Annals of Operations Research
https://doi.org/10.1007/s10479-022-05140-1

ORIGINAL RESEARCH

Stability of queueing-inventory systems with customers of different priorities

Sonja Otten[1] · Hans Daduna[2]

Accepted: 12 December 2022

Abstract
We study a production-inventory system with two customer classes with different priorities which are admitted to the system following a flexible admission control scheme. The inventory management is according to a base stock policy and arriving demand which finds the inventory depleted is lost (lost sales). We analyse the global balance equations of the associated Markov process and derive structural properties of the steady state distribution which provide insights into the equilibrium behaviour of the system. We derive a sufficient condition for ergodicity using the Foster-Lyapunov stability criterion. For a special case we show that the condition is necessary as well.

Keywords Queueing networks · Inventory control · Priority customer

Mathematics Subject Classification 60K25 · 68M20 · 90B22

1 Introduction

Adequate and differentiated service of customers becomes more and more important and product availability is an important aspect of service (cf. Isotupa [2006, p. 687]). There are many situations where it would be financially beneficial to provide different service levels to different customer classes (cf. Isotupa [2015, p. 411], Liu et al. [2013, pp. 1544f.] and Yadavalli et al. [2015, p. 637]): Companies often need to provide different service levels to different customers based on their contracts. This means that orders with long term contracts have higher priority than unscheduled orders since they may bear lower shortage cost than the booked orders. Another example is the case where different customers pay different prices for the same product. Then the company has an incentive to meet more of the demand

✉ Sonja Otten
sonja.otten@tuhh.de

Hans Daduna
hans.daduna@uni-hamburg.de

[1] Institute of Mathematics, Hamburg University of Technology, Am Schwarzenberg-Campus 3, 21073 Hamburg, Germany

[2] Department of Mathematics, Universität Hamburg, Bundesstraße 55, 20146 Hamburg, Germany

Published online: 28 December 2022

of the customer who pays a higher price than a customer who pays a lower price. Theory and application of priority queues are well established parts of classical queueing theory (see e.g. Jaiswal (1968) and Wang et al. (2015)). Similarly, theory and application of inventory holding for streams of differentiated demand has found much interest in research (see e.g. Isotupa (2015, 2011, 2007, 2006), Isotupa and Samanta (2013), Chen et al. (2012), Liu et al. (2013, 2014)). However, up to now, one of the key assumptions of queueing-inventory models in literature is that customers are indistinguishable.

The research communicated in this article is devoted to queueing-inventory systems where customers of two classes of different priorities are served. For serving a customer a piece of raw material from an associated finished goods inventory is needed. The inventory management is according to a standard base stock policy. Replenishment orders are fulfilled by an external supplier and the lead time is positive. Customers arriving when the inventory is depleted are rejected (lost sales). Additionally arrivals are regulated by a flexible admission control which respects priorities.

Our main interest are stability conditions for this system and the stationary behaviour of a stable system. We derive structural properties of the steady state distribution which provide insights into the equilibrium behaviour of the system. An explicit expression of the complete stationary distribution is still an open problem. Inspection of the literature shows that up to now there is no result available which provides rigorously proved conditions for stability in the case that both customer classes have unbounded queues. Our main result is therefore to derive a sufficient condition for ergodicity by the Foster-Lyapunov stability criterion. For a subclass of the admission policies we show that this condition is necessary as well. We expect that this is in general not the case. Furthermore, we consider the case of instant service, where we determine the stationary distribution explicitly.

The paper is organised as follows. In Sect. 2 we describe the related literature. In Sect. 3 we introduce our integrated model for production and inventory management and provide a Markov process Z for the description of the time evolution of the integrated system. In Sect. 4.1 ergodicity is investigated in detail. In Sect. 4.2 we assume that the queueing-inventory process Z is ergodic to analyse the properties of the stationary system. In Sect. 5 we consider the case of zero service time. In Sect. 6 we summarise our main findings and indicate further research directions.

2 Related literature and own contributions

Our research is connected to various parts of queueing theory and inventory control and their interplay in queueing-inventory models. Our fundamental production system is a classical $M/M/1/\infty$ queueing system, see e.g. Wolff [1989, Chapter 5]. The specific features added to the $M/M/1/\infty$ in this article are the following:

- Service of a customer needs a piece of raw material available in an associated inventory (for a review of research on these models see e.g. Krishnamoorthy et al. (2021)).
- Arriving customers have different priorities which result in a complicated admission control problem (for fundamentals on priority queues see Jaiswal (1968)).

Although in our models the customers are of different priority classes,

- they require the same type of raw material stored in a single inventory and
- they experience the same service time distributions.

The first property means that the items of raw material in the inventory are exchangeable with respect to the requesting customers' services. This property is common in certain applications of e.g. maintenance and repair, see e.g. investigations by Ravid et al. (2013) and Daduna (1990) and references therein.

The second property is discussed and motivated by Wang et al. [2015, p. 733] who identify three main motivations for prioritization and argue that two of these lead to identical service times over classes:

- Different customers have different willingness to pay for the same product.
- Customers may require different products or services, where some of these products are more profitable than others.
- Different service levels may substantially affect long-term profitability.

In our model we follow their conclusion: "Modelling the effects of prioritization due to the first and third motivations can be achieved with identical service time distributions for different segments." [Wang et al. (2015), p. 733]

Literature on inventory theory related to our problems encompass the following problems and articles. Studies by Gruen et al. (2002) and Verhoef and Sloot (2006) analyse customers' behaviour in practice and show that in many retail settings most of the original demand can be considered to be lost in case of a stockout. For an overview of the literature on systems with lost sales we refer to Bijvank and Vis (2011). They present a classification scheme for the replenishment policies most often applied in literature and practice, and they review the proposed replenishment policies, including the base stock policy.

Tempelmeier [2005, p. 84] argued that base stock control is economically reasonable if the order quantity is limited because of technical reasons. The base stock policy is "(...) more suitable for item with low demand, including the case of most spare parts" Rego and Mesquita [2011, p. 661].

Morse [1958, p. 9] investigated (pure) inventory systems that operate under a base stock policy. He gives a very simple example where the concept "re-order for each item sold" is useful: Items in inventory are bulky, and expensive (automobiles or TV sets[1]). He uses queueing theory to model the inventory systems, analogously to e.g. Reed and Zhang (2017).

Using queueing-theoretical methods to solve inventory models is well established, often under the heading of "production-inventory systems". In this setting the production usually refers to the replenishment systems, early references are e.g. Morse (1958), Karush (1957) (lost sales), Kaplan (1970) (backordering).

More recent examples are Rubio and Wein (1996), Lee and Zipkin (1992, 1995), Zazanis (1994), and Li and Arreola-Risa (2021) who investigated classical single item and multi-item inventory systems. Especially they evaluate the performance of base stock control policies in complex situations.

For a review of literature on inventory control systems (= queueing-inventory systems with service time equal to 0, i.e. "instant service" Melikov et al. (2018a)) with multiple customer classes, we refer to Isotupa ([2011, Sect. 2, pp. 3ff.], [2015 Sect. 1, pp. 411ff.]) and Arslan et al. [2007, pp. 1486ff.]. It is differentiated between priority disciplines that regulate customer arrivals and priority disciplines that regulate customer services. We combine both features in our model.

Literature on integrated queueing-inventory models seemingly appeared only from 1992 on, see e.g. Sigman and Simchi-Levi (1992), Melikov and Molchanov (1992). For a recent extensive review we refer to Krishnamoorthy et al. (2021). Because the research described

[1] The paper is from 1958.

in the present article is on queueing-inventory systems with priority classes for customers we concentrate here on articles which deal with priority problems in queueing-inventory systems.

Importance of research on this topic is emphasized by e.g. Yadavalli et al. (2015) who remarked that patients with serious illnesses are given priority over patients opting for routine checks or else in multi-speciality hospitals. Similarly, Liu et al. [2013, pp. 1544f.] stated that orders with long term contracts have higher priority than unscheduled orders since they may bear lower shortage cost than the booked orders.

To the best of our knowledge, queueing-inventory systems (under the heading "inventory in counter-stream serving systems") with different classes of customers were first considered by Melikov and Fatalieva (1998) who formulated a Markov decision model to minimise a cost function which encompasses costs for waiting, inventory holding, loss of demand and for dispatching items. In this problem setting there is no direct prioritization of classes.

Zhao and Lian (2011) seemingly were the first to investigate a system with Poisson arrivals of different priority classes, exponentially distributed service and lead times under back-ordering. They found a priority service rule to minimise the long-run expected waiting cost by a dynamic programming method. They formulate the model as a level-dependent quasi-birth-and-death process such that the steady state probability distribution of their production-inventory systems can be computed by the Bright-Taylor algorithm.

Chen et al. (2012) and Liu et al. (2013, 2014) introduce a flexible admission control with a priority parameter $0 \leq p \leq 1$ "for controlling the application of priority" [Liu et al. (2014), p. 181]. The priority parameter p indicates the probability with which the arrivals of ordinary customers are treated like the arrivals of priority customers. If $p = 1$, there is no priority in regulating arrivals. If $p = 0$, there is a strict priority in regulating arrivals. In the last case, their model is the same as the model of Isotupa (2007). They derive the stationary distribution of the inventory levels and some performance measures. To obtain the optimal inventory control policy they construct a mixed integer optimization problem in Liu et al. (2013, 2014) and develop an efficient searching algorithm in Chen et al. (2012).

Similar randomized differentiated admission control is incorporated in many models and is discussed in-depth by Isotupa in a sequence of papers (Isotupa (2006, 2007, 2011, 2015); Isotupa and Samanta (2013)). Isotupa investigates inventory systems with (r, Q)- and base stock policy for two classes of customers. Especially, the lost sales property for ordinary customers is of interest, e.g. for inventories of spare parts in the airline or shipping industries, (cf. Isotupa [2011, pp. 1f.]).

Jeganathan and coauthors investigate in a sequence of papers (e.g. Jeganathan et al. (2016), Jeganathan (2015), Yadavalli et al. (2015)) queueing-inventory systems with two classes of customers. They consider models with impatient customers, an optional second service and a mixed priority service (non-preemptive priority and preemptive priority).

Li and Zhao (2009) investigate a preemptive priority queueing system (without inventory) with two classes of customers and an exponential single server who serves the two classes of customers at potentially different rates.

Krishnamoorthy and Manjunath (2015) investigate a two-queue service system (without inventory) where customers arriving in a Poisson stream enter the high-priority queue. An ongoing service of the customers in the high-priority queue may be interrupted and the interrupted customer is sent to the low-priority queue. Service times are priority dependent exponential. In Krishnamoorthy and Manjunath (2018) customers arriving in a marked Poisson process enter either the high or the low priority queue. Furthermore, after being served at the high priority queue the customers may decide to ask for an additional service at the low priority queue. Service times are phase-type distributed.

Melikov et al. (2018a, b) investigate variants of queueing-inventory systems with demand of two classes with high and low priority. They consider an admission control scheme which allows access for high priority customers in any case (backordering) and rejects low priority customers when the stock level is below a critical value. Using approximation methods they evaluate various performance measures of the systems.

Following the authors' description the papers of Melikov et al. (2022a, b) are (at a first glance) not on priority systems. But after a reinterpretation the models fit into our modelling framework. The authors consider a typical (exponential) queueing-inventory system with some additional features e.g. a special reorder policy with the possibility of emergency reorders, randomized admission in case of stockout, impatient customers. Using the terms of priority systems these customers are those having low priority. Additionally, there is a second incoming Poisson stream of (very) high priority customers which have interrupting priority over the low priority customers. The service time of the high priority customers are negligible (= 0) and in case of stockout they immediately depart (lost sales). In terms of Melikov et al. (2022a, b) the "high priority customers" are "destructive customers" which describe sudden deterioration of items in the inventory which resemble negative customers in pure queueing systems (cf. Gelenbe (1991), Gelenbe et al. (1991)).

The problem of stability for queueing-inventory systems with priority classes of demand has not found much interest in the literature. Most of the papers with priorities in queueing-inventory systems use state space truncation, i.e. consider finite waiting rooms for tackling the stability problem, either both customer queues are assumed to be finite (then stability is not an issue) or one of the queues is truncated (most often the low priority queue, but in some articles high priority customers never queue up). Then the system fits into the realm of QBD processes. Some representative examples are
for finite state space: Yadavalli et al. (2015), Liu et al. (2013), Chen et al. (2012) (service time = 0), Jeganathan (2015), Jeganathan et al. (2013, 2016), Melikov et al. (2018a), Wang (2015),
for countable state space: Shajin et al. (2020) and Baek et al. (2017) (only a queue for low priority customers), Melikov et al. (2018a, b) (steady state distribution obtained with approximation methods), (Zhao & Lian, 2011) (heuristic/intuitive criterion for ergodicity for two demand classes, extending the single demand class case of Schwarz et al. (2006)), Melikov et al. (2022b, a) (no queue for the "very high priority" customers = destructive customers).
The conclusion is that (best to our knowledge) there exists up to now no rigorous result (criterion) for stability of queueing-inventory systems with differentiated priorities for demand classes in case of state spaces which are two-dimensional infinite, i.e. including $\mathbb{N}_0^2$, counting for high and low priority customers (demands). This observation is not surprising because in a general context this problem can be termed as "ergodicity of random walks in the quarter plane" ($\mathbb{N}_0^2$), which is known to be a notoriously hard problem, see Fayolle et al. (1999). Even more, in the present problem setting this random walk is influenced by the additional (finite) dimension of the inventory. So our investigation is on ergodicity of random walks in the quarter plane in a random environment. A very recent investigation of such problems is from Dimitriou (2022).

Our main contributions are the following:
Starting from a Markovian description for a queueing-inventory system with two unbounded queues for customers of different priority classes and flexible admission control we derive (rigorously) sufficient conditions which guarantee stability of the system. For special cases

we even show that the condition is sharp (necessary for stability). Although the complete stationary distribution seems to be an open problem we are able to derive several partial balance properties of the system which are of independent interest. Furthermore, we consider the special case of zero service time and determine the stationary distribution explicitly.

Notations and Conventions:

- $\mathbb{N} := \{1, 2, 3, \ldots\}$, $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$.
- The notation $\subset$ between sets means "subset or equal" and $\subsetneq$ means "proper subset". For a set A we denote by $|A|$ the number of elements in A.
- $1_{\{expression\}}$ is the indicator function which is 1 if $expression$ is true and 0 otherwise.
- Empty sums are 0, and empty products are 1.
- All random variables are defined on a common probability space $(\Omega, \mathcal{F}, P)$.
- By Markov process we mean time-homogeneous continuous-time strong Markov process with discrete state space (= Markov jump process). Markov processes occurring are assumed to be regular and having paths which are right-continuous with left limits (cadlag). A Markov process is regular if it is non-explosive (i.e. the sequence of jump times of the process diverges almost surely), its transition intensity matrix is conservative (i.e. row sums are 0) and stable (i.e. all diagonal elements of the transition intensity matrix are finite).

3 Description of the model

The supply chain of interest is depicted in Fig. 1 and consists of two arrival streams of priority and ordinary customers, a production system (a single server with two unlimited waiting rooms), an inventory and a supplier.

The production system manufactures units according to customers' demand on a make-to-order basis. There are two types of customers—priority customers (type 1) and ordinary customers (type 2). $\overline{C} = \{1, 2\}$ is the set of customer classes. Priority customers arrive

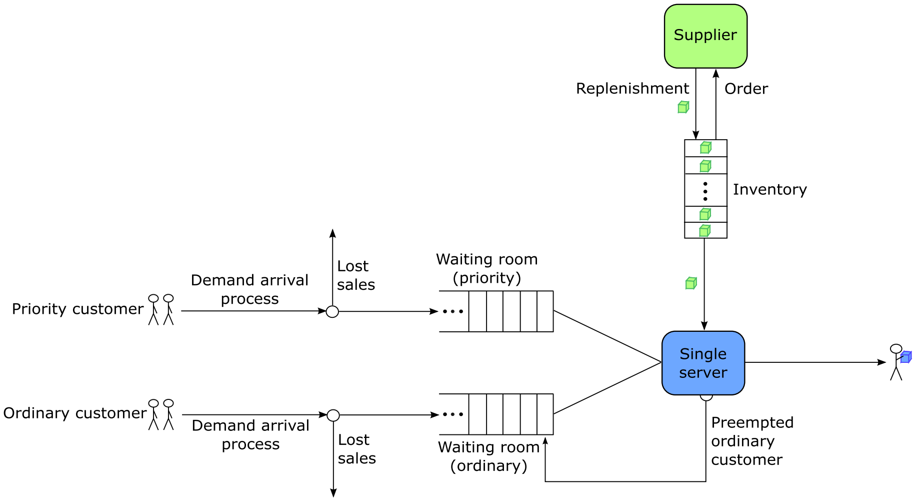

Fig. 1 The production-inventory system with two customer classes

according to a Poisson process with rate $\lambda_1 > 0$ and ordinary customers arrive according to a Poisson process with rate $\lambda_2 > 0$.

Each customer needs exactly one item from the inventory for service. The service time for both types of customers is exponentially distributed with parameter $\mu > 0$. If the server is ready to serve a customer, who is at the head of the line, and the inventory is not depleted, the service begins immediately. Otherwise, the service starts at the instant of time when the next replenishment arrives at the inventory. A served customer departs from the system immediately and the associated item is removed from the inventory at this time instant.

An outside supplier replenishes raw material to the inventory according to a base-stock policy. Hence, each item taken from the inventory results in a direct order sent to the supplier. This means, if a served customer departs from the system, an order for one item of the consumed raw material is placed at the supplier at this instant of time. The base stock level $b \geq 2$ is the maximal size of the inventory. The replenishment lead time is exponentially distributed with parameter $\nu > 0$. (Note that there can be more than one outstanding order.) Customers' arrivals are regulated by a flexible admission control with priority parameter p, $0 \leq p \leq 1$: If the inventory is depleted all arriving customers are rejected ("lost sales"). If the on-hand inventory is greater than a prescribed threshold level s, $0 < s < b$, customers of both classes are admitted to enter the system. If the on-hand inventory reaches or falls below the threshold level s, priority customers still enter the system but ordinary customers are allowed to enter only with probability p and are rejected with probability $1 - p$.

There is a single server with two separate infinite waiting rooms—one waiting room for priority customers (priority queue) and one waiting room for ordinary customers (ordinary queue) both under an FCFS regime. If both customer queues are not empty, the server needs to decide which one of them should be served. The choice is made according to the preemptive resume discipline. An overview of various priority disciplines can be found in Jaiswal [1968, p. 53].

According to the preemptive resume discipline a newly arriving priority customer interrupts immediately an ongoing service of an ordinary customer. The preempted ordinary customer is put at the head of the ordinary customer queue and has to wait until the priority queue is exhausted before he re-enters service. The preempted customer resumes service from the point of interruption so that his service time upon re-entry has been reduced by the amount of time the customer has already spent in service (cf. Miller [1958, p. 1]). Since it is assumed that the service time is exponential, the ordinary customer requires on its re-entry stochastically the same amount of service as it required on its earlier entry. Thus, the preemptive resume discipline is equal to the preemptive repeat-identical discipline where the preempted customer requires the same amount of service on its re-entry as he required on his earlier entry (cf. Jaiswal [1968, p. 53]).

It is assumed that transmission times for orders are negligible and set to zero and that the transportation time between the production system and the inventory is negligible. All service times, inter-arrival times and replenishment lead times constitute an independent family of random variables.

A Markovian process description of the integrated queueing-inventory system is obtained as follows. Denote by $X_1(t)$ the number of priority customers present in the system at time $t \geq 0$, and by $X_2(t)$ the number of ordinary customers in the system at time $t \geq 0$. So $X_j(t)$, $j = 1, 2$, counts the respective number of customers either waiting or in service. Since the customer in service will always be of the priority class when at least one priority customer is present, the value of the vector $(X_1(t), X_2(t))$ determines uniquely the type of the customer in service at time $t \geq 0$, if any. By $Y(t)$ we denote the on-hand inventory at time $t \geq 0$.

We define the joint queueing-inventory process of this system by

$$Z = ((X_1(t), X_2(t), Y(t)) : t \geq 0).$$

Then, due to the usual independence and memoryless assumptions Z is a homogeneous Markov process. The state space of Z is

$$E = \left\{(n_1, n_2, k) : (n_1, n_2) \in \mathbb{N}_0^2,\ k \in \{0, \ldots, b\}\right\},$$

where b is the maximal size of the inventory. Z has an infinitesimal generator $\mathbf{Q} = (q(z; \tilde{z}) : z, \tilde{z} \in E)$ with the following transition rates for $(n_1, n_2, k) \in E$:

$$\begin{aligned}
q((n_1, n_2, k); (n_1 + 1, n_2, k)) &= \lambda_1 \cdot 1_{\{k>0\}}, \\
q((n_1, n_2, k); (n_1, n_2 + 1, k)) &= p\,\lambda_2 \cdot 1_{\{0<k\leq s\}} + \lambda_2 \cdot 1_{\{k>s\}}, \\
q((n_1, n_2, k); (n_1 - 1, n_2, k - 1)) &= \mu \cdot 1_{\{n_1>0\}} \cdot 1_{\{k>0\}}, \\
q((n_1, n_2, k); (n_1, n_2 - 1, k - 1)) &= \mu \cdot 1_{\{n_1=0\}} \cdot 1_{\{n_2>0\}} \cdot 1_{\{k>0\}}, \\
q((n_1, n_2, k); (n_1, n_2, k + 1)) &= \nu \cdot 1_{\{k<b\}}.
\end{aligned}$$

Furthermore, $q(z; \tilde{z}) = 0$ for any other pair $z \neq \tilde{z}$, and

$$q\,(z; z) = -\sum_{\substack{\tilde{z}\in E,\\ z\neq\tilde{z}}} q\,(z; \tilde{z}) \qquad \forall z \in E.$$

Definition 1 If the queueing-inventory process Z on state space E is ergodic, we denote its limiting and stationary distribution by

$$\begin{aligned}
\pi &:= (\pi\,(n_1, n_2, k) : (n_1, n_2, k) \in E) \\
&\quad \text{with} \quad \pi\,(n_1, n_2, k) := \lim_{t\to\infty} P\,(Z(t) = (n_1, n_2, k))\,.
\end{aligned}$$

Discussion of modelling assumptions

(1) Our admission control is a modification of the admission schemes described in the related literature in Sect. 2. The strict rejection of ordinary customers below the control limit s is weakened by a randomized decision. If the inventory level is zero, demands due to both types of customers are lost, cf. Isotupa (2007, 2011, 2015).

(2) Preemptive resume regime of priority customers over ordinary customers is a common scheme in queueing models, c.f. Jaiswal (1968). For queueing-inventory systems coupling each service with obtaining an item from the inventory implies that an item dedicated at first to an ordinary customer can be attached to an interrupting priority customer as well. Here the property that items from stock are exchangeable is necessary.

4 Stability and stationary behaviour

In this section we investigate stability of the queueing-inventory system. This especially means to find conditions on the system which guarantee that the system approaches in the long run a steady state (i.e. stabilises). For queueing systems with priority classes of customers this is a non-trivial problem because in general (e.g. in our setting of a queue without inventory) a solution of the global balance equations for the stationary distribution is not available and the standard ergodicity criterion for QBDs (Quasi-Birth-Death processes) is not applicable, see e.g. Latouche and Ramaswami (1999).

In Sect. 4.1 we find natural conditions for Z to be ergodic by constructing a suitable Lyapunov function in the spirit of the classical Foster-Lyapunov stability criterion. Foster (1953) introduced a technique for proving stability of Markov chains which was generalised in several directions. Our proof relies on Kelly and Yudovina [2014, Proposition D.3].

Proposition 1 *Let $X := (X(t) : t \geq 0)$ be an irreducible regular Markov process with countable state space E and transition rate matrix $\mathbf{Q} := (q(x, y) : x, y \in E)$. Suppose that $\mathcal{L} : E \to [0, \infty)$ is a function such that for constants $\varepsilon > 0$ and $b \in \mathbb{R}$, and some finite exception set $F \subset E$ and all $x \in E$ it holds*

$$\sum_{y \in E \setminus \{x\}} q(x, y) \, [\mathcal{L}(y) - \mathcal{L}(x)] \leq \begin{cases} -\varepsilon & x \notin F, \\ b - \varepsilon & x \in F. \end{cases} \tag{1}$$

Then X is ergodic.

Although we do not find a solution of the global balance equations for Z in Sect. 4.2, we present interesting results of the behaviour of Z in steady state. These results rely on the presence of partial balance properties inherent in the dynamics of the queueing-inventory system in equilibrium. Some of the resulting properties are surprising.

4.1 Ergodicity

Irreducibility of the state process Z can be seen directly from the transition rates at the end of Sect. 3. Our main result is the following theorem in the spirit of the Foster-Lyapunov stability criterion.

Theorem 2 *The queueing-inventory process Z is ergodic if $\lambda_1 + \lambda_2 < \mu$.*

Proof Positive recurrence will be shown by the Foster-Lyapunov stability criterion with $\mathcal{L} : E \to \mathbb{R}_0^+$ as Lyapunov function with

$$\mathcal{L}(n_1, n_2, k) := n_1 + n_2 + \alpha(k), \quad (n_1, n_2, k) \in E, \tag{2}$$

where $\alpha : \{0, 1, \ldots, b\} \to [0, \infty)$ is strictly decreasing with

$$\alpha(k) = (b - k) \cdot \frac{\mu - \lambda_1 - \lambda_2}{2\mu} \tag{3}$$

and the finite exception set is given by

$$F := \{(n_1, n_2, k) : n_1 + n_2 = 0\}.$$

Furthermore, we define

$$\eta := \mu - \lambda_1 - \lambda_2 \tag{4}$$

and

$$\varepsilon := \frac{\eta}{2} \cdot \min\left\{1, \frac{\nu}{\mu}\right\}.$$

Due to the assumption, that $\mu > \lambda_1 + \lambda_2$ it holds $\varepsilon > 0$.

► First, we will check $(\mathbf{Q} \cdot \mathcal{L})\,(n_1, n_2, k) < \infty$ for $(n_1, n_2, k) \in F$.
Since $0 < \lambda_1 < \infty$, $0 < \lambda_2 < \infty$ and $0 < \nu < \infty$,
for $k = 0$ it holds

$$(\mathbf{Q} \cdot \mathcal{L})\,(0,0,0) = \nu \cdot (\mathcal{L}(0,0,1) - \mathcal{L}(0,0,0)) \overset{(2)}{=} \nu \cdot (\alpha(1) - \alpha(0)) < \infty,$$

for $k = 1, \ldots, s$ it holds

$$\begin{aligned}(\mathbf{Q} \cdot \mathcal{L})\,(0,0,k) &= \lambda_1 \cdot (\mathcal{L}(1,0,k) - \mathcal{L}(0,0,k)) + p\,\lambda_2 \cdot (\mathcal{L}(0,1,k) - \mathcal{L}(0,0,k)) \\ &\quad + \nu \cdot (\mathcal{L}(0,0,k+1) - \mathcal{L}(0,0,k)) \\ &\overset{(2)}{=} \lambda_1 \cdot [(1 + \alpha(k)) - \alpha(k)] + p\,\lambda_2 \cdot [(1 + \alpha(k)) - \alpha(k)] \\ &\quad + \nu \cdot (\alpha(k+1) - \alpha(k)) < \infty,\end{aligned}$$

for $k = s+1, \ldots, b-1$ it holds

$$\begin{aligned}(\mathbf{Q} \cdot \mathcal{L})\,(0,0,k) &= \lambda_1 \cdot (\mathcal{L}(1,0,k) - \mathcal{L}(0,0,k)) + \lambda_2 \cdot (\mathcal{L}(0,1,k) - \mathcal{L}(0,0,k)) \\ &\quad + \nu \cdot (\mathcal{L}(0,0,k+1) - \mathcal{L}(0,0,k)) \\ &\overset{(2)}{=} \lambda_1 \cdot [(1 + \alpha(k)) - \alpha(k)] + \lambda_2 \cdot [(1 + \alpha(k)) - \alpha(k)] \\ &\quad + \nu \cdot (\alpha(k+1) - \alpha(k)) < \infty,\end{aligned}$$

for $k = b$ it holds

$$\begin{aligned}(\mathbf{Q} \cdot \mathcal{L})\,(0,0,b) &= \lambda_1 \cdot (\mathcal{L}(1,0,b) - \mathcal{L}(0,0,b)) + \lambda_2 \cdot (\mathcal{L}(0,1,b) - \mathcal{L}(0,0,b)) \\ &\overset{(2)}{=} \lambda_1 \cdot [(1 + \alpha(b)) - \alpha(b)] + \lambda_2 \cdot [(1 + \alpha(b)) - \alpha(b)] < \infty.\end{aligned}$$

► Second, we will check $(\mathbf{Q} \cdot \mathcal{L})\,(n_1, n_2, k) \le -\varepsilon$ for $z = (n_1, n_2, k) \notin F$ with

$$-\varepsilon = -\frac{\eta}{2} \cdot \max\left\{1, \frac{\nu}{\mu}\right\}. \tag{5}$$

For $k = 0$, $n_1 \in \mathbb{N}$ and $n_2 \in \mathbb{N}_0$ it holds

$$\begin{aligned}(\mathbf{Q} \cdot \mathcal{L})\,(n_1, n_2, 0) &= \nu \cdot (\mathcal{L}(n_1, n_2, 1) - \mathcal{L}(n_1, n_2, 0)) \\ &\overset{(2)}{=} \nu \cdot ((n_1 + n_2 + \alpha(1)) - (n_1 + n_2 + \alpha(0))) = \nu \cdot (\alpha(1) - \alpha(0)) \\ &\overset{(3)}{=} \nu \cdot [(b-1) - (b-0)] \cdot \frac{\mu - \lambda_1 - \lambda_2}{2\mu} \overset{(4)}{=} -\frac{\eta}{2} \cdot \frac{\nu}{\mu} \overset{(5)}{\le} -\varepsilon.\end{aligned}$$

For $k = 0$, $n_1 = 0$ and $n_2 \in \mathbb{N}$ it holds

$$\begin{aligned}(\mathbf{Q} \cdot \mathcal{L})\,(0, n_2, 0) &= \nu \cdot (\mathcal{L}(0, n_2, 1) - \mathcal{L}(0, n_2, 0)) \\ &\overset{(2)}{=} \nu \cdot ((n_2 + \alpha(1)) - (n_2 + \alpha(0))) = \nu \cdot (\alpha(1) - \alpha(0)) \\ &\overset{(3)}{=} \nu \cdot [(b-1) - (b-0)] \cdot \frac{\mu - \lambda_1 - \lambda_2}{2\mu} \overset{(4)}{=} -\frac{\eta}{2} \cdot \frac{\nu}{\mu} \overset{(5)}{\le} -\varepsilon.\end{aligned}$$

For $k = 1, \ldots, s$, $n_1 \in \mathbb{N}$ and $n_2 \in \mathbb{N}_0$ it holds

$$\begin{aligned}&(\mathbf{Q} \cdot \mathcal{L})\,(n_1, n_2, k) \\ &= \lambda_1 \cdot (\mathcal{L}(n_1 + 1, n_2, k) - \mathcal{L}(n_1, n_2, k)) + p\,\lambda_2 \cdot (\mathcal{L}(n_1, n_2 + 1, k) - \mathcal{L}(n_1, n_2, k))\end{aligned}$$

$$
\begin{aligned}
&+\mu\cdot(\mathcal{L}(n_1-1,n_2,k-1)-\mathcal{L}(n_1,n_2,k))+\nu\cdot(\mathcal{L}(n_1,n_2,k+1)-\mathcal{L}(n_1,n_2,k))\\
&\overset{(2)}{=}\lambda_1\cdot(n_1+1+n_2+\alpha(k)-n_1-n_2-\alpha(k))\\
&\quad+p\,\lambda_2\cdot(n_1+n_2+1+\alpha(k)-n_1-n_2-\alpha(k))\\
&\quad+\mu\cdot(n_1-1+n_2+\alpha(k-1)-n_1-n_2-\alpha(k))\\
&\quad+\nu\cdot(n_1+n_2+\alpha(k+1)-n_1-n_2-\alpha(k))\\
&\overset{(3)}{=}\lambda_1+p\,\lambda_2-\mu+\mu\cdot[(b-(k-1))-(b-k)]\cdot\frac{\mu-\lambda_1-\lambda_2}{2\mu}\\
&\quad+\nu\cdot[(b-(k+1))-(b-k)]\cdot\frac{\mu-\lambda_1-\lambda_2}{2\mu}\\
&=\lambda_1+p\,\lambda_2-\mu-\frac{\mu-\lambda_1-\lambda_2}{\mu}-\frac{\mu-\lambda_1-\lambda_2}{2\mu}\cdot\nu\overset{(4)}{\leq}-\eta+\frac{\eta}{2}-\frac{\eta}{2}\cdot\frac{\nu}{\mu}\overset{(5)}{\leq}-\varepsilon.
\end{aligned}
$$

For $k=1,\dots,s$, $n_1=0$ and $n_2\in\mathbb{N}$ it holds

$$
\begin{aligned}
&(\mathbf{Q}\cdot\mathcal{L})\,(0,n_2,k)\\
&=\lambda_1\cdot(\mathcal{L}(1,n_2,k)-\mathcal{L}(0,n_2,k))+p\,\lambda_2\cdot(\mathcal{L}(0,n_2+1,k)-\mathcal{L}(0,n_2,k))\\
&\quad+\mu\cdot(\mathcal{L}(0,n_2-1,k-1)-\mathcal{L}(0,n_2,k))+\nu\cdot(\mathcal{L}(0,n_2,k+1)-\mathcal{L}(0,n_2,k))\\
&\overset{(2)}{=}\lambda_1\cdot(1+n_2+\alpha(k)-n_2-\alpha(k))+p\,\lambda_2\cdot(n_2+1+\alpha(k)-n_2-\alpha(k))\\
&\quad+\mu\cdot(n_2-1+\alpha(k-1)-n_2-\alpha(k))+\nu\cdot(n_2+\alpha(k+1)-n_2-\alpha(k))\\
&=\lambda_1+p\,\lambda_2-\mu+\mu\cdot(\alpha(k-1)-\alpha(k))+\nu\cdot(\alpha(k+1)-\alpha(k))\\
&\overset{(3)}{=}\lambda_1+p\,\lambda_2-\mu+\mu\cdot[(b-(k-1))-(b-k)]\cdot\frac{\mu-\lambda_1-\lambda_2}{2\mu}\\
&\quad+\nu\cdot[(b-(k+1))-(b-k)]\cdot\frac{\mu-\lambda_1-\lambda_2}{2\mu}\\
&\overset{(4)}{\leq}-\eta+\frac{\eta}{2}-\frac{\eta}{2}\cdot\frac{\nu}{\mu}\overset{(5)}{\leq}-\varepsilon.
\end{aligned}
$$

For $k=s+1,\dots,b-1$, $n_1\in\mathbb{N}$ and $n_2\in\mathbb{N}_0$ it holds

$$
\begin{aligned}
&(\mathbf{Q}\cdot\mathcal{L})\,(n_1,n_2,k)\\
&=\lambda_1\cdot(\mathcal{L}(n_1+1,n_2,k)-\mathcal{L}(n_1,n_2,k))+\lambda_2\cdot(\mathcal{L}(n_1,n_2+1,k)-\mathcal{L}(n_1,n_2,k))\\
&\quad+\mu\cdot(\mathcal{L}(n_1-1,n_2,k-1)-\mathcal{L}(n_1,n_2,k))+\nu\cdot(\mathcal{L}(n_1,n_2,k+1)-\mathcal{L}(n_1,n_2,k))\\
&\overset{(2)}{=}\lambda_1\cdot(n_1+1+n_2+\alpha(k)-n_1-n_2-\alpha(k))\\
&\quad+\lambda_2\cdot(n_1+n_2+1+\alpha(k)-n_1-n_2-\alpha(k))\\
&\quad+\mu\cdot(n_1-1+n_2+\alpha(k-1)-n_1-n_2-\alpha(k))\\
&\quad+\nu\cdot(n_1+n_2+\alpha(k+1)-n_1-n_2-\alpha(k))\\
&=\lambda_1+\lambda_2-\mu+\mu\cdot(\alpha(k-1)-\alpha(k))+\nu\cdot(\alpha(k+1)-\alpha(k))\\
&\overset{(3)}{=}\lambda_1+\lambda_2-\mu+\mu\cdot[(b-(k-1))-(b-k)]\,\frac{\mu-\lambda_1-\lambda_2}{2\mu}\\
&\quad+\nu\cdot[(b-(k+1))-(b-k)]\,\frac{\mu-\lambda_1-\lambda_2}{2\mu}
\end{aligned}
$$

$$\overset{(4)}{=} -\eta + \frac{\eta}{2} - \frac{\eta}{2} \cdot \frac{\nu}{\mu} \overset{(5)}{\leq} -\varepsilon.$$

For $k = s+1, \ldots, b-1$, $n_1 = 0$ and $n_2 \in \mathbb{N}$

$$\begin{aligned}
&(\mathbf{Q} \cdot \mathcal{L})\,(0, n_2, k) \\
&= \lambda_1 \cdot (\mathcal{L}(1, n_2, k) - \mathcal{L}(0, n_2, k)) + \lambda_2 \cdot (\mathcal{L}(0, n_2+1, k) - \mathcal{L}(0, n_2, k)) \\
&\quad + \mu \cdot (\mathcal{L}(0, n_2-1, k-1) - \mathcal{L}(0, n_2, k)) + \nu \cdot (\mathcal{L}(0, n_2, k+1) - \mathcal{L}(0, n_2, k)) \\
&\overset{(2)}{=} \lambda_1 \cdot (1 + n_2 + \alpha(k) - n_2 - \alpha(k)) + \lambda_2 \cdot (n_2 + 1 + \alpha(k) - n_2 - \alpha(k)) \\
&\quad + \mu \cdot (n_2 - 1 + \alpha(k-1) - n_2 - \alpha(k)) + \nu \cdot (n_2 + \alpha(k+1) - n_2 - \alpha(k)) \\
&= \lambda_1 + \lambda_2 - \mu + \mu \cdot (\alpha(k-1) - \alpha(k)) + \nu \cdot (\alpha(k+1) - \alpha(k)) \\
&\overset{(3)}{=} \lambda_1 + \lambda_2 - \mu + \mu \cdot [(b - (k-1)) - (b-k)]\, \frac{\mu - \lambda_1 - \lambda_2}{2\mu} \\
&\quad + \nu \cdot [(b - (k+1)) - (b-k)]\, \frac{\mu - \lambda_1 - \lambda_2}{2\mu} \\
&\overset{(4)}{=} -\eta + \frac{\eta}{2} - \frac{\eta}{2} \cdot \frac{\nu}{\mu} \overset{(5)}{\leq} -\varepsilon.
\end{aligned}$$

For $k = b$, $n_1 \in \mathbb{N}$ and $n_2 \in \mathbb{N}_0$ it holds

$$\begin{aligned}
&(\mathbf{Q} \cdot \mathcal{L})\,(n_1, n_2, b) \\
&= \lambda_1 \cdot (\mathcal{L}(n_1+1, n_2, b) - \mathcal{L}(n_1, n_2, b)) + \lambda_2 \cdot (\mathcal{L}(n_1, n_2+1, b) - \mathcal{L}(n_1, n_2, b)) \\
&\quad + \mu \cdot (\mathcal{L}(n_1-1, n_2, b-1) - \mathcal{L}(n_1, n_2, b)) \\
&\overset{(2)}{=} \lambda_1 \cdot (n_1 + 1 + n_2 + \alpha(b) - n_1 - n_2 - \alpha(b)) \\
&\quad + \lambda_2 \cdot (n_1 + n_2 + 1 + \alpha(b) - n_1 - n_2 - \alpha(b)) \\
&\quad + \mu \cdot (n_1 - 1 + n_2 + \alpha(b-1) - n_1 - n_2 - \alpha(b)) \\
&\overset{(3)}{=} \lambda_1 + \lambda_2 - \mu + \mu \cdot ((b - (b-1)) - (b-b)) \cdot \frac{\mu - \lambda_1 - \lambda_2}{2\mu} \\
&\overset{(4)}{=} -\eta + \frac{\eta}{2} \overset{(5)}{\leq} -\varepsilon.
\end{aligned}$$

For $k = b$, $n_1 = 0$ and $n_2 \in \mathbb{N}$ it holds

$$\begin{aligned}
&(\mathbf{Q} \cdot \mathcal{L})\,(0, n_2, b) \\
&= \lambda_1 \cdot (\mathcal{L}(1, n_2, b) - \mathcal{L}(0, n_2, b)) + \lambda_2 \cdot (\mathcal{L}(0, n_2+1, b) - \mathcal{L}(0, n_2, b)) \\
&\quad + \mu \cdot (\mathcal{L}(0, n_2-1, b-1) - \mathcal{L}(0, n_2, b)) \\
&\overset{(2)}{=} \lambda_1 \cdot (1 + n_2 + \alpha(b) - n_2 - \alpha(b)) + \lambda_2 \cdot (n_2 + 1 + \alpha(b) - n_2 - \alpha(b)) \\
&\quad + \mu \cdot (n_2 - 1 + \alpha(b-1) - n_2 - \alpha(b)) \\
&= \lambda_1 + \lambda_2 - \mu + \mu \cdot (\alpha(b-1) - \alpha(b)) \\
&\overset{(3)}{=} \lambda_1 + \lambda_2 - \mu + \mu \cdot ((b - (b-1)) - (b-b)) \cdot \frac{\mu - \lambda_1 - \lambda_2}{2\mu} \\
&\overset{(4)}{=} -\eta + \frac{\eta}{2} \overset{(5)}{\leq} -\varepsilon.
\end{aligned}$$

□

We do not expect that the result of the theorem is sharp for general $p \in [0, 1]$, i.e. the condition $\lambda_1 + \lambda_2 < \mu$ is in general only sufficient for stability. We have only

Theorem 3 *Consider the system from Theorem 2 for $p = 1$, i.e. low priority customers are admitted as long as the inventory is not depleted. Then Z is ergodic if and only if $\lambda_1 + \lambda_2 < \mu$ holds.*

The result of Theorem 3 at a first glance seems to be intuitive because no distinction is made between the arriving customers. The problem is that the influence of the interrupts of arrivals and service due to stockout is not accessible to intuition. The theorem states especially that the interrupts of arrivals and service balance over time. The proof of Theorem 3 is postponed to the next section because we need some preparations which we consider to be of independent interest.

4.2 Properties of the stationary system

In this section we assume that the queueing-inventory process Z is ergodic. So the limiting and stationary distribution π of Z (see Definition 1) exists but seems to be not accessible directly. Nevertheless we are able to present results on the stationary behaviour of Z. We will exploit the structural information inherent in the global balance equations of Z for π.

The global balance equations $\pi \cdot \mathbf{Q} = \mathbf{0}$ of the ergodic queueing-inventory process Z are for $(n_1, n_2, k) \in E$ given by

$$\begin{aligned}
&\pi(n_1, n_2, k) \cdot \big((\lambda_1 + p\,\lambda_2) \cdot 1_{\{0<k\le s\}} + (\lambda_1 + \lambda_2) \cdot 1_{\{k>s\}} \\
&\quad + \mu \cdot 1_{\{n_1+n_2>0\}} \cdot 1_{\{k>0\}} + \nu \cdot 1_{\{k<b\}}\big) \\
&= \pi(n_1 - 1, n_2, k) \cdot \lambda_1 \cdot 1_{\{n_1>0\}} \cdot 1_{\{k>0\}} \\
&\quad + \pi(n_1, n_2 - 1, k) \cdot p\,\lambda_2 \cdot 1_{\{n_2>0\}} \cdot 1_{\{0<k\le s\}} \\
&\quad + \pi(n_1, n_2 - 1, k) \cdot \lambda_2 \cdot 1_{\{n_2>0\}} \cdot 1_{\{k>s\}} \\
&\quad + \pi(n_1 + 1, n_2, k + 1) \cdot \mu \cdot 1_{\{k<b\}} + \pi(n_1, n_2 + 1, k + 1) \cdot \mu \cdot 1_{\{n_1=0\}} \cdot 1_{\{k<b\}} \\
&\quad + \pi(n_1, n_2, k - 1) \cdot \nu \cdot 1_{\{k>0\}}. \qquad (6)
\end{aligned}$$

Let (X_1, X_2, Y) be a random vector that is distributed according to the stationary distribution π. So, Y is distributed as the stationary distribution of the inventory process in equilibrium and X_1 resp. X_2 are respectively distributed as the processes counting for the number of priority resp. ordinary customers in equilibrium.

The proofs of the main results in this section rely on partial balance relations inherent in the global balance equations of Z. These relations are extracted and formulated in the next lemma.

Lemma 1 *For the queueing-inventory process Z it holds*

$$\begin{aligned}
&P(X_1 = n_1, Y > 0) \cdot \lambda_1 = P(X_1 = n_1 + 1, Y > 0) \cdot \mu, \quad n_1 \in \mathbb{N}_0, \qquad (7) \\
&P(X_2 = n_2, 0 < Y \le s) \cdot p\,\lambda_2 + P(X_2 = n_2, Y > s) \cdot \lambda_2 \\
&= P(X_1 = 0, X_2 = n_2 + 1, Y > 0) \cdot \mu, \quad n_2 \in \mathbb{N}_0, \qquad (8) \\
&P(X_1 + X_2 = n, 0 < Y \le s) \cdot (\lambda_1 + p\,\lambda_2) \\
&\quad + P(X_1 + X_2 = n, Y > s) \cdot (\lambda_1 + \lambda_2) \\
&= P(X_1 + X_2 = n + 1, 0 < Y \le b) \cdot \mu, \quad n \in \mathbb{N}_0. \qquad (9)
\end{aligned}$$

All the equations can be proven by applying the cut-criterion for positive recurrent processes (see Kelly [1979, Lemma 1.4, p. 8]). This is

Lemma 2 *For ergodic* Z *with stationary distribution* π *it holds for any non-empty proper subset* $A \subset E$

$$\sum_{z\in A}\sum_{\tilde z\in A^c}\pi(z)q(z,\tilde z)=\sum_{\tilde z\in A^c}\sum_{z\in A}\pi(\tilde z)q(\tilde z,z)$$

Proof of Lemma 1 For $n_1 \in \mathbb{N}_0$, equation (7) can be proven by a cut, which divides E into complementary sets according to the queue length of priority customers that is less than or equal to n_1 or greater than n_1. These are the sets

$$A := \Big\{(m_1, m_2, k) : m_1 \in \{0, 1, \dots .n_1\},\ m_2 \in \mathbb{N}_0,\ k \in \{0, \dots, b\}\Big\},$$

and A^c. Then for $n_1 \in \mathbb{N}_0$ it holds by direct evaluation

$$\underbrace{\sum_{m_1=n_1}^{n_1}\sum_{m_2=0}^{\infty}\sum_{k=1}^{b}\pi(m_1,m_2,k)\cdot\lambda_1}_{=P(X_1=n_1,Y>0)\cdot\lambda_1}=\underbrace{\sum_{\widetilde m_1=n_1+1}^{n_1+1}\sum_{\widetilde m_2=0}^{\infty}\sum_{\widetilde k=1}^{b}\pi(\widetilde m_1,\widetilde m_2,\widetilde k)\cdot\mu}_{=P(X_1=n_1+1,Y>0)\cdot\mu}.$$

Hence, for $n_1 \in \mathbb{N}_0$ it holds (7).

For $n_2 \in \mathbb{N}_0$, equation (8) can be proven by a cut, which divides E into complementary sets according to the queue length of ordinary customers that is less than or equal to n_2 or greater than n_2. These are the sets

$$A = \Big\{(m_1, m_2, k) : m_1 \in \mathbb{N}_0, m_2 \in \{0, 1, \dots .n_2\},\ k \in \{0, \dots, b\}\Big\},$$

and A^c. Then for $n_2 \in \mathbb{N}_0$ it holds by direct evaluation

$$\underbrace{\sum_{m_1=0}^{\infty}\sum_{m_2=n_2}^{n_2}\sum_{k=1}^{s}\pi(m_1,m_2,k)\cdot p\,\lambda_2}_{=P(X_2=n_2,0<Y\leq s)\cdot p\,\lambda_2}+\underbrace{\sum_{m_1=0}^{\infty}\sum_{m_2=n_2}^{n_2}\sum_{k=s+1}^{b}\pi(m_1,m_2,k)\cdot\lambda_2}_{=P(X_2=n_2,Y>s)\cdot\lambda_2}$$
$$=\underbrace{\sum_{\widetilde m_1=0}^{0}\sum_{\widetilde m_2=n_2+1}^{n_2+1}\sum_{\widetilde k=1}^{b}\pi(\widetilde m_1,\widetilde m_2,\widetilde k)\cdot\mu}_{=P(X_1=0,X_2=n_2+1,Y>0)\cdot\mu}.$$

Thus, for $n_2 \in \mathbb{N}_0$ it holds (8).

For $n \in \mathbb{N}_0$, equation (9) can be proven by a cut, which divides E into complementary sets according to the size of the total queue length that is less than or equal to n or greater than n. These are the sets

$$A = \Big\{(m_1, m_2, k) : m_1 \in \mathbb{N}_0,\ m_2 \in \mathbb{N}_0,\ (m_1 + m_2) \in \{0, 1, \dots, n\},\ k \in \{0, \dots, b\}\Big\},$$

and A^c. Then for $n \in \mathbb{N}_0$ holds by direct evaluation

$$\underbrace{\sum_{m_1+m_2=n}^{n}\sum_{k=1}^{s}\pi(m_1,m_2,k)\cdot(\lambda_1+p\,\lambda_2)}_{=P(X_1+X_2=n,0<Y\leq s)\cdot(\lambda_1+p\,\lambda_2)}+\underbrace{\sum_{m_1+m_2=n}^{n}\sum_{k=s+1}^{b}\pi(m_1,m_2,k)\cdot(\lambda_1+\lambda_2)}_{=P(X_1+X_2=n,Y>s)\cdot(\lambda_1+\lambda_2)}$$

$$= \underbrace{\sum_{\widetilde{m}_1+\widetilde{m}_2=n+1}^{n+1} \sum_{\widetilde{k}=1}^{b} \pi(\widetilde{m}_1, \widetilde{m}_2, \widetilde{k}) \cdot \mu}_{=P(X_1+X_2=n+1, Y>0)\cdot\mu} .$$

Thus, for $n \in \mathbb{N}_0$ it holds (9). □

Consider the queueing-inventory system with lead time zero for the replenishment orders. Then Y is constant b and the distribution of (X_1, X_2) is the equilibrium of the priority system as described in Sect. 3 without inventory. A little reflection shows that the distribution of X_1 is geometrical with parameter λ_1/μ, i.e. the priority customers are served as in a standard $M/M/1/\infty$ with parameters λ_1 and μ. (Indeed, this observation applies to a classical $M/M/c$ queue with two priority classes under a preemptive priority discipline as well. The priority customers behave as customers in a classic $M/M/c$ queue (cf. Wang et al. (2015)).)

Our first proposition embeds a similar observation into the setting of the model in Sect. 3. The problem with our queueing-inventory system is that the two customer classes have to share the same inventory. Therefore, the ordinary (= non-priority) customers impose restrictions on the priority customers' service because they consume items from the inventory and generate more stock-outs experienced by the priority customers.

Proposition 4 *For the queueing-inventory system from Sect. 3 the stationary number of priority customers in the system conditioned on positive inventory is geometrical with parameter λ_1/μ, i.e. for $n_1 \in \mathbb{N}_0$ it holds*

$$P(X_1 = n_1 \mid Y > 0) = P(X_1 = 0 \mid Y > 0) \cdot \left(\frac{\lambda_1}{\mu}\right)^{n_1}, \tag{10}$$

with normalisation constant

$$P(X_1 = 0 \mid Y > 0) = 1 - \frac{\lambda_1}{\mu}.$$

Proof From (7) it follows by iteration for $n_1 \in \mathbb{N}_0$

$$P(X_1 = n_1, Y > 0) = P(X_1 = 0, Y > 0) \cdot \left(\frac{\lambda_1}{\mu}\right)^{n_1}. \tag{11}$$

Conditioning and exploiting $\lambda_1 < \mu$ for summation ends the proof. □

A further consequence of Lemma 1 are the following intuitive rate equations. These equalize the mean admitted arrivals per customer classes to class dependent throughputs (= effective departure rates). The proof is in any case by direct summation of the respective formulas in Lemma 1.

Proposition 5 *For the queueing-inventory process it holds the following equilibrium of probability flows*

$$\underbrace{P(Y > 0) \cdot \lambda_1}_{\substack{\textit{effective departure rate of priority customers}\\ \textit{of priority customers}}} = \underbrace{P(X_1 > 0, Y > 0) \cdot \mu}_{\substack{\textit{effective departure rate of ordinary customers}\\ \textit{of priority customers}}}, \tag{12}$$

$$\underbrace{P(0 < Y \le s) \cdot p\lambda_2 + P(Y > s) \cdot \lambda_2}_{\substack{\textit{effective arrival rate}\\ \textit{of ordinary customers}}} = \underbrace{P(X_1 = 0, X_2 > 0, Y > 0) \cdot \mu}_{\substack{\textit{effective departure rate}\\ \textit{of ordinary customers}}}, \tag{13}$$

$$\underbrace{P(Y>0)\cdot\lambda_1+P(0<Y\le s)\cdot p\,\lambda_2+P(Y>s)\cdot\lambda_2}_{\substack{\textit{effective arrival rate}\\ \textit{of customers}}},$$

$$=\underbrace{P(X_1+X_2>0,Y>0)\cdot\mu}_{\substack{\textit{effective departure rate}\\ \textit{of customers}}} \tag{14}$$

Similar to the flow equations in Lemma 1 it is possible to derive flow equations with respect to the inventory level. The derivation is more tedious and can be found in Otten [2018, Proposition 11.1.6]. The main result is that for $k = 0, 1, \dots, b-1$ it holds

$$P(Y=k)\cdot\nu = P(Y=k+1, X_1+X_2>0)\cdot\mu. \tag{15}$$

Note that the statement of (15) exhibits an insensitivity property with respect to variation of the parameters of the system. More specifically it is independent of the threshold level s and the arrival intensities λ_1 and λ_2.

Remark 1 **(1)** The results in this section can be generalised in a direct way to the case of a system with C customer classes, where $\overline{C} = \{1,\dots,C\}$ is the set of customer classes. Customers of type c arrive with rate $\lambda_c > 0$, a priority parameter p_c $(0 \le p_c \le 1)$ and a threshold level s_c, $c \in \overline{C}$.
(2) In the models of Isotupa (2015), the replenishment rate depends on the number of pending orders. We can extend our model so that the replenishment lead time depends on the number of orders at the supplier. If there are $k > 0$ orders present at the supplier, the intensity of the replenishment lead time is $\nu(k) > 0$. For more details see Otten [2018, Chapter 11.1].

Proof of Theorem 3 Recall that the global balance equations for Z usually are stated with unknown positive $x = (x(z) : z \in E)$. In case of irreducibility (under Q) of E, Z is ergodic if and only if there exist a strictly positive solution x which is summable, i.e. $\sum_{z\in E} x(z) < \infty$. The first observation is that the partial balance relations of Lemma 1 hold for any solution x, even if the solution is not summable. We shall exploit equation (9) which reads in case of $p = 1$ for a general solution

$$\sum_{n_1+n_2=n}\left(\sum_{k=1}^{b} x(n_1,n_2,k)\right)\cdot(\lambda_1+\lambda_2)$$
$$=\sum_{n_1+n_2=n+1}\left(\sum_{k=1}^{b} x(n_1,n_2,k)\right)\cdot\mu, \qquad n\in\mathbb{N}_0\,.$$

It follows for $n \in \mathbb{N}_0$

$$\sum_{n_1+n_2=n}\left(\sum_{k=1}^{b} x(n_1,n_2,k)\right)=\left(\sum_{k=1}^{b} x(0,0,k)\right)\cdot\left(\frac{\lambda_1+\lambda_2}{\mu}\right)^{n}.$$

Assuming ergodicity of Z the summability condition yields

$$\infty > \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\sum_{k=0}^{b} x(n_1,n_2,k)$$
$$=\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\left(\sum_{k=1}^{b} x(n_1,n_2,k)\right)+\sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} x(n_1,n_2,0)$$

$$= \sum_{n=0}^{\infty} \left\{ \sum_{n_1+n_2=n} \left(\sum_{k=1}^{b} x(n_1, n_2, k) \right) \right\} + \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} x(n_1, n_2, 0)$$

$$= \sum_{n=0}^{\infty} \left\{ \left(\sum_{k=1}^{b} x(0, 0, k) \right) \cdot \left(\frac{\lambda_1 + \lambda_2}{\mu} \right)^n \right\} + \underbrace{\sum_{n_1=0}^{\infty} \sum_{n_2=n}^{\infty} x(n_1, n_2, 0)}_{\in(0,\infty)} .$$

Because of $\sum_{k=1}^{b} x(0, 0, k) \in (0, \infty)$ the last term is finite if and only if $\lambda_1 + \lambda_2 < \mu$ holds. This finishes the proof. □

5 The case of instant service

In this section we consider the case of instant service (zero production time), which turns the model into the formalism of classical inventory theory. The supply chain of interest is depicted in Fig. 2 and consists of priority and ordinary customers, an inventory and a supplier.

A short reflection shows that due to the lost sales assumption no customer queues will arise. Therefore, the on-hand inventory process $Y = (Y(t) : t \geq 0)$ carries all information for a Markovian description of the system, which is a slight generalisation of the pure inventory model in Isotupa (2015). Isotupa derives the stationary distribution but her model does not incorporate the priority parameter p. Chen et al. (2012), Liu et al. (2013, 2014) determine the stationary distribution for a pure inventory model with priority parameter but with a different replenishment policy.

The state space of Y is

$$K = \{0, \ldots, b\},$$

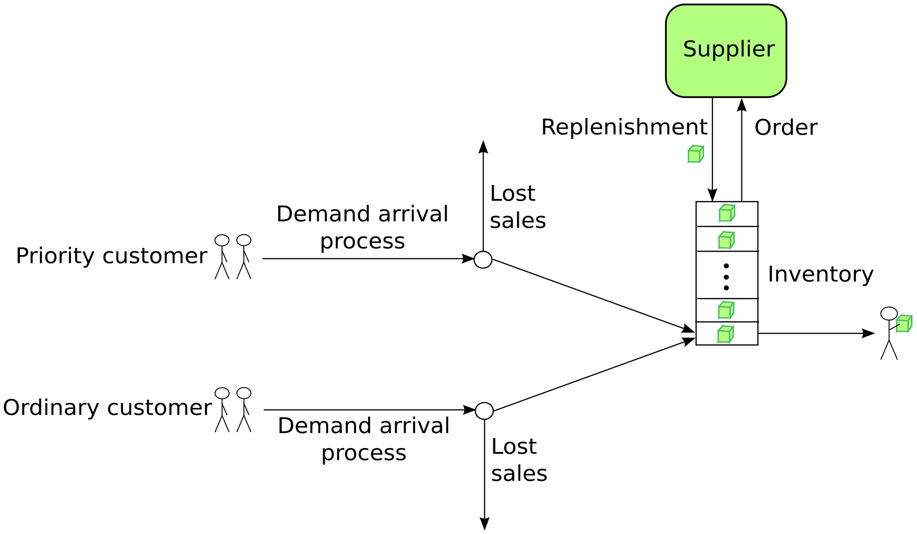

Fig. 2 The pure inventory system with two customer classes

where b is the maximal size of the inventory. Y is irreducible and therefore ergodic and we define the limiting and stationary distribution of Y by

$$\theta := (\theta(k) : k \in K), \quad \theta(k) := \lim_{t\to\infty} P(Y(t) = k).$$

θ satisfies the following global balance equations in case of $s < b$.

$$\begin{aligned}
&\theta(0)\cdot\nu = \theta(1)\cdot(\lambda_1 + p\,\lambda_2),\\
&\theta(k)\cdot(\lambda_1 + p\,\lambda_2 + \nu) = \theta(k+1)\cdot(\lambda_1 + p\,\lambda_2) + \theta(k-1)\cdot\nu, \quad k = 1,\ldots,s-1,\\
&\theta(s)\cdot(\lambda_1 + p\,\lambda_2 + \nu) = \theta(s+1)\cdot(\lambda_1 + \lambda_2) + \theta(s-1)\cdot\nu,\\
&\theta(k)\cdot(\lambda_1 + \lambda_2 + \nu) = \theta(k+1)\cdot(\lambda_1 + \lambda_2) + \theta(k-1)\cdot\nu, \qquad k = s+1,\ldots,b-1,\\
&\theta(b)\cdot(\lambda_1 + \lambda_2) = \theta(b-1)\cdot\nu.
\end{aligned}$$

It is direct to see that these are the balance equations for a finite birth-death process and we have immediately the following result.

Proposition 6 *The inventory process* $Y = (Y(t) : t \geq 0)$ *has the following limiting and stationary distribution*

$$\begin{aligned}
\theta(0) &= \left[\sum_{j=0}^{s}\left(\frac{\nu}{\lambda_1 + p\,\lambda_2}\right)^j + \left(\frac{\nu}{\lambda_1 + p\,\lambda_2}\right)^s\cdot\sum_{j=1}^{b-s}\left(\frac{\nu}{\lambda_1 + \lambda_2}\right)^j\right]^{-1}\\
&= \left[\frac{1-\left(\frac{\nu}{\lambda_1+p\,\lambda_2}\right)^{s+1}}{1-\left(\frac{\nu}{\lambda_1+p\,\lambda_2}\right)} + \left(\frac{\nu}{\lambda_1 + p\,\lambda_2}\right)^s\cdot\left(\frac{1-\left(\frac{\nu}{\lambda_1+\lambda_2}\right)^{b-s+1}}{1-\left(\frac{\nu}{\lambda_1+\lambda_2}\right)} - 1\right)\right]^{-1}\\
&= \left[\frac{1-\left(\frac{\nu}{\lambda_1+p\,\lambda_2}\right)^{s+1}}{1-\left(\frac{\nu}{\lambda_1+p\,\lambda_2}\right)} + \left(\frac{\nu}{\lambda_1 + p\,\lambda_2}\right)^s\cdot\left(\frac{\nu}{\lambda_1+\lambda_2}\right)\cdot\left(\frac{1-\left(\frac{\nu}{\lambda_1+\lambda_2}\right)^{b-s}}{1-\left(\frac{\nu}{\lambda_1+\lambda_2}\right)}\right)\right]^{-1},\\
\theta(k) &= \left(\frac{\nu}{\lambda_1 + p\,\lambda_2}\right)^k\cdot\theta(0), \qquad k = 1,\ldots,s,\\
\theta(k) &= \left(\frac{\nu}{\lambda_1 + p\,\lambda_2}\right)^s\left(\frac{\nu}{\lambda_1 + \lambda_2}\right)^{k-s}\cdot\theta(0), \quad k = s+1,\ldots,b.
\end{aligned}$$

6 Conclusion

We investigated ergodicity for queueing-inventory systems with two priority classes in case of unbounded queues for both customer classes. The admission control to the queues is flexible and incorporates two parameters (s, p) which mainly restrict entrance of low priority customers. The main result is a stability condition which is proved by construction of a suitable Lyapunov function. Although we proved that in case of parameter constellation $(s, 1)$ the condition is sufficient and necessary we believe that in general the condition is not sharp. To clarify the situation in general is part of our ongoing research.

Another direction of our research in this field will be to investigate the system with priority classes where arriving customers which find the inventory depleted are backlogged. This means the lost sales behaviour of customers is replaced by backordering unsatisfied demand.

Acknowledgements We thank two reviewers for their careful reading of the first version of the article and their constructive critics which helped us to enhance the paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The authors do not analyse or generate any data sets, because the work proceeds within a theoretical and mathematical approach.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

References

Arslan, H., Graves, S. C., & Roemer, T. A. (2007). A single-product inventory model for multiple demand classes. *Management Science, 53*(9), 1486–1500. https://doi.org/10.2139/ssrn.725803.

Baek, J. W., Dudina, O. S., & Kim, C. S. (2017). A queueing system with heterogeneous impatient customers and consumable additional items. *International Journal of Applied Mathematics and Computer Science, 27,* 367–384. https://doi.org/10.1515/amcs-2017-0026.

Bijvank, M., & Vis, F. A. (2011). Lost-sales inventory theory: A review. *European Journal of Operational Research, 215,* 1–13. https://doi.org/10.1016/j.ejor.2011.02.004.

Chen, H., Zhou, Z., & Liu, M. (2012). A queueing-inventory system with two classes demand and subject to selected service. *Journal of Information and Computational Science, 9*(11), 3081–3089.

Daduna, H. (1990). Exchangeable items in repair systems: Delay times. *Operations Research, 38*(2), 349–354. https://doi.org/10.1287/opre.38.2.349.

Dimitriou, I. (2022). Stationary analysis of certain Markov-modulated reflected random walks in the quarter plane. *Annals of Operations Research, 310,* 355–387. https://doi.org/10.1007/s10479-020-03676-8.

Fayolle, G., Iasnogorodski, R., & Malyshev, V. (1999). Random walks in the quarter-plane. Applications of Mathematics, vol. 40. Springer, Berlin, Heidelberg, New York. https://doi.org/10.1007/978-3-319-50930-3

Foster, F. G. (1953). On the stochastic matrices associated with certain queuing processes. *The Annals of Mathematical Statistics, 24*(3), 355–360. https://doi.org/10.1214/aoms/1177728976.

Gelenbe, E. (1991). Product form queueing networks with negative and positve customers. *Journal of Applied Probability, 28,* 656–663.

Gelenbe, E., Glynn, P., & Sigman, K. (1991). Queues with negative arrivals. *Journal of Applied Probability, 28,* 245–250.

Gruen, T. W., Corsten, D., & Bharadwaj, S. (2002). *Retail out-of-stocks: A worldwide examination of extent causes and consumer responses*. Washington: Grocery Manufacturers of America.

Isotupa, K. P. S. (2006). An (s, Q) Markovian inventory system with lost sales and two demand classes. *Mathematical and Computer Modelling, 43,* 687–694. https://doi.org/10.1016/j.mcm.2005.09.027.

Isotupa, K. P. S. (2007). Continuous review (s, Q) inventory system with two types of customers. *International Journal of Agile Manufacturing, 9*(1), 79–85.

Isotupa, K. P. S. (2011). An (s, Q) inventory system with two classes of customers. *International Journal of Operational Research, 12,* 1–19. https://doi.org/10.1504/IJOR.2011.041856.

Isotupa, K. P. S. (2015). Cost analysis of an $(S-1, S)$ inventory system with two demand classes and rationing. *Annals of Operations Research, 233,* 411–421. https://doi.org/10.1007/s10479-013-1407-3.

Isotupa, K. P. S., & Samanta, S. K. (2013). A continuous review (s, Q) inventory system with priority customers and arbitrarily distributed lead times. *Mathematical and Computer Modelling, 57,* 1259–1269. https://doi.org/10.1016/j.mcm.2012.10.029.

Jaiswal, N. K. (1968). *Priority queues. Mathematics in science and engineering* (Vol. 50). New York, London: Academic Press.

Jeganathan, K. (2015). Linear retrial inventory system with second optional service under mixed priority service. *TWMS Journal of Applied and Engineering Mathematics, 5*(2), 249–268.

Jeganathan, K., Anbazhagan, N., & Kathiresan, J. (2013). A retrial inventory system with non-preemptive priority service. *International Journal of Information and Management Sciences, 24*(1), 57–77.

Jeganathan, K., Kathiresan, J., & Anbazhagan, N. (2016). A retrial inventory system with priority customers and second optional service. *OPSEARCH, 53,* 808–834. https://doi.org/10.1007/s12597-016-0261-x.

Kaplan, R. S. (1970). A dynamic inventory model with stochastic lead times. *Management Science, 16*(7), 491–507.

Karush, W. (1957). A queuing model for an inventory problem. *Operations Research, 5*(5), 693–703.

Kelly, F. P., & Yudovina, E. (2014). *Stochastic Networks*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139565363

Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. Chichester - New York - Brisbane - Toronto: John Wiley and Sons.

Krishnamoorthy, A., Shajin, D., & Narayanan, V. C. (2021). Inventory with positive service time: a survey. In Anisimov, V., Limnios, N. (eds.) Queueing theory 2, Chap. 6, pp. 201–237. Wiley, London. https://doi.org/10.1002/9781119755234.ch6

Krishnamoorthy, A., & Manjunath, A. S. (2015). On priority queues generated through customer induced service interruption. *Neural, Parallel, and Scientific Computations, 23,* 459–486.

Krishnamoorthy, A., & Manjunath, A. S. (2018). On queues with priority determined by feedback. *Calcutta Statistical Association Bulletin, 70*(1), 33–56. https://doi.org/10.1177/0008068318767271.

Latouche, G., & Ramaswami, V. (eds.) (1999). Introduction to matrix analytic methods in stochastic modeling. ASA-SIAM series on statistics and applied probability. SIAM, Philadelphia. https://doi.org/10.1137/1.9780898719734

Lee, Y. J., & Zipkin, P. (1992). Tandem queues with planned inventories. *Operations Research, 40,* 936–947.

Lee, Y. J., & Zipkin, P. (1995). Processing networks with inventories: Sequential refinement systems. *Operations Research, 43,* 1025–1036.

Li, B., & Arreola-Risa, A. (2021). On minimizing downside risk in make-to-stock, risk averse firms. *Naval Research Logistics, 68,* 199–213.

Li, H., & Zhao, Y. Q. (2009). Exact tail asymptotics in a priority queue - characterizations of the preemptive model. *Queueing Systems, 63,* 355–381. https://doi.org/10.1007/s11134-009-9142-9.

Liu, M., Xi, F., & Chen, H. (2013). Control policies for a Markov queueing-inventory system with two demand classes. In International Asia conference on industrial engineering and management innovation (IEMI2012) proceedings, pp. 1543–1550. https://doi.org/10.1007/978-3-642-38445-5_162

Liu, M., Feng, M., & Wong, C. Y. (2014). Flexible service policies for a Markov inventory system with two demand classes. *International Journal of Production Economics, 151,* 180–185. https://doi.org/10.1016/j.ijpe.2013.10.010.

Melikov, A. Z., Mirzayev, R. R., & Nair, S. S. (2022a). Double sources queueing-inventory system with hybrid replenishment policy. Σ *Mathematics, 10,* 2423. https://doi.org/10.3390/math10142423.

Melikov, A. Z., & Fatalieva, M. R. (1998). Situational inventory in counter-stream serving systems. *Engineering Simulation, 15,* 839–848.

Melikov, A. Z., Mirzayev, R. R., & Nair, S. S. (2022b). Numerical study of a queueing-inventory system with two supply sources and destructive customers. *Journal of Computer and Systems Sciences International, 61*(4), 581–598. https://doi.org/10.1134/S1064230722030091.

Melikov, A. Z., & Molchanov, A. A. (1992). Stock optimization in transportation/storage systems. *Cybernetics and Systems Analysis, 28*(3), 484–487.

Melikov, A. Z., Ponomarenlo, L. A., & Aliyev, I. A. (2018a). Markov models of systems with demands of two types and different restocking policies. *Cybernetics and Systems Analysis, 54*(6), 900–917. https://doi.org/10.1007/s10559-018-0093-1.

Melikov, A. Z., Ponomarenlo, L. A., & Aliyev, I. A. (2018b). Analysis and optimization of models of queueing-inventory systems with two types of requests. *Journal of Automation and Information Sciences, 50*(12), 34–50. https://doi.org/10.1615/JAutomatInfScien.v50.i12.30.

Miller, R. G. (1958). Priority queues. Technical report, Stanford University, California.

Morse, P. M. (1958). Queues, inventories and maintenance. Wiley, New York https://doi.org/10.2307/2342909

Otten, S. (2018). Integrated models for performance analysis and optimization of queueing-inventory systems in logistic networks. PhD thesis, Universität Hamburg, Department of Mathematics, Hamburg, Germany.

Ravid, R., Boxma, O. J., & Perry, D. (2013). Repair systems with exchangeable items and the longest queue mechanism. *Queueing Systems, 73*, 295–316. https://doi.org/10.1007/s11134-012-9319-5.

Reed, J., & Zhang, B. (2017). Managing capacity and inventory for multi-server make-to-stock queues. *Queueing Systems, 86,* 61–94. https://doi.org/10.1007/s11134-017-9519-0.

Rego, J. R. D., & Mesquita, M. A. D. (2011). Spare parts inventory control: A literature review. *Produção, 21*(4), 656–666. https://doi.org/10.1590/S0103-65132011005000002.

Rubio, R., & Wein, L. M. (1996). Setting base stock levels using product-form queueing networks. *Management Science, 42,* 259–268. https://doi.org/10.1287/mnsc.42.2.259.

Schwarz, M., Sauer, C., Daduna, H., Kulik, R., & Szekli, R. (2006). $M/M/1$ queueing systems with inventory. *Queueing Systems: Theory and Applications, 54,* 55–78. https://doi.org/10.1007/s11134-006-8710-5.

Shajin, D., Dudin, A. N., Dudina, O. S., & Krishnamoorthy, A. (2020). A two-priority single server retrial queue with additional items. *Journal of Industrial and Management Optimization, 16*(6), 2891–2912. https://doi.org/10.3934/jimo/2019085.

Sigman, K., & Simchi-Levi, D. (1992). Light traffic heuristic for an M/G/1 queue with limited inventory. *Annals of Operations Research, 40,* 371–380.

Tempelmeier, H. (2005). Bestandsmanagement in Supply Chains. Norderstedt: Books on Demand.

Verhoef, P., & Sloot, L. M. (2006). Out-of-stock: Reactions, antecedents, management solutions, and a future perspective. In Krafft, M., Mantrala, M.K. (eds.) Retailing in the 21st century: Current and future trends, pp. 239–253. Springer, Philadelphia, Pennsylvania, USA. https://doi.org/10.1007/978-3-540-72003-4_18

Wang, F.-F. (2015). Approximation and optimization of a multi-server impatient retrial inventory-queueing system with two demand classes. *Quality Technology and Quantitative Management, 12*(3), 269–292. https://doi.org/10.1080/16843703.2015.11673381.

Wang, J., Baron, O., & Scheller-Wolf, A. (2015). $M/M/c$ Queue with Two Priority Classes. *Operations Research, 63*(3), 733–749. https://doi.org/10.1287/opre.2015.1375.

Wolff, R. W. (1989). *Stochastic modeling and the theory of queues*. Englewood Cliffs: Prentice-Hall International Editions.

Yadavalli, V. S. S., Anbazhagan, N., & Jeganathan, K. (2015). A retrial inventory system with impatient customers. *Applied Mathematics and Information Sciences, 9*(2), 637–650. https://doi.org/10.12785/amis/090212.

Zazanis, M. (1994). Push and pull systems with external demands. In *Proceedings of the 32nd Allerton conference on communication, control, and computing*, Allerton, Illinois.

Zhao, N., & Lian, Z. (2011). A queueing-inventory system with two classes of customers. *International Journal of Production Economics, 129*(1), 225–231. https://doi.org/10.1016/j.ijpe.2010.10.011.