{y}Y_2\mathbf{e},$$

where the row vector $\mathbf{y}$ is the unique solution to the system of linear algebraic equations

$$\mathbf{y}(Y_0 + Y_1 + Y_2) = \mathbf{y},\ \ \mathbf{ye} = 1.$$

4. Performance indices

Under the assumption that the ergodicity condition is satisfied, the following stationary probability exists:

$$\pi(i, n, l, \nu^{(1)}, \xi, \nu^{(2)}) =$$

$$\lim_{t\to\infty} P\{i_t = i,\ n_t = n,\ l_t = l,\ \nu_t^{(1)} = \nu^{(1)}, \xi_t = \xi, \nu_t^{(2)} = \nu^{(2)}\},$$

$$i \geq 0,\ n = \overline{0, N}, l = \overline{0, \min\{n, M\}},$$

$$\nu^{(1)} = \overline{0, W_1}, \xi = 0, 1, \nu^{(2)} = \overline{0, W_2}\ .$$

The row vectors of the invariant probabilities π_i of the levels are defined as follows:

$$\pi_i = (\pi(i, 0), \pi(i, 1), \dots, \pi(i, N)),\ i \geq 0,$$

where the row vectors $\pi(i, n)$ of the macro-states are:

$$\pi(i, n) = (\pi(i, n, 0), \pi(i, n, 1), \dots, \pi(i, n, \min\{n, M\})),$$

$$n = \overline{0, N},$$

with row vectors $\pi(i, n, l)$ that, in turn, consist of the invariant probabilities $\pi(i, n, l, \nu^{(1)}, \xi, \nu^{(2)})$ enumerated in the lexicographic order of the components $(\nu^{(1)}, \xi, \nu^{(2)})$.

In what follows, indicate by $\otimes$ the symbol of the Kronecker product and by I the identity matrix whose dimension is indicated by a suffix.

From the vectors of the stationary probabilities $\pi_i,\ i \geq 0$, some performance indices, listed as follows, are easily derived.

The distribution of the number of the requests in the orbit is

$$\lim_{t\to\infty} P\{i_t = i\} = \pi_i\mathbf{e},\ i \geq 0.$$

The average number of requests in the orbit is

$$L_{orbit} = \sum_{i=1}^{\infty} i\pi_i\mathbf{e}. \tag{1}$$

The average number of requests inside the system is

$$L = \sum_{i=0}^{\infty}\sum_{n=0}^{N}(i + n)\pi(i, n)\mathbf{e}. \tag{2}$$

The average number of busy servers is

$$N_{server} = \sum_{i=0}^{\infty}\sum_{n=1}^{N} n\pi(i, n)\mathbf{e}. \tag{3}$$

The average number of busy servers providing service to type-1 requests is

$$N_{server}^{(1)} = \sum_{i=0}^{\infty}\sum_{n=1}^{N}\sum_{l=0}^{\min\{n,M\}} (n - l)\pi(i, n, l)\mathbf{e}. \tag{4}$$

The average number of busy servers providing service to type-2 requests is

$$N_{server}^{(2)} = \sum_{i=0}^{\infty}\sum_{n=1}^{N}\sum_{l=1}^{\min\{n,M\}} l\pi(i, n, l)\mathbf{e} = N_{server} - N_{server}^{(1)}. \tag{5}$$

The intensity of output of type-i requests is

$$\lambda_{out}^{(i)} = \mu_1 N_{server}^{(i)}, i = 1, 2. \tag{6}$$

The intensity of output of requests from the system is

$$\lambda_{out} = \lambda_{out}^{(1)} + \lambda_{out}^{(2)}. \tag{7}$$

The loss probability of type-1 requests is

$$P_1^{(loss)} = \lambda_1^{-1}\sum_{i=0}^{\infty}\pi(i, N, 0)(D_1^{(1)} \otimes I_{2\bar{W}_2})\mathbf{e} = 1 - \frac{\lambda_{out}^{(1)}}{\lambda_1}. \tag{8}$$

The loss probability of type-2 requests is

$$P_2^{(loss)} = 1 - \frac{\lambda_{out}^{(2)}}{\lambda_2}. \tag{9}$$

The loss probability of an arbitrary request is

$$P^{(loss)} = 1 - \frac{\lambda_{out}}{\lambda}, \tag{10}$$

where $\lambda = \lambda_1 + \lambda_2$.

The probability of type-2 request loss upon arrival is computed by

$$P^{(ent-loss)} =$$

$$= (1-q)\lambda_2^{-1}\sum_{i=0}^{\infty}\left[\sum_{n=M_1}^{M-1}\pi(i,n)(I_{(n+1)\bar{W}_1}\otimes\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\otimes D_1^{(2)})\mathbf{e}+\right.$$

$$\left.+\sum_{n=M}^{N}\pi(i,n)(I_{(M+1)2\bar{W}_1}\otimes D_1^{(2)})\mathbf{e}\right]. \tag{11}$$

The probability that a type-2 request enters the orbit upon arrival is computed by

$$P^{(ent-to-orbit)} =$$

$$= q\lambda_2^{-1}\sum_{i=0}^{\infty}\left[\sum_{n=M_1}^{M-1}\pi(i,n)(I_{(n+1)\bar{W}_1}\otimes\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\otimes D_1^{(2)})\mathbf{e}+\right.$$

$$\left.+\sum_{n=M}^{N}\pi(i,n)(I_{(M+1)2\bar{W}_1}\otimes D_1^{(2)})\mathbf{e}\right]. \tag{12}$$

The rate of type-2 requests lost upon arrival is $\tilde{\lambda}_2 = P^{(ent-loss)}\lambda_2$.

The probability that an arbitrary type-2 request is forced to terminate service and goes into orbit is

$$P^{(termination-to-orbit)} = p\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{l=1}^{M}\pi(i,N,l)(D_1^{(1)}\otimes I_{2\bar{W}_2})\mathbf{e}. \tag{13}$$

The probability that an arbitrary type-2 request is forced to terminate service and is lost is

$$P^{(termination-loss)} = (1-p)\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{l=1}^{M}\pi(i,N,l)\otimes(D_1^{(1)}\otimes I_{2\bar{W}_2})\mathbf{e}. \tag{14}$$

The probability of an arbitrary type-2 request loss from the orbit is

$$P^{(loss-from-orbit)} = P_2^{(loss)} - P^{(ent-loss)} - P^{(termination-loss)}. \tag{15}$$

Notice that performance indices offer a wide spectrum for the analysis of the system. For a suitable explanation of the derivation of some indices, readers are addressed to (16).

5. Numerical results

This section deals with some numerical results, that show the dependence of the performances indices on the parameters of the hysteresis admission. The system performances in the case of non correlated flows and flows with different correlation levels are compared.

Consider a system with $N = 15$ servers. In what follows, we deal with three different cases, characterized by different correlations of the arrival flows and the same mean arrival intensity, $\lambda_1 = 2$ and $\lambda_2 = 5$:

- Case 1: Poisson flows.
- Case 2: MAP2, where matrices for the arrival processes $D_0^{(1)}$, $D_1^{(1)}$, $D_0^{(2)}$ and $D_1^{(2)}$ have order 2.
- Case 3: MAP5, where the previous matrices are of order 5 and higher correlation.

For case 2, for the arrival process of type-1 requests we have:

$$D_0^{(1)} = \begin{pmatrix} -7.44517 & 0.0000228358\\ 0.0000111687 & -0.000919517\end{pmatrix},$$

$$D_1^{(1)} = \begin{pmatrix} 7.44509 & 0.0000569601\\ 0.0000181491 & 0.0008902\end{pmatrix}.$$

Here, the squared coefficient of variation of inter-arrival times is $c_{var} = 7.51006$, while the coefficient of correlation of two neighbouring inter-arrival times is $c_{cor} = 0.483901$.

For the arrival process of type-2 requests, we set:

$$D_0^{(2)} = \begin{pmatrix} -19.6065 & 0.000590961\\ 0.00374586 & -0.203639\end{pmatrix},$$

$$D_1^{(2)} = \begin{pmatrix} 19.5781 & 0.027842\\ 0.00559836 & 0.194294\end{pmatrix}.$$

The arrival process of type-2 requests has $c_{var} = 2.45108$ and $c_{cor} = 0.463108$.

In case 3, the arrival process of type-1 requests is characterized by:

$$D_0^{(1)} = \begin{pmatrix} -4.5 & 4.5 & 0 & 0 & 0\\ 0 & -4.5 & 4.5 & 0 & 0\\ 0 & 0 & -4.5 & 4.5 & 0\\ 0 & 0 & 0 & -4.5 & 0\\ 0 & 0 & 0 & 0 & -9\end{pmatrix},$$

$$D_1^{(1)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 4.455 & 0 & 0 & 0 & 0.045\\ 0.09 & 0 & 0 & 0 & 8.91\end{pmatrix}.$$

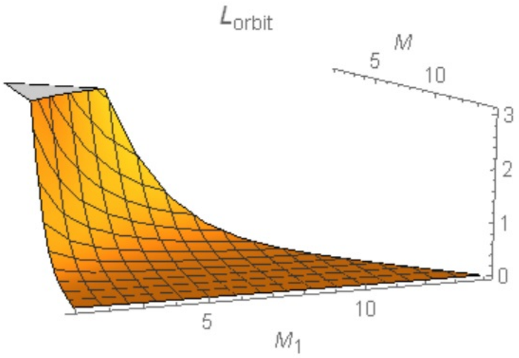

Figure 1. L_{orbit} versus M_1 and M in case 1.

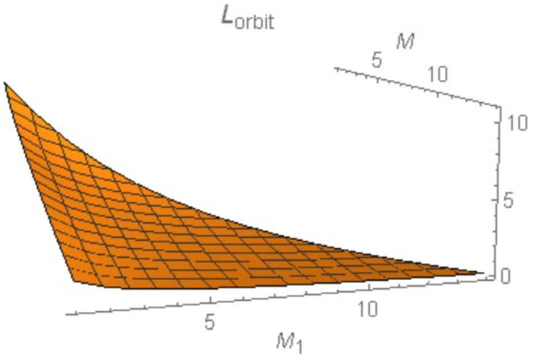

Figure 2. L_{orbit} versus M_1 and M in case 2.

The arrival process of type-2 requests presents:

$$D_0^{(2)} = \begin{pmatrix} -5.97 & 5.97 & 0 & 0 & 0 \\ 0 & -5.97 & 5.970 & 0 & 0 \\ 0 & 0 & -5.97 & 5.97 & 0 \\ 0 & 0 & 0 & -5.597 & 0 \\ 0 & 0 & 0 & 0 & -11.194 \end{pmatrix},$$

$$D_1^{(2)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 5.47267 & 0 & 0 & 0 & 0.124379 \\ 0.0572143 & 0 & 0 & 0 & 11.1369 \end{pmatrix}.$$

The arrival process of type-1 requests has $c_{var} = 1.02469$, and $c_{cor} = 0.578554$ while, for the one of type-2 requests, $c_{var} = 2.03612$ and $c_{cor} = 0.635934$.

In all three cases, the other parameters are: $\mu_1 = 2.5$, $\mu_2 = 1.5$, $q = 0.8$, $p = 0.2$, $\alpha = 1$ and $\gamma = 0.1$.

Notice that the performances of the system for type-1 requests do not depend on the values of M and M_1. Precisely, we have: for cases 1, 2 and 3, $N_{server}^{(1)} = 0.7999$, $\lambda_{out}^{(1)} = 1.9999$; the value of P_1^{loss} is: $P_1^{loss} = 1.20896 \cdot 10^{-14}$ in case 1; $P_1^{loss} = 4.999 \cdot 10^{-7}$ in case 2; $P_1^{loss} = 1.87134 \cdot 10^{-6}$ in case 3.

For variations of M over the interval $[1, N-1]$ and M_1 over the interval $[1, M]$, we consider Figures 1, 2, 3, 4, 5, 6, 7 and 8 that represent, respectively, for three different simulation cases, L_{orbit}, $N_{server}^{(2)}$, $P^{(ent-loss)}$ and $P^{(loss)}$.

Notice that L_{orbit}, i.e. the mean number of type-2 requests in the orbit, is very high for low values of M and M_1. Hence, type-2 requests have to wait for service in the orbit. When M and M_1 both grow, the access of type-2 requests is easier, and L_{orbit} decreases.

As for $N_{server}^{(2)}$, namely the mean number of busy servers that offer service to type-2 requests, it is low for low values of M and M_1. When M and M_1 increase, $N_{server}^{(2)}$ increases as more type-2 requests have access to the service upon arrival.

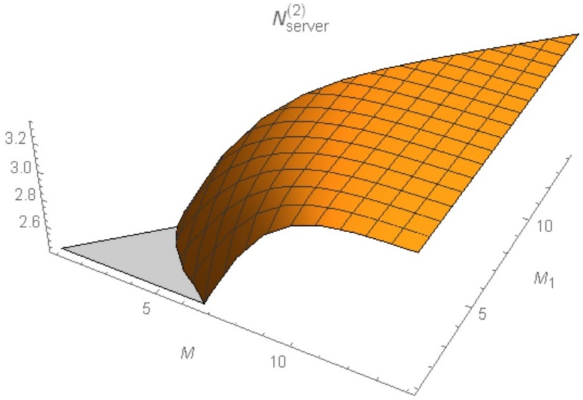

Figure 3. $N_{server}^{(2)}$ versus M_1 and M in case 1.

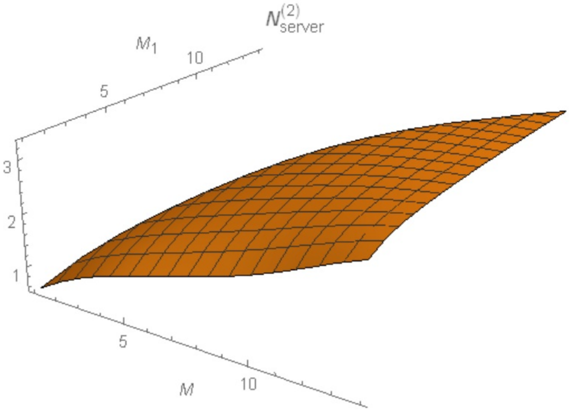

Figure 4. $N_{server}^{(2)}$ versus M_1 and M in case 3.

The probability $P^{(ent-loss)}$ of type-2 requests loss upon arrival is meaningful for small values of M and M_1. When M and M_1 get higher, $P^{(ent-loss)}$ decreases.

Finally, the loss probability $P^{(loss)}$ of an arbitrary request has a behaviour that depends on the number N of servers, chosen so that P_1^{loss} is very low (of order 10^{-4}). The loss probability decreases for high values of M and M_1.

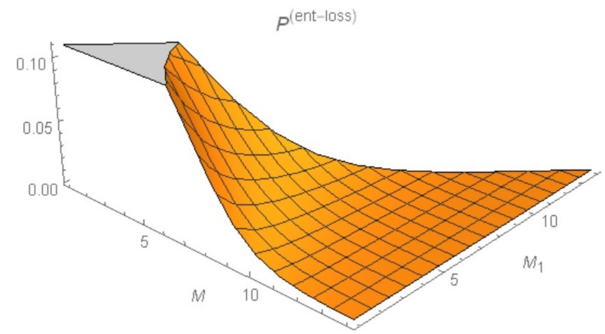

Figure 5. $P^{(ent-loss)}$ versus M_1 and M in case 1.

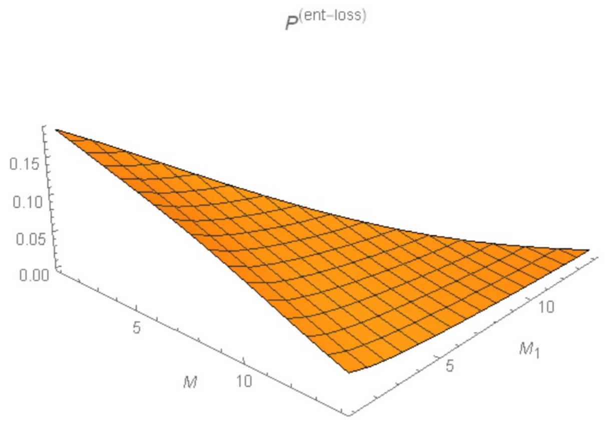

Figure 6. $P^{(ent-loss)}$ versus M_1 and M in case 2.

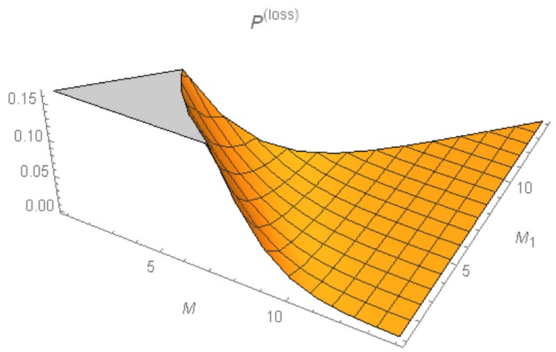

Figure 7. $P^{(loss)}$ versus M_1 and M in case 1.

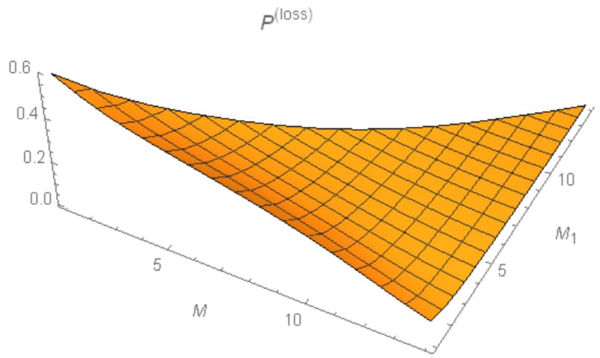

Figure 8. $P^{(loss)}$ versus M_1 and M in case 3.

Table 1. Values of $P_2^{(loss)}$ for different correlations

M	M_1	Case 1	Case 2	Case 3
1	1	0.814	0.911	0.826
2	2	0.600	0.838	0.667
3	2	0.485	0.802	0.595
4	3	0.295	0.728	0.479
5	4	0.162	0.655	0.379
6	5	0.082	0.584	0.291
7	6	0.038	0.515	0.215
8	8	0.014	0.423	0.134
9	7	8.848E-03	0.414	0.120
10	9	2.723E-03	0.323	0.067
11	10	0.975E-03	0.266	0.042
12	9	0.489E-03	0.260	0.036
13	12	0.101E-03	0.168	0.015
14	13	0.030E-03	0.130	0.099

Table 2. Values of $P^{(loss-from-orbit)}$ for different correlations

M	M_1	Case 1	Case 2	Case 3
1	1	0.626	0.718	0.639
2	2	0.433	0.650	0.500
3	2	0.334	0.618	0.440
4	3	0.181	0.549	0.340
5	4	0.087	0.484	0.256
6	5	0.039	0.421	0.184
7	6	1.64E-02	0.360	0.127
8	8	0.53E-02	0.282	0.071
9	7	0.34E-02	0.275	0.064
10	9	0.095E-02	0.202	0.032
11	10	0.33E-03	0.158	0.018
12	9	0.179E-03	0.155	0.016
13	12	0.326E-04	0.088	0.006
14	13	9.284E-06	0.062	0.003

Tables 1 and 2 report some values for indices, respectively, $P_2^{(loss)}$ and $P^{(loss-from-orbit)}$. There is a clear evidence of differences in the various simulation cases, as a consequence of correlated flows.

As for $P_2^{(loss)}$ and $P^{(loss-from-orbit)}$, they decrease when values of M and M_1 increase. The behaviour is very similar to $P^{(loss)}$. This is also a consequence of the probability $P^{(termination-loss)}$ of a generic type-2 request. Such probability increases when values of M and M_1 grow in all simulation cases. In fact, more type-2 requests are accepted inside the system despite many servers are busy.

6. Conclusions

In this paper, a multi-server queueing system, that provides service to two types of requests and models the operation of the cell of cognitive radio systems, has been presented. Requests of type-1 have preemptive priority over type-2 ones. Requests of type-2 obtain service through a hysteresis mechanism defined by two thresholds, that allow a possible rejection from the service. Rejected type-2 requests may abandon the system or retry for service after random intervals of time. Assuming fixed values for the thresholds:

- The dynamics of the system is described by a level dependent six-dimensional Quasi-Birth-and-Death process.
- Formulas for the main performance indices of the system are presented.
- Numerical results show the effectiveness of the strategy for the restriction of access of the requests of type-2 and the necessity of a careful account correlation in the arrival process. In particular, meaningful differences in performance indices can be noticed in the case of different correlation levels in the arrival flows.

Indeed, the obtained results show a practical relevance of the presented work, namely the possibility of defining a theoretical and numerical approach for the representation of real input flows for systems and networks. On the other hand, the analysis still presents some modelling limitations, that suggest suitable future research activities.

Some future work activities aim to extend the presented model. In this direction, the following possible different alternatives arise:

- The possibility of using different thresholds for the acceptance/rejection of requests of type-2 arriving from outside and from the orbit.
- A randomized procedure to drop requests of type-2 and to retry users in case of possible lack of servers.
- The substitution of fixed servers with a processor sharing discipline.
- The adoption of phase type distributions for service times.

References

1. Alipour-Vaezi M., Aghsami A. and Jolai F. (2022). Prioritizing and queueing the emergency departments patients using a novel data-driven decision-making methodology, a real case study. *Expert Systems with Applications*, 116568.
2. Bocharov, P.P., D'Apice, C., Manzo, R. and Pechinkin, A.V. (2007). Analysis of the Multi-server Markov Queuing System with Unlimited Buffer and Negative Customers. *Automation and Remote Control*, 68(1):85-94.
3. Brugno, A., D'Apice C., Dudin A.N., Manzo R. (2017). Analysis of an MAP/PH/1 Queue with Flexible Group Service. *International Journal of Applied Mathematics and Computer Science*, 27(1):119-131.
4. Brugno A., Dudin A.N. and Manzo R. (2017). Retrial Queue with Discipline of Adaptive Permanent Pooling. *Applied Mathematical Modelling (AMM)*, 50:1-16.
5. Brugno A., Dudin A.N. and Manzo R. (2018). Analysis of a Strategy of Adaptive Group Admission of Customers to Single Server Retrial System. *Journal of Ambient Intelligence & Humanized Computing*, 9(1):123-135.
6. Chen S., Wyglinski A., Vuyyuru R. and Altinas O. (2011). Feasibility analysis of vehicular dynamic spectrum access via queueing theory model. *IEEE Communications Magazine*, 49(11):156-163.
7. D'Apice C., Manzo R. and Piccoli B. (2013). Numerical schemes for the optimal input flow of a supply-chain, *SIAM Journal on Numerical Analysis (SINUM)*, DOI 10.1137/120889721, 51(5):2634-2650.
8. D'Apice C., Dudin A., Dudin S. and Manzo R. (2022). Priority queueing system with many types of requests and restricted processor sharing. *Journal of Artificial Intelligence and Humanized Computing*. DOI 10.1007/s12652-022-04233-w
9. D'Apice C., D'Arienzo M.P., Dudin A.N. and Manzo R. (2023). Admission control in priority queueing system with servers reservation and temporal blocking admission of low priority users, *IEEE Access*, 11:44425-44443.
10. D'Arienzo M.P., Dudin A.N., Dudin S.A. and Manzo R. (2020). Analysis of a Retrial Queue with Group Service of Impatient Customers. *Journal of Ambient Intelligence Humanized Computing*, 11(6), 2591-2599.
11. D'Arienzo M. P., Rarità L. (2019). Management of Supply Chains for the Wine Production. In: AIP Conference Proceedings, Volume 2293, 24 November 2020, Article number 420042, International Conference on Numerical Analysis and Applied Mathematics 2019.
12. de Falco M., Mastrandrea N. and Rarità L. (2017). A Queueing Networks-Based Model for Supply Systems, in: Sforza A., Sterle C. (eds) Optimization and Decision Science: Methodologies and Applications. ODS 2017. Springer Proceedings in Mathematics & Statistics, 217, Springer, Cham.
13. Dudin A. (1998). Optimal multithreshold control for a $BMAP/G/1$ queue with N service modes. *Queueing Systems*, 30(3): 273-287.
14. Dudin S., Dudin A., Kostyukova O. and Dudina, O. (2020). Effective algorithm for computation of the stationary distribution of multi-dimensional level-dependent Markov chains with upper block-Hessenberg structure of the generator. *Journal of Computational and Applied Mathematics*, 66(112425).
15. Dudin S., Kim C. and Dudina O. (2013). $MMAP/M/N$ queueing system with impatient heterogeneous customers as a model of a contact center. *Computers and Operations Research*, 40(7):1790-1803.
16. Dudin A.N., Klimenok V.I. and Vishnevsky V.M. (2020)

The theory of queuing systems with correlated flows. Springer Nature, Cham.

17. Dudin A.N., Lee M.H., Dudina O. and Lee S. K. (2016). Analysis of priority retrial queue with many types of customers and servers reservation as a model of cognitive radio system. *IEEE Transactions on Communications*: 65(1):186–199.
18. Dudin A.N., Manzo R. and Piscopo R. (2015). Single Server Retrial Queue with Group Admission of Customers. *Computers Operations Research (COR)*, 61:89-99.
19. Dudin A.N., Dudin S.A., Manzo R. and Rarità L. (2022). Analysis of Multi-Server Priority Queueing System with Hysteresis Strategy of Servers Reservation and Retrials. *Mathematics*, 10(3747):1-19, 2022.
20. Falin G. and Templeton J.G. (1997). *Retrial queues*, 75, CRC Press.
21. Goel S., Kulshrestha R. (2022). Dependability-Based Analysis for Ultra-reliable Communication in Heterogeneous Traffic Cognitive Radio Networks with Spectrum Reservation. *Wireless Personal Communications*, 1-25.
22. Kim C.S., Klimenok V., Birukov A. and Dudin A. (2006). Optimal multi-threshold control by the *BMAP*/*SM*/1 retrial system. *Annals of Operations Research*, 141(1):193-210.
23. Kim C., Dudin A., Dudin S. and Dudina O. (2016). Hysteresis control by the number of active servers in queueing system *MMAP*/*PH*/*N* with priority service. *Performance Evaluation*, 101:20-33.
24. Kim C., Dudin S., Dudin A. and Samouylov K. (2008). Multi-threshold control by a single-server queuing model with a service rate depending on the amount of harvested energy. *Performance Evaluation*, 127:1-20.
25. Klemm A., Lindermann C. and Lohmann M. (2003). Modelling IP traffic using the batch Markovian arrival process. *Performance Evaluation*, 54:149-173.
26. Klimenok V.I. and Dudin A.N. (2006). Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. *Queueing Systems*, 54:245-259.
27. Klimenok V., Dudin A. and Vishnevsky V. (2020). Priority multi-server queueing system with heterogeneous customers. *Mathematics*, 8(9):1501.
28. Lucantoni D. (1991). New results on the single server queue with a batch Markovian arrival process. *Communication in Statistics-Stochastic Models*, 7:1-46.
29. Maharaj B.T.J. and Awoyemi B.S. (2022). Developments in Cognitive Radio Networks. *Future Directions for Beyond 5G*. Springer.
30. Rarità L. (2022). A genetic algorithm to optimize dynamics of supply chains. In: L. Amorosi et al. (eds) Optimization in Artifical Intelligence and Data Sciences. AIRO Springer Series 8, 107 – 115, ISBN: 978-3-030-95380-5 (Ebook).
31. Sun B., Lee M.H., Dudin S.A. and Dudin A.N. (2014). Analysis of multiserver queueing system with opportunistic occupation and reservation of servers. *Mathematical Problems in Engineering*, 178108:1-13.
32. Zahed S., Awan I. and Cullen A. (2013). Analytical modeling for spectrum handoff decision in cognitive radio networks. *Simulation modelling practice and theory*, 38:98-114.
33. Zhu X.A., Shen L.A. and Yum T.-S. (2007). Analysis of cognitive radio spectrum access with optimal channel reservation. *IEEE Communications Letters*, 11(4):304-306.