Inventory Rationing on a One-for-One Inventory Model with Backorders and Lost Sales

Oguzhan VICIL
oguzhan.vicil[at]bilkent.edu.tr

October 31, 2019

Abstract

In this study, we are primarily motivated by the research problem of recognizing heterogeneous customer behavior towards waiting for order fulfillment under the *threshold rationing policy* (also known as the *critical level policy*), and aim to find its effect on system stock levels and performance measures. We assume a continuous review one-for-one ordering policy with generally distributed lead times. In the first model, we consider the case in which low-priority customer class exhibits zero patience for waiting if the demand is not satisfied immediately (a lost sale), while the demand of high-priority customer class can be backordered. This is the first study in the literature to consider this model. We provide an exact analysis for the derivation of the steady-state probability distribution and the average infinite horizon cost per unit time. We then develop an efficient procedure to minimize the average expected cost rate by deriving bounds on the optimal cost, which reduces the enumeration space considerably. In the numerical study, we show the relative savings achieved by implementing the proposed threshold rationing policy over some of the traditional industry practices. In the second model, we study the opposite case in which the high-priority customer class exhibits zero patience for waiting. The existing study in the literature uses simulation to demonstrate that the performance of the *critical level policy* is robust with respect to the form and variability of the lead time distribution provided the mean lead times are identical. Then it proposes using the Continuous-Time Markov Chain (CTMC) equivalent of the model as an approximation. In this study, we establish a theoretical basis for the rationale of using the CTMC approach as an approximation. We show that under certain assumptions the steady-state probabilities of the system with generally distributed lead times are identical to the steady-state probabilities of the CTMC system with the same mean. This result enables us to link the dynamics of the studied model to the CTMC model, which then allows us to provide a theoretical explanation to the empirically observed phenomenon that why the steady-state probabilities and performance measures are near-insensitive to the form and variability of the lead time distribution as long as mean lead times are identical.

Keywords: Threshold rationing policy; Critical level policy; Continuous-Time Markov Chain; Priority-customer classes; Backorder systems; Lost sales systems; Stochastic Lead Times.

1 Introduction

Inventory rationing, which has been an active subject of research for several decades, serves as a powerful tool for managing inventories in environments where customers are differentiated in terms of their priority. This differentiation is mostly originated from their differences in service level requirements (imposed through contractual obligations) or class-specific penalty costs during stock-out situations. If an organization maintains a separate inventory for each customer class, there can be severe diseconomies of scope. Rationing strategy helps organizations to limit the growth in inventory in support of this differentiation through inventory pooling while providing differentiated services among different customer classes.

However, the customers might not only differ in terms of their importance to the product/service provider or their service level requirements, but also might differ in terms of their advance demand information structures (available/not available, perfect/imperfect), delivery options (exact/flexible), due dates (immediate/after a demand lead time), or their responses when an order is not fulfilled immediately, etc.

In terms of customers' willingness to wait for order fulfillment, a common approach in the literature is to assume homogeneous customer classes: either all customer classes - whose orders are backordered - are assumed to be infinitely patient, or all customer classes have zero patience for waiting and therefore their orders are lost if not fulfilled immediately.

However, there may be environments where customers are not homogeneous in terms of their patience for order fulfillment, and therefore the inventory management system should be adopted to take into account this difference for cost reduction.

In this study, we are primarily motivated by the research problem of recognizing heterogeneous customer behavior towards waiting for order fulfillment under the *threshold rationing policy* (also known as the *critical level policy*), and aim to find its effect on system stock levels and performance measures. We consider an inventory system with two priority-customer classes: *non-critical* and *critical*. Both priority classes exhibit mutually independent, stationary, Poisson demand processes. We assume a continuous review one-for-one ordering policy with generally distributed replenishment lead times. Rather than dedicating completely separate inventories to these two types of customers, the service parts provider opts to use an inventory pool common to both demand classes, and provides differentiated levels of service among them by means of preferential stock allocation policies. In particular, a reserve (threshold) level of inventory is held for use by *critical* customers only, in anticipation of future demands.

We study two models. In the first model (Model-1), we assume that the *critical* (high-priority) customer class has a long term relationship with the service provider. Demands can be backordered since there exists a contract between them. On the other hand, non-contractual relations with the *non-critical* (low-priority) customer class leads the customer to be more ready to find an alternative if the requested order cannot be immediately satisfied, which would result in a lost sale. Due to the

followed threshold rationing policy, both class demands are satisfied on a first-come, first-served basis (FCFS) as long as on-hand stock is greater than the threshold level, and immediately a replenishment order of size 1 is placed through the supplier. All incoming *non-critical* demands are lost (and hence no replenishment order is placed) if on-hand stock is less than or equal to the threshold level at the time of the arrival. The *critical* demands are backordered only in the cases of stock-out situation.

We are the first in the literature to study this model. We aim to find the optimal policy parameters that minimizes the average infinite horizon cost of the given threshold rationing policy. The summary of our contributions for Model-1 is as follows:

1. Our derivations for the steady-state probability distribution and the average infinite horizon cost per unit time are exact.

2. We develop an efficient optimization algorithm by deriving bounds on the optimal cost which reduces the enumeration space considerably.

3. One important aspect of the optimization search routine is that, as the base-stock level is increased by one at each iteration, the steady-state probabilities need to be computed only once. For a given base-stock level, we are able to determine the majority of the expected cost measures for different threshold levels from the knowledge of the steady-state probabilities obtained from the previous iterations.

In the second model (Model-2), we study an inventory system in which the high-priority class exhibit zero patience and therefore any demand that cannot be satisfied immediately from the physical stock is lost. On the other hand, the low-priority class demands can be backordered. Our primary motivation towards this study originated from finding theoretical explanations to the following research questions arose from the empirical results of the Enders et al. (2014) study: (i) why do the steady-state probabilities and performance measures are near-insensitive to the form and variability of the lead time distribution as long as mean lead times are identical? (ii) Why does the CTMC approximation provide quality approximations for the generally distributed lead time model for many system settings? What is the link between the dynamics of the studied model and the CTMC model?

There is no exact solution yet in the literature to either stochastic or deterministic lead times. A closed form analytical solution to determine steady-state probabilities is very difficult, if not impossible, for this model. For the approximation, Enders et al. (2014) first assume exponential lead times and then generalize to degenerate hyperexponential. Using matrix analytic methods, they develop an exact evaluation procedure under the *critical level* policy and prove monotonicity properties of the main performance measures via sample path analysis. Then they develop an optimization procedure that minimizes the average infinite horizon cost by using the monotonicity properties. Enders et al. (2014) state that their approach cannot be generalized to other types

of distributions. However, using simulation they are able to obtain similar results for a variety of distributions (i.e. Weibull, deterministic, Erlang-k, hyperexponential, and lognormal distributions). Using simulation, they also show the near-insensitivity of the performance of the optimal *critical level* policy to lead time distribution variability.

Our study for Model-2 is in fact complementary to the Enders et al. (2014) study. The summary of our contributions is as follows:

1. We establish a theoretical basis for the rationale of using the CTMC approach as an approximation. We show that under certain approximation assumptions the steady-state probabilities of the system with generally distributed lead times are identical to the steady-state probabilities of the CTMC system with the same mean. This result is significant because it enables us to link the dynamics of the studied model to the CTMC model, which then allows us to provide a theoretical explanation to the empirically observed phenomenon that why the steady-state probabilities and performance measures are near-insensitive to the form and variability of the lead time distribution as long as mean lead times are identical. The theoretical rationale contributes to our understanding of the system dynamics, which may further open new doors for future research.

2. Enders et al. (2014) use matrix analytic methods to solve the CTMC model. We propose an alternative approach in which we exploit the special structure of state transitions and then derive a numerical solution method to solve the balance equations.

Another implication of the main result is that the quality of the CTMC approximation is a direct indicator of the quality/accuracy of the approximation assumptions. In other words, the quality of the CTMC approximation is a direct indicator of how weak or strong the approximation assumptions are. Via extensive simulation study, Vicil (2019) evaluates the aggregate effect of the approximation assumptions on the class-specific service levels. He shows that the approximation assumptions hold well for a wide range of system parameters and the performance measures can be estimated with high quality. As long as expected lead times are identical, he also shows empirically the near-insensitivity of the effect of the type of lead time distribution and the lead time variability on the achieved service levels. Vicil (2019) considers cases with deterministic lead times, and Erlang, lognormal and gamma distributed lead times. For the scenarios with $\hat{\beta}_n \geq 60\,\%$ and $\hat{\beta}_c \geq 90\,\%$, levels which we expect to see in practice, and coefficient of variations up to 1.5, the average absolute errors for the CTMC approximations range between 0.05% and 0.09% (for the *low-priority* class fill rate) and range between 0.03% and 0.11% (for the *high-priority* class fill rate) for all the lead time distributions we considered; while the maximum absolute errors for the CTMC approximations are observed as 0.37% (for the *low-priority* fill rate) and 0.47% (for the *high-priority* fill rate).

It is important to note that there are fundamental differences between the backorder models and lost sales models within the one-for-one ordering policy with inventory rationing. Generally

speaking, in conventional inventory models, analysis of the lost sales models are much more complicated than the backorder models. However, as seen in one-for-one policies, this may not be the case for rationing policies. Rationing policy changes the evolution of system states over time such that the sequence of demand arrivals as well as the sequence of receipts from resupply do matter and affect system states. This is a phenomenon which we do not expect to see in conventional backorder models. For the pure backorder systems (when both priority class demands are backordered), exact steady-state analysis seems intractable (except for the special case of exponentially distributed lead times). There is no exact analysis yet in the literature, though several authors provide heuristics (i.e., Dekker et al. (1998), Koçağa and Şen (2007), Vicil and Jackson (2016), Gabor et al. (2018)). Infinite system state and higher dimensionality are other factors which increases the complexity of the analysis, since system states should be represented with additional $n-1$ variables for n-priority demand classes.

For the models that consider pure lost sales (the case in which demands of all priority classes are lost if not satisfied immediately), since there is no backorder clearing process, the queuing results in the literature can be used, and therefore the analysis becomes tractable. In addition, the single-dimensional finite system state representation allows to write steady-state probabilities in a simple closed form which in turn makes it relatively easier for further optimization analysis, including inventory holding and penalty costs (i.e., Dekker et al. (2002), Kranenburg and van Houtum (2007)).

Considering the two models (Model-1 and Model-2) that have been studied in this article, the existence of an exact solution depends on whether the system can be modeled as one of the queuing models for which the exact analysis is possible. Model-1 can be modeled as a $M/G/\infty$ queue with mean service time T and state-dependent arrival rates. Within the scope of this queuing model, for state-dependent arrival rates that allow a representation of the certain form, there is a queuing theory result which allows us to determine the steady-state probabilities exactly. On the other hand, due to the followed threshold rationing policy, Model-2 has similar characteristics as the pure backorder models, and therefore does not fit into one of the conventional queuing models. It has an infinite two-dimensional system state representation and closed form expressions for steady-state probabilities are not readily available. The approximation routine requires recursive numerical solution. It is also interesting to note that even determining the *non-critical class* service levels is challenging. For both pure backorder and pure lost sales models in this setting, the low-priority class service level can be determined exactly. However, the dynamics of the hybrid structure in Model-2 setting does not allow us to achieve an exact solution for the low-priority class service levels. Therefore, computations of both high and low-priority class service levels are based on the heuristic.

This paper is organized as follows. In Section 2, we review the relevant literature. In Section 3, we introduce Model-1 in which the low-priority customers have zero patience for waiting. We provide exact derivation of steady-state probabilities and infinite horizon average cost performance. We also establish several structural results which are crucial for the optimization routine. Then we

provide an efficient optimization algorithm which reduces the enumeration space significantly due to the established bounds on the optimal cost. In Section 4, we introduce Model-2 in which the high-priority customers have zero patience for waiting. We analyze the limiting behavior of state transition probabilities during infinitesimal time intervals under generally distributed lead times. We then present our main theorem showing that under certain approximation assumptions, the steady-state probabilities of the system with generally distributed lead times are identical to the steady-state probabilities of the CTMC system with the same mean. Then we provide a solution strategy to solve the CTMC which exploits the special structure of state transitions. Section 5 is dedicated to comparing the policy performance for the cost optimization model for Model-1 with some of the traditional industry practices.

2 Literature Review

Several authors investigate the form of the optimal policy for capacitated models with multiple demand classes (assemble-to-order end product demands) and a make-to-stock component production environment. The demand classes differ in their demand rates and backorder costs. Ha (1997) shows that a critical level policy (i.e., a policy with a single threshold rationing level for each demand class) is optimal in the case of a Poisson demand process, exponential production times for a single shared component, and lost sales. For the two-demand class backorder case with Erlangian distributed production times, Ha (2000) shows that the optimal policy can be described by a monotone switching curve based on a state variable measuring *work in storage*. For the n-demand class, backorder case, de Véricourt et al. (2000), following the initial work by Ha (1997), prove that a critical level policy is optimal provided production times are exponentially distributed. Pang et al. (2014) study inventory rationing in a lost sales make-to-stock production system with batch ordering and multiple demand classes. For general production time distributions, with the assumption that at any time at most one order outstanding, they show that the optimal order policy is characterized by a reorder point and the optimal rationing policy is characterized by time-dependent rationing levels.

In the continuous review framework, Nahmias and Demmy (1981) consider a (c, s, Q) model with Poisson demand and constant lead time for two-demand-classes. In their model, backorders are allowed and low-priority customers are not served once on-hand inventory drops below critical level c. They assume there is at most one order outstanding in the pipeline. Moon and Kang (1998) extend the analysis to a compound Poisson demand process and Melchiors et al. (2000) study the model within the lost sales framework. Deshpande et al. (2003) study a model similar to Nahmias and Demmy (1981) but allow multiple replenishment orders to be present in the pipeline. They analyze this static policy first under the assumption that backorders are filled according to a special threshold clearing mechanism. They provide approximations for key performance measures and present an efficient algorithm for computing the optimal policy parameters. In a similar model setting, Arslan et al. (2007) analyzed a single location, single product inventory rationing

problem for multiple demand classes that are characterized by different shortage costs or service requirements. They first show how to map their inventory system into a serial inventory system and then use this mapping to develop a model for cost evaluation and optimization.

Dekker et al. (1998) is the first to study the critical level policy within the continuous review $(S-1,S)$ inventory model framework. They consider two priority-demand classes and assume that both class demands can be backordered. They provide approximations for estimating the service level for the critical demand, while the analysis for non-critical demand is exact. Koçağa and Şen (2007) study an environment similar to Dekker et al. (1998), but their model allows either one of the customer class demands to be filled in a given lead time. They provide an approximation for the critical service level while the service level for the non-critical demand is exact. The Vicil and Jackson (2016) study is the only study in the literature that considers general lead time distributions in a similar model as Dekker et al. (1998). They provide an efficient optimization algorithm for finding the minimal stock required to satisfy demand class-specific fill-rate constraints, which requires computation of the steady-state probabilities only once. To determine steady-state probabilities, they analyze the limiting behavior of transition probabilities during an infinitesimal time interval. Under the proposed *independence condition* that the delivery times are unaffected by the level of low-priority demand class backorders, they prove that the CTMC approach is exact for general lead time distributions. They also provide an exact procedure for solving the CTMC model. Then they use the results of the CTMC model to estimate the steady-state probabilities of system states under stochastic lead times. Later Vicil and Jackson (2018) extend the Vicil and Jackson (2016) analysis to include class-specific expected waiting-time requirements along with fill-rate constraints. However, the introduction of the expected waiting-time constraints imposes fundamental challenges and results of Vicil and Jackson (2016) do not directly extend to this model. They need to derive new results regarding the properties of the steady-state probability distribution and service level measures to decrease the computational complexity of the optimization routine. They also characterize the form of the optimal solution in this model setting and propose a simple two step solution strategy that leads to an optimal base-stock and threshold levels. Gabor et al. (2018) consider a similar inventory system as Dekker et al. (1998) but aim to optimize stock levels with respect to class-specific service levels. Their model assumes that the service level of the high-priority customers is measured by the fill rate, while the service level of low-priority customers is measured by a response time guarantee. They use basic lattice path combinatorics to derive an explicit expression of the response time constraint for low-priority customers. They propose a simple approximation for the fill rate calculation for high-priority customers.

The following studies consider pure lost sales environments, meaning that if a demand from any priority demand class cannot be met immediately, then it is considered a lost sale. Dekker et al. (2002) consider a $(c,S-1,S)$ lost sales inventory model for multiple demand classes with Poisson demand processes under general lead time distributions. The single-dimension finite system state representation allowed them to base their analysis on the queueing theory result for state dependent Poisson arrival rates. They derive the exact steady-state distribution of on-hand inventory and

develop techniques to find optimal policy parameters. Kranenburg and van Houtum (2007) study the same environment as Dekker et al. (2002). Their approach relies on solving a series of n-dimensional subproblems, where n is the number of demand classes. To solve the subproblems, they present three efficient heuristic algorithms to find optimal values for the critical levels at a given base stock level, which minimize inventory holding and penalty costs. Song and Zipkin (2009) consider an inventory system with multiple supply sources and Poisson demand. In one section of their study, they consider pure lost sales model within the framework of $(S-1, S)$ policy with rationing. Isotupa (2015) studies a continuous review, lost sales $(S-1, S)$ inventory system for two priority-demand classes with Poisson demand processes. The model is restricted to the independently and identically distributed exponential order lead times. In the study, several conditions are proposed and then under these conditions it is proven that there is a sub-optimal policy where differentiating between customers and using a threshold rationing policy when compared to the case of no customer differentiation.

In different model settings, there are several studies in the literature that consider a mixture of lost sales and backorders. Zhou and Zhao (2010) consider a periodic review inventory system with two priority demand classes. Unsatisfied high priority class demands are lost while low-priority class demands can be backordered. Tang et al. (2008) study a periodic review, capacitated, make-to-stock system with two demand classes. Unsatisfied first class demands can be backordered while second class demands are lost. Wang and Tang (2014) study the dynamic inventory rationing policy with mixed backorders and lost sales. Their model assumes that the penalty cost of backorders varies with time. Therefore, the priorities of demand classes vary with time.

3 Model-1: Low-priority Customers Have Zero Patience for Waiting

In this model, we consider an inventory system with two priority-customer classes: *non-critical* (low-priority) and *critical* (high-priority). The service provider opts to use an inventory pool common to both customer classes, and provides differentiated levels of service among them by means of preferential stock allocation policies. In particular, a reserve level of inventory, denoted by S_c, is held for use by *critical* customers only, in anticipation of future demands.

We assume the demand streams for *critical* and *non-critical* customers are independent Poisson processes with rates λ_c and λ_n, respectively. Service is differentiated using a threshold level, S_c. As long as on-hand inventory is higher than the threshold level, both class demands are satisfied on a first-come, first-served basis. On the other hand, an incoming *non-critical* demand is rejected (hence lost) if on-hand inventory is smaller than or equal to the threshold level S_c. *Critical* demand is backordered only in the case of stock-out situation. Since a continuous review $(S-1, S)$ policy is followed, the acceptance of any customer order (when an incoming demand is not rejected), either by a *critical* or a *non-critical* class, triggers an immediate replenishment order of size 1, which will

be received after a random lead time of L time units. We assume that lead times are independent and identically distributed positively valued random variables, which have no atom at zero. We also assume that the expected lead time, T, is finite.

Furthermore, an incoming replenishment orders are first used to clear *critical* class backorders, if there exists any. Otherwise, it is added to the common inventory pool.

Our main objective is to solve the two demand class cost optimization problem that minimizes the infinite horizon average total cost per unit time, $TC(S, S_c)$, and determine the optimal policy parameters (S, S_c). To do so, we need to determine the steady-state probabilities under general lead time distributions for given policy parameters (S, S_c), and then calculate the expected total cost per unit time $TC(S, S_c)$.

At any time t, let $OH(t)$ denote the number of units on-hand, $R(t)$ denote the number of units in resupply, and $B_c(t)$ denote the number of outstanding *critical* class backorders. At a random point t in time, the following holds:

$$S = OH(t) + R(t) - B_c(t). \tag{1}$$

$$OH(t) = [S - R(t)]^+ \tag{2}$$

$$B_c(t) = [R(t) - S]^+ . \tag{3}$$

Hence, at a random point in time, the system state information required to implement the overall policy can be reduced to the single state variable $R(t)$, the number of units in the resupply system. These relations are also valid for the steady-state distribution of these quantities, denoted by OH, R, and B_c.

3.1 Deriving the Steady-State Probabilities

Let $\mathbb{Z}_0 = \{0, 1, 2, ...\}$ be the set of non-negative integers and $\pi_r(S, S_c), r \in \mathbb{Z}_0$, denote the steady-state distribution of $R(t)$ when the policy parameters are (S, S_c).

Let $\lambda = \lambda_n + \lambda_c$. The following proposition establishes an exact solution for computing the steady-state probabilities under stochastic lead times.

Proposition 1 *For policy parameters (S, S_c) with an independent and identically distributed stochastic lead times with mean T, the steady-state probability of being in state r is given by*

$$\pi_r = \begin{cases} \frac{(\lambda T)^r}{r!}\pi_0, & 0 \le r \le S - S_c, \\ \frac{\left(\lambda^{(S-S_c)}\lambda_c^{r-(S-S_c)}\right)T^r}{r!}\pi_0, & S - S_c + 1 \le r, \end{cases} \tag{4}$$

where

$$\pi_0 = \left[\sum_{r=0}^{S-S_c} \frac{(\lambda T)^r}{r!} + \sum_{r=S-S_c+1}^{\infty} \frac{\left(\lambda^{(S-S_c)} \lambda_c^{r-(S-S_c)}\right) T^r}{r!} \right]^{-1}. \tag{5}$$

Proof: See Appendix A.

■

Corollary 1 *The steady-state probabilities $\pi_r, r \in \mathbb{Z}_0$, are invariant to changes in S provided $\Delta = S - S_c$ is constant.*

Proof: Due to Proposition 1, the steady-state probabilities depend only on Δ, the difference between the target inventory S and the threshold level S_c.

■

3.2 Structural Properties

Lemma 1 *For $S_c < S$, consider two systems: (S, S_c) with steady-state probabilities π_r, and (S, S_c+1) with steady-state probabilities π'_r. Then,*

$$\begin{aligned} \pi'_r &> \pi_r \text{ for } 0 \le r \le S - S_c - 1, \\ \pi'_r &< \pi_r \text{ for } S - S_c \le r. \end{aligned}$$

Proof: See Appendix B.

■

Let $\beta_n(S, S_c)$ (respectively $\beta_c(S, S_c)$) denote the fill rate for the *non-critical* (respectively *critical*) demand class as functions of (S, S_c). Let $\pi_r(S, S_c)$ denote the steady-state probability distribution as a function of (S, S_c). Due to the Poisson Arrivals See Time Averages principle (see, e.g., Tijms (1986)), *non-critical* and *critical* demand class fill rates are as follows:

$$\beta_n(S, S_c) = 1 - P_\infty(OH \le S_c \mid (S, S_c)) = \sum_{r=0}^{S-S_c-1} \pi_r(S, S_c), \tag{6}$$

$$\beta_c(S, S_c) = 1 - P_\infty(OH = 0 \mid (S, S_c)) = \sum_{r=0}^{S-1} \pi_r(S, S_c). \tag{7}$$

Let $\mathrm{B}(S, S_c)$ be the expected number of *critical* class backorders, $\mathrm{I}(S, S_c)$ be the expected on-

hand stock, and $\mathrm{X}(S, S_c)$ be the expected pipeline stock. Then,

$$\mathrm{B}(S, S_c) = \sum_{r=S}^{\infty}(r-S)\,\pi_r\,(S, S_c)\,, \tag{8}$$

$$\mathrm{I}(S, S_c) = \sum_{r=0}^{S}(S-r)\pi_r\,(S, S_c)\,, \tag{9}$$

$$\mathrm{X}(S, S_c) = \sum_{r=0}^{\infty} r\,\pi_r\,(S, S_c)\,. \tag{10}$$

The following proposition establishes important monotonicity results.

Proposition 2 *The performance measures depend on the threshold level $S_c, S_c < S$, in the following manner:*

$$\beta_n(S, S_c+1) < \beta_n(S, S_c) \tag{11}$$

$$\beta_c(S, S_c+1) > \beta_c(S, S_c) \tag{12}$$

$$\mathrm{I}(S, S_c+1) > \mathrm{I}(S, S_c) \tag{13}$$

$$\mathrm{B}(S, S_c+1) < \mathrm{B}(S, S_c) \tag{14}$$

$$\mathrm{X}(S, S_c+1) < \mathrm{X}(S, S_c) \tag{15}$$

Proof: See Appendix C.

■

As long as Δ is fixed, the following proposition allows us to compute several performance measures of the system $(\Delta+k, k), k = 1, 2, ...$, from the knowledge of $\pi_r(\Delta, 0), r \in Z_0$. These relations significantly reduce the computational complexity of the optimization search routine described in the next section.

Proposition 3 *For fixed Δ and $k = 1, 2, ...$, the following relations hold:*

$$\beta_n(\Delta+k, k) = \beta_n(\Delta, 0), \tag{16}$$

$$\beta_c(\Delta+k, k) = \beta_c(\Delta, 0) + \sum_{u=0}^{k-1}\pi_{\Delta+u}(\Delta, 0), \tag{17}$$

$$\mathrm{B}(\Delta+k, k) = \sum_{u=k}^{\infty}(u-k)\,\pi_{\Delta+u}(\Delta, 0), \tag{18}$$

$$\mathrm{I}(\Delta+k, k) = \sum_{u=0}^{\Delta+k}(\Delta+k-u)\,\pi_u(\Delta, 0), \tag{19}$$

$$\mathrm{X}(\Delta+k, k) = \mathrm{X}(\Delta, 0). \tag{20}$$

Proof: See Appendix D.

■

3.3 Cost Optimization

The expected total cost rate (the infinite horizon average total cost per unit time), $C(S, S_c)$, consists of three parts. First, there is a one-time penalty cost, ρ_n for the *non-critical* class and ρ_c for the *critical* class, whenever a demand is not immediately satisfied from on-hand stock. Second, for each *critical* demand that is backordered, there is an associated penalty cost of b per unit per unit time of delay. Third, there is a unit holding costs of h (per unit time) for each unit being held in stock. In steady-state, let $C_P(S, S_c)$ be the expected penalty cost rate, $C_B(S, S_c)$ be the expected backorder cost rate, and $C_H(S, S_c)$ be the expected holding cost rate. Then,

$$\begin{aligned} C_P(S, S_c) &= \rho_n \lambda_n (1 - \beta_n(S, S_c)) + \rho_c \lambda_c (1 - \beta_c(S, S_c)), && (21) \\ C_B(S, S_c) &= b \sum_{i=0}^{\infty} i \, \pi_{S+i}(S, S_c), && (22) \\ C_H(S, S_c) &= h \sum_{i=0}^{S} (S - i) \pi_i(S, S_c). && (23) \end{aligned}$$

Hence, the total expected cost rate is given by

$$\begin{aligned} C(S, S_c) &= C_P(S, S_c) + C_B(S, S_c) + C_H(S, S_c) \\ &= \rho_n \lambda_n (1 - \beta_n(S, S_c)) + \rho_c \lambda_c (1 - \beta_c(S, S_c)) + b \, \mathrm{B}(S, S_c) + h \, \mathrm{I}(S, S_c). \quad (24) \end{aligned}$$

The cost optimization model can be written as:

$$\begin{aligned} \min_{S, S_c} C(S, S_c) &= \min_{S, S_c} \left\{ \rho_n \lambda_n (1 - \beta_n(S, S_c)) + \rho_c \lambda_c (1 - \beta_c(S, S_c)) + b \, \mathrm{B}(S, S_c) + h \, \mathrm{I}(S, S_c) \right\} \\ s.t. \quad & S_c \le S, \\ & S, S_c \in \mathbb{Z}_0. \quad (25) \end{aligned}$$

Proposition 4 *A lower bound function for $C(S, S_c)$ is given by:*

$$C(S, S_c) \;\ge\; C_{LB}(S, S_c) := \rho_n \lambda_n (1 - \beta_n(S, S_c)) + h \; (S - \mathrm{X}(S, S_c))$$

Proof: See Appendix E.

■

Let $\tilde{C}(S)$ be the minimum cost for a given value of S:

$$\tilde{C}(S) := \min_{S_c} \{ C(S, S_c) : \; 0 \le S_c \le S, \; S_c \in \mathbb{Z}_0 \}.$$

Corollary 2 *A lower bound function for $\tilde{C}(S)$ is given by*

$$\tilde{C}(S) \geq \tilde{C}_{LB}(S) := h\,(S - \lambda T)$$

Proof:

$$\begin{aligned} C\,(S, S_c) &\geq C_{LB}\,(S, S_c) := \rho_n \lambda_n (1 - \beta_n(S, S_c)) + h\ (S - \mathrm{X}(S, S_c)) \\ &\geq h\ (S - \mathrm{X}(S, S_c)) \qquad (26) \\ &\geq h\ (S - \mathrm{X}(S, 0)) \qquad (27) \\ &\geq h\ (S - \lambda T)\,. \qquad (28) \end{aligned}$$

(27) follows from (26) due to the monotonicity results in Proposition 2. To prove (28), let us consider a second system, System-2, with policy parameters $(S, 0)$ such that both class demands are satisfied on a first-come, first-served basis, both demand types can be backordered, and there is no customer differentiation. Then clearly, the expected number of units in the resupply for this system will be higher than the expected number of units in the resupply, $\mathrm{X}(S, 0)$, for System-1 (original $(S, 0)$ system with lost sales and backorders). Then System-2 is equivalent to the $(S-1, S)$ system with a single customer type with demand rate $\lambda = \lambda_n + \lambda_c$. Then by Palm's Theorem, the number of units in the resupply for this system is Poisson distributed with mean λT. Hence, the expected number of units in the resupply is λT. Consequently, (28) follows from (27). Since the RHS of (28) holds for all S_c, the proof is completed.

■

Although it was given for a different model (equivalent to Model-2 of our study), we are going to follow a similar optimization search routine as provided by Enders et al. (2014) for determining the optimal policy parameters. For a given S, $C_{LB}\,(S, S_c)$ is increasing in S_c; and $\tilde{C}_{LB}(S)$ is increasing in S. Our search algorithm, which is presented in Table 1, uses these monotonicity properties and established bounds to reduce the enumeration space considerably. Let $\tilde{C}^*(S)$ be the minimum cost up to a certain value of S. Starting from $S = 0$, we increase S, one unit at a time. We continue increasing S until $\tilde{C}^*(S) \leq \tilde{C}_{LB}(S + 1)$. This condition is checked via "Step 6. WHILE loop execution". When the condition $\tilde{C}^*(S) \leq \tilde{C}_{LB}(S+1)$ is satisfied for the first time, there is no need to search for higher values of S to find the optimal policy parameters. Because for all (S', S_c) pairs such that $S' > S$, $\tilde{C}^*(S)$ is guaranteed to be less than or equal to $\tilde{C}(S')$.

The second part of the search algorithm involves searching for $\tilde{C}(S)$, the minimum cost for a given value of S. This is performed via the "Step 6.g. WHILE loop execution". For a given value of S, starting from $S_c = 0$ we increase S_c one unit at a time as long as $\{S_c < S\}$ AND $\{\tilde{C}(S) > C_{LB}(S, S_c+1)\}$. When the condition $\tilde{C}(S) \leq C_{LB}(S, S_c+1)$ is satisfied for the first time, $\tilde{C}(S)$ is guaranteed to be the minimum cost for a given value of S, and hence there is no need to check for higher values of S_c. The first condition of the "Step 6.g. WHILE loop execution", $S_c < S$, is necessary for terminating the execution because, since the algorithm relies on the lower bounds,

it may be possible that the condition $\tilde{C}(S) \leq C_{LB}(S, S_c + 1)$ might never be satisfied during the search. We demonstrate the efficiency of the optimization algorithm in the numerical study, Section 5. Our search algorithm needs to iterate, on average, only 0.926 values for S beyond S^*.

However, our search algorithm differs from the cost optimization algorithm of Enders et al. (2014) in the way cost performance measures are calculated at each iteration. An important aspect of our optimization search routine is that, as the base-stock level is increased by one at each iteration, the steady-state probability need to be computed only once for $(S, 0)$. For a given base-stock level, for all $S_c > 0$, $C(S, S_c)$ can be easily determined from the knowledge of $\pi_r(S - S_c, 0)$ (which has already been computed in the previous iterations) using Proposition 3.

Table 1: Cost optimization algorithm for the two demand class model

Step Inputs: $\lambda_n, \lambda_c, T, p_n, p_c, b, h$
1. Set $S = 0, S_c = 0$;
2. Set $S^* = 0, S_c^* = 0$;
3. Compute $\pi_r(0, 0)$ for all $r \in \mathbb{Z}_0$ using Proposition 1;
4. Initialize $\tilde{C}^*(0) = C(0, 0)$ using (24);
5. Compute $\tilde{C}_{LB}(S + 1)$ using Corollary 2;
6. WHILE $\tilde{C}^*(S) > \tilde{C}_{LB}(S + 1)$:
   a. Set $S = S + 1$;
   b. Set $S_c = 0$;
   c. Compute $\pi_r(S, S_c)$ for all $r \in \mathbb{Z}_0$ using Proposition 1;
   d. Initialize $\tilde{C}(S) = C(S, S_c)$ using (24);
   e. IF $\tilde{C}(S) < \tilde{C}^*(S - 1)$:
      - Set $\tilde{C}^*(S) = \tilde{C}(S)$;
      - Set $S^* = S, S_c^* = S_c$;
      ELSE:
      - Set $\tilde{C}^*(S) = \tilde{C}^*(S - 1)$;
   f. Compute $C_{LB}(S, S_c + 1)$ from the knowledge of $\pi_r(S - S_c - 1, 0)$ using Propositions 3 and 4;
   g. WHILE $\{S_c < S\}$ AND $\{\tilde{C}(S) > C_{LB}(S, S_c + 1)\}$:
      - Set $S_c = S_c + 1$;
      - Compute $C(S, S_c)$ from the knowledge of $\pi_r(S - S_c, 0)$ using Proposition 3;
      - IF $C(S, S_c) < \tilde{C}(S)$:
        * Set $\tilde{C}(S) = C(S, S_c)$;
        * IF $\tilde{C}(S) < \tilde{C}^*(S)$:
          ◇ Set $\tilde{C}^*(S) = \tilde{C}(S)$;
          ◇ Set $S^* = S, S_c^* = S_c$;
      - IF $S_c < S$:
        * Compute $C_{LB}(S, S_c + 1)$ from the knowledge of $\pi_r(S - S_c - 1, 0)$ using Propositions 3 and 4;
   h. Compute $\tilde{C}_{LB}(S + 1)$ using Corollary 2;
7. Return (S^*, S_c^*);
8. Return $\tilde{C}^*(S)$.

4 Model-2: High-priority Customers Have Zero Patience for Waiting

In this model, we consider a similar inventory system as in Model-1, but this time demands of the *non-critical* customer class can be backordered while the *critical* class demands are lost if not met immediately. The *critical* customer class experiences higher service level (fill rate) than the *non-critical* customer class.

Service is differentiated using a threshold level, S_c. As long as on-hand inventory is higher than the threshold level, both class demands are satisfied on a first-come, first-served basis. On the other hand, when on-hand stock is at or below S_c all incoming *non-critical* class demands are backordered. *Critical* class demands are lost only if the on-hand inventory is zero.

The delivery of a replenishment order is first used to replenish the *critical* reserve inventory (up to S_c). Only when on-hand inventory would otherwise exceed the level S_c is a delivery used to satisfy *non-critical* backorders. If on-hand inventory is at or above level S_c and there are no further *non-critical* backorders, then deliveries are added to the common inventory pool and the on-hand inventory level is allowed to exceed S_c.

Because the replenishment policy is a simple order-up-to policy and the rationing policy is static (rather than being dynamic), at a random point in time, the system state information required to implement the overall policy can be reduced to two state variables $\big(R(t), B_n(t)\big)$, the number of units in the resupply system, and the number of *non-critical* backorders, respectively. At a random point t in time, the following holds:

$$OH(t) \quad = \quad S - R(t) + B_n(t). \tag{29}$$

The relation (29) holds for the steady-state distributions as well.

4.1 Determining the Steady-State Probabilities

Let $\mathbb{Z}_0 = \{0, 1, 2, ...\}$ be the set of non-negative integers and $\xi_t = (r, b_n) \in \mathbb{F}_{(S,S_c)}$ 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). And let $\pi_{(r,b_n)}(S, S_c), (r, b_n) \in \mathbb{F}_{(S,S_c)}$, denote the steady-state distribution of $(R(t), B_n(t))$ for given policy parameters (S, S_c).

Determining the steady-state probability distribution is a challenge. This is because, for the case of generally distributed lead times, *"not only total demand arrivals and deliveries are important, but also the sequence of arrivals and deliveries do matter"*. Therefore, a closed form analytical solution to determine steady-state probabilities is very difficult, if not impossible, for this model. To show this, let us consider two scenarios with constant lead times L, and $S = 4$, $S_c = 1$. In the first scenario, let us suppose $OH = 4$, $R = 0$, and $B_n = 0$ at t=0. Suppose also that during the next L time periods there are exactly three *non-critical* demands followed by two *critical* demands.

At $t = L$, the system state will be $OH = 0$, $R = 4$, and $B_n = 0$. On the other hand, if the sequence of arrivals had been reversed (two *critical* demands followed by three *non-critical* demands) the resulting state would have been $OH = 1$, $R = 5$, and $B_n = 2$. In the second scenario, let us suppose $OH = 3$, $R = 1$, and $B_n = 0$ at $t = L$. Suppose also that during the next L time periods there are exactly one delivery, three *non-critical* demands followed by two *critical* demands. At $t = 2L$, the system state will be $OH = 0$, $R = 4$, and $B_n = 0$. On the other hand, if the delivery were to be the last (three *non-critical* demands followed by two *critical* demands and a single delivery), the resulting state would have been $OH = 1$, $R = 4$, and $B_n = 1$.

Exact determination of steady-state probabilities for arbitrary lead time distributions, including constant lead times, is an unsolved problem in the literature. However, the results of the Vicil and Jackson (2016) study triggered our motivation to analyze the limiting behavior of infinitesimal state transition probabilities in this model setting. But before exploring this approach in our model setting, we note that for the special case $(S, 0)$ the following proposition allows us to have an exact solution to steady-state probabilities.

Proposition 5 *For the special case* $(S, 0)$ *with an independent and identically distributed stochastic lead times with finite mean* T*, the steady-state probabilities are*

$$\pi_{(i,j)} = \begin{cases} \frac{(\lambda T)^i}{i!}\pi_{(0,0)}, & 0 \le i \le S, \\ \frac{\left(\lambda^S \lambda_n^j\right)T^i}{i!}\pi_{(0,0)}, & S+1 \le i, \end{cases} \tag{30}$$

where

$$\pi_{(0,0)} = \left[\sum_{v=0}^{S} \frac{(\lambda T)^v}{v!} + \sum_{v=S+1}^{\infty} \frac{\left(\lambda^S \lambda_n^{v-S}\right)T^v}{v!}\right]^{-1}. \tag{31}$$

Proof: See Appendix F. ■

Before analyzing the limiting behavior of state-transitions during infinitesimal time units, first we need to examine the demand arrival and replenishment processes. Although the demand process is Poisson and a one-for-one replenishment policy is followed, order cancelations occur due to the lost sale structure of the process. If there is a physical stock at the time of a *critical* demand arrival, the order is satisfied from the on-hand stock and immediately a replenishment order is given through the supplier. Replenishment orders are placed at a rate of $\lambda_n + \lambda_c$ in these situations. However, if there is a stock-out situation at the time of a *critical* demand arrival, then the order is rejected (a lost sale), and therefore no replenishment order is placed. Thus, replenishment orders are placed at a rate of λ_n in stock-out situations. Furthermore, in steady-state, an arriving customer demand encounters a stock-out situation with probability $1 - \beta_c = P_\infty(OH = 0)$. Hence, in steady-state, the expected number of replenishment orders placed through the supplier over a finite interval, for

example $(t, t+\tau]$, is $(\lambda_n + \beta_c\lambda_c)\tau$. However, the process is not a non-homogeneous Poisson Process because the acceptance of orders are state-dependent and the number of replenishment orders placed in any finite set of nonoverlapping intervals are not necessarily mutually independent random variables. But still we have a valid starting point for the analysis. This is because, interarrival times of replenishment orders due to the *non-critical* demands are exponentially distributed with mean $1/\lambda_n$, while interarrival times of replenishment orders due to the *critical* demands are exponentially distributed with mean $1/\lambda_c$ provided no stock-out occurs. The irregularities occur only during stock-out situations. As achieved service levels for the *critical* demand class increases, we expect to see these irregularities lesser. Since a one-for-one policy is implemented, if we ignore these irregularities, then given that a replenishment order is placed during $(0, t]$, the time of the order placement is expected to be uniformly distributed over this interval. This will serve as our first approximation assumption for the analysis.

On the other hand, there is another dynamic that affects the uniformity of the time of the order placement. Let us explore the system dynamics a little further. The state of the system changes if a replenishment order is placed or a unit is received from the resupply system. Demands are independent and identically distributed so the resupply process is independent of the subsequent demand processes. In addition, replenishment lead times are independent and identically distributed, hence orders may cross. Since demands are Poisson processes and orders from suppliers are triggered whenever a *non-critical* demand occurs or a *critical* demand occurs and is satisfied from on-hand stock, at a random point t in time, if only $R(t)$ is known (or observed), one may claim that the knowledge of the number of units in the resupply would be sufficient to characterize the delivery process via using the previous approximation assumption: *unordered times of the replenishment order placements are independent and uniformly distributed over* $(0, t]$. However, this claim is not correct. Our intuition may play us a trick here at first sight, because the replenishment process is governed by the base-stock policy and orders are placed no matter whether the inventory is rationed or not. Hence, the replenishment process may seem independent of the rationing policy.

Nevertheless, the above reasoning is not correct. If we condition on being at some system state, say $(R(t), B_n(t))$, then the delivery process will also depend on the knowledge of $B_n(t)$. This is because; due to the threshold rationing policy there is a correlation between the pipeline vector and the number *non-critical* class backorders. In other words, there is a difference between conditioning on only the knowledge of $R(t)$ and conditioning on the complete system state information $(R(t), B_n(t))$. To make the analysis tractable, we relax the dependency of these two variables via using another approximation assumption (Note that according to our model the system state information is $(R(t), B_n(t))$. However, there are other ways to represent system states so that $OH(t), B_n(t), B_c(t), R(t)$ information could be extracted. If this were to be the case, then the variant of the approximation assumption would be such that we relax the dependency of the pipeline vector to the current system sate information).

For our analysis, we will use the following approximation assumptions.

Approximation Assumptions:

1. Given that a replenishment order is placed during $(0, t]$, the time of the order is uniformly distributed over this interval.

2. Conditioned on being at the system state $\xi_t = (r, b_n) \in \mathbb{F}_{(S,S_c)}$ at an arbitrary point in time t, the probability of a unit delivery in the interval $(t, t+\tau)$ for an infinitesimally small $\tau, \tau > 0$, does not depend on the value of b_n.

But the question is how weak or strong these assumptions are. Because the quality of the approximation will depend on the aggregate effect of these approximation assumptions on system state transition probabilities. We will explore these dynamics later.

The following lemma establishes an important result which will be used in steady-state analysis.

Lemma 2 *Conditioned on being at the system state $\xi_t = (r, b_n) \in \mathbb{F}_{(S,S_c)}$ at an arbitrary point in time t, under the Approximation Assumptions, the probability of more than one state transition happening in the interval $(t, t+\tau)$ for an infinitesimally small $\tau, \tau > 0$, is $o(\tau)$.*

Proof: See Appendix G. ■

The following theorem establishes the relationship between steady-state probabilities under the *Approximation Assumptions* for generally distributed lead times.

Theorem 1 *For a system with a general, positively-valued lead time distribution having finite mean, T, with no probability mass at zero, then under the Approximation Assumptions the steady-state distribution of (R, B_n) satisfies the same balance equations as the system with an exponential lead time distribution with the same mean.*

Proof: See Appendix H. ■

For a given (S, S_c) pair such that $S > S_c \geq 2$, the steady-state distribution of (R, B_n) satisfies the following balance equations in Table 2. Special cases with $S_c \in \{0, 1\}$ can be written similarly. For notational simplicity, we use $\pi_{(r,b_n)}$ instead of $\pi_{(r,b_n)}(S, S_c)$.

Table 2: Balance equations for the system (S, S_c), such that $S > S_c \geq 2$

State classification	Balance Equation
$r = 0,\ b_n = 0$	$\lambda\pi_{(0,0)} = \mu\pi_{(1,0)}$
$0 < r < S - S_c, b_n = 0$	$(r\mu + \lambda)\,\pi_{(r,0)} = \lambda\pi_{(r-1,0)} + (r+1)\,\mu\pi_{(r+1,0)}$
$r = S - S_c,\ b_n = 0$	$(r\mu + \lambda)\,\pi_{(r,0)} = \lambda\pi_{(r-1,0)} + (r+1)\,\mu\pi_{(r+1,0)} + (r+1)\,\mu\pi_{(r+1,1)}$
$S - S_c < r < S,\ b_n = 0$	$(r\mu + \lambda)\,\pi_{(r,0)} = \lambda_c\pi_{(r-1,0)} + (r+1)\,\mu\pi_{(r+1,0)}$
$r = S,\ b_n = 0$	$(S\mu + \lambda_n)\,\pi_{(S,0)} = \lambda_c\pi_{(S-1,0)}$
$r > S - S_c,\ b_n = r - (S - S_c)$	$(r\mu + \lambda)\,\pi_{(r,b_n)} = \lambda_n\pi_{(r-1,b_n-1)} + (r+1)\,\mu\pi_{(r+1,b_n)} + (r+1)\,\mu\pi_{(r+1,b_n+1)}$
$(S - S_c) + b_n < r < S + b_n,\ b_n > 0$	$(r\mu + \lambda)\,\pi_{(r,b_n)} = \lambda_n\pi_{(r-1,b_n-1)} + \lambda_c\pi_{(r-1,b_n)} + (r+1)\,\mu\pi_{(r+1,b_n)}$
$r > S,\ b_n = r - S$	$(r\mu + \lambda_n)\,\pi_{(r,b_n)} = \lambda_n\pi_{(r-1,b_n-1)} + \lambda_c\pi_{(r-1,b_n)}$

The result of Theorem 1 is very significant for few reasons. First, although the analysis is quite complex we are able to study the steady-state behavior of state transitions with only two approximation assumptions, which have been reasoned using the system dynamics. We theoretically justified that under the two *Approximation Assumptions*, CTMC approach is exact for the generally distributed lead time model. Second, things might have turned out differently: under the *Approximation Assumptions* we might have ended up with some recursive equations for the balance equations that may depend on the form or the variability of the lead time distribution. But instead, we have been able to reach nicely expressed, simple closed form solutions to the balance equations of the steady-state distribution of (R, B_n), which does not depend on the form and variability of the lead time distribution. Therefore, this result allows us to provide a theoretical explanation to the empirically observed phenomenon that why the steady-state performance measures (i.e. class-specific service levels or cost measures) are near-insensitive to the form and variability of lead time distributions as long as mean lead times are identical.

So, using Theorem 1, if we determine the stationary probabilities of the alternate Continuous-Time Markov Chain model, we can use these probabilities to approximate the steady-state probabilities of the original system. And then we can estimate the class-specific service levels. But the question is how well this CTMC approximation is. Therefore, to evaluate the quality of the CTMC approximation for the service level calculations (and hence the quality/accuracy of the *Approximation Assumptions*), we refer to the Vicil (2019) study in which he performs extensive numerical study under a variety of lead time probability distributions and system parameters. He shows that the CTMC approximation works well for many model settings. As long as expected lead times are identical, it is also shown empirically the near-insensitivity of the effect of the type of lead time distribution and the lead time variability on the achieved service levels. Vicil (2019) considers cases with deterministic lead times, and Erlang, lognormal and gamma distributed lead times. For the scenarios with $\hat{\beta}_n \geq 60\,\%$ and $\hat{\beta}_c \geq 90\,\%$, levels which we expect to see in practice, and coefficient of variations (CV) up to 1.5, the summary of the main findings for the CTMC approximations for both customer classes are presented in Table 3 (for more discussion, see Vicil, 2019). Note that CV of the Erlang distribution always varies between 0 and 1.

Table 3: Average and Maximum Absolute Errors of the CTMC approximation ($\hat{\beta}_n \geq 60\,\%$ and $\hat{\beta}_c \geq 90\,\%$).

	Average		Maximum	
Distribution	AE_n	AE_c	AE_n	AE_c
Constant	0.09%	0.05%	0.37%	0.18%
Erlang	0.05%	0.04%	0.20%	0.17%
Lognormal ($CV \leq 1.5$)	0.05%	0.03%	0.28%	0.11%
Gamma ($CV \leq 1.5$)	0.09%	0.11%	0.31%	0.47%

4.2 Exact Solution: Exponentially Distributed Lead Times

In this section, we show that the balance equations for the exponentially distributed lead times can be numerically solved and then the steady-state probabilities can be exactly computed.

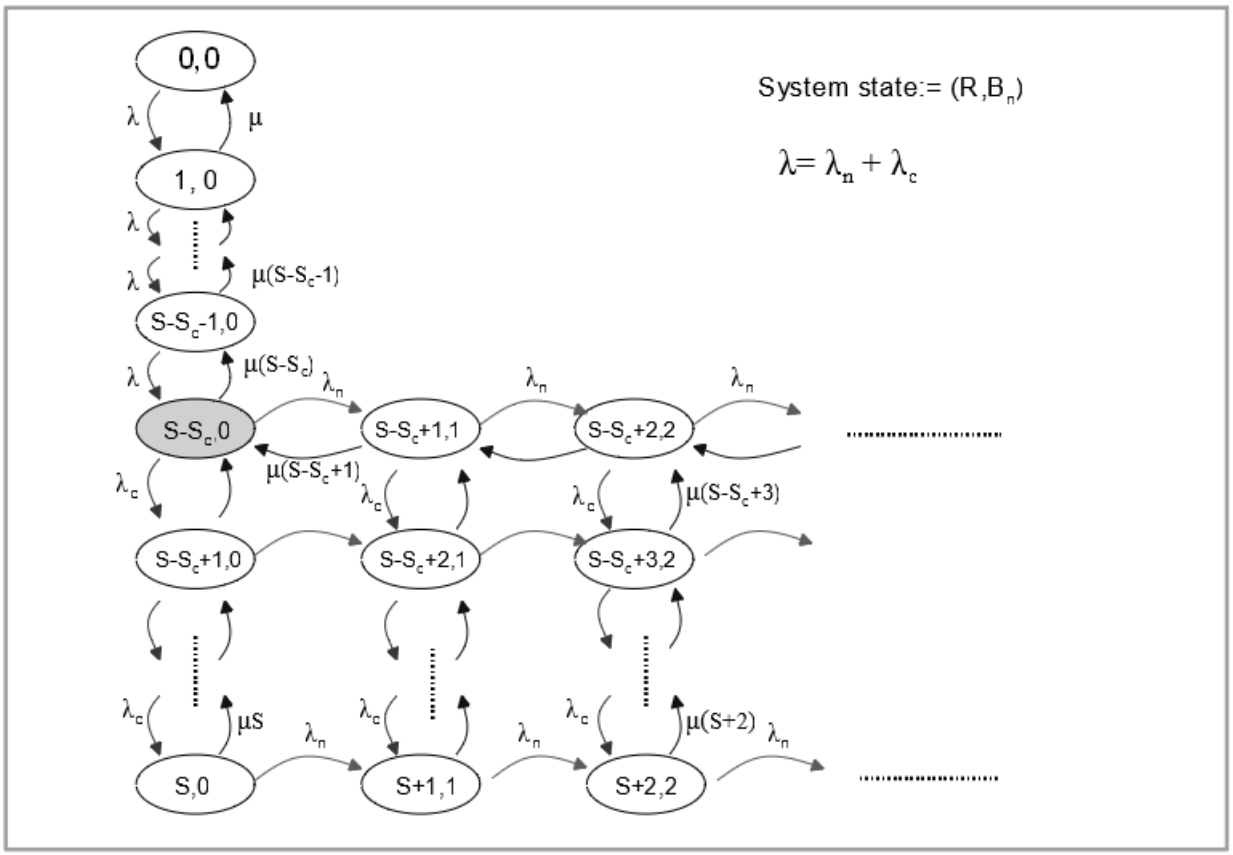

Figure 1: State transition diagram for the CTMC model

Figure 1 displays a state transition diagram for the policy parameters (S, S_c), $S > S_c \geq 2$, with rates for this alternate CTMC system in terms of the state variable pair (R, B_n). Note that all

the transitions between columns corresponding to deliveries are one-way, from left to right, unless $R - B_n = S - S_c$ (which means $OH = S_c$). That is, a delivery cannot cause a transition to the left unless $OH = S_c$. When $R - B_n > S - S_c$ (so that $OH < S_c$), an incoming resupply unit will decrease R by 1 unit which causes an upward transition.

As can be seen in Figure 1, the state transitions have a special structure such that every stationary probability can be written in terms of the single state $(0, 0)$ stationary probability. Therefore, the balance equations can be solved by truncating the infinite state transition matrix. Note that the system regenerates itself every time state $(0, 0)$ is reached, when there are no orders outstanding.

We are going to prove that any steady-state probability $\pi_{(i,j)}$ can be written in terms of the single steady-state probability $\pi_{(0,0)}$. Hence, since the sum of stationary probabilities will be equal to 1, steady-state probabilities can be computed easily. We consider the balance equations for the models with $S > S_c \geq 2$ which are provided in Table 2. Balance equations for models with other policy parameters (S, S_c) can be solved similarly.

STEP 1: For all $\pi_{(i,0)}$ s.t. $1 \leq i \leq S - S_c$;

$$\begin{aligned} \pi_{(1,0)}\mu &= \pi_{(0,0)}\lambda, \\ \pi_{(i,0)}\mu i &= \pi_{(i-1,0)}[\lambda + \mu(i-1)] - \pi_{(i-2,0)}\lambda \text{ for } i \geq 2. \end{aligned}$$

Solving the above two equations leads to the following recursion:

$$\pi_{(i,0)} = \frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i \pi_{(0,0)}. \tag{32}$$

Hence, for $1 \leq i \leq S - S_c$, all $\pi_{(i,0)}$ can be written in terms of $\pi_{(0,0)}$.

STEP 2: For all $\pi_{(i,0)}$ s.t. $S - S_c \leq i \leq S - 1$;

$$\begin{aligned} \pi_{(S,0)}(\lambda_n + \mu S) &= \pi_{(S-1,0)}\lambda_c, \\ \Longrightarrow \quad \pi_{(S-1,0)} &= \pi_{(S,0)}\left(\frac{\lambda_n + \mu S}{\lambda_c}\right). \end{aligned} \tag{33}$$

From the balance equations, we first write $\pi_{(S-2,0)}$ in terms of $\pi_{(S-1,0)}$ and $\pi_{(S,0)}$ as,

$$\pi_{(S-2,0)} = \left\{\pi_{(S-1,0)}[\lambda + \mu(S-1)] - \pi_{(S,0)}\mu S\right\}\frac{1}{\lambda_c}. \tag{34}$$

Note that since $\pi_{(S-1,0)}$ has already been determined in terms of $\pi_{(S,0)}$, $\pi_{(S-2,0)}$ can be written in terms of only $\pi_{(S,0)}$. And then recursively, all the remaining stationary probabilities can be

written in terms of only $\pi_{(S,0)}$:

$$\pi_{(S-k,0)} = \left\{\pi_{(S-k+1,0)}[\lambda + \mu(S-k+1)] - \pi_{(S-k+2,0)}\mu(S-k+2)\right\}\frac{1}{\lambda_c}, \text{ for } k = 3,4,...,S_c, \quad (35)$$

with the last one being,

$$\pi_{(S-S_c,0)} = \left\{\pi_{(S-S_c+1,0)}[\lambda + \mu(S-S_c+1)] - \pi_{(S-S_c+2,0)}\mu(S-S_c+2)\right\}\frac{1}{\lambda_c}. \quad (36)$$

But from the Equation (32), we also have

$$\pi_{(S-S_c,0)} = \frac{1}{(S-S_c)!}\left(\frac{\lambda}{\mu}\right)^{S-S_c}\pi_{(0,0)}. \quad (37)$$

Hence, numerical solution to (36) and (37) exist and therefore, $\pi_{(S,0)}$ can be written in terms of only $\pi_{(0,0)}$; which also implies that for $S - S_c \leq i \leq S-1$, all $\pi_{(i,0)}$ can be written in terms of only $\pi_{(0,0)}$.

STEP 3: For all $\pi_{(i,j)}$ s.t. $S - S_c + j \leq i \leq S + j$, $1 \leq j$;
In this step, we will show that all $\pi_{(i,j)}$ can be written recursively in terms of only $\pi_{(0,0)}$ for $j = 1, 2, ...$

For $j = 1$, the balance equation for state $(S - S_c + j - 1, j - 1)$ is,

$$\pi_{(S-S_c,0)}\left[\mu(S-S_c) + \lambda\right] = \pi_{(S-S_c-1,0)}\lambda + \pi_{(S-S_c+1,0)}\mu\left(S-S_c+1\right) + \pi_{(S-S_c+1,1)}\mu\left(S-S_c+1\right). \quad (38)$$

For $j \geq 2$, the balance equation for state $(S - S_c + j - 1, j - 1)$ is,

$$\begin{aligned}\pi_{(S-S_c+j-1,j-1)}\left[\mu(S-S_c+j-1) + \lambda\right] &= \pi_{(S-S_c+j-2,j-2)}\lambda_n + \pi_{(S-S_c+j,j-1)}\mu\left(S-S_c+j\right) \\ &\quad + \pi_{(S-S_c+j,j)}\mu\left(S-S_c+j\right). \quad (39)\end{aligned}$$

In the above equations, $\pi_{(S-S_c+j,j)}$ for $j \geq 1$ can be written exclusively in terms of $\pi_{(0,0)}$, because the steady-state probabilities $\pi_{(i,k)}$, $k \leq j-1$, were already written in terms of $\pi_{(0,0)}$ in previous steps and recursions.

And then, from the balance equations we first write $\pi_{(S+j-1,j)}$ in terms of $\pi_{(S+j,j)}$ and $\pi_{(S+j-1,j-1)}$ as,

$$\pi_{(S+j-1,j)} = \left\{\pi_{(S+j,j)}[\lambda_n + \mu(S+j)] - \pi_{(S+j-1,j-1)}\lambda_n\right\}\frac{1}{\lambda_c}, \quad (40)$$

where $\pi_{(S+j-1,j-1)}$ has already been written in terms of $\pi_{(0,0)}$ in the previous steps and recursions.

And then recursively, for $k = 1, ..., S_c - 1$, all the remaining stationary probabilities can be

written in terms of $\pi_{(S+j,j)}, \pi_{(S+j-1,j)}$ and $\pi_{(S+j-k,j-1)}$ as:

$$\pi_{(S+j-1-k,j)} = \Big\{\pi_{(S+j-k,j)}[\lambda+\mu(S+j-k)] - \pi_{(S+j+1-k,j)}\mu(S+j+1-k) - \pi_{(S+j-1-k,j-1)}\lambda_n\Big\}\frac{1}{\lambda_c}, \quad (41)$$

with the last one being,

$$\pi_{(S-S_c+j,j)} = \Big\{\pi_{(S-S_c+1+j,j)}[\lambda+\mu(S-S_c+1+j)] - \pi_{(S-S_c+2+j,j)}\mu(S-S_c+2+j) - \pi_{(S-S_c+j,j-1)}\lambda_n\Big\}\frac{1}{\lambda_c}. \quad (42)$$

$\pi_{(S-S_c+j,j)}$ had already been written in terms of only $\pi_{(0,0)}$ in (38) and (39). It has also been written in terms of $\pi_{(S+j,j)}$ and $\pi_{(0,0)}$ in (40) and (41). Hence numerical solution to those equations exist and, therefore $\pi_{(S+j,j)}$ can be solved and written in terms of only $\pi_{(0,0)}$; which also implies that for $S - S_c + j \leq i \leq S + j$, $j \geq 2$, all $\pi_{(i,j)}$ can be written in terms of only $\pi_{(0,0)}$.

STEP 4: Write all $\pi_{(i,j)}$ in terms of the single unknown $\pi_{(0,0)}$ and then solve the following equation

$$\sum_i \sum_j \pi_{(i,j)} = 1. \quad (43)$$

Solving the above equation, we can first determine $\pi_{(0,0)}$ and then all $\pi_{(i,j)}$ can be determined successively.

■

Note that for coding purpose, the infinite sum in (43) can be solved by truncating the infinite state space and solving for

$$\sum_i \sum_{j \leq K} \pi_{(i,j)} = 1, \quad (44)$$

where K denotes an upper limit on the number of *non-critical* customer class backorders to compute.

Let us denote the cumulative distribution function of Poisson distribution as $F(\cdot)$. To truncate the infinite state space according to a desired precision, one possible upper bound on the number of *non-critical* customer class backorders, K, can be found as

$$K = F^{-1}(\alpha, \frac{\lambda}{\mu}), \quad (45)$$

where $\frac{\lambda}{\mu}$ is the mean of Poisson distribution and α is the captured cumulative probability of the number of *non-critical* class backorders, representing the precision. For the purpose of this study, we set $\alpha = 99.999\%$. However, depending on the desired precision level, α can be chosen appropriately. To show this, for given arbitrary policy parameters (S, S_c) and demand stream, let us consider two systems. The first system represents Model-2 of this study. The second system represents a very similar inventory system except that backorders are allowed for both customer classes. Let system

states be $(R(t), B_n(t))$ and $(R'(t), B'_n(t))$ for the first and second systems, respectively. At any point in time, $B_n(t) \leq R(t) - (S - S_c)$. In addition, $R(t) \leq R'(t)$. This is because, although replenishment process is governed by the base-stock policy for both systems, some of the *critical* demands are lost in the first system during stock out situations, and therefore do not trigger replenishment orders. These relations are also valid in the steady-state. Hence we have $B_n \leq R - (S - S_c) \leq R \leq R'$. Furthermore, Vicil and Jackson (2016) showed that in steady-state R' is Poisson distributed with mean $\frac{\lambda}{\mu}$. Hence, if K is set according to $K = F^{-1}(\alpha, \frac{\lambda}{\mu})$, then

$$P(B_n \leq K) \geq P(R \leq K) \geq P(R' \leq K) \geq \alpha. \tag{46}$$

Note that $P(R' \leq K) \geq \alpha$, instead of $P(R' \leq K) = \alpha$, because of the discreteness of Poisson distribution. If we denote the maximum value of B_n as $B_n^{\max}$, this result implies that $P(B_n^{\max} \leq K) \geq \alpha$. Hence, for Model-2, $K = F^{-1}(\alpha, \frac{\lambda}{\mu})$ is a legitimate upper bound on the number of *non-critical* class backorders with precision level α. Furthermore, since the infinite sum is truncated to the finite sum in (44), the error associated with the truncation is bounded by $\sum_i \sum_{j>K} \pi_{(i,j)} \leq 1 - \alpha$.

In our setting, $P(B_n^{\max} \leq K) \geq 99.999\,\%$, and hence the error associated with the truncation is bounded by $0.001\,\%$.

Let $\hat{\pi}_{(i,j)}$ be the estimator of the steady-state probabilities of the original system under the generally distributed lead times. Then, the CTMC approximation algorithm can be implemented as in Table 4.

Table 4: The CTMC Approximation Algorithm for estimating the steady-state probabilities of the two demand class hybrid model

Step	Inputs: $S, S_c, T, \lambda_n, \lambda_c$
1.	Initialize $\mu = 1/T,\ \lambda = \lambda_n + \lambda_c,\ K = F^{-1}(99.999\%, \frac{\lambda}{\mu})$;
2.	Write every state $\pi_{(i,j)}, i \geq 1$, in terms of $\pi_{(0,0)}$ using STEP 1 through STEP 4;
3.	Compute $\pi_{(0,0)}$ from the equation $\sum_i \sum_{j \leq K} \pi_{(i,j)} = 1$;
4.	Determine all $\pi_{(i,j)}, i \geq 1$, from Step 2 (of Table 4);
5.	Set $\hat{\pi}_{(i,j)} = \pi_{(i,j)}$, for all i, j.

5 Comparing Policy Performance of Model-1 for the Cost Optimization Model

In this section, we compare the performance of the *threshold rationing policy* with the *pooling without rationing policy* and the *separate-stock policy*, which are common in practice (for more discussion see e.g., Deshpande (2003)). In the *pooling without rationing policy*, the advantage of inventory pooling is taken but there is no differentiation among customer classes and, therefore demands from both classes are satisfied on a first-come, first-served basis.

To illustrate the benefit of implementing the threshold rationing policy among the heterogeneous customer classes with different impatience for waiting, we study 27 instances with varying system parameters. For all the instances, we fix $h = 1, b = 10, p_n = 10, T = 1$, and vary λ_n, λ_c and p_c. The results are provided in Table 5. In the first two columns, the demand rates are provided for the *non-critical* and *critical* customer classes, respectively. In the third column, one-time penalty cost p_c are provided for the *critical* customer demands that are not immediately satisfied from on-hand stock. In columns $4-5$, we provide the percentage savings of using the *threshold rationing policy* over the *separate-stock policy* and the *pooling without rationing policy*, respectively.

We first observe that *threshold rationing policy* provides significant savings with respect to the *separate-stock policy* in all the instances. The percentage savings vary between 16.01% and 30.89%. We also observe that for $\lambda_n = \lambda_c$, as λ increases, the percentage savings increases. Furthermore, we see that for a fixed λ, the higher the ratio λ_n/λ_c, the higher the percentage savings. However, besides these relationships, there is no other regular pattern showing the relation between the percentage savings and variables λ_n, λ_c and p_c.

Although the savings are not as significant as in the *separate-stock policy* comparison, the *threshold rationing policy* still provides considerable savings over the *pooling without rationing policy*. The threshold rationing policy provides savings in 27 of the 29 instances, which can be as high as 11.33%. We observe that while keeping λ_n and λ_c fixed, as p_c increases the percentage saving increases. We observe a similar pattern as in the *separate-stock policy*. For $\lambda_n = \lambda_c$, as λ increases the percentage savings increases. We also observe that for a fixed λ, the higher the ratio λ_n/λ_c, the higher the percentage savings.

We note that the (total) savings associated with implementing the *threshold rationing* policy may be more pronounced in environments where hundreds to tens of thousands of stock units are being managed.

Furthermore, we can conclude that our optimization algorithm is quite efficient. Due to using the bounds in our optimization algorithm, we are able to eliminate considerable parts of the enumeration space. Considering all 27 instances in Table 5, our search algorithm needs to iterate, on average, only 0.926 values for S beyond S^* (The maximum and the minimum number of required iterations for S beyond S^* are observed as 2 and 0). This also implies that, on average, the steady-state probabilities need to be computed only $S^* + 1.926$ times for all possible enumeration of (S, S_c) pairs due to Proposition 3 (Recall that as the base-stock level is increased by one at each iteration, the steady-state probabilities need to be computed only once in our proposed optimization routine. For a given base-stock level S, $C(S, S_c)$ can be immediately determined for all $S_c > 0$ from the knowledge of the steady-state probabilities already computed in previous iterations using Proposition 3. Also note that when the optimization routine terminates with base-stock level, say S, the steady-state probabilities are calculated in total $S + 1$ times for the policy parameters $(0, 0), (1, 0), \ldots, (S, 0)$).

Table 5: Benefit of Threshold Rationing Policy vs. Pooling without Rationing and Separate-Stock Policies

λ_n	λ_c	p_c	% Saving (sep-stock)	% Saving (pooling w/o rat.)
1	1	50	27.54 %	0.74 %
1	1	100	26.40 %	4.49 %
1	1	200	25.27 %	4.86 %
1	5	50	20.77 %	1.07 %
1	5	100	20.09 %	3.16 %
1	5	200	19.73 %	3.48 %
1	10	50	17.11 %	0.94 %
1	10	100	16.50 %	1.73 %
1	10	200	16.01 %	2.61 %
5	1	50	27.04 %	0 %
5	1	100	26.96 %	3.69 %
5	1	200	27.52 %	7.64 %
5	5	50	29.49 %	3.53 %
5	5	100	29.46 %	6.68 %
5	5	200	29.08 %	8.76 %
5	10	50	28.11 %	4.52 %
5	10	100	27.52 %	6.14 %
5	10	200	26.80 %	7.45 %
10	1	50	24.07 %	0 %
10	1	100	23.46 %	0.84 %
10	1	200	24.97 %	6.05 %
10	5	50	30.61 %	4.66 %
10	5	100	30.13 %	7.39 %
10	5	200	30.79 %	11.33 %
10	10	50	30.78 %	5.41 %
10	10	100	30.67 %	8.12 %
10	10	200	30.89 %	11.26 %

6 Conclusion and Future Research

Inventory rationing in different settings have been studied in the past but for most, it remains a challenging problem because of the difficulty of computing exact or accurate performance measures. In this study, we explore a specific threshold rationing policy for a single service part with two priority-demand classes in concert with a continuous review $(S-1, S)$ replenishment policy. Demand classes are not only differentiated in terms of their importance to the service parts provider, but also differentiated in terms of their willingness to wait for backordered demand.

This is the first study in the literature to consider Model-1. We provide an exact analysis for the

derivation of the steady-state probability distribution and the average infinite horizon cost per unit time. We then develop a computationally efficient optimization algorithm to minimize the average expected cost rate by deriving bounds on the optimal cost, which reduces the enumeration space considerably. To make the optimization algorithm coding friendly for practitioners, we present every step in detail. Then in the numerical study section, we demonstrate that inventory rationing may provide considerable savings over some of the traditional approaches in practice.

In Model-2, we aim to analyze the dynamics behind the empirically observed phenomenon that why the CTMC approximation provides quality approximations for the generally distributed lead time model in many system settings. By analyzing the limiting behavior of state transition probabilities during infinitesimal time intervals, we establish a theoretical basis for the rationale of using the CTMC approach as an approximation. We show that under certain approximation assumptions, the steady-state probabilities of the system with generally distributed lead times are identical to the steady-state probabilities of the CTMC system with the same mean. This result allows us to provide a theoretical explanation to the empirically observed phenomenon that why the steady-state probabilities and performance measures are near-insensitive to the form and variability of the lead time distribution as long as mean lead times are identical. The theoretical rationale contributes to our understanding of the system dynamics, which may further open new doors for future research.

One direct extension of Model-1 is to consider a service level optimization model. As a suggestion for future research for both Model-1 and Model-2, it may be interesting to consider an inventory system such that rather than showing a zero patience towards waiting for order fulfillment, the corresponding demand class is willing to wait but only up to a certain point. The orders are lost only if the waiting time exceeds a certain patience time.

References

[1] Arslan, H., Graves, S.C., and Roemar, T. (2007) A single-product inventory model for multiple demand classes. *Management Science*, 53(9): 1486-1500

[2] Dekker, R., Kleijn, M. J., and de Rooij, P. J. (1998) A spare parts stocking policy based on equipment criticality. *International Journal of Production Economics*, 56-57: 69-77

[3] Dekker, R., Hill, R. M., Kleijn, M. J., and Teunter, R. H. (2002) On the (S-1, S) lost sales inventory model with priority demand classes. *Naval Research Logistics*, 49(6): 593-610.

[4] Deshpande, V., Cohen, M. A., and Donohue, K. (2003) A threshold rationing policy for service differentiated demand classes. *Management Science*, 49(6): 683-703.

[5] Enders, P., Adan, I., Scheller-Wolf, A. et al. (2014) Inventory rationing for a system with heterogeneous customer classes. *Flex Serv Manuf J* 26:344-386.

[6] Gabor, A.F., van Vianen, L., Yang, G. et al. (2018) A base-stock inventory model with service differentiation and response time guarantees. *European Journal of Operational Research*, 269(3): 900-908.

[7] Ha, A.Y. (1997) Inventory rationing in a make-to-stock environment with several demand classes and lost sales. *Management Science* 43(8): 1093-1103.

[8] Ha, A.Y. (2000) Stock rationing in an $M/E_k/1$ make-to-stock queue.*Management Science* 46(1): 77-87.

[9] Isotupa, K.P.S. (2015) Cost Analysis of an (S-1,S) Inventory system with two demand classes and rationing. *Annals of Operations Research*, 233(1): 411-421.

[10] Koçağa, Y. L. and Şen, A. (2007) Spare parts inventory management with demand lead times and rationing. *IIE Transactions*, 39(9): 879-898.

[11] Kranenburg, A. A. and van Houtum, G. J. (2007) Cost optimization in the (S- 1,S) lost sales inventory model with multiple demand classes. *Operations Research Letters*, 35(4), 493-502.

[12] Melchiors, P., R. Dekker, M. J. Kleijn. (2000) Inventory rationing in an (s,Q) inventory model with lost sales and two demand classes. *Journal of the Operational Research Society*, 51(1): 111-122.

[13] Moon, I., S. Kang. 1998. Rationing policies for some inventory policies. *Journal of the Operational Research Society* 49: 509-518.

[14] Nahmias, S. and Demmy, W. S. (1981) Operating characteristics of an inventory system with rationing. *Management Science*, 27(11): 1236-1245.

[15] Pang, Z., Shen, H., and Chengm, T.C.E. (2014) Inventory rationing in a make-to-stock system with batch production and lost sales. *Production and Operations Management*, 23(7): 1243-1257.

[16] Smeitink, E. (1990) A Note on "Operating Characteristics of the S-1, S Inventory System with Partial Backorders and Constant Resupply Times". *Management Science*, 36(11): 1413-1414.

[17] Song, J-S. and Zipkin P. (2009) "Inventories with Multiple Supply Sources and Networks of Queues with Overflow Bypasses". *Management Science*, 55(3): 362-372.

[18] Tang, Y., Xu, D., and Zhou, W. (2008) Inventory Rationing in a Capacitated System with Backorders and Lost Sales. *IEEM 2007: 2007 IEEE International Conference on Industrial Engineering and Engineering Management*, 1579 - 1583.

[19] Tijms, H.C. (1986) *Stochastic Modelling and Analysis: A Computational Approach*, John Wiley & Sons, Chichester, UK.

[20] Véricourt, F.D.,Karaesmen,F., Dallery, Y. (2000) Dynamic scheduling in a make-to-stock system: A partial characterization of optimal policies. *Operations Research*, 48(5): 811-819.

[21] Vicil, O. and Jackson, P. (2016) Computationally efficient optimization of stock pooling and allocation levels for two-demand-classes under general lead time distributions. *IIE Transactions*, 48(10): 955-974.

[22] Vicil, O. and Jackson, P. (2018) Stock optimization for service differentiated demands with fill rate and waiting time requirements. *Operations Research Letters*, 46(3): 367-372.

[23] Vicil, O. (2019) Numerical Validation of the Continuous Time Markov Chain Approximation as a Service Level Estimation Method for Heterogeneous Customer Demand Classes. *Working Paper, available at SSRN (October 31, 2019):* https://ssrn.com/abstract=

[24] Wang, D. and Tang, O. (2014) Dynamic inventory rationing with mixed backorders and lost sales. *International Journal of Production Economics*, 149: 56-67.

[25] Zhou, Y. and Zhao, X. (2010) A two-demand-class inventory system with lost-sales and back-orders. *Operations Research Letters*, 38(4): 261-266.

A Proof of Proposition 1

At any point in time, the knowledge of the number of units in the resupply system is sufficient to characterize the system state. Therefore, the system can be modeled as a $M/G/\infty$ queue with mean service time T and state-dependent arrival rates

$$\lambda_r = \begin{cases} \lambda, & 0 \leq r \leq S - S_c - 1, \\ \lambda_c, & S - S_c \leq r. \end{cases} \tag{47}$$

Smeitink (1990) studied a $(S-1, S)$ inventory system with partial backordering such that there is no demand differentiation. In our rationing model with two priority-demand classes, if we define $\alpha_{r-(S-S_c)} = \frac{\lambda_c}{\lambda}$ for $r \geq S - S_c$, and $\Psi(r)$ as

$$\Psi(r) = \begin{cases} 1, & 0 \leq r \leq S - S_c, \\ \Pi_{j=0}^{r-(S-S_c)-1} \alpha_j, & r \geq S - S_c + 1, \end{cases}$$

and then use the steady-state probabilities derived by Smeitink (1990) where the maximum stock level S is replaced by $S - S_c$, we obtain the steady-state probabilities presented in Proposition 1.

■

B Proof of Lemma 1

Considering the state-dependent arrival rates in (47), the steady-state probabilities in Proposition 1 are also equivalent to

$$\pi_r = \left\{\prod_{i=0}^{r-1}\lambda_i\right\}\frac{T^r}{r!}\pi_0, \quad r \geq 1, \tag{48}$$

where

$$\pi_0 = \left[\sum_{r=0}^{\infty}\left\{\prod_{i=0}^{r-1}\lambda_i\right\}\frac{T^r}{r!}\right]^{-1}. \tag{49}$$

Let

$$\begin{aligned}
A_r &= \left\{\prod_{i=0}^{r-1}\lambda_i\right\}\frac{T^r}{r!}, \quad 0 \leq r \leq S - S_c - 1,\\
A'_r &= \left\{\prod_{\substack{i=0\\ i\neq S-S_c-1}}^{r-1}\lambda_i\right\}\frac{T^r}{r!}, \quad S - S_c \leq r,\\
B &= \sum_{j=0}^{S-Sc-1}\left\{\prod_{i=0}^{j-1}\lambda_i\right\}\frac{T^j}{j!},\\
C &= \sum_{j=S-Sc}^{\infty}\left\{\prod_{\substack{i=0\\ i\neq S-S_c-1}}^{j-1}\lambda_i\right\}\frac{T^j}{j!}.
\end{aligned}$$

Results are immediately established by using the approach provided in the proof of *Lemma* 1, Dekker et al. (2002), which is given for the pure lost sales model with finitely valued system state information r.

■

C Proof of Proposition 2

(i) $\beta_n(S, S_c + 1) < \beta_n(S, S_c)$:

Due to Corollary 1, $\beta_n(S, S_c) = \beta_n(S - S_c, 0)$ and $\beta_n(S, S_c + 1) = \beta_n(S - S_c - 1, 0)$. If we consider the policy parameters $(S - S_c - 1, 0)$, there is no reserve stock for the *critical* class demands and demands of both classes are satisfied on a first-come, first-served basis as long as there is physical stock. Hence, there is no difference in terms the experienced service levels and therefore $\beta_n(S - S_c - 1, 0) = \beta_c(S - S_c - 1, 0)$. If we consider the policy parameters $(S - S_c, 0)$, there is more

inventory for common use, and therefore considering the infinite horizon average performance there will be less stock-out situations for both demand classes than the case would be for $(S - S_c - 1, 0)$. Consequently, $\beta_n(S - S_c, 0) = \beta_c(S - S_c, 0) > \beta_n(S - S_c - 1, 0) = \beta_c(S - S_c - 1, 0)$. As a result, $\beta_n(S, S_c) > \beta_n(S, S_c + 1)$.

(ii) $\beta_c(S, S_c + 1) > \beta_c(S, S_c)$:

$$\begin{aligned}
\beta_c(S, S_c + 1) - \beta_c(S, S_c) &= \left(1 - \sum_{r=S}^{\infty} \pi_r(S, S_c + 1)\right) - \left(1 - \sum_{r=S}^{\infty} \pi_r(S, S_c)\right) \\
&= \sum_{r=S}^{\infty} (\pi_r(S, S_c) - \pi_r(S, S_c + 1)) \\
&> 0 \quad \text{(due to Lemma 1).}
\end{aligned}$$

(iii) $\mathrm{I}(S, S_c + 1) > \mathrm{I}(S, S_c)$:

$$\begin{aligned}
\mathrm{I}(S, S_c + 1) - \mathrm{I}(S, S_c) &= \sum_{r=0}^{S} (S - r)\ (\pi_r(S, S_c + 1) - \pi_r(S, S_c)) \\
&= \sum_{r=0}^{S-S_c-1} (S - r)\,(\pi_r(S, S_c + 1) - \pi_r(S, S_c)) \\
&\quad - \sum_{r=S-S_c}^{S} (S - r)\,(\pi_r(S, S_c) - \pi_r(S, S_c + 1))\,. \qquad (50)
\end{aligned}$$

Due to Lemma 1, the terms of both summation in (50) are positive. Let

$$\begin{aligned}
\delta_1 &= \sum_{r=0}^{S-S_c-1} \pi_r(S, S_c + 1) - \pi_r(S, S_c), \\
\delta_2 &= \sum_{r=S-S_c}^{S} \pi_r(S, S_c) - \pi_r(S, S_c + 1), \\
\delta_3 &= \sum_{r=S+1}^{\infty} \pi_r(S, S_c) - \pi_r(S, S_c + 1).
\end{aligned}$$

Due to Lemma 1, $\delta_1, \delta_2, \delta_3 > 0$. In addition, since the sum of steady-state probabilities is 1 for both systems, we should have $\delta_1 = \delta_2 + \delta_3$.

Furthermore,

$$\begin{aligned}
\sum_{r=0}^{S-S_c-1} (S - r)\,(\pi_r(S, S_c + 1) - \pi_r(S, S_c)) &\geq \sum_{r=0}^{S-S_c-1} (S_c + 1)\,(\pi_r(S, S_c + 1) - \pi_r(S, S_c)) \\
&= (S_c + 1) \sum_{r=0}^{S-S_c-1} (\pi_r(S, S_c + 1) - \pi_r(S, S_c)) \\
&= (S_c + 1)\,\delta_1.
\end{aligned}$$

For the second term in the RHS of (50),

$$\begin{aligned}\sum_{r=S-S_c}^{S}(S-r)\left(\pi_r(S,S_c)-\pi_r(S,S_c+1)\right) &\leq \sum_{r=S-S_c}^{S} S_c\ \left(\pi_r(S,S_c)-\pi_r(S,S_c+1)\right)\\ &= S_c\sum_{r=S-S_c}^{S}\pi_r(S,S_c)-\pi_r(S,S_c+1)\\ &= S_c\,\delta_2.\end{aligned}$$

Since $\delta_1 > \delta_2$, $(S_c+1)\,\delta_1 > S_c\,\delta_2$. Therefore, the RHS of (50) is positive. Consequently,

$$\mathrm{I}(S,S_c+1)-\mathrm{I}(S,S_c) \;>\; 0.$$

(iv) $\mathrm{B}(S,S_c+1) < \mathrm{B}(S,S_c)$:

$$\begin{aligned}\mathrm{B}(S,S_c)-\mathrm{B}(S,S_c+1) &= \sum_{r=S}^{\infty}(r-S)\left(\pi_r(S,S_c)-\pi_r(S,S_c+1)\right)\\ &> 0 \quad \text{(due to Lemma 1).}\end{aligned}$$

(v) $\mathrm{X}(S,S_c+1) < \mathrm{X}(S,S_c)$:

$$\begin{aligned}\mathrm{X}(S,S_c)-\mathrm{X}(S,S_c+1) &= \mathrm{I}(S,S_c+1)-\mathrm{I}(S,S_c)+\mathrm{B}(S,S_c)-\mathrm{B}(S,S_c+1) \quad \text{(using (1))}\\ &> 0 \quad \text{(With reference to (13) and (14)).}\end{aligned}$$

D Proof of Proposition 3

The *non-critical* customer class fill rate for the system $(\Delta+k,k)$ is given by

$$\begin{aligned}\beta_n(\Delta+k,k) &= \sum_{r=0}^{\Delta-1}\pi_r(\Delta+k,k)\ \text{(by (6))}\\ &= \sum_{r=0}^{\Delta-1}\pi_r(\Delta,0)\ \text{(by Corollary 1)}\\ &= \beta_n(\Delta,0).\end{aligned}$$

The *critical* customer class fill rate for the system $(\Delta + k, k)$ is given by

$$\begin{aligned}
\beta_c(\Delta + k, k) &= \sum_{r=0}^{\Delta+k-1} \pi_r(\Delta + k, k) \text{ (by (7))} \\
&= \sum_{r=0}^{\Delta+k-1} \pi_r(\Delta, 0) \text{ (by Corollary 1)} \\
&= \sum_{r=0}^{\Delta-1} \pi_r(\Delta, 0) + \sum_{r=\Delta}^{\Delta+k-1} \pi_r(\Delta, 0) \\
&= \beta_c(\Delta, 0) + \sum_{u=0}^{k-1} \pi_{\Delta+u}(\Delta, 0).
\end{aligned}$$

The expected number of the *critical* class backorders is given by

$$\begin{aligned}
\mathrm{B}(\Delta + k, k) &= \sum_{i=0}^{\infty} i\, \pi_{\Delta+k+i}\,(\Delta + k, k) \\
&= \sum_{i=0}^{\infty} i\, \pi_{\Delta+k+i}\,(\Delta, 0) \text{ (by Corollary 1)} \\
&= \sum_{u=k}^{\infty} (u - k)\, \pi_{\Delta+u}\,(\Delta, 0) \text{ (by change of variables).}
\end{aligned}$$

The expected on-hand stock is given by

$$\begin{aligned}
\mathrm{I}(\Delta + k, k) &= \sum_{u=0}^{\Delta+k} (\Delta + k - u)\, \pi_u\,(\Delta + k, k) \\
&= \sum_{u=0}^{\Delta+k} (\Delta + k - u)\, \pi_u\,(\Delta, 0) \text{ (by Corollary 1).}
\end{aligned}$$

The expected number of units in the resupply is given by

$$\begin{aligned}
\mathrm{X}(\Delta + k, k) &= \sum_{u=0}^{\infty} u\, \pi_u\,(\Delta + k, k) \\
&= \sum_{u=0}^{\infty} u\, \pi_u\,(\Delta, 0) \text{ (by Corollary 1)} \\
&= \mathrm{X}(\Delta, 0).
\end{aligned}$$

E Proof of Proposition 4

The cost function $C(S, S_c)$ for the proposed policy can be bounded from below as follows:

$$\begin{aligned} C(S,S_c) &= \rho_n\lambda_n(1-\beta_n(S,S_c)) + \rho_c\lambda_c(1-\beta_c(S,S_c)) + b\ \mathrm{B}(S,S_c) + h\ \mathrm{I}(S,S_c) \\ &= \rho_n\lambda_n(1-\beta_n(S,S_c)) + \rho_c\lambda_c(1-\beta_c(S,S_c)) + (b+h)\ \mathrm{B}(S,S_c) + h\ (\mathrm{I}(S,S_c) - \mathrm{B}(S,S_c)) \\ &\geq \rho_n\lambda_n(1-\beta_n(S,S_c)) + (b+h)\ \mathrm{B}(S,S_c) + h\ (\mathrm{I}(S,S_c) - \mathrm{B}(S,S_c)) \\ &\geq \rho_n\lambda_n(1-\beta_n(S,S_c)) + h\ (\mathrm{I}(S,S_c) - \mathrm{B}(S,S_c)) \\ &= \rho_n\lambda_n(1-\beta_n(S,S_c)) + h\ (S - \mathrm{X}(S,S_c))\,. \end{aligned}$$

■

F Proof of Proposition 5

For the special case $(S, 0)$, the knowledge of the number of units in the resupply system is sufficient to characterize the system states as follows:

$$(i,j) = \begin{cases} (i,0), & 0 \leq i \leq S, \\ (i, i-S), & S+1 \leq i. \end{cases} \tag{51}$$

So, at any point in time, there is one-to-one mapping between the current system state and the current number of units in the resupply system. Therefore, the system can be modeled as a $M/G/\infty$ queue with mean service time T and state dependent arrival rates

$$\lambda^{(i,j)} = \begin{cases} \lambda, & 0 \leq i \leq S-1, \\ \lambda_n, & S \leq i. \end{cases} \tag{52}$$

In our rationing model with two priority demand classes, if we define $\alpha_{i-S} = \frac{\lambda_n}{\lambda}$ for $i \geq S$, then follow similar arguments as in the proof of Proposition 1, we obtain the results presented in the proposition.

■

G Proof of Lemma 2

In the proof, we will use some of the established probabilities in the proof of Theorem 1 in Appendix H, as well as the results previously established by Vicil and Jackson (2016) for the pure backorder system in a similar model setting. Let $G(\cdot)$ denote the probability distribution of the lead time.

Let $\tilde{p}_{t,t+\tau}$ be the probability that a unit in resupply at time t will still be in the resupply system at time $t+\tau$. And let

$$\begin{aligned} q_{t,t+\tau}(x|n) &= P[x \text{ of those } n \text{ units remain in the resupply at } t+\tau \mid R(t)=n] \\ &= \binom{n}{x} \tilde{p}_{t,t+\tau}^{\,x} \, (1-\tilde{p}_{t,t+\tau})^{n-x} \,. \end{aligned}$$

Starting from the initial system state $\xi_0 = (0,0)$ in which there are no units in the resupply at time 0, and assuming $R(t) = r$ is finite,

$$\begin{aligned} & P\big[\text{there are two or more state-transitions during } (t,t+\tau] \mid \xi_t = (r,b_n)\big] \\ &= 1 - P\big[\text{there is at most one state-transition during } (t,t+\tau] \mid \xi_t = (r,b_n)\big] \\ &= 1 - P\big[\text{no demand occurs during } (t,t+\tau];\ \text{all } r \text{ units in the resupply} \\ &\qquad\qquad \text{at time } t \text{ are still in the resupply at time } t+\tau \mid \xi_t = (r,b_n)\big] \\ &\quad - P\big[\text{only one demand occurs during } (t,t+\tau];\ \text{all } r \text{ units in the resupply} \\ &\qquad\qquad \text{at time } t \text{ are still in the resupply at time } t+\tau \mid \xi_t = (r,b_n)\big] \\ &\quad - P\big[\text{no demand occurs during } (t,t+\tau];\ \text{among the } r \text{ units in the resupply} \\ &\qquad\qquad \text{at time } t, \text{ only one of them is received during } (t,t+\tau] \mid \xi_t = (r,b_n)\big] \\ &= 1 - e^{-\lambda\tau} \, q_{t,t+\tau}(r \mid r) - \lambda\tau e^{-\lambda\tau} \, q_{t,t+\tau}(r \mid r) \\ &\qquad\qquad - e^{-\lambda\tau} \, q_{t,t+\tau}(r \mid r-1) \text{ (due to } \textit{Approximation Assumption 2}) \\ &= 1 - e^{-\lambda\tau} \binom{r}{r} \tilde{p}_{t,t+\tau}^{\,r} \, (1-\tilde{p}_{t,t+\tau})^0 - \lambda\tau e^{-\lambda\tau} \binom{r}{r} \tilde{p}_{t,t+\tau}^{\,r} \, (1-\tilde{p}_{t,t+\tau})^0 \\ &\qquad\qquad - e^{-\lambda\tau} \binom{r}{r-1} \tilde{p}_{t,t+\tau}^{\,r-1} \, (1-\tilde{p}_{t,t+\tau}) \,, \quad \text{(with reference to (54))} \\ &= 1 - e^{-\lambda\tau} \, \tilde{p}_{t,t+\tau}^{\,r} - \lambda\tau e^{-\lambda\tau} \, \tilde{p}_{t,t+\tau}^{\,r} - e^{-\lambda\tau} \, r \, \left(\tilde{p}_{t,t+\tau}^{\,r-1} - \tilde{p}_{t,t+\tau}^{\,r}\right) \end{aligned} \tag{53}$$

A function $f(h)$ is $o(h)$ if $\lim_{h\to 0} \frac{f(h)}{h} = 0$. Hence, to prove whether the probability expression on the right hand side of (53) is $o(\tau)$, we need to determine the limit of the expression as $\tau \to 0$. The individual limits are presented below.

$$\lim_{\tau\to 0} \frac{1 - e^{-\lambda\tau} \, \tilde{p}_{t,t+\tau}^{\,r}}{\tau} = \lambda + r \frac{G(t)}{\int_0^t [1-G(u)]\mathrm{d}u}, \quad \text{(due to Lemma A1 of Vicil and Jackson, 2016).}$$

$$\lim_{\tau\to 0} \frac{\lambda\tau e^{-\lambda\tau} \, \tilde{p}_{t,t+\tau}^{\,r}}{\tau} = \lambda.$$

$$\lim_{\tau\to 0}\frac{e^{-\lambda\tau}\, r\, \left(\tilde{p}^{r-1}_{t,t+\tau} - \tilde{p}^{r}_{t,t+\tau}\right)}{\tau} = \lim_{\tau\to 0}\left[e^{-\lambda\tau}\, r\right]\, \lim_{\tau\to 0}\left[\frac{\left(\tilde{p}^{r-1}_{t,t+\tau} - \tilde{p}^{r}_{t,t+\tau}\right)}{\tau}\right]$$
$$= r\frac{G(t)}{\int_0^t [1-G(u)]\mathrm{d}u}, \quad \text{(due to Lemma A1 of Vicil and Jackson, 2016).}$$

If we plug these individual limits into the right hand side of (53), the terms cancel each other and we get zero.

■

H Proof of Theorem 1

For the analysis, we are going to rely on the notation and results established earlier by Vicil and Jackson (2016) for the pure backorder system in a similar model setting. Let $G(\cdot)$ denote the probability distribution of the lead time. Let p_t be the common probability that any replenishment order which is placed during $[0,t)$ remains in the resupply system at time t. Conditioning on the time of the order placement and using *Approximation Assumption* 1, we have:

$$p_t = \int_0^t [1-G(t-s)]\frac{1}{t}\mathrm{d}s.$$

Let $\tilde{p}_{t,t+\tau}$ be the probability that a unit in resupply at time t will still be in the resupply system at time $t+\tau$. Conditioning on the time of the order placement, which belongs in $[0,t)$, this probability is given by:

$$\tilde{p}_{t,t+\tau} = \frac{P\left\{\text{an order is placed in } [0,t) \text{ and is in resupply at time } t+\tau\right\}}{P\left\{\text{an order is placed in } [0,t) \text{ and is in resupply at time } t\right\}}$$
$$= \frac{\int_0^t [1-G(t+\tau-s)]\frac{1}{t}\mathrm{d}s}{\int_0^t [1-G(t-s)]\frac{1}{t}\mathrm{d}s}$$
$$= \frac{\int_0^t [1-G(t+\tau-s)]\mathrm{d}s}{\int_0^t [1-G(t-s)]\mathrm{d}s}.$$

Note that $\tilde{p}_{t,t+\tau}$ represents the probability for unordered replenishment orders in the resupply.

Let

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

Each replenishment order that is in the resupply at time t, has a probability $\tilde{p}_{t,t+\tau}$ that it will be still in the resupply at time $t+\tau$. Therefore, the probability that among the n replenishment

orders in the resupply at time t, x of them will remain in the resupply at time $t+\tau$ is given by

$$q_{t,t+\tau}(x|n) = \binom{n}{x} \tilde{p}_{t,t+\tau}^{x} \ (1-\tilde{p}_{t,t+\tau})^{n-x}\,. \tag{54}$$

Let us denote the system state at time t by $\xi_t = (r, b_n)$. Starting from the initial state $(0,0)$ at time $t=0$, let

$$P_{(\bar{r},\bar{b}_n),(r,b_n)}(t,t') \ \ = \ \ P\big[\xi_{t'} = (r,b_n) \,|\, \xi_0 = (0,0),\ \xi_t = (\bar{r},\bar{b}_n)\big].$$

By conditioning on the state of the system at time t, the probability of being at system state (r,b_n) at time $t' = t+\tau$ can be written as

$$P_{(0,0),(r,b_n)}(0,t+\tau) \ = \sum_{(\bar{r},\bar{b}_n)\in\mathbb{F}_{(S,S_c)}} P_{(0,0),(\bar{r},\bar{b}_n)}(0,t)\cdot P_{(\bar{r},\bar{b}_n),(r,b_n)}(t,t+\tau). \tag{55}$$

One-step transition probabilities will be solved for a general system state $(r,b_n) \in \mathbb{F}_{(S,S_c)}$ with $S-S_c+b_n < r < S+b_n$ and $b_n \geq 1$ (hence, $OH < S_c$). Other system states can be solved similarly. By conditioning on the state of the system at time t, there are four possible ways to reach state (r,b_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, or nothing happens. Since τ is an infinitesimal time, the probability of two or more events happening during $(t,t+\tau]$ is captured within the term $o(\tau)$ due to Lemma 2. For the infinitesimal transition probability calculations, we assume $R(t) = r$ is finite for all t. Hence, we have

$$\begin{aligned} P_{(0,0),(r,b_n)}(0,t+\tau) \ \ = \ \ \Big\{ & P_{(0,0),(r-1,b_n-1)}(0,t)\cdot P_{(r-1,b_n-1),(r,b_n)}(t,t+\tau) \\ & + P_{(0,0),(r-1,b_n)}(0,t)\cdot P_{(r-1,b_n),(r,b_n)}(t,t+\tau) \\ & + P_{(0,0),(r+1,b_n)}(0,t)\cdot P_{(r+1,b_n),(r,b_n)}(t,t+\tau) \\ & + P_{(0,0),(r,b_n)}(0,t)\cdot P_{(r,b_n),(r,b_n)}(t,t+\tau) \\ & + o(\tau) \Big\}. \end{aligned} \tag{56}$$

Next, let us determine the one-step transition probabilities in (56).

a) A *non-critical* demand occurs:

$$\begin{aligned}
P_{(r-1,b_n-1),(r,b_n)}(t,t+\tau) &= P\big[\text{only a } \textit{non-critical} \text{ demand occurs during } (t,t+\tau];\ \text{all } r-1 \text{ units in} \\
&\qquad \text{the resupply at time } t \text{ are still in the resupply at time } t+\tau \\
&\qquad \mid \xi_t = (r-1, b_n-1),\ \xi_0 = (0,0)\big] \ + \ o(\tau) \\
&= \lambda_n \tau e^{-\lambda\tau}\ q_{t,t+\tau}(r-1 \mid r-1) \ + \ o(\tau) \ \text{(due to } \textit{Approximation Assumption } 2) \\
&= \lambda_n \tau e^{-\lambda\tau} \binom{r-1}{r-1} \tilde{p}_{t,t+\tau}^{\,r-1}\ (1-\tilde{p}_{t,t+\tau})^0 \ + \ o(\tau) \ \text{(with reference to (54))} \\
&= \lambda_n \tau e^{-\lambda\tau}\ \tilde{p}_{t,t+\tau}^{\,r-1} \ + \ o(\tau). \qquad (57)
\end{aligned}$$

b) A *critical* demand occurs:

$$\begin{aligned}
P_{(r-1,b_n),(r,b_n)}(t,t+\tau) &= P\big[\text{only a } \textit{critical} \text{ demand occurs during } (t,t+\tau];\ \text{all } r-1 \text{ units in} \\
&\qquad \text{the resupply at time } t \text{ are still in the resupply at time } t+\tau \\
&\qquad \mid \xi_t = (r-1, b_n),\ \xi_0 = (0,0)\big] \ + \ o(\tau) \\
&= \lambda_c \tau e^{-\lambda\tau}\ q_{t,t+\tau}(r-1 \mid r-1) \ + \ o(\tau) \ \text{(due to } \textit{Approximation Assumption } 2) \\
&= \lambda_c \tau e^{-\lambda\tau} \binom{r-1}{r-1} \tilde{p}_{t,t+\tau}^{\,r-1}\ (1-\tilde{p}_{t,t+\tau})^0 \ + \ o(\tau) \ \text{(with reference to (54))} \\
&= \lambda_c \tau e^{-\lambda\tau}\ \tilde{p}_{t,t+\tau}^{\,r-1} \ + \ o(\tau). \qquad (58)
\end{aligned}$$

c) Delivery from the resupply:

$$\begin{aligned}
P_{(r+1,b_n),(r,b_n)}(t,t+\tau) &= P\big[\text{no demand occurs during } (t,t+\tau];\ \text{among the } r+1 \text{ units in the} \\
&\qquad \text{resupply at time } t, \text{ only one of them is received during } (t,t+\tau] \\
&\qquad \mid \xi_t = (r+1, b_n),\ \xi_0 = (0,0)\big] \ + \ o(\tau) \\
&= e^{-(\lambda_n+\lambda_c)\tau}\ q_{t,t+\tau}(r \mid r+1) \ + \ o(\tau) \ \text{(due to } \textit{Approximation Assumption } 2) \\
&= e^{-(\lambda_n+\lambda_c)\tau} \binom{r+1}{r} \tilde{p}_{t,t+\tau}^{\,r}\ (1-\tilde{p}_{t,t+\tau}) \ + \ o(\tau) \ \text{(with reference to (54))} \\
&= e^{-(\lambda_n+\lambda_c)\tau}\ (r+1)\left(\tilde{p}_{t,t+\tau}^{\,r} - \tilde{p}_{t,t+\tau}^{\,r+1}\right) \ + \ o(\tau). \qquad (59)
\end{aligned}$$

d) Nothing happens:

$$\begin{aligned}
P_{(r,b_n),(r,b_n)}(t,t+\tau) &= P\big[\text{no demand occurs during } (t,t+\tau];\ \text{all } r \text{ units in the} \\
&\qquad \text{resupply at time } t \text{ are still in the resupply at time } (t,t+\tau] \\
&\qquad \mid \xi_t = (r,b_n),\ \xi_0 = (0,0)\big] \ + \ o(\tau) \\
&= e^{-(\lambda_n+\lambda_c)\tau}\ q_{t,t+\tau}(r \mid r) \ + \ o(\tau) \ \text{(due to } \textit{Approximation Assumption } 2) \\
&= e^{-(\lambda_n+\lambda_c)\tau}\binom{r}{r}\ \tilde{p}^{\,r}_{t,t+\tau}\ (1-\tilde{p}_{t,t+\tau})^0 \ + \ o(\tau) \ \text{(with reference to (54))} \\
&= e^{-(\lambda_n+\lambda_c)\tau}\ \tilde{p}^{\,r}_{t,t+\tau} \ + \ o(\tau). \qquad (60)
\end{aligned}$$

Subtracting $P_{(0,0),(r,b_n)}(0,t)$ from both sides of (56) and taking the limits as $\tau \to 0$:

$$\begin{aligned}
\lim_{\tau\to 0}\frac{P_{(0,0),(r,b_n)}(0,t+\tau)-P_{(0,0),(r,b_n)}(0,t)}{\tau} = {} & P_{(0,0),(r-1,b_n-1)}(0,t)\cdot\lim_{\tau\to 0}\frac{P_{(r-1,b_n-1),(r,b_n)}(t,t+\tau)}{\tau} \\
& + P_{(0,0),(r-1,b_n)}(0,t)\cdot\lim_{\tau\to 0}\frac{P_{(r-1,b_n),(r,b_n)}(t,t+\tau)}{\tau} \\
& + P_{(0,0),(r+1,b_n)}(0,t)\cdot\lim_{\tau\to 0}\frac{P_{(r+1,b_n),(r,b_n)}(t,t+\tau)}{\tau} \qquad (61) \\
& - P_{(0,0),(r,b_n)}(0,t)\cdot\lim_{\tau\to 0}\frac{\big(1-P_{(r,b_n),(r,b_n)}(t,t+\tau)\big)}{\tau} \\
& + \lim_{\tau\to 0}\frac{o(\tau)}{\tau}.
\end{aligned}$$

The left-hand side of Equation (61) is $P'_{(0,0),(r,b_n)}(0,t)$. The limit terms on the right-hand side are provided by Vicil and Jackson (2016). Plugging their values, we have

$$\begin{aligned}
P'_{(0,0),(r,b_n)}(0,t) &= P_{(0,0),(r-1,b_n-1)}(0,t)\lambda_n \\
&\quad + P_{(0,0),(r-1,b_n)}(0,t)\lambda_c \\
&\quad + P_{(0,0),(r+1,b_n)}(0,t)(r+1)\frac{G(t)}{\int_0^t[1-G(u)]\,\mathrm{d}u} \\
&\quad - P_{(0,0),(r,b_n)}(0,t)\Big(\lambda_n+\lambda_c \ + \ r\frac{G(t)}{\int_0^t[1-G(u)]\,\mathrm{d}u}\Big).
\end{aligned}$$

Taking the limits as $t \to \infty$:

$$\begin{aligned}\lim_{t\to\infty} P'_{(0,0),(r,b_n)}(0,t) &= \lim_{t\to\infty} P_{(0,0),(r-1,b_n-1)}(0,t)\lambda_n \\ &+ \lim_{t\to\infty} P_{(0,0),(r-1,b_n)}(0,t)\lambda_c \\ &+ \lim_{t\to\infty} P_{(0,0),(r+1,b_n)}(0,t)(r+1)\frac{G(t)}{\int_0^t [1-G(u)]\,\mathrm{d}u} \\ &- \lim_{t\to\infty} P_{(0,0),(r,b_n)}(0,t)\Big(\lambda_n + \lambda_c + r\frac{G(t)}{\int_0^t [1-G(u)]\,\mathrm{d}u}\Big). \qquad (62)\end{aligned}$$

$P_{(0,0),(r,b_n)}(0,t)$ is bounded by 0 and 1 for all t. Therefore, if $\lim_{t\to\infty} P'_{(0,0),(r,b_n)}(0,t)$ converges, then it must converge to 0. Since $\lim_{t\to\infty} \frac{G(t)}{\int_0^t [1-G(u)]\,\mathrm{d}u} = \frac{1}{T}$ and assuming the steady-state distributions exist, the limit exists for the RHS of (62). Hence $\lim_{t\to\infty} P'_{(0,0),(r,b_n)}(0,t) = 0$. After reordering the terms, (62) becomes:

$$\Big(\lambda_n + \lambda_c + \frac{r}{T}\Big)\pi_{(r,b_n)} = \lambda_n \pi_{(r-1,b_n-1)} + \lambda_c \pi_{(r-1,b_n)} + \frac{r+1}{T}\pi_{(r+1,b_n)}.$$

■