Flexible Services and Manufacturing Journal (2021) 33:381–424
https://doi.org/10.1007/s10696-020-09378-4

Optimizing stock levels for service-differentiated demand classes with inventory rationing and demand lead times

Oguzhan Vicil[1]

Published online: 19 March 2020

Abstract

In this paper, we study a service parts inventory management system for a single product at a parts distribution center serving two priority-demand classes: *critical* and *non-critical*. Distribution center keeps a common inventory pool to serve the two demand classes, and provides differentiated levels of service by means of inventory rationing. We assume a continuous review one-for-one ordering policy with backorders and Poisson demand arrivals. Only one demand class provides advance demand information whose orders are due after a deterministic demand lead time, whereas the orders of the other demand class need to be satisfied immediately. The problem has been studied before, but remained a challenging problem. The quality of the existing heuristic for estimating the *critical* class service levels can diminish significantly in some settings and the search routine for the service level optimization model relies on a brute force approach. Our contribution to the literature is twofold. For the given class of inventory replenishment and allocation policies, first we determine the form of the optimal solution to the service level optimization model, and then we derive an exact optimization routine to determine the optimal policy parameters provided the steady-state distribution is available. The computation of steady-state probabilities is needed only once. Second, we propose an alternative approach to estimate steady-state probabilities. By analyzing the limiting behavior of transition probabilities during infinitesimal time intervals, we are able to characterize the relationships between the steady-state probabilities, which satisfy nicely formed balance equations under the so-called *Independence Assumption*. In the numerical study section, we show that our approach provides superior performance in estimating service levels than the existing heuristic for all the examples considered. We also compare the performance of using the *critical* class service levels computed according to our method against the service levels computed by the existing heuristic, and show that our method can provide inventory savings up to 16.67%.

Keywords Threshold rationing policy · Advance demand information · Priority-demand classes · Service level · Backorder system · Demand lead time

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10696-020-09378-4) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article

1 Introduction

To better manage customer classes with different service level requirements, managers are constantly seeking new ways to improve system performance. For the last two decades, in addition to incorporating inventory rationing strategy in demand pooling, there is also a growing literature in integrating advance demand information (ADI) into inventory allocation and replenishment decisions. Numerous studies show that ADI improves system performance when used effectively and it may be possible to reduce the need for inventory or excess capacity.

Two types of ADI have been studied in the literature. In the "perfect ADI", customers provide exact information about their orders. The orders are to be delivered at a certain time in the future. Therefore, the time between order placement and due date, which is called demand lead time (DLT), is deterministic. There are no variations in the size of the order, and cancelations are not allowed. On the other hand, in the "imperfect ADI", early signals (or estimates) about prospective future orders are provided. Order sizes and due dates are subject to change, and cancelations might also be allowed.

Continuous advancements in information technology make it easier and inexpensive to collect and process prospective demand information in a timely manner, which leads to higher availability of advance demand information. Additionally, with increased cooperation between suppliers and customers, customers are more willing to share their advance demand information with suppliers in return for lower costs and higher service levels, which also leads to potential improvements in supply chain performance. However, the availability of advance demand information raised important managerial and research questions in the literature such as: How to incorporate ADI into the current policy? How beneficial is perfect/imperfect advance demand information for suppliers and customers? How do the system parameters affect the value of perfect/imperfect ADI? What is the optimal strategy to allocate inventory among different customer classes under perfect/imperfect ADI? How does the availability of ADI affect operation decisions and supply chain performance?

There may be several reasons why customers differ in their ADI structures. For example, only one of the customer classes might have the ability to accurately plan for repairs and scheduled maintenance. The customer knows exactly the time when that service part will be used in the repair/maintenance, and inform the service parts supplier as soon as this information is available. For another example, due to the information technologies in use, one of the customer classes can quickly diagnose the failed part via sensors and inform the supplier in advance before the actual repair/maintenance starts.

In this paper, we are motivated by a generic problem in incorporating perfect advance demand information into the threshold level based inventory rationing within the framework of continuous review one-for-one ordering policy (also known as the $(S-1, S)$ policy), and aim to find its effect on system stock levels and performance measures. We study a service parts inventory management system for a single product at a parts distribution center serving two priority-demand

classes: *critical* and *non-critical*. We assume that both customer classes have long term relations with the service provider and therefore both customer classes are willing to wait for the backordered demands. However, minimum service level requirements are imposed through the existing contracts. According to the *service level agreements*, the *critical* demand class has contracted for a higher service level than the *non-critical* demand class. The distribution center keeps a common inventory pool to serve the two demand classes, and provides differentiated levels of service by means of inventory rationing. A reserve level of inventory is held for use by *critical* demand class only, in anticipation of future demands. The two demand classes also differ in terms of their ADI structures. Only one demand class provides ADI whose orders are due after a deterministic demand lead time, whereas the orders of the other demand class need to be satisfied immediately. According to our model setting, the demand class, which provides the perfect ADI, does not accept early deliveries and wants to receive the part as soon as it is needed in a just-in-time fashion. (The same assumption is also used by, i.e., Hariharan and Zipkin (1995), Wang et al. (2002), Koçağa and Şen (2007), Benjaafar et al. (2011).) Hence, early fulfillment of orders before due dates are not allowed. We consider both cases. In the first one, *critical* demands are due immediately, whereas *non-critical* demands are due after a fixed DLT. In the second one, we consider the opposite case in which *non-critical* demands are due immediately while *critical* demands are due after a fixed DLT. Both priority classes exhibit mutually independent, stationary, Poisson demand processes. Whenever a demand of any type occurs, immediately a replenishment order is given through the supplier, which will be received after a constant lead time. Existing backorders are cleared according to the *priority clearing mechanism*.

Our objective is to find the optimal policy parameters (base-stock and threshold levels) that minimize the system stock (base-stock level) while satisfying service level constraints for both demand classes.

This model setting has also a practical importance, which has been studied before by Koçağa and Şen (2007) but remained a challenging problem. Their research originated from their real life experience with a leading capital equipment manufacturer which is at the top of the supply chain for many high technology products. The company has an extensive spare parts network (with more than 50,000 active parts need to be managed) which consists of more than 70 company owned distribution centers and depots across the world. The depots and regional distribution centers face two types of demand streams as *down orders* which that need to be satisfied immediately, and *lead time orders* which need to be satisfied at a future date.

The exact analysis for computing the steady-state probabilities seems intractable for this model. Even for the "DLT = 0" model, Vicil and Jackson (2016) show the difficulty of the exact analysis indicating that to predict the system state after a lead time, both the current system state information and the knowledge of the sequence of events over the lead time become relevant, which tremendously increase the complexity of the solution. In addition, the standard inventory balance equations do not hold: The physical stock and backorders can exist at the same time. To estimate service levels for this model, Koçağa and Şen (2007) use the same strong assumption as Dekker et al. (1998): A lead time ago, there are no existing orders and

on-hand inventory is equal to the base-stock level. To determine the *critical* demand class service level, Koçağa and Şen (2007) use a similar hitting time approach as employed by Dekker et al. (1998). However, as we show in the numerical study section, the absolute errors for the *critical* demand class service level can be significant in this approach in some settings (as high as 34.65%).

To improve the quality of approximation, we choose an unorthodox method and base our analysis on the limiting behavior of state transition probabilities during infinitesimal time intervals (rather than the traditional approaches used in the inventory theory and control literature, i.e., choosing an arbitrary point in time and predict the system state after a lead time). Then conditioned on being at a certain state, by relaxing the dependency of both the age-of-pipeline vector and the age-of-order-due date vector to the number of *non-critical* backorders (the so-called *Independence Assumption*), we are able to characterize the relationships between the steady-state probabilities. We show that the steady-state probabilities satisfy nicely formed balance equations, which can be easily solved via numerical methods. Furthermore, after establishing several structural properties regarding the properties of the steady-state distribution and service levels, we are able to provide a computationally efficient optimization routine which requires the computation of steady-state probabilities only once.

The remainder of the paper is organized as follows. In Sect. 2, we review the literature on related inventory systems and summarize our main contributions. In Sect. 3, we study the service level optimization model for the system setting in which *critical* demands are due immediately, whereas *non-critical* demands are due after a fixed DLT. In Sect. 4, we propose a method to estimate the steady-state probabilities and the *critical* demand class service level. In Sect. 5, we study the alternate model in which *non-critical* demands are due immediately, whereas *critical* demands are due after a fixed DLT. In the numerical study section, Sect. 6, we compare the performance of our method with the existing heuristic using simulation.

2 Literature review

This study is related to two streams of work. The first one consists of operations management papers which study the use of ADI/DLT in inventory and production decisions. The second one consists of inventory management papers that study the inventory allocation and replenishment decisions among different priority customers via demand pooling and inventory rationing.

Hariharan and Zipkin (1995) are among the first to study DLT in inventory/distribution systems. They demonstrate that demand lead times improve performance, in precisely the same way the replenishment lead times degrade it. Donselaar et al. (2001) consider two types of demand as regular demand from small orders, and very irregular lumpy demand from infrequent, large orders. They analyze the inventory reduction that could be achieved if the advance demand information could be provided to the manufacturer. Gallego and Özer (2001) consider the portfolio of customers with different demand lead times. They show that state-dependent (s, S) and base-stock policies are optimal for stochastic inventory systems with and without

fixed costs in a finite-horizon setting in which customers place advance orders. They also determine conditions under which advance demand information has no operational value. Karaesmen and Buzacott (2002) investigate the structure of optimal control policies for a discrete-time-make-to-stock queue with advance demand information. They propose a heuristic policy based on an extension of the base-stock that integrates advance demand information through a release lead time parameter. Özer and Wei (2004) consider a periodic-review, stochastic, capacitated, finite and infinite horizon production system, and show how advance demand information can be a substitute for capacity and inventory. Wang and Toktay (2008) analyze an inventory management with advance demand information and flexible delivery. Their model is closely related to the discrete-time, uncapacitated, advance demand information model of Gallego and Özer (2001), except that they allow for delivery flexibility. In a two-period setting, Tan et al. (2009) investigate the impact of using imperfect ADI in a production/inventory system with two priority-demand classes and inventory rationing. They aim to minimize the expected total costs, under the assumption that unmet demand is lost. Boyacı and Özer (2010) study a profit-maximization model in which a manufacturer collects advance sales information periodically prior to the regular sales season for a capacity decision. Benjaafar et al. (2011) analyze a production-inventory systems with imperfect advance demand information. Customers are allowed to update the status of their orders and may request an order fulfillment prior to or later than the expected due date. Bernstein and DeCroix (2015) study the impact of different types of advance demand information on optimal capacities and profit. According to their model, the firm receives information revealing either the total volume of demand across products or the mix of demand between products. Topan et al. (2018) examine a single-item, single-location, periodic-review lost sales inventory model with a general representation of imperfect ADI. Their model allows for returning excess stock built up due to imperfections to the upstream supplier.

In the continuous review inventory management framework, there are numerous studies in the literature which consider static rationing models for differentiated demand classes. For the (Q,R) models, Nahmias and Demmy (1981) are among the first to consider rationing. They consider a critical level policy with Poisson demand processes and constant lead times for two priority demand classes. Demands can be backordered. They provide approximations for the expected number of backorders under the assumption that at most one order is outstanding. Their model is extended by Moon and Kang (1998) to a compound Poisson demand process. Melchiors et al. (2000) study the same model as Nahmias and Demmy (1981), but with the lost sales assumption. They present an exact formulation of the average inventory cost and then provide a simple optimization procedure. Deshpande et al. (2003) study a model similar to Nahmias and Demmy (1981) but allow multiple outstanding orders in the pipeline, which increases the complexity of the analysis. They propose approximations on the expected number of backorders and provide an efficient algorithm for computing the optimal policy parameters. Arslan et al. (2007) analyze a similar model as Deshpande et al. (2003), but allow multiple demand classes that are characterized by different shortage costs or service requirements. They show that there is sample-path equivalence between their backorder clearing rule and the serial inventory system. For the $(S-1, S)$ backorder models, Dekker et al. (1998) study a

critical level policy within the framework of one-for-one inventory model with two demand classes and backorders. The demand process is a Poisson process and order lead times are constant. They provide approximations for estimating the service level for the critical demand, while the analysis for non-critical demand is exact. Wang et al. (2002) study distribution systems that provide two classes of service that differ in their demand lead times. An emergency service demand is to be filled immediately upon its arrival, while a non-emergency demand is to be filled after a deterministic demand lead time. Vicil and Jackson (2016) study a similar model as Dekker et al. (1998) but allow general lead time distributions. By exploring the limiting behavior of state transitions during infinitesimal time intervals, under the certain approximation assumption, they show that the steady-state distributions of the model is identical to the steady-state distributions of the model with exponentially distributed lead times with the same mean. Vicil and Jackson (2018) study the same model as Vicil and Jackson (2016) but include class-specific expected waiting-time requirements in addition to the fill-rate constraints. They characterize the form of the optimal solution in this model setting and propose a simple two step solution strategy to determine optimal base-stock and threshold levels. In a recent work, Gabor et al. (2018) consider a similar inventory system as Dekker et al. (1998) but differs in terms of the service level measures. Their model assumes that the service level of low-priority customers is measured by a response time guarantee, while the service level of the high-priority customers is measured by the fill rate. For the lost sales models within the continuous review $(S-1, S)$ framework, Dekker et al. (2002), Kranenburg and van Houtum (2007), and Isotupa (2015) all study a critical level policy for multiple demand classes with Poisson demand processes.

In the stream of the literature that consider both ADI/DLT and differentiated customer classes in terms of their priority, Koçağa and Şen (2007) is the first study to simultaneously consider demand lead times and rationing. They provide an approximation for the critical service level while the service level for the non-critical demand class is exact. Recently, Basten and Ryan (2019) consider a periodic review inventory system with zero replenishment lead times and two demand classes. They study the impact of maintenance delay flexibility on the optimal inventory control policies.

The problem we consider in this article is identical to the model studied by Koçağa and Şen (2007), and closely related to the models described in Dekker et al. (1998), Wang et al. (2002), and Vicil and Jackson (2016, 2018). We simultaneously consider inventory rationing and demand lead times. The exact analysis for computing the steady-state probabilities seems intractable. To overcome this obstacle, we use a similar approach employed by Vicil and Jackson (2016) and base our analysis on the limiting behavior of state transition probabilities during infinitesimal time intervals under the certain approximation assumption. Our main contributions are summarized as follows:

- We provide structural results for the steady-state distribution and performance measures of the original model setting studied in this article. Koçağa and Şen (2007) provide structural results (with limited scope compared to ours) based on their proposed approximation for the service level measures.

- We are able to determine the form of the optimal solution to the service level optimization model for the given class of inventory replenishment and allocation policies. We also provide an exact search algorithm to determine optimal policy parameters which requires the computation of steady-state probabilities only once. In their study, Koçağa and Şen (2007) found the optimal policy parameters through the brute force search approach (though, they were able to limit the number of possible policy parameters to consider).
- Our optimization search routine can be used with any approach; it doesn't matter whether the steady-state probabilities are determined via our proposed approximation, simulation study or any other heuristic. These (approximated) steady-state probabilities can be used as an input to our optimization algorithm, then the algorithm provides the optimal policy parameters as an output for the heuristic being used.
- Since our method allows us to determine the limiting distribution of being at any given system state, other performance measures such as expected on-hand stock, and expected number of *critical* and *non-critical* class backorders can be easily estimated. This information would be useful especially in cost optimization models.
- In the numerical study section, we show that our approach provides superior performance in estimating the *critical* class service level than the existing heuristic for all the examples considered. In the numerical study, we show that the average absolute errors of the existing heuristic are 2.30% (the *non-critical* class has a DLT) and 3.50% (the *critical* class has a DLT), while the average absolute errors are 0.38% and 0.09%, respectively, for our approach. Furthermore, the maximum absolute errors of the existing heuristic are 19.46% and 34.64% for the two cases, while they are limited to 4.79% and 0.52% in our approach.
- In the optimization study, we also compare the performance of using the *critical* class service levels computed according to our method against the service levels computed by the existing heuristic, and show that our method can provide inventory savings up to 16.67%.

As a final remark, incorporation of the DLT into the current threshold rationing policy imposes fundamental challenges in the analysis. Although both inventory systems may seem to be similar at first sight, the DLT model is not a simple extension to the "DLT = 0" model. This is mainly because the demand process and the due date process are not identical in the DLT model. As a result, the evolution of changes in system states in both models differ from each other. This in turn affects the system dynamics, and consequently the steady-state probabilities. Therefore, the structural results of the Vicil and Jackson (2016, 2018) studies, which are valid for the "DLT = 0" model, cannot be directly used in the DLT model. The structural properties should be defined and proved for the DLT model with the necessary changes in expressions and definitions. Furthermore, Vicil and Jackson (2016) proved that under the so-called independence assumption, the steady-state distribution of system states with deterministic or stochastic lead time distribution satisfies the same balance equations as the system with an exponential lead time distribution with the same mean. On the other hand, under a similar independence assumption,

the balance equations we derived in this paper are different than would be the balance equations of the system with exponential lead times. This also shows that the model we study in this paper has different dynamics than the model studied by Vicil and Jackson (2016).

3 Service level optimization model

We consider two priority-demand classes: *critical* and *non-critical*. *Critical* demand class requires a higher service level than the *non-critical* demand class. Both priority classes exhibit mutually independent, stationary, Poisson demand processes, with rates λ^n and λ^c. We assume a continuous review one-for-one policy with a base-stock level S. Whenever a demand of any type occurs, immediately a replenishment order is given through the supplier, which will be received after a constant lead time of L.

Distribution center keeps a common inventory pool to serve the two demand classes, and provides differentiated levels of service by means of inventory rationing. A reserve level of inventory, denoted by S_c, is held for use by *critical* demand class only, in anticipation of future demands. According to the model, *critical* demands are due immediately, whereas *non-critical* demands are due after a deterministic DLT of H. At their due dates, orders of *non-critical* class are backordered if on-hand stock is at or below a certain threshold level S_c, while *critical* class orders are backordered only if on-hand stock is zero. We assume that $H \leq L$, so that the DLT is not quoted longer than the replenishment lead time. In our model, service levels are measured in terms of fill rate, which is defined as the percentage of demands satisfied immediately from on-hand stock at their "due dates".

So, the proposed policy works as follows: an incoming *critical* demand (which is due immediately at the time of its arrival) is satisfied as long as there is physical stock. Otherwise, it is backordered. A *non-critical* class order is accepted upon its arrival, which is due after H time units. At its due date, the *non-critical* demand is satisfied only if on-hand stock is above the threshold level S_c. Otherwise, it is backordered. Since a one-for-one policy is followed, the arrival of any demand, either by a *critical* or a *non-critical* class, triggers an immediate replenishment order of size 1, which will be received after a constant lead time of L. Existing backorders are cleared according to the *priority clearing mechanism*. Incoming replenishment orders are first used to clear *critical* class backorders, if there exists any. Otherwise, they are used to restore the reserve stock as long as on-hand stock is less than the threshold level S_c. Existing *non-critical* backorders are only cleared if on-hand stock is at the threshold level S_c at the delivery times. Only after all *non-critical* backorders are cleared, deliveries are used to increase on-hand stock beyond S_c.

At any time t, let $OH(t)$ denote the number of units on-hand, $R(t)$ denote the number of units in resupply, $B^c(t)$ denote the number of outstanding *critical* backorders, $B^n(t)$ denote the number of outstanding *non-critical* backorders, and Y^n be the number of *non-critical* class orders that have been accepted but not yet due. Under the *one-for-one replenishment* and *threshold level based raioning and backorder clearing policy*, the following relations hold:

$$S = OH(t) + R(t) - B^c(t) - B^n(t) - Y^n(t). \tag{1}$$

$$OH(t) = \left[S - R(t) + B^n(t) + Y^n(t)\right]^+ \tag{2}$$

$$B^c(t) = \left[R(t) - B^n(t) - Y^n(t) - S\right]^+. \tag{3}$$

These relations are also valid for the steady-state distribution of these quantities, denoted by OH, R, B^c, B^n, and Y^n. Therefore, it is sufficient to capture the stationary distribution of (R, B^n, Y^n).

Note that although the *one-for-one replenishment policy* is in use, inventory position can be higher than the base-stock level S in this model due to the demand lead time effect. For example, for a given (S, S_c) pair, if there are exactly two *non-critical* demands during $[0, H)$, at time H on-hand stock will be S while the inventory position will be $S + 2$. This property also adds to the complexity of the steady-state analysis.

Furthermore, at any point t in time, the difference $[R(t) - Y^n(t)]$ has an effect on the net inventory level. $Y^n(t)$ represents the number of *non-critical* demands arrived during $(t - H, t]$ and whose order due dates have not yet come by time t. Since a *one-for-one replenishment policy* is implemented, each demand arrival triggers a replenishment order of size 1. Therefore, $Y^n(t)$ portion of the $R(t)$ does not affect the net inventory level at time t. This result also leads to the following implicit conditions:

$$Y^n(t) \leq R(t), \tag{4}$$

$$B^n(t) \leq \left[R(t) - Y^n(t) - (S - S_c)\right]^+. \tag{5}$$

Our objective is to determine the optimal policy parameters (S, S_c) that minimize the total inventory investment (base-stock level) S while satisfying all service level constraints for each demand class.

Let $\beta^n(S, S_c)$ (respectively $\beta^c(S, S_c)$) denote the fill rate achieved for the *non-critical* (respectively *critical*) class demands as functions of (S, S_c). Due to the *Poisson Arrivals See Time Averages* principle, arriving demands face the steady-state distribution of on-hand inventory (see, e.g., Tijms 1986). At their corresponding due dates, *non-critical* class demands are served if and only if on-hand stock is greater than S_c, and *critical* class demands are served if and only if on-hand stock is non-zero. Hence, if we denote $P_\infty(\cdot)$ as the steady-state probability distribution of a random process, then the provided fill rates will be as follows:

$$\beta^n(S, S_c) = 1 - P_\infty(OH \leq S_c \mid (S, S_c)), \tag{6}$$

and

$$\beta^n(S, S_c) = 1 - P_\infty(OH = 0 \mid (S, S_c)). \tag{7}$$

The optimization problem can be written as:

$$\begin{aligned} &\min S \\ &s.t. \\ &\quad \beta^n(S, S_c) \geq \bar{\beta}^n \\ &\quad \beta^c(S, S_c) \geq \bar{\beta}^c \\ &\quad S > S_c \quad \geq 0, \end{aligned}$$

for contractually-specified service levels $\bar{\beta}^c$ and $\bar{\beta}^n$, $\bar{\beta}^c > \bar{\beta}^n > 0$.

Although the *one-for-one replenishment policy* is implemented and the model is a pure backorder model, computing accurate service levels is still a challenging task due to the effects of inventory rationing and demand lead times. The steady-state analysis is very difficult, if not impossible. Even if the knowledge of a system state at any point t in time is known, it is still very difficult to probabilistically determine the system state after an order lead time L. This is because, not only the number of demand arrivals and deliveries over a lead time affect the system state at time $t + L$, but also the sequence of those demand arrivals and deliveries have effect. Furthermore, the demand lead time has also an effect on the system state since *non-critical* class orders are due H time units later upon their occurrence. Those dynamics contribute significantly to the complexity of the analysis. To show this, let us consider a model with policy parameters $(S, S_c) = (4, 2)$, order lead time $L = 2$, and demand lead time $H = 0.5$. At a random point t in time, let the system state be $OH = 2, R = 3, B^n = 1, B^c = 0, Y^n = 0$, and consider the following four scenarios with the corresponding event lists. Timings of events are also indicated within the parentheses. In each scenario, there are exactly two *critical* demand arrivals, two *non-critical* demand arrivals and three deliveries. However, we change the sequence and/or timings of events while keeping number of demand arrivals of each type and deliveries unchanged.

Scenario I: Event List = {*non-critical* demand $(t + 0.1)$, *critical* demand $(t + 0.2)$, delivery $(t + 0.3)$, delivery $(t + 0.4)$, delivery $(t + 1.2)$, *critical* demand $(t + 1.3)$, *non-critical* demand $(t + 1.6)$ }. The resulting system state at time $t + L$ will be $OH = 1, R = 4, B^n = 0, B^c = 0, Y^n = 1$.

Scenario II: Event List = {*non-critical* demand $(t + 0.1)$, *critical* demand $(t + 0.2)$, delivery $(t + 0.3)$, *critical* demand $(t + 1.3)$, *non-critical* demand $(t + 1.6)$, delivery $(t + 1.7)$, delivery $(t + 1.9)$}. The resulting system state at time $t + L$ will be $OH = 2, R = 4, B^n = 1, B^c = 0, Y^n = 1$.

Scenario III: Event List = {*non-critical* demand $(t + 0.1)$, *critical* demand $(t + 0.2)$, delivery $(t + 0.3)$, *non-critical* demand $(t + 1.3)$, *critical* demand $(t + 1.6)$, delivery $(t + 1.7)$, delivery $(t + 1.9)$}. The resulting system state at time $t + L$ will be $OH = 2, R = 4, B^n = 2, B^c = 0, Y^n = 0$.

Scenario IV: Event List = {delivery $(t + 0.1)$, delivery $(t + 0.2)$, delivery $(t + 0.3)$, *non-critical* demand $(t + 1.3)$, *non-critical* demand $(t + 1.4)$, *critical* demand $(t + 1.6)$, *critical* demand $(t + 1.7)$}. The resulting system state at time $t + L$ will be $OH = 2, R = 4, B^n = 2, B^c = 0, Y^n = 0$.

As can be observed from those examples, depending on the sequence of events and their timings, the resulting system state after an order lead time may differ greatly. Although this is a pure backorder model, keeping track of every possible permutation of events enormously increases the complexity of the solution. However, to overcome this obstacle, we are going to adapt a similar approach used by Vicil and Jackson (2016, 2018), and then base our analysis on the limiting behavior of state transitions over an infinitesimal time interval.

First, we establish structural properties for the steady-state distribution and service level measures, which hold regardless of the lead time distribution. Then, provided a method for computing stationary probabilities is available, we present a computationally efficient optimization algorithm to determine optimal policy parameters (S, S_c) which requires the computation of steady-state probabilities only once. And finally we present a method to compute the steady-state probabilities. Our approach is exact for the calculation of the *non-critical* class service level, while it is a high quality approximation for the *critical* class service level.

3.1 Structural results and properties of the steady-state distribution

For given (S, S_c) pair and system parameters λ^n, λ^c, L, and H, beginning from a regeneration point in which there is no unit in the resupply system, let (m, T_m, E_m) describe the mth event in the system: T_m is the time of the mth event, and E_m is the type of event where $E_m \in \{\text{“}v\text{”}, \text{“}n\text{”}, \text{“}c\text{”}, \text{“}y\text{”}\}$ representing events “delivery”, “*non-critical* demand arrival”, “*critical* demand arrival”, and “*non-critical* order due date”, respectively. After the mth event, let R_m denote the number of units in resupply, B^n_m denote the number of *non-critical* backorders, and Y^n_m denote the number of *non-critical* orders received but not yet due. Clearly at the regeneration point $m = 0, R_0 = 0, B^n_0 = 0$ and $Y^n_0 = 0$.

Proposition 1 *The dynamics of* (R_m, B^n_m, Y^n_m) *can be described completely in terms of the sample path* $\left\{(m, T_m, E_m); m = 1, 2, 3, \ldots\right\}$:

$$R_m = \begin{cases} R_{m-1} + 1, & E_m = \text{“}n\text{” or “}c\text{”}, \\ R_{m-1} - 1, & E_m = \text{“}v\text{”}, \\ R_{m-1}, & E_m = \text{“}y\text{”}. \end{cases}$$

$$B^n_m = \begin{cases} B^n_{m-1} + 1, & E_m = \text{“}y\text{”}, R_{m-1} - Y^n_{m-1} \geq S - S_c, \\ B^n_{m-1} - 1, & E_m = \text{“}v\text{”}, R_{m-1} - Y^n_{m-1} > S - S_c, B^n_{m-1} = R_{m-1} - Y^n_{m-1} - (S - S_c), \\ B^n_{m-1}, & \text{otherwise}. \end{cases}$$

$$Y^n_m = \begin{cases} Y^n_{m-1} + 1, & E_m = \text{“}n\text{”}, \\ Y^n_{m-1} - 1, & E_m = \text{“}y\text{”}, \\ Y^n_{m-1}, & \text{otherwise}. \end{cases}$$

Proof See “Appendix 1”. □

Proposition 2 *For a fixed S, the following relations hold for all $S_c' > S_c$:*

$$\begin{aligned}\beta^n(S, S_c) &\geq \beta^n(S, S_c'),\\ \beta^c(S, S_c) &\leq \beta^c(S, S_c').\end{aligned}$$

Proof See "Appendix 2". □

Let $Z_0 = \{0, 1, 2, \dots\}$, the set of non-negative integers and $\xi_t = (r, b^n, y^n) \in \mathbb{F}_{(S,S_c)}$ denote the system state at time t, $t \geq 0$, where $\mathbb{F}_{(S,S_c)}$ represents the set of feasible states when the policy parameters are given by (S, S_c). Clearly,

$$\mathbb{F}_{(S,S_c)} = \left\{(r, b^n, y^n) : \ y^n \leq r, b^n \leq [r - y^n - (S - S_c)]^+, (r, b^n, y^n) \in Z_0 \times Z_0 \times Z_0\right\}.$$

Let $\pi_{(r,b^n,y^n)}(S, S_c), (r, b^n, y^n) \in \mathbb{F}_{(S,S_c)}$, denote the steady-state distribution of $(R(t), B^n(t), Y^n(t))$. The following proposition establishes an important invariance result.

Proposition 3 *The steady-state probabilities $\pi_{(r,b^n,y^n)}, (r, b^n, y^n) \in \mathbb{F}_{(S,S_c)}$, are invariant to changes in S provided $\Delta = S - S_c$ is constant.*

Proof See "Appendix 3". □

A similar invariance result has been previously established by Vicil and Jackson (2016) for the "DLT = 0" model.

Let $\varphi_h\left(S, S_c\right)$ denote the stationary distribution of on-hand inventory when policy parameters are given by (S, S_c):

$$\varphi_h\left(S, S_c\right) = P_\infty\left(OH(t) = h \mid (S, S_c)\right), \ h = 0, 1, \dots, S.$$

Then, using (2), we have

$$\varphi_h\left(S, S_c\right) = \begin{cases} \sum\limits_{\substack{(r,0,y^n) \in \mathbb{F}_{(S,S_c)} \\ (S - r + y^n) = h}} \pi_{(r,0,y^n)}(S, S_c), & S_c < h \leq S,\\ \sum\limits_{\substack{(r,b^n,y^n) \in \mathbb{F}_{(S,S_c)} \\ (S - r + b^n + y^n) = h}} \pi_{(r,b^n,y^n)}(S, S_c), & 0 < h \leq S_c,\\ \sum\limits_{\substack{(r,b^n,y^n) \in \mathbb{F}_{(S,S_c)} \\ (S - r + b^n + y^n) \leq 0}} \pi_{(r,b^n,y^n)}(S, S_c), & h = 0. \end{cases} \tag{8}$$

At a random point t in time, let us suppose $R(t) = i$ and $Y^n(t) = j$; that is there are i replenishment orders outstanding, and there are j *non-critical* orders outstanding that are not due yet. Let $\mathrm{u}_{[k]}$ denote the age of the kth oldest replenishment order and $\left(\mathrm{u}_{[1]}, \mathrm{u}_{[2]}, \dots, \mathrm{u}_{[i]}\right)$ denote the age-of-pipeline vector. Similarly, let $\mathrm{v}_{[k]}$ denote the age of the kth oldest *non-critical* order that is not due yet, and $\left(\mathrm{v}_{[1]}, \mathrm{v}_{[2]}, \dots, \mathrm{v}_{[j]}\right)$ denote the age of *non-critical* order due date vector. Furthermore, let $\mathbb{U}(t) = \left\{\mathrm{u}_{[1]}, \mathrm{u}_{[2]}, \dots, \mathrm{u}_{[i]}\right\}$ denote the set of age of replenishment orders in the

resupply and $\mathbb{V}(t) = \left\{\mathrm{v}_{[1]}, \mathrm{v}_{[2]}, \ldots, \mathrm{v}_{[j]}\right\}$ denote the set of age of *non-critical* outstanding orders that are not yet due. Let $\mathcal{P}(\lambda)$ denote the Poisson probability distribution with mean λ, and let $p(k;\lambda) = e^{-\lambda}\lambda^k/k!$. Then, the cardinality of the difference of these two sets has Poisson distribution with mean $\lambda^c L + \lambda^n(L-H)$

$$|\mathbb{U}(t)\backslash\mathbb{V}(t)| \sim R(t) - Y^n(t) \sim \mathcal{P}(\lambda^c L + \lambda^n(L-H)). \tag{9}$$

The reasoning is as follows. Replenishment orders, which are originated from the demand arrivals before $t-L$, should have already arrived by t. Therefore at any point t in time, $\mathbb{U}(t)$ represents the set of timings of all the demand arrivals during $(t-L, t]$. Since total demand process is a Poisson process, $|\mathbb{U}(t)|$ is Poisson distributed with mean $(\lambda^c + \lambda^n)L$. Similarly, $\mathbb{V}(t)$ represents the set of timings of *non-critical* demand arrivals during $(t-H, t]$, and therefore $|\mathbb{V}(t)|$ is Poisson distributed with mean $\lambda^n H$. Hence, $\mathbb{U}(t)\backslash\mathbb{V}(t)$ represents the order process excluding the *non-critical* orders that are not due yet, but has impact on the net inventory level. Therefore, $|\mathbb{U}(t)\backslash\mathbb{V}(t)|$ has the same distribution as $R(t) - Y^n(t)$, which is Poisson distributed with mean $\lambda^c L + \lambda^n(L-H)$. It is also important to note that the events represented in $\mathbb{U}(t)\backslash\mathbb{V}(t)$ are independent of the events in $\mathbb{V}(t)$. (*Note: $R(t)$ and $Y^n(t)$ are dependent Poisson random variables. If they were to be independent, the difference of these two random variables would be a Skellam distribution. See e.g.,* Skellam (1946) *for more information.*)

Using (8) and summing over all possible (r, y^n) pairs, we can immediately calculate $\varphi_h\left(S, S_c\right)$ for $h > S_c$

$$\begin{aligned}\varphi_h\left(S, S_c\right) &= \sum_{\substack{(r,0,y^n) \in \mathbb{F}_{(S,S_c)} \\ r - y^n = S - h}} \pi_{(r,0,y^n)}(S, S_c) \\ &= p(S-h;\lambda^c L + \lambda^n(L-H)), \quad \text{for } S_c < h \le S.\end{aligned} \tag{10}$$

Therefore, we can determine the achieved fill rate exactly for the *non-critical* demand class

$$\begin{aligned}\beta^n\left(S, S_c\right) &= \sum_{h=S_c+1}^{S} \varphi_h\left(S, S_c\right) \\ &= \sum_{k=0}^{S-S_c-1} p(k;\lambda^c L + \lambda^n(L-H)).\end{aligned} \tag{11}$$

This result has also been established by Koçağa and Şen (2007), but they followed a different approach.

One immediate result of (11) is the following corollary.

Corollary 1 *The non-critical demand class fill rate $\beta^n\left(S, S_c\right)$ is invariant to changes in S provided $\Delta = S - S_c$ is constant.*

This corollary establishes that under the proposed rationing policy, the *non-critical* demand class service level is a function of Δ, and therefore can be calculated without the knowledge of S_c. This significant result will be exploited during the optimization procedure in Sect. 3.2.

However, the challenge is to determine the steady-state distribution of states for which $OH \leq S_c$ (equivalently $R - Y^n \geq S - S_c$), which we need them to determine the *critical* demand class fill rate. In Sect. 4, we will develop an approximation procedure to compute these probabilities based on the limiting behavior of an infinitesimal analysis.

Let $\psi_u\left(S, S_c\right)$ denote the steady-state marginal distribution of the number of *critical* backorders excluding the cases for which on-hand inventory is positive. Defining the distribution this way will be useful later in the optimization routine.

$$\begin{aligned}\psi_u\left(S, S_c\right) &= P_\infty\left(B^c(t) = u,\ OH(t) = 0 \mid (S, S_c)\right) \\ &= \sum_{\substack{(r, b^n, y^n) \in \mathbb{F}_{(S,S_c)} \\ (S - r + b^n + y^n)^+ = 0 \\ (r - b^n - y^n - S)^+ = u}} \pi_{(r,b^n,y^n)}(S, S_c).\end{aligned} \tag{12}$$

The next two lemmas will be used to reduce the computational complexity of the search routine for finding the optimal (S, S_c) pairs. For fixed Δ, the following lemma allows us to determine steady-state probabilities for different (S, S_c) pairs directly from the knowledge of $\psi_u(\Delta, 0)$ probabilities.

Lemma 1 *For fixed* $\Delta, k = 1, 2, \ldots,$ *and* $u = 0, 1, \ldots,$

$$\psi_u(\Delta + k, k) = \psi_{u+k}(\Delta, 0).$$

Proof See "Appendix 4". □

The next lemma establishes very useful property for the optimization search routine. As long as Δ is fixed, *critical* class service levels at different policy parameters (S, S_c) can be computed from the knowledge of $\psi_u(\Delta, 0)$ without any further computation of steady-state probabilities.

Lemma 2 *For fixed* Δ, *and* $k = 1, 2, \ldots,$

$$\beta^c(\Delta + k, k) = \beta^c(\Delta, 0) + \sum_{u=0}^{k-1} \psi_u(\Delta, 0).$$

Proof See "Appendix 5". □

3.2 Optimization algorithm

Let Δ^* be the smallest value of $\Delta = S - S_c$ that satisfies the required *non-critical* demand class fill rate:

$$\Delta^* = \text{argmin}\left\{ \Delta \in \{1, 2, \ldots\} \,:\, \sum_{k=0}^{\Delta-1} p(k; \lambda^c L + \lambda^n (L - H)) \,\geq\, \bar{\beta}^n \right\}. \tag{13}$$

By Corollary 1, (13) also implies that for any (S, S_c) pair to satisfy the *non-critical* demand class service level constraint, $S - S_c$ should be at least Δ^*.

Let S_c^* be the smallest value of S_c that satisfies the required *critical* demand class fill rate under the condition that $S^* = S_c^* + \Delta^*$:

$$S_c^* = \text{argmin}\ \left\{ S_c \in \{0, 1, \ldots\} \,:\, \beta^c\left(S_c + \Delta^*, S_c\right) \geq \bar{\beta}^c \right\}. \tag{14}$$

Theorem 1 *The parameters* $\left(S, S_c\right) = \left(S^*, S_c^*\right)$ *are optimal for the fill rate optimization model.*

Proof Suppose there exists another solution $\left(S', S_c'\right)$ that is feasible such that $S' < S^*$. In order this solution to be feasible with respect to the *non-critical* class fill rate constraint, we must have $S' - S_c' \geq \Delta^*$. Let us consider the solution $\left(S', S' - \Delta^*\right)$. Due to Corollary 1, this solution satisfies the *non-critical* demand class fill rate requirement. Since the *critical* demand class fill rate is nondecreasing in S_c for fixed S (by Proposition 2), and since $S' - \Delta^* \geq S_c'$, we should have $\beta^c\left(S', S' - \Delta^*\right) \geq \bar{\beta}^c$. But this implies $S' - \Delta^* \geq S_c^*$ by the definition of S_c^*; which in turn implies $S' \geq \Delta^* + S_c^* = S^*$, a contradiction. Therefore, there cannot be another feasible solution with a smaller value of S than S^*. □

Note that we are not omitting the possibility that there may be multiple optimal solutions $\left(S^*, S_c'\right)$ such that $S_c' \neq S_c^*$. However, none of the optimal solutions can have $S' < S^*$. The theorem guarantees to find one of the optimal $\left(S^*, S_c\right)$ pairs.

Next, in Table 1 we present a computationally efficient approach to determine the optimal $\left(S^*, S_c^*\right)$ pairs that requires the computation of steady-state probabilities only once. For now, we assume there is a method to compute stationary probabilities for states $OH \leq S_c$. In the next section, we will present a method to approximate those probabilities.

4 Estimating steady-state probabilities and determining the service level for the *critical* demand class

For states $(r, 0, y^n) \in \mathbb{F}_{(S,S_c)}$, $r - y^n \leq S - S_c$, we can immediately determine the steady-state probabilities. Because the knowledge of $R(t) - Y^n(t)$ and $Y^n(t)$ fully represent the system state at any point t in time, provided $R(t) - Y^n(t) \leq S - S_c$: for a

Table 1 Optimization algorithm for the two demand class

1.	Compute Δ^* using (13), and $\beta^n(\Delta^*, 0)$ using (11)
2.	Compute $\pi_{(r,b^n,y^n)}(\Delta^*, 0)$, for all $(r, b^n, y^n) \in \mathbb{F}_{(\Delta^*,0)}$
3.	Compute $\psi_u(\Delta^*, 0)$ for $u = 0, 1, \ldots$ using (12)
4.	(a) Set $\beta^c = \beta^n(\Delta^*, 0)$
	(b) Set $u = 0$
	(c) While $\beta^c < \bar{\beta}^c$
	Set $\beta^c = \beta^c + \psi_u(\Delta^*, 0)$ using Lemma 2
	Set $u = u + 1$
	(d) Set $S_c^* = u$
	(e) Set $S^* = S_c^* + \Delta^*$
	(f) Set
	$\beta^c\left(S^*, S_c^*\right) = \beta^c$
	$\beta^n\left(S^*, S_c^*\right) = \beta^n(\Delta^*, 0)$
6.	Return $\left(S^*, S_c^*\right)$
7.	Return $\beta^n\left(S^*, S_c^*\right)$, $\beta^c\left(S^*, S_c^*\right)$

given (r, y^n) pair there is only one possible system state, $(r, 0, y^n)$. Hence, steady-state probabilities for these states are given by

$$\pi_{(r,0,y^n)}(S, S_c) = p(r - y^n; \lambda^c L + \lambda^n (L - H)).p(y^n; \lambda^n H),\ r - y^n \leq S - S_c. \quad (15)$$

However, this is not the case for states $(r, b^n, y^n) \in \mathbb{F}_{(S,S_c)}$ such that $r - y^n > S - S_c$. Because the knowledge of only $R(t) - Y^n(t)$ and $Y^n(t)$ is not sufficient to determine the system state at a random point t in time: For a given (r, y^n) pair, there are multiple states (r, b^n, y^n) with different on-hand inventory, *critical* and *non-critical* backorder levels. Therefore, for these cases, we can only write the the left-hand side of (15) as the sum of the steady-state probabilities:

$$\sum_{\substack{(r,b^n,y^n) \in \mathbb{F}_{(S,S_c)} \\ r - y^n - (S - S_c) = u \\ y^n = v}} \pi_{(r,b^n,y^n)}(S, S_c) = p(r - y^n; \lambda^c L + \lambda^n (L - H)).p(y^n; \lambda^n H), \quad (16)$$

$u = 1, 2, \ldots; v = 0, 1, \ldots$.

But unfortunately individual stationary probabilities $\pi_{(r,b^n,y^n)}(S, S_c)$ in (16) are not readily available and, therefore we need a method to determine these probabilities, which is a challenging task.

In the literature, regarding the backorder models with rationing in the continuous review framework, the *hitting time approach* is often used to estimate the stationary probabilities. This approach is based on conditioning on the time that on-hand inventory first hits the threshold level S_c, but generally requires the strong assumption that there are no existing orders in the pipeline at the beginning of time interval in consideration. For the model studied in this article, Koçağa and Şen (2007) also

use this assumption in the *hitting time approach* to estimate the *critical* class service levels. However, as shown by Vicil and Jackson (2016) for the "DLT = 0" model, this approach does does not necessarily work under a variety of arrival rates and lead time values. Furthermore, even if the system state information is known at a random point t in time, it is still very difficult to determine the system state after a lead time. As we have shown in the previous section that not only the number of class-specific demand arrivals have impact on the system state a lead time later, but also the sequence of those demand arrivals, pipeline vector and *non-critical* order due date vector have impact. Hence, every possible permutation of events should be taken into account to analyze the system behavior over a lead time. Therefore, the classical approach in which only the number of demand arrivals over a lead time is taken into account would be misleading for this model setting.

To overcome those obstacles, we exploit another approach with the aim of decreasing the complexity of the analysis and providing more accurate results. To do so, starting from the initial system state (0, 0, 0) in which there are no orders in the resupply and on-hand inventory is equal to S, by conditioning on the system state at a random point t in time, we analyze the limiting behavior of transition probabilities for this process during an infinitesimal time interval τ. This infinitesimal analysis greatly reduces the number of events that can happen during $(t, t+\tau]$, and hence the analysis becomes much tractable.

But first, we need establish some background knowledge regarding the stochastic behavior of a general unit in the resupply system. At a random point t in time, suppose exactly one order has occurred during $(t-L, t]$, which has an impact on the net inventory level. Clearly this event belongs to the set $\mathbb{U}(t)\backslash\mathbb{V}(t)$, and $R(t) - Y^n(t) = 1$. Given that this order has occurred, let us establish the distribution of the time at which this order has occurred. Let T be the elapsed time from $t-L$ until this order occurs, and $N(s, s+u)$ be the number of orders $\in \mathbb{U}(t)\backslash\mathbb{V}(t)$ that has been received in $(s, s+u]$. Then, for $\tau < L - H$,

$$\begin{aligned} P[T < \tau \mid N(t-L,t) = 1] &= \frac{P[T < \tau; N(t-L,t) = 1]}{P[N(t-L,t) = 1]} \\ &= \frac{P[N(t-L,t-L+\tau) = 1; N(t-L+\tau,t) = 0]}{P[N(t-L,t) = 1]} \\ &= \frac{P[N(t-L,t-L+\tau) = 1] \,.\, P[N(t-L+\tau,t) = 0]}{P[N(t-L,t) = 1]}. \end{aligned} \tag{17}$$

Since $\tau < L - H$, $(t-L, t-L+\tau] \cap (t-H, t] = \emptyset$. Therefore, all demand arrivals in $(t-L, t-L+\tau]$ belong to the set $\mathbb{U}(t)\backslash\mathbb{V}(t)$. Therefore,

$$P[N(t-L,t-L+\tau) = 1] = (\lambda^n + \lambda^c)\tau e^{-(\lambda^n+\lambda^c)\tau}. \tag{18}$$

Furthermore, if we consider the interval $(t-L+\tau, t]$, all the *critical* demand arrivals in $(t-L+\tau, t]$ belongs to the set $\mathbb{U}(t)\backslash\mathbb{V}(t)$. However, among the *non-critical* demands, only the ones occurred in $(t-L+\tau, t-H]$ belongs to the set $\mathbb{U}(t)\backslash\mathbb{V}(t)$. Therefore, $N(t-L+\tau, t) = 0$ if and only if there is no *critical* demand arrival in $(t-L+\tau, t]$ and there is no *non-critical* demand arrival in $(t-L+\tau, t-H]$. Hence,

$$P[N(t-L+\tau,t)=0] = e^{-\lambda^c(L-\tau)}\, e^{-\lambda^n(L-H-\tau)}. \tag{19}$$

Using the results (18) and (19) in (17), we have

$$\begin{aligned} P[T<\tau \mid N(t-L,t)=1] &= \frac{(\lambda^n+\lambda^c)\tau e^{-(\lambda^n+\lambda^c)\tau}\, e^{-\lambda^c(L-\tau)}\, e^{-\lambda^n(L-H-\tau)}}{(\lambda^c L+\lambda^n(L-H))\, e^{-(\lambda^c L+\lambda^n(L-H))}} \\ &= \frac{(\lambda^n+\lambda^c)\tau\, e^{-(\lambda^c L+\lambda^n(L-H))}}{(\lambda^c L+\lambda^n(L-H))\, e^{-(\lambda^c L+\lambda^n(L-H))}} \\ &= \frac{(\lambda^n+\lambda^c)\tau}{\lambda^c L+\lambda^n(L-H)}. \end{aligned} \tag{20}$$

Let $p(\tau)$ be the common probability that any replenishment order in the resupply at a random point t in time, which belongs to the set $\mathbb{U}(t)\backslash\mathbb{V}(t)$, is still in the resupply system at time $t+\tau$, $\tau < L-H$. This order will be in the resupply at time $t+\tau$ if and only if it arrived during $(t-L+\tau, t]$, since all arrivals prior to $t+\tau-L$ should have arrived by $t+\tau$. Therefore, $p(\tau)$ is given by:

$$\begin{aligned} p(\tau) &= 1 - P[T<\tau \mid N(t-L,t)=1] \\ &= 1 - \frac{(\lambda^n+\lambda^c)\tau}{\lambda^c L+\lambda^n(L-H)}, \end{aligned} \tag{21}$$

which is independent of time t. Note that $p(\tau)$ represents the probability for unordered replenishment orders in the resupply.

Let

$$q_{t,\tau}(x|n) = P[x \text{ of those } n \text{ units remain in the resupply at } t+\tau \mid N(t-L,t)=n].$$

Each replenishment order occurred during $(t-L,t]$ belongs to the set $\mathbb{U}(t)\backslash\mathbb{V}(t)$, and has a probability $p(\tau)$ that it is still in the resupply at time $t+\tau$. Therefore, the probability that x of the n replenishment orders, which belong to the set $\mathbb{U}(t)\backslash\mathbb{V}(t)$, will remain in the resupply at time $t+\tau$ is given by

$$q_{t,\tau}(x|n) = \binom{n}{x} p(\tau)^x \left[1-p(\tau)\right]^{n-x}, \tag{22}$$

which is also independent of time t.

Similarly, let us derive system dynamics for the *non-critical* replenishment orders belonging to the set $\mathbb{V}(t)$. At a random point t in time, suppose exactly one *non-critical* order has occurred during $(t-H,t]$, which has an impact on the set of *non-critical* orders that are not yet due. This event belongs to the set $\mathbb{V}(t)$, and therefore $Y^n(t)=1$. Given that this event has occurred, let us establish the distribution of the time at which this order has occurred. Let $\tilde{T}$ be the elapsed time from $t-H$ until this order occurs, and $\tilde{N}(s,s+u)$ be the number of orders $\in \mathbb{V}(t)$ that has been received in $(s,s+u]$. Clearly $\tilde{N}(s,s+u)$ has a Poisson distribution with mean $\lambda^n u$. Then, for $\tau < H$,

$$
\begin{aligned}
P[\tilde{T} < \tau \mid \tilde{N}(t-H,t) = 1] &= \frac{P[\tilde{T} < \tau; \tilde{N}(t-H,t) = 1]}{P[\tilde{N}(t-H,t) = 1]} \\
&= \frac{P[\tilde{N}(t-H,t-H+\tau) = 1; \tilde{N}(t-H+\tau,t) = 0]}{P[\tilde{N}(t-H,t) = 1]} \\
&= \frac{P[\tilde{N}(t-H,t-H+\tau) = 1] \,.\, P[\tilde{N}(t-H+\tau,t) = 0]}{P[\tilde{N}(t-H,t) = 1]} \\
&= \frac{\lambda^n \tau e^{-\lambda^n \tau}\, e^{-\lambda^n (H-\tau)}}{\lambda^n H e^{-\lambda^n H}} \\
&= \frac{\tau}{H}.
\end{aligned}
\tag{23}
$$

Hence, the time at which the *non-critical* order belonging to the set $\mathbb{V}(t)$ occurs is uniformly distributed over the interval $(t-H,t]$.

Let $\tilde{p}(\tau)$ be the common probability that any order that is "not yet due" at a random point t in time, which belongs to the set $\mathbb{V}(t)$, is still "not yet due" at time $t+\tau$, $\tau < H$. This order will be "not yet due" at time $t+\tau$ if and only if it occurred during $(t+\tau-H,t]$, since all the *non-critical* orders prior to $t+\tau-H$ should have been due by $t+\tau$. Therefore, $p(\tau)$ is given by:

$$
\begin{aligned}
\tilde{p}(\tau) &= 1 - P[\tilde{T} < \tau \mid \tilde{N}(t-H,t) = 1] \\
&= 1 - \frac{\tau}{H},
\end{aligned}
\tag{24}
$$

which is independent of time t.

Let

$$
\begin{aligned}
\tilde{q}_{t,\tau}(x|n) &= P[x \text{ of those } n \text{ units remain "not yet due" at } t+\tau \mid \tilde{N}(t-L,t) = n] \\
&= \tilde{q}_{t,\tau}(x|n) = \binom{n}{x} \tilde{p}(\tau)^x \left[1-\tilde{p}(\tau)\right]^{n-x}.
\end{aligned}
\tag{25}
$$

Since class-specific demands are Poisson processes and a one-for-one policy is followed, both the resupply process and the due date process are mirror reflections of the demand realizations during the last L time units. In addition, resupply processes and due date processes are also independent of the demand processes after time t. Hence, during an infinitesimal time interval dt, the probability of more than one event to occur is $o(dt)$ due to the Poisson nature of the processes. Furthermore, conditioned on being at state (r, b^n, y^n) at time t, the delivery process over the next dt time units only depends on the elements of set $\mathbb{U}(t)\backslash\mathbb{V}(t)$, because $dt < H$ and no orders in the resupply that belongs to the set $\mathbb{V}(t)$ could be received by time $t+dt$. Hence, probability of a delivery process is determined by (22).

The brief summary of our approach for determining the steady-state probabilities is as follows. Starting from the initial state (0, 0, 0) at time $t = 0$, let

$$
P_{(\bar{r},\bar{b}^n,\bar{y}^n),(r,b^n,y^n)}(t,t') = P\left[\xi_{t'} = (r,b^n,y^n) \mid \xi_0 = (0,0,0),\ \xi_t = (\bar{r},\bar{b}^n,\bar{y}^n)\right].
$$

By conditioning on the state of the system at time t:

$$P_{(0,0,0),(r,b^n,y^n)}(0,t+\tau) = \sum_{(\bar{r},\bar{b}^n,\bar{y}^n)\in\mathbb{F}_{(S,S_c)}} P_{(0,0,0),(\bar{r},\bar{b}^n,\bar{y}^n)}(0,t)\cdot P_{(\bar{r},\bar{b}^n,\bar{y}^n),(r,b^n,y^n)}(t,t+\tau). \tag{26}$$

Since τ is an infinitesimal time duration, the states $(\bar{r},\bar{b}^n,\bar{y}^n)$ are chosen such that the state (r,b^n,y^n) can be reached by at most one transition over the next infinitesimal τ time duration. Then we explore the limiting behavior of above transitions. Therefore, first we need to determine one-step transition probabilities over an infinitesimal τ time duration.

At a random point t in time, since a one-for-one policy is implemented for the replenishment process, the system state information $(R(t),B^n(t),Y^n(t))$ contains essential information about what has happened over $(t-L,t]$ and $(t-H,t]$ time intervals. From the system state information, first we can understand that a a total of $R(t)$ number of replenishment orders have been placed over the last L periods which have been triggered by both class demand processes. Second, among those $R(t)$ replenishment orders $Y^n(t)$ of them has been triggered by the *non-critical* customer class demands realized during $(t-H,t]$. Both $R(t)$ and $Y^n(t)$ are Poisson distributed. However, conditioned on being at the system state $(R(t),B^n(t),Y^n(t))$, the placements of replenishment orders during $(t-L,t]$ or $(t-H,t]$ are no longer Poisson processes. This is because, due to the implemented rationing policy, there is a dependency between the $B^n(t)$ and the demand process over the past L periods (Note that this wouldn't be the case for the simple one-for-one policy without customer differentiation. The knowledge of $R(t)$ would be sufficient to characterize the order replenishment process, which is still a Poisson process). This in turn affect the delivery and due date processes after time t.

Therefore, to solve the one-step transition probabilities we relax the dependency of both the age-of-pipeline vector and the age-of-order-due date vector to the number of *non-critical* backorders, and make the following approximation assumption:

Independence Assumption Conditioned on being at state $(R(t),B^n(t),Y^n(t))$ at a random point t in time, both the age-of-pipeline vector $\left(\mathrm{u}_{[1]},\mathrm{u}_{[2]},\ldots,\mathrm{u}_{[i]}\right)$ and the age-of-*non-critical*-order-due date vector $\left(\mathrm{v}_{[1]},\mathrm{v}_{[2]},\ldots,\mathrm{v}_{[j]}\right)$ are independent of the number of *non-critical* backorders $B^n(t)$.

The *Independence Assumption* allows us to probabilistically determine the delivery process and the due date process from the knowledge of $R(t)$ and $Y^n(t)$. After solving the one-step transition probabilities, we take the limits as $\tau\to 0$ and $t\to\infty$ and achieve the results in the following theorem. For notational simplicity, we use $\pi_{(r,b^n,y^n)}$ instead of $\pi_{(r,b^n,y^n)}(S,S_c)$.

Theorem 2 *Let*

$$\mu = \frac{(\lambda^n+\lambda^c)}{\lambda^c L+\lambda^n(L-H)}, \tag{27}$$

and

$$\vartheta = \frac{1}{H}. \tag{28}$$

Under the Independence Assumption, for a given (S, S_c) pair, the steady-state distribution of (R, B^n, Y^n) satisfies the following balance equations:

$$
\begin{aligned}
&r = 0, b^n = 0, y^n = 0;\\
&\quad \lambda\pi_{(0,0,0)} = \mu\pi_{(1,0,0)}.\\
&r \geq 1, r - y^n = 0, b^n = 0;\\
&\quad (\lambda + y^n\vartheta)\pi_{(r,0,y^n)} = \lambda^n\pi_{(r-1,0,y^n-1)} + (r - y^n + 1)\mu\pi_{(r+1,0,y^n)}.\\
&0 < r < \Delta, b^n = 0, y^n = 0;\\
&\quad (\lambda + r\mu)\pi_{(r,0,0)} = \lambda^c\pi_{(r-1,0,0)} + \vartheta\pi_{(r,0,1)} + (r+1)\mu\pi_{(r+1,0,0)}.\\
&1 \leq r - y^n < \Delta, b^n = 0, y^n \geq 1;\\
&\quad [\lambda + (r - y^n)\mu + y^n\vartheta]\pi_{(r,0,y^n)} = \lambda^n\pi_{(r-1,0,y^n-1)} + \lambda^c\pi_{(r-1,0,y^n)}\\
&\qquad\qquad + (y^n + 1)\vartheta\pi_{(r,0,y^n+1)} + (r - y^n + 1)\mu\pi_{(r+1,0,y^n)}.\\
&r = \Delta, b^n = 0, y^n = 0;\\
&\quad (\lambda + \Delta\mu)\pi_{(r,0,0)} = \lambda^c\pi_{(r-1,0,0)} + \vartheta\pi_{(r,0,1)} + (\Delta+1)\mu\pi_{(r+1,0,0)} + (\Delta+1)\mu\pi_{(r+1,1,0)}.\\
&r - y^n = \Delta, b^n = 0, y^n > 0;\\
&\quad (\lambda + \Delta\mu + y^n\vartheta)\pi_{(r,0,y^n)} = \lambda^c\pi_{(r-1,0,y^n)} + \lambda^n\pi_{(r-1,0,y^n-1)} + (y^n+1)\vartheta\pi_{(r,0,y^n+1)}\\
&\qquad\qquad + (\Delta+1)\mu\pi_{(r+1,0,y^n)} + (\Delta+1)\mu\pi_{(r+1,1,y^n)}.\\
&r - b^n = \Delta, b^n > 0, y^n = 0;\\
&\quad (\lambda + r\mu)\pi_{(r,b^n,0)} = \vartheta\pi_{(r,b^n-1,1)} + (r+1)\mu\pi_{(r+1,b^n+1,0)} + (r+1)\mu\pi_{(r+1,b^n,0)}.\\
&r - b^n - y^n = \Delta, b^n > 0, y^n > 0;\\
&\quad [\lambda + (r - y^n)\mu + y^n\vartheta]\pi_{(r,b^n,y^n)} = \lambda^n\pi_{(r-1,b^n,y^n-1)} + (y^n+1)\vartheta\pi_{(r,b^n-1,y^n+1)}\\
&\qquad\qquad + (r - y^n + 1)\mu\pi_{(r+1,b^n+1,y^n)} + (r - y^n + 1)\mu\pi_{(r+1,b^n,y^n)}.\\
&r > \Delta, b^n = 0, y^n = 0;\\
&\quad (\lambda + r\mu)\pi_{(r,0,0)} = \lambda^c\pi_{(r-1,0,0)} + (r+1)\mu\pi_{(r+1,0,0)}.\\
&r - y^n > \Delta, b^n = 0, y^n > 0;\\
&\quad [\lambda + (r - y^n)\mu + y^n\vartheta]\pi_{(r,0,y^n)} = \lambda^n\pi_{(r-1,0,y^n-1)} + \lambda^c\pi_{(r-1,0,y^n)}\\
&\qquad\qquad + (r - y^n + 1)\mu\pi_{(r+1,0,y^n)}.\\
&r - b^n > \Delta, b^n > 0, y^n = 0;\\
&\quad (\lambda + r\mu)\pi_{(r,b^n,0)} = \lambda^c\pi_{(r-1,b^n,0)} + \vartheta\pi_{(r,b^n-1,1)} + (r+1)\mu\pi_{(r+1,b^n,0)}.\\
&r - b^n - y^n > \Delta, b^n > 0, y^n > 0;\\
&\quad [\lambda + (r - y^n)\mu + y^n\vartheta]\pi_{(r,b^n,y^n)} = \lambda^n\pi_{(r-1,b^n,y^n-1)} + \lambda^c\pi_{(r-1,b^n,y^n)}\\
&\qquad\qquad + (y^n+1)\vartheta\pi_{(r,b^n-1,y^n+1)} + (r - y^n + 1)\mu\pi_{(r+1,b^n,y^n)}.
\end{aligned}
$$

Proof See "Appendix 6". □

The balance equations in Theorem 2 show the in and out flows from a state in the steady-state. Depending on the system state in consideration, the inflows show the rates in steady-state from the adjacent neighbors from which one-step transitions are possible. The outflow shows the total rate of leaving the state in consideration. Those balance equations are linear system of equations and thus can be easily solved numerically, i.e., via the method of LU factorization. We also have a valid reference checking point since we can compute the stationary probabilities exactly for states $(r, 0, y^n) \in \mathbb{F}_{(S,S_c)}$, $r - y^n \leq S - S_c$ by (15).

Now, the question is how strong or weak is the dependence of age-of-pipeline vector and age-of-due date vector on the number of *non-critical* backorders. The quality of the approximation will depend on this answer. In the numerical study section, we explore how well the *Independence Assumption* performs under a variety of lead time values, and show the quality of our approximation.

Another important contribution of this study is to allow us to estimate other performance measures such as expected on-hand stock, and expected number of *critical* and *non-critical* class backorders, through the knowledge of $\pi_{(r,b^n,y^n)}$. This is because our approximation permits us to capture full information for the steady-state probabilities of system states (r, b^n, y^n) rather than only estimating a certain performance measure, such as fill rate.

$$E[OH] = \sum_{h=0}^{S} h\, \varphi_h\big(S, S_c\big), \text{using (8).}$$

$$E[B^c] = \sum_{u=0}^{\infty} u\, \psi_u\big(S, S_c\big), \text{with reference to (12).}$$

$$E[B^n] = \sum_{b^n=1}^{\infty} b^n \sum_{\substack{r,\, y^n \\ (r,\, b^n,\, y^n) \in \mathbb{F}_{(S,S_c)}}} \pi_{(r,b^n,y^n)}(S, S_c).$$

5 Alternate model: *critical* class orders are due after a demand lead time

With minor modifications, similar ideas can easily be applied to the model in which *critical* demand class orders are due after a demand lead time of H, while *non-critical* class orders are due immediately. The optimization model and the fill rate equations remain the same as in the original model. We leave the rest of the derivations and analysis to the online supplement. There, we first present the modified definitions and equations according to the alternate model. We then show that the structural results offered by Proposition 2, Proposition 3, Lemmas 1, 2, and Corollary 1 remain valid in this alternate model. We also provide the balance equations for this alternate model in Theorem 3 in the online supplement.

6 Numerical study

Since there is no exact solution yet in the literature for computing the steady-state probabilities, we compare the performance of our method with the only heuristic in the literature, Koçağa and Şen (2007), against the simulation results. Therefore, for comparison we use the same examples and results provided in their study. This section is divided into three subsections. In Sect. 6.1, we compare the performance of our method with the Koçağa and Şen (2007) heuristic in terms of service level approximations for the *critical* demand class (since service level for the *noncritical* demand class can be computed exactly using (11)). In Sect. 6.2, for the fill rate optimization model we compare the optimal policy parameters (S^*, S_c^*) found by using the *critical* class service levels calculated via our method against the Koçağa and Şen (2007) heuristic and the simulation study. In Sect. 6.3, we investigate the benefit of integrating DLT into the threshold rationing policy.

6.1 Service level calculations

First, we compare the performance of both heuristics at high service level requirements for the *critical* demands class. As mentioned in Koçağa and Şen (2007), such high service levels are quite common in the industry. In Table 2, the parameters are chosen such that the critical class service level is around 99%. The replenishment lead time L is 0.5 and the demand lead time H is 0.1 for all the instances. In columns 5–10, we study the case in which the *non-critical* demand class has a DLT. In columns 11–16, we study the case where the *critical* demand class a DLT. In column 5 and column 11, exact service levels for the *non-critical* demand class are computed according to (11). Absolute errors with respect to the simulation results are provided in columns 8 and 14 for the Koçağa and Şen heuristic, and in columns 10 and 16 for our method. We observe that although the Koçağa and Şen heuristic seems to work well for these scenarios, our method considerably provides better quality approximations for all the instances. For the Koçağa and Şen heuristic, the average absolute errors are 1.09% and 1.18% for the two cases, while the average absolute errors are 0.10% and 0.06% for our method. We also observe that, for the last seven instances, as the *non-critical* class service level decreases while the *critical* class service level is kept around 99%, the quality of both approximations are negatively affected. However, the maximum absolute errors for the Koçağa and Şen heuristic can be as high as 5.53% and 5.89% for the two cases, while the maximum absolute errors for our method are limited to 0.53% and 0.25%. We also note that the Koçağa and Şen heuristic consistently underestimates the achieved (simulated) critical class service level while our method consistently overestimates.

The situation is similar in Table 3, in which the analysis is repeated for ten different instances. But this time the parameters are chosen such that the critical class service level is between 90 and 99%. Although the performance of approximations for both methods are not as good as for the 99% service level scenarios, the

Independence Assumption holds well and our method still considerably provides better quality approximations for all the instances. For the Koçağa and Şen heuristic, the average absolute errors are 1.34% and 3.36% for the two cases, while the average absolute errors are 0.35% and 0.13% for our method. The maximum absolute errors are achieved when the *non-critical* service levels are very low: $\beta^n_{exact} = 18.51\%$ ("DLT: *non-critical*" option), and $\beta^n_{exact} = 28.54\%$ ("DLT: *critical*" option). For the Koçağa and Şen heuristic, the absolute errors can be as high as 3.69% and 8.44% for the two cases, while they are limited to 1.56% and 0.50% for our method. We also observe that, as in Table 2, the quality of the Koçağa and Şen heuristic for the "DLT: *non-critical*" option is better than the quality of the Koçağa and Şen heuristic for the "DLT: *critical*" option. The situation is opposite for our method.

Next, the performance of the approximations are tested against the varying system parameters. The results are provided in Table 4. There are four different parts. In each part, a single parameter is varied at a time among the following list while keeping other system parameters fixed: base stock level, the arrival rate for the *critical* demand class, the arrival rate for the *non-critical* demand class and the DLT, in sequential order. In the first part, as in line with previous results, we observe that as both demand class service levels increase, the quality of both approximations gets better. But again, our method provides better approximations for all the examples. The absolute error for the Koçağa and Şen heuristic can be as high as 2.61% and 4.20% while for our method they are limited to 0.31% and 0.18% for the two cases. On the other hand, in the second part, we observe that the lower service levels significantly affect the quality of the Koçağa and Şen heuristic. As the *critical* service level decreases below 90%, we observe that the Koçağa and Şen heuristic deviates significantly from the simulation results for both cases. For the "DLT: *non-critical*" option, the maximum absolute error is 19.46%, which is extremely high for an approximation. However, our approach still provides reasonable approximations: The maximum absolute error is 4.79%. Furthermore, we see that the quality of the Koçağa and Şen heuristic diminishes even more in the case of "DLT: *critical* option". The maximum absolute error is 34.65%. On the other hand, our approach provides superb performance and the maximum absolute error is only 0.52%. From these results, we can conclude that the *Independence Assumption* holds quite well when the *critical* demand class has a DLT option, and therefore even for the very low service levels, our approach provides high quality approximations. In the third part, although the *critical* service levels are higher than the 98%, the absolute errors for the Koçağa and Şen heuristic continue to be considerably high. The low levels of *non-critical* service levels seems to affect the quality of their heuristic and the absolute errors can be as high as 8.49% and 8.75% for the two cases. On the other hand, our approach again provides high quality approximations for all the examples and the maximum absolute errors are 0.18% and 0.26% for the two cases. It is also interesting to note that although the *non-critical* demand rate increases while keeping other system parameters fixed, we observe that the *critical* demand class is not affected much and still receives high levels of service. This might be

counter-intuitive because the total demand rate increases. One possible explanation is that *non-critical* class demands might be frequently in backorder situation and most of the replenishment orders due to the *non-critical* class orders are used to restore reserve stock up to S_c, which are used to satisfy the *critical* class demands. This situation might offset the negative effect of an increase in the total demand rate on the *critical* class service level (*Remark: We couldn't replicate the results of Table 3 of the* Koçağa and Şen (2007) *article for the "DLT: critical" option regarding the second (simulation) and third (heuristic) parts. Therefore, in this study we performed the numerical analysis for the simulation and Koçağa and Şen heuristic for these parts rather than directly using their results, and used the correct values in Table* 4). In the fourth part, the effect of the DLT is studied. We observe that as DLT increases, both demand classes receive higher services. The parameters are chosen such that both demand class service levels are high. Therefore, as we might expect, both methods perform high quality approximations. The absolute error for the Koçağa and Şen heuristic are 0.37% and 0.47% while they are limited to 0.03% and 0.01% for the two cases.

Considering Tables 2, 3 and 4, we can conclude that the *Independence Assumption* holds well for most of the scenarios. The quality of our approximation is more significant especially for the scenarios with $\beta^n \geq 70\%$, the levels which are no less than what we would expect in practice. For those, we observe that the average absolute errors are only 0.04% and 0.04% for our approach for the two cases ("DLT: *non-critical*" and "DLT: *critical*" options, respectively) compared to 0.53% and 1.35% for the existing heuristic; while the maximum absolute errors for our approximation are limited to only 0.17% and 0.18% compared to 1.18% and 4.20% for the existing heuristic. On the other hand, based on the empirical results, we may conclude that the quality of the *Independence Assumption* diminishes as the *non-critical* class service level gets lower, and therefore the overall effect on the system of balance equations is more pronounced. Consequently, this leads to higher absolute errors in estimating the *critical* class service levels.

6.2 Optimization study

In this subsection, we compare the optimal policy parameters (S^*, S_c^*) found by using the *critical* class service levels calculated via our method against the Koçağa and Şen heuristic and the simulation study. In their study, Koçağa and Şen (2007) found the optimal policy parameters through the brute force search approach (though, they were able to limit the number of possible (S, S_c) pairs to consider). However, for both our method and the simulation study, we implement the optimization algorithm presented in Table 1, which requires the computation of steady-state probabilities only once. This is one of the main strengths of our algorithm. As in the previous subsection, we study two cases: "*non-critical* class has a DLT" and "*critical* class has a DLT". The results are provided in Tables 5 and 6. In their study, Koçağa and Şen (2007) showed that inventory rationing can result in significant inventory savings

Table 2 Performance of the approximation for a fixed service level of 99% ($L = 0.5$, $H = 0.1$)

λ^c	λ^n	S	S_c	DLT: *non-critical*						DLT: *critical*					
						Kocaga and Sen		Our method				Kocaga and Sen		Our method	
				β^n_{exact}(%)	$\beta^c_{simulation}$(%)	$\hat{\beta}^c$ (%)	AE^c (%)	$\hat{\beta}^c$ (%)	AE^c (%)	β^n_{exact}(%)	$\beta^c_{simulation}$(%)	$\hat{\beta}^c$ (%)	AE^c (%)	$\hat{\beta}^c$ (%)	AE^c (%)
1	4	5	3	37.96	99.95	99.76	0.19	99.93	0.02	30.84	99.93	99.76	0.17	99.94	0.01
2	4	6	3	51.84	99.81	99.27	0.54	99.83	0.02	46.95	99.77	99.27	0.50	99.79	0.02
3	4	7	4	62.48	99.68	98.92	0.76	99.73	0.05	60.25	99.66	98.91	0.75	99.72	0.06
4	4	8	3	70.64	99.62	98.77	0.85	99.66	0.04	70.64	99.63	98.77	0.86	99.67	0.04
5	4	9	3	76.93	99.58	98.76	0.82	99.62	0.04	78.51	99.64	98.77	0.87	99.68	0.04
6	4	10	3	81.80	99.58	98.84	0.74	99.62	0.04	84.36	99.69	98.85	0.84	99.72	0.03
7	4	11	3	85.60	99.60	98.96	0.64	99.63	0.03	88.67	99.73	98.98	0.75	99.76	0.03
8	4	12	3	88.57	99.64	99.09	0.55	99.66	0.02	91.81	99.79	99.13	0.66	99.80	0.01
9	4	13	3	90.90	99.67	99.22	0.45	99.69	0.02	94.09	99.83	99.27	0.56	99.84	0.01
10	4	14	3	92.74	99.71	99.34	0.37	99.72	0.01	95.74	99.87	99.40	0.47	99.88	0.01
11	4	15	3	94.20	99.75	99.45	0.30	99.76	0.01	96.93	99.90	99.51	0.39	99.90	0
12	4	16	3	95.36	99.78	99.54	0.24	99.79	0.01	97.79	99.92	99.61	0.31	99.93	0.01
2	4	8	1	99.28	99.83	99.63	0.20	99.83	0	97.56	99.73	99.57	0.16	99.73	0
3	4	8	2	90.57	99.74	98.28	0.46	99.75	0.01	89.46	99.69	99.27	0.42	99.72	0.03
4	4	8	3	70.64	99.62	98.77	0.85	99.66	0.04	70.64	99.63	98.77	0.86	99.67	0.04
5	4	8	4	41.42	99.43	98.02	1.41	99.59	0.16	43.35	99.54	98.02	1.52	99.65	0.11
6	4	8	5	16.26	99.23	96.97	2.26	99.57	0.34	18.51	99.48	96.97	2.51	99.66	0.18
7	4	8	6	3.72	99.10	95.54	3.56	99.63	0.53	4.77	99.47	95.54	3.93	99.72	0.25
8	4	8	7	0.37	99.21	93.68	5.53	99.74	0.53	0.55	99.56	93.67	5.89	99.80	0.24
Average absolute error (%)							1.09		0.10				1.18		0.06
Maximum absolute error (%)							5.53		0.53				5.89		0.25

(up to 30%) compared to one without rationing. Therefore, in order to prevent any repetition, we don't include the relative savings due to inventory rationing in these tables. Readers may refer to their study for a more detailed discussion.

In both tables, columns 5 and 9 show the relative percentage savings in inventory investment S due to using our method over the Koçağa and Şen heuristic for calculating the *critical* class service levels. For all the examples presented in Tables 5 and 6, the optimal policy parameters found by using our approach are identical to the ones found by the simulation studies.

In Table 5, we first fix $\lambda^c = 1$, $L = 0.5$, $H = 0.1$, $\bar{\beta}^n = 0.80$, $\bar{\beta}^c = 0.99$ and vary the *non-critical* demand rate λ^n between 1 and 10. For each case, ten different scenarios are considered. We observe that out of ten instances, our method provides considerable amount of savings in six instances for the first case, and in five instances for the second case. We also observe that although the pattern is not regular, there is a tendency that percentage savings decreases as λ^n increases.

Next, we fix $\lambda^c = 5$, $\lambda^n = 10$, $L = 2$, $H = 0.5$, $\bar{\beta}^n = 0.80$ and vary $\bar{\beta}^c$ between 0.90 and 0.995. The results are provided in Table 6. We see that in all the instances, our approach achieves the optimal policy parameters correctly and provides inventory savings over the Koçağa and Şen heuristic. As $\bar{\beta}^c$ increases, we also observe that the percentage savings of using our method tends to increase in both cases, and can be as high as 10.81% for the first case, and 7.69% for the second case. From these results, we may conclude that as the gap between the required service levels of the two demand class increases, the quality of Koçağa and Şen heuristic diminishes considerably, while our method continues to provide high quality approximations.

It is important to note that since there is no exact solution yet in the literature for computing the steady-state probabilities, we have to rely on approximation methods. Therefore it should be kept in mind that there may be cases of infeasibility due to not meeting the service level requirements for the *critical* customer class. Hence, at the end of the optimization routine, it may be essential to compare the gap between the estimated service level and the fill rate constraint for the *critical* customer class. However, as shown in the numerical study of the previous section, our proposed approximation provides high quality approximation for the majority of the instances. Apart from the extreme cases such that *non-critical* class experiences very low service levels, the absolute errors are considerably very low for many instances for the *critical* customer class. When Tables 2, 3 and 4 are combined, if we consider the cases with $\beta^n \geq 70\%$ (the levels which are no less than what we would expect in practice), we see that the maximum absolute errors for our approximation are limited to 0.17% and 0.18% for the two cases ("DLT: *non-critical*" and "DLT: *critical*" options, respectively). As a result, we may conclude that unless the gap between the estimated service level and the fill rate constraint for the *critical* customer class is very close, our proposed approximation may be used conveniently in the optimization routine. Otherwise, a single simulation study might be performed to determine the optimal policy parameters (S^*, S_c^*) using our optimization routine, which also requires the computation of steady-state probabilities only once (note that our optimization algorithm is provided for the general case, independent of the

Table 3 Performance of the approximation for service levels between 90% and 99% ($L = 0.5$, $H = 0.1$)

λ^c	λ^n	S	S_c	DLT: *non-critical*						DLT: *critical*					
						Kocaga and Sen		Our method				Kocaga and Sen		Our method	
				β^n_{exact}(%)	$\beta^c_{simulation}$(%)	$\hat{\beta}^c$ (%)	AE^c (%)	$\hat{\beta}^c$ (%)	AE^c (%)	β^n_{exact}(%)	$\beta^c_{simulation}$(%)	$\hat{\beta}^c$ (%)	AE^c (%)	$\hat{\beta}^c$ (%)	AE^c (%)
4	1	5	2	56.97	93.80	91.90	1.90	94.24	0.44	64.96	96.09	92.08	4.01	96.32	0.23
5	1	6	2	66.96	94.81	93.39	1.42	95.09	0.28	75.76	97.12	93.68	3.44	97.21	0.09
6	1	7	2	74.42	95.73	94.67	1.06	95.90	0.17	83.18	97.90	95.05	2.85	97.97	0.07
7	1	8	2	80.06	96.52	95.73	0.79	96.63	0.11	88.29	98.50	96.17	2.33	98.54	0.04
8	1	9	2	84.36	97.18	96.58	0.60	97.24	0.06	91.82	98.92	97.06	1.86	98.94	0.02
9	1	10	2	87.69	97.72	97.26	0.46	97.75	0.03	94.27	99.23	97.76	1.47	99.24	0.01
5	1	7	1	92.58	97.61	97.22	0.39	97.63	0.02	95.80	98.83	97.85	0.98	98.83	0
6	1	7	2	74.42	95.73	94.67	1.06	95.90	0.17	83.18	97.90	95.05	2.85	97.97	0.07
7	1	7	3	45.32	93.21	91.18	2.03	93.85	0.64	58.03	96.66	91.30	5.36	96.90	0.24
8	1	7	4	18.51	90.40	86.71	3.69	91.96	1.56	28.54	95.17	86.73	8.44	95.67	0.50
Average absolute error (%)							1.34		0.35				3.36		0.13
Maximum absolute error (%)							3.69		1.56				8.44		0.50

Table 4 Performance of the approximation with varying system parameters

λ^c	λ^n	S	S_c	L	H	DLT: *non-critical*						DLT: *critical*					
								Koçağa and Şen		Our method				Koçağa and Şen		Our method	
						$\beta^n_{exact}(\%)$	$\beta^c_{simulation}(\%)$	$\hat{\beta}^c$ (%)	AE^c (%)	$\hat{\beta}^c$ (%)	AE^c (%)	$\beta^n_{exact}(\%)$	$\beta^c_{simulation}(\%)$	$\hat{\beta}^c$ (%)	AE^c (%)	$\hat{\beta}^c$ (%)	AE^c (%)
6	2	7	2	0.5	0.1	66.78	94.86	92.25	2.61	95.17	0.31	74.42	96.78	92.58	4.20	96.96	0.18
6	2	8	2	0.5	0.1	81.56	97.73	96.55	1.18	97.83	0.10	87.05	98.71	96.78	1.93	98.77	0.06
6	2	9	2	0.5	0.1	90.91	99.09	98.61	0.48	99.12	0.03	94.21	99.54	98.74	0.80	99.55	0.01
6	2	10	2	0.5	0.1	95.99	99.67	99.49	0.18	99.68	0.01	97.69	99.85	99.55	0.30	99.85	0
6	2	11	2	0.5	0.1	98.40	99.89	99.83	0.06	99.89	0	99.17	99.95	99.86	0.09	99.96	0.01
1	1	5	2	1	0.5	80.88	99.50	98.60	0.90	99.57	0.07	80.88	99.58	98.56	1.02	99.57	0.01
2	1	5	2	1	0.5	54.38	94.81	90.08	4.73	95.62	0.81	67.67	97.78	91.47	6.31	97.86	0.08
3	1	5	2	1	0.5	32.08	83.77	73.78	9.99	86.22	2.45	54.38	94.20	78.47	15.73	94.37	0.17
4	1	5	2	1	0.5	17.36	69.61	54.38	15.23	73.57	3.96	42.32	88.85	62.69	26.16	89.15	0.30
5	1	5	2	1	0.5	8.84	56.14	36.68	19.46	60.93	4.79	32.08	82.02	47.37	34.65	82.54	0.52
1	1	5	2	1	0.5	80.88	99.50	98.60	0.90	99.57	0.07	80.88	99.58	98.56	1.02	99.57	0.01
1	2	5	2	1	0.5	67.67	99.36	96.86	2.50	99.47	0.11	54.38	98.94	96.43	2.51	98.98	0.04
1	3	5	2	1	0.5	54.38	99.28	94.84	4.44	99.42	0.14	32.08	98.51	94.02	4.49	98.63	0.12
1	4	5	2	1	0.5	42.32	99.23	92.74	6.49	99.39	0.16	17.36	98.40	91.70	6.69	98.56	0.16
1	5	5	2	1	0.5	32.08	99.21	90.72	8.49	99.39	0.18	8.84	98.40	89.65	8.75	98.66	0.26
10	4	14	3	0.5	0.1	92.74	99.71	99.34	0.37	99.72	0.01	95.74	99.87	99.40	0.47	99.88	0.01
10	4	14	3	0.5	0.2	94.86	99.83	99.53	0.30	99.84	0.01	98.63	99.98	99.75	0.23	99.98	0
10	4	14	3	0.5	0.3	96.81	99.90	99.73	0.17	99.91	0.01	99.72	100	99.96	0.04	100	0
10	4	14	3	0.5	0.4	97.75	99.94	99.86	0.08	99.95	0.01	99.97	100	100	0	100	0
10	4	14	3	0.5	0.5	98.63	99.95	99.93	0.02	99.98	0.03	100	100	100	0	100	0
Average absolute error (%)									3.93		0.66				5.77		0.10
Maximum absolute error (%)									19.46		4.79				34.65		0.52

approximation method being used. It can be used with any other heuristic, assuming the computed steady-state probabilities are accurate).

6.3 Benefit of integrating DLT into the threshold rationing policy

In this part of the numerical study, we investigate the benefit of integrating DLT into the threshold rationing policy. For comparison, we use the same examples studied in Tables 5 and 6. The results are provided in Tables 7 and 8 . In the second columns of both tables, we provide the optimal policy parameters found by the simulation study for the inventory system without incorporating the DLT into the current inventory rationing policy. None of the priority-demand classes shares advance demand information, therefore the distribution center is able to see demand realizations only at their corresponding due dates.

In Table 7, we observe that the benefit is realized in six instances for the "DLT: *non-critical*" option, which can be as high as 16.67%. On the other hand, we don't observe any benefit for the for the "DLT: *critical*" option. However, the situation changes in Table 8. In all the instances and both cases, we observe that integrating DLT into the current policy provides considerable savings. The average savings for the "DLT: *non-critical*" and "DLT: *critical*" options are 15.57% and 7.62%, respectively. Furthermore, although the pattern is irregular, the associated savings tend to decrease for both cases as $\bar{\beta}^c$ increases.

7 Conclusion

We have numerically demonstrated that our approach provides superior performance in estimating service levels than the existing heuristic for all the examples considered. When Tables 2, 3 and 4 are combined, the average absolute errors of the existing heuristic are 2.30% and 3.50% for the two cases ("DLT: *non-critical*" and "DLT: *critical*" options, respectively), while the average absolute errors are 0.38% and 0.09% for our approach. Furthermore, the maximum absolute errors of the existing heuristic are 19.46% and 34.64% for the two cases, while they are limited to 4.79% and 0.52% in our approach. On the other hand, when we consider the settings with $\beta^n \geq 70\%$ (the levels which are no less than what we would expect in practice), we observe that the average absolute errors are only 0.04% and 0.04% for our approach for the two cases compared to 0.53% and 1.35% for the existing heuristic; while the maximum absolute errors for our approximation are limited to only 0.17% and 0.18% compared to 1.18% and 4.20% for the existing heuristic.

In the service level optimization study, we show that our method can provide considerable inventory savings over the existing heuristic in most of the examples, which can be as high as 16.67% and 14.29% for the two cases. The overall effect of savings may even be more pronounced in practice, especially in environments (such that manufacturing industry or retail businesses) where hundreds to tens of thousands of stock units are being managed.

Table 5 Optimal policy parameters: heuristics versus simulation ($\lambda^c = 1$, $L = 0.5$, $H = 0.1$, $\bar{\beta}^n = 0.80$, $\bar{\beta}^c = 0.99$)

λ^n	DLT: *non-critical*				DLT: *critical*			
	$(S^*, S_c^*)_{simulation}$	$(S^*, S_c^*)_{\text{Koçağa \& Şen}}$	$(S^*, S_c^*)_{\text{our method}}$	% Saving	$(S^*, S_c^*)_{simulation}$	$(S^*, S_c^*)_{\text{Koçağa \& Şen}}$	$(S^*, S_c^*)_{\text{our method}}$	% Saving
1	(4,1)	(4,1)	(4,1)	–	(4,1)	(4,1)	(4,1)	–
2	(5,2)	(5,2)	(5,2)	–	(5,2)	(5,2)	(5,2)	–
3	(5,1)	(6,0)	(5,1)	16.67	(6,2)	(6,2)	(6,2)	–
4	(6,2)	(6,2)	(6,2)	–	(6,1)	(7,2)	(6,1)	14.29
5	(6,1)	(7,2)	(6,1)	14.29	(7,2)	(7,2)	(7,2)	–
6	(7,2)	(7,2)	(7,2)	–	(7,1)	(8,2)	(7,1)	12.50
7	(7,1)	(8,2)	(7,1)	12.50	(8,2)	(8,2)	(8,2)	–
8	(7,1)	(8,2)	(7,1)	12.50	(8,1)	(9,2)	(8,1)	11.11
9	(8,1)	(9,2)	(8,1)	11.11	(9,1)	(10,2)	(9,1)	10.00
10	(8,1)	(9,2)	(8,1)	11.11	(9,1)	(10,2)	(9,1)	10.00

Table 6 Optimal policy parameters: heuristics versus simulation ($\lambda^c = 5,\ \lambda^n = 10,\ L = 2,\ H = 0.5,\ \bar{\beta}^n = 0.80$)

$\bar{\beta}^c$	DLT: *non-critical*				DLT: *critical*			
	$(S^*, S_c^*)_{simulation}$	$(S^*, S_c^*)_{\text{Koçağa \& Şen}}$	$(S^*, S_c^*)_{\text{our method}}$	% Saving	$(S^*, S_c^*)_{simulation}$	$(S^*, S_c^*)_{\text{Koçağa \& Şen}}$	$(S^*, S_c^*)_{\text{our method}}$	% Saving
0.900	(31,1)	(32,2)	(31,1)	3.13	(34,1)	(35,0)	(34,1)	2.86
0.925	(31,1)	(33,0)	(31,1)	6.06	(34,1)	(35,2)	(34,1)	2.86
0.950	(31,1)	(34,0)	(31,1)	8.82	(34,1)	(36,3)	(34,1)	5.56
0.970	(32,2)	(35,5)	(32,2)	8.57	(35,2)	(36,3)	(35,2)	2.78
0.980	(32,2)	(35,5)	(32,2)	8.57	(35,2)	(37,4)	(35,2)	5.41
0.985	(32,2)	(36,6)	(32,2)	11.11	(35,2)	(37,4)	(35,2)	5.41
0.990	(33,3)	(36,6)	(33,3)	8.33	(36,3)	(38,5)	(36,3)	5.26
0.995	(33,3)	(37,7)	(33,3)	10.81	(36,3)	(39,6)	(36,3)	7.69

Table 7 Optimal policy parameters: with/without DLT ($\lambda^c = 1$, $L = 0.5$, $H = 0.1$, $\bar{\beta}^n = 0.80$, $\bar{\beta}^c = 0.99$)

λ^n	Without DLT	DLT: *non-critical*		DLT: *critical*	
	(S^*, S_c^*)	(S^*, S_c^*)	% Saving	(S^*, S_c^*)	% Saving
1	(4,1)	(4,1)	–	(4,1)	–
2	(5,2)	(5,2)	–	(5,2)	–
3	(6,2)	(5,1)	16.67	(6,2)	–
4	(6,1)	(6,2)	–	(6,1)	–
5	(7,2)	(6,1)	14.29	(7,2)	–
6	(7,1)	(7,2)	–	(7,1)	–
7	(8,1)	(7,1)	12.50	(8,2)	–
8	(8,1)	(7,1)	12.50	(8,1)	–
9	(9,1)	(8,1)	11.11	(9,1)	–
10	(9,1)	(8,1)	11.11	(9,1)	–

Table 8 Optimal policy parameters: with/without DLT ($\lambda^c = 5$, $\lambda^n = 10$, $L = 2$, $H = 0.5$, $\bar{\beta}^n = 0.80$)

$\bar{\beta}^c$	Without DLT	DLT: *non-critical*		DLT: *critical*	
	(S^*, S_c^*)	(S^*, S_c^*)	% Saving	(S^*, S_c^*)	% Saving
0.900	(37,1)	(31,1)	16.22	(34,1)	8.11
0.925	(37,1)	(31,1)	16.22	(34,1)	8.11
0.950	(37,1)	(31,1)	16.22	(34,1)	8.11
0.970	(38,2)	(32,2)	15.79	(35,2)	7.89
0.980	(38,2)	(32,2)	15.79	(35,2)	7.89
0.985	(38,2)	(32,2)	15.79	(35,2)	7.89
0.990	(38,2)	(33,3)	13.16	(36,3)	5.26
0.995	(39,3)	(33,3)	15.38	(36,3)	7.69

The limiting behavior of an infinitesimal probabilistic analysis has not caught much attention in the literature. However, as shown in this study, it allows us to study complex system dynamics which arose due to inventory rationing and demand lead times. Therefore, for the continuous review one-for-one policies, studying the limiting behavior of an infinitesimal analysis may open new research possibilities in the future.

One possible extension of the model would be to consider the case in which both priority demand classes have their own demand lead times, which might lead to additional savings in inventory management costs. Although the dimensionality of the state space will increase, our analysis can be directly extended to this setting as well. As suggestions for future research, it would be useful to extend the model to generally distributed lead times, and/or allow flexible delivery (early fulfillment of

orders before due dates). Furthermore, since the quality of our approach for estimating the steady-state probabilities is promising, this might further enable us to study the cost optimization model as well, which also has a practical importance.

Appendix 1: Proof of Proposition 1

Since a one-for-one policy is implemented, whenever a demand of any type occurs, it triggers a replenishment order of size one, and therefore R is incremented by one. If there is a delivery, number of units in the resupply is decremented by one. If it is a due date of the *non-critical* order, then R is unaffected and hence remains the same.

The dynamics for Y^n are also straightforward. Whenever a *non-critical* demand occurs, Y^n is incremented by one. On the other hand, Y^n is decremented by one only if it is a due date of the *non-critical* order. For all other cases, Y^n is unaffected.

For the *non-critical* backorders, the only situation in which B^n can be decremented is with the arrival of a delivery ($E_m =$ "v") when on-hand inventory prior to the delivery is S_c and there is at least one *non-critical* backorder. If on hand inventory equals S_c, then, by (1), $S = S_c + R_{m-1} - B^n_{m-1} - Y^n_{m-1}$. From this equation, by rearranging the terms we have $B^n_{m-1} = R_{m-1} - Y^n_{m-1} - (S - S_c)$. Since the number of *non-critical* backorders should be at least one, we also have $R_{m-1} - Y^n_{m-1} > S - S_g$. On the other hand, when a due date of a *non-critical* order comes ($E_m =$ "y") and on-hand inventory just prior to the due date is less than or equal to S_c, then B^n is incremented by one. On-hand inventory in this case is given by $\left[S - R_{m-1} + B^n_{m-1} + Y^n_{m-1}\right]^+ \le S_c$, due to (2); which is also equivalent to stating $S - R_{m-1} + B^n_{m-1} + Y^n_{m-1} \le S_c$. Rearranging the terms, we have $R_{m-1} - Y^n_{m-1} \ge S - S_c + B^n_{m-1}$. But B^n_{m-1} is a non-negative variable. Hence we should have $R_{m-1} - Y^n_{m-1} \ge S - S_c$. □

Appendix 2: Proof of Proposition 2

The result $\beta^n(S, S_c) \ge \beta^n(S, S'_c)$ is immediately follows from (11).

To prove $\beta_c(S, S_c) \le \beta_c(S, S'_c)$, let us consider two systems with identical event sequences $\left\{(m, T_m, E_m); m = 1, 2, 3, \ldots\right\}$. In the first system, the policy parameters are (S, S_c) and the resulting states are given by $\left\{(R_m, B^n_m, Y^n_m); m = 1, 2, 3, \ldots\right\}$. In the second system, the policy parameters are (S, S'_c) with $S'_c > S_c$ and the resulting states are given by $\left\{(R'_m, B^{n\prime}_m, Y^{n\prime}_m); m = 1, 2, 3, \ldots\right\}$. We conjecture that $R'_m = R_m$, $Y^{n\prime}_m = Y^n_m$ and $B^{n\prime}_m \ge B^n_m$ for all m. With reference to Proposition 1, we can immediately establish $R'_m = R_m$ and $Y^{n\prime}_m = Y^n_m$ for all m. We will prove $B^{n\prime}_m \ge B^n_m$ by induction. For $m = 1, \ldots, S - S_c$, it is clear that $B^{n\prime}_m = B^n_m = 0$. For induction, we first assume that $B^{n\prime}_m \ge B^n_m$ is true for some m, and then validate the result for $m + 1$. Since $(S - 1, S)$ policy is followed, the number of backorders can change at most by one unit per event. Therefore, if $B^{n\prime}_m > B^n_m$, $B^{n\prime}_{m+1} \ge B^n_{m+1}$ is immediate. Else, $B^{n\prime}_m = B^n_m$ and it is left to prove $B^{n\prime}_{m+1} \ge B^n_{m+1}$. It suffices to show that $B^{n\prime}_{m+1} < B^n_{m+1}$ is not possible. First, let us suppose $B^n_{m+1} = B^n_m + 1$. This can only

happen if $R_m - Y_m^n \geq S - S_c$ and $E_{m+1} =$ "y" (the *non-critical* due date comes). Since $S_c' > S_c$, we have $R_m' - Y_m^{n\prime} = R_m - Y_m^n \geq S - S_c'$. Since $E_{m+1} =$ "y", we will have $B_{m+1}^{n\prime} = B_m^{n\prime} + 1 = B_m^n + 1 = B_{m+1}^n$. Now let us suppose $B_{m+1}^{n\prime} = B_m^{n\prime} - 1$. This is possible only if $B_m^{n\prime} = R_m' - Y_m^{n\prime} - (S - S_c')$ and $E_{m+1} =$ "v" (a delivery occurs). But this would imply $B_m^n = B_m^{n\prime} = R_m' - Y_m^{n\prime} - (S - S_c') > R_m - Y_m^n - (S - S_c)$. However, the condition $B_m^n > R_m - Y_m^n - (S - S_c)$ is not possible due to (5). Therefore we should have $B_{m+1}^{n\prime} \geq B_{m+1}^n$. Proof by induction is completed. Furthermore, conditioned on $B_m^{n\prime} \geq B_m^n$, by (2), at any point in time on-hand inventory in the second system will always be greater than or equal to the on-hand inventory in the second system. Hence, the *critical* class fill rate for the second system must be at least as high as for the first system. □

Appendix 3: Proof of Proposition 3

Let us consider two cases with (S, S_c) and (S', S_c') such that $\Delta = S - S_c = S' - S_c'$. For any given sample path $\{(m, T_m, E_m); m = 1, 2, 3, \ldots\}$, (R_m, B_m^n, Y_m^n) will always be identical to $(R_m', B_m^{n\prime}, Y_m^{n\prime})$ for an arbitrary m. This is because, due to Proposition 1 the sample path dynamics depend only on Δ, the difference between the target inventory S and the threshold level S_c. Therefore, at any point t in time, system state ξ_t will be identical to ξ_t'. Since this is also true for all sample paths, we have $\pi_{(r,b^n,y^n)}(S, S_c) = \pi_{(r,b^n,y^n)}(S', S_c'), (r, b^n, y^n) \in Z_0 \times Z_0 \times Z_0$. □

Appendix 4: Proof of Lemma 1

By definition,

$$
\begin{aligned}
\psi_u(\Delta + k, k) &= \sum_{\substack{(r,b^n,y^n) \in \mathbb{F}_{(\Delta+k,k)} \\ (\Delta + k - r + b^n + y^n)^+ = 0 \\ (r - b^n - y^n - \Delta - k)^+ = u}} \pi_{(r,b^n,y^n)}(\Delta + k, k) \\
&= \sum_{\substack{(r,b^n,y^n) \in \mathbb{F}_{(\Delta,0)} \\ (\Delta + k - r + b^n + y^n)^+ = 0 \\ (r - b^n - y^n - \Delta - k)^+ = u}} \pi_{(r,b^n,y^n)}(\Delta, 0) \text{ (by Proposition 3)}, \\
&= \sum_{\substack{(r,b^n,y^n) \in \mathbb{F}_{(\Delta,0)} \\ (\Delta - r + b^n + y^n)^+ = 0 \\ (r - b^n - y^n - \Delta)^+ = u + k}} \pi_{(r,b^n,y^n)}(\Delta, 0) \\
&= \psi_{u+k}(\Delta, 0).
\end{aligned}
$$

□

Appendix 5: Proof of Lemma 2

By definition,

$$\begin{aligned}
\beta^c(\Delta+k,k) &= 1-\varphi_0(\Delta+k,k) \text{ (with reference to Eqs.(7) and (8))}\\
&= 1-\sum_{u=0}^{\infty}\psi_u(\Delta+k,k)\\
&= 1-\sum_{u=0}^{\infty}\psi_{u+k}(\Delta,0) \text{ (by Lemma 1)},\\
&= 1-\sum_{u=0}^{\infty}\psi_u(\Delta,0)\ +\ \sum_{u=0}^{k-1}\psi_u(\Delta,0)\\
&= 1-\varphi_0(\Delta,0)\ +\ \sum_{u=0}^{k-1}\psi_u(\Delta,0)\\
&= \beta^c(\Delta,0)\ +\ \sum_{u=0}^{k-1}\psi_u(\Delta,0).
\end{aligned}$$

□

Appendix 6: Proof of Theorem 2

Starting from the initial state (0, 0, 0) at time $t=0$, let

$$P_{(\bar{r},\bar{b}^n,\bar{y}^n),(r,b^n,y^n)}(t,t') = P\big[\xi_{t'}=(r,b^n,y^n)\mid \xi_0=(0,0,0),\ \xi_t=(\bar{r},\bar{b}^n,\bar{y}^n)\big].$$

One-step transition probabilities will be solved for a general system state $(r,b^n,y^n)\in\mathbb{F}_{(S,S_c)}$ with $b^n\geq 1$, $y^n\geq 1$, and $r-b^n-y^n>S-S_c$ (hence, $OH<S_c$). Other system states can be solved similarly. By conditioning on the state of the system at time t, there are five possible ways to reach state (r,b^n,y^n) in at most one transition over the next infinitesimal τ time units: a *non-critical* demand occurs, a *critical* demand occurs, a delivery is received from the resupply, a due date of a *non-critical* order comes, or nothing happens. Probabilities of two or more events happening during $(t,t+\tau]$ are captured within the term $o(\tau)$.

(a) A *non-critical* demand occurs:

$$\begin{aligned}
&P_{(r-1,b^n,y^n-1),(r,b^n,y^n)}(t,t+\tau)\\
&\quad= P\big[\text{only a } \textit{non-critical} \text{ demand occurs during } (t,t+\tau];\ \text{all } r-1 \text{ units in}\\
&\qquad\text{the resupply at time } t \text{ are still in the resupply at time } t+\tau;\\
&\qquad\text{all } y^n-1 \text{ orders that are "not yet due" at time } t\\
&\qquad\text{are still "not yet due" at time } t+\tau \mid \xi_t=(r-1,b^n,y^n-1),\\
&\qquad\xi_0=(0,0,0)\big]\ +\ o(\tau).
\end{aligned}\tag{29}$$

By the *Independence Assumption* and using (22) and (25), right-hand side of (29) becomes

$$\begin{aligned}
&P_{(r-1,b^n,y^n-1),(r,b^n,y^n)}(t,t+\tau)\\
&\quad= \lambda^n \tau e^{-\lambda\tau}\, q_{t,\tau}(r-y^n \mid r-y^n)\, \tilde{q}_{t,\tau}(y^n-1 \mid y^n-1) \,+\, o(\tau)\\
&\quad= \lambda^n \tau e^{-\lambda\tau} \begin{pmatrix} r-y^n \\ r-y^n \end{pmatrix} p(\tau)^{r-y^n} \left[1-p(\tau)\right]^0\\
&\qquad\cdot \begin{pmatrix} y^n-1 \\ y^n-1 \end{pmatrix} \tilde{p}(\tau)^{y^n-1} \left[1-\tilde{p}(\tau)\right]^0 \,+\, o(\tau)\\
&\quad= \lambda^n \tau e^{-\lambda\tau}\, p(\tau)^{r-y^n}\, \tilde{p}(\tau)^{y^n-1} \,+\, o(\tau).
\end{aligned} \tag{30}$$

Note that we used $q_{t,\tau}(r-y^n \mid r-y^n)$ rather than $q_{t,\tau}(r-1 \mid r-1)$. This is because as discussed earlier, conditioned on being at state $\xi_t = (r, b^n, y^n)$, delivery process over the next τ time units is determined by the elements of the set $\mathbb{U}(t)\backslash\mathbb{V}(t)$.

(b) A *critical* demand occurs:

$$\begin{aligned}
&P_{(r-1,b^n,y^n),(r,b^n,y^n)}(t,t+\tau)\\
&\quad= P\big[\text{only a } \textit{critical} \text{ demand occurs during } (t,t+\tau];\ \text{all } r-1 \text{ units in}\\
&\qquad \text{the resupply at time } t \text{ are still in the resupply at time } t+\tau;\\
&\qquad \text{all } y^n \text{ orders that are "not yet due" at time } t\\
&\qquad \text{are still "not yet due" at time } t+\tau \mid \xi_t = (r-1, b^n, y^n),\\
&\qquad \xi_0 = (0,0,0)\big] \,+\, o(\tau).
\end{aligned} \tag{31}$$

By the *Independence Assumption* and using (22) and (25), right-hand side of (31) becomes

$$\begin{aligned}
&P_{(r-1,b^n,y^n),(r,b^n,y^n)}(t,t+\tau)\\
&\quad= \lambda^c \tau e^{-\lambda\tau}\, q_{t,\tau}(r-y^n-1 \mid r-y^n-1)\, \tilde{q}_{t,\tau}(y^n \mid y^n) \,+\, o(\tau)\\
&\quad= \lambda^c \tau e^{-\lambda\tau} \begin{pmatrix} r-y^n-1 \\ r-y^n-1 \end{pmatrix} p(\tau)^{r-y^n-1} \left[1-p(\tau)\right]^0\\
&\qquad\cdot \begin{pmatrix} y^n \\ y^n \end{pmatrix} \tilde{p}(\tau)^{y^n} \left[1-\tilde{p}(\tau)\right]^0 \,+\, o(\tau)\\
&\quad= \lambda^c \tau e^{-\lambda\tau}\, p(\tau)^{r-y^n-1}\, \tilde{p}(\tau)^{y^n} \,+\, o(\tau).
\end{aligned} \tag{32}$$

(c) The delivery from the resupply:

$$
\begin{aligned}
&P_{(r+1,b^n,y^n),(r,b^n,y^n)}(t,t+\tau)\\
&\quad = P\Big[\text{no demand occurs during } (t,t+\tau];\ \text{among the } r+1 \text{ units in}\\
&\qquad \text{the resupply at time } t,\ \text{only one of them is received during}\\
&\qquad (t,t+\tau];\ \text{all } y^n \text{ orders that are "not yet due" at time } t\\
&\qquad \text{are still "not yet due" at time } t+\tau \mid \xi_t = (r+1,b^n,y^n),\\
&\qquad \xi_0 = (0,0,0)\Big] \ + o(\tau).
\end{aligned} \tag{33}
$$

By the *Independence Assumption* and using (22) and (25), right-hand side of (33) becomes

$$
\begin{aligned}
&P_{(r+1,b^n,y^n),(r,b^n,y^n)}(t,t+\tau)\\
&\quad = e^{-(\lambda^n+\lambda^c)\tau}\ q_{t,\tau}(r-y^n \mid r-y^n+1)\ \tilde{q}_{t,\tau}(y^n \mid y^n) \ + o(\tau)\\
&\quad = e^{-(\lambda^n+\lambda^c)\tau} \begin{pmatrix} r-y^n+1 \\ r-y^n \end{pmatrix} p(\tau)^{r-y^n}\ \big[1-p(\tau)\big]\\
&\qquad \cdot \begin{pmatrix} y^n \\ y^n \end{pmatrix} \tilde{p}(\tau)^{y^n}\ \big[1-\tilde{p}(\tau)\big]^0 \ + o(\tau)\\
&\quad = e^{-(\lambda^n+\lambda^c)\tau}\ (r-y^n+1)p(\tau)^{r-y^n}\big[1-p(\tau)\big]\ \tilde{p}(\tau)^{y^n} \ + o(\tau).
\end{aligned} \tag{34}
$$

Note that b^n has not changed because $OH(t) < S_c$ and therefore none of the existing *non-critical* backorders are cleared, if any.

(d) The *non-critical* order is due:

$$
\begin{aligned}
&P_{(r,b^n-1,y^n+1),(r,b^n,y^n)}(t,t+\tau)\\
&\quad = P\Big[\text{no demand occurs during } (t,t+\tau];\ \text{all } r \text{ units in the}\\
&\qquad \text{resupply at time } t \text{ are still in the resupply at time } t+\tau;\\
&\qquad \text{among the } y^n+1 \text{ orders that are "not yet due" at time } t,\\
&\qquad y^n \text{ of them are still "not yet due" at } t+\tau\\
&\qquad \mid \xi_t = (r,b^n-1,y^n+1),\ \xi_0 = (0,0,0)\Big] \ + o(\tau).
\end{aligned} \tag{35}
$$

By the *Independence Assumption* and using (22) and (25), right-hand side of (35) becomes

$$
\begin{aligned}
&P_{(r,b^n-1,y^n+1),(r,b^n,y^n)}(t,t+\tau)\\
&\quad = e^{-(\lambda^n+\lambda^c)\tau}\ q_{t,\tau}(r-y^n-1 \mid r-y^n-1)\ \tilde{q}_{t,\tau}(y^n \mid y^n+1) \ + o(\tau)\\
&\quad = e^{-(\lambda^n+\lambda^c)\tau} \begin{pmatrix} r-y^n-1 \\ r-y^n-1 \end{pmatrix} p(\tau)^{r-y^n-1}\ \big[1-p(\tau)\big]^0\\
&\qquad \cdot \begin{pmatrix} y^n+1 \\ y^n \end{pmatrix} \tilde{p}(\tau)^{y^n}\ \big[1-\tilde{p}(\tau)\big]^1 \ + o(\tau)\\
&\quad = e^{-(\lambda^n+\lambda^c)\tau}\ p(\tau)^{r-y^n-1}\ (y^n+1)\tilde{p}(\tau)^{y^n}\ \big[1-\tilde{p}(\tau)\big] + o(\tau).
\end{aligned} \tag{36}
$$

(e) Nothing happens:

$$
\begin{aligned}
&P_{(r,b^n,y^n),(r,b^n,y^n)}(t,t+\tau) \\
&\quad = P\Big[\text{no demand occurs during } (t,t+\tau]\text{; all } r \text{ units in} \\
&\qquad \text{the resupply at time } t \text{ are still in the resupply at } t+\tau \\
&\qquad \text{units; all } y^n \text{ orders that are "not yet due" at time } t \\
&\qquad \text{are still "not yet due" at time } t+\tau \mid \xi_t = (r,b^n,y^n), \\
&\qquad \xi_0 = (0,0,0)\Big] \; + \; o(\tau).
\end{aligned} \tag{37}
$$

By the *Independence Assumption* and using (22) and (25), right-hand side of (37) becomes

$$
\begin{aligned}
P_{(r,b^n,y^n),(r,b^n,y^n)}(t,t+\tau) &= e^{-(\lambda^n+\lambda^c)\tau}\; q_{t,\tau}(r-y^n \mid r-y^n)\; \tilde{q}_{t,\tau}(y^n \mid y^n) \; + \; o(\tau) \\
&= e^{-(\lambda^n+\lambda^c)\tau} \begin{pmatrix} r-y^n \\ r-y^n \end{pmatrix} p(\tau)^{r-y^n} \left[1-p(\tau)\right]^0 \\
&\qquad \cdot \begin{pmatrix} y^n \\ y^n \end{pmatrix} \tilde{p}(\tau)^{y^n} \left[1-\tilde{p}(\tau)\right]^0 \; + \; o(\tau) \\
&= e^{-(\lambda^n+\lambda^c)\tau}\, p(\tau)^{r-y^n}\, \tilde{p}(\tau)^{y^n} \; + \; o(\tau).
\end{aligned} \tag{38}
$$

By conditioning on the state of the system at time t:

$$
P_{(0,0,0),(r,b^n,y^n)}(0,t') \; = \sum_{(\bar{r},\bar{b}^n,\bar{y}^n) \in \mathbb{F}_{(S,S_c)}} P_{(0,0,0),(\bar{r},\bar{b}^n,\bar{y}^n)}(0,t) \cdot P_{(\bar{r},\bar{b}^n,\bar{y}^n),(r,b^n,y^n)}(t,t'). \tag{39}
$$

Then using (39), the probability of being at system state (r,b^n,y^n) at time $t' = t+\tau$ can be written as

$$
\begin{aligned}
&P_{(0,0,0),(r,b^n,y^n)}(0,t+\tau) \\
&\quad = P_{(0,0,0),(r-1,b^n,y^n-1)}(0,t) \cdot P_{(r-1,b^n,y^n-1),(r,b^n,y^n)}(t,t+\tau) \\
&\qquad + \; P_{(0,0,0),(r-1,b^n,y^n)}(0,t) \cdot P_{(r-1,b^n,y^n),(r,b^n,y^n)}(t,t+\tau) \\
&\qquad + \; P_{(0,0,0),(r+1,b^n,y^n)}(0,t) \cdot P_{(r+1,b^n,y^n),(r,b^n,y^n)}(t,t+\tau) \\
&\qquad + \; P_{(0,0,0),(r,b^n-1,y^n+1)}(0,t) \cdot P_{(r,b^n-1,y^n+1),(r,b^n,y^n)}(t,t+\tau) \\
&\qquad + \; P_{(0,0,0),(r,b^n,y^n)}(0,t) \cdot P_{(r,b^n,y^n),(r,b^n,y^n)}(t,t+\tau) \\
&\qquad + \; o(\tau).
\end{aligned} \tag{40}
$$

Under the *Independence Assumption*, each of the one-step transition probabilities on the right-hand side of (40) can be determined using the results in (30), (32), (34), (36), and (38).

Subtracting $P_{(0,0,0),(r,b^n,y^n)}(0,t)$ from both sides and taking the limits as $\tau \to 0$:

$$
\begin{aligned}
&\lim_{\tau\to 0}\frac{P_{(0,0,0),(r,b^n,y^n)}(0,t+\tau)-P_{(0,0,0),(r,b^n,y^n)}(0,t)}{\tau}\\
&\quad= P_{(0,0,0),(r-1,b^n,y^n-1)}(0,t)\cdot\lim_{\tau\to 0}\frac{P_{(r-1,b^n,y^n-1),(r,b^n,y^n)}(t,t+\tau)}{\tau}\\
&\qquad+ P_{(0,0,0),(r-1,b^n,y^n)}(0,t)\cdot\lim_{\tau\to 0}\frac{P_{(r-1,b^n,y^n),(r,b^n,y^n)}(t,t+\tau)}{\tau}\\
&\qquad+ P_{(0,0,0),(r+1,b^n,y^n)}(0,t)\cdot\lim_{\tau\to 0}\frac{P_{(r+1,b^n,y^n),(r,b^n,y^n)}(t,t+\tau)}{\tau}\\
&\qquad+ P_{(0,0,0),(r,b^n-1,y^n+1)}(0,t)\cdot\lim_{\tau\to 0}\frac{P_{(r,b^n-1,y^n+1),(r,b^n,y^n)}(t,t+\tau)}{\tau}\\
&\qquad- P_{(0,0,0),(r,b^n,y^n)}(0,t)\cdot\lim_{\tau\to 0}\frac{\left(1-P_{(r,b^n,y^n),(r,b^n,y^n)}(t,t+\tau)\right)}{\tau}\\
&\qquad+\lim_{\tau\to 0}\frac{o(\tau)}{\tau}.
\end{aligned}
\tag{41}
$$

The left-hand side of Eq. (41) is

$$
\lim_{\tau\to 0}\frac{P_{(0,0,0),(r,b^n,y^n)}(0,t+\tau)-P_{(0,0,0),(r,b^n,y^n)}(0,t)}{\tau}=P'_{(0,0,0),(r,b^n,y^n)}(0,t). \tag{42}
$$

To determine the right-hand side, first we need to determine the limits as $\tau\to 0$. Limits as $\tau\to 0$:
(a) Using (30),

$$
\lim_{\tau\to 0}\frac{P_{(r-1,b^n,y^n-1),(r,b^n,y^n)}(t,t+\tau)}{\tau}=\lim_{\tau\to 0}\frac{\lambda^n\tau e^{-\lambda\tau}\,p(\tau)^{r-y^n}\,\tilde{p}(\tau)^{y^n-1}}{\tau}+\lim_{\tau\to 0}\frac{o(\tau)}{\tau},
$$

substituting values of $p(\tau)$ from (21) and $\tilde{p}(\tau)$ from (24),

$$
\begin{aligned}
&=\lim_{\tau\to 0}\frac{\lambda^n\tau e^{-\lambda\tau}\left[1-\frac{(\lambda^n+\lambda^c)\tau}{\lambda^c L+\lambda^n(L-H)}\right]^{r-y^n}\left[1-\frac{\tau}{H}\right]^{y^n-1}}{\tau}\\
&=\lambda^n.
\end{aligned}
\tag{43}
$$

(b) Using (32),

$$
\lim_{\tau\to 0}\frac{P_{(r-1,b^n,y^n),(r,b^n,y^n)}(t,t+\tau)}{\tau}=\lim_{\tau\to 0}\frac{\lambda^c\tau e^{-\lambda\tau}\,p(\tau)^{r-y^n-1}\,\tilde{p}(\tau)^{y^n}}{\tau}+\lim_{\tau\to 0}\frac{o(\tau)}{\tau},
$$

substituting values of $p(\tau)$ from (21) and $\tilde{p}(\tau)$ from (24),

$$
\begin{aligned}
&=\lim_{\tau\to 0}\frac{\lambda^c\tau e^{-\lambda\tau}\left[1-\frac{(\lambda^n+\lambda^c)\tau}{\lambda^c L+\lambda^n(L-H)}\right]^{r-y^n-1}\left[1-\frac{\tau}{H}\right]^{y^n}}{\tau}\\
&=\lambda^c.
\end{aligned}
\tag{44}
$$

(c) Using (34),

$$\lim_{\tau\to 0}\frac{P_{(r+1,b^n,y^n),(r,b^n,y^n)}(t,t+\tau)}{\tau}$$
$$=\lim_{\tau\to 0}\frac{e^{-(\lambda^n+\lambda^c)\tau}\,(r-y^n+1)p(\tau)^{r-y^n}[1-p(\tau)]\,\tilde{p}(\tau)^{y^n}}{\tau}+\lim_{\tau\to 0}\frac{o(\tau)}{\tau},$$

substituting values of $p(\tau)$ from (21) and $\tilde{p}(\tau)$ from (24),

$$=\lim_{\tau\to 0}\frac{e^{-(\lambda^n+\lambda^c)\tau}\,(r-y^n+1)\left[1-\frac{(\lambda^n+\lambda^c)\tau}{\lambda^c L+\lambda^n(L-H)}\right]^{r-y^n}\left[\frac{(\lambda^n+\lambda^c)\tau}{\lambda^c L+\lambda^n(L-H)}\right]\left[1-\frac{\tau}{H}\right]^{y^n}}{\tau},$$

replacing the corresponding terms with μ from (27) and ϑ from (28),

$$\begin{aligned}&=\lim_{\tau\to 0}\frac{e^{-(\lambda^n+\lambda^c)\tau}\,(r-y^n+1)[1-\mu\tau]^{r-y^n}\mu\tau\,[1-\vartheta\tau]^{y^n}}{\tau}\\&=(r-y^n+1)\mu\lim_{\tau\to 0}e^{-(\lambda^n+\lambda^c)\tau}\lim_{\tau\to 0}[1-\mu\tau]^{r-y^n}\lim_{\tau\to 0}[1-\vartheta\tau]^{y^n}\\&=(r-y^n+1)\mu.\end{aligned}\tag{45}$$

(d) Using (36),

$$\lim_{\tau\to 0}\frac{P_{(r,b^n-1,y^n+1),(r,b^n,y^n)}(t,t+\tau)}{\tau}$$
$$=\lim_{\tau\to 0}\frac{e^{-(\lambda^n+\lambda^c)\tau}\,p(\tau)^{r-y^n-1}\,(y^n+1)\tilde{p}(\tau)^{y^n}\left[1-\tilde{p}(\tau)\right]}{\tau}+\lim_{\tau\to 0}\frac{o(\tau)}{\tau},$$

substituting values of $p(\tau)$ from (21) and $\tilde{p}(\tau)$ from (24),

$$=\lim_{\tau\to 0}\frac{e^{-(\lambda^n+\lambda^c)\tau}\left[1-\frac{(\lambda^n+\lambda^c)\tau}{\lambda^c L+\lambda^n(L-H)}\right]^{r-y^n-1}(y^n+1)\left[1-\frac{\tau}{H}\right]^{y^n}(\frac{\tau}{H})}{\tau},$$

replacing the corresponding terms with μ from (27) and ϑ from (28),

$$\begin{aligned}&=\lim_{\tau\to 0}\frac{e^{-(\lambda^n+\lambda^c)\tau}\,[1-\mu\tau]^{r-y^n-1}\,(y^n+1)[1-\vartheta\tau]^{y^n}\vartheta\tau}{\tau}\\&=(y^n+1)\vartheta\lim_{\tau\to 0}e^{-(\lambda^n+\lambda^c)\tau}\lim_{\tau\to 0}[1-\mu\tau]^{r-y^n-1}\lim_{\tau\to 0}[1-\vartheta\tau]^{y^n}\\&=(y^n+1)\vartheta.\end{aligned}\tag{46}$$

(e) Using (38),

$$\lim_{\tau\to 0}\frac{\left(1-P_{(r,b^n,y^n),(r,b^n,y^n)}(t,t+\tau)\right)}{\tau}$$
$$=\lim_{\tau\to 0}\frac{1-e^{-(\lambda^n+\lambda^c)\tau}\,p(\tau)^{r-y^n}\,\tilde{p}(\tau)^{y^n}}{\tau}+\lim_{\tau\to 0}\frac{o(\tau)}{\tau},$$

substituting values of $p(\tau)$ from (21) and $\tilde{p}(\tau)$ from (24),

$$= \lim_{\tau \to 0} \frac{1 - e^{-(\lambda^n+\lambda^c)\tau} \left[1 - \frac{(\lambda^n+\lambda^c)\tau}{\lambda^c L + \lambda^n (L-H)}\right]^{r-y^n} \left[1 - \frac{\tau}{H}\right]^{y^n}}{\tau},$$

replacing the corresponding terms with μ from (27) and ϑ from (28),

$$= \lim_{\tau \to 0} \frac{1 - e^{-(\lambda^n+\lambda^c)\tau} [1 - \mu\tau]^{r-y^n} [1 - \vartheta\tau]^{y^n}}{\tau}.$$

Since there is no point of discontinuity, we can apply the *L'Hôpital's rule*. Then the right-hand side becomes

$$\begin{aligned} &= \lim_{\tau \to 0} (\lambda^n + \lambda^c) e^{-(\lambda^n+\lambda^c)\tau} [1 - \mu\tau]^{r-y^n} [1 - \vartheta\tau]^{y^n} \\ &\quad + \lim_{\tau \to 0} -e^{-(\lambda^n+\lambda^c)\tau} \big\{ (r - y^n)[1 - \mu\tau]^{r-y^n-1} (-\mu)[1 - \vartheta\tau]^{y^n} \\ &\quad + (y^n)[1 - \vartheta\tau]^{y^n-1} (-\vartheta)[1 - \mu\tau]^{r-y^n} \big\} \\ &= (\lambda^n + \lambda^c) + (r - y^n)\mu + y^n \vartheta. \end{aligned} \tag{47}$$

Plugging the values of these limits into Eq. (41), we have

$$\begin{aligned} P'_{(0,0,0),(r,b^n,y^n)}(0,t) = &\, P_{(0,0,0),(r-1,b^n,y^n-1)}(0,t)\, \lambda^n \\ &+ P_{(0,0,0),(r-1,b^n,y^n)}(0,t)\, \lambda^c \\ &+ P_{(0,0,0),(r+1,b^n,y^n)}(0,t)\, (r - y^n + 1)\mu \\ &+ P_{(0,0,0),(r,b^n-1,y^n+1)}(0,t)\, (y^n + 1)\vartheta \\ &- P_{(0,0,0),(r,b^n,y^n)}(0,t) \left[(\lambda^n + \lambda^c) + (r - y^n)\mu + y^n \vartheta\right]. \end{aligned} \tag{48}$$

Assuming the steady-state distributions exist, taking the limits as $t \to \infty$:

$$\begin{aligned} &\lim_{t \to \infty} P'_{(0,0,0),(r,b^n,y^n)}(0,t) \\ &\quad = \lim_{t \to \infty} \big\{ P_{(0,0,0),(r-1,b^n,y^n-1)}(0,t)\, \lambda^n \\ &\qquad + P_{(0,0,0),(r-1,b^n,y^n)}(0,t)\, \lambda^c \\ &\qquad + P_{(0,0,0),(r+1,b^n,y^n)}(0,t)\, (r - y^n + 1)\mu \\ &\qquad + P_{(0,0,0),(r,b^n-1,y^n+1)}(0,t)\, (y^n + 1)\vartheta \\ &\qquad - P_{(0,0,0),(r,b^n,y^n)}(0,t) \left[(\lambda^n + \lambda^c) + (r - y^n)\mu + y^n \vartheta\right] \big\} \\ &\quad = \pi_{(r-1,b^n,y^n-1)}\, \lambda^n + \pi_{(r-1,b^n,y^n)}\, \lambda^c \\ &\qquad + \pi_{(r+1,b^n,y^n)}\, (r - y^n + 1)\mu + \pi_{(r,b^n-1,y^n+1)}\, (y^n + 1)\vartheta \\ &\qquad - \pi_{(r,b^n,y^n)} \left[(\lambda^n + \lambda^c) + (r - y^n)\mu + y^n \vartheta\right]. \end{aligned} \tag{49}$$

$P_{(0,0,0),(r,b^n,y^n)}(0,t)$ is bounded by 0 and 1 for all t. Hence, if $\lim_{t \to \infty} P'_{(0,0,0),(r,b^n,y^n)}(0,t)$ converges, it must converge to 0. But as shown, the right-hand side of (49) has a fixed value. Hence the left-hand side of (49) is zero. Rearranging the terms, we have

$$\begin{aligned}&\left[(\lambda^n + \lambda^c) + (r - y^n)\mu + y^n\vartheta\right]\pi_{(r,b^n,y^n)}\\&\quad = \lambda^n\pi_{(r-1,b^n,y^n-1)} + \lambda^c\pi_{(r-1,b^n,y^n)}\\&\qquad + (r - y^n + 1)\mu\pi_{(r+1,b^n,y^n)} + (y^n + 1)\vartheta\pi_{(r,b^n-1,y^n+1)}.\end{aligned}$$

□

References

Arslan H, Graves SC, Roemer T (2007) A single-product inventory model for multiple demand classes. Manag Sci 53(9):1486–1500

Basten RJI, Ryan JK (2019) The value of maintenance delay flexibility for improved spare parts inventory management. Eur J Oper Res 278(2):646–657

Benjaafar S, Cooper WL, Mardan S (2011) Production-inventory systems with imperfect advance demand information and updating. Naval Res Log 58(2):88–106

Bernstein F, DeCroix GA (2015) Advance demand information in a multiproduct system. Manuf Serv Oper Manag 17(1):52–65

Boyacı T, Özer Ö (2010) Information acquisition for capacity planning via pricing and advance selling: when to stop and act? Oper Res 58(5):1328–1349

Dekker R, Kleijn MJ, de Rooij PJ (1998) A spare parts stocking policy based on equipment criticality. Int J Prod Econ 56–57:69–77

Dekker R, Hill RM, Kleijn MJ, Teunter RH (2002) On the (S-1, S) lost sales inventory model with priority demand classes. Naval Res Log 49(6):593–610

Deshpande V, Cohen MA, Donohue K (2003) A threshold rationing policy for service differentiated demand classes. Manag Sci 49(6):683–703

Donselaar KH, van Kopczak LR, Wouters MJF (2001) The use of advance demand information in a project-based supply chain. Eur J Oper Res 130(3):519–538

Gabor AF, van Vianen L, Yang G et al (2018) A base-stock inventory model with service differentiation and response time guarantees. Eur J Oper Res 269(3):900–908

Gallego G, Özer Ö (2001) Integrating replenishment decisions with advance demand information. Manag Sci 47(10):1344–1360

Hariharan R, Zipkin P (1995) Customer-order information, leadtimes, and inventories. Manag Sci 41:1599–1607

Isotupa KPS (2015) Cost Analysis of an (S-1, S) Inventory system with two demand classes and rationing. Ann Oper Res 233(1):411–421

Karaesmen F, Buzacott J et al (2002) Integrating advance order information in make-to-stock production systems. IIE Trans 34(8):649–662

Koçağa YL, Şen A (2007) Spare parts inventory management with demand lead times and rationing. IIE Trans 39(9):879–898

Kranenburg AA, van Houtum GJ (2007) Cost optimization in the (S- 1, S) lost sales inventory model with multiple demand classes. Oper Res Lett 35(4):493–502

Melchiors P, Dekker R, Kleijn MJ (2000) Inventory rationing in an (s, Q) inventory model with lost sales and two demand classes. J Oper Res Soc 51(1):111–122

Moon I, Kang S (1998) Rationing policies for some inventory policies. J Oper Res Soc 49:509–518

Nahmias S, Demmy WS (1981) Operating characteristics of an inventory system with rationing. Manag Sci 27(11):1236–1245

Özer Ö, Wei W (2004) Inventory control with limited capacity and advance demand information. Oper Res 52(6):988–1000

Skellam JG (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. J R Stat Soc 109(3):296–296

Tan T, Güllü R, Erkip N (2009) Using imperfect advance demand information in ordering and rationing decisions. Int J Prod Econ 121(2):665–677

Tijms HC (1986) Stochastic modelling and analysis: a computational approach. Wiley, Great Britain

Topan E, Tan T, van Houtum GJ, Dekker R (2018) Using imperfect advance demand information in lost-sales inventory systems with the option of returning inventory. IISE Trans 50(3):246–264
Vicil O, Jackson P (2016) Computationally efficient optimization of stock pooling and allocation levels for two-demand-classes under general lead time distributions. IIE Trans 48(10):955–974
Vicil O, Jackson P (2018) Stock optimization for service differentiated demands with fill rate and waiting time requirements. Oper Res Lett 46(3):367–372
Wang Y, Cohen MA, Zheng YS (2002) Differentiating customer service on the basis of delivery lead times. IIE Trans 34:979–989
Wang T, Toktay BL (2008) Inventory management with advance demand information and flexible delivery. Manag Sci 54(4):716–732

Oguzhan Vicil is an operations research and technology management consultant working in the private sector. He received his Ph.D. in Operations Research and Information Engineering from Cornell University, New York in 2006. Since then, he has been managing projects in both the public and private sectors, especially within the scope of applied operations research, information systems, and technological innovation management. His major research interests are in stochastic modeling and optimization in supply chain management, inventory theory and control. He is also active in contributing to the dissemination of scientific literacy among public and bridging the gap between scientific community and policy makers.

Affiliations

Oguzhan Vicil[1]

✉ Oguzhan Vicil
oguzhan.vicil@bilkent.edu.tr; oguzhanvicil@yahoo.com; ov24@cornell.edu

[1] Sharptron, Inc., 13612, Midway Rd, Dallas, TX 75244, USA