computation

MDPI

Article

Algorithmic Analysis of Queuing System with Varying Number of Servers, Phase-Type Service Time Distribution, and Changeable Arrival Process Depending on Random Environment

Alexander Dudin *, Olga Dudina and Sergei Dudin

Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus; dudina@bsu.by (O.D.); dudins@bsu.by (S.D.)
* Correspondence: dudin@bsu.by

Abstract

An $MAP/PH/N$-type queuing system functioning within a finite-state Markovian random environment is studied. The random environment's state impacts the number of available servers, the underlying processes of customer arrivals and service, and the impatience rate of customers. The impact on the state space of the underlying processes of customer arrivals and of the more general, as compared to exponential, service time distribution defines the novelty of the model. The behavior of the system is described by a multidimensional Markov chain that belongs to the classes of the level-independent quasi-birth-and-death processes or asymptotically quasi-Toeplitz Markov chains, depending on whether or not the customers are absolutely patient in all states of the random environment or are impatient in at least one state of the random environment. Using the tools of the corresponding processes or chains, a stationary analysis of the system is implemented. In particular, it is shown that the system is always ergodic if customers are impatient in at least one state of the random environment. Expressions for the computation of the basic performance measures of the system are presented. Examples of their computation for the system with three states of the random environment are presented as 3-D surfaces. The results can be useful for the analysis of a variety of real-world systems with parameters that may randomly change during system operation. In particular, they can be used for optimally matching the number of active servers and the bandwidth used by the transmission channels to the current rate of arrivals, and vice versa.

Keywords: queuing analysis; performance characteristics; random environment; varying number of channels

MSC: 60K25; 60K30; 68M20; 90B22

Academic Editor: Francesco Cauteruccio

Received: 3 June 2025
Revised: 18 June 2025
Accepted: 25 June 2025
Published: 29 June 2025

Citation: Dudin, A.; Dudina, O.; Dudin, S. Algorithmic Analysis of Queuing System with Varying Number of Servers, Phase-Type Service Time Distribution, and Changeable Arrival Process Depending on Random Environment. *Computation* **2025**, *13*, 154. https://doi.org/10.3390/computation13070154

1. Introduction

Queuing theory is an effective mathematical instrument for optimally scheduling restricted resources in a wide range of real-world systems and networks, like contact centers, manufacturing and logistical systems, wired and wireless telecommunication networks, healthcare, emergency help, administrative systems, banks, etc.

A lot of the research that has been carried out on the edges of queuing theory has focused on creating and using methods to analyze and improve systems that have a fixed configuration, such as a set number of channels, buffer capacity, and admission and service rules, as well as a fixed pattern of arrivals, service, retrials, breakdowns, repairs, impatience,

and other processes. However, in many real-world systems, these configurations can vary over time. In particular, the information arrival rate is usually not constant during a day, and the transmission rate in many wireless networks may essentially vary due to various reasons (noise, interference, rain, fog, obstacles in the transmission thread, etc.). An important class of queuing models that account for the fact that the parameters of a system can randomly vary during its operation are so-called queuing systems operating in a random environment (RE). Usually, we refer to an RE as an external random process that influences the parameters of a queuing system. Usually, this process is Markovian (sometimes semi-Markovian) with a finite state space. The jump of an RE between states implies the instantaneous change in certain parameters of the distributions characterizing the system's dynamics.

Motivation of the necessity of investigating queues operating in an *RE*. The simplest examples of queues operating in an RE ($QORE$s) are unreliable queues, where one of the two states of the RE corresponds to the system's working state, and the second state corresponds to the system state when the channels can be broken. Another trivial example of $QORE$s is queues with a Markov arrival process (MAP) and/or a phase-type (PH) distribution of service times. The directing process of arrivals and/or services plays the role of an RE in these examples. It is more common to speak about a $QORE$ when more than one directing process is influenced by the RE and when the jump of the RE process has a synchronous impact on the directing processes, e.g., simultaneously on the directing processes of service and arrivals.

Because the standard assumption that the system's parameters are fixed and not altered during system operation is quite artificial, rarely holds in real-world systems, and is imposed only to make the analysis simpler and obtain nicer analytical or algorithmic results, it is evident that the investigation of $QORE$s is significant for queuing theory, as well as its applications. This explains why the analysis of $QORE$s has quite a long history.

A short overview of the relevant literature. As the earliest work in the field of $QORE$s, the works cited in [1–6] can be mentioned. A brief history of the theory of $QORE$s, a list of references, and examples of the application of $QORE$s in real-world systems are presented, for example, in the publications cited in [7–15].

Most of the findings that have been obtained for $QORE$s relate to the queues having a single channel or infinitely many channels. To simplify the survey of the existing extensive literature, we will mention here only publications related to the analysis of $QORE$s that are closest to the system considered in our paper.

In particular, we mention only papers devoted to *multi-channel* queues having a finite number of channels larger than one.

***QORE*s with a Markov arrival process and phase-type service time distribution.** Here we limit our study to the investigation of systems where the service time has a phase-type (PH) distribution and the arrival flow is characterized by a Markov arrival process (MAP) in a given state of the RE; see, e.g., [6]. This decision is justified as follows.

A stationary Poisson arrival process that has been extensively explored in the literature over the years is already acknowledged to be inferior to the MAP as a model of actual arrival processes. The MAP was introduced by M. Neuts as a versatile arrival process.

The MAP allows fitting not only the mean arrival rate (or inter-arrival time) of a real flow, e.g., information flow in telecommunication networks, but also the values of the variance and higher moments and the possible correlation of inter-arrival times. The MAP is able to capture characteristic properties of arrival processes in contemporary communication networks and contact centers, such as overdispersion and positive correlations between arrival moments. This has made popular the use of the MAP for modeling real-world flows. For further details on the MAP, its attributes, and particular cases, see, e.g., [16–20].

The phase-type PH distribution (see [6,21–23]) is the essential generalization of the exponential distribution for which the analysis of queues with such a distribution in an analytical form is still possible. In contrast to the exponential distribution, which allows fitting only the expectation of a random variable, the PH distribution allows fitting also the higher moments, the variance in particular.

The choice of the PH distribution for characterization of service times in real-world systems is justified by (i) the relative simplicity of the analysis of queues (especially multi-channel queues) with such a distribution, compared to the analogous systems with an arbitrary distribution, and (ii) the suitability of this distribution in order to suit any given distribution.

In the papers [8,9], an $MAP/PH/N$ retrial $QORE$ with a finite source of requests is considered. In ref. [8], the existence of an additional MAP arrival process for negative requests is supposed. In ref. [8,9], the analysis of the systems is implemented via consideration of a $CTMC$ with the finite state space. The stationary probabilities of the $CTMC$ are calculated. Primary performance metrics are acquired, and numerical examples are given.

In paper [11], a $BMAP/PH/N$ type $QORE$ with retrials is investigated. The system lacks a buffer. When a request is unable to catch a free channel upon arrival, it is still able to try again to obtain a service by retrying from orbit. It is suggested that the $RE's$ state also affects the rate of repeated attempts. In contrast to [8,9], the state space of the considered $CTMC$ in [11] is infinite. Deriving a condition for the existence of the MC's stationary distribution becomes necessary as a result. The stationary distribution is computed, and the sufficient condition for ergodicity is proven. The numerical experiment's findings demonstrate that simpler queuing models do not accurately represent the system's key performance metrics. Both the impact of the retry intensity and the potential for traffic smoothing are displayed.

The above-mentioned models assume that all requests belong to one class. In some real-world setups, requests can be heterogeneous, having various requirements for service and different priorities. In ref. [24], a $MMAP/PH_1, PH_2/N$ type $QORE$ with two types of requests is under study. Under the fixed state of an RE, the arrivals occur in an $MMMAP$ (marked Markov arrival process); for details, see, e.g., [25,26]. One of the two types has non-preemptive priority.

In [27], an $MAP/PH/N$ inventory-$QORE$ was considered. Service to a request can be implemented only when there are inventory units that are stored in a storage facility of finite capacity.

***QORE*s with a varying number of channels.** Most of the existing papers on multi-channel $QORE$s do not assume a change in the number of channels. Such a change is supposed to occur in the unreliable multi-channel queuing systems, where channels can be broken and require repair. As a result, a random number of channels can be available at an arbitrary time.

However, usually, a simultaneous random change in the number of available channels and rates of the arrival flow and service process is not suggested. This is a significant flaw in the body of the current literature from the point of view of potential applications. This is because it is evident that an effective system manager has to immediately react to the increase or decrease in the arrival or service rate by appropriately increasing or decreasing the number of channels providing service. Consequently, the account of the potential to alter the number of channels along with changes in the traffic rate and service rate is vital for making correct managerial decisions. Unfortunately, the existing literature devoted to the analysis of $QORE$s with a variable quantity of channels is sparse. Only studies in [28,29] where the arrival and service process characteristics, as well as the number of accessible channels, are changed simultaneously with the state of an RE changing can be mentioned.

In ref. [28], an $MAP/M/N/\infty$ type $QORE$ is under study. The number of channels, arrival, service, and impatience rates for requests are among the system's characteristics that are contingent on an RE's state and fluctuate in value throughout its leaps. The dynamics of the system are determined by a multidimensional asymptotically quasi-Toeplitz MC ($AQTMC$), see [19]. Using the results for $AQTMC$s, a steady-state distribution computation algorithm is given, along with the derivation of an ergodicity criterion. The system's primary performance metrics are computed.

A variation in the number of channels was also assumed in the paper [29]. An $MMAP/M/N$ type $QORE$ with heterogeneous requests and non-preemptive priority is analyzed. In addition to the available number of channels, the thresholds defining the strategy of channel reservation for priority requests depend on an RE as well.

Models considered in [28,29], and several other papers of the authors of this paper assuming the possibility of a random alternation of the channel number are valuable from the perspective of solving a wide range of problems in the examination of different real-world systems. The well-known occurrence of hours of low, intermediate, and peak load in communication and transportation networks explains fluctuations in the intensities of request arrivals and retrials in real-world systems, including contact centers, supermarkets, etc. Alternation of the number of channels and service rates can be triggered by the change in the bandwidth currently available for transmission and the used modulation schemes.

Contributions of the paper. The paper's contributions consist of an exact algorithmic study of a more advanced model compared to [28], where the dependence of a channel number on the state of an RE is also assumed. The two essential advantages of this model are as follows.

A. In the listed models, the exponential distribution of service time, at a given state of an RE, is assumed. Here, we suggest that such time has an essentially more general PH distribution. The reason for the consideration of only the exponential distribution in [28] is the following. The possibility of a random change in the channel number implies the necessity to sometimes break ongoing services or, conversely, to start several new services. When the service time is exponentially distributed, due to the famous memoryless property of this distribution, it can be carried out rather easily. In the case of a PH distribution, essential difficulties arise due to the need to account for, along with transitions of directing processes in busy channels, additional transitions of these directing processes caused by the alternation of the channel number. Because of the potential for huge block sizes in a generator, there are challenges when constructing the MC's generator explicitly, as well as when implementing the computation of the stationary distribution using a computer. Here, we succeeded in overcoming these difficulties.

B. In the existing papers, it is suggested that in every state of the RE, the state space of the directing process of the MAP is permanent. The state of this process is not changed at the jump moment of the RE. Only the intensities of transitions in the directing process become different after this jump. Oppositely, we assume that the directing process may jump at the moment of the jump of the RE. Accordingly, in various RE states, the state spaces of the directing process may have different cardinalities. An account of this possibility is important for applications because traces of real traffic at different states of the RE can be fitted by the MAPs of different dimensions. For example, in some states, when inter-arrival times have a low correlation and a coefficient of variation of about 1, the flow may be fitted by a stationary Poisson process. In other states, when inter-arrival times have higher correlation and variation, the flow may be fitted by an MAP. A more adaptable and sufficient model of the arriving process in the actual world may be constructed by taking into consideration the flexibility of the cardinality of the state space of the directing process.

Possible applications of the model. Some potential applications of *QORE*s are as follows.

1. Various systems for the transmission of heterogeneous information. Certain kinds of data are tolerant of partial losses yet extremely sensitive to jitter and delay. Some other varieties can handle delays, but they become sensitive when a packet is lost. Elastic traffic can tolerate bandwidth fluctuations, but non-elastic traffic needs a steady bandwidth. As a result, several systems for dynamic bandwidth sharing across various information types are constantly evolving. They make the assumption that transmission of the delay-tolerant flows is delayed in the event of congestion to improve circumstances for delay-sensitive flows. A probabilistic analysis of the random processes describing the transmission process of various flows is necessary for the study of such systems. The intricacy of the mathematics frequently makes this analysis unfeasible.

2. Various systems with varying service rates. Such a fluctuation may result from the transmission thread's random deterioration for technical reasons, as well as from external factors including weather, inversions in the atmosphere, interference, noise from enterprises and transportation, users' mobility within a cell using different modulation techniques based on how far they are from a base station, etc.

3. Hybrid Free Space Optics/Radio transmission channels. In recent years, free space optics (*FSO*) technology has been extensively employed; for example, see [30,31]. The foundation of this technology is the transmission of data via the atmosphere using modulated radiation in the visible or infrared spectrum, followed by an optical photodetector device's detection of the data. The two primary benefits of an optical communication link are (i) excellent digital communication quality and bandwidth and (ii) good channel security against stealth and illegal access.

Consequently, laser systems are frequently employed for diverse applications necessitating high data transmission confidentiality, as well as elevated noise immunity and non-interference.

The *FSO* technology has two major disadvantages in addition to its advantages: (i) the channel's accessibility is weather-dependent, and (ii) direct sight between a transmitter and a receiver is necessary. The range of a laser atmospheric communication line can be significantly reduced by unfavorable weather conditions like snow or fog. Therefore, to achieve acceptable dependability values for an *FSO* communication channel, hybrid approaches are required.

The foundation of a hybrid radio–optical technology is the combination of an optical channel with redundant radio channels operating in the centimeter and/or millimeter range of radio waves. The weather has essentially no effect on how the radio channel operates in the centimeter band of radio waves. Fog has no effect on a millimeter-wave wireless channel's performance. At the same time, severe rain significantly lowers the signal/noise ratio, which is a measure of how well a channel performs. The complementary characteristics of broadband radio channels and optical channels have allowed for the development of hybrid carrier systems, which are capable of dependable operation in all weather conditions.

In response to the significant need for high-speed and stable communication channels, several architectures of hybrid systems are employed to address the "last mile" issue. They utilize a high-speed laser channel allocated by a broadband radio channel functioning under the IEEE 802.11n protocol within the centimeter band of radio waves; a radio channel in the millimeter-wave E-band (71–76 GHz, 81–86 GHz); and a parallel-operating radio channel in the millimeter band reserved by an IEEE 802.11n channel. The functionality of such channels in diverse meteorological situations is accurately delineated by *QORE*s.

4. Technologies 5G and 6G New Radio, in which significant variation of transmission rates may occur due to a dynamic blockage of propagation paths by various blockers (vehicles, human bodies, etc.) and the multipath transmission, see, e.g., [32–35].

5. Wireless local area networks serve people living in a certain area, such as a collection of buildings. It is popular among residents to use, along with access to external resources, the relatively inexpensive access to service provider's internal resources, including computer games, music, video, and multimedia. A provider can collect statistics about the rate of generation of user requests that may essentially vary during the day and nighttime (which can be interpreted as the influence of an *RE*). A provider may control the cost of access by raising it during times of high demand and lowering it at times when the servers (and wireless channels for access) are not being used to maintain a high standard of customer care. This should make the traffic smoother and increase the revenue of the service provider.

6. Wireless sensor networks with energy harvesting are designed for different needs, e.g., providing security for some objects. The rates of signal generation from the objects may fluctuate. Signals are generated more frequently during the dark period, while the rate of energy harvesting (if solar energy is used, see, e.g., [36]) and, correspondingly, the signal transmission rate are lower during this period. Naturally, the performance analysis of such a network should account for these fluctuations and can be performed based on *QORE*s.

The rest of the text follows this structure. The mathematical model is explained in Section 2. A *CTMC* describing the system operation is constructed, and its block-structured infinitesimal generator is obtained. Considering that customers are impatient in at least one state of the *RE*, this process is analyzed in Section 3.

Specifically, it is demonstrated here that the considered *MC* is always ergodic. In Section 4, this *MC* is analyzed in the scenario when customers are patient in all states of the *RE*. In particular, the ergodicity constructive criterion is introduced. In Section 5, expressions for the system's performance measures are provided. In Section 6, numerical examples are shown and examined. An analysis is conducted on the three-state system that operates in the *RE*. In state 1 of the *RE*, services are not rendered. It is demonstrated how performance metrics rely on the quantity of servers that are available in states 2 and 3 of the *RE*. Section 7 concludes the paper.

2. Mathematical Model

We consider a multi-channel *QORE* with an infinite buffer. Figure 1 shows the layout of the system that is being studied.

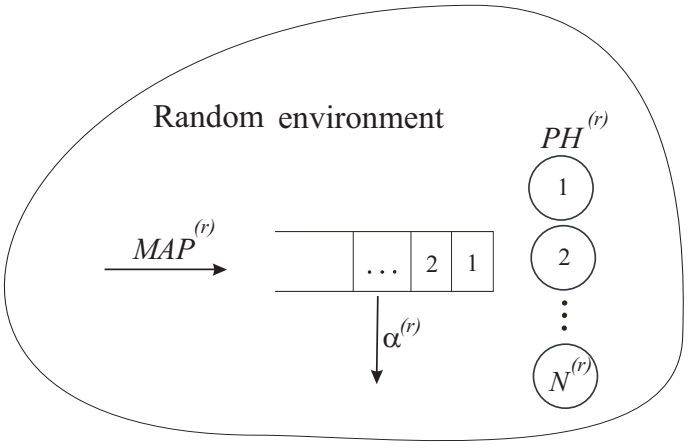

Figure 1. Queuing system under study.

The stochastic process $r_t,\ t \geq 0$, which is an irreducible $CTMC$, is called an RE. The RE's current state determines the system's dynamics. The process r_t is defined by the state space $\{1, 2, \ldots, R\}$ and an infinitesimal generator H, which is a square matrix of size R. $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_R)$ is the invariant probability vector of the RE that represents the unique solution to the system of linear algebraic equations $\boldsymbol{\varphi} H = \mathbf{0},\ \boldsymbol{\varphi}\mathbf{e} = 1$. In this case, $\mathbf{0}$ is a zero-row vector, and $\mathbf{e}$ is a column vector of suitable size made out of 1s.

The system operates as a typical $MAP/PH/N$ type queuing system, i.e., a multi-channel system with an infinite buffer, request arrival defined by an MAP, and a phase-type distribution of service times while the RE is in its fixed state. The following system parameters are instantaneously altered at the RE jump moments: (i) a quantity of available channels; (ii) the state space, state, and transition rates of the directing process of arrivals; (iii) parameters of the PH distribution of service times; and (iv) the impatience rate of requests waiting in the buffer. Here is a more thorough explanation of how the RE affects the system parameters.

For every state r of the RE, $N^{(r)}$ channels are available for service provision. We assume, without losing generality, that the states of the RE are enumerated according to the number of available channels in ascending order, i.e., the numbers $N^{(r)},\ r = \overline{1, R}$ are ordered as $0 \leq N^{(1)} \leq N^{(2)} \leq \cdots \leq N^{(R)}$. Note that the presented analysis is also valid when the number $N^{(r)}$ can be equal to 0. Therefore, the considered model includes, as a particular case, unreliable systems where all channels can be temporarily out of operation.

The following MAP extension defines the arrival of requests. Here, the directing process of the MAP is $\{r_t, \nu_t^{(r_t)}\},\ t \geq 0$, where r_t is the RE's state, and the process $\nu_t^{(r_t)}$ has a state space $\{1, \ldots, W^{(r_t)}\},\ W^{(r_t)} < \infty$, depending on the state r_t of the RE. The process $\nu_t^{(r)}$ walks as an irreducible $CTMC$ in the fixed state r of the process r_t. With a parameter of $\lambda_\nu^{(r)},\ 0 < \lambda_\nu^{(r)} < \infty$, the period of stay of this chain in the state ν is exponentially distributed. At the end of the stay in the state ν, the process $\nu_t^{(r)}$ proceeds to the state ν' without generating a request with the probability $p_0^{(r)}(\nu, \nu'), \nu, \nu' = \overline{1, W^{(r)}}\ \nu \neq \nu', r = \overline{1, R}$. With the probability $p_1^{(r)}(\nu, \nu')$, the process $\nu_t^{(r)}$ moves to the state ν', and a request is generated, $\nu, \nu' = \overline{1, W^{(r)}}$, $r = \overline{1, R}$.

Thus, at a given state r, $r = \overline{1, R}$, of the RE,

- the arrival process is characterized by a set of matrices $D_0^{(r)}$ and $D_1^{(r)}$ having the entries

$$(D_0^{(r)})_{\nu,\nu} = -\lambda_\nu^{(r)}, \nu = \overline{1, W^{(r)}},$$

$$(D_0^{(r)})_{\nu,\nu'} = \lambda_\nu^{(r)} p_0^{(r)}(\nu, \nu'),\ \nu, \nu' = \overline{1, W^{(r)}}, \nu \neq \nu',$$

$$(D_1^{(r)})_{\nu,\nu'} = \lambda_\nu^{(r)}\ p_1^{(r)}(\nu, \nu'),\ \nu, \nu' = \overline{1, W^{(r)}};$$

- the matrix $D^{(r)}(1) = D_0^{(r)} + D_1^{(r)}$ is a generator of the process $\nu_t^{(r)}, t \geq 0$;
- the mean arrival rate $\lambda^{(r)}$ is $\lambda^{(r)} = \boldsymbol{\theta}^{(r)} D_1^{(r)} \mathbf{e}$ where the vector $\boldsymbol{\theta}^{(r)}$ satisfies the system $\boldsymbol{\theta}^{(r)} D^{(r)}(1) = \mathbf{0},\ \boldsymbol{\theta}^{(r)}\mathbf{e} = 1$;
- the inter-arrival times' squared coefficient of variation, $c_{var}^{(r)}$, is provided by $c_{var}^{(r)} = 2\lambda^{(r)}\boldsymbol{\theta}^{(r)}(-D_0^{(r)})^{-1}\mathbf{e} - 1$;
- the coefficient of correlation $c_{cor}^{(r)}$ of the neighboring inter-arrival times is given by $c_{cor}^{(r)} = (\lambda^{(r)}\boldsymbol{\theta}^{(r)}(-D_0^{(r)})^{-1}(D^{(r)}(1) - D_0^{(r)})(-D_0^{(r)})^{-1}\mathbf{e} - 1)/c_{var}^{(r)}$.

Note that a new process defining request arrivals starts at the moment of the RE transition. If it jumps to the state r, then a new state of the process $\nu_t^{(r)}$ is chosen according to the vector $\boldsymbol{\theta}^{(r)},\ r = \overline{1, R}$.

Let us provide the matrices below:

$$\tilde{D}_1 = \mathrm{diag}\{D_1^{(r)},\ r = \overline{1,R}\}, \tilde{D}_0 = \tilde{H} + \mathrm{diag}\{D_0^{(r)},\ r = \overline{1,R}\}$$

where $\mathrm{diag}\{D_k^{(r)},\ r = \overline{1,R}\}$ indicates a diagonal matrix with the diagonal blocks $D_k^{(r)}$, $r = \overline{1,R},\ k = 0,1$; and $\tilde{H}$ is a square matrix of size $\sum_{r=1}^{R} W^{(r)}$ with the block structure. The blocks $(\tilde{H})_{r,r'}$, $r,\ r' = \overline{1,R}$, have the following form:

$$(\tilde{H})_{r,r} = (H)_{r,r} I_{W^{(r)}},\ r = \overline{1,R},$$

$$(\tilde{H})_{r,r'} = (H)_{r,r'} \mathbf{e}_{W^{(r)}} \boldsymbol{\theta}^{(r')},\ r,\ r' = \overline{1,R}, r \neq r'.$$

The rate λ of the requests input flow, averaged over all RE states, is defined as

$$\lambda = \boldsymbol{\theta} \tilde{D}_1 \mathbf{e}$$

where $\boldsymbol{\theta}$ is the solution of the system $\boldsymbol{\theta}\tilde{D}(1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$ where $\tilde{D}(1) = \tilde{D}_0 + \tilde{D}_1$.

It can be checked that the vector $\boldsymbol{\theta}$ has the form $\boldsymbol{\theta} = (\varphi_1 \boldsymbol{\theta}^{(1)}, \varphi_2 \boldsymbol{\theta}^{(2)}, \ldots, \varphi_R \boldsymbol{\theta}^{(R)})$, and an alternative formula for the computation of the averaged intensity λ is

$$\lambda = \sum_{r=1}^{R} \varphi_r \lambda^{(r)}.$$

It should be noted that there has already been extensive discussion of the issue of fitting real-world flows by an MAP in the literature. Popular methods for the MAP construction based on traces of such flows, see, e.g., [20], usually fit several initial moments of inter-arrival times, the coefficient of correlation, and, sometimes, the correlation function of the counting process.

The service time of a request at a given state r of the RE has a PH distribution that is defined by the directing $CTMC$ η_t, $t \geq 0$. This $CTMC$ has the state space $\{1, \ldots, M\}$ and is defined by the pair of the probabilistic row vector $\boldsymbol{\beta}^{(r)}$ and the sub-generator $S^{(r)}$. The vector $\boldsymbol{\beta}^{(r)}$ determines the state (phase) of the chain η_t at the service beginning moment, and the matrix $S^{(r)}$ defines the transition rates of the process η_t inside the set $\{1, \ldots, M\}$ that do not lead to the end of service. The entries of the column vector $\mathbf{S}_0^{(r)} = -S^{(r)}\mathbf{e}$ determine the rates of transitions leading to the end of service. See [6,19] for further details about the PH, its properties, and its applicability to the approximation of any distribution. We suggest that the states of the $CTMC$ η_t do not change at the moments of the RE transitions; only the intensities of subsequent transitions of the $CTMC$ η_t vary.

An incoming request initiates service if some channel is idle at an arrival point. The request is put in the buffer if every channel is busy.

If the RE jumps from r state to r' and $N^{(r')} < N^{(r)}$, and the number of requests in service exceeds $N^{(r')}$, then the excessive requests interrupt service and return to the buffer.

To analyze the system, we have to explicitly specify service of which requests will be interrupted in such a situation at first. Without the loss of generality, we suggest that requests currently processed at phases having the minimal number return to the buffer at first. E.g., if the current phase of service of one request is 1 while the current phase of service of another request is 2, the service of, namely, the first request will be interrupted.

The intuitive explanation of such a choice of requests, which will be deleted from the service at first, is as follows. In such particular cases of the PH distribution as Erlangian or hypo-exponential distribution, the minimal phase of the service implies, on average, the minimal loss of system resources already spent for service of a request if its service is

terminated. In case of an arbitrary PH distribution, any desired preference of the choice of requests to be deleted at first is achieved simply by enumeration of phases of service in the desired order. Therefore, the imposed assumption that, namely, the requests currently processed at phases having the minimal number return to the buffer at first does not restrict the generality of the model.

A request, the service of which was interrupted, enters the service again; its service will start from the beginning. If the quantity of channels that have to interrupt service is less than the quantity of channels having the minimal phase of service, the choice of channels that interrupt service is made with equal probabilities.

If the RE jumps from the state r to the state r' such that $r' > r$, additional $N^{(r')} - N^{(r)}$ channels, $r = \overline{1, R-1}$, activate and the corresponding number of requests from the buffer, if any, begins service.

The requests that are currently in the buffer are impatient. When the state of the RE is $r,\ r = \overline{1, R}$, each request leaves the system (is lost) independently of other requests with the rate $\alpha_r,\ \alpha_r > 0$.

Let us examine the queuing model that has been presented.

3. The Markov Process That Characterizes the Behavior of the System

Let, at the instant $t,\ t \geq 0$,

- i_t, $i_t \geq 0$, be the number of requests in the system,
- r_t, $r_t = \overline{1, R}$, be the state of the RE,
- $\nu_t^{(r_t)}$, $\nu_t^{(r_t)} \in \{1, \ldots, W^{(r_t)}\}$, be the state of the second entry of the directing process of request arrivals,
- $\eta_t^{(m)}$ denotes the number of requests receiving service at the m-th phase, $\eta_t^{(m)} = \overline{0, \min\{i_t, N^{(r_t)}\}}$, $\sum\limits_{m=1}^{M} \eta_t^{(m)} = \min\{i_t, N^{(r_t)}\}$, $\boldsymbol{\eta}_t = (\eta_t^{(1)}, \ldots, \eta_t^{(M)})$.

It is simple to see that $\xi_t = \{i_t, r_t, \nu_t^{(r_t)}, \boldsymbol{\eta}_t\},\ t \geq 0$, is a multidimensional irreducible $CTMC$. We enumerate the states of this $CTMC$ $\xi_t,\ t \geq 0$, in the components $(i_t, r_t, \nu_t^{(r_t)})$ exact lexicographic order and the backward lexicographic order of entries of the vector $\boldsymbol{\eta}_t$. By a macro-state, we mean the set of states with the value (i, r) of the first two entries of the MC.

Now, let us introduce the following notation:

- the number J_n is a cardinality of the state space of the process $\boldsymbol{\eta}_t$ when n requests simultaneously receive service. It is calculated as

$$J_n = \frac{(n+M-1)!}{n!(M-1)!},\ n = \overline{1, N^{(R)}},\ J_0 = 1;$$

- the matrix X_n of size $J_n \times J_{n-1}$ defines the probabilities of the process $\boldsymbol{\eta}_t$ transition at the instant when the service of one of n requests that is at the minimal phase of service is terminated, $n = \overline{1, N^{(R)}}$.

  The matrices X_n have the following form:

$$X_1 = \mathbf{e}_M,\quad X_n = \left(\begin{array}{c} I_{T_{n-1}^{(M)}} \\ \hline O \quad I_{T_{n-1}^{(M-1)}} \\ \hline O \quad I_{T_{n-1}^{(M-2)}} \\ \hline \vdots \\ \hline O \quad I_{T_{n-1}^{(2)}} \\ \hline \mathbf{0} \quad 1 \end{array}\right),\ n = \overline{2, N^{(R)}},$$

where

$$T_n^{(m)} = \frac{(n+m-1)!}{n!(m-1)!},\ n = \overline{1, N^{(R)}},\ m = \overline{1, M};$$

- The symbol δ_A is an indicator of the event A. If the event A is true, it equals 1, and if not, it equals 0;
- The Kronecker product and sum of matrices (for definition and properties, see [37]) are denoted by the symbols $\otimes$ and $\oplus$, respectively;
- If the RE is in the state r and n requests receive service, then
  the matrix $L_n^{(r)} = L_n(\mathbf{S}_0^{(r)})$ defines the intensities of the process $\boldsymbol{\eta}_t$ transitions during the epoch when one of the busy channels finishes service, $n = \overline{1, N^{(r)}}$;
  the matrix $A_n^{(r)} = A_n(S^{(r)})$ defines the intensities of the process $\boldsymbol{\eta}_t$ transitions during the epoch when a service phase is changed in one of the busy channels, $n = \overline{1, N^{(r)}}$;
  the matrix $P_n^{(r)} = P_n(\boldsymbol{\beta}^{(r)})$ defines the probabilities of the process $\boldsymbol{\eta}_t$ transitions during the epoch of new service beginning, $n = \overline{0, N^{(r)} - 1}$;
  the matrix

$$\Delta_n^{(r)} = -\mathrm{diag}\{A_n^{(r)}\mathbf{e} + L_n^{(r)}\mathbf{e}\},\ n = \overline{1, N^{(r)}}.$$

  is a diagonal matrix whose entries define the intensities of the process $\boldsymbol{\eta}_t$ leaving its states.

The matrices $P_n(\boldsymbol{\beta}^{(r)})$, $L_n(\mathbf{S}_0^{(r)})$, and $A_n(S^{(r)})$ can be founded based on algorithms elaborated on in [38,39].

Let Q be the $CTMC$ ξ_t generator. The MC ξ_t transition rates from the macro-state (i, r) to the macro-state (j, r'), $r, r = \overline{1, R}$, are defined by the entries of the matrices $(Q_{i,j})_{r,r'}$ in the blocks $Q_{i,j}$. The matrix $Q_{i,i}$ has negative diagonal elements. The intensity of departure from the corresponding state of the MC ξ_t is defined by the modulus of each diagonal entry.

Lemma 1. *The generator Q has a block-tridiagonal form. The following defines the non-zero blocks $Q_{i,j}$, $i, j \geq 0$:*

$$Q_{i,i} = (Q_{i,i})_{r,r'}, r, r' = \overline{1, R},$$

where

$$(Q_{0,0})_{r,r} = D_0^{(r)} + (H)_{r,r} I_{W^{(r)}},\ r = \overline{1, R}, \tag{1}$$

$$(Q_{i,i})_{r,r} = D_0^{(r)} \oplus (\Delta^{(r)}_{\min\{i,N^{(r)}\}} + A^{(r)}_{\min\{i,N^{(r)}\}}) -$$
$$(\delta_{i>N^{(r)}}(i - N^{(r)})\alpha^{(r)} - (H)_{r,r}) I_{W^{(r)} J_{\min\{i,N^{(r)}\}}}, i > 0, r = \overline{1, R}, \tag{2}$$

$$(Q_{i,i})_{r,r'} = (H)_{r,r'}\mathbf{e}_{W^{(r)}}\boldsymbol{\theta}^{(r')} \otimes I_{J_{\min\{i,N^{(r)}\}}}, r' < r, i \leq N^{(r')},\ or\ N^{(r)} = N^{(r')}, r = \overline{2, R}, \tag{3}$$

$$(Q_{i,i})_{r,r'} = (H)_{r,r'}\mathbf{e}_{W^{(r)}}\boldsymbol{\theta}^{(r')} \otimes X_i X_{i-1} \ldots X_{N^{(r')}+1}, r' < r,\ N^{(r')} < i < N^{(r)},\ r = \overline{2, R}, \tag{4}$$

$$(Q_{i,i})_{r,r'} = (H)_{r,r'}\mathbf{e}_{W^{(r)}}\boldsymbol{\theta}^{(r')} \otimes X_{N^{(r)}} X_{N^{(r)}-1} \times \ldots X_{N^{(r')}+1},\ r' < r,\ i \geq N^{(r)},\ N^{(r)} \neq N^{(r')}, \tag{5}$$

$$(Q_{i,i})_{r,r'} = (H)_{r,r'}\mathbf{e}_{W^{(r)}}\boldsymbol{\theta}^{(r')} \otimes I_{J_{\min\{i,N^{(r)}\}}}, r' > r, i \leq N^{(r)},\ or\ N^{(r)} = N^{(r')},\ r = \overline{1, R-1}, \tag{6}$$

$$(Q_{i,i})_{r,r'} = (H)_{r,r'}\mathbf{e}_{W^{(r)}}\boldsymbol{\theta}^{(r')} \otimes P^{(r')}_{N^{(r)}} P^{(r')}_{N^{(r)}+1} \times \ldots P^{(r')}_{i-1},\ r' > r, N^{(r')} > i > N^{(r)},\ r = \overline{1, R-1}, \tag{7}$$

$$(Q_{i,i})_{r,r'} = (H)_{r,r'}\mathbf{e}_{W^{(r)}}\boldsymbol{\theta}^{(r')} \otimes P^{(r')}_{N^{(r)}} P^{(r')}_{N^{(r)}+1} \ldots P^{(r')}_{N^{(r')}-1},$$
$$r' > r, i \geq N^{(r')},\ N^{(r)} \neq N^{(r')},\ r = \overline{1, R-1}. \tag{8}$$

$$Q_{i,i+1} = \mathrm{diag}\{C_i^{(r)},\ r = \overline{1,R}\},\ i \geq 0, \tag{9}$$

$$C_i^{(r)} = \begin{cases} D_1^{(r)} \otimes P_i^{(r)}, & i = \overline{1, N^{(r)} - 1}, \\ D_1^{(r)} \otimes I_{J_N^{(r)}}, & i \geq N^{(r)}. \end{cases} \tag{10}$$

$$Q_{i,i-1} = \mathrm{diag}\{B_i^{(r)},\ r = \overline{1,R}\}, i \geq 1. \tag{11}$$

$$B_i^{(r)} = I_{W^{(r)}} \otimes L_i^{(r)},\ 1 \leq i \leq N^{(r)},$$
$$B_i^{(r)} = (i - N^{(r)})\alpha^{(r)} I_{W^{(r)} J_{N^{(r)}}} + I_{W^{(r)}} \otimes L_{N^{(r)}}^{(r)} P_{N^{(r)}-1}^{(r)}, i > N^{(r)},\ if\ \ N^{(r)} \neq 0. \tag{12}$$

If $N^{(r)} = 0$, the second summand in (12) is omitted.

Proof of Lemma 1. This is evident if the above-explained probabilistic interpretation of the matrices participating in Formulas (1)–(12) is taken into account. □

The study of the stationary distribution of this chain may be implemented after the explicit form of its generator is known.

Such a study for any *MC* usually includes two phases: verification of whether or not the stationary distribution exists and analytical or algorithmic computation of this distribution, conditional if it exists in the given values of the transition probabilities. In the study of the *MC* ξt, implementation of these phases is different depending on values of the impatience rates $\alpha_r,\ r = \overline{1,R}$.

The case when at least one of the impatience rates $\alpha_r,\ r = \overline{1,R}$, is strictly positive is essentially more difficult to study. However, due to the extensive experience of the authors in the analysis of the asymptotically quasi-Toeplitz Markov chains (*AQTMC*s), see [19], this analysis is implemented easily and is presented in the short Section 4.

The case when impatience rates $\alpha_r,\ r = \overline{1,R}$, are equal to zero, i.e., requests never depart from the system due to the impatience, formally is much easier to study because *MC* ξ_t belongs to the class of level-independent quasi-birth-and-death processes (*QBD*s) exhaustively analyzed by M. Neuts (see [6]). However, the ergodicity condition for *QBD*s is given in the form of an inequality up to the value of some vector that is a solution of the system of linear algebraic equations. Thus, the ergodicity condition can be verified only by means of the solution of this system on a computer. In other words, the ergodicity condition does not have an analytical form.

At the same time, due to the specific features of the generator of the *MC*, sometimes it is possible to succeed in solving the above-mentioned system of linear algebraic equations analytically and obtain a nice scalar form of the ergodicity condition. Such a condition is obtained for the considered *MC* ξ_t in the case of the exponential distribution of service time in Section 5.

4. Ergodicity Condition and the Stationary Distribution Calculation: Case of Impatient Requests

As it is outlined above, at first, we address the situation in which there is a strictly positive impatience rate $\alpha_r,\ r = \overline{1,R}$, at least once. In this situation, the blocks $Q_{i,i}$ and $Q_{i,i-1}$ depend on i, and therefore, the *MC* ξ_t is a level-dependent *QBD* process and cannot be analyzed using the results by M. Neuts.

Fortunately, it may be demonstrated that the following limits exist:

$$Y_k = \lim_{i\to\infty} U_i^{-1} Q_{i,i+k-1} + \delta_{k=1} I,\ k = 0, 1, 2,$$

where $U_i = -I \circ Q_{i,i},\ i \geq 0$; and $\circ$ is the Hadamard product of the matrices symbol, see [40]. It stems from this fact that the $MC\ \xi_t$ is a member of the class of the $AQTMC$s.

The sufficient condition for ergodicity of the $AQTMC$ has different forms depending on whether or not the matrix $Y_0 + Y_1 + Y_2$ is irreducible.

It is easy to verify that for the MC under study, the matrix $Y_0 + Y_1 + Y_2$ is reducible. In this case, the ergodicity condition is defined as follows.

Let $\mathcal{R}$ represent the collection of integers corresponding to the states of the RE, for which the impatience rate is strictly positive. Then, the number of irreducible stochastic blocks in the canonical normal form of this matrix is equal to the cardinality of the set $\mathcal{R}$. Denote these blocks as $(Y_0 + Y_1 + Y_2)^{(r)},\ r \in \mathcal{R}$. For the matrices Y_0, Y_1, and Y_2, let the corresponding blocks be $Y_0^{(r)}, Y_1^{(r)}$, and $Y_2^{(r)}$. The sufficient condition for the ergodicity of the $AQTMC$ provided in [19] is the realization of disparities

$$Y_0^{(r)} \mathbf{y}^{(r)} \mathbf{e} > Y_2^{(r)} \mathbf{y}^{(r)} \mathbf{e}, r \in \mathcal{R},$$

where the vector $\mathbf{y}^{(r)}$ satisfies the system

$$\mathbf{y}^{(r)}(Y_0^{(r)} + Y_1^{(r)} + Y_2^{(r)}) = \mathbf{y}^{(r)},\ \mathbf{y}^{(r)}\mathbf{e} = 1,\ r \in \mathcal{R}.$$

It is easy to verify that for $r \in \mathcal{R}$, $Y_0^{(r)} = I, Y_1^{(r)} = O$, and $Y_2^{(r)} = O$. Thus, the required inequalities are fulfilled. Therefore, we proved the following statement.

Theorem 1. *If $\alpha^{(r)} > 0$ for at least one value of r, $r = \overline{1, R}$, then the MC ξ_t is ergodic for any values of the system's parameters.*

If the $MC\ \xi_t$ is ergodic, then for $i \geq 0$, the stationary probabilities of the MC's states exist:

$$\pi(i, r, \nu, \eta^{(1)}, \ldots, \eta^{(M)}) =$$

$$\lim_{t\to\infty} P\{i_t = i, r_t = r, \nu_t^{(r)} = \nu^{(r)}, \eta_t^{(1)} = \eta^{(1)}, \ldots, \eta^{(M)}\}, \tag{13}$$

$$r = \overline{1, R},\ \nu^{(r)} = \overline{1, W^{(r)}},\ \eta^{(m)} = \overline{0, \min\{i, N^{(r)}\}},\ m = \overline{1, M}.$$

By enumerating these probabilities in the backward lexicographic order of the entries $(\eta^{(1)}, \ldots, \eta^{(M)})$, we form the row vectors $\boldsymbol{\pi}(i, r, \nu^{(r)}),\ i \geq 0,\ r = \overline{1, R},\ \nu^{(r)} = \overline{1, W^{(r)}}$. Subsequently, we denote $\boldsymbol{\pi}(i, r) = (\boldsymbol{\pi}(i, r, 1), \boldsymbol{\pi}(i, r, 2), \ldots, \boldsymbol{\pi}(i, r, W^{(r)}))$, and $\boldsymbol{\pi}(i) = (\boldsymbol{\pi}(i, 1), \boldsymbol{\pi}(i, 2), \ldots, \boldsymbol{\pi}(i, R)), i \geq 0$. The set of equilibrium (balance) equations

$$(\boldsymbol{\pi}(0), \boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots)Q = \mathbf{0},\ (\boldsymbol{\pi}(0), \boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots)\mathbf{e} = 1 \tag{14}$$

is satisfied by the vectors $\boldsymbol{\pi}(i),\ i \geq 0$.

In the considered case when some impatience rates $\alpha^{(r)}$ are positive, the MC is a level-dependent QBD process, and the solution of the infinite system of equilibrium equations is a difficult task. Most frequently, such systems in the literature are solved approximately via truncation of the system.

Instead of using this rough method or its very popular clever modification in the existing literature by Neuts and Rao from [41], we apply for computing the vectors $\boldsymbol{\pi}(i)$, $i \geq 0$, the numerically stable algorithms for $AQTMC$s presented in [19,42].

5. Ergodicity Condition and the Stationary Distribution's Calculation: Case of the Patient Requests

Now, let us examine the scenario in which all impatience rates $\alpha^{(r)}$ are equal to zero. In this scenario, all the blocks $Q_{i,i}$, $Q_{i,i-1}$, and $Q_{i,i+1}$ do not depend on i, and therefore, the *MC* ξ_t is a level-independent *QBD* process. The probability vectors $\boldsymbol{\pi}(i)$ for $i \geq N^{(R)}$ are recursively computed via multiplication of the previous vector by a constant matrix. The vectors $\boldsymbol{\pi}(i)$, $i = \overline{0, N^{(R)}}$, are computed as the solution of a finite system of equations. Therefore, the theoretical interest in this case is only the criterion of ergodicity for the considered *MC*.

Let us introduce the matrices $\mathcal{Q} = Q^0 + Q^- + Q^+$ as follows:

The matrix Q^0 consists of the blocks $(Q^0)_{r,r'}$ where $r, r' = \overline{1, R}$, that are defined as

$$(Q^0)_{r,r} = D_0^{(r)} \oplus (\Delta^{(r)}_{N^{(r)}} + A^{(r)}_{N^{(r)}}) + (H)_{r,r} I_{W^{(r)} J_{N^{(r)}}},\ r = \overline{1, R},$$

$$(Q^0)_{r,r'} = (H)_{r,r'} \mathbf{e}_{W^{(r)}} \boldsymbol{\theta}^{(r')} \otimes X^{(r,r')},\ r' < r,$$

$$(Q^0)_{r,r'} = (H)_{r,r'} \mathbf{e}_{W^{(r)}} \boldsymbol{\theta}^{(r')} \otimes P^{(r,r')},\ r' > r,$$

where

$$X^{(r,r')} = X_{N^{(r)}} X_{N^{(r)}-1} \cdots X_{N^{(r')}+1},\ r' < r,$$

$$P^{(r,r')} = P^{(r')}_{N^{(r)}} P^{(r')}_{N^{(r)}+1} \cdots P^{(r')}_{N^{(r')}-1},\ r' > r.$$

The matrices $X^{(r,r')}$ and $P^{(r,r')}$ are equal to an identity matrix I of size $J_{N^{(r)}}$ if $N^{(r)} = N^{(r')}$.

$$Q^- = \mathrm{diag}\{I_{W^{(r)}} \otimes L^{(r)}_{N^{(r)}} P^{(r)}_{N^{(r)}-1},\ r = \overline{1, R}\}.$$

If some values of $N^{(r)}$ are equal to zero, the corresponding diagonal block in the matrix Q^- is equal to O.

$$Q^+ = \mathrm{diag}\{D_1^{(r)} \otimes I_{J_{N^{(r)}}},\ r = \overline{1, R}\}.$$

Theorem 2. *If all impatience rates $\alpha^{(r)}$ are equal to zero, the MC ξ_t has a stationary distribution of the states if*

$$\mathbf{u} Q^- \mathbf{e} > \mathbf{u} Q^+ \mathbf{e} \tag{15}$$

where the vector $\mathbf{u}$ *can be found as a single solution to the system*

$$\mathbf{u}\mathcal{Q} = \mathbf{0},\ \mathbf{u}\mathbf{e} = 1. \tag{16}$$

Proof of Theorem 2. This immediately stems from the theory of the level-independent *QBD*s, see [6]. □

Under any fixed set of system parameters, system (16) of the finite size $\sum_{r=1}^{R} W^{(r)} J_{N^{(r)}}$ can be solved on a computer. By substituting this solution into (15), it is simple to verify whether or not the *MC* is ergodic.

Remark 1. *It would be tempting to try to find the solution of (16) analytically. The matrix $\mathcal{Q}$ can be shown as*

$$\mathcal{Q} = \mathrm{diag}\{(D_0^{(r)} + D_1^{(r)}) \otimes I_{J_{N^{(r)}}},\ r = \overline{1, R}\} +$$

$$\mathrm{diag}\{I_{W^{(r)}} \otimes (\Delta^{(r)}_{N^{(r)}} + A^{(r)}_{N^{(r)}} + L^{(r)}_{N^{(r)}} P^{(r)}_{N^{(r)}-1}),\ r = \overline{1, R}\} +$$

$$\begin{pmatrix} \bar{H}_1 & H_{1,2}^{+} & \dots & H_{1,R}^{+} \\ H_{2,1}^{-} & \bar{H}_2 & \dots & H_{2,R}^{+} \\ \vdots & \vdots & \ddots & \vdots \\ H_{R,1}^{-} & H_{R,2}^{-} & \dots & \bar{H}_R \end{pmatrix}, \tag{17}$$

where

$$\bar{H}_r = (H)_{r,r} I_{W^{(r)} J_{N^{(r)}}},$$

$$H_{r,r'}^{+} = (H)_{r,r'} \mathbf{e}_{W^{(r)}} \boldsymbol{\theta}^{(r')} \otimes P^{(r,r')},$$

$$H_{r,r'}^{-} = (H)_{r,r'} \mathbf{e}_{W^{(r)}} \boldsymbol{\theta}^{(r')} \otimes X^{(r,r')}.$$

Looking at this representation of the matrix $\mathcal{Q}$, we may guess that the solution of system (15) has the form

$$\mathbf{u} = (\varphi_1 \boldsymbol{\theta}^{(1)} \otimes \boldsymbol{\psi}^{(1)}, \varphi_2 \boldsymbol{\theta}^{(2)} \otimes \boldsymbol{\psi}^{(2)}, \dots, \varphi_R \boldsymbol{\theta}^{(R)} \otimes \boldsymbol{\psi}^{(R)}) \tag{18}$$

where the vectors $\boldsymbol{\psi}^{(r)}$, $r = \overline{1,R}$, satisfy equations

$$\boldsymbol{\psi}^{(r)} (\Delta_{N^{(r)}}^{(r)} + A_{N^{(r)}}^{(r)} + L_{N^{(r)}}^{(r)} P_{N^{(r)}-1}^{(r)}) = \mathbf{0}, \quad \boldsymbol{\psi}^{(r)} \mathbf{e} = 1. \tag{19}$$

Let us substitute this vector into system (16) with the matrix $\mathcal{Q}$ given by Formula (17). Utilizing the mixed product rule and the relation $\boldsymbol{\theta}^{(r)} (D_0^{(r)} + D_1^{(r)}) = \mathbf{0}$, we obtain a zero vector from multiplication of the first summand in (17) by the vector $\mathbf{u}$. Analogously, taking into account Formula (19) for the vector $\boldsymbol{\psi}^{(r)}$, we obtain a zero vector from multiplication of the second summand in (17) by the vector $\mathbf{u}$. Now, we need to obtain a zero vector from multiplication of the third summand in (17) by the vector $\mathbf{u}$ with account of the equation $\boldsymbol{\varphi} H = \mathbf{0}$. However, we succeed in doing this only if the following relations hold:

$$\boldsymbol{\psi}^{(r)} X^{(r,r')} = \boldsymbol{\psi}^{(r')}, \; r' < r, \; \boldsymbol{\psi}^{(r)} P^{(r,r')} = \boldsymbol{\psi}^{(r')}, \; r' > r. \tag{20}$$

In general, these relations do not hold, and therefore, our guess that the vector $\mathbf{u}$ is defined by Formula (18) is wrong. We do not have an analytical formula for the vector $\mathbf{u}$ and need to calculate it numerically as a solution to system (16).

However, relations (20) become true if the service time distribution in all states of the *RE* is a particular case of *PH* distribution, namely, the exponential distribution. In this case, expression (18) for the vector $\mathbf{u}$ transforms to

$$\mathbf{u} = (\varphi_1 \boldsymbol{\theta}^{(1)}, \varphi_2 \boldsymbol{\theta}^{(2)}, \dots, \varphi_R \boldsymbol{\theta}^{(R)}).$$

By substituting this vector into inequality (15), we obtain the following statement.

Corollary 1. *In the case of the exponential service time distribution with the parameter $\mu^{(r)}$ in the state r of the RE for all $r = \overline{1,R}$, the MC ξ_t is ergodic if and only if*

$$\sum_{r=1}^{R} \varphi_r N^{(r)} \mu^{(r)} > \lambda = \sum_{r=1}^{R} \varphi_r \lambda^{(r)}. \tag{21}$$

This means that the averaged (over the states of the RE) service rate in the overloaded system is greater than the mean arrival rate.

Therefore, instead of complicated matrix form (15) and (16) of the ergodicity condition, we succeeded in obtaining the ergodicity condition in a nice analytical form that is easily tractable intuitively.

If the chain is ergodic, its stationary distribution is easily computed, as it is outlined at the beginning of this section, in a matrix-geometric form with the probabilities of several boundary states computed numerically.

6. System Performance Metrics

It is possible to implement the computation of the queuing system's essential performance metrics after computing the probability vectors $\boldsymbol{\pi}(i),\ i \geq 0$. The formulas needed to calculate a few of them are presented below.

The mean quantity of requests in the system is

$$L = \sum_{i=1}^{\infty} i\boldsymbol{\pi}(i)\mathbf{e}.$$

The mean quantity of requests in the buffer is

$$N_{buffer} = \sum_{r=1}^{R} \sum_{i=N^{(r)}+1}^{\infty} (i - N^{(r)})\boldsymbol{\pi}(i,r)(I_W \otimes I_{J_{N^{(r)}}})\mathbf{e}.$$

The mean quantity of requests in the buffer, conditional that the RE is in the r-th state, $r = \overline{1,R}$, is

$$N_{buffer}^{(r)} = \frac{1}{\varphi_r} \sum_{i=N^{(r)}+1}^{\infty} (i - N^{(r)})\boldsymbol{\pi}(i,r)(I_W \otimes I_{J_{N^{(r)}}})\mathbf{e},$$

where φ_r is the rth entry of the vector $\boldsymbol{\varphi} = (\varphi_1, \varphi_2 \ldots, \varphi_R)$ of the stationary distribution of the RE.

The mean quantity of busy channels is

$$N_{channel} = \sum_{r=1}^{R} \sum_{i=1}^{\infty} \min\{i, N^{(r)}\}\boldsymbol{\pi}(i,r)(I_W \otimes I_{J_i})\mathbf{e}.$$

The mean quantity of busy channels conditioned that the RE is in the r-th state is

$$N_{channel}^{(r)} = \frac{1}{\varphi_r} \sum_{i=1}^{\infty} \min\{i, N^{(r)}\}\boldsymbol{\pi}(i,r)(I_W \otimes I_{J_i})\mathbf{e},\ r = \overline{1,R}.$$

The rate of the output flow of serviced requests is

$$\lambda_{out} = \sum_{i=1}^{\infty} \sum_{r=1}^{R} \delta_{N^{(r)} \neq 0}\boldsymbol{\pi}(i,r)(I_{W^{(r)}} \otimes L^{(r)}_{\min\{i,N^{(r)}\}})\mathbf{e}.$$

The rate of output of serviced requests in the r-th state of the RE is

$$\lambda_{out}^{(r)} = \sum_{i=1}^{\infty} \delta_{N^{(r)} \neq 0}\boldsymbol{\pi}(i,r)(I_{W^{(r)}} \otimes L^{(r)}_{\min\{i,N^{(r)}\}})\mathbf{e},\ r = \overline{1,R}.$$

The loss probability for a request is

$$P_{loss} = 1 - \frac{\lambda_{out}}{\lambda} = \frac{\sum_{r=1}^{R} \alpha^{(r)} N_{buffer}^{(r)} \varphi_r}{\lambda}.$$

The joint probability that a request loss occurs during the stay of the RE in the state r is

$$P_{loss}^{(r)} = \frac{\alpha^{(r)} N_{buffer}^{(r)} \varphi_r}{\lambda}, \; r = \overline{1, R}.$$

The rate of the service termination of requests due to the decrease in the number of available channels caused by the change in the state of the RE is

$$\mu_{term} = \sum_{r=2}^{R} \sum_{r'=1}^{r-1} \sum_{i=N^{(r')}}^{\infty} (H)_{r,r'} (\min\{i, N^{(r)}\} - N^{(r')}) \boldsymbol{\pi}(i, r) \mathbf{e}.$$

7. Numerical Results

Assume that the queuing system operates in a three-state RE with the generator

$$H = \begin{pmatrix} -0.03 & 0.02 & 0.01 \\ 0.01 & -0.02 & 0.01 \\ 0.007 & 0.003 & -0.01 \end{pmatrix}.$$

The invariant vector of the RE's distribution is $\boldsymbol{\varphi} = (0.2125, 0.2875, 0.5)$.

We suppose the following:

- In state 1 of the RE, the MAP is given by the matrices $D_0^{(1)} = (-0.5)$, $D_1^{(1)} = (0.5)$. This arrival flow is a stationary Poisson arrival flow. It has the arrival rate 0.5, and the coefficients of correlation and variation are $c_{cor}^{(1)} = 0$ and $c_{var}^{(1)} = 1$;
- in state 2 of the RE, the MAP flow is determined by the matrices

$$D_0^{(2)} = \text{diag}\{-2.5, \, -0.5\}, \; D_1^{(2)} = \begin{pmatrix} 2.4 & 0.1 \\ 0.05 & 0.45 \end{pmatrix}.$$

  This MAP has the mean arrival rate 1.16667, and the coefficients of correlation and variation are $c_{cor}^{(2)} = 0.252477$ and $c_{var}^{(2)} = 2.42222$;
- in state 3 of the RE, the MAP flow is determined by the matrices

$$D_0^{(3)} = \text{diag}\{ \, -1.6, \, -0.5, \, -2.1\},$$

$$D_1^{(3)} = \begin{pmatrix} 1.5 & 0.05 & 0.05 \\ 0.01 & 0.29 & 0.2 \\ 0.1 & 0.2 & 1.8 \end{pmatrix}.$$

  This MAP has the mean arrival rate 1.34818, $c_{cor}^{(3)} = 0.117002$, and $c_{var}^{(3)} = 1.91823$. The averaged (over all states of the RE) mean arrival rate is $\lambda = 1.11575$.
- the service time distribution of a request in state 2 of the RE is determined by the sub-generator

$$S^{(2)} = \begin{pmatrix} -0.8 & 0.2 & 0.2 \\ 0.2 & -0.9 & 0.1 \\ 0.2 & 0.1 & -0.7 \end{pmatrix}$$

  and the stochastic vector $\boldsymbol{\beta}^{(2)} = (0.3, 0.5, 0.2)$. The mean service time is $b_1^{(2)} = 2.10849$.
- the service time distribution of a request in state 3 of the RE is determined by the sub-generator

$$S^{(3)} = \begin{pmatrix} -0.5 & 0.1 & 0.1 \\ 0.1 & -0.7 & 0.1 \\ 0.2 & 0.1 & -0.8 \end{pmatrix}$$

and the stochastic vector $\boldsymbol{\beta}^{(3)} = (0.4, 0.4, 0.2)$. The mean service time is $b_1^{(3)} = 2.464$.

- The impatience intensities are fixed as $\alpha^{(1)} = 0.001$, $\alpha^{(2)} = 0.005$, and $\alpha^{(3)} = 0.01$.

In this numerical example, we fix $N^{(1)} = 0$ (all channels are not available, e.g., are broken or under maintenance) and investigate the impact of the numbers $N^{(2)}$ and $N^{(3)}$ of available channels in states 2 and 3 of the *RE*. To carry this out, we alter $N^{(3)}$ throughout the range $[1, 15]$ and $N^{(2)}$ along the interval $[1, \min\{10, N^{(3)}\}]$ with step 1.

Note that the obtained theoretical and numerical results are new, and we do not have experience of their application to a real communication system yet, while such an application is planned to be carried out in the future. The choice of the parameters in this example is made from common sense reasoning, taking into account the computer realization aspects.

The size of the blocks $Q_{i,i}$ of the generator Q of infinite size for $i \geq N^{(R)}$ is equal to $\sum_{r=1}^{R} W^{(r)} \frac{(N^{(r)}+M)!}{N^{(r)}!M!}$. In this example, for $N^{(3)} = 15$ and $N^{(2)} = 10$, the size of the blocks is pretty large (equal to 7916). The increase in the value of R, M, $N^{(r)}, W^{(r)}$, $r = \overline{1, R}$, should cause the larger size of the blocks and the longer computation time. Therefore, we restricted ourselves by the values of the parameters given above.

The dependencies of the mean number of requests in the system L, the rate of the flow of serviced requests λ_{out}, and the intensity μ_{term} of the service termination of requests due to the decrease in the number of channels on the parameters $N^{(2)}$ and $N^{(3)}$ are presented in Figures 2–4.

To draw these figures, it was necessary to compute the stationary distribution of the system and the values of the performance measures at 105 various points $(N^{(2)}, N^{(3)})$. Using Wolfram Mathematica 12.2, the calculations were performed on a Lenovo notebook with an Intel(R) Core(TM) i7-1165G7 2.80 GHz and 16 GB RAM. The total computation time was 2226.4 s (about 37 min).

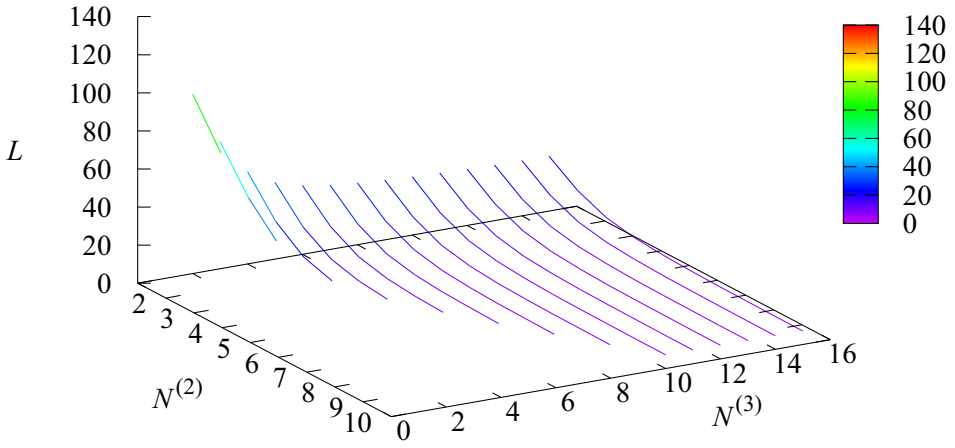

Figure 2. The dependence of L on $N^{(2)}$ and $N^{(3)}$.

The mean quantity L of requests represented in Figure 2 is very large (about 122.5) when only one channel is available in states 2 and 3 of the *RE*. The increase in $N^{(2)}$ and $N^{(3)}$ implies a sharp decrease in L. When $N^{(3)} = 2$ and $N^{(2)} = 1$, the value of L is 94.3. When both $N^{(2)}$ and $N^{(3)}$ are equal to 2, the value of L is 71.17. When $N^{(2)}$ and $N^{(3)}$ admit the maximal values ($N^{(3)} = 15$ and $N^{(2)} = 10$), $L = 7.04$. Figure 2 (and the results of the computations made to plot this figure) helps us easily find the values of $N^{(2)}$ and $N^{(3)}$ for which L will not exceed a number fixed in advance. For example, if $N^{(3)}$ is less than 5, it is impossible to guarantee that L will be less than 10. If $N^{(3)} = 5$, then only $N^{(2)} = 5$ will provide the value of L less than 10. If $N^{(3)} = 6$, the minimal possible value of $N^{(2)}$ is 5. If $N^{(3)} = 15$, the minimal possible value of $N^{(2)}$ is 4.

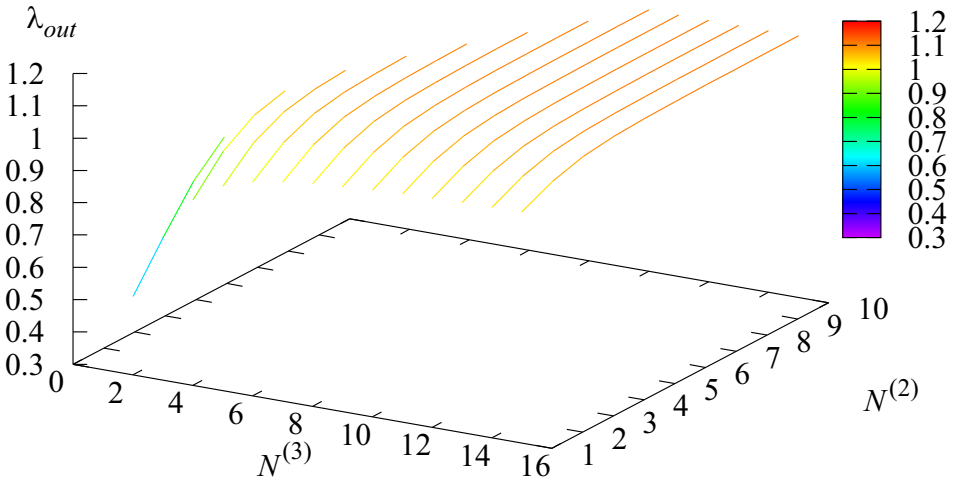

Figure 3. The dependence of λ_{out} on $N^{(2)}$ and $N^{(3)}$.

The rate λ_{out} of the flow of serviced requests is one of the most important economic characteristics of the system.

It is obvious to us instinctively that λ_{out} is small if the numbers of channels are small (because requests renege due to long waiting). The rate λ_{out} quickly increases with the growth of these numbers. Therefore, the shape of the surface in Figure 3 is as anticipated. However, this figure supports the intuitive consideration with concrete numbers. If only one channel is used in both states of the RE, then $\lambda_{out} = 0.3395$. If $N^{(3)} = 15$ and $N^{(2)} = 10$, then λ_{out} is about three times larger: $\lambda_{out} = 1.11$. Recalling that the arrival rate of λ is 1.11575, we see that if $N^{(3)} = 15$ and $N^{(2)} = 10$, only 0.5 percent of requests are lost. If $N^{(3)} = N^{(2)} = 1$, then 69.5 percent of requests are lost.

Using the data corresponding to Figure 3, various problems related to choosing the number of channels can also be solved. For example, if one would like to have $\lambda_{out} > 0.75$, he/she has to have $N^{(3)} > 3$ and any $N^{(2)}$ or $N^{(3)} = 3$ and $N^{(2)} > 1$.

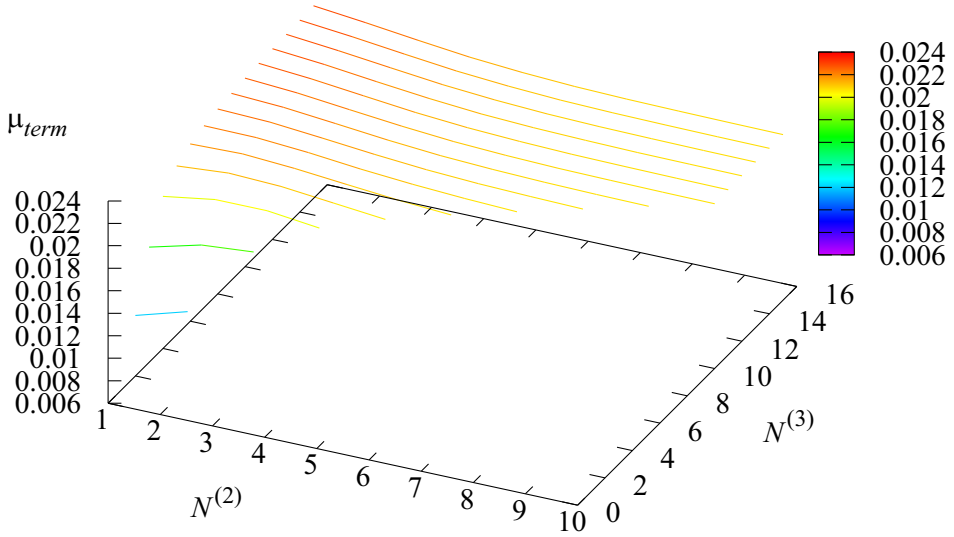

Figure 4. The dependence of μ_{term} on $N^{(2)}$ and $N^{(3)}$.

As anticipated, the rate μ_{term} presented in Figure 4 is minimal when $N^{(2)} = N^{(3)} = 1$. In this case, service terminations happen only when the RE jumps to state 1, and it does not occur during the transition of the RE from state 3 to state 2. The rate μ_{term} increases when $N^{(2)}$ and $N^{(3)}$ and the difference between them increase.

Figures 5–8 illustrate the dependence of the mean number of requests in the buffer N_{buffer} and the mean numbers of requests in the buffer conditional that the RE resides in the r-th state, $N^{(r)}_{buffer}$, $r = 1, 2, 3$.

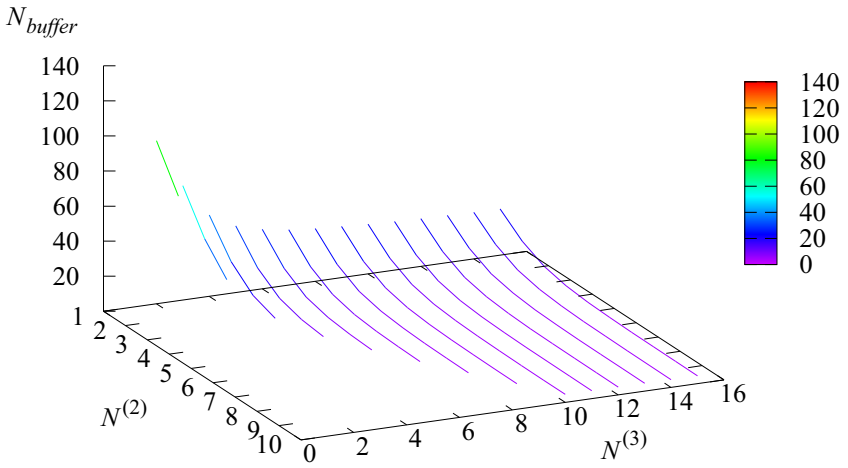

Figure 5. The dependence of N_{buffer} on $N^{(2)}$ and $N^{(3)}$.

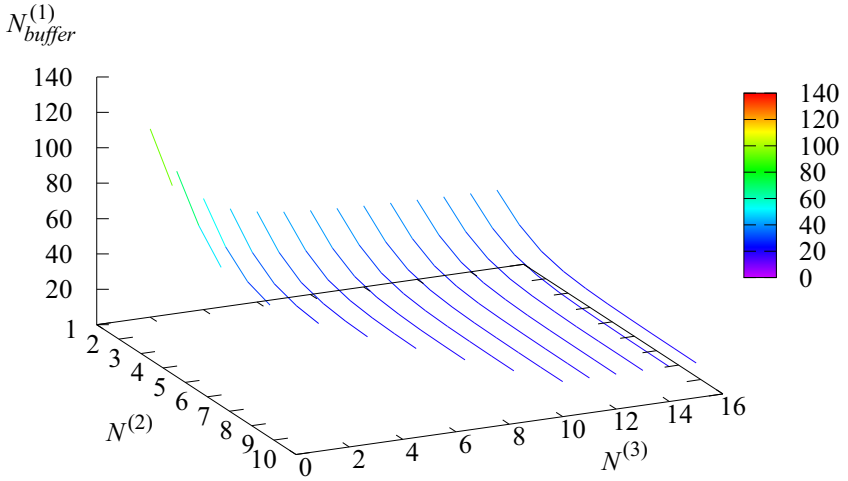

Figure 6. The dependence of $N^{(1)}_{buffer}$ on $N^{(2)}$ and $N^{(3)}$.

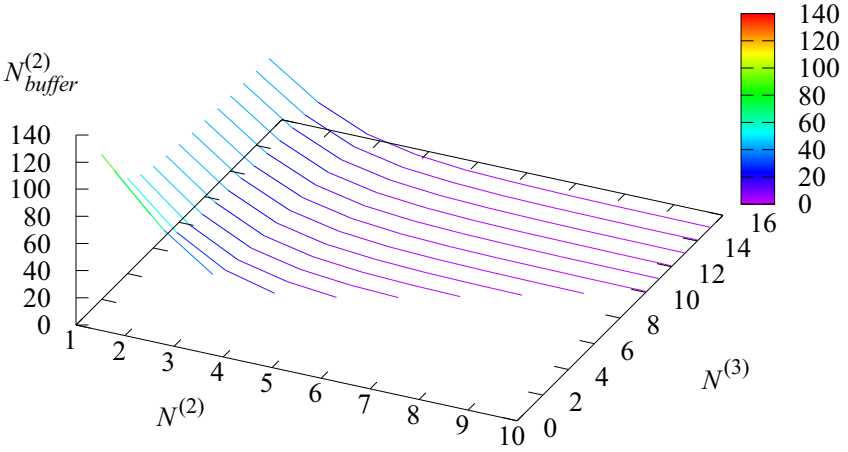

Figure 7. The dependence of $N^{(2)}_{buffer}$ on $N^{(2)}$ and $N^{(3)}$.

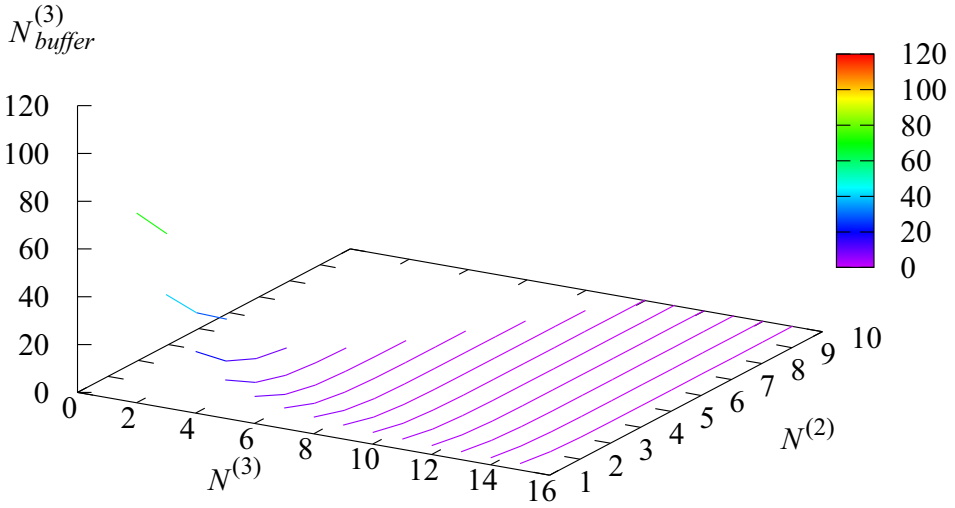

Figure 8. The dependence of $N^{(3)}_{buffer}$ on $N^{(2)}$ and $N^{(3)}$.

As expected, the values of N_{buffer} and $N^{(r)}_{buffer}$ are maximal when $N^{(3)} = N^{(2)} = 1$ and quickly decrease when the values of $N^{(2)}$ and $N^{(3)}$ increase. For $N^{(2)} = N^{(3)} = 1$, the system is permanently overloaded, and the relative difference between the values of N_{buffer} and $N^{(r)}_{buffer}$, $r = \overline{1,3}$, is not very large.

The maximal value is reached when the state of the RE is 1. This is expected because, although in this state the arrival rate is minimal, the service is not provided.

For $N^{(3)} = 15$ and $N^{(2)} = 10$, the relative difference between the values of N_{buffer} and $N^{(r)}_{buffer}$ becomes more essential: $N_{buffer} = 4.45$, $N^{(1)}_{buffer} = 19.65$, $N^{(2)}_{buffer} = 0.73$, $N^{(3)}_{buffer} = 0.12$. Such a large spread of values has to be taken into account via the differentiation of the quality of the request's service promised in service level agreements.

Figures 9–11 demonstrate the dependence of the mean number of busy channels $N_{channel}$ and the mean number of busy channels $N^{(r)}_{channel}$, $r = 2, 3$, under the condition that the RE is in the r-th state, on the parameters $N^{(2)}$ and $N^{(3)}$.

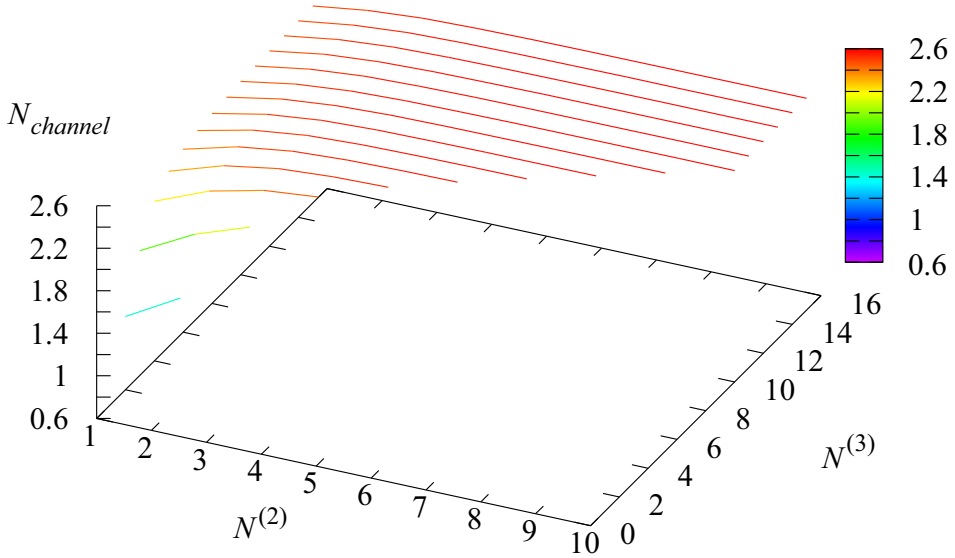

Figure 9. The dependence of $N_{channel}$ on $N^{(2)}$ and $N^{(3)}$.

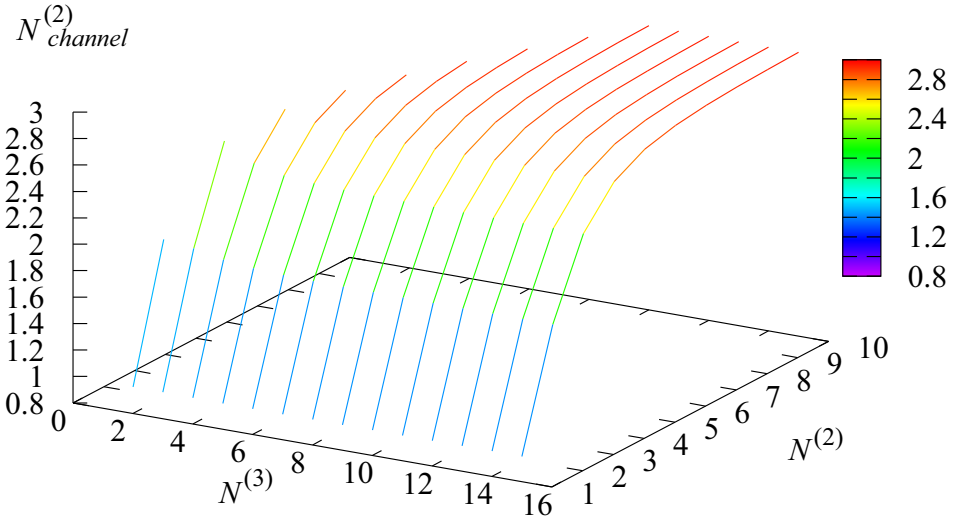

Figure 10. The dependence of $N^{(2)}_{channel}$ on $N^{(2)}$ and $N^{(3)}$.

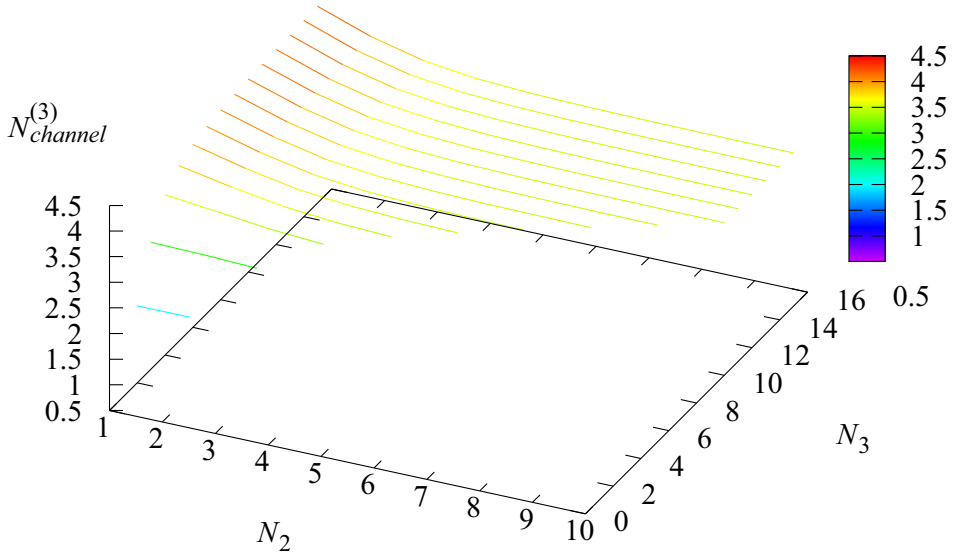

Figure 11. The dependence of $N^{(3)}_{channel}$ on $N^{(2)}$ and $N^{(3)}$.

The range of values of $N_{channel}$ is not very wide. The following information clarifies this: (i) a channel does not work during the stay of the *RE* in state 1; (ii) when the numbers $N^{(2)}$ and $N^{(3)}$ are small, the number of busy channels cannot be large; (iii) when the numbers $N^{(2)}$ and $N^{(3)}$ become larger, the growth of the number of busy channels is not very large because the request's arrival rate is not very high, and available channels may stay idle. The conditional mean number of busy channels in states 2 and 3 of the *RE* varies in a wider range as the account is not used during the times when all channels are unavailable and the number of busy channels is set to zero by default.

Figures 12–15 illustrate the dependencies of the loss probability P_{loss} and the joint loss probabilities $P^{(r)}_{loss}$, $r = 1, 2, 3$, that the *RE* stays in the state r and a request is lost on the parameters $N^{(2)}$ and $N^{(3)}$. The probability P_{loss} is the sum of the joint probabilities $P^{(r)}_{loss}$, $r = 1, 2, 3$.

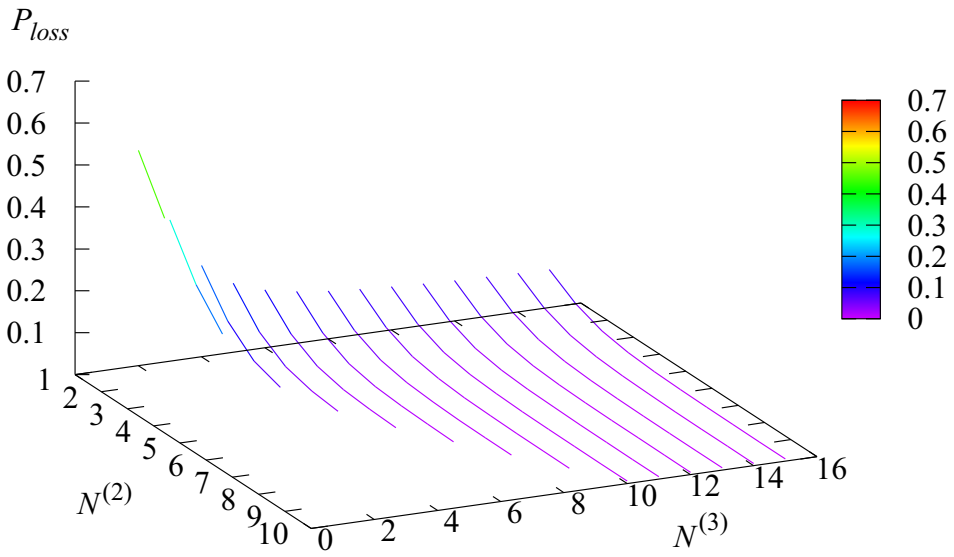

Figure 12. The dependence of P_{loss} on $N^{(2)}$ and $N^{(3)}$.

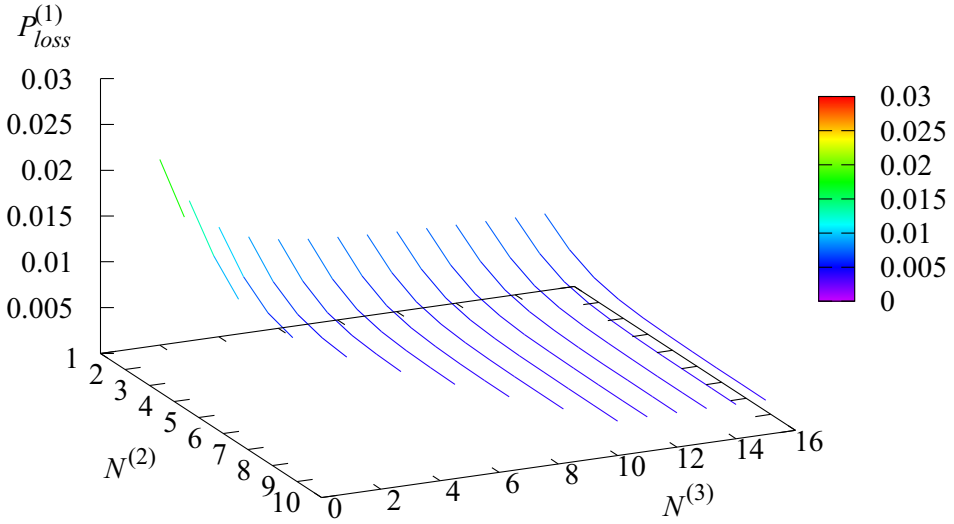

Figure 13. The dependence of $P^{(1)}_{loss}$ on $N^{(2)}$ and $N^{(3)}$.

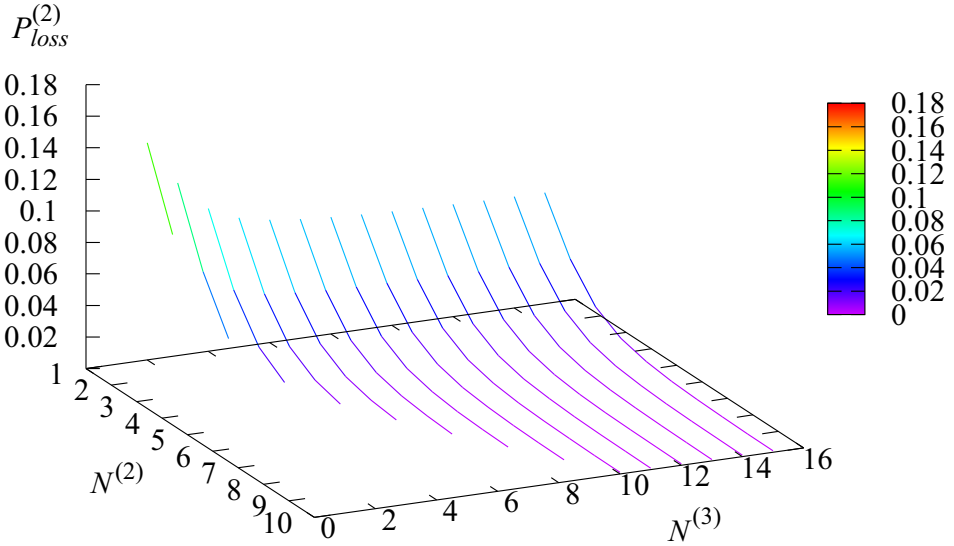

Figure 14. The dependence of $P^{(2)}_{loss}$ on $N^{(2)}$ and $N^{(3)}$.

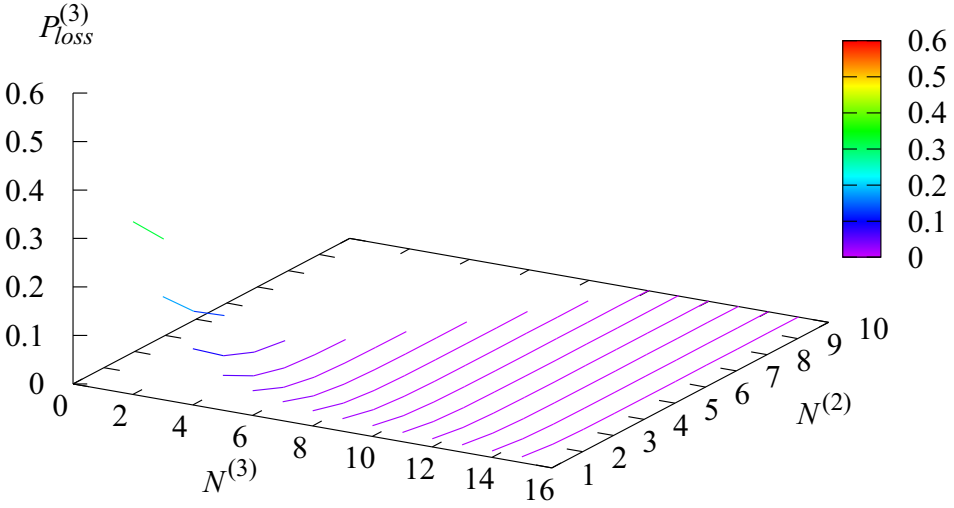

Figure 15. The dependence of $P_{loss}^{(3)}$ on $N^{(2)}$ and $N^{(3)}$.

Above, we stated that the economical quality of the system's functioning is mainly defined by the output rate λ_{out} of the serviced requests. We noted the obvious fact that λ_{out} increases when the numbers of channels $N^{(2)}$ and $N^{(3)}$ increase. In applications of the considered model, it is necessary to take into account that the purchasing or leasing as well as the maintenance of the channels require expenditures. As a result, the following revenue criterion may be used to assess the system's operational quality:

$$E = E(N^{(2)}, N^{(3)}) = a\lambda_{out} - b\lambda P_{loss} - c_2 N^{(2)} - c_3 N^{(3)}$$

where c_l, $l = 2, 3$, are the costs for one channel maintained in states 2 and 3 of the RE, b is the fee paid by the system in the event that a request is lost, and a is the system's income for one request's service.

A pair $(N_*^{(2)}, N_*^{(3)})$ that optimizes the criteria E must be defined. In this numerical example, we set the cost coefficients' values to the following: $a = 3$, $b = 2$, $c_2 = c_3 = 0.1$.

Figure 16 shows the surface that relates the criterion's dependency on $N^{(2)}$ and $N^{(3)}$ values.

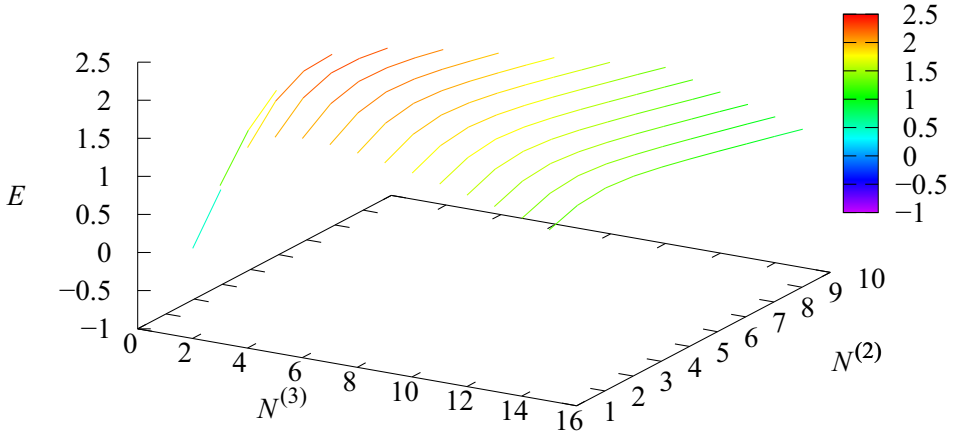

Figure 16. The dependence of E on $N^{(2)}$ and $N^{(3)}$.

For all values of $N^{(3)}$ from 1 to 4, the optimal value of $N^{(2)}$ is equal to $N^{(3)}$. For $N^{(3)}$ from 5 to 15, the optimal value of $N^{(2)}$ is equal to 3. The optimal value of the cost criterion E is 2.28548 when $N^{(2)} = 3$ and $N^{(3)} = 5$. Note that $E(1, 1) = -0.7339$ and $E(15, 10) = 0.818$. Thus, the optimal value of the cost criterion is essentially greater than its values in cases

where only one channel is exploited in states 2 and 3 of the RE and when 10 and 15 channels are used, respectively.

Note also that similar optimization problems can be easily solved in more general settings, e.g., when some restrictions on the values of some performance measures, e.g., the mean number of requests in the buffer or the rate of request service termination, are imposed. Such types of problems arise in many real-world service systems, and the outcomes acquired can be used to achieve managerial objectives.

8. Conclusions

In this paper, an $MAP/PH/N$ type $QORE$ with possible impatient requests is analyzed. An RE may have an influence on the number of available channels, request arrival and service processes, as well as impatience rates. The results can be useful for a performance analysis of this type of queue under various sets of queue parameters. It is important to remember that occasionally, only a part of the system parameters randomly changes. The other parameters can be varied by a system manager aiming to obtain more revenue or a better service experience. For example, under the known pattern of request arrivals, the service provider can dynamically vary the number of active channels. The number of channels can be increased during periods of peak arrival rates and decreased during periods when the arrival rate is low. Conversely, if the random mechanism of changing the number of channels is known (e.g., if the channels are shared with another service provider and his schedule of channels is known), the service provider can manage the procedures of requests, admitting or rejecting them. Requests can be offered (e.g., via the flexible tariffs) to reduce their arrival rate or service time when the number of channels is small. The findings of the analysis can be useful for optimal decision-making in such and similar situations.

The results can be extended, e.g., to systems with different strategies of admission control and request retrials, systems with heterogeneous requests, priorities, and various channel reservation schemes.

Author Contributions: Conceptualization, A.D. and O.D.; methodology, A.D. and O.D.; software, O.D. and S.D.; validation, O.D.; formal analysis, A.D., S.D., and O.D.; investigation, S.D., A.D., and O.D.; writing—original draft preparation, S.D., A.D., and O.D.; writing—review and editing, investigation, A.D. and O.D.; supervision, S.D.; project administration, A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yechialy, U.; Naor, P. Queueing Problems with Heterogeneous Arrivals and Service. *Oper. Res.* **1971**, *19*, 722–734. [CrossRef]
2. Yadin, M.; Syski, R. Randomization of intensities in a Markov chain. *Adv. Appl. Probab.* **1979**, *11*, 397–421. [CrossRef]
3. O'Cinneide C.; Purdue, P. The $M/M/\infty$ queue in a random environment. *J. Appl. Probab.* **1986**, *23*, 175–184.
4. Purdue, P. The $M/M/1$ queue in a Markovian environment. *Oper. Res.* **1974**, *22*, 562–569. [CrossRef]
5. Neuts, M.F. The $M/M/1$ queue with randomly varying arrival and service rates. *Opsearch* **1974**, *15*, 139–157.
6. Neuts, M. *Matrix-Geometric Solutions in Stochastic Models*; The Johns Hopkins University Press: Baltimore, MD, USA, 1981.
7. Cordeiro, J.D.; Kharoufeh, J.P. The unreliable $M/M/1$ retrial queue in a random environment. *Stoch. Model.* **2012**, *28*, 29–48. [CrossRef]
8. Wu, J.; Liu, Z.; Yang, G. Analysis of the finite source $MAP/PH/N$ retrial G-queue operating in a random environment. *Appl. Math. Model.* **2011**, *35*, 1184–1193. [CrossRef]

9. Yang, G.; Yao, L.G.; Ouyang, Z.S. The $MAP/PH/N$ retrial queue in a random environment. *Acta Math. Appl. Sin.* **2013**, *29*, 725–738. [CrossRef]
10. Sztrik, J. Modelling of a multiprocessor system in a randomly changing environment. *Perform. Eval.* **1993**, *17*, 1–11. [CrossRef]
11. Kim, C.S.; Klimenok, V.; Mushko, V.; Dudin, A. The $BMAP/PH/N$ retrial queueing system operating in Markovian random environment. *Comput. Oper. Res.* **2010**, *37*, 1228–1237. [CrossRef]
12. Naumov, V.; Samouylov, K. Resource system with losses in a random environment. *Mathematics* **2021**, *9*, 2685. [CrossRef]
13. Houalef, M.; Bouchentouf, A.A.; Yahiaoui, L. A multi-server queue in a multi-phase random environment with waiting servers and customers' impatience under synchronous working vacation policy. *J. Oper. Res. Soc. China* **2023**, *11*, 459–487. [CrossRef]
14. Otten, S.; Krenzler, R.; Daduna, H.; Kruse, K. Exponential single server queues in an interactive random environment. *Stoch. Syst.* **2023**, *13*, 271–319.
15. Zhu, J.; Hu, L.; Xie, H.; Li, K. A $PH(i)/PH(i,n)/C/C$ Queuing Model in Randomly Changing Environments for Traffic Circulation Systems. *J. Adv. Transp.* **2022**, *2022*, 6533567.
16. Chakravarthy, S.R. The Batch Markovian Arrival Process: A Review and Future Work. In *Advances in Probability Theory and Stochastic Processes*; Krishnamoorthy, A., Raju, N.,Ramaswami, V., Eds.; Notable Publications, Inc.: Neptune, NJ, USA, 2001; pp. 21–49. [CrossRef]
17. Chakravarthy, S.R. *Introduction to Matrix-Analytic Methods in Queues 1: Analytical and Simulation Approach—Basics*; ISTE Ltd.: London, UK; John Wiley and Sons: New York, NJ, USA, 2022.
18. Lucantoni, D. New results on the single server queue with a batch Markovian arrival process. *Commun. Stat. Stoch. Model.* **1991**, *7*, 1–46. [CrossRef]
19. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer Nature: Cham, Switzerland, 2020. [CrossRef]
20. Yera, Y.G.; Lillo, R.E.; Ramírez-Cobo, P. Fitting procedure for the two-state Batch Markov modulated Poisson process. *Eur. J. Oper. Res.* **2019**, *279*, 79–92.
21. O'Cinneide, C.A. Phase-type distributions: Open problems and a few properties. *Stoch. Model.* **1999**, *15*, 731–757.
22. Asmussen, S. *Applied Probability and Queues*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2003. [CrossRef]
23. Horvath, A.; Telek, M. *Phase Type Distributions: Theory and Application*; John Wiley & Sons: Hoboken, NJ, USA, 2024. [CrossRef]
24. Kim, C.; Dudin, A.; Dudin, S.; Dudina, O. Analysis of an $MMAP/PH_1, PH_2/N/\infty$ queueing system operating in a random environment. *Int. J. Appl. Math. Comput. Sci.* **2014**, *24*, 485–501. [CrossRef]
25. He, Q.M. Queues with marked customers. *Adv. Appl. Probab.* **1996**, *28*, 567–587. [CrossRef]
26. Buchholz, P.; Kemper, P.; Kriege, J. Multi-class Markovian arrival processes and their parameter fitting. *Perform. Eval.* **2010**, *67*, 1092–1106.
27. Anbazhagan, N.; Acharya, S.; Vinitha, V.; Amutha, S.; Jeganathan, K.; Seo, C.; Kim, H.I. The $MAP/PH/N/\infty$ Queueing-Inventory System With Demands From a Random Environment. *IEEE Access* **2022**, *10*, 47371–47383. [CrossRef]
28. Kim C.S.; Dudin, A.; Dudin, S.; Dudina, O. Multi-server queueing system $MAP/M/N_R/\infty$ operating in random environment. *Commun. Comput. Inf. Sci.* **2015**, *522*, 306–315. [CrossRef]
29. Dudin, A.; Kim, C.; Dudin, S.; Dudina, O. Priority retrial queueing model operating in random environment with varying number and reservation of servers. *Appl. Math. Comput.* **2015**, *269*, 674–690. [CrossRef]
30. Jeon, H.B.; Kim, S.M.; Moon, H.J.; Kwon, D.H.; Lee, J.W.; Chung, J.M.; Alouini, M.S. Free-space optical communications for 6G wireless networks: Challenges, opportunities, and prototype validation. *IEEE Commun. Mag.* **2023**, *61*, 116–121. [CrossRef]
31. Cui, X.; Park, K.H.; Alouini, M.S. Effect of Random Misalignment in the Capacity of Millimeter-wave OAM. *IEEE Open J. Commun. Soc.* **2024**, *5*, 1141–1154. [CrossRef]
32. Begishev, V.; Sopin, E.; Moltchanov, D.; Kovalchukov, R.; Samuylov, A.; Andreev, S.; Samouylov, K. Joint use of guard capacity and multiconnectivity for improved session continuity in millimeter-wave 5G NR systems. *IEEE Trans. Veh. Technol.* **2021**, *70*, 2657–2672. [CrossRef]
33. Ostrikova, D.; Beschastnyi, V.A.; Moltchanov, D.; Gaidamaka, Y.; Koucheryavy, Y.; Samouylov, K.E. System-level analysis of energy and performance trade-offs in mmWave 5G NR systems. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 7304–7318. [CrossRef]
34. Moltchanov, D.; Sopin, E.; Begishev, V.; Samuylov, A.; Koucheryavy, Y.; Samouylov, K. A tutorial on mathematical modeling of 5G/6G millimeter wave and terahertz cellular systems. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 1072–1116. [CrossRef]
35. Zeng, Y.; Chen, J.; Xu, J.; Wu, D.; Xu, X.; Jin, S.; Zhang, R. A tutorial on environment-aware communications via channel knowledge map for 6G. *IEEE Commun. Surv. Tutor.* **2024**, *26*, 1478–1519.
36. Kristiansen, B.A.; Gravdahl, J.T.; Gros, S.; Johansen, T.A. Energy optimal attitude control and task execution for a solar-powered spacecraft. *IEEE Trans. Control Syst. Technol.* **2024**, *32*, 1212–1225. [CrossRef]

37. Graham, A. *Kronecker Products and Matrix Calculus with Applications*; Courier Dover Publications: Garden City, NY, USA, 2018. [CrossRef]
38. Ramaswami, V. Independent Markov processes in parallel. *Commun. Stat. Stoch. Model.* **1985**, *1*, 419–432.
39. Kim, C.; Dudin, A.; Dudin, S.; Dudina, O. Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users. *IEEE Access* **2021**, *9*, 106933–106946. [CrossRef]
40. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge University Press: Cambridge, UK, 2012. [CrossRef]
41. Neuts, M.F.; Rao, B.M. Numerical investigation of a multiserver retrial model. *Queueing Syst.* **1990**, *7*, 169–189.
42. Dudin, S.; Dudin, A.; Kostyukova, O.; Dudina, O. 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.* **2020**, *366*, 112425.