mathematics

MDPI

Article

Analysis of a Multi-Server Queue with Group Service and Service Time Dependent on the Size of a Group as a Model of a Delivery System

Sergei Dudin * and Olga Dudina

Department of Applied Mathematics and Computer Science, Belarusian State University, 4, Nezavisimosti Ave., 220030 Minsk, Belarus; dudina@bsu.by

* Correspondence: dudins@bsu.by or dudin85@mail.ru

Abstract: In this paper, we consider a multi-server queue with a finite buffer. Request arrivals are defined by the Markov arrival process. Service is provided to groups of requests. The minimal and maximal group sizes are fixed. The service time of a group has a phase-type distribution with an irreducible representation depending on the size of the group. The requests are impatient. The patience time for an arbitrary request has an exponential distribution. After this time expires, the request is lost if all servers are busy or, if some server is idle, with a certain probability, all requests staying in the buffer start their service even if their number is below the required minimum. The behavior of the system is described by a multi-dimensional continuous-time Markov chain that does not belong to the class of level-independent quasi-birth-and-death processes. The algorithm for the computation of the stationary distribution of this chain is presented, and expressions for the computation of the queuing system's performance characteristics are derived. The description of a delivery system operation in terms of the analyzed queuing model is given, and the problem of the optimization of its operation is numerically solved. Multi-server queues with a phase-type distribution for the group service time that are dependent on the size of the group, the account of request impatience, and the correlated arrival process have not previously been analyzed in the existing literature. However, they represent a precise model of many real-world objects, including delivery systems.

Keywords: group service; multi-server queue; *MAP*; phase-type distribution; delivery system

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

Citation: Dudin, S.; Dudina, O. Analysis of a Multi-Server Queue with Group Service and Service Time Dependent on the Size of a Group as a Model of a Delivery System. *Mathematics* **2023**, *11*, 4587. https://doi.org/10.3390/math11224587

Academic Editor: Steve Drekic

Received: 16 October 2023
Revised: 6 November 2023
Accepted: 7 November 2023
Published: 9 November 2023

1. Introduction

The problem of the optimal management of various goods and grocery delivery has become especially important due to the quick development of ordering via the Internet over the last few years. In particular, COVID-19 caused this development due to the better isolation of members of society achieved via online shopping. The business model of delivery systems suggests, as a rule, the collection of customer orders via the Internet and then the delivery of the ordered items to the customers via some transport vehicles. Typically, the size and weight of orders may not be very large, and a vehicle can deliver several orders simultaneously. Therefore, because the orders are generated at random moments, the number of orders uploaded into a vehicle is random, and the delivery times are random, adequate mathematical modeling of a delivery system can be done within the borders of the theory of queues with group service.

In this paper, we build and analyze a queuing model for a delivery system. Generally speaking, the process of providing a service to a customer in a delivery system consists of two phases. The first phase corresponds to registering the orders obtained via the Internet or phone and packing them into some containers designed for item delivery. The second

phase corresponds to the delivery of these containers (throughout the paper, we call them requests) to the customers by the vehicles. A corresponding two-phase (tandem) queuing model was recently analyzed in [1]. The essential restriction of the model considered in [1] is the assumption that a single vehicle performs a delivery. This suggestion does not hold in many large real-world delivery systems. In this paper, we consider only the second phase of the delivery process, with the assumption that the fleet of available vehicles can have many items.

1.1. Short Literature Review

Queuing models with a group service have been a focus of research for about seventy years (see, e.g., the papers [2,3]). As early works, we can also mention [4–8]. Some surveys of relevant research can be found, e.g., in [9–17]. The overwhelming majority of papers devoted to queues with a group service assume that the stationary Poisson process is an arrival flow. But, this process poorly describes flows in many real-world systems. Essential restrictions of the stationary Poisson process are assumptions of a constant arrival rate that does not fluctuate during the system's operation and has values of 1 and 0 for the coefficients of variation and correlation of the inter-arrival times, respectively. Models with an arbitrary distribution of inter-arrival times have been analyzed, e.g., in [18–20].

Another possibility for modeling in a more general way than the stationary Poisson arrival process is the model of the Markov arrival process (MAP) mentioned above, which has been known since the early 1990s. Queuing systems with the MAP and group service have been considered, e.g., in papers [1,11,15–17,21–32].

The majority of these papers consider single-server queues. To the best of our knowledge, multi-server queues with a group service and the MAP flow (or generalization of the MAP to the case of batch arrivals, that is, the batch Markov arrival process) were considered only in Refs. [33,34], where systems with an infinite and a finite buffer were dealt with, respectively; Ref. [35], where the number of servers in a system from [34] can vary; Ref. [36], where a two-server queue was under study; and [12], where the retrial queuing model $BMAP/M/c$ with a constant retrial rate was considered.

1.2. Disadvantages of Previous Research

The models of multi-server queues with the MAP (or $BMAP$) and group service considered in the papers [12,33–36] have some disadvantages in terms of the following aspects:

- The service time distribution is assumed to be exponential, while a more general PH distribution should be considered. This would allow significantly better fitting of the actual distributions of service time in real-world systems;
- The possible impatience of requests, which is an inherent feature of many real-world systems, see, e.g., [37], is not taken into account;
- Service can only be provided to groups with fixed minimum and maximum sizes in [12,33–36]. In some real-world systems, the service of groups of sizes that are less than the fixed minimum is allowed if a request loss due to impatience can occur;
- Definitely, the service time of a group may depend on the size of the group. Such a dependence is assumed in this paper. Among the papers [12,33–36], such a dependence was considered only for a two-server queue in [36].

Due to the possible applicability of the considered queuing model to a variety of real-world systems, it is important to analyze models that are free from these disadvantages.

1.3. Aim of the Study

Taking into account the listed disadvantages, the aim of this paper is formulated as follows. We intend to analytically and numerically build and investigate a queuing model of a delivery system that satisfies the following requirements:

- A finite number of vehicles provide delivery, and this number can be greater than one. The minimal and maximal values of the sizes of groups to which a service can be provided are fixed;

- The process of request arrival has to take into account the typical features of real flows of orders. Namely, inter-arrival times can have a wide range of initial moments (mean value, variance, the third, fourth, etc.) and a wide range of values for the coefficient of correlation of successive inter-arrival times. The instantaneous arrival rate may essentially fluctuate during the system's operation. This account will be done via the assumption that the requests arrive in the MAP. For more information about such a process, see, e.g., [38–43];
- The possible impatience (perishability, obsolescence, etc.) of requests has to be taken into account. To satisfy this requirement, we assume that after a random time interval (with an exponential distribution), the waiting request is canceled (lost) if all vehicles are busy. If at least one vehicle is idle, the system makes a randomized choice among the options to lose a request or to start the delivery process, even if the number of requests in the buffer is below the preassigned minimum;
- The delivery time of a group of requests has to depend on the number of requests in the group. Typically, the delivery of a group of requests from a warehouse to customers residing in some area consists of a permanent part, which is the transportation time between the warehouse and this area, a time for each container to upload into the vehicle, and an individual delivery time inside the area. To satisfy this requirement, we assume that the delivery (service) time has a phase-type (PH) distribution with an irreducible representation depending on the size of the group. For more information about the PH distribution and its usefulness in stochastic modeling, see, e.g., [44];
- Algorithms for the computation of the stationary distribution of the system's states and its main performance characteristics should provide a high computation speed for a not very large number of vehicles and buffer size. The number of vehicles (servers) and the capacity of the buffer for waiting requests should not be very small and should match the corresponding parameters of real-world systems. This requirement is explained by the necessity of using algorithms to solve various optimization problems, which require the computation of the performance measures of the system for different values of its parameters, including the capacity of the buffer and the number of servers.

1.4. Contributions of the Paper

The formulated aims are achieved, and the main contributions of the paper are as follows:

- We consider a multi-server queue with a group service of the $MAP/PH/N$ type, which allows us to model the typical bursty character of traffic in real-world systems and a wide range of service time distributions. Previously, queues with a group service, MAP, and PH distribution of service times were analyzed only in single-server settings;
- We account for the possible impatience of the requests waiting in the queue and apply more flexible control for the service initiation. The most common strategy of control allows the service to begin only when the number of requests in the queue is not less than a preassigned threshold. We considered the strategy that allows us to start servicing in situations where the number of requests is below this threshold, but some requests decide to leave the system without service due to impatience;
- We assume that the service time of a group depends on its size, while the majority of the papers consider a more simple case of service time that is independent of this size;
- The elaborated algorithms and software allow for the computation of the main performance measures and the solving of optimization problems for realistic system parameter values. Numerical results are presented for a system with 50 servers and a capacity of 300 for the input buffer.

1.5. Brief Outline of the Content of the Paper

Section 2 contains a mathematical description of the constructed queuing model and the necessary denotations. A multi-dimensional continuous-time Markov chain that is suitable for the description of the constructed model is defined in Section 3. The infinitesimal

generator of this chain is obtained. An algorithm for the computation of its steady-state distribution is presented here. Formulas for the computation of the key performance characteristics of the system are presented in Section 4. Some numerical results that provide insight into the system's behavior and an example of a solution to the optimization problem are given in Section 5. Section 6 concludes the paper.

2. Mathematical Model

We consider a queuing system with a finite buffer of size R with N independent identical servers. The structure of the system is shown in Figure 1.

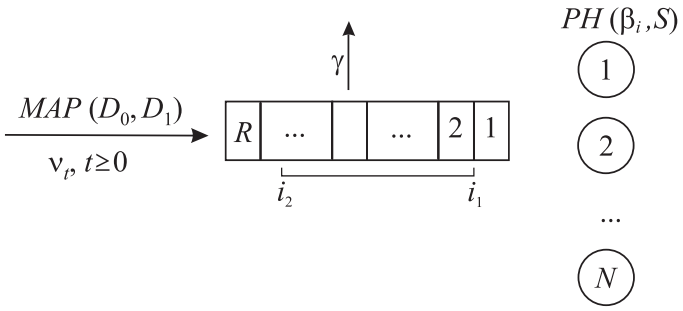

Figure 1. Structure of the system.

Requests enter the system in the MAP arrival flow. This flow is given by a Markov chain with continuous time ($CTMC$) ν_t, $t \geq 0$, having a state space $\{1, 2, ..., W\}$ and matrices D_0 and D_1 of size W. Here, W is a certain finite integer, and the matrix D_1 consists of the transition intensities of the chain ν_t, accompanied by the arrival of a request. The non-diagonal elements of the matrix D_0 determine the rate of the corresponding transition of the chain ν_t without the arrival of a request. The modules of the negative diagonal elements determine the intensity of the exit of the process ν_t from the corresponding state. The matrix $D(1) = D_0 + D_1$ is the generator of the $CTMC$ ν_t. This generator is assumed to be irreducible.

The average request arrival rate λ is determined by the formula $\lambda = \boldsymbol{\theta} D_1 \mathbf{e}$, where $\boldsymbol{\theta}$ is a row vector of stationary probabilities of the $CTMC$ ν_t. This vector is the only solution to the system $\boldsymbol{\theta} D(1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$. Here and below, $\mathbf{0}$ is a row vector of an appropriate size consisting of zeros, and $\mathbf{e}$ is a column vector of an appropriate size consisting of ones.

The requests can be served in groups of varying sizes. All requests from the group accepted for service go to one server and complete their services simultaneously. We assume that if, at the service completion epoch, there are less than i_1 requests in the buffer, then a new service does not start. Otherwise, the server starts its service for all available requests if their number does not exceed the parameter i_2, such that $\max\{1, i_1 - 1\} < i_2 \leq R$. In other words, if there are more than i_2 requests in the buffer at the service completion epoch, then the first i_2 requests go for servicing, and the rest remain in the buffer. Thus, the parameters i_1 and i_2, $1 \leq i_1 < i_2 \leq R$, determine the minimum and maximum sizes of the groups that can be taken for service.

If, at the arrival moment, the number of requests in the buffer is less than or equal to $i_1 - 2$, then the incoming request becomes buffered and awaits service. If the number of requests in the buffer is $i_1 - 1$ and there is a free server, then the entire group of size i_1 starts its service. If all servers are busy, the request becomes buffered to wait. If the buffer is full, the request permanently leaves the system and is lost.

It is also assumed that requests in the buffer can become impatient and try to leave the system without service at random time intervals, the lengths of which are exponentially distributed with the intensity of γ. In this case, if there is a free server with the probability q_i, where i is the number of requests in the buffer, all available requests go to the server and begin servicing, even though their number is less than the parameter i_1, and with a complimentary probability, the request leaves the system forever. If all servers are busy, the impatient request is lost.

We assume that the service time of a group has a phase-type distribution (PH-service). The $CTMC$ m_t, $t \geq 0$, with transient states $\{1, 2, \ldots, M\}$ and a unique absorbing state $M+1$ specify this distribution. The irreducible representation of the $CTMC$ m_t, $t \geq 0$, is given as $(\boldsymbol{\beta}_i, S)$, $i = \overline{1, i_2}$, where i is the number of requests taken for service. Note that $\boldsymbol{\beta}_i$ is a stochastic row vector of dimension M, and the square matrix S of dimension M is a sub-generator.

The average service time for a group of requests of size i is defined as $b_1^{(i)} = \boldsymbol{\beta}_i(-S)^{-1}\mathbf{e}$. Note that, assuming that the initial probability vector of the service time depends on the size of the group, we take into account the dependence of the service process on the size of the group.

To apply the formulated queuing model for the description and optimization of some real-world delivery system, it is necessary to determine the matrices D_0 and D_1 that define the MAP of the requests and representations $(\boldsymbol{\beta}_i, S)$, $i = \overline{1, i_2}$, to define the service of groups of requests.

The problem of building the matrices D_0 and D_1 based on observations of the traces of the actual arrival process (or time stamps giving the arrival epochs) is well addressed in the existing literature. For references, see, e.g., the book [45] and papers [45–54].

The problem of choosing representations $(\boldsymbol{\beta}_i, S)$, $i = \overline{1, i_2}$, to fit the known average service times of groups of different sizes can be solved, e.g., as follows:

Let the average service time of a group consisting of i requests be estimated based on the available samples as w_i, $i = \overline{1, i_2}$. It is natural to assume that the values w_i, $i = \overline{1, i_2}$, are such that $w_1 \leq w_2 \leq \cdots \leq w_{i_2}$ (a larger group does not imply a shorter average service time) and $w_1 \neq w_{i_2}$ (dependence of the service time on the group size exists).

The number M of transient states of the $CTMC$ m_t, $t \geq 0$, has a significant impact on the feasibility of the analysis of the $CTMC$, as shown in the next section. Thus, one should construct the irreducible representations $(\boldsymbol{\beta}_i, S)$, $i = \overline{1, i_2}$, of the minimal possible size, namely, $M = 2$.

It is possible to fix the sub-generator S as a diagonal matrix of size two with the diagonal entries $-w_1^{-1}$ and $-w_{i_2}^{-1}$. It is easy to check that the desired values w_i, $i = \overline{1, i_2}$, can be achieved via the following choice of vectors $\boldsymbol{\beta}_i$ of size 2,

$$\boldsymbol{\beta}_i = (\varphi_i, 1 - \varphi_i)$$

where

$$\varphi_i = \frac{w_{i_2} - w_i}{w_{i_2} - w_1}, \; i = \overline{1, i_2}.$$

Let us analyze the constructed queuing system under the assumption that the inter-arrival, service, and patience times are mutually independent.

3. The Process of System States and Its Stationary Distribution

Let i_t, $i_t = \overline{0, R}$, be the number of requests in the buffer; n_t, $n_t = \overline{0, N}$, be the number of busy servers; ν_t, $\nu_t = \overline{1, W}$, be the state of the underlying process; and MAP, $m_t^{(l)}$ be the number of servers on the l-th service phase. $m_t^{(l)} = \overline{0, n_t}$, $l = \overline{1, M}$, $\sum\limits_{l=1}^{M} m_t^{(l)} = n_t$, at an arbitrary moment t, $t \geq 0$.

The behavior of the system under study is described by a regular irreducible $CTMC$ with continuous time

$$\xi_t = \{i_t, n_t, \nu_t, m_t^{(1)}, \ldots, m_t^{(M)}\}, \; t \geq 0.$$

Let us renumber the states of the $CTMC$ ξ_t in reverse order for the components $m_t^{(1)}, \ldots, m_t^{(M)}$ and in direct order for the components n_t and ν_t and denote the set of states of the chain with the value i of the first component of the $CTMC$ as level i, $i \geq 0$.

Theorem 1. *The generator Q of the CTMC ξ_t, $t \geq 0$, has the following lower-Hessenberg block structure:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \dots & O & O \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & \dots & O & O \\ Q_{2,0} & Q_{2,1} & Q_{2,2} & Q_{2,3} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ Q_{R-1,0} & Q_{R-1,1} & Q_{R-1,2} & Q_{R-1,3} & \dots & Q_{R-1,R-1} & Q_{R-1,R} \\ Q_{R,0} & Q_{R,1} & Q_{R,2} & Q_{R,3} & \dots & Q_{R,R-1} & Q_{R,R} \end{pmatrix}$$

where the non-zero blocks $Q_{i,j}$, $j \leq i+1$, containing the intensities of transitions from level i to level j are defined as follows:

$$Q_{i,i} = \mathrm{diag}\{D_0, D_0 \oplus (A_n + \Delta_n),\ n = \overline{1,N}\} +$$

$$+\mathrm{diag}^{-}\{I_W \otimes L_n,\ n = \overline{1,N}\} - i\gamma I_{W\sum_{n=0}^{N} T_n} +$$

$$+\delta_{i_1,1}\mathrm{diag}^{+}\{D_1 \otimes P_n(\boldsymbol{\beta}_1),\ n = \overline{0,N-1}\},\ i = \overline{0,i_1-1},$$

$$Q_{i,i} = D_0 \oplus (A_N + \Delta_N) - i\gamma I_{WT_N},\ i = \overline{i_1,R-1},$$

$$Q_{R,R} = D(1) \oplus (A_N + \Delta_N) - R\gamma I_{WT_N},$$

$$Q_{i,i+1} = \mathrm{diag}\{D_1 \otimes I_{T_n},\ n = \overline{0,N}\},\ i = \overline{0,i_1-2},$$

$$Q_{i_1-1,i_1} = \begin{pmatrix} O_{W\sum_{n=0}^{N-1} T_n \times WT_N} \\ D_1 \otimes I_{T_N} \end{pmatrix},$$

$$Q_{i,i+1} = D_1 \otimes I_{T_N},\ i = \overline{i_1,R-1},$$

$$Q_{1,0} = \begin{pmatrix} O_{WT_N \times W\sum_{n=0}^{N-1} T_n} & \gamma I_{WT_N} + I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_1) \end{pmatrix},\ \textit{if } i_1 = 1,$$

$$Q_{1,0} = \gamma\mathrm{diag}\{\{(1-q_1)I_{WT_n},\ n = \overline{0,N-1}\}, I_{WT_N}\} +$$

$$+\mathrm{diag}^{+}\{D_1 \otimes P_n(\boldsymbol{\beta}_2),\ n = \overline{0,N-1}\} +$$

$$+\gamma q_1\mathrm{diag}^{+}\{I_W \otimes P_n(\boldsymbol{\beta}_1),\ n = \overline{0,N-1}\},\ \textit{if } i_1 = 2,$$

$$Q_{1,0} = \gamma\mathrm{diag}\{\{(1-q_1)I_{WT_n},\ n = \overline{0,N-1}\}, I_{WT_N}\} +$$

$$+\gamma q_1\mathrm{diag}^{+}\{I_W \otimes P_n(\boldsymbol{\beta}_1),\ n = \overline{0,N-1}\},\ \textit{if } i_1 > 2,$$

$$Q_{i,i-1} = i\gamma\mathrm{diag}\{\{(1-q_i)I_{WT_n},\ n = \overline{0,N-1}\}, I_{WT_N}\},\ i = \overline{2,i_1-1},$$

$$Q_{i_1,i_1-1} = i_1\gamma\begin{pmatrix} O_{WT_N \times W\sum_{n=0}^{N-1} T_n} & I_{WT_N} \end{pmatrix},\ i_1 \neq 1,$$

$$Q_{i,i-1} = i\gamma I_{WT_N},\ i = \overline{i_1+1,R},$$

$$Q_{i,0} = i\gamma q_i\mathrm{diag}^{+}\{I_W \otimes P_n(\boldsymbol{\beta}_i),\ n = \overline{0,N-1}\},\ i = \overline{2,i_1-2},$$

$$Q_{i_1-1,0} = \mathrm{diag}^{+}\{D_1 \otimes P_n(\boldsymbol{\beta}_{i_1}),\ n = \overline{0,N-1}\} +$$

$$+(i_1-1)\gamma q_{i_1-1}\mathrm{diag}^{+}\{I_W \otimes P_n(\boldsymbol{\beta}_{i_1-1}),\ n = \overline{0,N-1}\},\ i_1 \neq 1,$$

$$Q_{i,0} = \begin{pmatrix} O_{WT_N \times W\sum_{n=0}^{N-1} T_n} & I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_i) \end{pmatrix},\ i = \overline{i_1,i_2},\ \textit{if } i_1 \neq 1,$$

$$\textit{and } i = \overline{i_1+1,i_2},\ \textit{if } i_1 = 1,$$

$$Q_{i,i-i_2} = \begin{pmatrix} O_{WT_N \times W \sum_{n=0}^{N-1} T_n} & I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_{i_2}) \end{pmatrix}, \text{ if } i - i_2 < i_1,$$

$$Q_{i,i-i_2} = I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_{i_2}), \text{ if } i - i_2 \geq i_1, \ i = \overline{i_2 + 1, R}.$$

Here,

$\otimes$ *and* $\oplus$ *are the symbols of the Kronecker product and the sum of matrices (see, for example, [55]);*

I *is the identity matrix;* O *is the zero matrix, the dimension of which is indicated by a subscript if necessary;*

$\delta_{i,j}$ *is the Kronecker symbol, that is,* $\delta_{i,j} = \begin{cases} 1, & i = j; \\ 0, & i \neq j; \end{cases}$

$\mathrm{diag}\{d_1, d_2, \ldots, d_n\}$ *is the diagonal matrix with the diagonal elements* $d_1, d_2, \ldots, d_n$;

$\mathrm{diag}^+\{d_1, d_2, \ldots, d_n\}$ *is the square matrix with the non-zero overdiagonal elements* $d_1, d_2, \ldots, d_n$;

$\mathrm{diag}^-\{d_1, d_2, \ldots, d_n\}$ *is the square matrix with the non-zero subdiagonal elements* $d_1, d_2, \ldots, d_n$; *and*

the number T_n *is equal to the cardinality of the state space of the process* $\{m_t^{(1)}, \ldots, m_t^{(M)}\}$ *when a service is simultaneously provided to n groups of requests. It is calculated as*

$$T_n = \frac{(n + M - 1)!}{n!(M - 1)!}, \ n = \overline{1, N}.$$

For convenience, we put $T_0 = 1$.

The matrix L_n *defines the transition intensities of the process* $\{m_t^{(1)}, \ldots, m_t^{(M)}\}$ *at the moment when service in one of n busy servers is completed,* $n = \overline{1, N}$. *The matrix* A_n *contains the transition intensities of the process* $\{m_t^{(1)}, \ldots, m_t^{(M)}\}$ *at the moment of the change in the phase of service in one of n busy servers,* $n = \overline{1, N}$. *The matrix* $P_n(\boldsymbol{\beta}_i)$ *defines the transition probabilities of the process* $\{m_t^{(1)}, \ldots, m_t^{(M)}\}$ *at the moment when the group of i requests begins service in the presence of n busy servers,* $n = \overline{0, N-1}$. *The diagonal elements of the diagonal matrix* Δ_n *determine the rates of the exit of the process* $\{m_t^{(1)}, \ldots, m_t^{(M)}\}$ *from the corresponding states. When the matrices* L_n *and* A_n *are computed, the matrices* Δ_n *are computed by the formula*

$$\Delta_n = -\mathrm{diag}\{A_n \mathbf{e} + L_n \mathbf{e}\}.$$

Detailed descriptions of the matrices $P_n(\boldsymbol{\beta}_i)$ $n = \overline{0, N-1}$, $i = \overline{1, i_2}$, L_n, A_n, Δ_n, $n = \overline{1, N}$, *and algorithms for their calculation are presented in [56].*

Proof. The proof of the theorem was carried out by analyzing the intensities of all possible transitions of the *CTMC* ξ_t over a time interval of infinitesimal length.

The generator has a block lower Hessenberg structure, since requests can enter the buffer strictly one at a time and leave it in groups, the size of which is up to i_2.

Let us explain the form of the diagonal blocks $Q_{i,i}$, $i = \overline{0, R}$. All diagonal elements of the block $Q_{i,i}$ are negative, and the absolute values of these elements determine the intensities of the *CTMC* ξ_t exiting the corresponding states. The *CTMC* ξ_t can exit the current state in the following cases:

a. The underlying process ν_t of the request arrivals leaves its current state. The corresponding transition intensities are determined up to their signs by the diagonal elements of the matrix $D_0 \otimes I_{\sum_{n=0}^{N} T_n}$, if $i = \overline{0, i_1 - 1}$, $D_0 \otimes I_{T_N}$ if $i = \overline{i_1, R-1}$, and $(D_0 + D_1) \otimes I_{T_N}$ if $i = R$.

b. The service process on one of the busy servers changes its state. In this case, the transition intensities are determined by the diagonal elements of the matrices $\mathrm{diag}\{O_{W \times W}, I_W \otimes \Delta_n, \ n = \overline{1, N}\}$ for $i = \overline{0, i_1 - 1}$ and $I_W \otimes \Delta_N$ for $i = \overline{i_1, R}$.

c. A request from the buffer tries to leave the system due to impatience; the corresponding intensities are specified by the matrices $i\gamma I_{W\sum_{n=0}^{N}T_n}$ if $i = \overline{0, i_1 - 1}$ and $i\gamma I_{WT_N}$, if $i = \overline{i_1, R}$.

The non-diagonal elements of the matrix $Q_{i,i}$ determine the transition intensities of the *CTMC* ξ_t without changing the value i of the first component. The following statements specify these transitions:

(a) For non-diagonal elements of matrices $D_0 \otimes I_{\sum_{n=0}^{N}T_n}$ if $i = \overline{0, i_1 - 1}$*and* $D_0 \otimes I_{T_N}$ if $i = \overline{i_1, R}$ when the underlying process ν_t makes a transition without generating a request.

(b) For non-diagonal elements of the matrix $D_1 \otimes I_{T_N}$, when the underlying process ν_t makes a transition, a generated request is lost due to the buffer being full (case $i = R$).

(c) For elements of the matrix $\mathrm{diag}^{-}\{I_W \otimes L_n, n = \overline{1, N}\}$, when the process $\{m_t^{(1)}, \ldots, m_t^{(M)}\}$ makes a transition leading to the end of the service but a new service does not start because there are fewer than i_1 requests in the buffer;

(d) For elements of the matrices $I_W \otimes A_n$, $i = \overline{0, i_1 - 1}$, and $I_W \otimes A_N$, $i = \overline{i_1, R}$, when the process $\{m_t^{(1)}, \ldots, m_t^{(M)}\}$ makes a transition that does not lead to the end of the service.

(e) In the case of $i_1 = 1$, for the elements of the matrix $\mathrm{diag}^{+}\{D_1 \otimes P_n(\boldsymbol{\beta}_1), n = \overline{0, N - 1}\}$, when a new request arrives and the buffer is empty. In this case, the arriving request is immediately processed for service.

As a result, we obtain blocks $Q_{i,i}$, $i = \overline{0, R}$, presented above.

The form of the blocks $Q_{i,i+1}$, $i = \overline{0, R}$ is explained as follows. These blocks contain the transition rates of the *CTMC* ξ_t as the number of requests in the buffer increases by one. This can only happen when a new request arrives in the system. The transition intensities of the process ν_t at the moment of the request arrival are determined by the elements of the matrix D_1; therefore, the blocks $Q_{i,i+1}$ are specified by the matrix $\mathrm{diag}\{D_1 \otimes I_{T_n}, n = \overline{0, N}\}$ if $i < i_1 - 1$, the matrix $\begin{pmatrix} O_{W\sum_{n=0}^{N-1}T_n \times WT_N} \\ D_1 \otimes I_{T_N} \end{pmatrix}$ if there are $i_1 - 1$ requests in the buffer, and by the matrix $D_1 \otimes I_{T_N}$ in all other cases.

Next, consider the blocks $Q_{i,j}$, $j < i$, $i = \overline{1, R}$. First, let us explain the case where $i = 1$. The form of the block $Q_{1,0}$ significantly depends on the value of the parameter i_1. Let us consider the following three cases:

(1) If $i_1 = 1$ (the server starts servicing, even if there is only one request in the buffer), then reducing the number of requests in the buffer is possible only if the service is completed (the matrix $\begin{pmatrix} O_{WT_N \times W\sum_{n=0}^{N-1}T_n} & I_W \otimes L_N P_{N-1}(\boldsymbol{\beta}_1) \end{pmatrix}$ specifies the intensity of this event) or the request is loss due to impatience (the matrix $\begin{pmatrix} O_{WT_N \times W\sum_{n=0}^{N-1}T_n} & \gamma I_{WT_N} \end{pmatrix}$ gives the corresponding intensities).

(2) If $i_1 = 2$, then a decrease in the number of requests in the buffer results in (a) the loss of a request due to impatience. The intensities of this event are given by the matrix $\gamma\mathrm{diag}\{\{(1 - q_1)I_{WT_n}, n = \overline{0, N - 1}\}, I_{WT_N}\}$. Event (b) is also possible when there is a free server and a single request in the buffer. This request wants to leave the system due to impatience, but the system decides to service this single request. The corresponding intensities are given by the matrix $\gamma q_1 \mathrm{diag}^{+}\{I_W \otimes P_n(\boldsymbol{\beta}_1), n = \overline{0, N - 1}\}$. In addition, case (c) is added. When there is a free server and a single request in the buffer, a new request arrives, and a group of two requests begins service $\mathrm{diag}^{+}\{D_1 \otimes P_n(\boldsymbol{\beta}_2), n = \overline{0, N - 1}\}$.

(3) If $i_1 > 2$, then, as in the previous case, a decrease in the number of requests in the buffer is caused by events (a), (b), and (c), which is impossible. The corresponding intensity matrices are determined in a similar way.

Then, consider the blocks $Q_{i,i-1}$, $i = \overline{2, R}$. The transition of *CTMC* ξ_t from the state with the value i for the first component in the state with a value $i - 1$ for this component is possible only if the request is lost due to impatience. The corresponding intensities are

given by the matrix $i\gamma\mathrm{diag}\{\{(1-q_i)I_{WT_n},\ n=\overline{0,N-1}\},I_{WT_N}\}$ for $i<i_1$, by the matrix $i_1\gamma\begin{pmatrix} O_{WT_N\times W\sum_{n=0}^{N-1}T_n} & I_{WT_N}\end{pmatrix}$ when there are i_1, $i_1\neq 1$, requests in the buffer, and by the matrix $i\gamma I_{WT_N}$ when there are more than i_1 requests in the buffer.

Let us explain the form of the blocks $Q_{i,0}$, $i=\overline{2,i_2}$. The transition of the $CTMC$ ξ_t from the state with a value i for the first component of the chain to the state where a value of zero occurs when a group of size i is taken over for service. The following statements give the corresponding intensities:

a. For the matrix $i\gamma q_i\mathrm{diag}^+\{I_W\otimes P_n(\boldsymbol{\beta}_i),\ n=\overline{0,N-1}\}$, when there is a free server and i, $i=\overline{2,i_1-1}$ requests in the buffer, one of the requests wants to leave the buffer due to impatience, but the system decides to take all requests from the buffer for servicing, although their number is less than the minimum number allowable;

b. For the matrix $\mathrm{diag}^+\{D_1\otimes P_n(\boldsymbol{\beta}_{i_1}),\ n=\overline{0,N-1}\}$, when, at the moment of the arrival of a new request, a free server is available and i_1-1, $i_1\neq 1$, requests are already in the buffer, and a group of i_1 requests is taken for servicing;

c. For the matrix $\begin{pmatrix} O_{WT_N\times W\sum_{n=0}^{N-1}T_n} & I_W\otimes L_NP_{N-1}(\boldsymbol{\beta}_i)\end{pmatrix}$, when, at the moment when one of the busy servers is released, a group of i requests waiting for service in the buffer is taken over for service, $i=\overline{i_1,i_2}$ if $i_1\neq 1$ and $i=\overline{i_1+1,i_2}$ if $i_1=1$.

The blocks $Q_{i,i-i_2}$, $i=\overline{i_2+1,R}$, contain the transition intensities of the $CTMC$ ξ_t from the state with the value i for the first component of the chain to a state with the value $i-i_2$ for this component in the case when a group of requests of size i_2 is taken for service at the moment of service completion on one of the busy servers. These intensities are determined by the matrix $\begin{pmatrix} O_{WT_N\times W\sum_{n=0}^{N-1}T_n} & I_W\otimes L_NP_{N-1}(\boldsymbol{\beta}_{i_2})\end{pmatrix}$ if the size of the group of requests remaining in the buffer does not exceed the parameter i_1 or the matrix $I_W\otimes L_NP_{N-1}(\boldsymbol{\beta}_{i_2})$ otherwise.

Theorem 1 is proven. □

Because the $CTMC$ ξ_t is regular and irreducible and its state space is finite, the following invariant probabilities of the states of the chain exist:

$$\pi(i,n,\nu,m^{(1)},\dots,m^{(M)})=\lim_{t\to\infty}P\{i_t=i,n_t=n,\nu_t=\nu,m_t^{(1)}=m^{(1)},\dots,m_t^{(M)}=m^{(M)}\},$$

$$i=\overline{0,R},\ n=\overline{0,N},\ \nu=\overline{1,W},\ m^{(l)}=\overline{0,n},\ l=\overline{1,M},\ \sum_{l=1}^{M}m^{(l)}=n.$$

Then, let us form the vectors $\boldsymbol{\pi}_i$, $i=\overline{0,R}$ of these probabilities enumerated in the reverse order of the components $m^{(1)},\dots,m^{(M)}$ and in the direct order of the components n and ν.

It is well known that the probability vectors $\boldsymbol{\pi}_i$, $i=\overline{0,R}$ satisfy the following system of linear algebraic equations:

$$(\boldsymbol{\pi}_0,\boldsymbol{\pi}_1,\dots,\boldsymbol{\pi}_R)Q=\mathbf{0},\ \ (\boldsymbol{\pi}_0,\boldsymbol{\pi}_1,\dots,\boldsymbol{\pi}_R)\mathbf{e}=1,$$

which are called an equilibrium or Chapman–Kolmogorov equations.

To solve this system, we developed the algorithm presented below Algorithm 1.

Remark 1. *Because the generator Q does not have a three-block diagonal form, the presented algorithm is not trivial. It effectively exploits the block structure of the generator. All matrices, the inversions of which are used in this algorithm, are non-singular due to O. Tausski theorem (see, e.g., [42,57]), because they are irreducible sub-generators and the diagonal entry strictly dominates in at least one row. If the matrix, say $\mathcal{A}$, is the irreducible sub-generator with the mentioned dominance, then (i) the matrix $\mathcal{A}^{-1}$ exists and (ii) the matrix $(-\mathcal{A})^{-1}$ is non-negative. Thus, the presented algorithm works with only non-negative matrices and does not use a subtraction operation.*

Hence, it is numerically stable. Therefore, it may be successfully exploited for the computation of the stationary distribution of the states of a variety of queuing systems with group service of requests.

Algorithm 1: Computation of the stationary probabilities

1. Calculate the matrices $\mathcal{X}_k^{(i)}$ by using the following formulas:

$$\mathcal{X}_k^{(0)} = \begin{cases} -Q_{k,0}Q_{0,0}^{-1}, & k = \overline{1, i_2}; \\ O_{WT_N \times W \sum_{n=0}^{N-1} T_n}, & k = \overline{i_2+1, R}, \end{cases}$$

$$\mathcal{X}_k^{(i)} = \begin{cases} -(\mathcal{X}_{i+1}^{(i-1)}Q_{i-1,i} + Q_{i+1,i})(Q_{i,i} + \mathcal{X}_i^{(i-1)}Q_{i-1,i})^{-1}, & k = i+1; \\ -\mathcal{X}_k^{(i-1)}Q_{i-1,i}(Q_{i,i} + \mathcal{X}_i^{(i-1)}Q_{i-1,i})^{-1}, & k = \overline{i+2, i_2+i-1}; \\ -Q_{i_2+i,i}(Q_{i,i} + \mathcal{X}_i^{(i-1)}Q_{i-1,i})^{-1}, & k = i_2+i; \\ \begin{cases} O_{WT_N \times W \sum_{n=0}^{N-1} T_n}, & i < i_1; \\ O_{WT_N \times WT_N}, & i \geq i_1; \end{cases} & k = \overline{i_2+i+1, R}; \end{cases}$$

$$i = \overline{1, R-i_2},$$

$$\mathcal{X}_k^{(i)} = \begin{cases} -(\mathcal{X}_{i+1}^{(i-1)}Q_{i-1,i} + Q_{i+1,i})(Q_{i,i} + \mathcal{X}_i^{(i-1)}Q_{i-1,i})^{-1}, & k = i+1; \\ -\mathcal{X}_k^{(i-1)}Q_{i-1,i}(Q_{i,i} + \mathcal{X}_i^{(i-1)}Q_{i-1,i})^{-1}, & k = \overline{i+2, R}; \end{cases},$$

$$i = \overline{R-i_2+1, R-2},$$

$$\mathcal{X}_R^{(R-1)} = -(\mathcal{X}_R^{(R-2)}Q_{R-2,R-1} + Q_{R,R-1})(Q_{R-1,R-1} + \mathcal{X}_{R-1}^{(R-2)}Q_{R-2,R-1})^{-1}.$$

2. Calculate the matrices Y_i as follows:

$$Y_R = I,\ Y_i = \sum_{k=i+1}^{R} Y_k \mathcal{X}_k^{(i)},\ i = R-1, R-2, \ldots, 0.$$

3. Find the vector $\boldsymbol{\pi}_R$ as the only solution to the system

$$\boldsymbol{\pi}_R(\mathcal{X}_R^{(R-1)} + Q_{R,R}) = \mathbf{0},$$

$$\boldsymbol{\pi}_R \sum_{k=0}^{R} Y_k \mathbf{e} = 1.$$

4. Calculate the vectors $\boldsymbol{\pi}_i$ as $\boldsymbol{\pi}_i = \boldsymbol{\pi}_R Y_i$, $i = \overline{0, R-1}$.

4. Performance Characteristics

Having calculated the probability vectors $\boldsymbol{\pi}_i$, $i = \overline{0, R}$, it is possible to present the formulas for the computation of the main characteristics of the performance of the considered system.

The average number of requests in the buffer is

$$L_{buffer} = \sum_{i=1}^{R} i\boldsymbol{\pi}_i \mathbf{e}.$$

The average number of busy servers is

$$N_{serv} = \sum_{i=0}^{i_1-1} \sum_{n=1}^{N} n\boldsymbol{\pi}(i,n)\mathbf{e} + \sum_{i=i_1}^{R} N\boldsymbol{\pi}_i \mathbf{e}.$$

The average intensity of the server releases is

$$\mu_{release} = \sum_{i=0}^{i_1-1} \sum_{n=1}^{N} \boldsymbol{\pi}(i,n)(I_W \otimes L_n)\mathbf{e} + \sum_{i=i_1}^{R} \boldsymbol{\pi}_i(I_W \otimes L_N)\mathbf{e}.$$

The probability that an incoming request will find the buffer full and leave the system is

$$P_{ent-loss} = \frac{1}{\lambda}\boldsymbol{\pi}_R(D_1 \otimes I_{T_N})\mathbf{e}.$$

The probability that a request will start its service immediately upon entering the system is

$$P_{to-serv} = \frac{1}{\lambda} \sum_{n=0}^{N-1} \boldsymbol{\pi}(i_1 - 1, n)(D_1 \otimes I_{T_n})\mathbf{e}.$$

The rate of requests entered for service is

$$\mu_{to-serv} = i_1 \sum_{n=0}^{N-1} \boldsymbol{\pi}(i_1 - 1, n)(D_1 \otimes I_{T_n})\mathbf{e} +$$

$$+ \sum_{i=i_1}^{R} \min\{i, i_2\}\boldsymbol{\pi}_i(I_W \otimes L_N)\mathbf{e} + \sum_{i=1}^{i_1-1} i^2 q_i \gamma \sum_{n=0}^{N-1} \boldsymbol{\pi}(i,n)\mathbf{e}.$$

The probability of losing a request from the buffer due to impatience is

$$P_{imp-loss} = \frac{1}{\lambda}\Big(\sum_{i=1}^{i_1-1} i(1-q_i)\gamma \sum_{n=0}^{N-1} \boldsymbol{\pi}(i,n)\mathbf{e} + \sum_{i=1}^{i_1-1} i\gamma\boldsymbol{\pi}(i,N)\mathbf{e} + \sum_{i=i_1}^{R} i\gamma\boldsymbol{\pi}_i\mathbf{e}\Big).$$

The average size of a group of requests taken for servicing is

$$N_{batch} = \frac{\mu_{to-serv}}{\mu_{release}}.$$

The probability of losing a request from the buffer due to impatience while there is a free server is

$$P^{imp-loss}_{idle-server} = \frac{1}{\lambda} \sum_{i=1}^{i_1-1} i\gamma(1-q_i) \sum_{n=0}^{N-1} \boldsymbol{\pi}(i,n)\mathbf{e}.$$

The probability of losing a request from the buffer due to impatience at a time when all servers are busy is

$$P^{imp-loss}_{all-busy-servers} = \frac{1}{\lambda}\Big[\sum_{i=1}^{i_1-1} i\gamma\boldsymbol{\pi}(i,N)\mathbf{e} + \sum_{i=i_1}^{R} i\gamma\boldsymbol{\pi}_i\mathbf{e}\Big].$$

The probability that, at an arbitrary moment, there is at least one free server in the system is

$$P_{idle-server} = \sum_{i=0}^{i_1-1} \sum_{n=0}^{N-1} \boldsymbol{\pi}(i,n)\mathbf{e}.$$

The probability that at an arbitrary moment there are requests in the buffer while there is at least one free server is

$$P^{requests}_{idle-server} = \sum_{i=1}^{i_1-1} \sum_{n=0}^{N-1} \boldsymbol{\pi}(i,n)\mathbf{e}.$$

The probability of losing an arbitrary request is

$$P_{loss} = 1 - \frac{\mu_{to-serv}}{\lambda} = P_{imp-loss} + P_{ent-loss}.$$

The probability that a group with a size of less than i_1 undergoes servicing is

$$P_{batch<i_1} = \frac{1}{\mu_{release}} \sum_{i=1}^{i_1-1} iq_i\gamma \sum_{n=0}^{N-1} \boldsymbol{\pi}(i,n)\mathbf{e}.$$

The probability that a group of the maximum size i_2 undergoes servicing is

$$P_{batch=i_2} = \frac{1}{\mu_{release}} (\sum_{i=i_2}^{R} \boldsymbol{\pi}_i(I_W \otimes L_N)\mathbf{e} + \delta_{i_1-i_2,0} \sum_{n=0}^{N-1} \boldsymbol{\pi}(i_1-1,n)(D_1 \otimes I_{T_n})\mathbf{e}).$$

The probability that a group of the maximum size from the interval $[i_1, i_2)$ undergoes servicing is

$$P_{i_1 \le batch<i_2} = \frac{(1-\delta_{i_1-i_2,0})}{\mu_{release}} (\sum_{n=0}^{N-1} \boldsymbol{\pi}(i_1-1,n)(D_1 \otimes I_{T_n})\mathbf{e} + \sum_{i=i_1}^{i_2-1} \boldsymbol{\pi}_i(I_W \otimes L_N)\mathbf{e}).$$

5. Numerical Examples

In this section, we consider the problem of the optimal choice of the number of required vehicles N and the value of the parameter i_1 that defines the minimal desirable group size for service in a delivery system.

To this end, let us consider a storehouse that can place up to $R = 300$ orders for delivery. We believe that one vehicle can accommodate up to $i_2 = 20$ orders. Orders are received at the storehouse in accordance with the MAP that is specified by the matrices

$$D_0 = \begin{pmatrix} -10.1599 & 0.32778 \\ 0.32778 & -2.76287 \end{pmatrix}, D_1 = \begin{pmatrix} 9.44979 & 0.382355 \\ 0.0491604 & 2.38593 \end{pmatrix},$$

and has an average intensity of $\lambda = 5$ orders per minute as well as the coefficient of variation $c_{var} = 1.8333$ and the correlation coefficient $c_{cor} = 0.183092$.

The delivery time of a group consisting of i orders by one vehicle has a phase-type distribution with the parameters $(\boldsymbol{\beta}_i, S)$, where the matrix S has the form

$$S = \begin{pmatrix} -0.01 & 0 \\ 0 & -0.05 \end{pmatrix},$$

and the vector $\boldsymbol{\beta}_i$ is given as

$$\boldsymbol{\beta}_i = (i/i_2, 1 - i/i_2),\ i = \overline{1, i_2}.$$

Thus, we assume that the average delivery time for one order is $b_1^{(1)} = 24$ min, and the average delivery time for $i_2 = 20$ orders is $b_1^{(i_2)} = 100$ min.

We assume that the intensity of impatience of orders from the buffer is determined by the parameter $\gamma = 0.01$. Thus, the average time that an arbitrary request can wait in the buffer is 100 min.

As noted earlier, the purpose of this experiment is to determine the optimal values for the parameters of the number of vehicles N and the threshold i_1, which defines the minimal desirable size for the group taken for service.

Let us vary the parameter N from 1 to 50 with a step of 1 and the parameter i_1 from 1 to $i_2 = 20$, also with a step of 1. We assume that the probabilities defining the tolerance to servicing a group smaller in size than the parameter i_1 in the case of the end of the patience time of one of the requests in the buffer are given as

$$q_i = i/i_1,\ i = \overline{1, i_1 - 1}.$$

First, let us analyze the dependence of the main characteristics of the system's performance on the parameters N and i_1. Figures 2 and 3 illustrate the dependence of the average number L_{buffer} of requests in the buffer and the average number N_{serv} of busy servers on the parameters N and i_1.

Figures 4 and 5 illustrate the dependence of the probabilities $P_{idle-server}$ and $P^{requests}_{idle-server}$ on the parameters N and i_1.

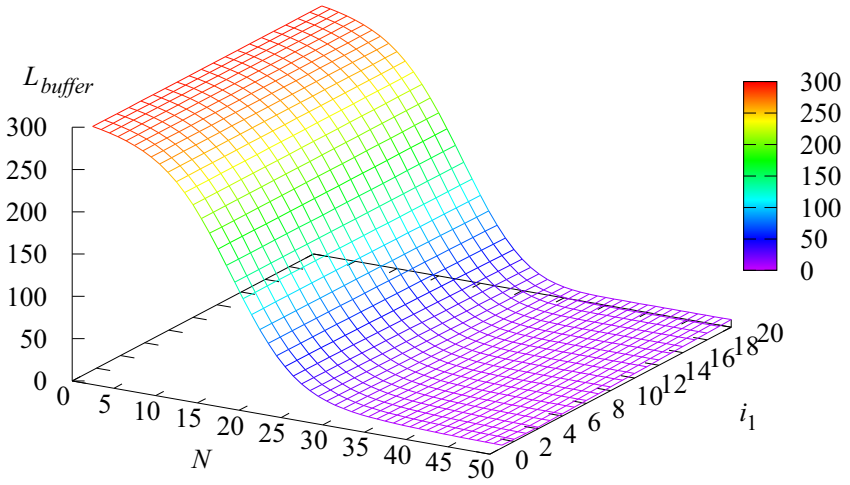

Figure 2. Dependence of the average number L_{buffer} of requests in the buffer on the parameters N and i_1.

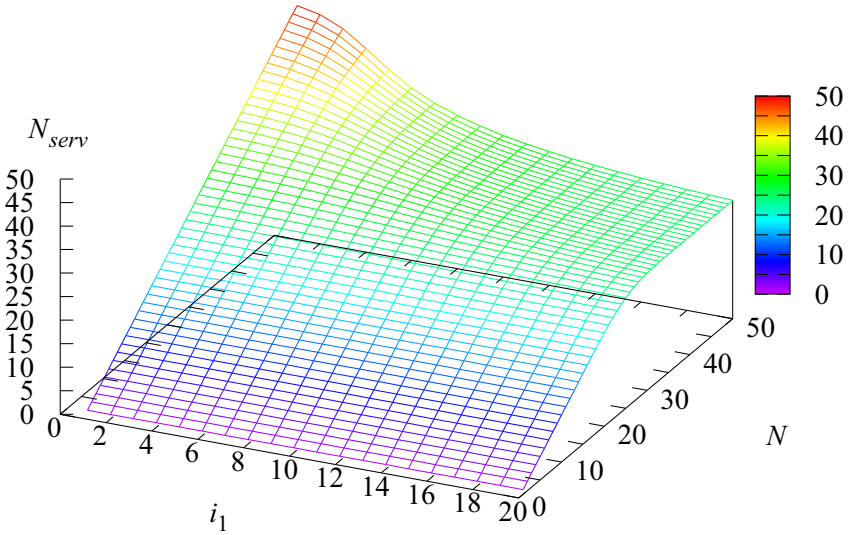

Figure 3. Dependence of the average number N_{serv} of busy servers on the parameters N and i_1.

As can be seen from Figure 2, the average number L_{buffer} of requests in the buffer decreases as the number of available servers increases. For small values of N, the values of L_{buffer} practically do not depend on the parameter i_1. For example, for $N = 5$, $L_{buffer} = 285.16345$ for $i_1 = 1$ and for $i_1 = 20$. This is explained by the fact that, for the given system parameters, for small values of N, the probability that there will be fewer than i_2 requests in the buffer is negligibly small, and the servers do not stand idle but begin servicing, as follows on from Figures 4 and 5, immediately after being released. For the same reason, other performance characteristics in the case under consideration also do not depend on i_1 for small values of N. For bigger values of N, in the considered case, the average number L_{buffer} of requests in the buffer increases with an increase in the parameter i_1. For example, for $N = 50$, $L_{buffer} = 3.05371$ for $i_1 = 1$ and $L_{buffer} = 8.95773$ for $i_1 = 20$.

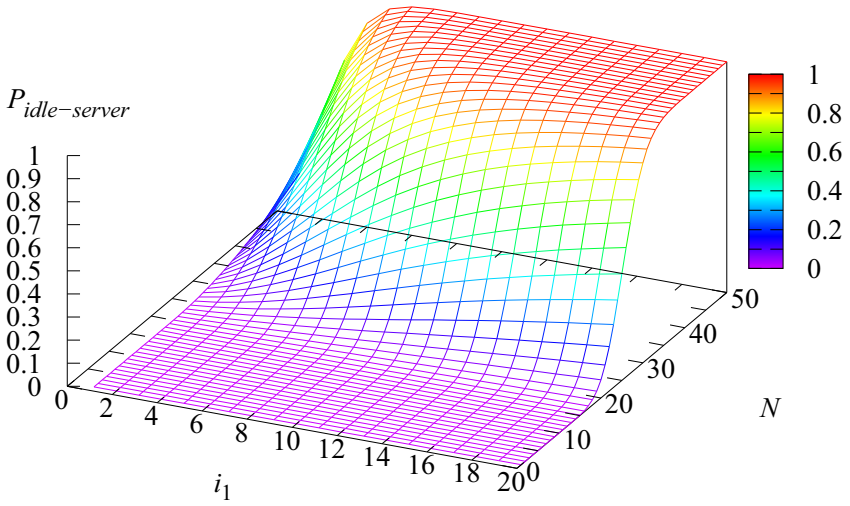

Figure 4. Dependence of the probability $P_{idle-server}$ that, at an arbitrary moment, there is at least one free server on the parameters N and i_1.

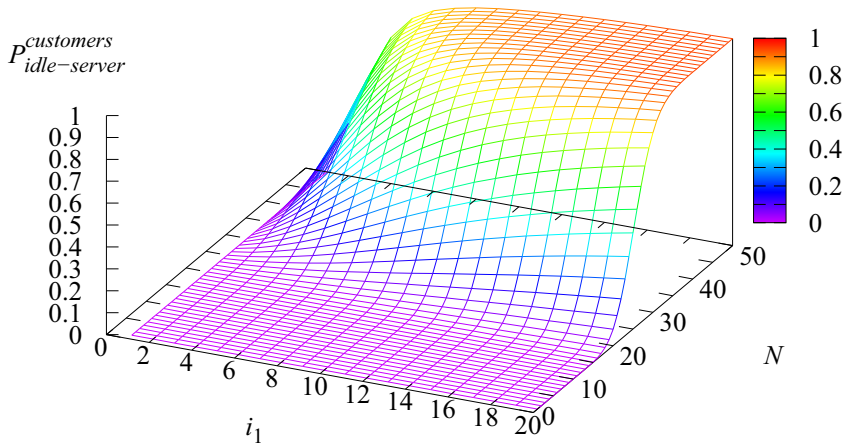

Figure 5. Dependence of the probability $P^{requests}_{idle-server}$ that, at an arbitrary moment, there are requests in the buffer, while there is at least one free server on the parameters N and i_1.

As shown in Figure 3, the average number of busy servers increases when the parameter N increases. For small values of N, the growth is linear; that is, an increase in N by one leads to an increase in N_{serv} by the same unit. With a further increase in N, the dependence on N weakens, while for larger values of i_1, this dependence weakens earlier. As N grows, N_{serv} depends on i_1. An increase in i_1 leads to a decrease in N_{serv}, since an increase in i_1 leads to an increase in the probability $P_{idle-server}$ that, at an arbitrary moment, there is at least one free server and the probability $P^{requests}_{idle-server}$ that there is at least one free server while there are requests in the buffer, which are presented in Figures 4 and 5. Thus, servers are idle more often, which leads to a decrease in the average number of occupied servers.

Figures 6 and 7 show the dependence of the average size N_{batch} of a group of requests taken for servicing and the probability $P_{batch<i_1}$ that a group with a size of less than i_1 undergoes servicing on the parameters N and i_1.

As one can see from Figure 6, for small N values, the service is provided only for groups of requests with a maximal size of $i_2 = 20$. As N increases, the values of N_{batch} decrease, and the rate of decrease significantly depends on the value of the parameter i_1: the higher i_1, the lower the rate of decrease of N_{batch}. For example, for $N = 50$, $N_{batch} = 3.33746$ for $i_1 = 1$, and $N_{batch} = 18.78027$ for $i_1 = 20$. Under the fixed value of N, the average size N_{batch} of a group taken for servicing increases with an increase in i_1. This is because, with an increase in i_1, the system tries to avoid serving small groups of requests.

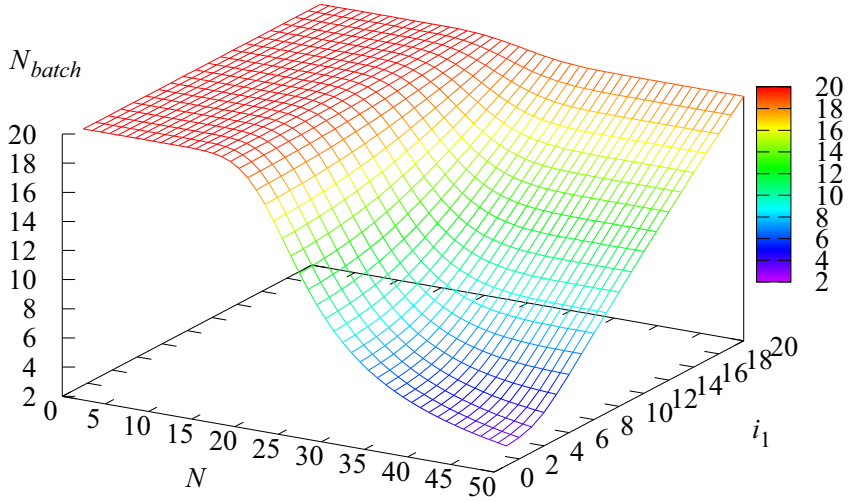

Figure 6. Dependence of the average size N_{batch} of a group of requests taken for servicing on the parameters N and i_1.

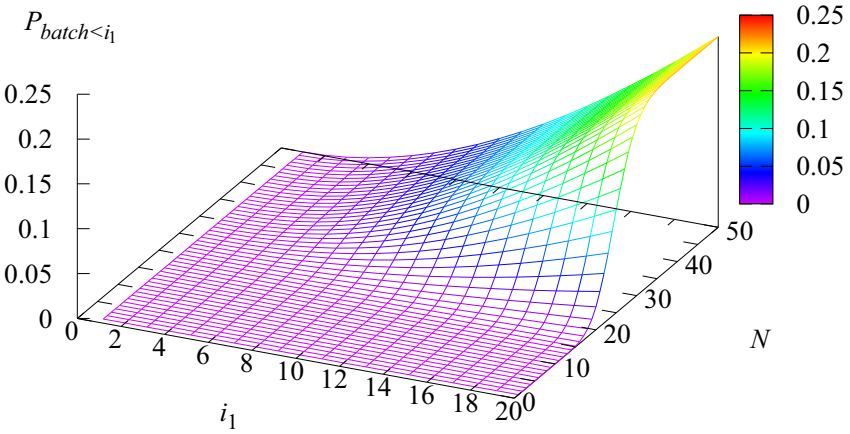

Figure 7. Dependence of the probability $P_{batch<i_1}$ that a group with a size of less than i_1 undergoes servicing on the parameters N and i_1.

The probability $P_{batch<i_1}$ that a group with a size of less than i_1 undergoes servicing, the shape of which is depicted in Figure 7, grows with increases in N and i_1. The growth of N is explained by the fact that, with an increase in N, the probability $P^{requests}_{idle-server}$ also grows (see Figure 5). Thus, when a request tries to leave the system due to impatience, a free server is more likely to be available, and the system has the possibility of providing service to a smaller group. The growth by i_1 can be explained by the same reasons and also by the fact that with an increase in i_1, the number of requests in the buffer increases, so more requests try to leave the system due to impatience.

Figures 8–10 show the dependence of the probability $P_{ent-loss}$ that an arriving request will find the buffer full and leave the system, the probability $P_{imp-loss}$ of losing a request from the buffer due to impatience, and the probability P_{loss} of losing an arbitrary request on the parameters N and i_1.

One can see from these figures that the loss probabilities essentially depend on the number of servers N. With an increase in N, all loss probabilities decrease. The impact of the parameter i_1 on the loss probabilities in the considered example is not so essential. This is mainly caused by the high capacity of the buffer $R = 300$. For example, to lose a request upon arrival, it is necessary to have a full buffer, and when the buffer is almost full, the service is provided to groups of the maximum possible size. Thus, we can mention that the probability $P_{ent-loss}$ increases slightly with an increase in the parameter i_1, since the average number of requests in the buffer grows. The dependence of the probability $P_{imp-loss}$ on i_1 under some fixed N can be non-monotonic. For example, for $N = 50$, $P_{imp-loss} = 0.0061$

for $i_1 = 1$, $P_{imp-loss} = 0.00195$ for $i_1 = 5$, and $P_{imp-loss} = 0.00667$ for $i_1 = 20$. This can be explained by the following reasoning. For small i_1, the servers are more frequently utilized, and when a request shows impatience, it is possible to have no idle service at this moment, and the request will leave the system with a probability of 1. Thus, the probability $P_{imp-loss}$ should decrease with an increase in i_1. However, on the other hand, an increase in i_1 leads to an increase in the number of requests in the buffer; thus, more requests show their impatience and leave the system. So, the probability $P_{imp-loss}$ should increase with an increase in i_1. All of these reasons explain the non-monotonic shape of the dependence of $P_{imp-loss}$ on i_1. The loss probability P_{loss} is the sum of the probabilities $P_{imp-loss}$ and $P_{ent-loss}$; thus, this probability can also behave non-monotonically with an increase in i_1. By the way, the minimal value of the loss probability is also equal to 0.00195, and this is reached for $N = 50$ and $i_1 = 5$. This proves the necessity of finding the optimal values of i_1 to improve the system's performance.

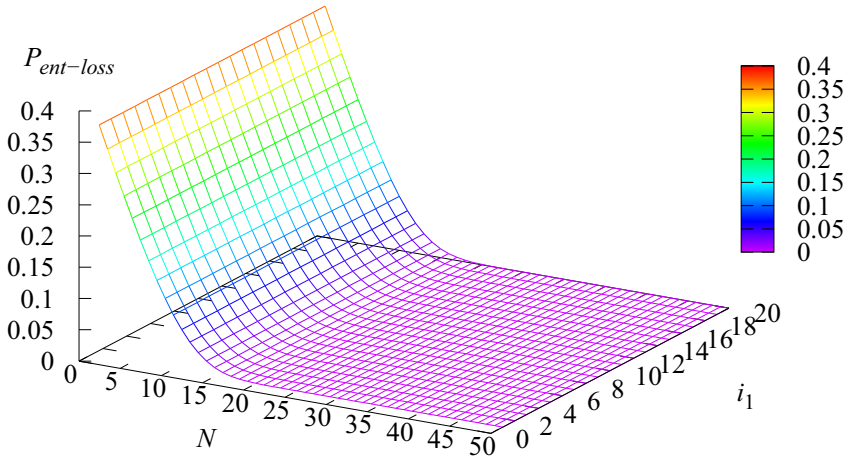

Figure 8. Dependence of the probability $P_{ent-loss}$ that an incoming request will find the buffer full and leave the system on the parameters N and i_1.

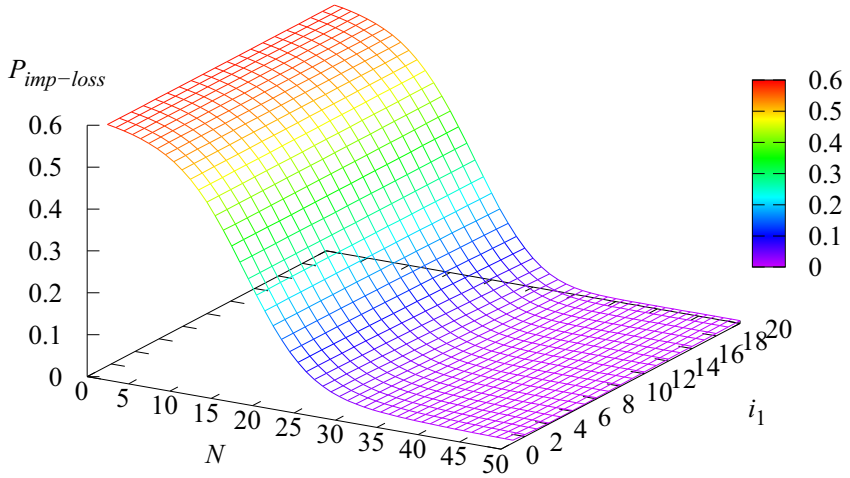

Figure 9. Dependence of the probability $P_{imp-loss}$ of losing a request from the buffer due to impatience on the parameters N and i_1.

Looking at Figure 10, we can conclude that, for large values of N, the loss probability P_{loss} becomes rather small. It is evident that the use of a server that corresponds to a vehicle in a real system costs money. It is necessary to buy vehicles, pay the costs of repairs, maintenance, and fuel, pay salaries to forwarders, etc., so it makes no sense to maintain an excessive number of vehicles. To this end, it is necessary to find the optimal number of vehicles N.

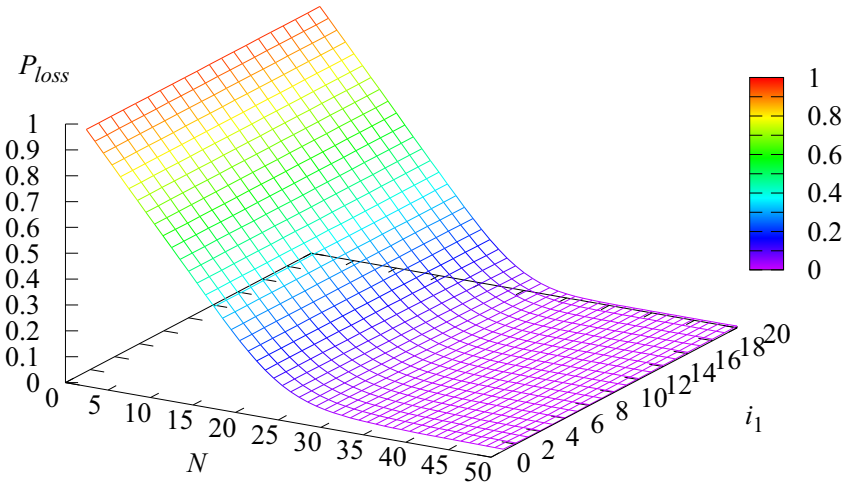

Figure 10. Dependence of the probability P_{loss} of losing an arbitrary request on the parameters N and i_1.

Let us assume that the quality of the system's operation can be determined in terms of the following cost criteria:

$$E = E(N, i_1) = a\mu_{to-serv} - c_1\lambda P_{ent-loss} - c_2\lambda P_{imp-loss} - dN.$$

Here, a is a profit for the service of one request, c_1 and c_2 are the charges for the loss of a request at the entrance to the system and due to impatience, respectively, and d is the cost for maintaining one server per unit of time. Therefore, the criterion E determines the average system's profit per unit of time, and our managerial goal is to obtain parameters such as N and i_1 under which the system's profit is maximal.

In this numerical example, let us fix the following cost coefficients:

$$a = 1,\ c_1 = 1,\ c_2 = 5,\ d = 0.02.$$

Figure 11 shows the dependence of the cost criterion E on the parameters N and i_1.

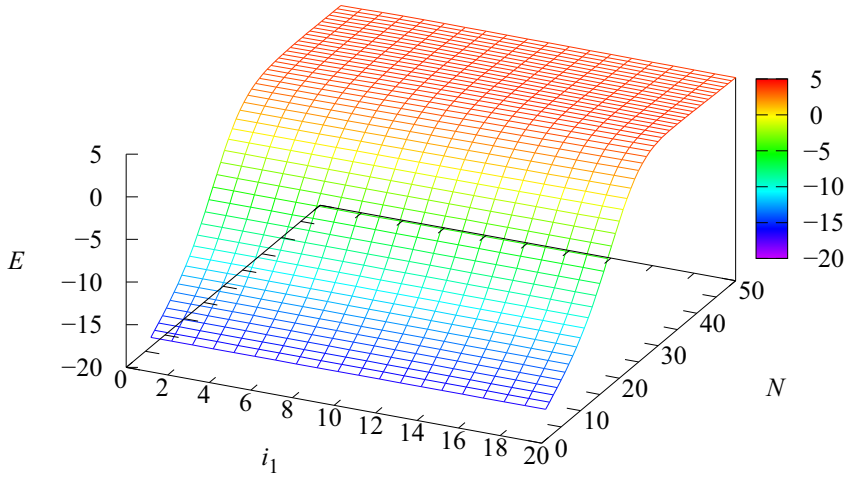

Figure 11. Dependence of the cost criterion E on the parameters N and i_1.

As we can observe from Figure 11, the system has essential losses if the number of servers N is small. With an increase in N, firstly, the system's profit grows sharply and then starts to decrease slightly. This decrease is not essential due to the fact that we chose the small cost coefficient $d = 0.02$ and the servers are cheap in the considered case. Otherwise, the decrease has to be more essential. In the considered example, the optimal value for the cost criterion is $E^*(N^*, i_1^*) = 4.1125$ and it is reached for $N^* = 36$ and $i_1^* = 12$. This means

that it is necessary to have 36 vehicles and not start a new service after the service has been completed on some server if the number of requests in the buffer is less than 12.

If we fix $N = 50$ and choose the parameter i_1, then the optimal value of the cost criterion is $E^*(50, i_1^*) = 3.94139$ and it is reached for $i_1^* = 5$. Thus, for different values of N, the optimal value of i_1 can differ. Our results can help to determine the optimal values for the control parameters.

6. Conclusions

The mathematical model of a delivery system that acts as a multi-server queuing system with group service of requests is built. The model suggests a reasonably general MAP flow for the requests and a phase-type distribution of the service time that depends on the size of the group. The possible impatience of waiting requests is taken into account. In contrast to the usual rule shown in the literature that the size of the serviced group must have values between certain fixed minimum and maximum values, it is suggested that groups of smaller size can be processed to reduce the probability of a request loss due to impatience.

The multi-dimensional $CTMC$ with a generator with the block lower Hessenberg structure describes the dynamics of the system. An effective algorithm for computing the steady-state distribution of this $CTMC$ is presented. This algorithm can be applied to the analysis of a variety of queueing models with group service of requests. Formulas for the computation of the key performance characteristics of the system were derived. The dependence of the basic performance characteristics on the number of servers and the minimum size of the group was numerically illustrated. The problems of the optimal values of the number of servers and the minimum size of the group were formulated and numerically solved.

The results can be extended in several directions, e.g., to account for the possibility of the batch arrival of the requests, the fluctuation of the number of available servers, the parameters of the service time distribution, etc. Another criterion related to the quality of the system's operation can be considered, taking into account restrictions on the values of some characteristics of the system. The maximum size of the group can also be used as a parameter with an impact on the value of the cost criterion, and the problem of its optimization can be numerically solved.

The analyzed model has essentially wider applications than the delivery systems. Possible applications in various manufacturing systems (see, e.g., [18]) and transportation networks (see, e.g., [58]), in particular, car/ride-share systems (see, e.g., [59]) look promising. The results can be used for the optimal choice of the number of servers to provide the required quality for the requested service, the maximum capacity of the servers and their delivery speed and cost, and the economically reasonable minimum size of a group.

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

Funding: This research received no external funding.

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

References

1. Dudin, S.A.; Dudin, A.N.; Dudina, O.S.; Chakravarthy, S.R. Analysis of a tandem queuing system with blocking and group service in the second node. *Int. J. Syst. Sci. Oper. Logist.* **2023**, *10*, 2235270. [CrossRef]
2. Bailey, N.T. On queueing processes with bulk service. *J. R. Stat. Soc. Ser. B (Methodol.)* **1954**, *16*, 80–87. [CrossRef]
3. Downton, F. Waiting time in bulk service queues. *J. R. Stat. Soc. Ser. B (Methodol.)* **1955**, *17*, 256–261. [CrossRef]
4. Miller, R.G., Jr. A contribution to the theory of bulk queues. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1959**, *21*, 320–337. [CrossRef]
5. Neuts, M.F. A general class of bulk queues with Poisson input. *Ann. Math. Stat.* **1967**, *38*, 759–770. [CrossRef]

6. Deb, R.; Serfozo, R. Optimal control of batch service queues. *Adv. Appl. Probab.* **1973**, 5, 340–361. [CrossRef]
7. Chaudhry, M.L.; Templeton, J.G.C. *A First Course in Bulk Queues*; John Wiley and Sons: New York, NY, USA, 1983.
8. Powell, W.B.; Humblet, P. The bulk service queue with a general control strategy: Theoretical analysis and a new computational procedure. *Oper. Res.* **1986**, *34*, 267–275. [CrossRef]
9. Sasikala, S.; Indhira, K. Bulk service queueing models-A survey. *Int. J. Pure Appl. Math.* **2016**, *106*, 43–56.
10. Niranjan, S.P.; Indhira, K. A review on classical bulk arrival and batch service queueing models. *Int. J. Pure Appl. Math.* **2016**, *106*, 45–51.
11. Brugno, A.; D'Apice, C.; Dudin, A.; Manzo, R. Analysis of an $MAP/PH/1$ queue with flexible group service. *Int. J. Appl. Math. Comput. Sci.* **2017**, *27*, 119–131. [CrossRef]
12. Chakravarthy, S.R.; Dudin, A.N. A multi-server retrial queue with BMAP arrivals and group services. *Queueing Syst.* **2002**, *42*, 5–31. [CrossRef]
13. Nakamura, A.; Phung-Duc, T. Equilibrium Analysis for Batch Service Queueing Systems with Strategic Choice of Batch Size. *Mathematics* **2023**, *11*, 3956. [CrossRef]
14. Claeys, D.; Steyaert, B.; Walraevens, J.; Laevens, K.; Bruneel, H. Analysis of a versatile batch-service queueing model with correlation in the arrival process. *Perform. Eval.* **2013**, *70*, 300–316. [CrossRef]
15. Chakravarthy, S.R. Analysis of a queueing model with MAP arrivals and heterogeneous phase-type group services. *Mathematics* **2022**, *10*, 3575. [CrossRef]
16. Banerjee, A.; Gupta, U.C.; Chakravarthy, S.R. Analysis of a finite-buffer bulk-service queue under Markovian arrival process with batch-size-dependent service. *Comput. Oper. Res.* **2015**, *60*, 138–149. [CrossRef]
17. Pradhan, S.; Gupta, U.C. Analysis of an infinite-buffer batch-size-dependent service queue with Markovian arrival process. *Ann. Oper. Res.* **2019**, *277*, 161–196. [CrossRef]
18. Chakravarthy, S.R. A finite capacity $GI/PH/1$ queue with group services. *Nav. Res. Logist. (NRL)* **1992**, *39*, 345–357. [CrossRef]
19. Baba, Y. A bulk service $GI/M/1$ queue with service rates depending on service batch size. *J. Oper. Res. Soc. Jpn.* **1996**, *39*, 25–34. [CrossRef]
20. Laxmi, P.V.; Gupta, U.C. On the finite-buffer bulk-service queue with general independent arrivals: $GI/M^{[b]}/1/N$. *Oper. Res. Lett.* **1999**, *25*, 241–245. [CrossRef]
21. Chakravarthy, S.R. Analysis of a finite $MAP/G/1$ queue with group services. *Queuing Syst. Theory Appl.* **1993**, *13*, 385–407. [CrossRef]
22. Chakravarthy, S.R. Two finite queues in series with nonrenewal input and group services. In *Seventh International Symposium on Applied Stochastic Models and Data Analysis*; 1995; pp. 78–87. Available online: https://digitalcommons.kettering.edu/industrialmanuf_eng_conference/38/ (accessed on 15 March 2012).
23. Chakravarthy, S.R.; Shruti, G.; Rumyantsev, A. Analysis of a queueing model with batch markovian arrival process and general distribution for group clearance. *Methodol. Comput. Appl. Probab.* **2021**, *23*, 1551–1579. [CrossRef] [PubMed]
24. Dudin, A.; Manzo, R.; Piscopo, R. Single server retrial queue with adaptive group admission of customers. *Comput. Oper. Res.* **2015**, *61*, 89–99. [CrossRef]
25. Brugno, A.; Dudin, A.N.; Manzo, R. Retrial queue with discipline of adaptive permanent pooling. *Appl. Math. Model.* **2017**, *50*, 1–16. [CrossRef]
26. Brugno, A.; Dudin, A.N.; Manzo, R. Analysis of a strategy of adaptive group admission of customers to single server retrial system. *J. Ambient. Intell. Humaniz. Comput.* **2018**, *9*, 123–135. [CrossRef]
27. D'Arienzo, M.P.; Dudin, A.N.; Dudin, S.A.; Manzo, R. Analysis of a retrial queue with group service of impatient customers. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 2591–2599. [CrossRef]
28. Singh, G.; Gupta, U.C.; Chaudhry, M.L. Computational analysis of bulk service queue with Markovian arrival process: $MAP/R(a,b)/1$ queue. *Opsearch* **2013**, *50*, 582–603. [CrossRef]
29. Avram, F.; Gomez-Corral, A. On bulk-service $MAP/P^{L,N}/1/N$ G-Queues with repeated attempts. *Ann. Oper. Res.* **2006**, *141*, 109–137. [CrossRef]
30. Banik, A.D. Queueing analysis and optimal control of $BMAP/G(a,b)/1/N$ and $BMAP/MSP(a,b)/1/N$ systems. *Comput. Ind. Eng.* **2009**, *57*, 748–761. [CrossRef]
31. Banik, A.D. Single server queues with a batch Markovian arrival process and bulk renewal or non-renewal service. *J. Syst. Sci. Syst. Eng.* **2015**, *24*, 337–363. [CrossRef]
32. Gupta, U.C.; Laxmi, P.V. Analysis of the $MAP/G^{a,b}/1/N$ queue. *Queueing Syst.* **2001**, *38*, 109–124. [CrossRef]
33. Chakravarthy, S.; Alfa, A.S. A multiserver queue with Markovian arrivals and group services with thresholds. *Nav. Res. Logist. (NRL)* **1993**, *40*, 811–827. [CrossRef]
34. Chakravarthy, S.R. Analysis of a multi-server queue with batch Markovian arrivals and group services. *Eng. Simul.* **2000**, *18*, 51–66.
35. Chakravarthy, S.R.; Dudin, A.N. A batch Markovian queue with a variable number of servers and group services. In *Matrix-Analytic Methods: Theory and Applications*; World Scientific Publishing Co.: Hackensack, NJ, USA, 2002; pp. 63–88.
36. Chakravarthy, S.; Alfa, A.S. A finite capacity queue with Markovian arrivals and two servers with group services. *J. Appl. Math. Stoch. Anal.* **1994**, *7*, 161–178. [CrossRef]

37. Swensen, A.R. Remaining loads in a $PH/M/c$ queue with impatient customers. *Methodol. Comput. Appl. Probab.* **2023**, *25*, 25. [CrossRef]
38. Chakravarthy, S.R. The batch Markovian arrival process: A review and future work. *Adv. Probab. Theory Stoch. Process.* **2001**, *1*, 21–49.
39. 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, NY, USA, 2022.
40. Chakravarthy, S.R. *Introduction to Matrix-Analytic Methods in Queues 2: Analytical and Simulation Approach–Queues and Simulation*; ISTE Ltd.: London, UK; John Wiley and Sons: New York, NY, USA, 2022.
41. Lucantoni, D.M. New results on the single server queue with a batch Markovian arrival process. *Commun. Stat. Stoch. Model* **1991**, *7*, 1–46. [CrossRef]
42. Dudin, A.N.; Klimenok, V.I.; Vishnevsky, V.M. *The Theory of Queuing Systems with Correlated Flows*; Springer Nature: Berlin/Heidelberg, Germany, 2020.
43. Vishnevskii, V.M.; Dudin, A.N. Queueing systems with correlated arrival flows and their applications to modeling telecommunication networks. *Autom. Remote. Control* **2017**, *78*, 1361–1403.
44. Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models*; The Johns Hopkins University Press: Baltimore, MD, USA, 1981.
45. Buchholz, P.; Kriege, J.; Felko, I. *Input Modeling with Phase-Type Distributions and Markov Models: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2014.
46. Buchholz, P. An EM-algorithm for MAP fitting from real traffic data. In *International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 218–236.
47. Kriege, J.; Buchholz, P. PH and MAP fitting with aggregated traffic traces. In *Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance*; Springer: Cham, Switzerland, 2014; pp. 1–15.
48. Horvath, A.; Telek, M. Markovian modeling of real data traffic: Heuristic phase type and MAP fitting of heavy tailed and fractal like samples. In *IFIP International Symposium on Computer Performance Modeling, Measurement and Evaluation*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 405–434.
49. Horvath, G.; Buchholz, P.; Telek, M.A. MAP fitting approach with independent approximation of the inter-arrival time distribution and the lag correlation. In Proceedings of the Second International Conference on the Quantitative Evaluation of Systems (QEST'05), Turin, Italy, 19–22 September 2005; pp. 124–133.
50. Kriege, J.; Buchholz, P. An empirical comparison of MAP fitting algorithms. In Proceedings of the International GI/ITG Conference on Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance, Essen, Germany, 15–17 March 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 259–273.
51. Buchholz, P.; Kriege, J. A heuristic approach for fitting MAPs to moments and joint moments. In Proceedings of the Sixth International Conference on the Quantitative Evaluation of Systems, Budapest, Hungary, 13–16 September 2009; pp. 53–62.
52. Buchholz, P.; Kemper, P.; Kriege, J. Multi-class Markovian arrival processes and their parameter fitting. *Perform. Eval.* **2010**, *67*, 1092–1106. [CrossRef]
53. Buchholz, P.; Panchenko, A. Two-Step EM Algorithm for MAP Fitting. *Lect. Notes Comput. Sci.* **2004**, 3280, 217–272.
54. Okamura, H.; Dohi, T. Mapfit: An R-Based Tool for PH/MAP Parameter Estimation. *Lect. Notes Comput. Sci.* **2015**, *9259*, 105–112.
55. Graham, A. *Kronecker Products and Matrix Calculus with Applications*; Ellis Horwood: Cichester, UK, 1981.
56. 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.
57. Gantmakher, F.R. *The Theory of Matrices*; Chelsea: New York, NY, USA, 1960.
58. Zharkov, M.L.; Kazakov, A.L.; Lempert, A.A. Transient process modeling in micrologistic transport systems. In *IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2021; Volume 629, p. 012023.
59. Nakamura, A.; Phung-Duc, T.; Ando, H. Queueing analysis of a Car/Ride-Share system. *Ann. Oper. Res.* **2022**, *310*, 661–682.