IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 April 2023, accepted 1 May 2023, date of publication 4 May 2023, date of current version 10 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3273148

RESEARCH ARTICLE

Admission Control in Priority Queueing System With Servers Reservation and Temporal Blocking Admission of Low Priority Users

CIRO D'APICE[1], MARIA PIA D'ARIENZO[2], ALEXANDER DUDIN[3], AND ROSANNA MANZO[4]

[1]Dipartimento di Scienze Aziendali-Management and Innovation Systems (DISA-MIS), University of Salerno, 84084 Fisciano, Italy
[2]Dipartimento di Scienze Umane, Filosofiche e della Formazione, University of Salerno, 84084 Fisciano, Italy
[3]Department of Applied Mathematics and Informatics, Belarusian State University, 220030 Minsk, Belarus
[4]Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, 84084 Fisciano, Italy

Corresponding author: Rosanna Manzo (rmanzo@unisa.it)

ABSTRACT We analyse a cell of Cognitive Radio Network (*CRN*) as the multiline queueing system supplying service to two Markovian arrival flows of users. Primary (or licensed) users called as High Priority Users (*HPU*s) have a preemptive priority over the secondary (cognitive) users called as Low Priority Users (*LPU*s). The *HPU*s are dropped upon the arrival only if all servers are occupied by *HPU*s. If at the arrival epoch all servers are busy but some of them provide service to *LPU*s, service of one *LPU* is immediately interrupted and service of the *HPU* begins in the released server. A *LPU* is accepted only if the number of busy servers at arrival epoch is less than the defined in advance threshold M. Otherwise, the *LPU* is permanently lost or becomes a retrial user. A retrial user repeats attempts to receive service later after random time intervals. The *LPU* whose service is interrupted is either lost or transferred to a virtual place called as orbit. The users placed in the orbit may be impatient and can renege the system. The service time follows an exponential probability distribution with the rate determined by the user's type. After loss of a *HPU*, admission of *LPU*s is blocked. *LPU*s are informed that their access is temporarily suspended and do not generate new requests until blocking expires. The purpose of the research is the optimization of threshold M and admission blocking period duration. Behavior of the system is described by a multidimensional continuous-time Markov chain. Its generator, ergodicity condition and invariant distribution are derived. Expressions for performance indicators are given. Numerical results demonstrating usefulness of blocking and significance of account of correlation in arrivals are presented. E.g., in the presented example of cost criterion optimization blocking gives 18 percent profit comparing to the system without blocking.

INDEX TERMS Cognitive radio system, Markov arrival process, preemptive priority queueing system, servers reservation.

I. INTRODUCTION

A multiline queueing system with many different types of users and priority may be used effectively for modeling and performance assessment, capacity planning, and optimization of numerous real-world systems, see, e.g., recent paper [38]. In particular, these systems are now popular for description of emergency rooms, see, e.g. [1], [18], [21], and cognitive radio networks, see, e.g., surveys [2] and [36] and papers [9], [10], [11], [12], [23], [25], [29], [30], [31], [32], [34], [40], [42], [44], [45], [47]. There are many different possible variants of information transmission in *CRN*s which require consideration of different queueing systems as their descriptors, see, e.g., [36]. Here we impose the following assumptions about the scenarios of operation of *CRN*.

The associate editor coordinating the review of this manuscript and approving it for publication was Ding Xu.

A. ASSUMPTIONS ABOUT THE MODEL OF A CELL OF CRN

We assume that the bandwidth of the cell of the network is divided to equal sub-bands (channels) each of which can be

used by either *HPU* or *LPU*. The number of channels may be more than one. The case of one server, which is very popular in the literature, is hardly relevant to real *CRN*s. There is no buffer for waiting in case if the bandwidth is completely occupied. Spectrum sensing is ideal: *LPU*s are able to exactly determine/estimate the channel occupancy by *HPU*s or other *LPU*s. The cell operates in the overlay mode. This means that the service of one *LPU* has to be interrupted in case of *HPU* arrival when all channels are busy. There is a spectrum handoff: once the channel being used by a *LPU* is re-occupied by a *HPU*, the *LPU* can immediately occupy another idle channel if it is available. If idle channel is not available, the *LPU* can try to enter the service later on, after some random time interval. In addition to the given above assumptions, we suggest that in this cell there is an opportunity to broadcast to all users the status of the bandwidth. This status has two possible states: the bandwidth is available for sensing by *LPU*s or it is temporarily unavailable (blocked) for *LPU*s. This suggestion looks non-restrictive for radio access networks.

Thus, it is suggested in our model that the *HPU*s have absolute priority over the *LPU*s. A *HPU* is lost only if during his/her arrival epoch all servers are busy by processing of *HPU*s. If all servers are busy but some are satisfying requests of *LPU*s, the service of one *LPU* is interrupted and the server is occupied by the *HPU*. Forced cutoff of *LPU*'s service may have negative implications, including *LPU*'s discontent with the quality of service and waste of throughput owing to the loss of work previously done for the interrupted user's service. Thus, to mitigate these consequences, it is desirable to control the admission of *LPU*s. For example, it appears that it makes sense to temporarily halt *LPU*s admittance when the number of occupied servers exceeds some set level and, as a result, the risks of forced termination of *SPU* service are significant. The known in the literature policy of *LPU*s admission is described in the following subsection.

B. POLICY OF CHANNELS RESERVATION

The admission restriction policy via channels (servers) reservation was offered in the paper [46]. In this paper, the admission strategy assumes that any arriving *LPU* is admitted for service only if the occupied servers number is less than the fixed in advance integer number M, such as $0 < M \leq N$ where N is the servers number. It is numerically demonstrated in [46] that the good choice of the number $N - M$ of reserved servers leads to improvement of system performance.

In paper [39], the models considered in [46] and in the most part of the published papers devoted to the study of *CRN* were essentially generalized by assuming that:

- *HPU*s and *LPU*s arrive according to a Marked Markov Arrival Process (*MMAP*). It is the generalization of well known versatile Markov Arrival Process (*MAP*), see [6], [7], [8], [13], [33], [41], to the case of many types of users. The stationary Poisson arrival processes of *HPU*s and *LPU*s suggested in the most part of the relevant literature is the very simple case of the *MAP*. Its disadvantage is that it does not allow to take into consideration the possible correlation and possible high variability of intervals between arrivals that are common in modern communications networks and other real-world systems. This leads to huge under-estimation of the required amount of resources for service (bandwidth) under the fixed requirements to the quality of users service, e.g., *HPU*'s dropping probability, *LPU*'s blocking and forced termination probability, traffic throughput, etc.;
- the *LPU*, which is not admitted for service, may decide either to abandon the system or to transit to the orbit, a virtual place, and attempt to get access after a random time. The repeated attempts (retrials) are an inherent phenomenon of many telecommunication networks, see, e.g., [3], [19];
- the *LPU*s can be not absolutely persistent (can leave the system after any retrial failure) and (or) impatient (can depart from the system without receiving service after a certain random time of residing in orbit).

It is worth noting that because the *HPU*s have an absolute priority over the *LPU*s, the purpose of reservation of certain number of servers for service of only *HPU*s creates better conditions for service of *LPU*s, not *HPU*s. This is important because the phenomenon of forced termination of service (expelling from service) is very unpleasant for *LPU*s. Frequent expelling of *LPU*s from service may lead to their permanent refusal from service in the queueing system. This, in turn, reduces the system throughput and its economic revenue.

C. CONTRIBUTIONS OF THIS PAPER

The main contributions of this paper are as follows.

- We supplement the known in the literature mechanism of *LPU*s protection via reservation of some part of servers for service of only *HPU*s by the mechanism of temporal blocking of admission of *LPU*s after each occurrence of the loss of *HPU*. During the blocking time, information about unavailability of service is broadcasted and arriving *LPU*s do not even try to sense the channels. They can retry after an interval having a random length. Duration of blocking time is also random having exponential distribution and the rate of this distribution should be properly chosen based on the results of the presented analysis. After the blocking time expires, information about availability of service is broadcasted and arriving or retrying *LPU*s can try to sense the channels to get access.
- The proposed mechanism of blocking admission of *LPU* after the loss of *HPU* is novel in the literature and can be applied simultaneously with the servers reservation or separately. Its separate usage significantly simplifies control by *LPU*s admission comparing to the servers reservation. It is not necessary to permanently keep track on the number of busy servers. A *LPU* does not sense the

channels if the signal about admission blocking is sent. If the blocking is not currently imposed, the *LPU* can sense the channels and occupy any idle channel. If all channels are busy, the *LPU* departs from the system or retries later on. As it is shown in the numerical example in our paper, the blocking is helpful also in the situation when the reservation is not applied.

- We assume *MAP* flows of both types of users what is essential advantage over the vast majority of existing research, except the papers [10], [12], [14], [32], [38], [39].
- Consideration of possibility of *LPU* retrials in case of unavailability of servers or service interruption. This possibility is realistic in the majority of *CRN*s but is rarely considered in the literature. E.g., it is stated in paper [31] published in 2016 that retrial multi-server queueing models of *CRN* are not previously considered in the literature. This is not true because such a model was considered in 2014 in [39]. But indeed, there are only a few papers considering retrial multi-server queueing models of *CRN*. The reasons of such a situation consist of mathematical difficulty of analysis of the corresponding random process describing behavior of the queueing system.

D. CHALLENGES ASSOCIATED WITH CONSIDERATION OF THE MODEL

The main challenges are as follows.

- The first, albeit non-principal but essential, difficulty of analysis of the model consists of the necessity of consideration of a six-dimensional continuous-time Markov chain (*MC*) describing behavior of the considered model. Corresponding experience of work with structured matrices, entries of which, in turn, also are matrices with several nesting levels (like nesting dolls), is required during this analysis. In particular, it is required for derivation of explicit expression for the generator of the *MC*. The use of operations of Kronecker product and sum of the matrices and knowledge of their properties is very helpful at this stage of analysis.
- The second, principal, difficulty of analysis of the model consists of the necessity of the: (i) derivation of the constructive conditions for existence of the stationary probabilities of the states of the constructed *MC* and (ii) solution of an infinite system of equilibrium equations for these probabilities. For the corresponding system without retrials (with buffers), this difficulty drastically reduces through the use of famous results by M. Neuts for so called **level independent** Quasi-Birth-and-Death (*QBD*) processes, see [35]. The stationary distribution of these processes is computed in the matrix geometric form. The account of retrials makes the considered six-dimensional process be **level dependent** *QBD* process. Such processes are complicated and not enough well investigated in the existing literature. As popular reference to the methods of analysis of such processes, usually the paper [4] is cited. However, indeed this paper does not give the solution of the problem. Presented solution is given in terms of an infinite sequence of some matrices say, $\mathcal{R}_i, i \geq 1$, of a finite size computed as a solution of some infinite system of recursive matrix equations. But, even in the case of level independent *QBD* processes solution of the corresponding equations exists only under fulfillment of so called ergodicity condition. Thus, definitely in the case of level dependent *QBD* processes fulfilment of ergodicity condition is also mandatory for existence of the sequence of matrices $\mathcal{R}_i, i \geq 1$. Such a condition is given in [4] in terms of these matrices. Therefore, to verify ergodicity of the considered process, one needs to compute the sequence of matrices $\mathcal{R}_i, i \geq 1$. But to compute these matrices, one needs to be sure that the *QBD* process is ergodic. There is a vicious circle.
  In such a situation to avoid the described difficulty, researchers usually either make certain unrealistic assumptions about the system like "the total retrial rate does not depend on the number of retrying users" or "the number of retrying users is finite" (the later assumption is imposed in [31]) or make the rough or soft truncation of the state space of the *MC*. In this paper, we successfully derive ergodicity condition and compute the stationary distribution of the constructed *MC* using our experience of application of results from [26] relating to so called Asymptotically Quasi-Toeplitz *MC* and more recent enhancements of these results.
- The third, technical but unpleasant, difficulty consists of the necessity to operate with infinite size matrices consisting of infinitely many matrices of a finite, but large size. E.g., in the considered below numerical example of the system with $N = 20$ servers, without servers reservation and *MAP* processes of *HPU* and *LPU* determined by the matrices of size two the size of the finite blocks is 1848. Thus, careful realization of numerically stable algorithm from [26] is required.

E. ORGANIZATION OF THE PAPER

The paper is organized as follows. The mathematical model is reported in detail in section II. In section III, a multi-dimensional continuous-time *MC* with space heterogeneous transitions that describes the dynamics of the system is constructed. Necessary notation is introduced and the generator of *MC* is obtained as the block matrix there. The requirements to system parameters sufficient for ergodicity (stability) of *MC* are presented in section IV. Information relating to the invariant distribution of the system states calculation is given in brief and the expressions for key performance indicators are derived in section V. Section VI contains some numerical results. Section VII contains conclusion. Possible directions for generalization of the considered model are outlined in Section VIII.

II. DETAILED DESCRIPTION OF THE SYSTEM OPERATION

We analyse a queueing system characterized by N servers and no waiting space. High priority (Type-1 or *HPU*s) and low priority (Type-2 or *LPU*s) users arrive to the system according to Markov Arrival Process MAP_r, $r = 1, 2$, respectively. Such a process is governed by continuous-time MC $\nu_t^{(r)}$, $t \geq 0$, having a state space $\{0, \ldots, W_r\}$. The transitions rates of the process $\nu_t^{(r)}$ within its state space are given by the components of $D_0^{(r)}$ and $D_1^{(r)}$ matrices. The off-diagonal components of the matrix $D_0^{(r)}$ define the rates of jumps that are not accompanied by Type-r user arrival. The diagonal components of the matrix $D_0^{(r)}$ define the departure rate of the process $\nu_t^{(r)}$ from its states, while the $D_1^{(r)}$ entries define the transitions rates at which arrivals of Type-r users occur.

The matrix

$$D^{(r)}(1) = D_0^{(r)} + D_1^{(r)}$$

is the generator of the MC $\nu_t^{(r)}$. It is supposed to be irreducible. The average intensity of Type-r users arrival (fundamental rate) λ_r is calculated as

$$\lambda_r = \boldsymbol{\theta}^{(r)} D_1^{(r)} \mathbf{e},$$

where $\boldsymbol{\theta}^{(r)}$ is the (row) vector of the invariant probabilities of the MC $\nu_t^{(r)}$. It is the unique solution to the system

$$\boldsymbol{\theta}^{(r)} D^{(r)}(1) = \mathbf{0}, \ \boldsymbol{\theta}^{(r)} \mathbf{e} = 1$$

where $\mathbf{e}$ is a column vector of 1's with suitable size and $\mathbf{0}$ is a row vector of zeroes with suitable size.

Formulas for computation of the coefficient of variation of inter-arrival times and the coefficient of correlation of neighboring inter-arrival times as well as their derivation can be found in [13]. The values of these coefficients for the stationary Poisson arrival process are equal to 1 and 0, correspondingly. If these coefficients for the real-world arrival process are essentially different from the values 1 and 0, application of results obtained under assumption that the flow is the stationary Poisson can lead to significant errors in computation of performance measures of the system. The error is especially huge when the coefficient of variation is much more than 1 and (or) the coefficient of correlation is positive and greater than, say, 0.1. This is easily explained intuitively because the flows with such values of the coefficients of variation and correlation are bursty. Periods of time when the users arrive often and the system becomes congested alternate with periods of rare arrivals when the servers may stay idle and the throughput of the system is under-utilized. This motivates analysis of queueing systems with the *MAP*.

Various methods for the estimation of MAP_r parameters, depending on the set of observed user arrival moments (time stamps) in a real-world system, are pretty well developed, see, e.g., the paper [5].

Probability distribution of Type-r users service time is of exponential type having service rate μ_r, $r = 1, 2$.

We suppose that *HPU*s have absolute priority over *LPU*s. If at the arrival moment of a *HPU* all servers are occupied by *HPU*s the arriving *HPU* is not admitted to the system. In this case, *HPU* is dropped without receiving service (is lost). In the opposite case, *HPU* is admitted. If all servers are occupied, one of *LPU* is expelled from service.

By analogy with [39], we suppose that the policy of *LPU*s admission is of a threshold type. An integer threshold M is assumed to be fixed in advance such that $0 < M \leq N$. If $M = N$ the system is without reservation. The arriving *LPU* is admitted for service only when the number of idle servers is greater than $N - M$.

If *LPU* is accepted to the system, he/she starts service. If *LPU* is not accepted, he/she abandons the system (is lost) with probability $1 - q$, $0 \leq q \leq 1$. *LPU* will repeat attempts to receive access later on from so-called as orbit, with probability q.

Each user residing in the orbit makes the trials to receive access, independently of other users staying in the orbit. Individual inter-trials times are exponentially distributed with rate α, $\alpha > 0$. The total retrial rate is $i\alpha$ when i *LPU*s stay in the orbit, $i > 0$. A trial is successful if the number of servers that are not in use is larger than $N - M$. In this case, the user immediately starts service. If the trial has no success, the *LPU* leaves the system forever with probability $1 - q$ or returns back to the orbit for further retrials with probability q.

Any *LPU* may be expelled from service due to *HPU* arrival. In such a case, *LPU* either departs from the system with probability $1 - p$, $0 \leq p \leq 1$, or moves to the orbit with probability p. The users residing in the orbit may show impatience and depart without service when a random patience time expires. The cumulative distribution function of this time is

$$1 - \exp\{-\gamma t\}, \ t \geq 0, \ \gamma \geq 0.$$

If $\gamma = 0$ the users are completely patient and can leave the system only after the service.

As it has been noted above, the rejection of *LPU* upon arrival may be less offensive than the expelling from the ongoing service. This is because in the former case the user has an opportunity to immediately move for service to some alternative system or try to enter to the system later on when it will be less congested. While in the latter case he/she wastes some amount of time during the failed service. In turn, the owner or the manager of the system has a loss of some share of the system resource that was dedicated to supplying the failed service. Therefore, as an additional precautionary measure, we extend admission control strategy with the following feature.

Let the loss of an arbitrary *HPU* occur. This happens when at his/her arrival moment all servers provide service to *HPU*s. At this moment the admission blocking time starts. This time finishes after an interval length which has cumulative distribution function

$$1 - \exp\{-\sigma t\}, \ t \geq 0, \ \sigma \geq 0.$$

When the blocking is imposed, the *LPU*s receive announcement that their admission is temporarily suspended and do not try to enter to service. They are not counted as the rejected ones. During the blocking time, all transitions of the underlying process $\nu_t^{(2)}$ of the MAP_2, including the ones governed by the matrix $D_1^{(2)}$, do not lead to new arrivals. Note that we do not make any assumption about prolongation or resuming from the early beginning of the blocking time in case of a loss of the new *HPU* during the ongoing blocking time. This is due to the memoryless property of the exponential distribution of blocking time.

For easier understanding, the processing of *HPU*s and *LPU*s in the system is schematically illustrated in Figures 1 and 2.

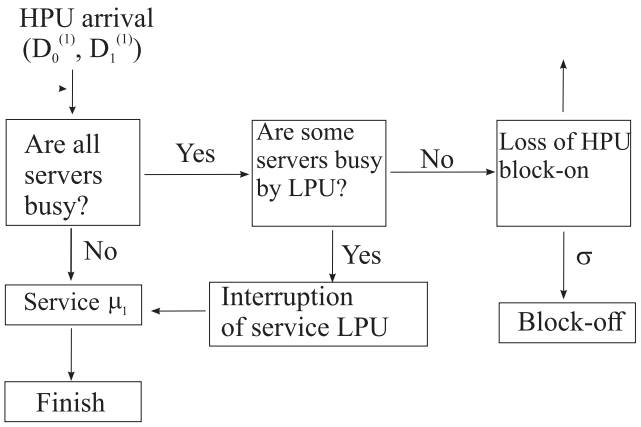


FIGURE 1. Scheme of processing of *HPU*.

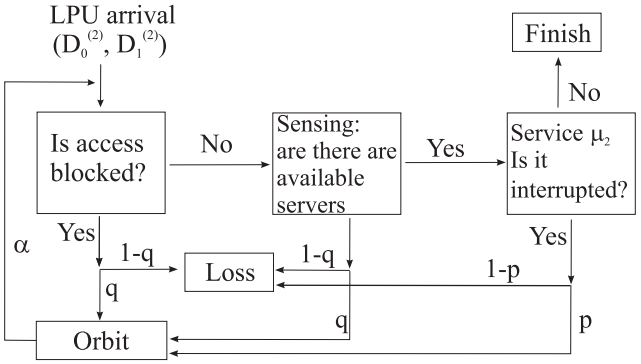


FIGURE 2. Scheme of processing of *LPU*.

To analyse the described queueing model of operation of the cell of *CRN*, in the next section we describe the dynamics of this model in terms of Markovian stochastic process.

III. RANDOM PROCESS OF THE SYSTEM STATES

Denote

- i_t, $i_t \geq 0$, the number of users placed in orbit,
- n_t, $n_t = \overline{0, N}$, the number of busy servers,
- l_t, $l_t = \overline{0, min\{n_t, M\}}$, the number of *LPU*s receiving service,
- $\nu_t^{(r)}$, $\nu_t^{(r)} = \overline{0, W_r}$, the state of the MAP_r, $r = 1, 2$, underlying process,
- ζ_t the blocking indicator: $\zeta_t = 0$ if admission of *LPU*s is blocked and $\zeta_t = 1$, otherwise

at an arbitrary epoch t, $t \geq 0$.

It can be verified that the stochastic process having one countable component and five finite components

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

is an irreducible continuous-time *MC*.

The states of the chain ξ_t are counted in the components' $(i, n, l, \nu^{(1)}, \zeta, \nu^{(2)})$ direct lexicographic order. A *sub-level* (i, n) is the set of states with the value (i, n) of two initial components and a *level* i is the set of sub-levels $((i, 0), \ldots, (i, N))$.

Let $\mathcal{Q}$ be the *MC* ξ_t, $t \geq 0$, generator. It is formed by the blocks $\mathcal{Q}_{i,j}$, which, in turn, contain the matrices $(\mathcal{Q}_{i,j})_{n,n'}$ consisting of the chain ξ_t transition rates from the sub-level (i, n) to the sub-level (j, n'), $n, n' = \overline{0, N}$. The matrices $\mathcal{Q}_{i,i}$ have negative diagonal elements. The diagonal entry modulus specifies the rate of the *MC*'s departure from the respective state that belongs to level i.

The following assertion is correct.

Lemma 1: Generator $\mathcal{Q}$ of the MC ξ_t has the following block-tridiagonal form:

$$\mathcal{Q} = \begin{pmatrix} \mathcal{Q}_{0,0} & \mathcal{Q}_{0,1} & O & O & \ldots \\ \mathcal{Q}_{1,0} & \mathcal{Q}_{1,1} & \mathcal{Q}_{1,2} & O & \ldots \\ O & \mathcal{Q}_{2,1} & \mathcal{Q}_{2,2} & \mathcal{Q}_{2,3} & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{1}$$

with blocks $\mathcal{Q}_{i,j}$, $i, j \geq 0$, which are not equal to zero, defined as follows:

- *The diagonal block $\mathcal{Q}_{i,i}$ is the sum of the matrix*

$$(1-q)\tilde{I} \otimes (I_{\bar{W}_1} \otimes \tilde{D}_1^{(2)}) + I_K \otimes (D_0^{(1)} \oplus \tilde{D}_0^{(2)})$$

and the block-tridiagonal matrix having the diagonal blocks $\mathbf{A}_i^{(n)}$, $n = \overline{0, N}$, given by formula

$$\mathbf{A}_i^{(n)} = \begin{cases} -(\mu_2\Omega_n + \mu_1\bar{\Omega}_n + i(\alpha+\gamma)I_{n+1}) \otimes I_{\bar{W}}, \\ \quad n = \overline{0, M-1}, \\ -(\mu_2\Omega_M + \mu_1\tilde{\Omega}_n + i((1-q)\alpha+\gamma)I_{M+1}) \\ \quad \otimes I_{\bar{W}} + \delta_{n,N}\Big[(1-p)E^- \otimes (D_1^{(1)} \otimes I_{2\bar{W}_2}) \\ \quad + \hat{I} \otimes (D_1^{(1)} \otimes \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \otimes I_{\bar{W}_2})\Big], \ n = \overline{M, N}, \end{cases}$$

the up-diagonal blocks $\mathbf{B}^{(n)}$, $n = \overline{0, N-1}$, given by formula

$$\mathbf{B}^{(n)} = \begin{cases} E_n^+ \otimes (I_{\bar{W}_1} \otimes \tilde{D}_1^{(2)}) \\ \quad + \hat{E}_n^+ \otimes (D_1^{(1)} \otimes I_{2\bar{W}_2}), & n = \overline{0, M-1}, \\ I_{M+1} \otimes (D_1^{(1)} \otimes I_{2\bar{W}_2}), & n = \overline{M, N}, \end{cases}$$

and the sub-diagonal blocks $\mathbf{F}^{(n)}$, $n = \overline{1,N}$, *given by formula*

$$\mathbf{F}^{(n)} = \begin{cases} (\mu_2\Omega_n\hat{E}_n^- + \mu_1\bar{\Omega}_n E_n^-) \otimes I_{\bar{W}}, & n = \overline{1,M}, \\ (\mu_2\Omega_M E^- + \mu_1\tilde{\Omega}_n) \otimes I_{\bar{W}}, & n = \overline{M+1,N}, \end{cases}$$

- *The up-diagonal block* $\mathcal{Q}_{i,i+1} = Q^+$ *is a diagonal matrix with diagonal blocks* $\mathbf{H}^{(n)}$, $n = \overline{0,N}$, *defined by formula*

$$\mathbf{H}^{(n)} = \begin{cases} O, & n = \overline{0,M-1}, \\ qI_{M+1} \otimes (I_{\bar{W}_1} \otimes \tilde{D}_1^{(2)}) & \\ \quad + \delta_{n,N} pE^- \otimes (D_1^{(1)} \otimes I_{2\bar{W}_2}), & n = \overline{M,N}; \end{cases}$$

- *The sub-diagonal block* $\mathcal{Q}_{i,i-1}$ *is a block two-diagonal matrix with diagonal blocks* $\mathbf{L}_i^{(0)}$, $n = \overline{0,N}$, *defined by formula*

$$\mathbf{L}_i^{(n)} = \begin{cases} i\gamma I_{n+1} \otimes I_{\bar{W}}, & n = \overline{0,M-1}, \\ i(\gamma + (1-q)\alpha) I_{M+1} \otimes I_{\bar{W}}, & n = \overline{M,N}, \end{cases}$$

and up-diagonal blocks $\mathbf{V}_i^{(n)}$, $n = \overline{0,N-1}$, *given by formula*

$$\mathbf{V}_i^{(n)} = \begin{cases} i\alpha E_n^+ \otimes I_{\bar{W}}, & n = \overline{0,M-1}, \\ O, & n = \overline{M,N-1}. \end{cases}$$

where we use the following notation:

- I *represents the identity matrix, while* O *represents a zero matrix. The suffix indicates the dimension of a matrix if it is not evident from context. E.g.,* $I_{\bar{W}_r}$ *means the identity matrix of size* $\bar{W}_r = W_r + 1$;
- $\mathbf{0}^T$ *represents the column vector generated by transposing the row vector* $\mathbf{0}$;
- $\otimes$ *and* $\oplus$ *denote matrices Kronecker product and sum, respectively, see [20], [22], and [43] and they are highly handy for determining the transition rates and probabilities of numerous independent MCs at the same time.*
- $\delta_{n,N}$ *are the Kronecker delta. If* $n = N$ *it is equal to 1, otherwise it is equal to 0;*
- $diag\{\ldots\}$ *are the diagonal matrix with diagonal elements given in parentheses;*
- Ω_l, $\bar{\Omega}_l$ *and* $\tilde{\Omega}_l$ *are the diagonal matrices defined as:*

$$\Omega_l = diag\{0, 1, \ldots, l\}, \ l = \overline{0,M},$$
$$\bar{\Omega}_l = diag\{l, l-1, \ldots, 0\}, \ l = \overline{0,M},$$

and

$$\tilde{\Omega}_l = diag\{l, l-1, \ldots, l-M+1, l-M\}, \ l = \overline{M,N};$$

- E_l^+ *and* $\hat{E}_l^+$, $l = \overline{0,M-1}$, *are the matrices obtained by attaching to the identity matrix* I_{l+1} *the zero column from the left and from the right, respectively;*
- E_l^- *and* $\hat{E}_l^-$, $l = \overline{0,M-1}$, *are the matrices obtained by attaching to the identity matrix* I_l *the zero row from below and from above, respectively;*
- E^- *is the matrix obtained by attaching to the matrix* $\hat{E}_M^-$ *the zero column from the right;*
- $\tilde{E}$ *is the matrix obtained by attaching to the matrix* E_M^- *the zero column from the left;*
- $\hat{I}$ *is the diagonal matrix of size* $M+1$ *defined as* $\hat{I} = diag\{1, 0, \ldots, 0\}$;
- $K = (M+1)(N+1-M/2)$;
- $\tilde{I}$ *is the diagonal matrix of size* K *with the first* $\frac{M(M+1)}{2}$ *diagonal components equal to 0 and the rest* $(M+1)(N-M+1)$ *diagonal components equal to 1;*
- $\bar{W} = 2\bar{W}_1\bar{W}_2$;
-

$$\tilde{D}_0^{(2)} = \begin{pmatrix} -\sigma I_{\bar{W}_2} + D_0^{(2)} + D_1^{(2)} & \sigma I_{\bar{W}_2} \\ O & D_0^{(2)} \end{pmatrix},$$
$$\tilde{D}_1^{(2)} = \begin{pmatrix} O & O \\ O & D_1^{(2)} \end{pmatrix}.$$

The validity of Lemma 1 is demonstrated by doing a comprehensive study of the transitions of the *MC* ξ_t within an infinitesimally tiny interval while taking into consideration the probabilistic meaning of the relevant matrices. The proof's sketch is below reported.

The block-tridiagonal form of the generator $\mathcal{Q}$ is explained by the fact that during a very short time the users number in the orbit may remain the same or change (either decrease or increase) by one.

The diagonal components of the diagonal blocks $\mathbf{A}_i^{(n)}$, $n = \overline{0,N}$, of the matrices $\mathcal{Q}_{i,i}$ added to the modules of the corresponding components of the blocks of the matrix

$$(1-q)\tilde{I}_K \otimes (I_{\bar{W}_1} \otimes \tilde{D}_1^{(2)}) + I_K \otimes (D_0^{(1)} \oplus \tilde{D}_0^{(2)})$$

define the processes' $\{n_t, l_t, \nu_t^{(1)}, \zeta_t, \nu_t^{(2)}\}$ departure rates from their states which are not accompanied by the transition of the component i_t into another state. Such exits are possible due to:

(i) finish of service of one *HPU* or *LPU*. The respective rates are determined by the elements of the matrices

$$\mu_2\Omega_n + \mu_1\bar{\Omega}_n \text{ when } n < M$$

or by the matrices

$$\mu_2\Omega_M + \mu_1\tilde{\Omega}_n \text{ when } M \le n \le N;$$

(ii) departure of one user from the orbit due to impatience or successful retrial or the failed retrial after which the user returns to the orbit but does not depart from the system (the respective rates are given by the value $i(\alpha+\gamma)$ when $n < M$ or the value $i((1-q)\alpha+\gamma)$ when $M \le n \le N$);

(iii) arrival of *LPU* which is rejected due to the supposed admission strategy. The respective rates are given by the elements of the matrix

$$(1-q)\tilde{I}_K \otimes (I_{\bar{W}_1} \otimes \tilde{D}_1^{(2)});$$

(iv) exit of the process $\{\nu_t^{(1)}, \zeta_t, \nu_t^{(2)}\}$ from its states. The respective rates are given by the diagonal elements of the matrix

$$I_K \otimes (D_0^{(1)} \oplus \tilde{D}_0^{(2)}).$$

Here, $\tilde{D}_0^{(2)}$ and $\tilde{D}_1^{(2)}$ describe the process of LPUs arrival defined by the two-dimensional MC $\{\zeta_t, \nu_t^{(2)}\}$. Their form, presented above, is explained as follows. Given the state 1 of the process ζ_t, arrivals of LPUs occur with the rates defined by the elements of the matrix $D_1^{(2)}$. Given the state 0 of the process ζ_t, arrivals of LPUs are blocked and all transitions of the underlying process $\nu_t^{(2)}$ defined by the matrix $D_0^{(2)} + D_1^{(2)}$ do not cause arrival of LPUs. Transition of the process ζ_t from the state 0 to the state 1 occurs with the rate σ when the blocking time expires. It is worth noting that the return of the process ζ_t from the state 1 to the state 0 occurs when the loss of HPU happens.

If $n = N$, there is also an additional option (v): arrival of HPU when all servers are occupied by service of HPUs and the arriving user is lost. The respective rates are given by the diagonal elements of the matrix

$$\hat{I} \otimes (D_1^{(1)} \otimes I_{\bar{W}_2}).$$

The off-diagonal elements of the matrices $\mathcal{Q}_{i,i}$ have the meaning of transition intensities of the components $\{n_t, l_t, \nu_t^{(1)}, \zeta_t, \nu_t^{(2)}\}$ without any jump of the component i_t of the MC ξ_t. They are defined by the off-diagonal elements of the blocks $\mathbf{A}_i^{(n)}$ as well as the corresponding elements of the blocks of the matrix

$$(1-q)\tilde{I}_K \otimes (I_{\bar{W}_1} \otimes \tilde{D}_1^{(2)}) + I_K \otimes (D_0^{(1)} \oplus \tilde{D}_0^{(2)}).$$

The non-zero off-diagonal elements of $\mathbf{A}_i^{(N)}$ are given by the off-diagonal elements of the matrix

$$(1-p)E^- \otimes (D_1^{(1)} \otimes I_{2\bar{W}_2}) + \hat{I} \otimes (D_1^{(1)} \otimes \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \otimes I_{\bar{W}_2}).$$

The first summand corresponds to situation when HPU arrives when all N servers are busy but some of them are occupied by LPUs. This service is interrupted and LPU does not join the orbit but abandons the system. The second summand reflects the situation when HPU arrives when all N servers provide service to HPUs. The arriving user is lost and simultaneously the state of the process ζ_t, which is responsible for keeping track whether or not arrivals of LPUs are blocked, admits the value 0 corresponding to the blocking state.

The up-diagonal blocks $\mathbf{B}^{(n)}$, $n = \overline{0, N-1}$, of the matrices $\mathcal{Q}_{i,i}$ contain transition rates of the process ξ_t that do not change the value of the component i_t but cause the increase of the number of busy servers. If $n < M$, this happens when HPU or LPU arrives to the system. The rates of the corresponding transitions are given by the elements of the matrices

$$E_n^+ \otimes (I_{\bar{W}_1} \otimes \tilde{D}_1^{(2)}) + \hat{E}_n^+ \otimes (D_1^{(1)} \otimes I_{2\bar{W}_2}).$$

The first summand here corresponds to LPU arrival. In this case, the component l_t of the MC ξ_t (the number of LPUs in service) increases by one as well as the component n_t (the total number of users in service). The second summand corresponds to HPU arrival. In this case, the component l_t does not change but the range of its possible values increases by one. If $M \le n \le N-1$, the increasing of the occupied servers number can occur only due HPU arrival.

The sub-diagonal blocks $\mathbf{F}^{(n)}$, $n = \overline{1, N}$, of the matrices $\mathcal{Q}_{i,i}$ contain rates of jumps of the process ξ_t that do not cause value change of the component i_t but imply the decrease of the number of occupied servers. Different form of these blocks in the cases $1 \le n < M$ and $M \le n \le N$ stems from the fact that in the former case the number of LPUs in service admits any value from the range $\{0, \dots, n\}$. In the latter case this range is $\{0, \dots, M\}$.

The up-diagonal blocks $\mathcal{Q}_{i,i+1}$ of the generator $\mathcal{Q}$ are diagonal matrices. This follows from the fact that the increasing of the users number in the orbit never occurs simultaneously with the occupation or release of servers. The increase can occur only due: (i) arrival of LPU when the occupied servers number is not less than M and this user joins the orbit but not permanently leaves the system. The respective rates of transitions are given by the elements of the matrix

$$qI_{M+1} \otimes (I_{\bar{W}_1} \otimes \tilde{D}_1^{(2)})$$

or (ii) arrival of HPU when all servers are occupied, but some of them provide service to LPUs. One of these users is pushed out and decides to join the orbit. The respective rates of transitions are given by the elements of the matrix

$$pE^- \otimes (D_1^{(1)} \otimes I_{2\bar{W}_2}).$$

The sub-diagonal blocks $\mathcal{Q}_{i,i-1}$ of the generator $\mathcal{Q}$ have the diagonal blocks $\mathbf{L}_i^{(n)}$, $n = \overline{0, N}$, and up-diagonal blocks $\mathbf{V}_i^{(n)}$, $n = \overline{0, N-1}$. The former blocks contain the rates of users departure from the orbit due to impatience (the rate is equal to $i\gamma$) or a failed retrial which are possible only when the number n of occupied servers is not less than M (the rate is equal to $i(1-q)\alpha$). The up-diagonal blocks contain the rates of successful retrials which lead to the increase by one in the number of serviced HPUs.

Lemma 1 is proven.

Remark 1: While the pair of matrices $(D_0^{(2)}, D_1^{(2)})$ describes the MAP_2, which defines the arriving flow of LPUs under control of the underlying process $\nu_t^{(2)}$, it is tempting to say that the pair of matrices $(\tilde{D}_0^{(2)}, \tilde{D}_1^{(2)})$ defines the MAP flow of unblocked LPUs under control of the two-dimensional underlying process $\{\zeta_t, \nu_t^{(2)}\}$. However, although the matrix $\tilde{D}_0^{(2)} + \tilde{D}_1^{(2)}$ is the generator, this would be not correct because the conventional definition of a MAP suggests that the underlying process does not have absorbing states. But the generator $\tilde{D}_0^{(2)} + \tilde{D}_1^{(2)}$ is reducible and the states $\{1, \nu^{(2)}\}$, $\nu^{(2)} = \overline{0, W_2}$, are the absorbing ones. In the considered queueing model, the exit from these states can occur only at the moments of the HPU's loss.

Remark 2: The discipline of finishing the blocking can be modified as follows. Admission of LPUs is resumed after expiration of this random blocking time or the decreasing of users number in the system to the value $M_1 - 1$ where $M_1 \le M$, whatever occurs first. In the case of such a modification,

only a small change of the generator is required. The block $\mathbf{F}^{(M_1)}$ should be defined by

$$\mathbf{F}^{(M_1)} = (\mu_2 \Omega_{M_1} \hat{E}_{M_1}^{-} + \mu_1 \bar{\Omega}_{M_1} E_{M_1}^{-}) \otimes I_{\bar{W}_1} \otimes \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \otimes I_{\bar{W}_2}.$$

Remark 3: The complete blocking of *LPU*s admission during the blocking time can be replaced with their partial (randomized) blocking. With the fixed probability u, $0 \leq u \leq 1$, an arbitrary *LPU* is admitted for the trial to receive service and with the complementary probability he/she is blocked. When $u = 0$, we have the model considered in this paper. When $u = 1$, we obtain the model studied in [39]. All the presented results remain valid in this case if expressions for the matrices $(\tilde{D}_0^{(2)}, \tilde{D}_1^{(2)})$ given above are replaced with the formulas

$$\tilde{D}_0^{(2)} = \begin{pmatrix} -\sigma I_{\bar{W}_2} + D_0^{(2)} + (1-u)D_1^{(2)} & \sigma I_{\bar{W}_2} \\ O & D_0^{(2)} \end{pmatrix},$$

$$\tilde{D}_1^{(2)} = \begin{pmatrix} uD_1^{(2)} & O \\ O & D_1^{(2)} \end{pmatrix}.$$

Having derived the explicit form of the generator $\mathcal{Q}$ of the *MC* ξ_t we can implement probabilistic analysis of the stationary behavior of this *MC*. As the first step in such an analysis, usually it is required to formulate conditions sufficient or necessary and sufficient for existence of the stationary (ergodic, invariant) distribution of the states of *MC*. We implement this step in the next Section.

IV. CONDITION OF EXISTENCE OF ERGODIC PROBABILITY DISTRIBUTION OF THE SYSTEM STATES

General methodology for analysis of multi-dimensional *MC*s with state inhomogeneous transitions suitable for analysis of *MC* ξ_t was elaborated in [26] in discrete and continuous time settings. In contrast to the mentioned above paper [4], generator of *MC* in [26] may be not triblockdiagonal matrix, but more general block-upper-Hessenberg matrix. In [4], no assumptions were imposed on the form of dependence of the blocks $\mathcal{Q}_{i,j}$ on i and j. Likely, this is the main reason why paper [4] does not present a constructive ergodicity condition for the considered chain. In [26], it is imposed a natural assumption about the asymptotic form of these blocks when i tends to infinity. We call this assumption natural because it is automatically fulfilled for *MC*s describing a wide range of retrial queues, queues with impatient users, tandem queues with infinite-server stations, etc. The class of the *MC*s introduced in [26] is called as asymptotically quasi-Toeplitz *MC*s (*AQTMC*s). To use results from [26] for derivation of ergodicity and non-ergodicity conditions of *MC* ξ_t and construction of the algorithm to compute its stationary distribution, firstly we have to verify that *MC* ξ_t belongs to the class of *AQTMC*.

Let us denote $R_i = -\mathcal{Q}_{i,i} \circ I$, $i \geq 0$, were $A_1 \circ A_2$ denotes Hadamard product (entrywise product or Schur product) of matrices A_1 and A_2. For definition and properties of Hadamard product of matrices see, e.g., [22]. This indicates that R_i is a diagonal matrix whose diagonal elements are supplied by the moduli of the matrix $\mathcal{Q}_{i,i}$, $i \geq 0$ corresponding diagonal components.

The presence of the following limits can by checked:

$$Z_0 = \lim_{i \to \infty} R_i^{-1} \mathcal{Q}_{i,i-1}, \; Z_1 = \lim_{i \to \infty} R_i^{-1} \mathcal{Q}_{i,i} + I,$$
$$Z_2 = \lim_{i \to \infty} R_i^{-1} \mathcal{Q}_{i,i+1}. \quad (2)$$

The values of the matrices Z_k, $k = 0, 1, 2$, distinguish here in two cases depending on the parameters of the system defining persistency and patience of users. These cases need a different treatment.

The Case 1 suggests that at least one of two relations is true: $\gamma > 0$ or $q < 1$, what means that some part of *LPU*s admitted to the system can leave it without receiving service. Case 2 suggests that $\gamma = 0$ and $q = 1$. This means that all *LPU*s admitted to the system must receive service.

In Case 1, it is possible to find that the limits Z_0, Z_1 and Z_2 are defined by:

$$Z_0 = \tilde{Z}_0 \otimes I_{\bar{W}}, \; Z_1 = O, \; Z_2 = O$$

where the matrix's $\tilde{Z}_0$ blocks are given as follows.

The diagonal block $(\tilde{Z}_0)_{n,n}$ of the matrix $\tilde{Z}_0$ equals $\frac{\gamma}{\gamma+\alpha} I_{n+1}$ for $n = \overline{0, M-1}$ and to I_{M+1} for $n = \overline{M, N}$.

The up-diagonal block $(\tilde{Z}_0)_{n,n+1}$ equals $\frac{\alpha}{\gamma+\alpha} E_n^+$ for $n = \overline{0, M-1}$ and to O_{M+1} for $n = \overline{M, N}$.

In Case 2, the limits Z_0, Z_1 and Z_2 are defined as follows.

- $$Z_0 = \mathbf{T}^{-1}(\hat{Z}_0 \otimes I_{\bar{W}})$$

  where

  $$\mathbf{T} = diag\{I_1, \ldots, I_M, T^{(M)}, \ldots, T^{(N)}\},$$

  the matrices $T^{(n)}$, $M \leq n \leq N$, are defined by

  $$T^{(n)} = (\mu_2 \Omega_M + \mu_1 \tilde{\Omega}_n) \otimes I_{\bar{W}} + I_{M+1} \otimes \Sigma_0 - \delta_{n,N} \hat{I} \otimes \Sigma_1$$

  where

  $$\Sigma_1 = -(D_1^{(1)} \otimes I_{2\bar{W}_2}) \circ I,$$
  $$\Sigma_0 = -(D_0^{(1)} \oplus \tilde{D}_0^{(2)}) \circ I.$$

  The blocks of the matrix $\hat{Z}_0$ are zero blocks except the up-diagonal blocks $(\hat{Z}_0)_{n,n+1} = E_n^+$ for $n = \overline{0, M-1}$.

- $$Z_1 = \tilde{I} \otimes I_{\bar{W}} + \mathbf{T}^{-1} \times \begin{pmatrix} O & \ldots & O & O & O & \ldots & O & O \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & \ldots & O & O & O & \ldots & O & O \\ O & \ldots & \mathbf{F}^{(M)} & \mathbf{A}^{(M)} & \mathbf{B}^{(M)} & \ldots & O & O \\ O & \ldots & O & \mathbf{F}^{(M+1)} & \mathbf{A}^{(M+1)} & \ldots & O & O \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & \ldots & O & O & O & \ldots & \mathbf{A}^{(N-1)} & \mathbf{B}^{(N-1)} \\ O & \ldots & O & O & O & \ldots & \mathbf{F}^{(N)} & \mathbf{A}^{(N)} \end{pmatrix}.$$

•

$$Z_2 = \mathbf{T}^{-1} diag\{O_{\frac{M(M+1)}{2}\bar{W}}, \mathbf{H}^{(M)}, \mathbf{H}^{(M+1)}, \dots, \mathbf{H}^{(N)}\}$$

where

$$\mathbf{A}^{(n)} = -(\mu_2\Omega_M + \mu_1\tilde{\Omega}_n)\otimes I_{\bar{W}} + I_{M+1}\otimes(D_0^{(1)} \oplus \tilde{D}_0^{(2)}) + \delta_{n,N}\Big[(1-p)E^- \otimes (D_1^{(1)} \otimes I_{\bar{W}_2}) + \hat{I} \otimes (D_1^{(1)} \otimes \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \otimes I_{\bar{W}_2})\Big],$$

$M \le n \le N$.

From existence of the limits Z_k, $k = 0, 1, 2$, it follows that MC ξ_t, $t \ge 0$, belongs to the class of continuous-time $AQTMC$ defined in [26]. As a consequence, the results obtained in [26] may be utilized to deduce the ergodicity condition for the MC ξ_t and compute its invariant distribution.

According to [26], the sufficient requirement for ergodicity of the $AQTMC$ ξ_t, $t \ge 0$, is the achievement of the relation

$$\mathbf{z}Z_0\mathbf{e} > \mathbf{z}Z_2\mathbf{e}, \tag{3}$$

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

$$\mathbf{z}(Z_0 + Z_1 + Z_2) = \mathbf{z}, \ \mathbf{z}\mathbf{e} = 1. \tag{4}$$

It is easy to check that in Case 1 inequality (3) reduces to inequality $1 > 0$ which is trivial. Therefore, the following assertion is valid.

Theorem 1: The fulfilment of at least one of the inequalities $q < 1$ or $\gamma > 0$ is sufficient for ergodicity of the MC ξ_t for any values of the system parameters.

In Case 2, the following statement is true.

Theorem 2: If $q = 1$ and $\gamma = 0$, then the MC ξ_t is ergodic under fulfilment of condition (3) in which $\mathbf{z}$ is the unique solution to the equations (4) where the explicit expressions for matrices Z_k, $k = 0, 1, 2$, are given above.

If

$$\mathbf{z}Z_0\mathbf{e} < \mathbf{z}Z_2\mathbf{e},$$

the chain is not ergodic.

The proof immediately follows from [26].

Remark 4: Verification of fulfilment of ergodicity condition can be performed via solution of the system (4) on computer and substitution of this vector to (3). Size of the vector $\mathbf{z}$, giving the solution of system (4), is equal to $K\bar{W}$. This size may be pretty large. In the case $\sigma = \infty$, i.e., blocking of arrival of LPUs is not performed, solution to equation (4) is found in [39] practically analytically as the Kronecker product of some vector of size $(N-M+1)(M+1)$, which gives the steady-state distribution of the number of servers satisfying requests of HPUs and LPUs when the system is overloaded, by the invariant probability vector $\boldsymbol{\theta}^{(2)}$ of underlying process of LPUs arrival.

In the model under study, further simplification of condition (3) - (4) by analogy with [39] is not possible due to the complex stochastic dependence of behavior of the underlying process of LPUs arrival $\{\zeta_t, \nu_t^{(2)}\}$ on the number of users in the system. This dependence is caused by the forced transition of the process ζ_t to state 0 after any loss of a HPU. Recall that such a loss occurs when HPU arrives and the component n_t of the MC ξ_t has the value N while the component l_t is equal to 0.

Remark 5: Because a smaller fraction of LPUs will successfully receive service in the considered model comparing to the one from [39] under the same values of the system parameters, the following is intuitively clear. The relatively simple sufficient ergodicity condition obtained in [39] is sufficient for the considered model as well. However, the opposite inequality in that condition can be not mandatory necessary for non-ergodicity of MC ξ_t.

V. COMPUTATION OF STEADY-STATE DISTRIBUTION OF THE *MC* AND PERFORMANCE INDICATORS OF THE SYSTEM

Suppose that the derived stability condition is satisfied. Then there are the following invariant probabilities of the MC ξ_t states:

$$p(i, n, l, \nu^{(1)}, \zeta, \nu^{(2)}) = \lim_{t\to\infty} P\{i_t = i, n_t = n, l_t = l, \nu_t^{(1)} = \nu^{(1)}, \zeta_t = \zeta, \nu_t^{(2)} = \nu^{(2)}\},$$
$$i \ge 0, \ n = \overline{0, N},$$
$$l = \overline{0, min\{n, M\}}, \ \nu^{(1)} = \overline{0, W_1}, \ \zeta = 0, 1, \ \nu^{(2)} = \overline{0, W_2}.$$

Let us form the steady-state probabilities row vectors p_i as follows: vector $p(i, n, l)$ combines the probabilities $p(i, n, l, \nu^{(1)}, \zeta, \nu^{(2)})$ enumerated in the lexicographic order,

$$p(i, n) = (p(i, n, 0), p(i, n, 1), \dots, p(i, n, min\{n, M\})), \quad n = \overline{0, N},$$
$$p_i = (p(i, 0), p(i, 1), \dots, p(i, N)), \ i \ge 0.$$

The probability vectors p_i, $i \ge 0$, satisfy the following system of linear algebraic equations:

$$(p_0, p_1, \dots)\mathcal{Q} = \mathbf{0}, \quad (p_0, p_1, \dots)\mathbf{e} = 1 \tag{5}$$

where the matrix $\mathcal{Q}$ is the generator of the MC ξ_t, $t \ge 0$.

Note, that in the case $q = 0$, $p = 0$ we have the model where the users never visit the orbit, i.e., a LPU is immediately lost if it arrives when the number of idle servers is less or equal to $N - M$ or its service is interrupted due to HPU arrival. In this case, instead of the infinite size generator $\mathcal{Q}$ one has the finite block $\mathcal{Q}_{0,0}$ and, thus, system (5) is finite. The probability vectors can be directly found via solving equation (5) on computer or using stable algorithms in [24] and [27].

In general case, the system (5) has infinitely many equations and unknowns and, therefore, cannot be solved directly on a computer without its truncation. If truncation is implemented, it is possible to obtain only an approximate solution to this system. Because the matrix $\mathcal{Q}$ has a tri-blockdiagonal structure, it is possible to recursively express all vectors

$p_i,\ i \geq 1$, via the vector p_0. However: (i) the recursion is numerically unstable and (ii) it is not clear how to compute the vector p_0. In [26], it was proposed not to solve the systems like (5) at all but to construct another system of equations for the unknown vectors $p_i,\ i \geq 0$. Such a construction is possible through the use of so called censored Markov chains. Thus, to compute vectors $p_i,\ i \geq 0$, we use the numerically stable techniques established in [16] and [26], which are geared to a more broad version of the generator $\mathcal{Q}$ (whose blocks above the up-diagonal blocks can be not equal to zero). A variant of the algorithm for a block-tridiagonal form of the generator $\mathcal{Q}$ is described, e.g., in [15] and [17]. A modification of this algorithm is also briefly reproduced in [14].

Having values of the vectors of the steady-state probabilities $\mathbf{p}_i,\ i \geq 0$, there is on opportunity to compute a variety of characteristics of the system's performance. Below we list some of them.

The probability distribution of the number of the *LPU*s in orbit is

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

The mean number of the *LPU*s in orbit is

$$L_{orbit} = \sum_{i=1}^{\infty} i\mathbf{p}_i\mathbf{e}.$$

The mean number of users in the system is

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

The mean number of occupied servers is

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

The mean number of occupied servers supplying service to *HPU*s is

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

The mean number of occupied servers supplying service to *LPU*s is

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

The departure rate of *HPU*s is

$$\lambda^{(1)}_{out} = \mu_1 N^{(1)}_{server}.$$

The departure rate of successful *LPU*s is

$$\lambda^{(2)}_{out} = \mu_2 N^{(2)}_{server}.$$

The total departure rate of serviced users is

$$\lambda_{out} = \lambda^{(1)}_{out} + \lambda^{(2)}_{out}.$$

The probability of *HPU*'s loss is

$$P^{(loss)}_1 = \lambda_1^{-1}\sum_{i=0}^{\infty}\mathbf{p}(i,N,0)(D^{(1)}_1 \otimes I_{2\bar{W}_2})\mathbf{e} = 1 - \frac{\lambda^{(1)}_{out}}{\lambda_1}.$$

Remark 6: Indeed, here we have two different formulas for calculation of the probability of *HPU*s' loss. One formula accounts that the loss occurs when the *HPU* arrives when all N servers are busy. Another formula accounts that this loss probability is the ratio of the rate of the lost *HPU*s to their arrival rate. Existence of two different formulas for calculation of the same probability is helpful for control of accuracy of computer realization of the algorithm for computation of the vectors of the stationary probabilities of the system states.

The probability of *LPU*'s loss or blocking is

$$P^{(loss)}_2 = 1 - \frac{\lambda^{(2)}_{out}}{\lambda_2}.$$

The probability of an arbitrary user loss is

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

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

The probability of arbitrary arriving *LPU* loss due to the reservation policy (the number of idle servers is not less than M) is

$$P^{(ent-loss)} = (1-q)\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{n=M}^{N}\mathbf{p}(i,n)(I_{(M+1)\bar{W}_1}\otimes\begin{pmatrix}0&0\\0&1\end{pmatrix}\otimes D^{(2)}_1)\mathbf{e}.$$

The probability that an arbitrary arriving *LPU* will join the orbit is

$$P^{(ent-to-orbit)} = q\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{n=M}^{N}\mathbf{p}(i,n)(I_{(M+1)\bar{W}_1}\otimes\begin{pmatrix}0&0\\0&1\end{pmatrix}\otimes D^{(2)}_1)\mathbf{e}.$$

The probability of blocking of an arbitrary *LPU* is

$$P^{(ent-block)} = \lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{n=0}^{N}\sum_{l=0}^{min\{n,M\}}\mathbf{p}(i,n,l)(I_{\bar{W}_1}\otimes\begin{pmatrix}1&0\\0&0\end{pmatrix}\otimes D^{(2)}_1)\mathbf{e}.$$

The rate of the blocking *LPU*s is equal to

$$\tilde{\lambda}_2 = P^{(ent-block)}\lambda_2.$$

The probability that an arbitrary *LPU* will be pushed out of the service and transit to orbit is

$$P^{(term-to-orbit)} = p\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{l=1}^{M}\mathbf{p}(i,N,l)(D^{(1)}_1\otimes I_{2\bar{W}_2})\mathbf{e}.$$

The probability that an arbitrary *LPU* will be pushed out of the service and will abandon the system is

$$P^{(term-loss)} = (1-p)\lambda_2^{-1}\sum_{i=0}^{\infty}\sum_{l=1}^{M}\mathbf{p}(i,N,l)\otimes(D^{(1)}_1\otimes I_{2\bar{W}_2})\mathbf{e}.$$

The loss probability of an arbitrary *LPU* from orbit is

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

The probability that an attempt of an arbitrary orbiting user will be not successful and he/she returns back to orbit is

$$P^{(return-to-orbit)} = q\tilde{\alpha}^{-1} \sum_{i=1}^{\infty} \sum_{n=M}^{N} i\alpha \mathbf{p}(i,n)\mathbf{e}$$

where $\tilde{\alpha} = \alpha L_{orbit}$.

The probability that an attempt of an arbitrary orbiting user will be not successful and he/she leaves the system without receiving service is

$$P_1^{(loss-from-orbit)} = (1-q)\tilde{\alpha}^{-1} \sum_{i=1}^{\infty} \sum_{n=M}^{N} i\alpha \mathbf{p}(i,n)\mathbf{e}.$$

The percentage of time, during which admission of *LPU*s is blocked, is

$$P_{block} = \sum_{i=0}^{\infty} \sum_{n=0}^{N} \sum_{l=0}^{min\{n,M\}} \mathbf{p}(i,n,l)(\mathbf{e}_{\bar{W}_1} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes \mathbf{e}_{\bar{W}_2}).$$

VI. OPTIMIZATION PROBLEM AND NUMERICAL EXAMPLE

A. GOALS OF THE NUMERICAL EXAMPLES AND INPUT DATA

The effectiveness of servers reservation for optimization of the model of a cognitive radio network cell was already illustrated by the numerical examples presented in [14] and [39]. The threshold strategy of reservation considered in [39] assumes that admission or rejection of a *LPU* is based on current relation of busy servers number and the fixed in advance threshold value. In [14], the hysteresis strategy was applied according to two thresholds. If the number of active servers rises beyond the upper threshold, admission of *LPU*s is halted; admission resumes when it falls below the lower level. The novel feature of the admission strategy considered in this paper is the possibility of additional temporal blocking of *LPU*s after the loss of a *HPU*.

The goals of this section are: (i) to show dependence of the main performance measures on the control parameters M and σ; (ii) to illustrate the profound effect of correlation in arrival processes; (iii) to show that the blocking can improve system operation quality even without servers reservation.

We consider three sets of the *MAP*s having the same average arrival rates of Type-1 and Type-2 users, respectively $\lambda_1 = 4/3$ and $\lambda_2 = 8/3$, but different values of the coefficient of correlation.

Set 1: Let MAP_1 be defined by the matrices

$$D_0^{(1)} = \begin{pmatrix} -1.8266 & 0.024 \\ 0.06515 & -0.12365 \end{pmatrix},$$
$$D_1^{(1)} = \begin{pmatrix} 1.7906 & 0.012 \\ 0.03257 & 0.02593 \end{pmatrix}.$$

This arrival process has a squared coefficient of variation of inter-arrival times $c_{var} = 2.7891$, and a coefficient of correlation of two neighbouring inter-arrival times $c_{cor} = 0.2874$.

The MAP_2 is defined by the matrices

$$D_0^{(2)} = \begin{pmatrix} -3.6174 & 0.012 \\ 0.03257 & -0.14957 \end{pmatrix},$$
$$D_1^{(2)} = \begin{pmatrix} 3.5814 & 0.024 \\ 0.06515 & 0.05185 \end{pmatrix}.$$

This arrival process has a squared coefficient of variation $c_{var} = 2.8781$, and a coefficient of correlation $c_{cor} = 0.1780$.

Set 2: Let MAP_1 be defined by the matrices

$$D_0^{(1)} = \begin{pmatrix} -4.687 & 0.122 \\ 0.0365 & -0.183 \end{pmatrix},$$
$$D_1^{(1)} = \begin{pmatrix} 4.544 & 0.021 \\ 0.016 & 0.1305 \end{pmatrix}.$$

This *MAP* has coefficient of variation of inter-arrival times $c_{var} = 2.982$, and coefficient of correlation of two neighbouring inter-arrival times $c_{cor} = 0.3993$.

The MAP_2 is defined by the matrices

$$D_0^{(2)} = \begin{pmatrix} -9.231, & 0.102 \\ 0.02 & -0.3135 \end{pmatrix},$$
$$D_1^{(2)} = \begin{pmatrix} 9.088 & 0.041 \\ 0.0325 & 0.261 \end{pmatrix}.$$

This *MAP* has the coefficient of variation of inter-arrival times $c_{var} = 3.0001$, and the coefficient of correlation $c_{cor} = 0.3830$.

Set 3: The MAP_1 and MAP_2 are defined as the stationary Poisson processes with rates λ_1 and λ_2 correspondingly. They have the coefficient of variation of inter-arrival times $c_{var} = 1$, and the coefficient of correlation $c_{cor} = 0$.

We set the number of servers N equal to 20. We vary the values of the threshold M from 16 to 20 and the values of the rate σ in the interval from 0.05 till 4.5 with step 0.05. We do not show the values for smaller values of σ which correspond to very long blocking time by two reasons. One is that the dynamics of the corresponding performance measures in this case is more or less clear. The second reason is the desire to avoid making the surfaces more flat due to the wide diapason of values of these measures.

The other system parameters are chosen as follows. Service rates are $\mu_1 = 1/10$ and $\mu_2 = 1/3$. Retrial and impatience rates are $\alpha = 3$ and $\gamma = 0.05$, respectively. The probabilities of moving to the orbit in case when arriving *LPU* sees no available to him/her server and returning to the orbit after unsuccessful retrial are fixed as $q = 0.7$ and $p = 0.3$, respectively.

B. ILLUSTRATION OF DEPENDENCE OF PERFORMANCE MEASURES ON CONTROL PARAMETERS M AND σ AND IMPORTANCE OF ACCOUNT OF CORRELATION IN ARRIVAL PROCESS

To evidently show high importance of account of correlation on arrival processes, first we present the values $P_1^{(loss)}$ of the probability of *HPU*'s loss. Due to preemptive priority of *HPU*s, this probability does not depend on control parameters M and σ. This probability is equal to 0.0000157 for the Set 3 of *MAP* (having zero correlation), 0.072379 for the Set 1 of *MAP* (having coefficients of correlation 0.2874 for *HPU*s and 0.1780 for *LPI*s), 0.323423 for the Set 2 of *MAP* (having coefficients of correlation 0.3993 for *HPU*s and 0.3830 for *LPU*s). Let us stress again that the corresponding *MAP*s have the same mean arrival rate but different values of the coefficients of variation and correlation of inter-arrival times. Huge difference in values of the loss probability clearly motivates analysis of queueing models with *MAP* arrival process.

Figures 3 - 5 show dependence of the average number L_{orbit} of *LPU*s in orbit on M and σ for flows from Sets 1 - 3. Figure 3 for the *MAP*s with zero correlation illustrates the evident fact that L_{orbit} decreases when the threshold M increases and *LPU*s have better access to service. The value of L_{orbit} is pretty small and is about 0.03 for $M = 16$, i.e. four servers are reserved for service of only *HPU*s. For Sets 1 and 2, this value is essentially larger. The value of L_{orbit} sharply decreases when σ is small, i.e., duration of blocking period is long and many *LPU*s do not try to enter the system and risk to go to the orbit.

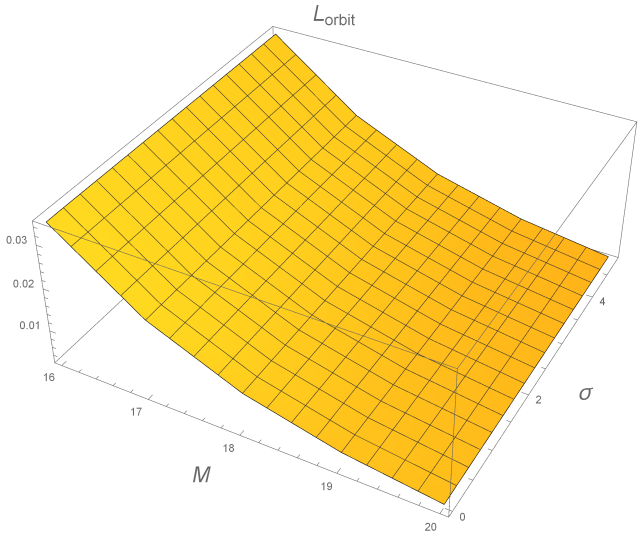


FIGURE 3. Dependence of average number L_{orbit} of *LPU*s in orbit on M and σ for flows with zero correlation.

Figures 6 - 8 show dependence of the average number N_{server} of busy servers on M and σ for flows from Sets 1 - 3. Figure 6 for the *MAP*s with zero correlation illustrates the evident fact that N_{server} increases when the threshold M increases and *LPU*s have better access to service and occupy servers. The value of N_{server} varies in the interval $[10.5 - 10.65]$. For Sets 1 and 2, again we observe the decreases of the considered performance measure with the decrease of σ. The value of N_{server} sharply decreases when σ is small, i.e., duration of blocking period is long and many *LPU*s do not try to enter the system and do not occupy servers. For the Set 2, the number of busy servers is essentially larger than for Set 3, it varies in the interval $[14.5 - 16.5]$. It might be expected that the number of busy servers will be larger for Set 2 with the highest correlation. However, this is not true. The number of busy servers for Set 2 varies in the interval $[11.5 - 12.5]$. Explanation of this fact follows from the increasing of loss probability $P_2^{(loss)}$ of *LPU*s loss or blocking for flows with higher correlation.

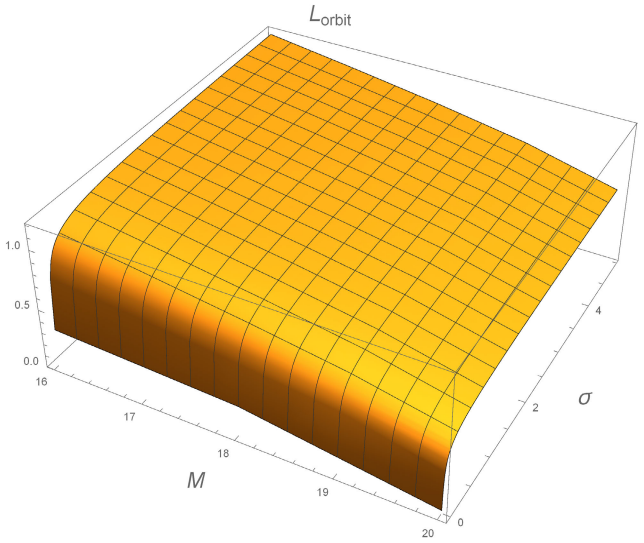


FIGURE 4. Dependence of average number L_{orbit} of *LPU*s in orbit on M and σ for flows with correlation 0.2874 in *HPU* arrival process.

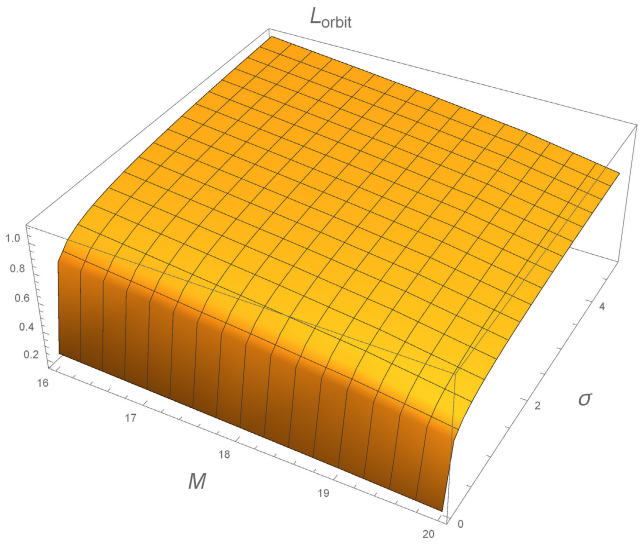


FIGURE 5. Dependence of average number L_{orbit} of *LPU*s in orbit on M and σ for flows with correlation 0.3993 in *HPU* arrival process.

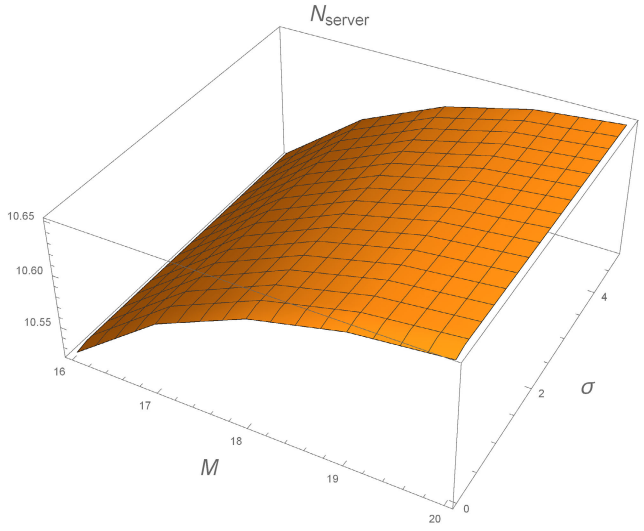


FIGURE 6. Dependence of average number N_{server} of busy servers on M and σ for flows with zero correlation.

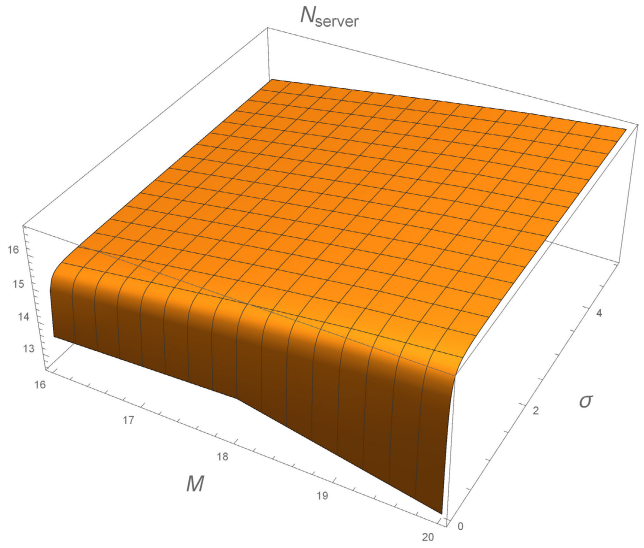


FIGURE 7. Dependence of average number N_{server} of busy servers on M and σ for flows with correlation 0.2874 in *HPU* arrival process.

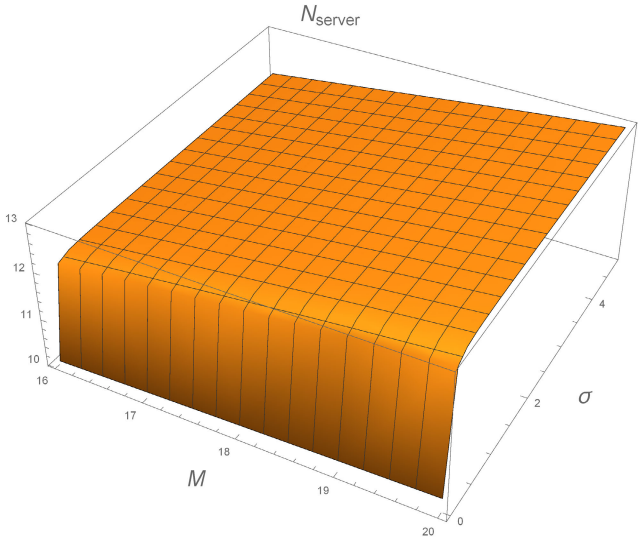


FIGURE 8. Dependence of average number N_{server} of busy servers on M and σ for flows with correlation 0.3993 in *HPU* arrival process.

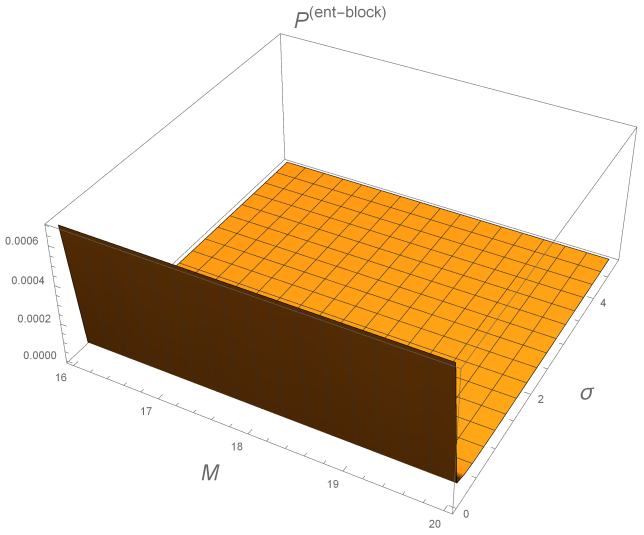


FIGURE 9. Dependence of the probability $P^{(ent-block)}$ on M and σ for flows with zero correlation.

Figures 9 - 11 show dependence of the probability $P^{(ent-block)}$ of blocking and arbitrary *LPU* upon arrival on M and σ for flows from Sets 1 - 3. It is clear from these figures that the probability $P^{(ent-block)}$ is very small (of order 10^{-4} for Set 3 and is quite essential (up to 0.1) for the Set 1. For the Set 2 it is twice larger. Therefore, as it is anticipated, introduction of blocking period causes blocking of some *LPU*s, while effect of blocking is small when the arrival flow is described by the stationary Poisson process.

Figures 12 - 14 show dependence of the loss probability $P^{(loss)}$ of an arbitrary user on M and σ for flows from Sets 1 - 3. Loss probability decreases with increase of M making easier access to service for *LPU*s. Again, loss probability $P^{(loss)}$ is very small (less than 0.02) for the case of the stationary Poisson arrival process and is essentially larger (up to 0.5 for Set 1 and up to 0.55 for Set 2). The loss probability $P^{(loss)}$ sharply increases when rate σ increases and the system becomes blocked during a long time.

Figures 15 - 17 show dependence of the loss probability $P^{(ent-loss)}$ of an arbitrary user upon arrival to the system on M and σ for flows from Sets 1 - 3. Loss probability decreases with increase of M making easier access to service for *LPU*s. Again, loss probability $P^{(ent-loss)}$ is very small (less than 0.015) for the case of the stationary Poisson arrival process and is essentially larger (up to 0.2 for Set 1 and Set 2). Smaller values of $P^{(ent-loss)}$ comparing to the probability $P^{(loss)}$ is easily explained by the possibility of the users loss not only at an arrival moment but also due to service interruption of

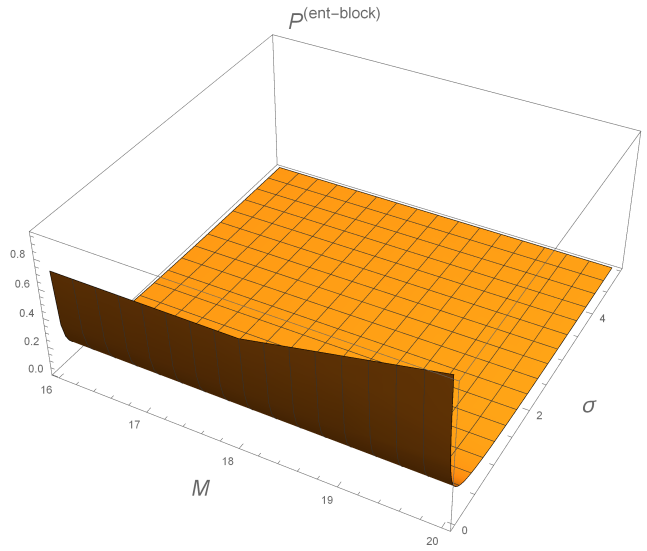


FIGURE 10. Dependence of the probability $P^{(ent-block)}$ on M and σ for flows with correlation 0.2874 in *HPU* arrival process.

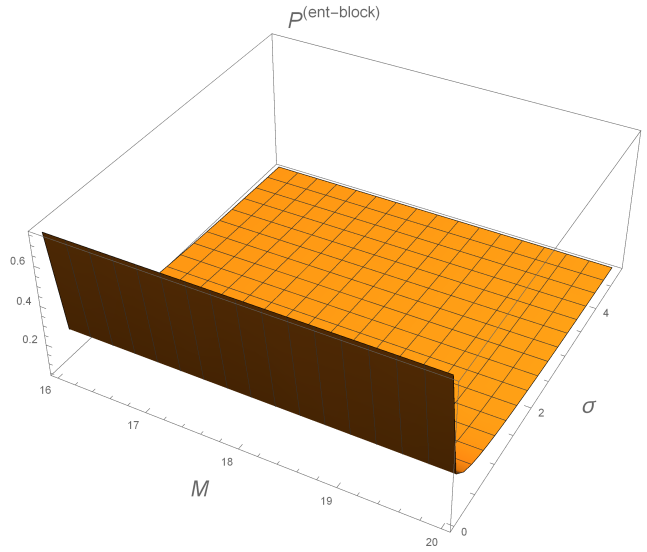


FIGURE 11. Dependence of the probability $P^{(ent-block)}$ on M and σ for flows with correlation 0.3993 in *HPU* arrival process.

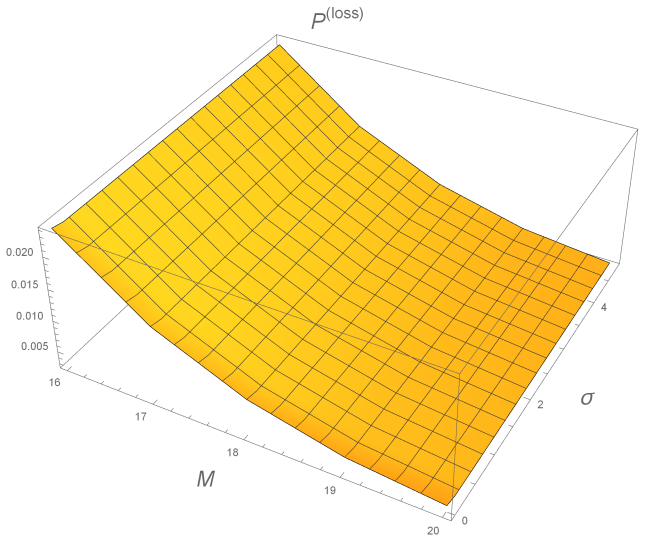


FIGURE 12. Dependence of the loss probability $P^{(loss)}$ on M and σ for flows with zero correlation.

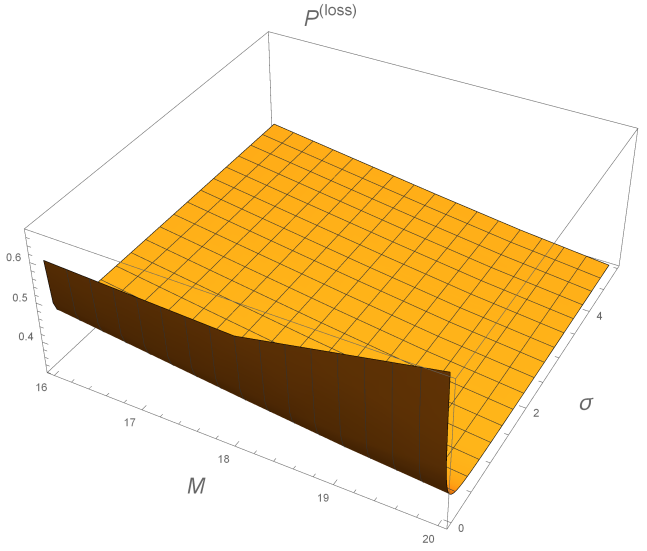


FIGURE 13. Dependence of the loss probability $P^{(loss)}$ on M and σ for flows with correlation 0.2874 in *HPU* arrival process.

LPU by *HPU* arrival and its immediate loss or loss after one or several retrials from the orbit due to servers unavailability.

Figures 18 - 20 show dependence of the loss probability $P_2^{(loss)}$ of an *LPU* on M and σ for flows from Sets 1 - 3. Behavior of this loss probability is similar to the behavior of the corresponding surfaces given on Figures 15 - 17. But the loss probability $P_2^{(loss)}$ is essentially larger than loss probability $P^{(ent-loss)}$ because the later probabilities relate to the loss of both *HPU*s and *LPU*s while, as we saw above, loss probability $P_1^{(loss)}$ of *HPU*s is small, especially for the Sets with low correlation in arrival process. Recall of one third of arriving users are *HPU*s.

Based on the presented in this subsection numerical results, we can summarize the following:

1) Correlation in arrival process drastically changes the values of performance measures of the system.

2) Blocking of servers has small effect in the case of the stationary Poisson arrival process and essential effect in case of flows with high correlation and large variation of inter-arrival times.

3) Long duration of blocking periods leads to the decrease of congestion in the system (smaller average number of *LPU*s retrying from the orbit) but leads to lower load of the servers and smaller throughput of the system.

C. ILLUSTRATION OF POSSIBILITY OF OPTIMAL CHOICE OF THE DURATION OF BLOCKING PERIOD

As it was stressed above, the main novelty of the considered model is introduction of the blocking of acceptance of *LPU*s

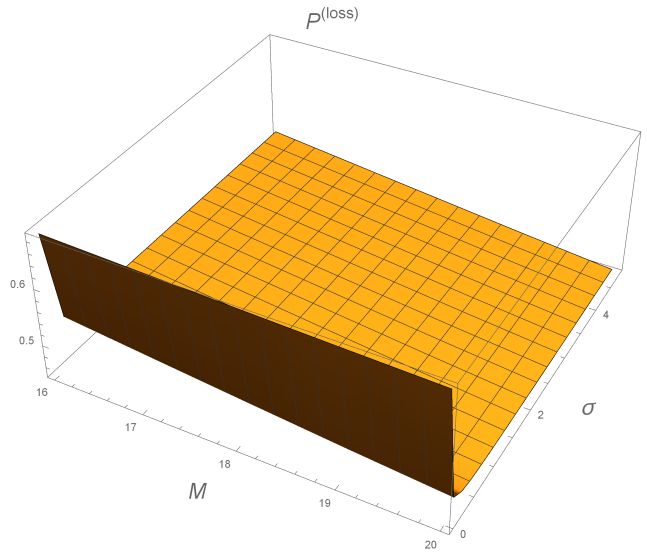


FIGURE 14. Dependence of the loss probability $P^{(loss)}$ on M and σ for flows with correlation 0.3993 in *HPU* arrival process.

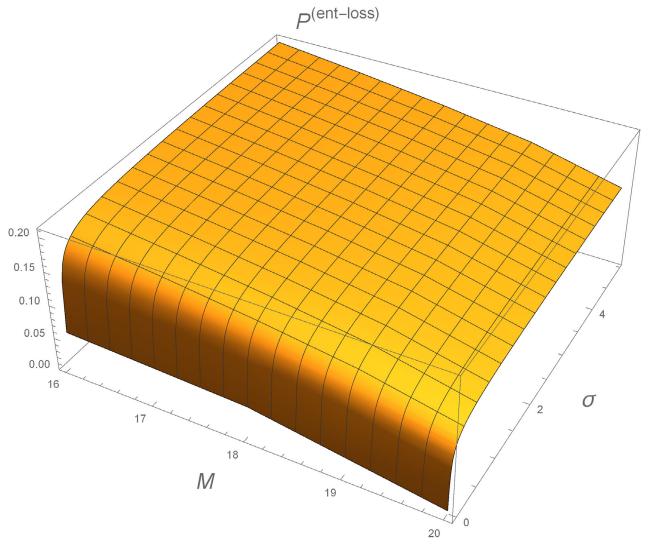


FIGURE 16. Dependence of the loss probability $P^{(ent-loss)}$ on M and σ for flows with correlation 0.2874 in *HPU* arrival process.

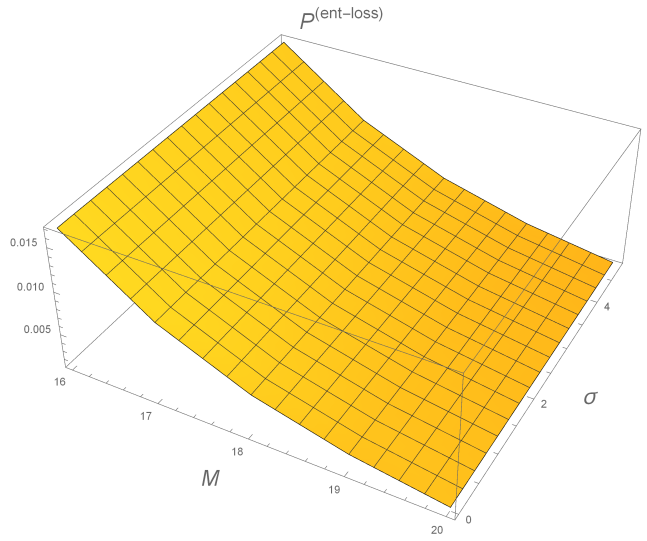


FIGURE 15. Dependence of the loss probability $P^{(ent-loss)}$ on M and σ for flows with zero correlation.

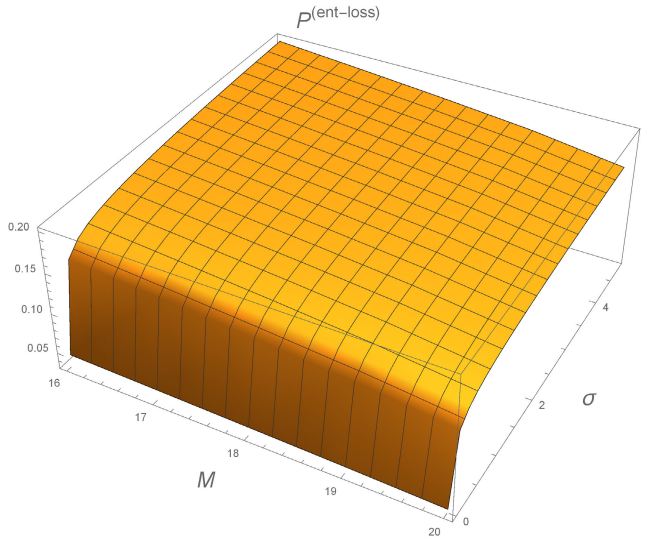


FIGURE 17. Dependence of the loss probability $P^{(ent-loss)}$ on M and σ for flows with correlation 0.3993 in *HPU* arrival process.

after each loss of a *HPU*. The goal of the blocking is to reduce the probability of interruption of ongoing service of *LPU* by the arrival of *HPU* to the fully occupied system. As indicator of a full occupation of the system, we consider the event of the loss of a *HPU* (because he/she met all servers busy by *HPU*s). It is intuitively clear that after this event occurrence it makes sense to temporarily block arrival of *LPU* because there is a high probability that new arriving *LPU*s will be lost. Even if they will be not lost immediately upon arrival, there is high chance that soon they will be lost due to the service interruption by arrival of *HPU*.

The question, which has to be answered based on the results of the implemented in this paper analysis is: "What is the optimal value of the mean duration of a blocking period?" Because *HPU*s have an absolute priority, service of *LPU*s does not effect service quality of *HPU*. Thus, the optimal duration of blocking period should be defined based on the performance characteristics of service of *LPU*s. From the point of view of the eventual result (access for *LPU* is denied), blocking of access and rejection upon arrival have the same effect. The difference between the blocking and rejection is essential from the psychological point of view. Blocking of *LPU*s access is assumed to be announced to *LPU*s which know that they should not try to obtain access. Existence of blocking periods can be directly indicated in Service Level Agreement between the service provider and *LPU*, if the *LPU* rents a channel along to the *HPU*s, as well as the possible mean duration or frequency of occurrence of

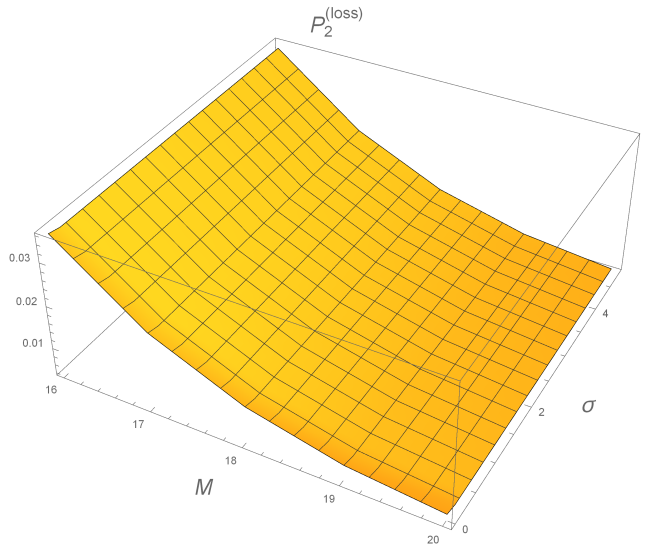


FIGURE 18. Dependence of the loss probability $P^{(ent-loss)}$ on M and σ for flows with zero correlation.

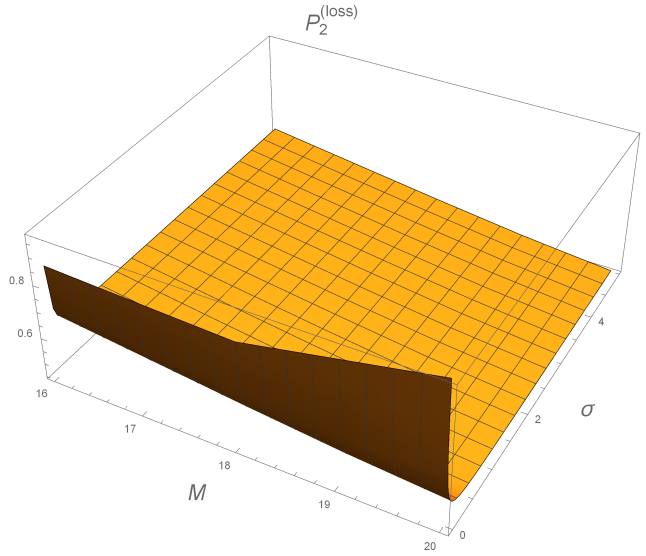


FIGURE 19. Dependence of the loss probability $P^{(ent-loss)}$ on M and σ for flows with correlation 0.2874 in *HPU* arrival process.

these periods. Therefore, the LPU must be tolerant to access denial due to the temporal blocking. Conversely, occasional rejection in access without the preliminary warning can have an irritating effect and the LPU can permanently stop the use of the service by a given provider.

It is clear that when the blocking period is short (this corresponds to a large value of the rate σ), more LPUs will be rejected than blocked. With the increase of the blocking period duration, the share of the blocked LPUs will increase. When the blocking period is long, more LPUs will be blocked than rejected.

To account the explained above different psychological and economical impact of blocking and rejecting, we will optimize the choice of the mean duration of the blocking

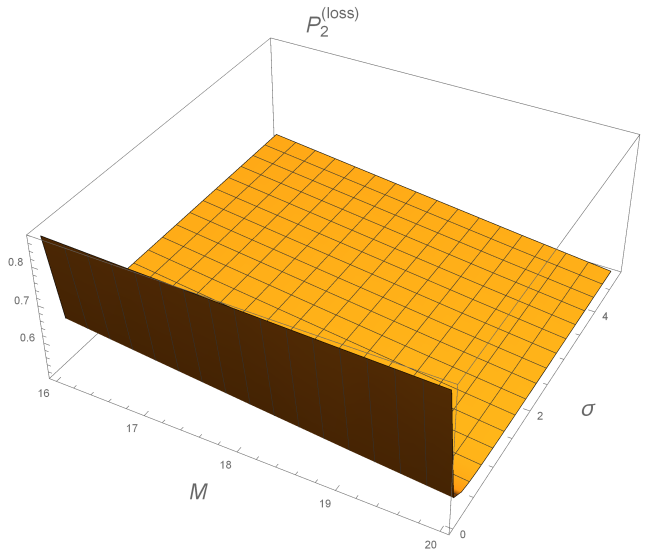


FIGURE 20. Dependence of the loss probability $P^{(ent-loss)}$ on M and σ for flows with correlation 0.3993 in *HPU* arrival process.

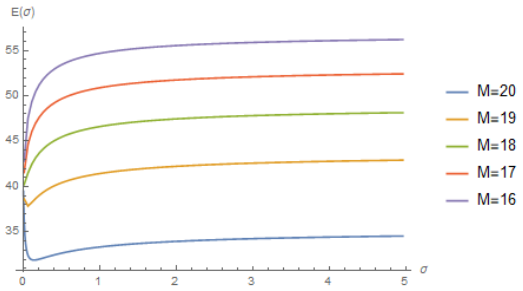


FIGURE 21. Dependence of cost criterion $E(\sigma)$ values for the Set 1 of *MAP*s for different values of M on σ.

period aiming to minimize the following cost criterion:

$$E(\sigma) = c_1 P^{(ent-loss)}(\sigma) + c_2 P^{(ent-block)}(\sigma),$$

where c_1 is the charge paid due to the blocking access to an arbitrary LPU and c_2 is the charge paid due to the rejection of an arbitrary LPU. The values of the cost coefficients are chosen as $c_1 = 400$ and $c_2 = 30$.

Below we illustrate behavior of the function $E(\sigma)$ for Sets 1 and 2 of the MAPs having the same rates of arrival of both types of users but different values of the coefficient of correlation of successive inter-arrival times in the arrival process.

Figure 21 illustrates the behaviour of cost criterion on the parameter σ for the values of commonly available servers $M = 16, 17, 18, 19, 20$ for the Set 1.

It is evidently seen from Figure 21 that the control by the quality of service via variation of duration of blocking period is more effective in case of small number $N - M$ of reserved servers. The most interesting case is $M = N$, i.e., permanent monitoring of the number of busy servers is not performed and only temporal blocking of access is the tool

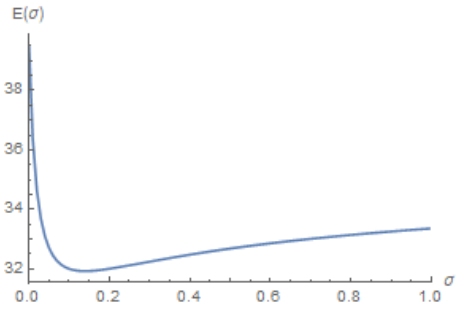


FIGURE 22. Cost criterion $E(\sigma)$ values for the Set 1 of *MAP*s for $\sigma \in [0.001, 1]$.

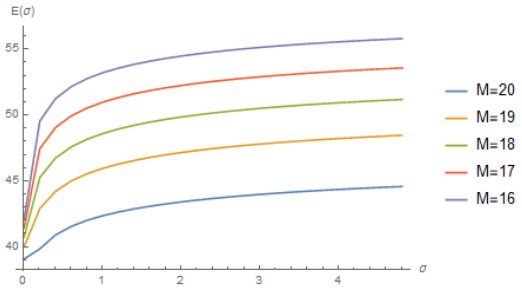


FIGURE 23. Dependence of cost criterion $E(\sigma)$ values for the Set 2 of *MAP*s for different values of M on σ.

to decrease probabilities of LPU loss upon arrival or due to service interruption.

Figure 24 shows the behaviour of cost criterion on the parameter σ for $M = N = 20$ for the Set 1 of *MAP*s.

The minimum value of the criterion $E(\sigma)$ is 31.9256. It is attained for $\sigma = 0.14$. The values of the criterion $E(\sigma)$ for small and large values of σ are as follows:

$$E(0.001) = 39.4771, \ E(100) = 35.1811.$$

Figure 23 illustrates the behaviour of cost criterion on the parameter σ for the values of commonly available servers $M = 16, 17, 18, 19, 20$ for the Set 2.

Figure 24 shows the behaviour of cost criterion on the parameter σ for $M = N = 20$ for the Set 2 of *MAP*s.

The minimum value of the criterion $E(\sigma)$ is 38.4793. It is attained for $\sigma = 0.04$. The values of the criterion $E(\sigma)$ for small and large values of σ are as follows:

$$E(0.001) = 39.8614, \ E(100) = 45.4051.$$

Based on the presented in this subsection numerical results, we can summarize the following:

1) When the average duration of the blocking period is suitably chosen, the use of blocking LPU arrival provides for a higher level of system operating quality. The profit might be greater than 10 percent.

2) It is crucial to take into account correlation in the arrival process. The *MAP*s from the second set have the same rates as the *MAP*s from the first set. But they have higher correlation. This implies worse quality of the system operation.

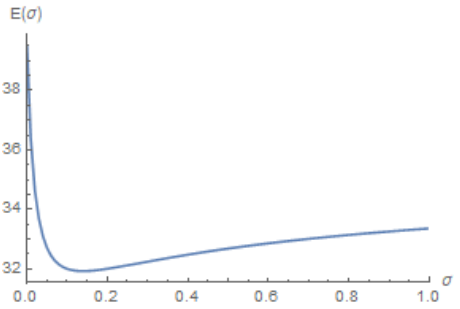


FIGURE 24. Cost criterion $E(\sigma)$ values for the Set 1 of *MAP*s for $\sigma \in [0.001, 1]$.

Minimal value of the cost criterion for more correlated second set is essentially larger.

3) Correlation has significant impact on the optimal value of σ (and average value σ^{-1} of duration period). For smaller correlation, the optimal duration is about 7. For larger correlation, the optimal duration is about 25. This is intuitively explained as follows. Larger correlation implies more irregular, bursty, arrival of LPUs. Periods of rare arrivals, during which the servers may be under-loaded and stay idle, alternate with periods of very frequent arrivals when congestion occurs. Therefore, to avoid many rejections during the periods of frequent arrivals it is reasonable to block the admission during the longer time.

4) Among two, not optimal, options: to have very long or very short blocking periods, in case of smaller correlation the later option is more preferable. In case of larger correlation, the former option is more preferable. Therefore, effectiveness of blocking is especially high in case of flows with high correlation of inter-arrival times. For the second set of the *MAP*s, the profit is about 18 percent.

5) It is not possible to formulate some simple recommendation (so-called "rule of thumb") relating the optimal choice of duration of blocking period. Such a choice can be made only based on computation, under the fixed set of the system parameters, with the use of the presented results.

VII. CONCLUSION

We have analyzed the priority retrial multiline queueing model of $MAP_2/M_2/N/N$ type with two types of users suitable for modeling, e.g., cognitive radio networks. The model accounts possible dependence of inter-arrival times of users and not equal service rate of two types of arriving users. HPUs have absolute priority over LPUs. To prevent frequent interruption of LPUs services, their access is restricted via threshold mechanism, which is known in the literature as servers reservation for HPUs, and temporal blocking of LPUs arrival with the purpose to provide maximum effective service to LPUs. The LPUs have an option to retry for service later on in the case of access deny or service interruption.

Under the fixed value of the reservation threshold and distribution of duration of an admission blocking period, the dynamics of the system is described by a level dependent MC.

This *MC*'s generator is derived. This chain's sufficient conditions for ergodicity and non-ergodicity are stated. The expressions for primary system performance indicators via the vectors defining the invariant distribution of the considered multi-dimensional *MC* are obtained.

Numerical results show the positive effect of using the blocking period for improvement of service quality of *LPU* even in absence of servers reservation. Essential effect of correlation of inter-arrival times on the optimal choice of duration of blocking period and preference of long and short blocking periods is demonstrated.

VIII. POSSIBLE DIRECTIONS OF RESEARCH

As possible directions for generalization of the considered model we suggest the following ones:

- consideration of possibility of the work in underlay mode when the *LPU*s have some own share of a bandwidth and more flexible strategy of *LPU*s expelling from the service similar to the one considered in [32];
- existence of more than two priority classes, e.g., separation to different classes of new and handover *HPU*s and *LPU*, distinguishing cognitive users obtaining opportunistic access without any payment to service provider and users that lease the channels and pay for service;
- possibility of temporal buffering of *HPU* arriving at the moments when the bandwidth is exhausted;
- possibility of temporal buffering of *SPU* that arrive at the moments the non-reserved servers are absent or service of which is interrupted;
- account of possibility of using by *LPU*s for service not channels but only sub-channels (like in [12]);
- batch arrival of *HPU*s and *HPU*s;
- phase type distribution of blocking times;
- processor sharing or limiting processor sharing discipline of service of *LPU*s which represent the elastic traffic;
- phase type distribution of service times (using the approach by Ramaswami and Lucantoni [37] and results from [25] and [28]);
- hysteresis strategy of servers reservation (using the results from [14]);
- possible breakdowns or vacations of the servers or disaster occurrence, etc.

ACKNOWLEDGMENT

(Ciro D'Apice, Maria Pia D'Arienzo, Alexander Dudin, and Rosanna Manzo equally contributed to this paper.) The authors highly appreciate the valuable comments by the anonymous reviewers account of which lead to the improvement of their manuscript.

REFERENCES

[1] M. Alipour-Vaezi, A. Aghsami, and F. Jolai, "Prioritizing and queueing the emergency departments' patients using a novel data-driven decision-making methodology, a real case study," *Expert Syst. Appl.*, vol. 195, Jun. 2022, Art. no. 116568.

[2] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Comput. Netw.*, vol. 50, no. 13, pp. 2127–2159, Sep. 2006.

[3] J. R. Artalejo and A. Gomez-Corral, *Retrial Queueing Systems: A Computational Approach*. Berlin, Germany: Springer-Verlag, 2008.

[4] L. Bright and P. G. Taylor, "Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes," *Commun. Statist. Stochastic Models*, vol. 11, no. 3, pp. 497–525, Jan. 1995.

[5] P. Buchholz, P. Kemper, and J. Kriege, "Multi-class Markovian arrival processes and their parameter fitting," *Perform. Eval.*, vol. 67, no. 11, pp. 1092–1106, 2010.

[6] S. R. Chakravarthy, "The batch Markovian arrival process: A review and future work," *Adv. Probab. Theory Stochastic Processes*, vol. 1, pp. 21–49, Jan. 2001.

[7] S. R. Chakravarthy, *Introduction to Matrix-Analytic Methods in Queues 1: Analytical and Simulation Approach Basics*. London, U.K.: Wiley, 2022.

[8] S. R. Chakravarthy, *Introduction to Matrix-Analytic Methods in Queues 2: Analytical and Simulation Approach Queues and Simulation*. London, U.K.: Wiley, 2022.

[9] S. Chen, A. M. Wyglinski, S. Pagadarai, R. Vuyyuru, and O. Altintas, "Feasibility analysis of vehicular dynamic spectrum access via queueing theory model," *IEEE Commun. Mag.*, vol. 49, no. 11, pp. 156–163, Nov. 2011.

[10] C. D'Apice, A. Dudin, S. Dudin, and R. Manzo, "Priority queueing system with many types of requests and restricted processor sharing," *J. Ambient Intell. Humanized Comput.*, vol. 14, Jul. 2022, doi: 10.1007/s12652-022-04233-w.

[11] K. Devarajan and M. Senthilkumar, "On the retrial-queuing model for strategic access and equilibrium-joining strategies of cognitive users in cognitive-radio networks with energy harvesting," *Energies*, vol. 14, no. 8, p. 2088, Apr. 2021.

[12] A. N. Dudin, M. H. Lee, O. S. Dudina, and S. K. Lee, "Analysis of priority retrial queue with many types of customers and servers reservation as a model of cognitive radio system," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 186–199, Jan. 2017.

[13] A. N. Dudin, V. I. Klimenok, and V. M. Vishnevsky, *The Theory of Queuing Systems With Correlated Flows*. Cham, Switzerland: Springer, 2020.

[14] A. Dudin, S. Dudin, R. Manzo, and L. Raritá, "Analysis of multi-server priority queueing system with hysteresis strategy of server reservation and retrials," *Mathematics*, vol. 10, no. 20, p. 3747, Oct. 2022.

[15] S. Dudin and O. Dudina, "Retrial multi-server queuing system with PHF service time distribution as a model of a channel with unreliable transmission of information," *Appl. Math. Model.*, vol. 65, pp. 676–695, Jan. 2019.

[16] S. Dudin, A. Dudin, O. Kostyukova, and O. Dudina, "Effective algorithm for computation of the stationary distribution of multi-dimensional level-dependent Markov chains with upper block-hessenberg structure of the generator," *J. Comput. Appl. Math.*, vol. 366, Mar. 2020, Art. no. 112425.

[17] S. A. Dudin, "The $MAP/M/N$ retrial queueing system with time-phased batch arrivals," *Problems Inf. Transmiss.*, vol. 45, no. 3, pp. 270–281, 2009.

[18] A. Elalouf and G. Wachtel, "Queueing problems in emergency departments: A review of practical approaches and research methodologies," *Oper. Res. Forum*, vol. 3, no. 1, pp. 1–46, 2022.

[19] G. Falin and J. G. Templeton, *Retrial Queues*, vol. 75. Boca Raton, FL, USA: CRC Press, 1997.

[20] A. Graham, *Kronecker Products and Matrix Calculus With Applications*. Chichester, U.K.: Ellis Horwood, 1981.

[21] Q. M. He, J. Xie, and X. Zhao, "Priority queue with customer upgrades," *Nav. Res. Logistics*, vol. 59, no. 5, pp. 362–375, 2012.

[22] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[23] S. Huang, D. Yuan, and A. Ephremides, "Bandwidth partition and allocation for efficient spectrum utilization in cognitive communications," *J. Commun. Netw.*, vol. 21, no. 4, pp. 353–364, Aug. 2019.

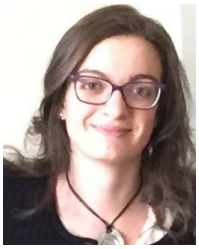
[24] C. S. Kim, S. Dudin, O. Taramin, and J. Baek, "Queueing system $MMAP/PH/N/N+R$ with impatient heterogeneous customers as a model of call center," *Appl. Math. Model.*, vol. 37, no. 3, pp. 958–976, 2013.

[25] C. Kim, A. Dudin, S. Dudin, and O. Dudina, "Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users," *IEEE Access*, vol. 9, pp. 106933–106946, 2021.

[26] V. I. Klimenok and A. N. Dudin, "Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory," *Queueing Syst.*, vol. 54, pp. 245–259, Dec. 2006.

[27] V. Klimenok, C. S. Kim, D. Orlovsky, and A. Dudin, "Lack of invariant property of the Erlang loss model in case of *MAP* input," *Queueing Syst.*, vol. 49, pp. 187–213, Feb. 2005.
[28] V. Klimenok, A. Dudin, and V. Vishnevsky, "Priority multi-server queueing system with heterogeneous customers," *Mathematics*, vol. 8, no. 9, p. 1501, 2020.
[29] Y. Konishi, H. Masuyama, S. Kasara, and Y. Takahashi, "Performance analysis of dynamic spectrum handoff scheme with variable bandwidth demand of secondary users for cognitive radio networks," *Wireless Netw.*, vol. 19, pp. 607–617, Jul. 2013.
[30] B. K. Kumar, R. N. Krishnan, R. Sankar, and R. Rukmani, "Performance analysis of cognitive wireless retrial queueing networks with admission control for secondary users," *Qual. Technol. Quant. Manage.*, vol. 20, pp. 1–38, Dec. 2022, doi: 10.1080/16843703.2022.2136360.
[31] M. S. Kumar, A. Dadlani, K. Kim, and R. O. Afolabi, "Overlay secondary spectrum sharing with independent re-attempts in cognitive radios," in *Proc. IEEE 37th Sarnoff Symp.*, Newark, NJ, USA, Sep. 2016, pp. 178–180.
[32] S. Lee, A. Dudin, O. Dudina, and C. Kim, "Analysis of a priority queueing system with the enhanced fairness of servers scheduling," *J. Ambient Intell. Humanized Comput.*, vol. 14, May 2022, doi: 10.1007/s12652-022-03903-z.
[33] D. M. Lucantoni, "New results on the single server queue with a batch Markovian arrival process," *Commun. Statist. Stochastic Models*, vol. 7, no. 1, pp. 1–46, 1991.
[34] B. T. J. Maharaj and B. S. Awoyemi, *Developments in Cognitive Radio Networks. Future Directions for Beyond 5G*. Cham, Switzerland: Springer, 2022.
[35] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models—An Algorithmic Approach*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1981.
[36] F. Paluncic, A. S. Alfa, B. T. Maharaj, and H. M. Tsimba, "Queueing models for cognitive radio networks: A survey," *IEEE Access*, vol. 6, pp. 50801–50823, 2018.
[37] V. Ramaswami and D. M. Lucantoni, "Algorithms for the multi-server queue with phase type service," *Commun. Statist.-Stochastic Models*, vol. 1, no. 3, pp. 393–417, 1985.
[38] R. Raj and V. Jain, "Optimization of traffic control in $MMAP[2]/PH[2]/S$ priority queueing model with *PH* retrial times and the preemptive repeat policy," *J. Ind. Manage. Optim.*, vol. 19, no. 4, pp. 2333–2353, 2023.
[39] B. Sun, M. H. Lee, S. A. Dudin, and A. N. Dudin, "Analysis of multi-server queueing system with opportunistic occupation and reservation of servers," *Math. Problems Eng.*, vol. 2014, May 2014, Art. no. 178108.
[40] K. Sun, Y. Liu, and K. Li, "Energy harvesting cognitive radio networks with strategic users: A two-class queueing model with retrials," *Comput. Commun.*, vol. 199, pp. 98–112, Feb. 2023.
[41] V. M. Vishnevskii and A. N. Dudin, "Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks," *Autom. Remote Control*, vol. 78, no. 8, pp. 1361–1403, Aug. 2017.
[42] S. Zahed, I. Awan, and A. Cullen, "Analytical modeling for spectrum handoff decision in cognitive radio networks," *Simul. Model. Pract. Theory*, vol. 38, pp. 98–114, Nov. 2013.
[43] H. Zhang and F. Ding, "On the Kronecker products and their applications," *J. Appl. Math.*, vol. 2013, Jun. 2013, Art. no. 296185.
[44] Y. Zhao, Z. Xiang, K. Chen, Z. Ye, and Q. Lu, "Modelling and optimization for cognitive radio networks with preemption backoff mechanism," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 10, pp. 9039–9051, 2022.
[45] Y. Zhao, W. Yue, and Z. Saffer, "Spectrum allocation strategy with a probabilistic preemption scheme in cognitive radio networks: Analysis and optimization," *Ann. Oper. Res.*, vol. 310, no. 2, pp. 621–639, Mar. 2022.
[46] X. Zhu, L. Shen, and T. S. P. Yum, "Analysis of cognitive radio spectrum access with optimal channel reservation," *IEEE Commun. Lett.*, vol. 11, no. 4, pp. 304–306, Apr. 2007.
[47] S. Zhu, J. Wang, and W. W. Li, "Optimal service rate in cognitive radio networks with different queue length information," *IEEE Access*, vol. 6, pp. 51577–51586, 2018.

CIRO D'APICE received the degree (cum laude) and Ph.D. degrees in mathematics from the University of Naples Federico II, in 1991 and 1997, respectively. He is currently a Full Professor of mathematical analysis with Dipartimento di Scienze Aziendali-Management and Innovation Systems (DISA-MIS). His research interests include variational calculus, homogenization, and optimal control; complex networks modeling; conservation laws and applications to traffic networks, telecommunication ones, and supply chains, queueing systems and networks; analytical aspects for the temporal and spatial behaviour of solutions of dynamic problems.



MARIA PIA D'ARIENZO received the degree (cum laude) and Ph.D. degrees in mathematics from the University of Salerno, in 2012 and 2016, respectively. She is currently a contract Professor with the University of Salerno. Her research interests include modeling of cardiovascular systems, and queueing theory and development and analysis of efficient and stable numerical methods for the solution of ordinary differential equations.

ALEXANDER DUDIN received the Ph.D. degree in probability theory and mathematical statistics from Vilnius University, in 1982, and the D.Sc. degree from Tomsk University, in 1992. He is currently the Head of the Laboratory of Applied Probabilistic Analysis, Belarusian State University. He is the author of more than 450 publications including five books and more than 130 articles in top level Journals. He coedited more than ten volumes of the Springer series. In 2013, he received the Scopus Award Belarus for outstanding contribution to the field of Mathematics. He has been the Chairperson of Belarusian Winter Workshops in Queueing Theory which are held since 1985 and has been the Chairperson of IPC of the conference named after A.F. Terpugov since 2014. He was invited for lecturing and research to USA, U.K., Germany, France, The Netherlands, Japan, South Korea, India, Russia, China, Italy, and Sweden.

ROSANNA MANZO received the degree (cum laude) in mathematics and the Ph.D. degree in information engineering from the University of Salerno, in 1996 and 2007, respectively. She is currently an Associate Professor of mathematical analysis with the Department of Information and Electrical Engineering and Applied Mathematics (DIEM), University of Salerno. Her research interests include fluid dynamic models for road networks, telecommunication networks, supply chains and blood flows, optimal control for hybrid systems, variational analysis and optimization in normed spaces, and queueing systems and networks.

Open Access funding provided by 'Università degli Studi di Salerno' within the CRUI CARE Agreement