mathematics

MDPI

Article

A Multi-Server Heterogeneous Queuing-Inventory System with Class-Dependent Inventory Access

Karumbathil Rasmi [1,†], Machuveettil Joseph Jacob [1,*,†], Alexander S. Rumyantsev [2,3,†] and Achyutha Krishnamoorthy [4,†]

[1] Department of Mathematics, National Institute of Technology Calicut, Kozhikode 673601, Kerala, India; rasmi_p180125ma@nitc.ac.in
[2] Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences, 185910 Petrozavodsk, Russia; ar0@krc.karelia.ru
[3] Department of Applied Mathematics and Cybernetics, Petrozavodsk State University, 185035 Petrozavodsk, Russia
[4] Centre for Research in Mathematics, Department of Mathematics, CMS College, Kottayam 686001, Kerala, India; krishnamoorthy@cmscollege.ac.in
* Correspondence: mjj@nitc.ac.in
† These authors contributed equally to this work.

Abstract: In this paper, we consider a queuing inventory system with heterogeneous customers of K types arriving according to a marked Markovian arrival process. Each class of customers differs by nature of the service they seek and different priorities are assigned for each class resulting in different levels of inventory admitted to exhaust for customers of each class. A single service node is provided for each class with exponential services having class-dependent service rates. All classes of customers are served from a single source of inventory replenished according to (s, S) policy with exponentially distributed lead time. Stability condition and steady state probabilities are obtained by matrix-analytic method. Some important performance measures are also derived. Inventory recycle time was analyzed in detail. Useful cost function and numerical illustrations are also given. The optimization problem is interesting and can be solved in similar real scenario.

Keywords: MMAP[K]; heterogeneous inventory access; multi-server system; queuing-inventory; matrix analytic model

Citation: Rasmi, K.; Jacob, M.J.; Rumyantsev, A.; Krishnamoorthy, A. A Multi-Server Heterogeneous Queuing-Inventory System with Class-Dependent Inventory Access. *Mathematics* **2021**, *9*, 1037. https://doi.org/10.3390/math9091037

Academic Editor: Vladimir V. Rykov

Received: 15 April 2021
Accepted: 28 April 2021
Published: 3 May 2021

1. Introduction

Queuing-inventory systems are in the focus of recent research due to practical applicability in many fields including social, biological and technical systems. Access to a finite consumable and refillable resource is a natural way of modeling interaction with retail shop customers, office visitors, hospital patients, and even packages in the telecommunication network. In such systems, various sophisticated models arise, including the models with random demand and/or number of customers served [1,2], random order grouping [3] or duplicate ordering from several facilities [4]. In operations research, queuing-inventory system is a natural way to model load leveraging techniques, such as the leaky bucket congestion avoidance scheme used in a wide range of systems, from large scale routers [5] to electric vehicle charging stations [6].

Queuing inventory systems with positive service time were first investigated in Reference [7], followed by the work of Reference [8], in which an optimal quantity of inventory to be ordered to minimize the cost rate was obtained. In queuing inventory framework with positive service time, customers' queue is formed even when some inventory is available. We point the reader to a detailed survey of inventory systems with positive service time given in Reference [9], which includes classical, retrial, and production inventory.

Many retailers and banks find it helpful to partition the customers into different categories (classes) according to specific characteristics and adopt an inventory management

policy based on this differentiation strategy. In particular, long-term customers can be treated as high priority as compared to walk-in customers. Given a limited resource, the low priority customers may have to wait while some amount of resource is still available, reserved for customers with higher priority. The reservation may be made by imposing typical inventory levels for each class of customers according to their priority level. When the inventory comes below the level, customers from corresponding classes may have to wait until the inventory replenishment. Such a critical level policy is introduced and studied in Reference [10]. Different classes experience congestion with a single inventory; thus, customer sojourn times in a system are intrinsically correlated.

In most cases priorities are accompanied with heterogeneity of customer classes, e.g., in terms of service time distribution either in single-server [11,12] or in multi-server case [13–15]. As such, the arrival process also becomes heterogeneous, and the promising candidate is the so-called Marked Markovian Arrival Process (MMAP) used, e.g., in the works [11,14,16].

MMAP[K] is a generalization of Markov arrival processes (MAP) which have been studied and used extensively in queuing theory. MAP was introduced in Reference [17] to model non-Markovian point processes. While MAP is a useful tool to model point processes with one class of customers, MMAP[K] introduced by Neuts (see Reference [18]) is useful when multiple types of customers are present, while the model remains analytically tractable [19]. The basic characteristics of MMAP, such as peakedness of the arrival process, the first passage time to the arrival of an item of a specific type, and the behavior of the MMAP during that first passage, are analyzed in Reference [18].

In this paper, we analyze a multi-server queuing-inventory system with K classes of customers served from a single inventory which is managed according to the (s, S) with a positive lead time. Following Reference [11,14,16], we use MMAP[K] to model the arrival process. The servers are class-dependent, each server dedicated to one specific class of customers. Only the highest priority customers are allowed to wait in an infinite buffer. All other class customers can wait in respective finite buffers. The service for each class of customers is carried out with different exponential rates, and the inventory item is consumed at the end of service. A class-specific boundary level in the inventory is defined, causing customers of this specific class to wait for inventory replenishment when this boundary is down crossed. To the best of our knowledge, this model is new.

The structure of the paper is as follows. In Section 2, we give a detailed description of the model. The example of $K = 2$ is also given for better illustration of the model. In Section 3, an intuitive stability criterion for the system has been derived. Section 4 analyzes the steady state of the system and expresses a few important performance measures. In Section 5, a detailed analysis of the inventory recycle time has been carried out. Numerical illustrations are provided in Section 6, in which an optimization problem of practical importance has also been stated.

2. Model Description

We consider a multi-server queuing inventory system with heterogeneous customers. A K-server station provides service to K classes of customers. Arrivals occur according to the Marked Markovian Arrival Process (MMAP) driven by an irreducible continuous time Markov chain (CTMC) $\{Z(t)\}_{t\geq 0}$ with finite state space $\mathcal{W}$. Let $|\mathcal{W}| = W$. The sojourn time in each state $z \in \mathcal{W}$ is distributed exponentially with rate σ_z and d.f.

$$F_{z\hat{z}}(x) = 1 - e^{-\sigma_z x}.$$

At sojourn time expiration epoch, with probability $P^{(k)}(z, \hat{z})$ the process $Z(t)$ moves from the state z to $\hat{z}$, generating a class k customer arrival, $z, \hat{z} \in \mathcal{W}$, $k = 0, 1, ..., K$ (conventionally, no customer arrival is generated if $k = 0$). Note that, for any $z \in \mathcal{W}$,

$$\sum_{k=0}^{K} \sum_{\hat{z}\in\mathcal{W}} P^{(k)}(z, \hat{z}) = 1. \tag{1}$$

Thus, the MMAP is characterized by a set of $K+1$ *transition rate* square matrices $\boldsymbol{D}_0, \ldots, \boldsymbol{D}_K$ of order W defined as follows:

$$(\boldsymbol{D}_k)_{z\hat{z}} = \begin{cases} \sigma_z P^{(k)}(z,\hat{z}), & z \neq \hat{z}, \\ -\sigma_z 1_{k=0}, & z = \hat{z}, \end{cases} \quad z, \hat{z} \in \mathcal{W}, \;\; k = 0, 1, \ldots, K, \tag{2}$$

where $1_{k=0}$ is the indicator of a non-random condition $k = 0$. The matrices $\boldsymbol{D}_k$ constitute a *generator* matrix $\boldsymbol{D}$ of the Markov process $\{Z(t)\}_{t\geq 0}$ such that

$$\boldsymbol{D} = \sum_{k=0}^{K} \boldsymbol{D}_k.$$

Recall that $\boldsymbol{D}\mathbf{e} = \mathbf{0}$, which also follows from (1) and (2). Hereinafter, $\mathbf{e}$ ($\mathbf{0}$, respectively) is the vector of ones (zeroes), and, if necessary, we designate the row vector with a transpose sign, e.g., $\mathbf{0}'$. As such, denoting by $\boldsymbol{\theta}$ the *stochastic* invariant vector of $\boldsymbol{D}$ (i.e., $\boldsymbol{\theta}\boldsymbol{D} = \mathbf{0}$, that is, $\boldsymbol{\theta}$ is the distribution of the corresponding Markov jump process governed by $\boldsymbol{D}$), the class k arrival rate, λ_k, equals

$$\lambda_k = \boldsymbol{\theta}\boldsymbol{D}_k\mathbf{e}.$$

We would like to note that the MMAP process being a versatile Markov process, is suitable for a wide range of applications due to variety of modeling features including correlated arrivals [20]. However, capturing sophisticated features, such as Long-Range Dependence (slow autocorrelation decay of the process that complicates application of the standard methods of performance estimation), may require infinite number of sources [21]. As such, it is necessary to balance the size W of the MMAP state space with practical capabilities of the model, including computational/storage capacity. Useful examples of MMAP processes may be found in Reference [19].

If the server k is busy, then the arriving class k customer can wait in a buffer space of capacity l_k, where $l_1 = \infty$ and $l_k < \infty$, $k = 2, \ldots, K$. The class k customer, finding the respective buffer completely filled on arrival, leaves the system forever.

The inventory is divided into K segments,

$$s_0 = 0 < s = s_1 < s_2 < \cdots < s_K = S.$$

Class k customer spends an exponentially distributed, with rate μ_k, time at the (class-specific) server $k = 1, \ldots, K$. At *service completion* epoch, one item from a single common inventory is consumed by the customer. However, class k customers are served only if the inventory level exceeds s_{k-1}, $k = 1, \ldots, K$. Thus, s is not only inventory replenishment boundary, but also the critical level, at or below which only the class 1 customers (highest priority) are served. If the inventory level at service completion epoch is insufficient, the corresponding customer of class k repeats an independent service time with the same rate, μ_k, *until a service completion coincides with a sufficient inventory level*.

The inventory is replenished under (s, S) policy in an exponentially distributed *lead* time with rate γ. That is, when the inventory hits level s, a request for replenishment is issued, and the inventory level S is restored after an independent exponentially distributed random time with rate γ. Finally, we note that the arrival process, service times and replenishment times are *independent*. The structure of the system is presented on Figure 1.

Let $N_k(t)$ be the number of class k customers in the system, $k = 1, \ldots, K$, $I(t)$ be the inventory level, $0 \leq I(t) \leq S$, and $Z(t) \in \mathcal{W}$ be the phase of the MMAP, at time $t \geq 0$. Then, the considered system can be modeled by the regular irreducible CTMC

$$\zeta(t) = \{N_1(t), \ldots, N_K(t), I(t), Z(t)\}, t \geq 0,$$

with state space $(n_1, n_2, \ldots, n_K, i, z), n_k \in \{0, \ldots, l_k\}, k = 1, \ldots, K, i \in \{0, \ldots, S\}$ and $z \in \mathcal{W}$.

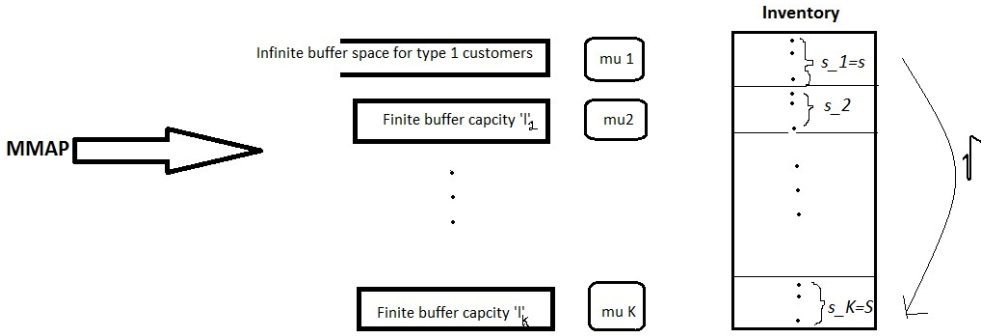

Figure 1. Structure of the system.

Fix the process $\zeta(t)$ at some state $(n_1, n_2, \ldots, n_k, i, z)$. Due to independence of the components of $\zeta(t)$, the transitions are now possible only to $2K+2$ states enumerated below:

- $(n_1, n_2, \ldots, n_j + 1_{n_j<l_j}, \ldots, n_K, i, \hat{z})$ with rate $(\boldsymbol{D}_j)_{z\hat{z}}$, $j = 1, \ldots, K$ (arrival of class j customer);
- $(n_1, n_2, \ldots, n_j - 1, \ldots, n_k, i-1, z)$ with rate μ_j, if $n_j > 0$ and $i > s_{j-1}$, $j = 1, \ldots, K$ (departure of class j customer);
- $(n_1, \ldots, n_K, i, \hat{z})$ with rate $(\boldsymbol{D}_0)_{z\hat{z}}$ (MMAP phase switch);
- $(n_1, \ldots, n_K, S, \hat{z})$ with rate γ, if $i \leq s$ (inventory replenishment).

This gives a very special structure of the infinitesimal generator matrix of the considered process $\{\zeta(t)\}_{t\geq 0}$. First, since the component $N_1(t)$ corresponds to the unbounded buffer, while other components of $\zeta(t)$ are finite, and due to the fact that the transitions of $N_1(t)$ are *skip-free* in both directions (i.e., it may be incremented/decremented by at most one), the process $\zeta(t)$ is the so-called Quasi-Birth-Death (QBD) process with *level* $N_1(t)$ and *phase* $(N_2(t), \ldots, N_K(t), I(t), Z(t))$. Lexicographically ordered state space allows one to write the generator matrix in block-tridiagonal form

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{A}_1^{(0)} & \boldsymbol{A}_0 & \mathbb{O} & \mathbb{O} & \ldots \\ \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \mathbb{O} & \ldots \\ \mathbb{O} & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \ldots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{3}$$

where the matrix $\mathbb{O}$ is a zero block of corresponding dimension (we give the dimension explicitly if and when necessary).

Define an integer-valued function of two arguments, i, j, such that $2 \leq i \leq j \leq K$:

$$\alpha(i,j) = \prod_{k=i}^{j} (l_k + 1), \tag{4}$$

and define $\alpha(i,j) = 0$ otherwise. Then, the blocks $\boldsymbol{A}_i, i = 0, 1, 2$, and $\boldsymbol{A}_1^{(0)}$ are square matrices having size

$$\alpha(2,K)(S+1)W.$$

The matrix $\boldsymbol{A}_0$ consists of the rates of transitions corresponding to the arrival of class 1 customer, $\boldsymbol{A}_2$ keeps the transition rates corresponding to departure of a class 1 customer, while $\boldsymbol{A}_1$ is related to $2K$ remaining possible transitions, such that the *level* remains unaffected. It is quite straightforward to define $\boldsymbol{A}_0$, since, upon arrival of a class 1 customer, only the MMAP phase is switched; thus,

$$\boldsymbol{A}_0 = \boldsymbol{I}_{\alpha(2,K)(S+1)} \otimes \boldsymbol{D}_1, \tag{5}$$

where $\otimes$ is the Kronecker product and $\boldsymbol{I}$ is the identity matrix of corresponding dimension.

We will need the following notation: hereafter $\mathbf{e}^k_{i:j}$ is the (column) vector of dimension $k \geq 0$ with ith to jth components equal 1, and zero otherwise, $i \leq j \leq k$ (conventionally we take $\mathbf{e}^k_{i:j} \equiv 1$ if $k = 0$). To shorten the notation, we use for any $i \leq k$

- $\mathbf{e}^k_i \equiv \mathbf{e}^k_{i:i}$ (a single non-zero component at ith row, dimension k);
- $\mathbf{e}_k \equiv \mathbf{e}^k_{k:k}$ (last non-zero component, dimension k);
- $\mathbf{e}^k \equiv \mathbf{e}^k_{1:k}$ (all non-zero components of dimension k).

To define $\boldsymbol{A}_2$ and for further use, we need to construct some auxiliary matrices. Define for $j = 1, \ldots, K$ the following square matrix of order $S + 1$:

$$\boldsymbol{L}_j = \begin{bmatrix} \mathbf{0}' & 0 \\ \operatorname{diag}(\mathbf{e}^S_{s_{j-1}+1,S}) & \mathbf{0} \end{bmatrix}, \tag{6}$$

where the matrix $\boldsymbol{L}_j$ has non-zero entries below main diagonal only for rows $s_{j-1} + 2, \ldots, S + 1$, and the zero vectors are of size S. The block matrix $\boldsymbol{L}_j$ corresponds to possible transitions of the inventory for class j customer, and is indexed from 0 to S. In particular, for class 1 customers the matrix $\boldsymbol{L}_1$ has the lower diagonal of ones (there is no constraint for the inventory to be decremented). Since, upon departure of class 1 customer, the *level*, as well as the inventory, are decremented by one, the matrix $\boldsymbol{A}_2$ has the following form:

$$\boldsymbol{A}_2 = \boldsymbol{I}_{\alpha(2,K)} \otimes \mu_1 \boldsymbol{L}_1 \otimes \boldsymbol{I}_W. \tag{7}$$

Now, to define $\boldsymbol{A}_1$, which contains the transition rates related to arrivals and departures of class j customers, $j \geq 2$, inventory replenishment and MMAP phase change, we need auxiliary matrices of an increment/decrement in the corresponding (phase) component:

$$\boldsymbol{N}^{(+)}_j = \begin{bmatrix} \mathbf{0} & \boldsymbol{I}_{l_j} \\ 0 & \mathbf{e}'_{l_j} \end{bmatrix}, \quad \boldsymbol{N}^{(-)}_j = \begin{bmatrix} \mathbf{0}' & 0 \\ \boldsymbol{I}_{l_j} & \mathbf{0} \end{bmatrix}. \tag{8}$$

Note that $\boldsymbol{N}^{(-)}_j$ is a square matrix of order $l_j + 1$ having lower diagonal of ones, while $\boldsymbol{N}^{(+)}_j$ is semi-upper diagonal. This asymmetry will be explained below.

Using these constructions, it is rather straightforward to define the transition rates of all possible transitions constituting the matrix $\boldsymbol{A}_1$:

$$\boldsymbol{A}_1 = \boldsymbol{A}^{(a)}_1 + \boldsymbol{A}^{(b)}_1 + \boldsymbol{A}^{(c)}_1 - \boldsymbol{\Delta}, \text{ where} \tag{9}$$

$$\boldsymbol{A}^{(a)}_1 = \sum_{j=2}^{K} \boldsymbol{I}_{\alpha(2,j-1)} \otimes \Big(\boldsymbol{N}^{(+)}_j \otimes \boldsymbol{I}_{\alpha(j+1,K)} \otimes \boldsymbol{I}_{S+1} \otimes \boldsymbol{D}_j \tag{10}$$

$$+ \boldsymbol{N}^{(-)}_j \otimes \boldsymbol{I}_{\alpha(j+1,K)} \otimes \mu_j \boldsymbol{L}_j \otimes \boldsymbol{I}_W \Big), \tag{11}$$

$$\boldsymbol{A}^{(b)}_1 = \boldsymbol{I}_{\alpha(2,K)} \otimes \gamma \mathbf{e}^{S+1}_{1:s+1} \mathbf{e}'_{S+1} \otimes \boldsymbol{I}_W, \tag{12}$$

$$\boldsymbol{A}^{(c)}_1 = \boldsymbol{I}_{\alpha(2,K)} \otimes \boldsymbol{I}_{S+1} \otimes \boldsymbol{D}_0. \tag{13}$$

Conventionally, in (10) and (11), we define the zero-size identity matrix $\boldsymbol{I}_0 = 1$. Note that (10) corresponds to arrival of class j customer, (11) is the corresponding class departure, (12) is a replenishment (where in fact the corresponding vector product gives a matrix with only last column being non-zero), while (13) is the MMAP phase change. It is worth noting that asymmetry in $\boldsymbol{N}^{(+)}_j$ (non-zero last row), defined in (8), is used in (10) to indicate an arrival of a class j customer that is lost, while the MMAP phase is changed according to $\boldsymbol{D}_j$. The matrix $\boldsymbol{\Delta}$ is a diagonal matrix that guarantees $(\boldsymbol{A}_0 + \boldsymbol{A}_1 + \boldsymbol{A}_2)\mathbf{e} = \mathbf{0}$; thus,

$$\boldsymbol{\Delta} = \operatorname{diag}\Big[\Big(\boldsymbol{A}_0 + \boldsymbol{A}^{(a)}_1 + \boldsymbol{A}^{(b)}_1 + \boldsymbol{A}^{(c)}_1 + \boldsymbol{A}_2 \Big) \mathbf{e} \Big]. \tag{14}$$

Straightforward algebra allows to obtain $\mathbf{\Delta}$, the diagonal matrix of dimension $\alpha(2,K)(S+1)W$, from (14) in a closed form as follows:

$$\mathbf{\Delta} = \text{diag}\left[\boldsymbol{\delta}^{(a)} + \boldsymbol{\delta}^{(b)} + \boldsymbol{\delta}^{(d)}\right],$$

where $\boldsymbol{\delta}^{(a)}$ is the vector of transition rates due to departure of customers of classes $2,\ldots,K$, $\boldsymbol{\delta}^{(d)}$ contains transition rates due to a class 1 customer departure, and $\boldsymbol{\delta}^{(b)}$ is the vector of transition rates due to replenishment given as follows:

$$\boldsymbol{\delta}^{(a)} = \sum_{j=2}^{K} \mu_j \mathbf{e}^{\alpha(2,j-1)} \otimes \mathbf{e}^{l_j+1}_{2:l_j+1} \otimes \mathbf{e}^{\alpha(j+1,K)} \otimes \mathbf{e}^{S+1}_{s_{j-1}+2:S+1} \otimes \mathbf{e}^{W}, \tag{15}$$

$$\boldsymbol{\delta}^{(b)} = \gamma \mathbf{e}^{\alpha(2,K)} \otimes \mathbf{e}^{S+1}_{1:s+1} \otimes \mathbf{e}^{W}, \tag{16}$$

$$\boldsymbol{\delta}^{(d)} = \mu_1 \mathbf{e}^{\alpha(2,K)} \otimes \mathbf{e}^{S+1}_{2:S+1} \otimes \mathbf{e}^{W}. \tag{17}$$

It remains to define the block $\boldsymbol{A}_1^{(0)}$ corresponding to possible transitions of the model from within the states having zero class 1 customers. Note that such a matrix is very similar to the matrix $\boldsymbol{A}_1$, and the only difference is in the diagonal balancing matrix (14). Indeed, from the condition $(\boldsymbol{A}_1^{(0)} + \boldsymbol{A}_0)\mathbf{e} = \mathbf{0}$, define

$$\mathbf{\Delta}_0 = \text{diag}\left[\left(\boldsymbol{A}_0 + \boldsymbol{A}_1^{(a)} + \boldsymbol{A}_1^{(b)} + \boldsymbol{A}_1^{(c)}\right)\mathbf{e}\right]. \tag{18}$$

The matrix $\boldsymbol{A}_1^{(0)}$ is then defined as follows:

$$\boldsymbol{A}_1^{(0)} = \boldsymbol{A}_1^{(a)} + \boldsymbol{A}_1^{(b)} + \boldsymbol{A}_1^{(c)} - \mathbf{\Delta}_0, \tag{19}$$

where it follows from (14) and (18) that $\mathbf{\Delta}_0 = \text{diag}\left[\boldsymbol{\delta}^{(a)} + \boldsymbol{\delta}^{(b)}\right]$.

We note that the definitions of subblocks may be rewritten in a more compact form using Kronecker sums. Moreover, by defining unbounded analogs of the matrices given in (8), the generator matrix itself may be rewritten similarly. However, we skip this possibility to keep parsimony of the notation.

To illustrate (3), we consider the case $K = 2$. Recall that class 1 customers have priority over class 2 customers. Class 1 customers are allowed to wait in an infinite buffer, whereas class 2 customers cannot enter the system if there are l_2 class 2 customers already in the system. Server 1 serves class 1 customers with rate μ_1 and server 2 serves class 2 customers with rate μ_2. Customers are served with an inventory from a common source running according to (s,S) policy with exponential lead time (rate γ). Even if the server 2 is free, class 2 customers will be served only if the inventory level is at least $s+1$.

We consider the CTMC $\zeta_t = \{N_1(t), N_2(t), I(t), Z(t)\}, t \geq 0$ with state space (n_1, n_2, i, z), where $n_1 \geq 0, 0 \leq n_2 \leq l_2, 0 \leq i \leq S, z \in \mathcal{W}$. Since $K = 2$, it follows from (4) that $\alpha(2,K) = l_2 + 1$. Corresponding to each *level* n_1, there will be $(l_2+1)(S+1)W$ *phase states*. The infinitesimal generator matrix of ζ_t is given by (3), where the non zero blocks $\boldsymbol{A}_1^{(0)}, \boldsymbol{A}_0, \boldsymbol{A}_1$, and $\boldsymbol{A}_2$ are of size $(l_2+1)(S+1)W$ and have the following forms:

$$\boldsymbol{A}_0 = \boldsymbol{I}_{(l_2+1)(S+1)} \otimes \boldsymbol{D}_1,$$
$$\boldsymbol{A}_2 = \boldsymbol{I}_{l_2+1} \otimes \mu_1 \boldsymbol{L}_1 \otimes \boldsymbol{I}_W$$

where $\boldsymbol{L}_1$ is a matrix of order $S+1$ with a non-zero lower diagonal, defined as

$$\boldsymbol{L}_1 = \begin{bmatrix} \mathbf{0}' & 0 \\ \boldsymbol{I}_S & \mathbf{0} \end{bmatrix}.$$

The subblocks $\boldsymbol{A}_1, \boldsymbol{A}_1^{(0)}$ also have the block-tridiagonal structure with l_2+1 blocks over the main diagonal indexed by the number of class 2 customers:

$$\boldsymbol{A}_1^{(0)}=\begin{bmatrix} \boldsymbol{A}_1^{(2)} & \boldsymbol{A}_0^{(1)} & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} \\ \boldsymbol{A}_2^{(1)} & \boldsymbol{A}_1^{(1)} & \boldsymbol{A}_0^{(1)} & \mathbb{O} & \dots & \mathbb{O} \\ \mathbb{O} & \boldsymbol{A}_2^{(1)} & \boldsymbol{A}_1^{(1)} & \boldsymbol{A}_0^{(1)} & \ddots & \mathbb{O} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \dots & \dots & \ddots & \boldsymbol{A}_1^{(1)} & \boldsymbol{A}_0^{(1)} \\ \vdots & \dots & \dots & \dots & \boldsymbol{A}_2^{(1)} & \widehat{\boldsymbol{A}}_1^{(1)} \end{bmatrix}, \quad \boldsymbol{A}_1=\begin{bmatrix} \boldsymbol{A}_1^{(4)} & \boldsymbol{A}_0^{(1)} & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} \\ \boldsymbol{A}_2^{(1)} & \boldsymbol{A}_1^{(3)} & \boldsymbol{A}_0^{(1)} & \mathbb{O} & \dots & \mathbb{O} \\ \mathbb{O} & \boldsymbol{A}_2^{(1)} & \boldsymbol{A}_1^{(3)} & \boldsymbol{A}_0^{(1)} & \ddots & \mathbb{O} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \dots & \dots & \ddots & \boldsymbol{A}_1^{(3)} & \boldsymbol{A}_0^{(1)} \\ \vdots & \dots & \dots & \dots & \boldsymbol{A}_2^{(1)} & \widehat{\boldsymbol{A}}_1^{(3)} \end{bmatrix}.$$

Submatrices $\boldsymbol{A}_0^{(1)}, \boldsymbol{A}_2^{(1)}, \boldsymbol{A}_1^{(i)}, i=1,\dots,4$, of order $(S+1)W$, are given below.

$$\boldsymbol{A}_0^{(1)} = \boldsymbol{I}_{S+1} \otimes \boldsymbol{D}_2,$$
$$\boldsymbol{A}_2^{(1)} = \mu_2 \boldsymbol{L}_2 \otimes \boldsymbol{I}_W,$$

where $\boldsymbol{L}_2$ has non-zero entries in the lower diagonal from $s+2$nd row onward,

$$\boldsymbol{L}_2 = \begin{bmatrix} 0 & 0 & \dots & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 & 0 \end{bmatrix}.$$

It remains to define $\boldsymbol{A}_1^{(i)}, i=1,\dots,4$. These matrices have semi-block-diagonal structure with main diagonal and last column containing $S+1$, possibly non-zero, blocks, indexed by the number of inventory items available, each block being a square matrix of order W. Indeed, since the departures of class 1 customers are not possible from *level* 0,

$$\boldsymbol{A}_1^{(2)} = \boldsymbol{I}_{S+1} \otimes \boldsymbol{D}_0 - \begin{bmatrix} \gamma \boldsymbol{I}_W & \mathbb{O} & \dots & \mathbb{O} & \dots & -\gamma \boldsymbol{I}_W \\ \vdots & \ddots & \dots & \dots & \dots & \vdots \\ \mathbb{O} & \dots & \gamma \boldsymbol{I}_W & \mathbb{O} & \dots & -\gamma \boldsymbol{I}_W \\ \mathbb{O} & \dots & \mathbb{O} & \mathbb{O} & \dots & \mathbb{O} \\ \vdots & \dots & \dots & \dots & \dots & \vdots \\ \mathbb{O} & \dots & \dots & \mathbb{O} & \dots & \mathbb{O} \end{bmatrix},$$

where, from $s+1$st row onward, the diagonal elements are $\boldsymbol{D}_0$ (only MMAP phase change is possible). Similarly, since $\boldsymbol{A}_1^{(1)}$ corresponds to states with positive number of class 2 customers,

$$\boldsymbol{A}_1^{(1)} = \boldsymbol{A}_1^{(2)} - \begin{bmatrix} \mathbb{O} & \dots & \mathbb{O} & \dots & \mathbb{O} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \mathbb{O} & \dots & \mu_2 \boldsymbol{I}_W & \dots & \mathbb{O} \\ \vdots & \dots & \vdots & \ddots & \vdots \\ \mathbb{O} & \dots & \mathbb{O} & \dots & \mu_2 \boldsymbol{I}_W \end{bmatrix}.$$

Since $\boldsymbol{A}_1$ corresponds to positive levels of the process ζ_t, i.e., the number of class 1 customers is positive,

$$\boldsymbol{A}_1^{(4)} = \boldsymbol{A}_1^{(2)} - \boldsymbol{I}_{S+1}^{(0)} \otimes \mu_1 \boldsymbol{I}_W, \quad \boldsymbol{A}_1^{(3)} = \boldsymbol{A}_1^{(1)} - \boldsymbol{I}_{S+1}^{(0)} \otimes \mu_1 \boldsymbol{I}_W,$$

where $\boldsymbol{I}_{S+1}^{(0)}$ is a square matrix of order $S+1$ consisting of a zero column, zero row, and identity matrix as follows:

$$\boldsymbol{I}_{S+1}^{(0)} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_S \end{bmatrix}.$$

Indeed, the matrices $\boldsymbol{A}_1^{(4)}$ and $\boldsymbol{A}_1^{(3)}$ correspond to positive number of class 1 customers; thus, their diagonals include the service rates of class 1 customers, except the case of the empty inventory. As a final note, the last diagonal blocks in matrices, $\widehat{\boldsymbol{A}}_1^{(1)}$ and $\widehat{\boldsymbol{A}}_1^{(3)}$, correspond to arrivals of class 2 customers that are lost; hence,

$$\widehat{\boldsymbol{A}}_1^{(i)} = \boldsymbol{A}_1^{(i)} + \boldsymbol{A}_0^{(1)}, \quad i = 1, 3.$$

3. Stability Condition

The necessary and sufficient condition for existence of the non-zero steady-state probability is the specific version of Foster ergodicity condition known as the Neuts ergodicity criterion [22],

$$\boldsymbol{\pi} \boldsymbol{A}_0 \mathbf{e} < \boldsymbol{\pi} \boldsymbol{A}_2 \mathbf{e}, \tag{20}$$

where the stochastic vector $\boldsymbol{\pi}$ is the solution of the system

$$\boldsymbol{\pi} \boldsymbol{A} = \mathbf{0}, \tag{21}$$

and

$$\boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{A}_1 + \boldsymbol{A}_2.$$

Note that, due to the properties of $\boldsymbol{A}_i$, $i = 0, 1, 2$, the matrix $\boldsymbol{A}$ is a generator matrix of a finite state space CTMC giving the projection of the phase transition at high levels; thus, the vector $\boldsymbol{\pi}$ may be considered as the steady-state probability of the phase at high levels [22]. It now follows from (5), (7), and (9) that $\boldsymbol{A}$ has block-tridiagonal structure indexed by the number of class-2 customers as follows:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{R}_1^{(0)} & \boldsymbol{R}_0 & 0 & 0 & \dots & 0 \\ \boldsymbol{R}_2 & \boldsymbol{R}_1 & \boldsymbol{R}_0 & 0 & \dots & 0 \\ 0 & \boldsymbol{R}_2 & \boldsymbol{R}_1 & \boldsymbol{R}_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{R}_2 & \boldsymbol{R}_1 & \boldsymbol{R}_0 \\ 0 & 0 & \dots & 0 & \boldsymbol{R}_2 & \widehat{\boldsymbol{R}}_1 \end{bmatrix}.$$

The blocks $\boldsymbol{R}_1^{(0)}$, $\widehat{\boldsymbol{R}}_1$, and $\boldsymbol{R}_i$, $i = 0, 1, 2$, are square matrices, and their structure follows from (5), (7), and (9). Indeed,

$$\begin{aligned} \boldsymbol{R}_0 =& \boldsymbol{I}_{\alpha(3,K)} \otimes \boldsymbol{I}_{S+1} \otimes \boldsymbol{D}_2, \\ \boldsymbol{R}_2 =& \boldsymbol{I}_{\alpha(3,K)} \otimes \mu_2 \boldsymbol{L}_2 \otimes \boldsymbol{I}_W, \\ \boldsymbol{R}_1 =& \boldsymbol{I}_{\alpha(3,K)} \otimes \left(\left(\gamma \mathbf{e}_{1:s+1}^{S+1} \boldsymbol{e}'_{S+1} + \mu_1 \boldsymbol{L}_1 \right) \oplus (\boldsymbol{D}_0 + \boldsymbol{D}_1) \right) \\ &+ \sum_{j=3}^{K} \boldsymbol{I}_{\alpha(3,j-1)} \otimes \Big(\boldsymbol{N}_j^{(+)} \otimes \boldsymbol{I}_{\alpha(j+1,K)} \otimes \boldsymbol{I}_{S+1} \otimes \boldsymbol{D}_j \\ &\qquad + \boldsymbol{N}_j^{(-)} \otimes \boldsymbol{I}_{\alpha(j+1,K)} \otimes \mu_j \boldsymbol{L}_j \otimes \boldsymbol{I}_W \Big) - \hat{\boldsymbol{\Delta}}, \end{aligned}$$

where $\oplus$ is the Kronecker sum defined as $\boldsymbol{A} \oplus \boldsymbol{B} = \boldsymbol{I} \otimes \boldsymbol{B} + \boldsymbol{A} \otimes \boldsymbol{I}$; $\widehat{\boldsymbol{\Delta}}$ being a diagonal matrix that guarantees $(\boldsymbol{R}_0 + \boldsymbol{R}_1 + \boldsymbol{R}_2)\mathbf{e} = \mathbf{0}$, and obvious convention $\sum_{j=3}^{K} = 0$ for $K < 3$ is used. It remains to note that, since $\boldsymbol{R}_1^{(0)}$ corresponds to states with no class-2

customers, $\boldsymbol{R}_1^{(0)} = \boldsymbol{R}_1 + \mathrm{diag}[\boldsymbol{R}_2\mathbf{e}]$, while the matrix $\widehat{\boldsymbol{R}}_1 = \boldsymbol{R}_1 + \boldsymbol{R}_0$ includes the lost arrivals of class-2 customers. Recall that the losses of class 2 customers appear here due to the specific indexing of the blocks of matrix $\boldsymbol{A}$, and corresponding losses of classes $3, \ldots, K$ are encoded in the components of matrix $\boldsymbol{R}_1$. It is worthwhile to note that, despite the expected independence of the MMAP component evolution on the system state, the changes in the phase occur upon arrival of customers of classes $2, \ldots, K$, both in the corresponding queue size and in MMAP phase; similar simultaneous changes occur upon departure of such customers (both counter and inventory). In the case $K = 2$, the subblocks of $\boldsymbol{A}$ have the following clear structure:

$$\begin{aligned}
&\boldsymbol{R}_0 = \mathbb{O} \oplus \boldsymbol{D}_2, \quad \boldsymbol{R}_2 = \mu_2 \boldsymbol{L}_2 \oplus \mathbb{O},\\
&\boldsymbol{R}_1 = \left(\gamma \mathbf{e}_{1:s+1}^{S+1} \mathbf{e}'_{S+1} + \mu_1 \boldsymbol{L}_1\right) \oplus (\boldsymbol{D}_0 + \boldsymbol{D}_1) - \hat{\boldsymbol{\Delta}},\\
&\widehat{\boldsymbol{R}}_1 = \left(\gamma \mathbf{e}_{1:s+1}^{S+1} \mathbf{e}'_{S+1} + \mu_1 \boldsymbol{L}_1\right) \oplus \boldsymbol{D} - \hat{\boldsymbol{\Delta}}.
\end{aligned}$$

Note that the Kronecker sums in the phase transition rate matrices appear due to the multidimensional structure of the phase state space. These sums highlight the independent changes of one of the components of the phase vector, and $\mathbb{O}$ is used if the corresponding component remains unchanged due to a transition.

To solve the system (21), the following numerically stable algorithm is suggested. Let the row vector $\boldsymbol{\pi}$ be presented in the form $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{l_2})$. The vectors $\boldsymbol{\pi}_m, m = 0, \ldots, l_2$, are considered to have the following form:

$$\boldsymbol{\pi}_m = \boldsymbol{\pi}_{m-1}\boldsymbol{U}_{m-1} = \boldsymbol{\pi}_0 \prod_{k=1}^{m} \boldsymbol{U}_{k-1}, \quad m = 1, \ldots, l. \tag{22}$$

The matrices $\boldsymbol{U}_m, m = 0, \ldots, l_2 - 1$, are obtained from the system (21) using (22), starting from the last column, i.e., calculated using the backward recursion,

$$\boldsymbol{U}_m = -\boldsymbol{R}_0(\boldsymbol{R}_1 + \boldsymbol{U}_{m+1}\boldsymbol{R}_2)^{-1}, \quad m = l_2 - 2, \ldots, 0, \tag{23}$$

under the initial condition (following from the last column of the matrix $\boldsymbol{A}$)

$$\boldsymbol{U}_{l_2-1} = -\boldsymbol{R}_0\left(\widehat{\boldsymbol{R}}_1\right)^{-1}.$$

Note that, since $\boldsymbol{A}$ is a generator matrix, $\widehat{\boldsymbol{R}}_1$ is diagonally dominant and, hence, invertible. Finally, the vector $\boldsymbol{\pi}_0$ is the unique solution to the system (obtained from the first column of $\boldsymbol{A}$ and the fact that $\boldsymbol{\pi}$ is stochastic)

$$\begin{cases} \boldsymbol{\pi}_0\left(\boldsymbol{R}_1^{(0)} + \boldsymbol{U}_0\boldsymbol{R}_2\right) = \mathbf{0}, \\ \boldsymbol{\pi}_0\left(\mathbf{e} + \sum_{m=1}^{l_2} \prod_{k=1}^{m} \boldsymbol{U}_{k-1}\mathbf{e}\right) = 1. \end{cases} \tag{24}$$

Now, let us consider the l.h.s. of (20). The vector $\boldsymbol{\pi}$ is the steady-state probability vector of the finite state space CTMC defined by the matrix $\boldsymbol{A}$ and having the state space

$$\mathcal{E} = \{(n_2, \ldots, n_K, i, z), n_k \in \{0, \ldots, l_k\}, k = 2, \ldots, K, i \in \{0, \ldots, S\}, z \in \mathcal{W}\}. \tag{25}$$

However, it is possible to shrink the state space $\mathcal{E}$ into the following subsets defined for all $z \in \mathcal{W}$:

$$\mathcal{E}_z = \{(n_2, \ldots, n_K, i, z_0) \in \mathcal{E} : z_0 = z\}.$$

Now, we show that, for each state from $\mathcal{E}_z$, the transition rate to the set $\mathcal{E}_{\hat{z}}$ (i.e., the sum of corresponding transition rates to individual states) equals $(\boldsymbol{D})_{z\hat{z}}$. Indeed, to obtain this transtion rate, the matrix $\boldsymbol{A}$ needs to be multiplied by the matrix that would sum up

the corresponding transition rates, that is, $\mathbf{e}^{\alpha(2,K)(S+1)} \otimes I_W$. As such, using the fact that, for any matrices $\boldsymbol{B}_i, i = 1, \ldots, 4$, the following equation holds good,

$$(\boldsymbol{B}_1 \otimes \boldsymbol{A}_2)(\boldsymbol{B}_3 \otimes \boldsymbol{B}_4) = \boldsymbol{B}_1\boldsymbol{B}_3 \otimes \boldsymbol{B}_2\boldsymbol{B}_4, \tag{26}$$

it can be obtained from (5) that

$$\boldsymbol{A}_0(\mathbf{e}^{\alpha(2,K)(S+1)} \otimes \boldsymbol{I}_W) = (\boldsymbol{I}_{\alpha(2,K)(S+1)} \otimes \boldsymbol{D}_1)(\mathbf{e}^{\alpha(2,K)(S+1)} \otimes \boldsymbol{I}_W) = \mathbf{e}^{\alpha(2,K)(S+1)} \otimes \boldsymbol{D}_1. \tag{27}$$

It follows from (7) and (17) that

$$\begin{aligned}&(\boldsymbol{A}_2 - \mathrm{diag}(\boldsymbol{\delta}^{(d)}))(\mathbf{e}^{\alpha(2,K)(S+1)} \otimes \boldsymbol{I}_W) = \\ &(\boldsymbol{I}_{\alpha(2,K)} \otimes \mu_1\boldsymbol{L}_1 \otimes \boldsymbol{I}_W)(\mathbf{e}^{\alpha(2,K)} \otimes \mathbf{e}^{S+1} \otimes \boldsymbol{I}_W) - \mu_1\mathbf{e}^{\alpha(2,K)} \otimes \mathbf{e}^{S+1}_{2:S+1} \otimes \mathrm{diag}(\mathbf{e}^W)) = \mathbf{0}.\end{aligned} \tag{28}$$

Finally, it follows from (9), (15), (16), after some algebra, that

$$\begin{aligned}&\left(\boldsymbol{A}_1^{(a)} + \boldsymbol{A}_1^{(b)} + \boldsymbol{A}_1^{(c)} - \mathrm{diag}(\boldsymbol{\delta}^{(a)} + \boldsymbol{\delta}^{(b)})\right)(\mathbf{e}^{\alpha(2,K)(S+1)} \otimes \boldsymbol{I}_W) \\ &= \mathbf{e}^{\alpha(2,K)(S+1)} \otimes \left(\boldsymbol{D}_0 + \textstyle\sum_{j=2}^{K} \boldsymbol{D}_j\right).\end{aligned} \tag{29}$$

Then, it follows from (27)–(29) that

$$\boldsymbol{A}(\mathbf{e}^{\alpha(2,K)(S+1)} \otimes \boldsymbol{I}_W) = \mathbf{e}^{\alpha(2,K)(S+1)} \otimes \boldsymbol{D}.$$

Thus, the subsets $\mathcal{E}_z, z \in \mathcal{W}$, are the states of a CTMC defined by the matrix $\boldsymbol{D}$. Now, we note that

$$\boldsymbol{\pi}\boldsymbol{A}(\mathbf{e}^{\alpha(2,K)(S+1)} \otimes \boldsymbol{I}_W) = \boldsymbol{\pi}(\mathbf{e}^{\alpha(2,K)(S+1)} \otimes \boldsymbol{I})\boldsymbol{D};$$

hence, $\boldsymbol{\theta} = \boldsymbol{\pi}(\mathbf{e}^{\alpha(2,K)(S+1)} \otimes \boldsymbol{I})$. Finally, from (5), obtain

$$\boldsymbol{\pi}\boldsymbol{A}_0\mathbf{1} = \boldsymbol{\theta}\boldsymbol{D}_1\mathbf{1}^W = \lambda_1.$$

Now, taking into account (7), (22), (24), and using the property (26) for the unit vector $\mathbf{e}$, the r.h.s. of the stability condition (20) becomes

$$\sum_{m=0}^{l_2} \boldsymbol{\pi}_m \left(\boldsymbol{I}_{\alpha(3,K)} \otimes \mu_1\boldsymbol{L}_1 \otimes \boldsymbol{I}_W\right)\mathbf{e} = \sum_{m=0}^{l_2} \boldsymbol{\pi}_m \mathbf{e}^{\alpha(3,K)} \otimes \mu_1\mathbf{e}^{S+1}_{2:S+1} \otimes \mathbf{e}^W.$$

Finally, (20) becomes

$$\rho := \lambda_1 \left[\sum_{m=0}^{l_2} \boldsymbol{\pi}_m \mathbf{e}^{\alpha(3,K)} \otimes \mu_1\mathbf{e}^{S+1}_{2:S+1} \otimes \mathbf{e}^W\right]^{-1} < 1. \tag{30}$$

Note that the stability condition (30) has a nice interpretation, since λ_1 is the upward drift, while the sum in brackets is the mean downward drift of class 1 customers at high levels obtained by aggregation of the components phase probabilities vector $\boldsymbol{\pi}$ corresponding to the states with positive departure probability of class 1 customers.

4. Steady State Performance

If the stability condition (30) holds, the stochastic vector of stationary probabilities, $\boldsymbol{q}$, exists and is the solution of the steady-state equation involving the infinitesimal generator $\boldsymbol{Q}$ given in (3):

$$\boldsymbol{q}\boldsymbol{Q} = \mathbf{0}. \tag{31}$$

Note that the components of the vector $\boldsymbol{q}$ are ordered lexicographically, that is, for $n_k \in \{0, \ldots, l_k\}, k = 1, \ldots, K, i \in \{0, \ldots, S\}$ and $z \in \mathcal{W}$,

$$\boldsymbol{q}(n_1, \ldots, n_K, i, z) = \lim_{t\to\infty} \mathrm{P}(N_1(t) = n_1, \ldots, N_K(t) = n_K, I(t) = i, Z(t) = z).$$

Since the matrix $\boldsymbol{Q}$ defines a level-independent QBD, the soluton can be obtained in matrix-geometric form [22]. Indeed, splitting the vector $\boldsymbol{q}$ into finite vectors $\boldsymbol{q}_0, \boldsymbol{q}_1, \ldots$ by the (value of the) first coordinate, we assume that

$$\boldsymbol{q}_i = \boldsymbol{q}_0 \boldsymbol{R}^i. \tag{32}$$

The matrix $\boldsymbol{R}$ is the so-called rate matrix, being the minimal non-negative solution of the matrix quadratic equation (which follows from (31) using the block-tridiagonal structure of $\boldsymbol{Q}$ and (32)),

$$\boldsymbol{R}^2 \boldsymbol{A}_2 + \boldsymbol{R}\boldsymbol{A}_1 + \boldsymbol{A}_0 = \boldsymbol{0}. \tag{33}$$

The boundary conditions for obtaining $\boldsymbol{q}_0$ follow from the first block-column of $\boldsymbol{Q}$ and are [23]

$$\boldsymbol{q}_0(\boldsymbol{A}_1^{(0)} + \boldsymbol{R}\boldsymbol{A}_2) = \boldsymbol{0}, \tag{34}$$

$$\boldsymbol{q}_0(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{1} = 1. \tag{35}$$

However, in order to avoid numerical difficulties, it is recommended to use the alternative matrix quadratic equation for the substochastic matrix $\boldsymbol{G}$ being the minimal non-negative solution of the system

$$\boldsymbol{A}_0 \boldsymbol{G}^2 + \boldsymbol{A}_1 \boldsymbol{G} + \boldsymbol{A}_2 = \boldsymbol{0}, \tag{36}$$

obtain the $\boldsymbol{R}$ matrix by the known relation [23]

$$\boldsymbol{R} = \boldsymbol{A}_0(-\boldsymbol{A}_1 - \boldsymbol{A}_0\boldsymbol{G})^{-1},$$

and, finally, use (32) to obtain the vector $\boldsymbol{q}$.

After obtaining the steady-state probability vector, it is straightforward to define the steady-state performance measures of interest. To do so, we use an auxiliary vector $\boldsymbol{j}^{(n)}$ of size $n+1$ containing the sequence $(0, \ldots, n)$, i.e.,

$$\boldsymbol{j}^{(n)} = (0, \ldots, n).$$

Average number of class 1 customers in the system:

$$\mathbf{E_1} = \sum_{i=1}^{\infty} i\boldsymbol{q}_i \mathbf{e}^{\alpha(2,K)(S+1)W} = \boldsymbol{q}_0 \boldsymbol{R}(\boldsymbol{I} - \boldsymbol{R})^{-2} \mathbf{e}^{\alpha(2,K)(S+1)W}. \tag{37}$$

Average number of class $k = 2, \ldots, K$ customers in the system:

$$\begin{array}{ll} \mathbf{E_k} & = \sum_{i=0}^{\infty} \boldsymbol{q}_i \mathbf{e}^{\alpha(2,k-1)} \otimes \boldsymbol{j}^{(l_k)} \otimes \mathbf{e}^{\alpha(k+1,K)} \otimes \mathbf{e}^{(S+1)W} \\ & = \boldsymbol{q}_0(\boldsymbol{I} - \boldsymbol{R})^{-1} \mathbf{e}^{\alpha(2,k-1)} \otimes \boldsymbol{j}^{(l_k)} \otimes \mathbf{e}^{\alpha(k+1,K)} \otimes \mathbf{e}^{(S+1)W} \end{array}. \tag{38}$$

Average inventory size in the system:

$$\bar{\mathbf{I}} = \sum_{i=0}^{\infty} \boldsymbol{q}_i \mathbf{e}^{\alpha(2,K)} \otimes \boldsymbol{j}^{(S)} \otimes \mathbf{e}^{W} = \boldsymbol{q}_0(\boldsymbol{I} - \boldsymbol{R})^{-1} \mathbf{e}^{\alpha(2,K)} \otimes \boldsymbol{j}^{(S)} \otimes \mathbf{e}^{W}. \tag{39}$$

Replenishment rate, where $\boldsymbol{\delta}^{(b)}$ is given in (16),

$$\bar{\mathbf{R}} = \sum_{i=0}^{\infty} \boldsymbol{q}_i \boldsymbol{\delta}^{(b)} = \boldsymbol{q}_0(\boldsymbol{I} - \boldsymbol{R})^{-1} \boldsymbol{\delta}^{(b)}. \tag{40}$$

Class k loss rate, $k = 2, \ldots, K$:

$$\mathbf{L}_k = \lambda_k \sum_{i=0}^{\infty} \boldsymbol{q}_i \mathbf{e}^{\alpha(2,k-1)} \otimes \mathbf{e}_{l_k+1} \otimes \mathbf{e}^{\alpha(k+1,K)} \otimes \mathbf{e}^{(S+1)W} = \lambda_k (\boldsymbol{I} - \boldsymbol{R})^{-1} \mathbf{e}^{\alpha(2,k-1)} \otimes \mathbf{e}_{l_k+1} \otimes \mathbf{e}^{\alpha(k+1,K)} \otimes \mathbf{e}^{(S+1)W}. \quad (41)$$

We note that the average number of queued customers can be obtained by the corresponding formula for the average number of customers in the system by replacing i to $(i-1)$. Finally, the corresponding sojourn times of class $k \geq 1$ customer are obtained by Little's law as follows:

$$\mathbf{E}_{\mathbf{s_k}} = \frac{\mathbf{E_k}}{\hat{\lambda}_k},$$

where $\hat{\lambda}_k$ is the effective arrival rate of class k, i.e., $\hat{\lambda}_1 = \lambda_1$ and

$$\hat{\lambda}_k = \lambda_k - \mathbf{L}_k.$$

5. Analysis of Inventory Recycle Time

Starting with inventory level S, the time taken to hit the level S again is called inventory cycle time, say, Γ. The distribution of inventory cycle time depends on the number of customers of all classes in the system and the phase of the arrival process. However, we note that, since the replenishment happens only at the event when the inventory hits the level s, the inventory cycle is essentially the lead time plus the time it takes to reach s from S by the steps decreasing the inventory, made by the process $\zeta(t)$. It is clear that such a process is a time until absorption of the process, where the absorption happens at the level s. It is known that such a time can be modeled by a phase-type distribution (for details on this type of distributions, see Reference [24]), and below we obtain the parameters of such a distribution.

Indeed, consider the inventory level $I(t) = S$. If $N_k(t) \geq S - s_{k-1}$, then the class k customers present in the system are capable of consuming all the inventory allowed for such a class, either before stopping service for this specific class at the boundary level s_{k-1}, or before hitting the absorbing inventory state $I(t) = s$, and in such case the time to absorption does not depend on the arrivals of class k customers after time t. Otherwise, arrivals of class k customers can be tracked until the condition $N_k(t) \geq S - s_{k-1}$ is met (if this happens before absorption). Moreover, if the level $N_k(t)$ hits the value $S - s_{k-1}$ from below, the departures of class k customers need not be tracked, but instead, only the inventory decreasing process should take into account the corresponding rate of the class k customer service, μ_k. Thus, the time it takes $I(t)$ to reach s (time to absorption) can be modeled by a finite state space absorbing CTMC, which is essentially a restriction of $\zeta(t)$ to the set

$$\{0, \ldots, S-s\} \times \{0, \ldots, S-s\} \times \{0, \ldots, S-s_2\} \times \ldots \{0, \ldots, S-s_{K-1}\} \times \{s+1, \ldots, S\} \times \mathcal{W}.$$

This means that the time to reach inventory level s has a phase-type distribution $PH(\boldsymbol{\beta}, \boldsymbol{T})$, where $\boldsymbol{\beta}$ is the initial distribution of the restricted chain $(\hat{N}_1(t), \ldots, \hat{N}_k(t), \hat{I}(t), Z(t))$ at time 0, and $\boldsymbol{T}$ is the transition matrix which follows from $\boldsymbol{Q}$ and the aforementioned restrictions. In particular, it is straightforward to define the transition rate matrix $\boldsymbol{T}$ as follows.

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{B}_1^{(0)} & \boldsymbol{B}_0 & \mathbb{O} & \mathbb{O} & \ldots & \mathbb{O} \\ \boldsymbol{B}_2 & \boldsymbol{B}_1 & \boldsymbol{B}_0 & \mathbb{O} & \ldots & \mathbb{O} \\ \mathbb{O} & \boldsymbol{B}_2 & \boldsymbol{B}_1 & \boldsymbol{B}_0 & \ldots & \mathbb{O} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathbb{O} & \ldots & \ldots & \ldots & \mathbb{O} & \widehat{\boldsymbol{B}}_1 \end{bmatrix}. \quad (42)$$

Note that $\boldsymbol{T}$ is a block-tridiagonal finite matrix, with $S-s+1$ blocks on the main diagonal, where the first block corresponds to states with $N_1(t)=0$, while the last block is for $N_1(t)=S-s$, respectively. The rows corresponding to the last block have zeroes except the block on the main diagonal, since upon reaching the level $N_1(t)=S-s$, neither the departures, nor the arrivals, of class 1 customers are tracked, but the rate μ_1 is taken into account in the value of the inventory decreasing rate. Moreover, the inventory states become in fact re-numbered so as to have $S-s$ states starting from inventory level $s+1$, numbered sequentially, that is, compared to the original inventory component $I(t)$, it holds that

$$\hat{I}(t)=I(t)-s, \tag{43}$$

and the absorption happens when $\hat{I}(t)$ makes a transition to 0. The size of the blocks is

$$\hat{\alpha}(2,K)(S-s)W,$$

where, for $2\le i\le j\le K$,

$$\hat{\alpha}(i,j)=\prod_{k=i}^{j}(S-s_{k-1}+1), \tag{44}$$

and $\hat{\alpha}(i,j)=0$ otherwise.

It is straightforward to see

$$\boldsymbol{B}_0=\boldsymbol{I}_{\hat{\alpha}(2,K)(S-s)}\otimes\boldsymbol{D}_1, \tag{45}$$

and

$$\boldsymbol{B}_2=\boldsymbol{I}_{\hat{\alpha}(2,K)}\otimes\mu_1\widehat{\boldsymbol{L}}_1\otimes\boldsymbol{I}_W, \tag{46}$$

where $\widehat{\boldsymbol{L}}_j$ is obtained from the matrix $\boldsymbol{L}_j$ defined in (6) by reducing the latter to rows and columns from $s_{j-1}+2$ to $S+1$, i.e.,

$$\widehat{\boldsymbol{L}}_j=\boldsymbol{J}_{s_{j-1}+1,S+1}\boldsymbol{L}_j\boldsymbol{J}'_{s_{j-1}+1,S+1},$$

where the matrix $\boldsymbol{J}_{a,b}$ removes the first a rows, $1\le a\le b$, that is,

$$\boldsymbol{J}_{a,b}=\begin{bmatrix}\mathbb{O}_{(b-a)\times a} & \boldsymbol{I}_{b-a}\end{bmatrix},$$

with an exception for $\widehat{\boldsymbol{L}}_1$ defined as

$$\widehat{\boldsymbol{L}}_1=\boldsymbol{J}_{s+1,S+1}\boldsymbol{L}_1\boldsymbol{J}'_{s+1,S+1}.$$

Let us restrict the matrices $\boldsymbol{N}_j^{(+)}$ and $\boldsymbol{N}_j^{(-)}$ defined in (8) to the finite state space of the absorbing CTMC, and modify the last row of $\boldsymbol{N}_j^{(-)}$ so as to force $\hat{N}_j(t)$ to stay at the level $S-s_{j-1}$ once reached, $j\ge 2$. To do so, denote

$$\boldsymbol{H}_{a,b}=\begin{bmatrix}\boldsymbol{I}_a & \mathbb{O}_{a\times(b-a)}\end{bmatrix}.$$

Multiplication by $\boldsymbol{H}_{a,b}$ on the left leaves only the first a rows in the resulting matrix and removes the last $b-a$ rows, while multiplication on the right by $\boldsymbol{H}'_{a,b}$ removes the last $b-a$ columns. Then, the matrices $\widehat{\boldsymbol{N}}_j^{(+)}$ and $\widehat{\boldsymbol{N}}_j^{(-)}$ are defined as follows:

$$\widehat{\boldsymbol{N}}_j^{(+)}=\boldsymbol{H}_{S-s_{j-1}+1,S+1}\boldsymbol{N}_j^{(+)}\boldsymbol{H}'_{S-s_{j-1}+1,S+1}+\mathrm{diag}(\mathbf{e}_{S-s_{j-1}+1}),$$

$$\widehat{\boldsymbol{N}}_j^{(-)}=\boldsymbol{H}_{S-s_{j-1}+1,S+1}\boldsymbol{N}_j^{(-)}\boldsymbol{H}'_{S-s_{j-1}+1,S+1}\mathrm{diag}(\mathbf{e}^{S-s_{j-1}+1}-\mathbf{e}_{S-s_{j-1}+1})+\mathrm{diag}(\mathbf{e}_{S-s_{j-1}+1}),$$

where the $S+1$-dimensional vector $\mathbf{e}_{S+1}$ is zero vector except the last component equal to one, while $\mathbf{e}^{S+1}$ is the $S+1$-dimensional vector of ones. Now, we are ready to define the matrix $\boldsymbol{B}_1$ as follows:

$$\boldsymbol{B}_1 = \sum_{j=2}^{K} \boldsymbol{I}_{\widehat{\alpha}(2,j-1)} \otimes \Big(\widehat{\boldsymbol{N}}_j^{(+)} \otimes \boldsymbol{I}_{\widehat{\alpha}(j+1,K)} \otimes \boldsymbol{I}_{S-s} \otimes \boldsymbol{D}_j + \widehat{\boldsymbol{N}}_j^{(-)} \otimes \boldsymbol{I}_{\widehat{\alpha}(j+1,K)} \otimes \mu_j \widehat{\boldsymbol{L}}_j \otimes \boldsymbol{I}_W \Big) + \boldsymbol{I}_{\widehat{\alpha}(2,K)} \otimes \boldsymbol{I}_{S-s} \otimes \boldsymbol{D}_0 - \tilde{\boldsymbol{\Delta}}.$$

Note that the diagonal matrix $\tilde{\boldsymbol{\Delta}}$ holds the exit rates from states which do not lead to absorption, that is, transitions from inventory levels $\hat{I}(t) = 2, \ldots, S-s$ according to enumeration (43). As such, this matrix can be defined explicitly as follows:

$$\tilde{\boldsymbol{\Delta}} = \mu_1 \mathbf{e}^{\widehat{\alpha}(2,K)} \otimes \mathbf{e}_{2:S-s}^{S-s} \otimes \mathbf{e}^{W} + \sum_{j=2}^{K} \mu_j \mathbf{e}^{\widehat{\alpha}(2,j-1)} \otimes \mathbf{e}_{2:S-s_{j-1}+1}^{S-s_{j-1}+1} \otimes \mathbf{e}^{\widehat{\alpha}(j+1,K)} \otimes \mathbf{e}_{s_{j-1}-s+2:S-s}^{S-s} \otimes \mathbf{e}^{W}.$$

Similarly to (19), the matrix $\boldsymbol{B}_1^{(0)}$ differs from $\boldsymbol{B}_1$ only on the diagonal (the states corresponding to $\boldsymbol{B}_1^{(0)}$ have no class 1 customers), so that

$$\boldsymbol{B}_1^{(0)} = \boldsymbol{B}_1 - \mu_1 \mathbf{e}^{\widehat{\alpha}(2,K)} \otimes \mathbf{e}_{2:S-s}^{S-s} \otimes \mathbf{e}^{W}.$$

Finally, we need to define the matrix $\widehat{\boldsymbol{B}}_1$ corresponding to the boundary states. Since at the *level* (number of class 1 customers) $S-s$, neither arrivals nor departures of the class 1 customers are tracked,

$$\widehat{\boldsymbol{B}}_1 = \boldsymbol{B}_0 + \boldsymbol{B}_1 + \boldsymbol{B}_2.$$

It remains to denote $\boldsymbol{t}_0 = -\boldsymbol{T}\mathbf{e}$ as the corresponding absorption rate vector, and note that the initial state probability vector should be taken so as to have initial inventory equal to S, that is, $\mathrm{P}(I(0) = S) = 1$.

To simplify comprehension, we outline the transitions possible for the chain

$$\{(\hat{N}_1(t), \ldots, \hat{N}_K(t), \hat{I}(t), Z(t))\}, t \geq 0,$$

according to the subgenerator matrix $\boldsymbol{T}$ and the absorption vector $\boldsymbol{t}_0$. For some $t \geq 0$, fix $(\hat{N}_1(t), \ldots, \hat{N}_K(t), \hat{I}(t), Z(t)) = (n_1, \ldots, n_K, i, z)$. Then, the state after transition is one of the following:

- $(n_1, \ldots, n_j + 1_{n_j < S - s_{j-1}}, \ldots, n_K, i, z')$, with rate $(\boldsymbol{D}_j)_{z,z'}$ (arrival of a class j customer),
- $(n_1, \ldots, n_j, \ldots, n_K, i, z')$, with rate $(\boldsymbol{D}_0)_{z,z'}$ (change of the MMAP phase),
- $(n_1, \ldots, n_j - 1_{n_j < S - s_{j-1}}, \ldots, n_K, i-1, z')$, with rate μ_j, if $n_j > 0$ and $i > s_{j-1} - s$ (departure of a class j customer, if allowed).

Note that the absorption happens if the transition is made from the inventory level $i = 1$ downward. In particular, this means that, since $s_2 > s$, the absorption happens only due to a transition caused by service completion of either class 1, or class 2 customer.

It is well known that time to absorption, say X, of an absorbing CTMC defined by a subgenerator $\boldsymbol{T}$ and initial state probability vector $\boldsymbol{\beta}$ has a phase-type distribution with mean [24]

$$\mathrm{E}X = \boldsymbol{\beta}(-\boldsymbol{T})^{-1}\mathbf{e}. \tag{47}$$

Thus, to define the mean inventory cycle time, EΓ, it only remains to convert the steady-state probability vector $\boldsymbol{q}$ obtained in (32) into the initial state probability vector for the corresponding phase-type distribution defined by the matrix (42).

We summarize our findings in the following Lemma.

Lemma 1. *Let* Γ *be the inventory cycle time. Then,*

$$\mathrm{E}\Gamma = \frac{1}{\gamma} + \sum_{n=0}^{\infty} \widehat{\boldsymbol{q}}_n (-\boldsymbol{T})^{-1}\mathbf{e}, \tag{48}$$

where the vector $\widehat{\boldsymbol{q}}$*, split into finite subvectors by the first component, is defined from the vector* $\boldsymbol{q}$ *in a component-wise manner as follows: if* $n_k < S - s_{k-1}, k = 1, \ldots, K$*, then*

$$\widehat{\boldsymbol{q}}(n_1, \ldots, n_K, S, z) = \boldsymbol{q}(n_1, \ldots, n_K, S, z),$$

and if for some $m \geq 1$ *and indices* $k_1, \ldots, k_m$*, holds* $n_{k_j} = S - s_{k_j - 1}, j = 1, \ldots, m$*, then*

$$\widehat{\boldsymbol{q}}(n_1, \ldots, n_K, S, z) = \sum_{\hat{n}_1, \ldots, \hat{n}_K \in \mathcal{N}} \boldsymbol{q}(\hat{n}_1, \ldots, \hat{n}_K, S, z),$$

where

$$\mathcal{N} = \Big\{ (\hat{n}_1, \ldots, \hat{n}_K) : \hat{n}_k = n_k, k \in \{1, \ldots, m\} \backslash \{k_1, \ldots, k_m\}; \hat{n}_{k_j} \geq S - s_{k_j - 1}, j = 1, \ldots, m \Big\}.$$

The vector $\widehat{\boldsymbol{q}}$ *is zero elsewhere.*

It remains to note that the expectation is taken so as to summarize the resulting times by the appropriate steady-state probabilities involving the inventory level S, and the term $\frac{1}{\gamma}$ is added to emphasize the exponentially distributed lead time.

6. Numerical Illustration

6.1. Stability Condition Parametric Sensitivity

In this section we illustrate the sensitivity of the stability condition (30) on the parameters. In particular, we take $K = 2$, fix the MMAP arrival process, inventory size S and queue size for the second class customers, l_2. We then vary the rates μ_1, μ_2 and γ for several values of replenishment level s, *ceteris paribus*, and plot the resulting dependency. In the experiment we use the following default settings:

$$S = 100,\ s = 7,\ l_2 = 15,\ \mu_1 = 7,\ \mu_2 = 15,\ \gamma = 10.$$

The MMAP arrival process with $W = 2$ is driven by the following matrices:

$$\boldsymbol{D}_0 = \begin{bmatrix} -13 & 1 \\ 2 & -14 \end{bmatrix},\ \boldsymbol{D}_1 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix},\ \boldsymbol{D}_2 = \begin{bmatrix} 5 & 4 \\ 3 & 2 \end{bmatrix}. \tag{49}$$

Thus, the arrival rate of class 1 customers may be calculated as $\lambda_1 \approx 4.867$.

In the first experiment we consider ρ defined in (30) to be the function of $\mu_2 \in [2, 15]$ for values $s = 5, 6, 7$. The resulting curves depict $\rho_s(\mu_2)$ versus μ_2 on Figure 2 (top). While in absolute values the variability of ρ is rather small, a non-linear dependency is clearly visible, and the dependency on s is motivated by the relatively higher impact of the second class customers on the system load for smaller values of s. Counter-intuitively though, the load *increases* (slightly) with an increasing service rate of the second class customers.

To investigate this interesting effect, we consider the probability $\mathrm{P}(N_2 = i)$ for two boundary values, $i = 0, l_2$, that is, the probability that second class customer queue is empty or full. We plot the corresponding estimates within the setup of the first experiment on Figure 3. It can be observed that, for smaller values of μ_2 (that is, for larger mean service times of class 2 customers), the queue is mostly overloaded (with high probability the queue is full), which causes relatively high loss of class 2 customers. In contrast, for $\mu_2 > 7$, the probability of an empty class 2 queue overtakes the probability of a queue completely occupied, and the former is increasing, while the latter is decreasing with increasing μ_2.

We note that, in the experiment, we used only $s = 7$ since, as numerical results show, the effect of s on empty/full probability of class 2 customers is negligible for $s = 5, 6, 7$.

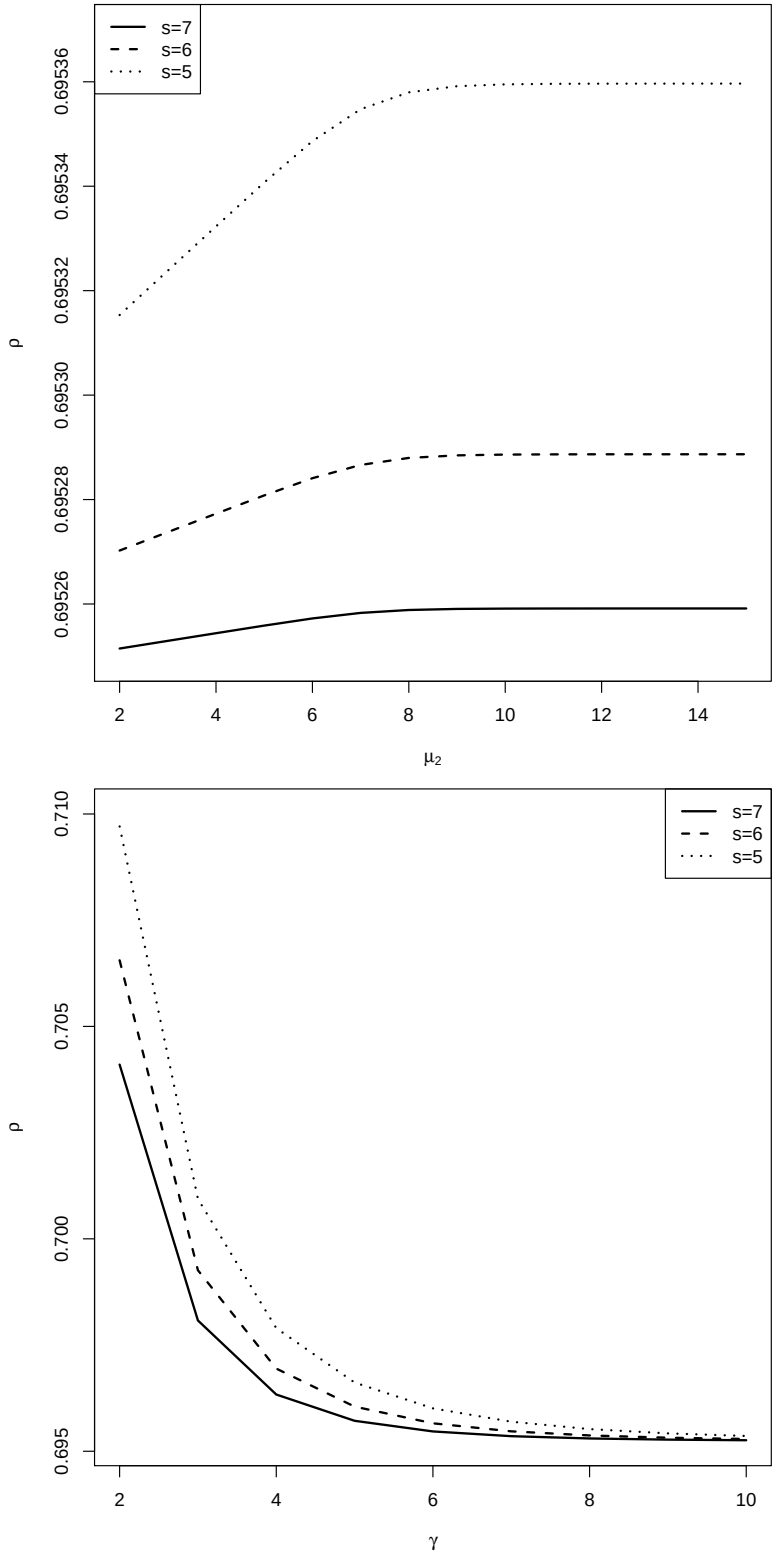

Figure 2. Dependency of ρ on $\mu_2 \in [2, 15]$ (**top**, fixed $\gamma = 10$) and on on $\gamma \in [2, 10]$ (**bottom**, fixed $\mu_2 = 15$) for $s = 5, 6, 7$ and other parameters fixed: $S = 100, l_2 = 15, \mu_1 = 7$, MMAP arrival process parameters given in (49).

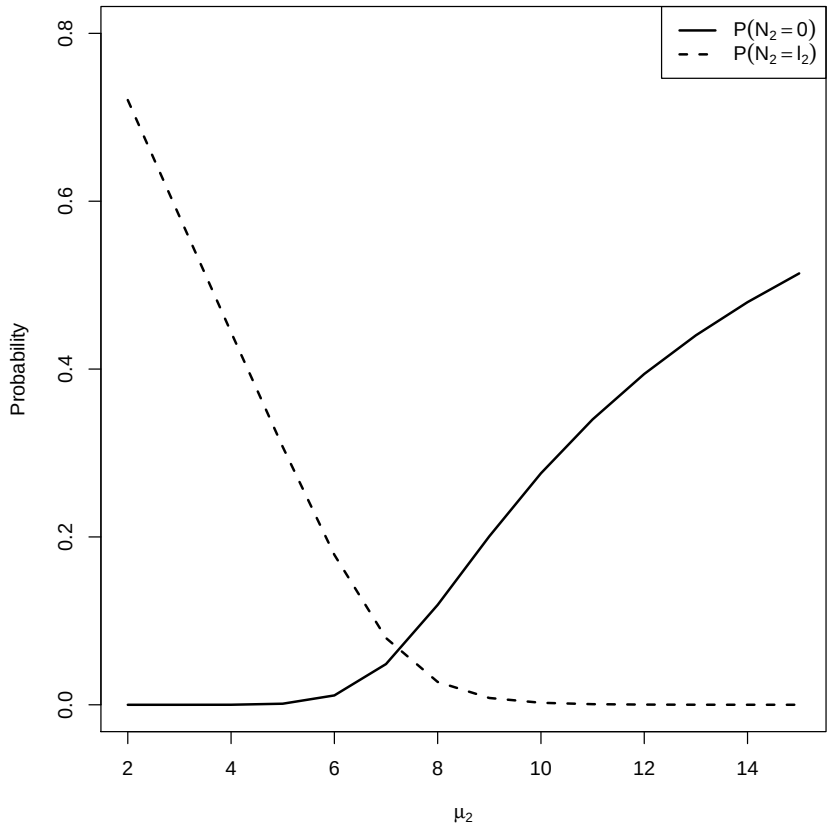

Figure 3. Dependency of the probabilities of an empty (full) class 2 queue on $\mu_2 \in [2, 15]$, fixed $\gamma = 10, s = 7, S = 100, l_2 = 15, \mu_1 = 7$, MMAP arrival process parameters given in (49).

In the second experiment, we fix $\mu_2 = 15$ and consider ρ defined in (30) to be the function of $\gamma \in [2, 10]$ for values $s = 5, 6, 7$. The resulting curves depict $\rho_s(\gamma)$ versus γ on Figure 2 (bottom). In this experiment, the non-linear dependency on γ follows the intuition: higher replenishment rate decreases the customer's queuing time and results in a lower system load.

Finally, we fix $\mu_2 = 15, \gamma = 10$ and consider ρ defined in (30) to be the function of $\mu_1 \in [7, 12]$. However, since the dependency on s is rather weak, we take more contrast values $s = 1$ and $s = 50$. The resulting curves depict $\rho_s(\mu_1)$ versus μ_1 on Figure 4. As expected, increasing μ_1, and, hence, decreasing the service time of class 1 customers, causes a dramatic decrease of the system load. The additional load caused by a "lazy" replenishment at the level $s = 1$ is also visible.

6.2. Steady-State Performance Sensitivity

In this section, we illustrate the sensitivity of the performance measures (37)–(41) described in Section 4 to the management parameter γ, that is, the lead time intensity. To do so, we take $K = 2$ and slightly modify the parameters used in the previous section, so as to make computations more convenient. Namely, we take

$$S = 50,\ s = 5,\ l_2 = 10,\ \mu_1 = 7,\ \mu_2 = 15.$$

We use the same MMAP defined in (49). We vary $\gamma = 1, \ldots, 10$ and obtain the performance measures (37)–(41) for given γ, *ceteris paribus*. Finally, we depict the obtained numerical results.

Figure 5 (top) describes the dependency of the mean number of class 1 and class 2 customers given in (37) and (38), on the replenishment rate γ. A decreasing pattern with increasing γ is caused by decreasing load ρ given in (30); see Figure 5 (bottom).

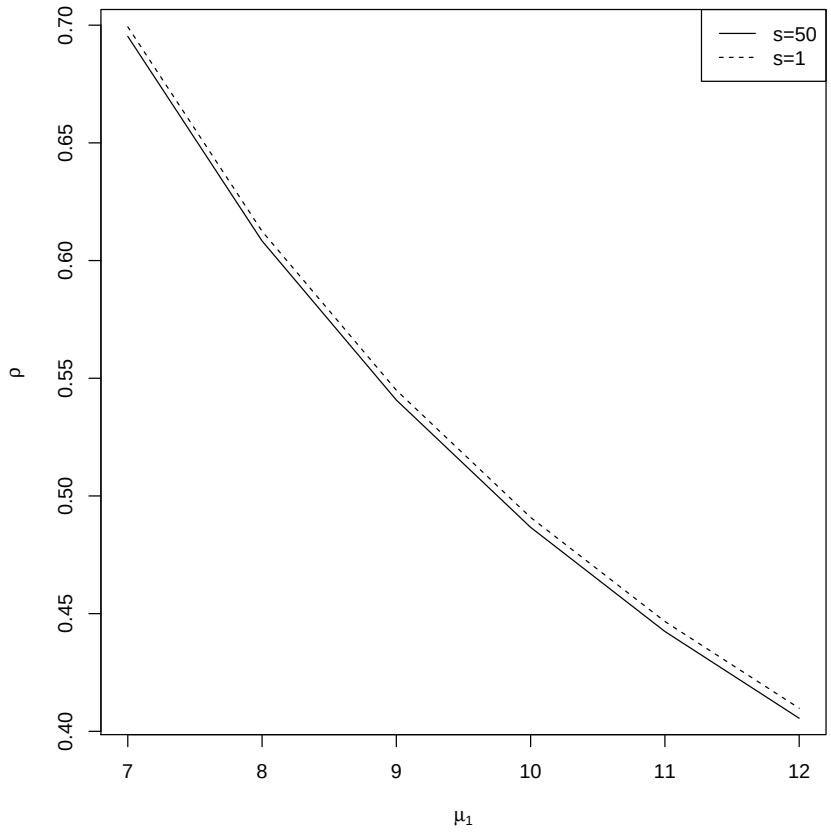

Figure 4. Dependency of ρ on $\mu_1 \in [7, 12]$ for $s = 1, 50$ and other parameters fixed: $S = 100$, $l_2 = 15, \mu_1 = 7, \mu_2 = 15, \gamma = 10$, MMAP arrival process parameters given in (49).

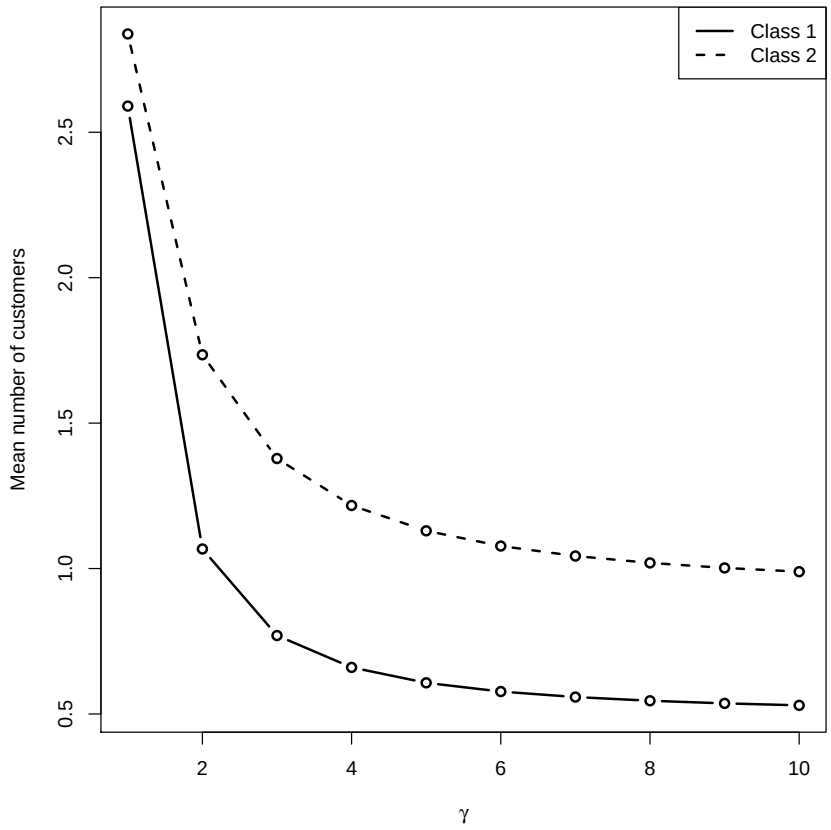

Figure 5. *Cont.*

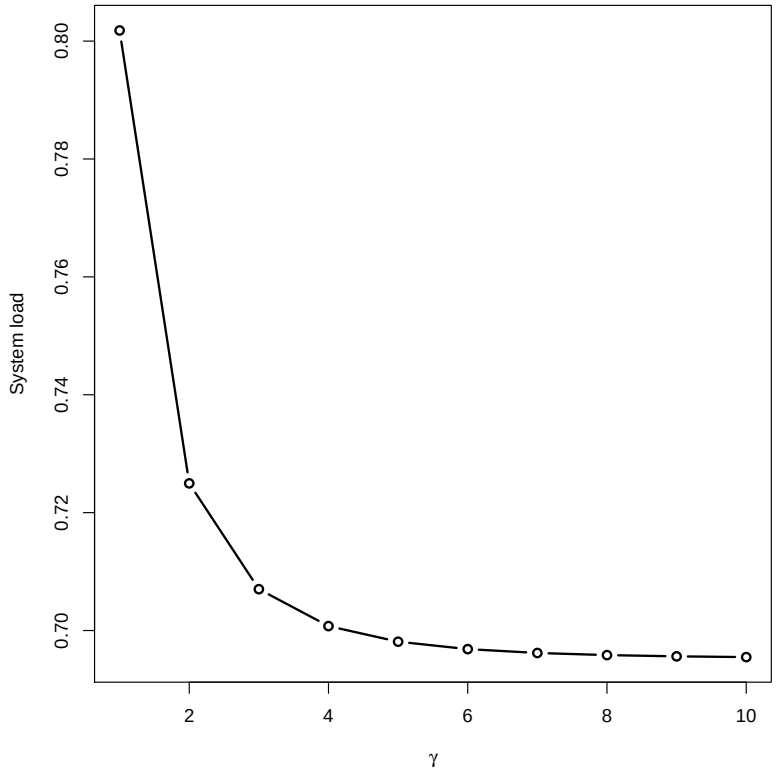

Figure 5. Dependency of $\mathbf{E}_i, i = 1,2$ (**top**) and ρ (**bottom**) on $\gamma = 1,\ldots,10$ for $S = 50$, $s = 5$, $l_2 = 10$, $\mu_1 = 7$, $\mu_2 = 15$, MMAP arrival process parameters given in (49).

Figure 6 (top) describes the dependency of the mean inventory level $\bar{\mathbf{I}}$ defined in (39), while, in Figure 6 (bottom), the replenishment rate $\bar{\mathbf{R}}$ is depicted, for $\gamma = 1,\ldots,10$. As expected, with increasing γ the inventory contents increases, along with the replenishment rate.

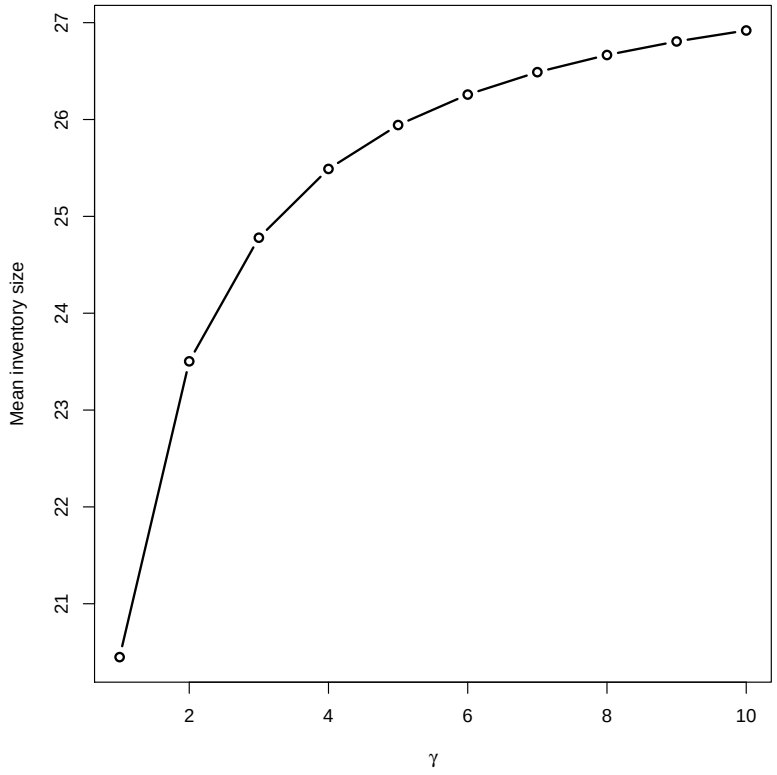

Figure 6. *Cont.*

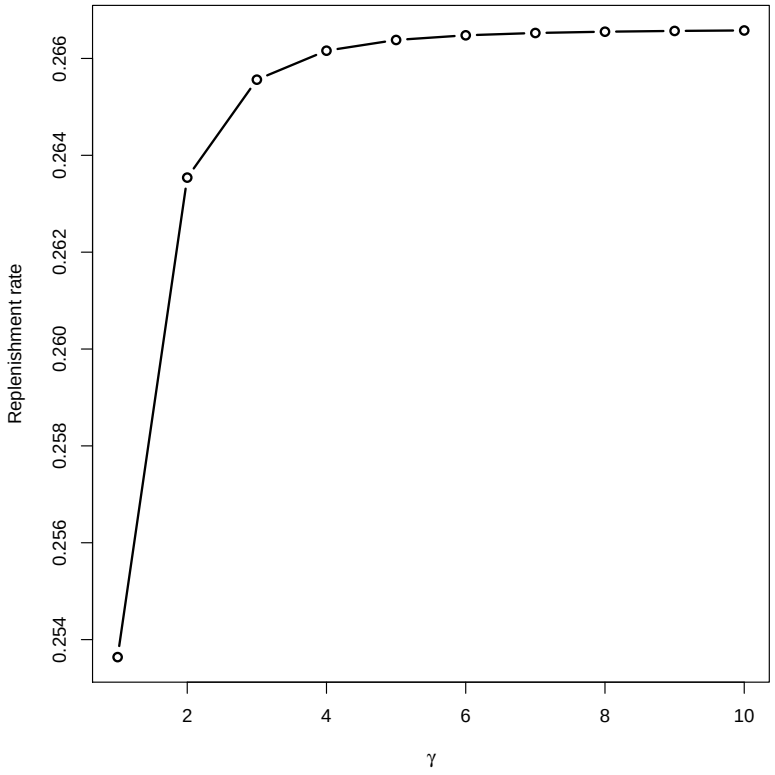

Figure 6. Dependency of $\bar{\mathbf{I}}$ (**top**) and $\bar{\mathbf{R}}$ (**bottom**) on $\gamma = 1, \ldots, 10$ for $S = 50,\ s = 5,\ l_2 = 10,$ $\mu_1 = 7,\ \mu_2 = 15$, MMAP arrival process parameters given in (49).

Finally, Figure 7 describes the dependency of the class 2 customer loss rate $\mathbf{L}_2$ defined in (41), on $\gamma = 1, \ldots, 10$. As expected, with increasing γ the customer loss rate is decreasing.

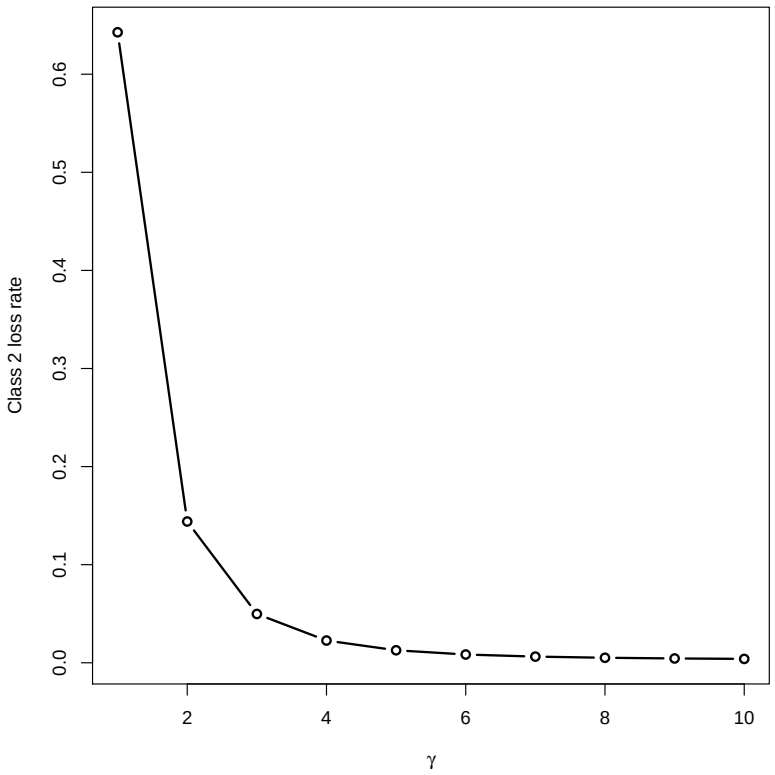

Figure 7. Dependency of $\mathbf{L}_2$ on $\gamma = 1, \ldots, 10$ for $S = 50,\ s = 5,\ l_2 = 10,\ \mu_1 = 7,\ \mu_2 = 15$, MMAP arrival process parameters given in (49).

6.3. Total Cost Optimization

Based on the above performance measures, we obtain expected total cost per unit time in the considered system as follows:

$$E_{TC} = c_I \bar{\mathbf{I}} + \sum_{k=2}^{K} c_{L_k} \mathbf{L}_k + c_R \bar{\mathbf{R}} + \sum_{k=1}^{K} c_{w_k} \mathbf{E}_\mathbf{k}, \tag{50}$$

where the non-negative coefficients are defined on a per system | customer | item per time unit basis:

- c_I is the holding cost of inventory (per item per unit time),
- c_{L_k} is the cost due to a class k customer loss (per system per unit time),
- c_R is the reorder cost (per system per unit time),
- c_{w_k} is the waiting cost of class k customer (per customer per unit time), $k \leq K$.

Using the same 2-class system with same parameters as in Section 6.2, that is,

$$S = 50,\ s = 5,\ l_2 = 10,\ \mu_1 = 7,\ \mu_2 = 15,$$

and MMAP defined in (49), we perform a numerical exploration of sensitivity of the cost function E_{TC} given in (50) on the parameter c_I. This is motivated as follows. Since the per-customer measures are decreasing, while the per-system measures are increasing, there should be a trade-off between keeping a high inventory level and compromising some second-class customer losses. This is regulated by the inventory holding cost parameter c_I. For the experiment, we take $c_I = 1, 5, 10$ and vary $\gamma = 1, \ldots, 10$, as in the previous section. We depict the resulting curves on Figure 8 and observe clearly that the optimal (minimal) cost shifts to lower values of γ with increasing c_I, which is intuitive.

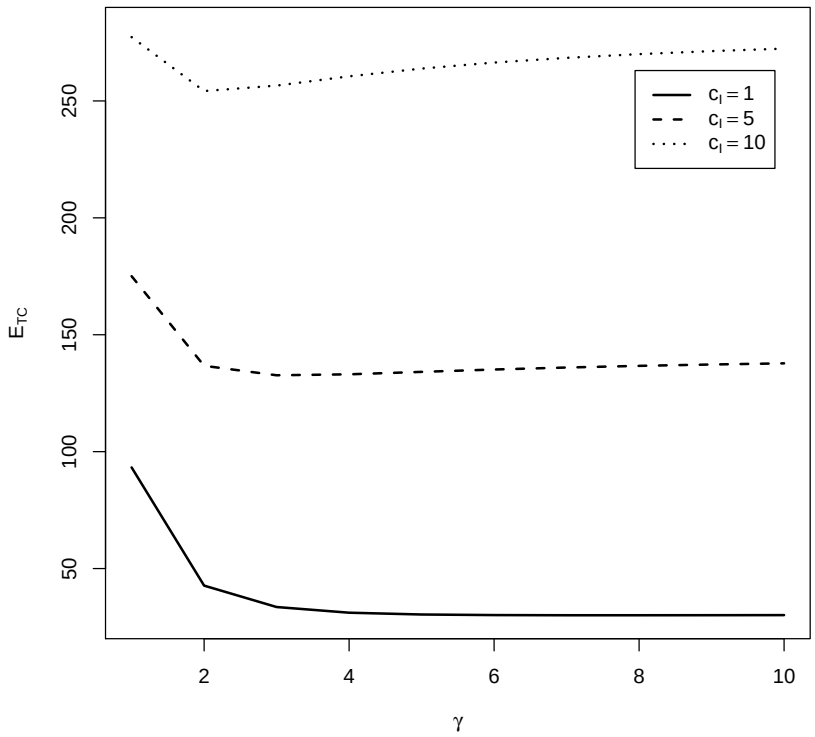

Figure 8. Total cost E_{TC} versus $\gamma = 1, \ldots, 10$ for various $c_I = 1, 5, 10$; the parameters are $S = 50$, $s = 5$, $l_2 = 10$, $\mu_1 = 7$, $\mu_2 = 15$, MMAP arrival process parameters given in (49).

7. Conclusions

We have analyzed a queuing inventory system with correlated arrivals of heterogeneous customers. The general case of K types of customers was investigated, where each

class of customers is assigned some priority according to which the inventory access is managed. We illustrated the general system and a special case of $K = 2$. An intuitive stability condition was derived and the steady state probability vector was obtained. Key performance measures of the system were obtained. The inventory cycle time was analyzed and it was shown that the inventory cycle follows a phase type distribution.

The large scale models have computational difficulties due to a large state space (while the matrices are sparse), so we cannot rely on numerical investigation only. Thus, we performed numerical experiments only to highlight interesting features of the model, and an interesting cost optimization problem has been stated which could be studied in future for practical applications.

As future directions of investigation we would like to point out the following possibilities. First, it is relatively easy to incorporate customer impatience (which will cause additional flow of customers towards decreasing the *level* and the first K components of the *phase* in the corresponding QBD process) and balking upon arrival finding the server busy (which modifies the rates related to corresponding class customer arrivals). Phase-type service times and replenishment times can also be incorporated leaving the model mathematically tractable. At the same time, the inventory replenishment discipline can be modified into booking-type, where the inventory items are booked for specific customers upon arrival; thus, arriving customers are rejected if no items are available for booking. Comparison of this type of model with the one analyzed in the present paper is one of the promising directions for future research.

Thinking beyond simple extensions of the model, the retrial queues instead of classical queues can be incorporated, while the stability part most likely can be extended towards a regenerative input. Finally, it might be interesting to consider the inventory with common lifetime of items and priorities related to item "freshness".

It might be also interesting to consider the model in transient regime, according to possible applications in social systems. A few methods are available for this type of analysis of QBD processes [25–27]. However, this might require inverting Laplace transforms which might cause numerical instability. In this regard, we refer to a recent work, Reference [28], where a novel method of numerically effective Laplace transform inversion is presented.

Author Contributions: Conceptualization, A.K., K.R., and M.J.J.; formal analysis, A.K., K.R., M.J.J., and A.S.R.; Writing—original draft preparation, K.R.; simulation and visualization, K.R. and A.S.R.; writing—review and editing, A.K., A.S.R., and M.J.J.; supervision, project administration, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Ministry of Science and Higher Education of the Russian Federation Grant 075-15-2019-1621. The work of AR is partially supported by RFBR, projects 19-57-45022, 19-07-00303.

Acknowledgments: The authors would like to thank anonymous referees for review and detailed comments that helped to improve the paper.

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

References

1. Chakravarthy, S.R.; Maity, A.; Gupta, U.C. An '(s, S)' inventory in a queueing system with batch service facility. *Ann. Oper. Res.* **2017**, *258*, 263–283. [CrossRef]
2. Chakravarthy, S.R.; Rumyantsev, A. Analytical and simulation studies of queueing-inventory models with MAP demands in batches and positive phase type services. *Simul. Models Pract. Theory* **2020**, *103*, 102092. [CrossRef]
3. Altendorfer, K. Influence of lot size and planned lead time on service level and inventory for a single-stage production system with advance demand information and random required lead times. *Int. J. Prod. Econ.* **2015**, *170*, 478–488. [CrossRef]
4. Armony, M.; Plambeck, E.L. The impact of duplicate orders on demand estimation and capacity investment. *Manag. Sci.* **2005**, *51*, 1505–1518. [CrossRef]
5. Sohraby, K.; Sidi, M. On the performance of bursty and modulated sources subject to leaky bucket rate-based access control schemes. *IEEE Trans. Commun.* **1994**, *42*, 477–487. [CrossRef]
6. Choi, D.I.; Lim, D.E. Analysis of the State-Dependent Queueing Model and Its Application to Battery Swapping and Charging Stations. *Sustainability* **2020**, *12*, 2343. [CrossRef]

7. Sigman, K.; Simchi-Levi, D. Light traffic heuristic for anM/G/1 queue with limited inventory. *Ann. Oper. Res.* **1992**, *40*, 371–380. [CrossRef]
8. Berman, O.; Kaplan, E.H.; Shevishak, D.G. Deterministic approximations for inventory management at service facilities. *IIE Trans.* **1993**, *25*, 98–104. [CrossRef]
9. Krishnamoorthy, A.; Lakshmy, B.; Manikandan, R. A survey on inventory models with positive service time. *Opsearch* **2011**, *48*, 153–169. [CrossRef]
10. Enders, P.; Adan, I.; Scheller-Wolf, A.; van Houtum, G.J. Inventory rationing for a system with heterogeneous customer classes. *Flex. Serv. Manuf. J.* **2014**, *26*, 344–386. [CrossRef]
11. Avrachenkov, K.; Dudin, A.; Klimenok, V. Retrial queueing model MMAP/M 2/1 with two orbits. In Proceedings of the International Workshop on Multiple Access Communications, Barcelona, Spain, 13–14 September 2010; pp. 107–118.
12. He, Q.M.; Stanford, D.A. Distributions of the interdeparture times in FCFS and nonpreemptive priority MMAP [2]/G [2]/1 queues. *Perform. Eval.* **1999**, *38*, 85–103. [CrossRef]
13. Klimenok, V.; Dudin, A.; Vishnevsky, V. Priority Multi-Server Queueing System with Heterogeneous Customers. *Mathematics* **2020**, *8*, 1501. [CrossRef]
14. Dudin, S.; Kim, C.; Dudina, O. MMAP | M | N queueing system with impatient heterogeneous customers as a model of a contact center. *Comput. Oper. Res.* **2013**, *40*, 1790–1803. [CrossRef]
15. He, Q.M. Quasi-birth-and-death Markov processes with a tree structure and the MMAP [K]/PH [K]/N/LCFS non-preemptive queue. *Eur. J. Oper. Res.* **2000**, *120*, 641–656. [CrossRef]
16. He, Q.M. Queues with marked customers. *Adv. Appl. Probab.* **1996**, *28*, 567–587. [CrossRef]
17. Neuts, M.F. A versatile Markovian point process. *J. Appl. Probab.* **1979**, *16*, 764–779. [CrossRef]
18. He, Q.M.; Neuts, M.F. Markov chains with marked transitions. *Stoch. Process. Their Appl.* **1998**, *74*, 37–52. [CrossRef]
19. He, Q.M. The Versatility of MMAP[K] and the MMAP[K]/G[K]/1 Queue. *Queueing Syst.* **2001**, *38*, 397–418. [CrossRef]
20. Chakravarthy, S.R. Markovian Arrival Processes. In *Wiley Encyclopedia of Operations Research and Management Science*; American Cancer Society: Atlanta, GA, USA, 2011; [CrossRef]
21. Geerts, F.; Blondia, C. Superposition of Markov Sources and Long Range Dependence. In *Broadband Communications*; Springer: Boston, MA, USA, 1998; pp. 550–561. [CrossRef]
22. He, Q.M. *Fundamentals of Matrix-Analytic Methods*; Springer: New York, NY, USA, 2014.
23. Latouche, G.; Ramaswami, V. *Introduction to Matrix Analytic Methods in Stochastic Modeling*; ASA-SIAM: Philadelphia, PA, USA, 1999.
24. Bladt, M.; Nielsen, B.F. *Matrix-Exponential Distributions in Applied Probability*; Probability Theory and Stochastic Modelling; Springer: Boston, MA, USA, 2017; Volume 81. [CrossRef]
25. Lucantoni, D. Further transient analysis of the BMAP/G/1 Queue. *Commun. Stat. Stoch. Models* **1998**, *14*, 461–478. [CrossRef]
26. Kulkarni, L.A.; Li, S.Q. Transient behaviour of queueing systems with correlated traffic. *Commun. Stat. Stoch. Models* **1998**, *14*, 933–978. [CrossRef]
27. Zhang, J.; Coyle, E.J. Transient analysis of quasi-birth-death processes. *Commun. Stat. Stoch. Models* **1989**, *5*, 459–496. [CrossRef]
28. Almousa, S.A.D.; Horváth, G.; Telek, M. Transient analysis of piecewise homogeneous QBD process. *Stoch. Models* **2021**, *37*, 59–84. [CrossRef]