future internet

MDPI

Article

Performance Modeling of Cloud Systems by an Infinite-Server Queue Operating in Rarely Changing Random Environment

Svetlana Moiseeva [1,†], Evgeny Polin [1,†], Alexander Moiseev [1,†] and Janos Sztrik [2,†,*]

[1] Institute of Applied Mathematics and Computer Science, National Research Tomsk State University, 36 Lenina Ave., 634050 Tomsk, Russia; smoiseeva@mail.ru (S.M.); polin_evgeny@mail.ru (E.P.); moiseev.tsu@gmail.com (A.M.)
[2] Faculty of Informatics, University of Debrecen, Egyetem tér 1, 4032 Debrecen, Hungary
* Correspondence: sztrik.janos@inf.unideb.hu
† These authors contributed equally to this work.

Abstract

This paper considers a heterogeneous queuing system with an unlimited number of servers, where the parameters are determined by a random environment. A distinctive feature is that the parameters of the exponential distribution of the request processing time do not change their values until the end of service. Thus, the devices in the system under consideration are heterogeneous. For the study, a method of asymptotic analysis is proposed under the condition of extremely rare changes in the states of the random environment. We consider the following problem. Cloud node accepts requests of one type that have a similar intensity of arrival and duration of processing. Sometimes an input scheduler switches to accept requests of another type with other intensity and duration of processing. We model the system as an infinite-server queue in a random environment, which influences the arrival intensity and service time of new requests. The random environment is modeled by a Markov chain with a finite number of states. Arrivals are modeled as a Poisson process with intensity dependent on the state of the random environment. Service times are exponentially distributed with rates also dependent on the state of the random environment at the time moment when the request arrived. When the environment changes its state, requests that are already in the system do not change their service times. So, we have requests of different types (serviced with different rates) present in the system at the same time. For the study, we consider a situation where changes of the random environment are made rarely. The method of asymptotic analysis is used for the study. The asymptotic condition of a rarely changing random environment (entries of the generator of the corresponding Markov chain tend to zero) is used. A multi-dimensional joint steady-state probability distribution of the number of requests of different types present in the system is obtained. Several numerical examples illustrate the comparisons of asymptotic results to simulations.

Keywords: infinite-server queue; asymptotic analysis; random environment

Academic Editors: Jerry Chou and Wu-Chun Chung

Received: 6 September 2025
Revised: 29 September 2025
Accepted: 7 October 2025
Published: 8 October 2025

Citation: Moiseeva, S.; Polin, E.; Moiseev, A.; Sztrik, J. Performance Modeling of Cloud Systems by an Infinite-Server Queue Operating in Rarely Changing Random Environment. *Future Internet* **2025**, *17*, 462. https://doi.org/10.3390/fi17100462

1. Introduction

To study real data transmission systems, cellular networks, and cloud computing systems, the mathematical apparatus of queuing theory is used. It includes models of the process of request arrivals, operation of service devices, as well as other features such as changes in parameters of the system over time, and so on. Queueing systems with variable service and arrival rates arise naturally in practice, and therefore, one can find many

classical works. Such systems in queuing theory are called queuing systems operating in a random environment [1–4]. These models more accurately reflect real processes compared to classical systems, taking into account the change in the external random environment over time and the system's response to these changes [5,6].

One of the first works devoted to the study of queuing systems in a random environment is the 1963 publication by M. Eisen and M. Taineter [7], where a single-server system was considered under the assumption that the external environment can be in only two states. Later, in the works of Naor and U. Yechiali [8], the results were generalized to the case of an arbitrary finite number of states of the external environment.

In 1981, M. F. Newts proposed a fundamental method for analyzing queuing systems in a random Markov environment [9]. He reduced the problem of determining the characteristics of the system to solving a matrix equation, which made it possible to study the parameters of systems with dependence on the state of a Markov chain with a finite number of states.

Further development was the expansion of the class of queuing systems in a changing external environment: single- and multi-channel queuing systems [10,11], retrial queues [12], resource queues [13,14], and inventory management systems [15–17].

Various works consider various options for the reaction of requests to the transition of the environment to a new state. For example, in [18], a case is presented in which at the moment of the environment state changed, all requests immediately leave the system. In papers [19,20] an option is considered in which, at the moment of the environment transition to a new state, the requests existing in the system switch to the corresponding new service mode. Infinite-server queues are one of the most common models in the queueing theory and are often used to model systems with instant start of the service, such as satellite communication lines or long communication cables, or to approximate the behavior of multi-server systems. There are many works devoted to the study of infinite-server systems in both Markov [1,19,21–23] and semi-Markov [2,3,20,24,25] random environments.

Let us present several significant results on the methods of studying infinite-server queues operating in a random environment. In [21], the authors consider queue $M/M/\infty$, where only the service rates are regulated by an external process, which is a Markov chain with continuous time and two states. It is shown that the number of requests in the system in the stationary mode can be written as the sum of two independent random variables, where one of these variables is Poisson. Further, B. D'Auria [24,25], using the stochastic decomposition method, showed that using the basic properties of Poisson processes, it is possible to obtain similar results for the number of requests in a similar system in the case of a non-Markov random environment. In the work of Falin G. [2], a model is considered in which a random environment affects both the rate of arrival and the service rate, which are influenced by a semi-Markov environment. To find the average number of requests in the system in the steady state, the method of supplementary variables is used. Namely, a three-dimensional Markov process is considered, the components of which are the number of customers in the system, the state of the environment, and the residual time until the next transition of the environment. This allows us to obtain an expression for the generating function of the number of customers in the system, and it can be used to calculate numerical characteristics in the steady state. In those cases when the study cannot be found by classical methods of the queueing system, asymptotic methods [19,20] are used. As a rule, these papers consider cases where the mode of servicing requests is not changed until they leave the system.

Cloud systems provide different services for users. Interaction with cloud servers is usually hidden from users under a specific user interface, so, actually, cloud servers receive only requests from this interface. There are many approaches for the mathematical

modeling of cloud server operations. One of the most promising and productive approaches is using the queueing theory. In [26–29] and many other papers, different queueing models of cloud systems were considered, and various methods were provided for their study. A good and actual analysis of the queueing theory application in modeling cloud systems was done in [30].

In this paper, we consider a specific case of the models where there are several types of requests incoming to a cloud server and some input device coordinates their arrivals. We propose to model such system in the form of a heterogeneous queueing system with an unlimited number of servers operating in a random environment, allowing one to take into account the dependence of the service time on the type of the request. For the study, the method of asymptotic analysis [31] is used under the condition of extremely rare changes in the random environment. We rely on the method of asymptotic analysis but extend it to the class of heterogeneous systems operating in a random environment. The main difference from the previous work is the different services provided to different types of clients.

The rest of the paper is organized as follows. System and mathematical models are introduced in Section 2. In Section 3 we derive the main system of equations for analysis of the model. In Section 4 expressions for the evaluation of exact moments of the target distribution are obtained and analyzed. In Section 5, we apply the method of the asymptotic analysis to solve the main system of equations under the asymptotic condition of rare changes in the random environment. Finally, we perform numerical analysis of the precision of the solution to find its error and applicability area in Section 6.

2. Mathematical Model of M/M/∞ in Markov Random Environment

Consider a cloud server that performs some operations according to incoming requests. The execution of these requests and further operations take some time and occupy some hardware resources of the server. In this paper, we consider only computational resources—CPU cores that are used for the execution of requests. We consider a situation where the server executes one type of request for a long time. The 'type' can be referred to as the main characteristics of the requests execution: average intensity of their arrivals, average execution time, and so on. Sometimes, an input device (e.g., balancer) switches the server to accept requests of another type. The problem is to find the distribution characteristics of the requests of different types present in the system. Such analysis can give some helpful information for the server owner, for example, how many cores can be required simultaneously for the execution of requests, how many other types of resources (like RAM, disk space, etc.) should be provided for the normal operation of the server if requirements for resources depend on the types of requests, and so on. We propose to use a heterogeneous infinite-server queue operating in a random environment for modeling the described system.

So, let us consider the following heterogeneous queueing system M/M/∞ operating in a Markov random environment (Figure 1).

The process of changing the states of the external environment is a continuous Markov process $s(t)$ with finite number of states $s = 1, \ldots, S$, which is determined by a generator matrix $\mathbf{Q} = [q_{ij}]$, $i, j = 1, \ldots, S$. The state of the environment determines the intensity of the arrival process, the values of which take the values $\lambda_s \geq 0$.

The discipline of servicing the incoming requests is also determined by the state of the environment at the time of the request arrival. If the environment is in the state $s(t) = s$, then the request is serviced during a random time distributed according to the exponential law with parameter μ_s : $F_n(x) = 1 - e^{-\mu_s x}$. The service time is not changed while the request is serviced, even if the random environment changes its state. Thus, the system simultaneously services requests with different service parameters, which is why such

systems are called heterogeneous [32,33]. Such systems more adequately describe the processes of information transfer, but their study is significantly complicated due to the multi-dimensionality of the process being studied. Let us denote the number of requests of type s (serviced with parameter μ_s) at time moment t by $i_s(t)$.

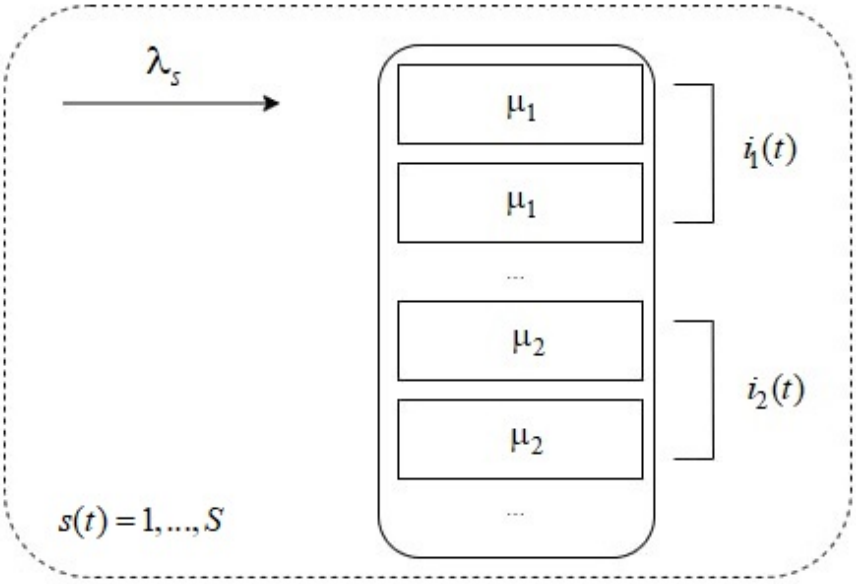

Figure 1. Heterogeneous queueing system in Markov random environment.

The problem is to study multi-dimensional Markov random process $\{i_1(t), i_2(t), \ldots, i_S(t), s(t)\}$, namely, to find its multi-dimensional joint probability distribution.

3. Kolmogorov Equations

Let us introduce the notation.

$$P(\mathbf{i}, s, t) = P(i_1, i_2, \ldots, i_S, s, t) = \Pr\{i_1(t) = i_1, i_2(t) = i_2, \ldots, i_S(t) = i_S, s(t) = s\},$$

and the notations for the row vectors

$$\mathbf{e}_1 = [1, 0, \ldots, 0], \mathbf{e}_2 = [0, 1, \ldots, 0], \ldots, \mathbf{e}_s = [0, 0, \ldots, 1],\ s = 1, \ldots, S.$$

Let us use the total probability formula and compose a system of equations for all $i \geq 0$ and $s = 1, \ldots, S$:

$$P(\mathbf{i}, s, t + \Delta t) = P(\mathbf{i}, s, t)(1 - \lambda_s \Delta t)(1 + q_{ss}\Delta t)(1 - i_1\mu_1\Delta t) \ldots (1 - i_S\mu_S\Delta t) +$$

$$P(\mathbf{i} - \mathbf{e}_s, s, t)\lambda_s\Delta t + \sum_{\nu=1}^{S} P(\mathbf{i} + \mathbf{e}_\nu, s, t)(i_\nu + 1)\mu_\nu\Delta t + \sum_{v \neq s} P(\mathbf{i}, v, t)q_{vs}\Delta t + o(\Delta t),$$

where

- $q_{vs}\Delta t$ is the probability that the environment changes its state from v to s during time interval Δt;
- $\lambda_s\Delta t$ is the probability of a new request arrival during time interval Δt when the environment is in state s;
- $i_s\mu_s\Delta t$ is the probability that one of the i_s requests of type s finishes its service;
- $(1 - \lambda_s\Delta t)(1 + q_{ss}\Delta t)(1 - i_1\mu_1\Delta t) \ldots (1 - i_S\mu_S\Delta t)$ is the probability that nothing happens in the system during time interval Δt when the environment is in state s and

the number of requests of the corresponding types present in the system are the components of vector **i**.

Let us divide each equation by Δt and make limit transition $\Delta t \to 0$; then the system takes the form

$$\frac{\partial P(\mathbf{i},s,t)}{\partial t} = -P(\mathbf{i},s,t)(\lambda_s + \sum_{\nu=1}^{S} \mu_\nu i_\nu) + \lambda_s P(\mathbf{i}-\mathbf{e}_s,s,t) +$$

$$+ \sum_{\nu=1}^{S} \mu_\nu (i_\nu+1) P(\mathbf{i}+\mathbf{e}_\nu,s,t) + \sum_{v\neq s} q_{vs} P(\mathbf{i},s,t)$$

with initial conditions

$$P(s,i_1,\ldots,i_S,t_0) = \begin{cases} r_s \text{ , if } i_1 = \cdots = i_S = 0, \\ 0, \text{otherwise}, \end{cases} \quad (1)$$

where $\mathbf{r} = [r_1, r_2, \ldots, r_S]$ is a row vector of the stationary probability distribution of the Markov chain $s(t)$, which can be evaluated from the system of linear equations

$$\begin{cases} \mathbf{rQ} = \mathbf{0}, \\ \mathbf{re} = 1. \end{cases} \quad (2)$$

Here and below **e** is a column vector which consists of ones.

Suppose that the system has a stationary regime, then we can write the system of linear equations for the stationary probabilities:

$$0 = -P(\mathbf{i},s)(\lambda_s + \sum_{\nu=1}^{S} \mu_\nu i_\nu) + \lambda_s P(\mathbf{i}-\mathbf{e}_s,s) + \sum_{\nu=1}^{S} \mu_\nu (i_\nu+1) P(\mathbf{i}+\mathbf{e}_\nu,s) + \sum_{v\neq s} q_{vs} P(\mathbf{i},v), \quad (3)$$

where $P(\mathbf{i},s)$ is the stationary probability distribution of the random process under consideration.

For further study, we use the method of characteristic functions. Let us denote

$$H(u_1,u_2,\ldots,u_s,s) = H(\mathbf{u},s) = \sum_{i_1} e^{ju_1 i_1} \sum_{i_2} e^{ju_2 i_2} \cdots \sum_{i_S} e^{ju_S i_S} P(\mathbf{i},s).$$

$H(u_1,u_2,\ldots,u_s,s)$ are the partial characteristic functions of the stationary probability distribution of the number of requests in the system. Using matrix notations

$$\mathbf{h}(\mathbf{u}) = [H(\mathbf{u},1),\ldots,H(\mathbf{u},S)],$$

$$\mathbf{m}(\mathbf{u}) = \left[\mu_1(e^{-ju_1}-1),\ldots,\mu_S(e^{-ju_S}-1)\right],$$

$$\frac{\partial \mathbf{h}(\mathbf{u})}{\partial \mathbf{u}} = \begin{bmatrix} \frac{\partial H(\mathbf{u},1)}{\partial u_1} & \cdots & \frac{\partial H(\mathbf{u},S)}{\partial u_1} \\ \cdots & \cdots & \cdots \\ \frac{\partial H(\mathbf{u},1)}{\partial u_S} & \cdots & \frac{\partial H(\mathbf{u},S)}{\partial u_S} \end{bmatrix}$$

$$\mathbf{\Lambda}(\mathbf{u}) = \begin{bmatrix} \lambda_1(e^{ju_1}-1) & 0 & \ldots & 0 \\ 0 & \lambda_2(e^{ju_2}-1) & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \ldots & \lambda_S(e^{ju_S}-1) \end{bmatrix}.$$

Finally, we can write the system of Kolmogorov Equation (3) in the matrix form for characteristic functions:

$$j\mathbf{m}(\mathbf{u})\frac{\partial \mathbf{h}(\mathbf{u})}{\partial \mathbf{u}} = \mathbf{h}(\mathbf{u})[\mathbf{\Lambda}(\mathbf{u}) + \mathbf{Q}]. \quad (4)$$

4. Evaluating Exact Moments of the Distribution

Using the properties of characteristic functions, we can determine the exact values of moments of the probability distribution of the number of requests of different types present in the system, including correlation coefficients.

To find the mathematical expectation of the number of requests of each type, we differentiate each equation of system (4) with respect to all variables $u_s, s = 1, 2, \ldots, K$. We obtain a system of differential equations, which we write in matrix form

$$j\frac{\partial \mathbf{m}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{h}(\mathbf{u})}{\partial \mathbf{u}} + j\mathbf{m}(\mathbf{u})\frac{\partial^2 \mathbf{h}(\mathbf{u})}{\partial \mathbf{u}^2} = [\mathbf{\Lambda}(\mathbf{u}) + \mathbf{Q}]\frac{\partial \mathbf{h}(\mathbf{u})}{\partial \mathbf{u}} + \frac{\partial \mathbf{\Lambda}(\mathbf{u})}{\partial \mathbf{u}}\mathbf{h}(\mathbf{u}), \quad (5)$$

where

$$\frac{\partial \mathbf{m}(\mathbf{u})}{\partial \mathbf{u}} = -j\begin{bmatrix} \mu_1 e^{-ju_1} & 0 & \ldots & 0 \\ 0 & \mu_2 e^{-ju_2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \mu_S e^{-ju_S} \end{bmatrix},$$

$$\frac{\partial \mathbf{\Lambda}(\mathbf{u})}{\partial \mathbf{u}} = j\begin{bmatrix} \lambda_1 e^{ju_1} & 0 & \ldots & 0 \\ 0 & \lambda_2 e^{ju_2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_S e^{ju_S} \end{bmatrix}.$$

Let us substitute $u_s = 0, s = 1, 2, \ldots, S$ and introduce the notations

$$\frac{\partial H(\mathbf{0}, s)}{\partial u_k} = \frac{\partial H(\mathbf{u}, s)}{\partial u_k}\bigg|_{u_1,\ldots,u_S=0} = \sum_{i_1}\cdots\sum_{i_S} i_k P(\mathbf{i}, s) = jm_1^k(s), k = 1, 2, \ldots, S.$$

Here $m_1^k(s)$ are partial moments of the first order, i.e., the average number of requests of the k-th type when the random environment is in state s.

$$\frac{\partial \mathbf{h}(\mathbf{0})}{\partial \mathbf{u}} = j\mathbf{M_1} = j\begin{bmatrix} m_1^1(1) & m_1^1(2)\ldots & m_1^1(S) \\ m_1^2(1) & m_1^2(2)\ldots & m_1^1(S) \\ \ldots & \ldots & \ldots \\ m_1^S(1) & m_1^S(2)\ldots & m_1^S(S) \end{bmatrix},$$

$$\mathbf{A} = \frac{\partial \mathbf{m}(\mathbf{0})}{\partial \mathbf{u}} = \begin{bmatrix} \mu_1 & 0 & \ldots & 0 \\ 0 & \mu_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \mu_S \end{bmatrix},$$

$$\mathbf{\Lambda} = \frac{\partial \mathbf{\Lambda}(\mathbf{0})}{\partial \mathbf{u}} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_S \end{bmatrix},$$

$$\mathbf{R} = \begin{bmatrix} r_1 & 0 & \ldots & 0 \\ 0 & r_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & r_S \end{bmatrix}.$$

Entries of matrix $\mathbf{R}$ are determined from the stationary probabilities of the states of the Markov chain (2).

As a result, we derive the following matrix equation for evaluating the first raw moments:

$$\mathbf{A}\mathbf{M}_1 - \mathbf{M}_1\mathbf{Q} = \mathbf{R}\boldsymbol{\Lambda}.$$

Let us multiply both parts of the equation by vector $\mathbf{e}$; we obtain

$$\mathbf{M}_1\mathbf{e} = \mathbf{A}^{-1}\mathbf{R}\boldsymbol{\Lambda}\mathbf{e}.$$

It follows that the mathematical expectation of the number of requests of the k-th type is calculated by the expression

$$\mathrm{E}\{i_k\} = m_1^{(k)} = \frac{\lambda_k r_k}{\mu_k}.$$

In a similar way, we can find raw moments of the second order. They are determined by expressions.

$$\mathrm{E}\{i_k^2\} = m_2^{(k)} = \frac{(2\lambda_k + \mu_k)m_1^{(k)}(k) + \lambda_k r_k}{2\mu_k}$$

for all $k = 1, \ldots, S$.

It is easy to obtain that the raw correlation moment for values of i_k and i_s is determined by the expression.

$$\mathrm{E}\{i_k i_s\} = m_2^{ks} = \frac{\lambda_k m_1^s(k) + \lambda_s m_1^k(s)}{\mu_s + \mu_k}.$$

A detailed description of the derivation of this formula for the special case of $S = 2$ can be found in [33].

Thus, the obtained formulas allow us to find exact values of the first- and second-order moments for the stationary distribution of the multi-dimensional random process of the number of requests of different types present in the system.

Obviously, the components of this multi-dimensional random process are dependent random variables. Let us consider how the system parameters affect the values of the correlation coefficient.

$$r^{(ks)} = \frac{m_2^{(ks)} - m_1^{(k)} m_1^{(s)}}{\sqrt{\mathrm{Var}^{(k)}\mathrm{Var}^{(s)}}}, \tag{6}$$

where

$$\mathrm{Var}^{(l)} = m_2^{(l)} - (m_1^{(l)})^2 \tag{7}$$

is a variation of component l.

Consider the following numerical example. Let the random environment have three possible states, which are determined by the generator matrix

$$\mathbf{Q} = \varepsilon \cdot \begin{bmatrix} -3 & 2 & 1 \\ 1. & -2 & 1 \\ 4 & 2 & -6 \end{bmatrix}, \tag{8}$$

where $\varepsilon > 0$ takes small values which reflect rare changes in the ransom environment (see Sections 5 and 6 for details). Let the intensities of arrivals be the following:

$$\lambda_1 = 1 \cdot N, \quad \lambda_2 = 2 \cdot N, \quad \lambda_3 = 10 \cdot N, \tag{9}$$

where $N > 0$ takes growing values which correspond to the case of growing intensity of arrivals. Let service rates for all types of requests be equal to $\mu_1 = \mu_2 = \mu_3 = 1$.

Numerical calculations show that the absolute value of the correlation coefficient (6) increases with an increase in the intensity of the arrival process (growing of N) and rarity of changes of states of the random environment (decreasing of ε). Values of the correlation coefficient $r^{(12)}$ of the number of requests of the first and second types present in the system are presented in Table 1.

Table 1. Dependence of values of correlation coefficient $r^{(12)}$ on the parameters of intensity of the arrival process N and rarity of the random environment changing ε.

N \ ε	**10**	**1**	**0.1**	**0.01**
1	−0.005	−0.044	−0.188	−0.304
10	−0.007	−0.063	−0.305	−0.582
100	−0.008	−0.066	−0.325	−0.641

5. Asymptotic Analysis Under Extremely Rare Changes of the Environment

The goal of this paper is to find the probability distribution $P(\mathbf{i}, s)$ of the random process under consideration in the steady-state regime. To do this, we need to solve the system of Equation (4). Unfortunately, it seems impossible to solve it in a direct way. So, we apply the method of asymptotic analysis similar to [31] to find its solution. In this paper, we will find the solution under the asymptotic condition of extremely rare changes in the states of the random environment. The difference from [31] is that here we have different service rates for customers of different types in the model, which leads us to the multi-dimensional model and multidimensionality of the system of equations in several components. So, we should try to find the solution taking into account this fact. To do this, we will expand (4) by components and solve the system as a series of equations with less multidimensionality.

The average duration of the Markov chain of stay in the k-th state is determined as $T_k = -1/q_{kk}$. So, a decrease of values q_{kk} leads us to rare changes of the states of the random environment. To perform asymptotic analysis, we represent the generator of the Markov chain in the form $\mathbf{Q}' = \varepsilon\mathbf{Q}$, where $\mathbf{Q}$ is some infinitesimal generator matrix with finite entries and parameter $\varepsilon > 0$ takes small values and tends to zero in theoretical derivations. Let us substitute $\mathbf{Q}'$ instead of $\mathbf{Q}$ in (4); we obtain the equation

$$j\mathbf{m}(\mathbf{u})\frac{\partial \mathbf{h}(\mathbf{u})}{\partial \mathbf{u}} = \mathbf{h}(\mathbf{u})(\mathbf{\Lambda}(\mathbf{u}) + \varepsilon\mathbf{Q}).$$

Let $\varepsilon \to 0$ here; then we obtain

$$j\mathbf{m}(\mathbf{u})\frac{\partial \mathbf{h}(\mathbf{u})}{\partial \mathbf{u}} = \mathbf{h}(\mathbf{u})\mathbf{\Lambda}(\mathbf{u}),$$

or in the expanded form:

$$j\begin{bmatrix} \sum_{k=0}^{S} \mu_k(e^{-ju_k} - 1)\dfrac{\partial H(\mathbf{u},1)}{\partial u_k} \\ \sum_{k=0}^{S} \mu_k(e^{-ju_k} - 1)\dfrac{\partial H(\mathbf{u},2)}{\partial u_k} \\ \cdots \\ \sum_{k=0}^{S} \mu_k(e^{-ju_k} - 1)\dfrac{\partial H(\mathbf{u},S)}{\partial u_k} \end{bmatrix} = \begin{bmatrix} \lambda_1(e^{ju_1} - 1)H(\mathbf{u},1) \\ \lambda_2(e^{ju_2} - 1)H(\mathbf{u},2) \\ \cdots \\ \lambda_S(e^{ju_S} - 1)H(\mathbf{u},S) \end{bmatrix}.$$

Consider the first equation of the system:

$$j\sum_{k=0}^{S}\mu_k(e^{-ju_k}-1)\frac{\partial H(\mathbf{u},1)}{\partial u_k}=\lambda_1(e^{ju_1}-1)H(\mathbf{u},1). \tag{10}$$

We can solve this partial differential equation by the method of characteristics:

$$\frac{du_1}{j\mu_1(e^{-ju_1}-1)}=\cdots=\frac{du_S}{j\mu_S(e^{-ju_S}-1)}=\frac{dH(\mathbf{u},1)}{\lambda_1(e^{ju_1}-1)H(\mathbf{u},1)}. \tag{11}$$

Consider the equations

$$\frac{du_1}{\mu_1(e^{-ju_1}-1)}=\frac{du_s}{\mu_s(e^{-ju_s}-1)}$$

for $s=2,\ldots,S$. From each of them, we derive

$$\left(\frac{e^{ju_1}-1}{C_s}\right)^{\frac{1}{\mu_1}}=(e^{ju_s}-1)^{\frac{1}{\mu_s}},$$

$$(e^{ju_s}-1)=\frac{1}{C_s}(e^{ju_1}-1)^{\frac{\mu_s}{\mu_1}},\ s=2,\ldots,S.$$

Then we consider the last equation from (11):

$$\frac{du_1}{j\mu_1(e^{-ju_1}-1)}=\frac{dH(\mathbf{u},1)}{\lambda_1(e^{ju_1}-1)H(\mathbf{u},1)}.$$

Performing the following derivations:

$$\frac{e^{ju_1}du_1}{j\mu_1(1-e^{ju_1})}=\frac{dH(\mathbf{u},1)}{\lambda_1(e^{ju_1}-1)H(\mathbf{u},1)},$$

$$\frac{d(e^{ju_1})}{j^2\mu_1(1-e^{ju_1})}=\frac{dH(\mathbf{u},1)}{\lambda_1(e^{ju_1}-1)H(\mathbf{u},1)},$$

$$\frac{d(e^{ju_1})}{\mu_1}=\frac{dH(\mathbf{u},1)}{\lambda_1 H(\mathbf{u},1)},$$

we obtain its solution in the form

$$H(\mathbf{u},1)=\Phi_1(C_1,\ldots,C_S)\exp\left(\frac{\lambda_1}{\mu_1}(e^{ju_1}-1)\right),$$

where $\Phi_1(C_1,\ldots,C_S)$ is some function which can be found from the initial condition:

$$\Phi_1(C_1,\ldots,C_S)=H(\mathbf{0},1)=r_1.$$

So, we obtain the following solution of (10):

$$H(\mathbf{u},1)=r_1\exp\left(\frac{\lambda_1}{\mu_1}(e^{ju_1}-1)\right).$$

In a similar way, we obtain

$$H(\mathbf{u},s)=r_s\exp\left(\frac{\lambda_s}{\mu_s}(e^{ju_s}-1)\right)$$

for each $s = 2, \ldots, S$. Summing up all these expressions, we find the characteristic function of the joint probability distribution of the number of requests of each type:

$$h(\mathbf{u}) = h(u_1, u_2, \ldots, u_S) = \sum_{s=1}^{S} r_s \exp\left(\frac{\lambda_s}{\mu_s}(e^{ju_s} - 1)\right). \quad (12)$$

One-dimensional characteristic functions of the stationary probability distribution of the number of requests of type $s, s \in \{1, \ldots, S\}$ have the form

$$h(u_s) = 1 - r_s + r_s \exp\left(\frac{\lambda_s}{\mu_s}(e^{ju_s} - 1)\right). \quad (13)$$

Therefore, the generating function has the form

$$F_s(z) = \mathrm{E}\{z^{i_s}\} = 1 - r_s + r_s \exp\left(\frac{\lambda_s}{\mu_s}(z - 1)\right).$$

It follows that the probability of the absence of requests of the s-th type is determined as

$$\pi_0^{(s)} = F_s(0) = 1 - r_s + r_s \exp\left(-\frac{\lambda_s}{\mu_s}\right).$$

6. Numerical Analysis

Because the result (12) is obtained under the asymptotic condition of extremely rare changes of the random environment ($\varepsilon \to 0$), it is necessary to check the accuracy and find the area of applicability of this result. To do this, we can compare distributions (12) and (13) with exact (not limiting) results.

First of all, we can notice that asymptotic and exact expressions (see Section 4) for the mathematical expectation coincide. Let us compare variations of these distributions. We can do this numerically, for example from Section 4.

For comparison, let us calculate relative errors of theoretical $\mathrm{Var}^{(s)}$ and asymptotical $\mathrm{Var}_{\mathrm{as}}^{(s)}$ variances of one-dimensional distributions of i_s, $s \in \{1, 2, 3\}$:

$$\Delta^{(s)} = \frac{\left|\mathrm{Var}^{(s)} - \mathrm{Var}_{\mathrm{as}}^{(s)}\right|}{\mathrm{Var}^{(s)}}.$$

The results of the evaluations for various values ε and $N = 1$ are presented in Table 2. So, we can draw a conclusion that the second moments of the asymptotic distribution become more precise with the decrease of ε and they have enough good accuracy (about 5% and less) for $\varepsilon \leq 0.01$.

Table 2. Relative errors of asymptotic variances of the number of requests of s-th type $\Delta^{(s)}$ for various values of asymptotic parameter ε.

ε	**1**	**0.1**	**0.01**	**0.001**
$\Delta^{(1)}$	0.481	0.293	0.059	0.007
$\Delta^{(2)}$	0.256	0.162	0.031	0.003
$\Delta^{(3)}$	1.967	0.464	0.053	0.005

Let us study why the error of the second moments of the asymptotic distribution is so big for greater values of ε. This follows from the probability distribution law (12) and (13): it has a jump in point $i = 0$ (see solid line in Figure 2). This jump can be big enough in relation to the nearest points, and the distribution has the second mode far from 0.

For example, for requests of the third type $\pi_0^{(3)} \approx 0.857$ while $\pi_1^{(3)} \approx 6.5 \cdot 10^{-5}$, and the distribution has the second mode in points 9, 10.

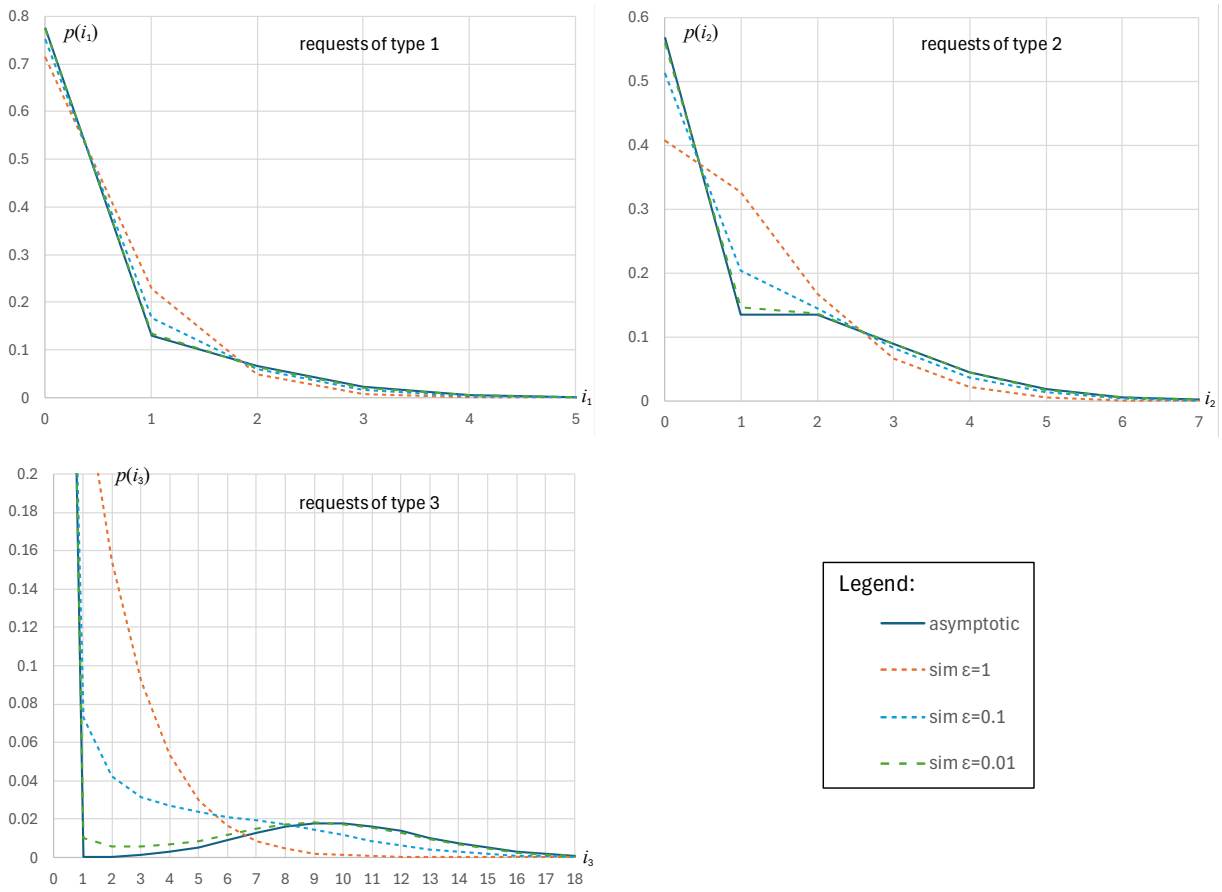

Figure 2. Probability distributions of the number of requests of different types obtained under asymptotic conditions and by using simulations.

For the analysis of the accuracy of the entire distribution, we performed experiments using simulation modeling. In Figure 2, you can see a graphical comparison of asymptotic distribution (it is not changed when ε changes its values) and probability distributions obtained from simulation results for various values of ε.

For numerical analysis of the accuracy of asymptotic distributions (13), we use Kolmogorov distances.

$$\Delta_\varepsilon^{(s)} = \max_{i \geq 0} \left| F_{\text{as}}^{(s)}(i) - F_\varepsilon^{(s)}(i) \right|,$$

where $F_{\text{as}}^{(s)}(i)$ is asymptotical one-dimensional cumulative distribution function of the number of requests of type s ($s \in \{1, 2, 3\}$), $F_\varepsilon^{(s)}(i)$ is cumulative distribution function of the number of requests of type s obtained from simulation for chosen value of ε. Values of the Kolmogorov distances evaluated for various values of ε are presented in Table 3.

The number of iterations in the simulation experiments is chosen from a criterion that an estimated error of the simulation result should not be greater 0.001. Such an error does not influence the evaluation of the asymptotic result accuracy. We estimate the error of the simulation by comparing the results of several runs of the simulation and evaluating the maximum difference between them in terms of the Kolmogorov distance. So, we discover that the number of iterations greater than 2 million is enough to reach such accuracy.

Table 3. Values of Kolmogorov distances $\Delta_\varepsilon^{(s)}$ for various values of asymptotic parameter ε for the distributions of the number of requests of different types s.

ε	**1**	**0.5**	**0.1**	**0.05**	**0.01**
$\Delta_\varepsilon^{(1)}$	0.061	0.050	0.024	0.014	0.003
$\Delta_\varepsilon^{(2)}$	0.161	0.140	0.056	0.033	0.009
$\Delta_\varepsilon^{(3)}$	0.469	0.379	0.163	0.093	0.027

Let us assume that the accuracy of the asymptotic distribution is good enough if its value of the Kolmogorov distance is equal to or less than 0.05. So, from the table, we see that such values of accuracy are achieved when $\varepsilon \leq 0.5$ for requests of the first type, $\varepsilon \leq 0.05$ for requests of the second type, and $\varepsilon \leq 0.01$ for requests of the third type.

So, we can draw a conclusion that asymptotic result (12) provides good accuracy for the cases when the average number of requests in the system is not too big (there are no two modes in the distribution) and intensities of changing states of the random environment are small enough. To be more precise, based on the comparative analysis conducted, it can be concluded that the asymptotic Formulas (12) and (13) can be applied when values of the average intensity of changes in the state of random environment do not exceed 0.03.

Let us consider practical usage of the obtained results. First, let us discuss what "rare changes of the states of random environment" mean. Consider the example with parameters (8) and (9), evaluated distributions presented in Figure 2 and estimations of their error presented in Table 3. The average intensity of arrivals in this model is about 2.7. Because we have found that the average intensity of the state changes should be 0.03 or less to the precision of approximation (12) and (13) be acceptable, then we can draw a conclusion that the frequency of the random environment state changes should be less than the intensity of arrivals by 100 times or greater. In practice, this means that the changes should occur after 100 requests or more on average. Such a picture is typical enough for real computing systems—in real systems, we usually encounter situations where thousands or millions of requests of one type arrive before the balancer swithces to receiving another type of request.

Let us discuss the implications of a probability distribution (12) and (13) for real-world problems. When we have a probability distribution represented analytically, we can use it to solve more complex problems of analyzing and improving system performance. In particular, such problems may be related to system optimization. Let us say, in the given example, requests of the second and third types are very demanding on computing resources, but their execution is cost-effective. We can formulate an optimality criterion that depends on the number of requests of these types simultaneously present in the system (a negative factor) and the number of requests of these types processed (a positive factor). The criterion can have the form

$$\Theta(\mathbf{Q}, \mathbf{\Lambda}) = -F^-\big(h(u_1, u_2, u_3)\big) + F^+\big(g(n_2, n_3)\big), \tag{14}$$

where $F^-(\cdot)$ is a cost function of the resources consumed by requests of types 2 and 3 during their execution, $F^+(\cdot)$ is a profit function of the number of the requests of the second and third types (n_2 and n_3 respectively) processed by the system during fixed time interval, $\mathbf{Q}$ and $\mathbf{\Lambda}$ are parameters of the system (intensities of the state changes and arrivals) which we can vary to reach maximum value of the criterion. By adjusting these parameters, for example, the time of the random environment being in states 2 and 3 (the average time when the balancer receives requests of these types), we can find conditions where

criterion (14) reaches its maximum value. Unfortunately, a specific solution to problems of this type is beyond the scope of this article.

7. Conclusions

In the paper, we consider the mathematical model of request processing in the cloud system in the form of an infinite-server queue with several types of customers and a random environment. The multi-dimensional probability distribution of the number of requests of different types present in the system is obtained under the condition of rare changes in the random environment. The numerical results and simulation modeling approve the result obtained. Some interesting properties of the distribution are considered.

The result obtained can be helpful in the analysis of real cloud systems and for understanding the processes of requests processing. It can help to answer questions, for example, why we have too many requests of one type in comparison to requests of other types, how many resources should be provided by the server for normal operations (if the resource requirements depend on the types of requests), solving optimization problems, and so on.

Author Contributions: Conceptualization, S.M.; methodology, E.P. and S.M.; software, A.M. and J.S.; writing, original draft preparation, E.P. and S.M.; writing, review and editing, A.M. and J.S.; supervision S.M. All authors have read and agreed to the published version of the manuscript.

Funding: The research is supported by the Russian Science Foundation according to the research project No. 24-21-00454, https://rscf.ru/project/24-21-00454/ (accessed on 5 September 2025).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: We are very grateful to the reviewers for their advice and comments. This allowed us to significantly improve the text of the article.

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

References

1. O'Cinneide, C.A.; Purdue, P. The M/M/∞ Queue in a Random Environment. *J. Appl. Prob.* **1986**, *23*, 175–184.
2. Falin, G. The M/M/∞ queue in a random environment. *Queueing Syst.* **2008**, *58*, 65–76. [CrossRef]
3. Fralix, B.H.; Adan, I.J.B.F. An infinite-server queue influenced by a semi-Markovian environment. *Queueing Syst.* **2009**, *61*, 65–84. [CrossRef]
4. Liu, Z.; Yu, S. The M/M/C queueing system in a random environment. *J. Math. Anal. Appl.* **2016**, *436*, 556–567. [CrossRef]
5. Bin Sun; Dudin, S.A.; Dudina, O.S.; Dudin, A.N. A Customer Service Model in an Adaptive-Modulation Mobile Communication Cell with Allowance for Random Environment. *Autom. Remote Control* **2021**, *82*, 812–826. [CrossRef]
6. Spiridovska, N. Markov-Modulated Processes, Their Applications and Big Data Cases: State of the Art. *Lect. Notes Netw. Syst.* **2020**, *117*, 100–109.
7. Eisen, M. Stochastic variations in queueing processes. *Oper. Res.* **1963**, *11*, 922–927. [CrossRef]
8. Naor, P.; Yechiali, U. Queueing problems with heterogeneous arrivals and service. *Opens. Res.* **1971**, *19*, 722–734.
9. Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models*; The John Hopkins University Press: Baltimore, MD, USA; London, UK, 1981; 351p.
10. Boxma, O.; Mandjes, M.; Heemskerk, M. Single-server queues under overdispersion in the heavy-traffic regime. *Stoch. Model.* **2020**, *37*, 197–230. [CrossRef]
11. Kim, C.; Dudin, A.; Dudina, O.; Kim, J. Queueing system operating in a random environment as a model of a cell operation. *Ind. Eng. Manag. Syst.* **2016** *15*, 131–142. [CrossRef]
12. Dudin, S.A.; Dudin, A.N.; Dudina, O.S. Analysis of a retrial queue with limited processor sharing operating in the random environment. *Lect. Notes Comput. Sci.* **2017**, *10372*, 38–49.
13. Krishtalev, N.; Lisovskaya, E.; Moiseev, A. Resource Queueing System M/GI/∞ in a Random Environment. *Lect. Notes Comput. Sci.* **2021**, *13144*, 211–225.
14. Naumov, V.; Samouylov, K. Resource system with losses in a random environment. *Mathematics* **2021**, *9*, 2685. [CrossRef]

15. Anbazhagan, N.; Vinitha, V.; Acharya, S.; Amutha, S.; Jeganathan, K.; Seo, C.; Kim, H.-I. The MAP/PH/N/∞ Queueing-Inventory System with Demands from a Random Environment. *IEEE Access* **2022**, *10*, 47371–47383. [CrossRef]
16. Jacob, J.; Shajin, D.; Krishnamoorthy, A.; Vishnevsky, V.; Kozyrev, D. Queueing-inventory with one essential and m optional items with environment change process forming correlated renewal process. *Mathematics* **2022**, *10*, 104. [CrossRef]
17. Joshua, V.C.; Mathew, A.P.; Krishnamoorthy, A. An MMAP/M/∞ Queueing System with an Offer Zone Working in a Random Environment. *Commun. Comput. Inf. Sci.* **2019**, *1141*, 244–257.
18. Linton, D. An M/G/∞ Queue with m Customer Types Subject to Periodic Clearing. *Opsearch* **1979**, *16*, 80–88.
19. Nazarov, A.; Baymeeva, G. The M/G/∞ Queue in Random Environment. *Commun. Comput. Inf. Sci.* **2014**, *487*, 312–324. [CrossRef]
20. Nazarov, A.; Baymeeva, G. The M/GI/∞ system subject to semi-Markovian random environment. *Commun. Comput. Inf. Sci.* **2015**, *564*, 128–140.
21. Baykal-Gursoy, M. Stochastic Decomposition in M/M/∞ Queues with Markov Modulated Service Rates. *Queueing Syst.* **2004**, *48*, 75–88. [CrossRef]
22. Blom, J. Markov-Modulated Infinite-Server Queues with General Service Times. *Queueing Syst.* **2014** *76*, 403–424. [CrossRef]
23. Kerobyan, K.; Enakoutsa, K.; Kerobyan R. Analysis of an Infinite-Server Queue $MAP_k|G_k|\infty$ in Random Environment with k Markov Arrival Streams and Random Volume of Customers. *Commun. Comput. Inf. Sci.* **2018**, *912*, 305–320.
24. D'Auria, B. M/M/∞ queues in semi-Markovian random environment. *Queueing Syst.* **2008**, *58*, 221–237. [CrossRef]
25. D'Auria, B. Stochastic decomposition of the M/G/∞ queue in a random environment. *Oper. Res. Lett.* **2007**, *35*, 805–812. [CrossRef]
26. Bai, W.H.; Xi, J.Q.; Zhu, J.X.; Huang, S.W. Performance analysis of heterogeneous data centers in cloud computing using a complex queuing model. *Math. Probl. Eng.* **2015**, *1*, 1–15. [CrossRef]
27. Vetha, S.; Devi, K.V. Dynamic resource allocation in cloud using queueing model. *J. Ind. Pollut. Control* **2017**, *33*, 1547–1554.
28. Hanini, M.; Kafhali, S.E.; Salah, K. Dynamic VM allocation and traffic control to manage QoS and energy consumption in cloud computing environment. *Int. J. Comput. Appl. Technol.* **2019**, *60*, 307–316. [CrossRef]
29. Fedorova, E.; Lapatin, I.; Lizyura, O.; Moiseev, A.; Nazarov, A.; Paul, S. Queueing System with Two Phases of Service and Service Rate Degradation. *Axioms* **2023**, *12*, 104. [CrossRef]
30. Jafarnejad Ghomi, E.; Rahmani, A.M.; Qader, N.N. Applying queue theory for modeling of cloud computing: A systematic review. *Concurr. Comput. Pract. Exp.* **2019**, *31*, e5186. [CrossRef]
31. Gorbatenko, A.E.; Nazarov, A.A. Research of the Semi-Markovian process in conditions of limitedly rare changes in its state. *Vestn. Sibsau. Aerosp. Tehnol. Control Syst.* **2010**, *7*, 44–48.
32. Pankratova, E.V.; Moiseeva, S.P.; Farhadov, M.P.; Moiseev, A.N. Heterogeneous System MMPP/GI(2)/∞ with Random Customers Capacities. *J. Sib. Fed. Univ. Math. Phys.* **2019**, *12*, 231–239. [CrossRef]
33. Polin, E.P.; Moiseeva, S.P.; Moiseev, A.N. Heterogeneous queueing system with Markov renewal arrivals and service times dependent on states of arrival process. *Discret. Contin. Model. Appl. Comput. Sci.* **2023**, *31*, 105–119. [CrossRef]