IIE Transactions

ISSN: 0740-817X (Print) 1545-8830 (Online) Journal homepage: http://www.tandfonline.com/loi/uiie20

Computationally efficient optimization of stock pooling and allocation levels for two-demand-classes under general lead time distributions

Oguzhan Vicil & Peter Jackson

To cite this article: Oguzhan Vicil & Peter Jackson (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, DOI: 10.1080/0740817X.2016.1146421

To link to this article: https://doi.org/10.1080/0740817X.2016.1146421

IIE TRANSACTIONS
2016, VOL. 48, NO. 10, 955–974
http://dx.doi.org/10.1080/0740817X.2016.1146421

Taylor & Francis
Taylor & Francis Group

Computationally efficient optimization of stock pooling and allocation levels for two-demand-classes under general lead time distributions

Oguzhan Vicil[a] and Peter Jackson[b]

[a]Department of Industrial Engineering, Bilkent University, Ankara, Turkey; [b]School of Operations Research & Information Engineering, Cornell University, Ithaca, NY, USA

ABSTRACT

In this article we develop a procedure for estimating service levels (fill rates) and for optimizing stock and threshold levels in a two-demand-class model managed based on a lot-for-lot replenishment policy and a static threshold allocation policy. We assume that the priority demand classes exhibit mutually independent, stationary, Poisson demand processes and non-zero order lead times that are independent and identically distributed. A key feature of the optimization routine is that it requires computation of the stationary distribution only once. There are two approaches extant in the literature for estimating the stationary distribution of the stock level process: a so-called single-cycle approach and an embedded Markov chain approach. Both approaches rely on constant lead times. We propose a third approach based on a Continuous-Time Markov Chain (CTMC) approach, solving it exactly for the case of exponentially distributed lead times. We prove that if the independence assumption of the embedded Markov chain approach is true, then the CTMC approach is exact for general lead time distributions as well. We evaluate all three approaches for a spectrum of lead time distributions and conclude that, although the independence assumption does not hold, both the CTMC and embedded Markov chain approaches perform well, dominating the single-cycle approach. The advantages of the CTMC approach are that it is several orders of magnitude less computationally complex than the embedded Markov chain approach and it can be extended in a straightforward fashion to three demand classes.

ARTICLE HISTORY
Received 30 September 2013
Accepted 23 December 2015

KEYWORDS
Threshold rationing policies; Markov process; inventory control; priority demand classes; service levels; backorder systems; service parts logistics; inventory allocation

1. Introduction

Inventory rationing among different customer classes arises in several contexts. Our primary motivation is the situation of managing service parts inventory in a parts distribution center serving multiple customer classes, each of which has contracted for a specific level of customer service, typically measured as fill rate. By pooling demand across classes, the parts manager can reduce safety stock requirements and, by setting threshold levels for allocation and backorder clearing, the parts manager can achieve differentiated service levels by demand class.

This area has been an active subject of research for several decades. It remains a challenging problem due to the difficulty of efficiently computing accurate service-level performance measures. In this article, we focus on a two-demand-class stocking problem, in which pooled inventory is managed with a continuous review order-up-to policy, together with a static rationing policy, under which low-priority customers are not served as long as the on-hand inventory is at or below a fixed threshold level. Essentially only two methods are proposed in the literature for this problem and both methods assume constant order lead times. The single-cycle approach (Dekker *et al.*, 1998; Deshpande *et al.*, 2003) assumes that no orders were outstanding a lead time ago. This dramatically simplifies the service-level calculation. The embedded Markov chain approach (Fadiloglu and Bulut, 2010b) samples the system at multiples of the lead time and approximates the transition probabilities assuming that delivery times are independent of the number of low-priority backorders. The embedded Markov chain approach dominates the single-cycle approach in accuracy of estimating service levels; however, it is much more computationally complex.

In this article, we introduce a third approach and evaluate all three methods in the context of general lead time distributions. We use a continuous-time Markov chain approach and solve for exact expressions of the service levels under the assumption of exponentially distributed lead times. We further show that if the independence assumption of the embedded Markov chain approach is true, then these same service-level expressions are true for generally distributed lead times. These expressions are computed using recursive procedures that are several orders of magnitude less computationally complex than the embedded Markov chain approach. We further show that the stock optimization problem can be solved with simple line searches and a single evaluation of the stationary probability distribution.

Our computational studies reveal that service levels are relatively insensitive to the form of the lead time distribution. Consequently, any good approximation algorithm, whether for constant or exponentially distributed lead times, may be used with general lead time distributions with likely good results. In particular, the continuous-time Markov chain approach is close in accuracy to the embedded Markov chain approach and both

CONTACT Oguzhan Vicil ✉ oguzhanvicil@yahoo.com
Supplemental data for this article can be accessed on the publisher's website at www.tandfonline.com/uiie.

methods dominate the single-cycle approach, performing well over a broad range of lead time distributions. The embedded Markov chain performs best with low coefficients of variation of the lead time and the continuous-time Markov chain performs better as the coefficient of variation increases. Furthermore, as shown in a companion paper (Vicil and Jackson, 2015), it is straightforward to extend the continuous-time Markov chain approach to three demand classes. In summary, this article presents an extensible and computationally efficient way of computing optimal stocking and threshold levels for a two-demand-class model and provides service level estimates comparable to the best existing heuristic.

The article is organized as follows. After reviewing the literature in Section 2, we introduce the stocking and allocation optimization model in Section 3 together with properties of the stationary probability distribution. We exploit these properties in Section 4 to present an algorithm to solve the optimization problem. The unique feature of this algorithm is that it requires only a single computation of the stationary probabilities. The remainder of this article focuses on this computation. In Section 5, we consider the special case of exponentially distributed lead times. This exact approach exploits the special structure of the probability transition matrix in a novel way and results in an efficient recursive procedure. In Section 6, we describe counterexamples that hamper traditional approaches to exact analysis of such problems under generally distributed lead time settings. We cast the essential difficulty as a theorem, highlighting the so-called Independence Condition, which, if satisfied, leads to a distribution-invariant result. In Section 7, we use numerical simulation to evaluate the quality of the Continuous-Time Markov Chain (CTMC) approximation under a variety of lead time probability distributions for two-demand-class models and compare the relative quality with the existing heuristics. In the conclusion, we note that the model can be extended to three-demand-class models but leave the development and discussion to a supplemental technical report (Vicil and Jackson, 2015).

2. Literature review

Kleijn and Dekker (1999) review the literature on inventory systems with multiple demand classes. Their taxonomy focuses on two characteristics: (i) periodic versus continuous review and (ii) the number of demand classes considered (two or more). We highlight several papers published after their work. Ding *et al.* (2006) consider a single-period, single-item, multiple-class model that allows the use of dynamic price discounts to encourage backlogging of demand for customers classes not immediately satisfied. They determine the optimal discounts to offer and characterize the optimal inventory allocation policy. In a subsequent paper, Ding *et al.* (2007) consider an infinite horizon, deterministic demand, economic-order-quantity-like environment with holding, backorder, lost demand, and setup costs. They determine the optimal policy in this deterministic environment, where partial backlogging of unfilled demand is possible, based on dynamic price discounts. They also study the effect of changes in various system parameters on performance measures such as profitability and customer service. Teunter and Haneveld (2008) consider a single-period, two-demand-class model with Poisson demand and backordering. They study a dynamic rationing strategy where the number of units reserved for critical demand depends on the remaining time until the next order arrives. They derive a set of expressions that enable calculation of the optimal rationing level based on the time remaining. In a later study, Fadiloglu and Bulut (2010a) conduct simulation studies for both backordering and lost sales environments, in order to compare the performance of the dynamic policy with the static critical level and common stock policies and to quantify the gain obtained from dynamic rationing.

The following papers employ a static threshold policy and develop procedures to determine the optimal threshold level(s). Nahmias and Demmy (1981) use a continuous review (c, s, Q) policy for two demand classes (the parameter c is the critical level for on-hand inventory below which low-priority customers are not served). In their model, backorders are allowed, demand is a Poisson process, and the order lead time is constant. They further assume that there is at most one order outstanding. Deshpande *et al.* (2003) consider a situation similar to that of Nahmias and Demmy (1981) but allow multiple replenishment orders to be present in the pipeline at the same time. They use a hitting time approach with a creative approximation to the distribution of backorders among customer classes at the time a replenishment order arrives. Empirical results demonstrate that the approximation is quite good for the parameters considered. Deshpande and Cohen (2005) extend their threshold clearing mechanism from two to N demand classes. Also, in a similar work, Arslan *et al.* (2007) analyze a single-location, single-product inventory rationing problem for N demand classes that are characterized by different shortage costs or service requirements. They assume a backorder clearing mechanism, in which a backorder for a lower-priority class is treated as equivalent to a reserve-stock shortfall for the higher-priority class. They propose a computationally efficient heuristic and develop a bound on its performance. They also show that there is sample-path equivalence between their backorder clearing rule and the threshold clearing rule in Deshpande *et al.* (2003) and Deshpande and Cohen (2005).

Dekker *et al.* (1998) use a continuous review $(c, S-1, S)$ policy for two demand classes. As with Nahmias and Demmy (1981), their model is based on the assumption that excess demand is backordered. The demand process is a Poisson process and order lead times are constant. They use a hitting time approach under the approximating assumption that there was no order outstanding a lead time ago. The accuracy of the approach can be increased by assuming, instead, that there was no order outstanding two lead times ago. Kocaga and Sen (2007) study a similar environment as Dekker *et al.* (1998); however, their model differs in the way the non-critical orders are satisfied. According to their model, critical orders are due immediately, whereas non-critical orders allow for a deterministic demand lead time. They provide an approximation for the critical service level while the service level for the non-critical demand is exact. Dekker *et al.* (2002) consider a $(c, S-1, S)$ replenishment policy for N demand classes (c is an N-dimensional vector in this case). Their model includes lost sales, Poisson demand processes, and a general lead time distribution. The lost sales assumption simplifies the state space. They derive the exact steady-state distribution of on-hand inventory and, from there, develop techniques to find optimal policy parameters.

The problem we consider is most closely related to the models in Dekker *et al.* (1998) and Dekker *et al.* (2002). For zero setup costs, it is also identical to the model of Deshpande *et al.* (2003). We focus on $(S-1, S)$ replenishment policies, as these are appropriate in the high-cost, low-demand-rate service parts distribution contexts of our applied work. We also assume a static threshold policy and seek to determine the service levels provided for each customer class.

Fadiloglu and Bulut (2010b) consider a model that is identical to the one developed in this article, but is restricted to a constant lead time. They suggest that an embedded Markov chain approach can be used to estimate the stationary probability distribution by sampling the system at multiples of the lead time. The transition probabilities are approximated under the assumption that delivery times are independent of the number of low-priority backorders. They provide a recursive procedure for computing the transition probabilities of the Markov chain. The stationary probabilities are computed as the limit of a convergent sequence of bounds using a sophisticated technique from computational algebra. They demonstrate through simulation that the approximation is quite good. We show that the assumption that delivery times are independent of the number of low-priority backorders permits the analysis of the same model under general lead time distributions. However, instead of a Markov chain approach, we are led to the analysis of a continuous-time Markov process. We refer to our approach as the CTMC approach to distinguish it from the embedded Markov chain approach of Fadiloglu and Bulut (2010b). The resulting algorithm is several orders of magnitude less computationally complex than the embedded Markov chain approach.

3. Stock optimization for a two-demand-class model

We consider a model with two priority demand classes: *gold* and *silver*. The *gold* customers have contracted for a service-level fill rate, c_g, and the silver customers have contracted for fill rate c_s, with $c_g > c_s$. We assume that the demand streams for *gold* and *silver* customers are independent Poisson processes with demand rates λ_g and λ_s, respectively, and that the demands for both classes can be backordered. We further assume that replenishment orders for the product are placed according to a continuous review $(S-1, S)$ policy based on inventory position. Hence, the arrival of any demand, by either a *gold* or a *silver* customer, triggers an immediate replenishment order of size 1. Service is differentiated using a threshold level, S_g. No *silver* demand or backorder is satisfied as long as the on-hand inventory, OH, is at or below S_g. *Gold* demands are backordered only if the on-hand inventory is zero. The overall policy is referred to as a *lot-for-lot replenishment and threshold allocation policy*.

The delivery lead times for successive orders form a sequence of independent and identically distributed random variables with mean T. In this article, we consider simulations of the system using a variety of lead time probability distributions including the constant, the exponential, the Erlang, the geometric, and the lognormal distributions. However, the service-level estimation techniques discussed rely only on the value of T, the mean lead time. Consequently, the model is parameterized by the vector $(S, S_g; \lambda_g, \lambda_s, T)$.

Let β_g (respectively, β_s) denote the steady-state fill rate for *gold* (respectively, *silver*) customers as functions of the parameters $(S, S_g; \lambda_g, \lambda_s, T)$. Denote the stationary probability distribution of a random process by $P_\infty(\cdot)$. By the PASTA principle (Poisson Arrivals See Time Averages), arriving demands face the stationary distribution of on-hand inventory, *OH* (Tijms, 1986). A *silver* customer arrival will be served if and only if $OH > S_g$, whereas a *gold* customer arrival will be served if and only if $OH > 0$.

Consequently,

$$\beta_s = 1 - P_\infty(OH \le S_g),$$

and

$$\beta_g = 1 - P_\infty(OH = 0).$$

A natural formulation of the optimization problem is

$$\begin{aligned} &\min\ S\\ &\text{subject to:}\\ &\quad \beta_s(S, S_g) \ge c_s\\ &\quad \beta_g(S, S_g) \ge c_g\\ &\quad S \ \ge\ S_g \ge 0 \end{aligned}$$

for management-specified service levels c_g and $c_s, c_g > c_s$. That is, we seek the minimum target level of inventory required to achieve the service level constraints.

At any time t, let $OH(t)$ denote the number of units on hand, let $R(t)$ denote the number of units in resupply, let $B_g(t)$ denote the number of outstanding *gold* backorders, and let $B_s(t)$ denote the number of outstanding *silver* backorders. Under the *lot-for-lot replenishment and threshold allocation policy*, the following relations hold:

$$S = OH(t) + R(t) - B_g(t) - B_s(t) \tag{1}$$

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

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

These relations will also apply to the stationary distribution of these quantities denoted by OH, R, B_g, and B_s. Consequently, it is sufficient to capture the stationary distribution of the pair (R, B_s), the number of units in resupply, and the number of outstanding *silver* backorders.

If we consider only the number of units in resupply, R, then only the replenishment policy has any effect and the resulting system can be analyzed according to a single-demand-class system with demand rate $\lambda = \lambda_g + \lambda_s$. Due to Palm's Theorem (Muckstadt and Sapra, 2010), for a general, positively valued lead time distribution with no probability mass at zero, the steady-state distribution of the units in resupply is Poisson distributed with mean λT. The importance of Palm's Theorem is that the form of the lead time probability distribution has no effect on the stationary behavior of the system beyond the mean of the distribution. It follows that the *silver* fill rate is given by

$$\begin{aligned} \beta_s &= P_\infty(OH > S_g)\\ &= P_\infty(R < S - S_g)\\ &= \sum_{k=0}^{S-S_g-1} \frac{(\lambda T)^k e^{-\lambda T}}{k!}. \end{aligned} \tag{4}$$

Since the *silver* fill rate is easily determined, the challenge in subsequent analysis is to estimate the steady-state distribution of states for which $R \geq S - S_g$ in order to determine β_g, the *gold* fill rate.

The first observation is that a stationary distribution does, in fact, exist over states $(R(t), B_s(t))$ for this system. Let $Z_0 = \{0, 1, 2, \ldots\}$, denote the set of non-negative integers and $\xi_t = (r, b_s) \in Z_0 \times Z_0$ denote the system state at time t, $t \geq 0$. Denote the transition probability by

$$P_{(r,b_s),(r',b_s')}(t, t') = P\left\{\xi_{t'} = (r', b_s') \mid \xi_t = (r, b_s)\right\}.$$

The following theorem establishes the existence of a stationary distribution under general lead time distributions.

Theorem 1. *For a general, positively valued lead time distribution with no probability mass at zero, and for any $(r, b_s) \in Z_0 \times Z_0$, $\lim_{t\to\infty} P_{(0,0),(r,b_s)}(0, t) = \pi_{(r,b_s)}$ exists and is well defined.*

Proof. Provided in the online supplement. □

3.1. Properties of the stationary distribution

In this section we present several invariance and monotonicity results that simplify the optimization problem. We have already seen one invariance result (4) that states that β_s, the *silver* fill rate, depends only on the difference $S - S_g$ and that this result is unaffected by the form of the lead time probability distribution. Let $\pi_{(r,b_s)}(S, S_g)$, $(r, b_s) \in Z_0 \times Z_0$, denote the stationary distribution of $(R(t), B_s(t))$ when the policy parameters are given by (S, S_g). Let $\pi_h(S, S_g)$ denote the stationary distribution of on-hand inventory: for $h = 0, 1, \ldots, S$:

$$\pi_h\left(S, S_g\right) \equiv P_\infty\left(OH(t) = h\right).$$

Recall that $OH(t) = [S - R(t) + B_s(t)]^+$. It follows that the stationary distribution of on-hand inventory is given by

$$\pi_h\left(S, S_g\right) = \sum_{r\in Z_0} \sum_{\substack{b_s\in Z_0 \\ (S-r+b_s)^+=h}} \pi_{(r,b_s)}(S, S_g). \tag{5}$$

For $h > 0$, this can be written as

$$\pi_h\left(S, S_g\right) = \sum_{r=S-h}^{\infty} \pi_{(r,r+h-S)}(S, S_g). \tag{6}$$

Finally, let $\beta_g(S, S_g)$ denote the *gold* fill rate as a function of the policy parameters. We make no claim that these quantities, $\pi_{(r,b_s)}$, π_h, and β_g, are invariant to the form of the lead time probability distribution. However, we do establish certain fundamental properties of these quantities that hold without regard to the lead time distribution.

We use sample path arguments to establish the subsequent results. Beginning from a regeneration point in which no orders are outstanding, let (n, T_n, E_n) describe the nth event in the system: T_n is the time of the nth event, and E_n is the type of event where $E_n \in \{\upsilon, g, s\}$ representing events "delivery," "*gold* demand," and "*silver* demand," respectively. Clearly, $T_n \geq 0$. Let R_n denote the number of units in resupply after the nth event and $B_{s,n}$ denote the number of *silver* backorders after the nth event.

Proposition 1. *The dynamics of $(R_n, B_{s,n})$ can be completely described in terms of the sample path $\{(n, T_n, E_n); n = 1, 2, 3, \ldots\}$:*

$$R_{n+1} = \begin{cases} R_n + 1 & E_n \neq \text{“}\upsilon\text{”}, \\ R_n - 1 & E_n = \text{“}\upsilon\text{”}. \end{cases}$$

$$B_{s,n+1} = \begin{cases} 0, & R_{n+1} \leq S - S_g, \\ B_{s,n} + 1, & E_n = \text{“}s\text{”}, R_n \geq S - S_g, \\ B_{s,n} - 1, & E_n = \text{“}\upsilon\text{”}, R_n > S - S_g, B_{s,n} = R_n - (S - S_g), \\ B_{s,n}\,, & \textit{otherwise}. \end{cases}$$

Proof. We have earlier noted the simplicity of the dynamics for the number of units in resupply. The only situation in which $B_{s,n}$ can be decremented is with the arrival of a delivery ($E_n = \upsilon$) when on-hand inventory prior to the delivery is S_g and there is at least one *silver* backorder. If the on-hand inventory equals S_g, then, by Equation (1), $S - R_n + B_{s,n} = S_g$. The number of *silver* backorders in this case is given by $B_{s,n} = R_n - (S - S_g)$. For there to be at least one *silver* backorder, we require $R_n - (S - S_g) \geq 1$, or equivalently $R_n > (S - S_g)$. This describes the situation where $B_{s,n}$ is decremented. If the on-hand inventory is less than or equal to S_g, then $R_n \geq (S - S_g)$ and any arriving *silver* demand is backordered. Hence, $B_{s,n}$ is incremented in this event. The remaining dynamics are straightforward. □

Let $\Delta = S - S_g$, the difference between the target inventory and the *gold* threshold. As is clear from the formulas, the dynamics governing the sample paths depend only on the value of Δ. Figure 1 illustrates the sample path dynamics by representing all reachable states for $(R_n, B_{s,n})$ and the transitions that can occur to $(R_{n+1}, B_{s,n+1})$ for an arbitrary n. In the figure we classify states based on how an arriving delivery is treated. Observe that if $R_n \leq \Delta$, then a delivery is used to replenish inventory. We classify these states as "deliver to stock." If $R_n > \Delta$, then the treatment of deliveries is restricted. *Silver* backorders can exist only if $R_n > \Delta$. A *silver* backorder is filled by a delivery if and only if $B_{s,n} = R_n - \Delta$. We classify these states as "deliver to *silver*." When $R_n > \Delta$ and $B_{s,n} < R_n - \Delta$, then deliveries are used to satisfy *gold* backorders or to replenish *gold* reserves (up to S_g). We classify these states as "deliver to *gold*." Among these latter states, we further distinguish those states that form the interface between deliver-to-*gold* states and deliver-to-stock or deliver-to-*silver* states. These interface or bridge states have the property $B_{s,n} = R_n - \Delta - 1$. Bridge states play an important role in a subsequent section.

Corollary 1. *The stationary probabilities $\pi_{(r,b_s)}$, $(r, b_s) \in Z_0 \times Z_0$, are invariant to changes in S provided $\Delta = S - S_g$ is constant.*

Proof. The sample path dynamics depend only on Δ, the difference between the target inventory and the *gold* threshold. □

Let $\pi_{(r,b_s)}(\Delta)$, $(r, b_s) \in Z_0 \times Z_0$, denote the stationary probabilities computed using the knowledge that $S - S_g = \Delta$. The remainder of this section assumes that a method for computing these probabilities is available. An example of such a method will be presented in a subsequent section.

We exploit this result to simplify the computation of *gold* fill rates. Suppose we have computed the stationary probability distribution of the on-hand inventory, π_h, $h = 0, 1, \ldots, S$, for

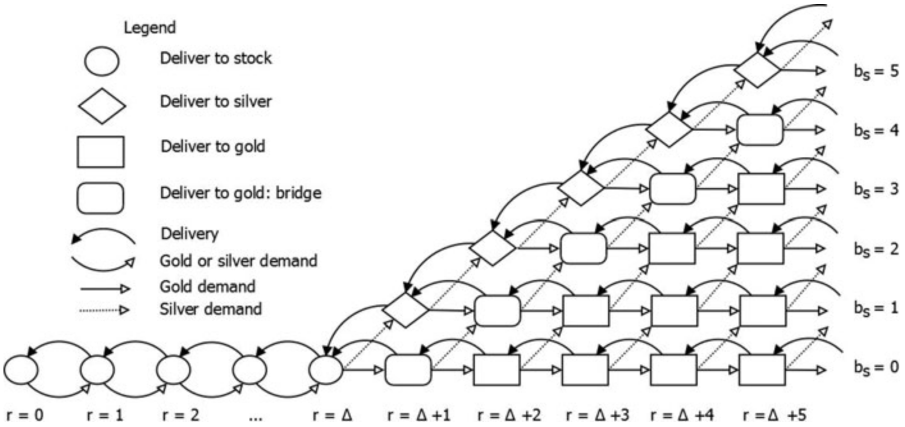

Figure 1. Sample path dynamics.

some combination of parameters S and $S_g = S - \Delta$. The following result shows how to extract the stationary on-hand distribution for other target inventory levels, $S - k$, provided Δ is fixed.

Corollary 2. *For fixed* Δ

$$\pi_h\,(S-k, S-k-\Delta) = \pi_{h+k}\,(S, S-\Delta)$$

for all $k = 0, 1, 2, \ldots, S-\Delta$ *and all* $h = 1, 2, \ldots, S-k$.

Proof. By Equation (5),

$$\begin{aligned}\pi_h(S-k, S-k-\Delta) &= \sum_{r \in Z_0} \sum_{\substack{b_s \in Z_0\\ (S-k-r+b_s)^+ = h}} \pi_{(r,b_s)}(S-k, S-k-\Delta)\\ &= \sum_{r \in Z_0} \sum_{\substack{b_s \in Z_0\\ (S-k-r+b_s)^+ = h}} \pi_{(r,b_s)}(S, S-\Delta),\end{aligned}$$

by the invariance of $\pi_{(r,b_s)}(S, S_g)$ to S when $S - S_g$ is fixed. The result follows easily. □

Corollary 3. *For fixed* Δ, *and* $k = 0, 1, \ldots, S-\Delta$:

$$\beta_g(S-k, S-k-\Delta) = \sum_{j=k+1}^{S} \pi_j(S, S-\Delta).$$

Proof. By definition

$$\begin{aligned}\beta_g(S-k, S-k-\Delta) &= 1 - \pi_0(S-k, S-k-\Delta)\\ &= \sum_{h=1}^{S-k} \pi_h(S-k, S-k-\Delta)\\ &= \sum_{h=1}^{S-k} \pi_{h+k}(S, S-\Delta)\end{aligned}$$

by the invariance result. A change of variables ($j = h + k$) completes the result. □

This last result gives rise to a recursive scheme to compute the *gold* fill rate.

Corollary 4. *When* $\Delta = S - S_g$ *is fixed, we have for* $r = \Delta + 1, \ldots, S$:

$$\beta_g(r, r-\Delta) = \beta_g(r-1, r-1-\Delta) + \pi_{S-r+1}\,(S, S-\Delta)\,.$$

All that is needed to initiate this scheme is $\beta_g(\Delta, 0)$ and $\pi_k(S, S-\Delta)$, $k = 1, \ldots, S-\Delta$. When $S_g = 0$, we have $\beta_g(\Delta, 0) = P_\infty(OH > 0) = P_\infty(OH > S_g) = \beta_s(\Delta)$. That is, for a given value of Δ, we can initiate the calculation of the *gold* fill rate with the *silver* fill rate.

Our final result is a monotonicity property of the *gold* fill rate.

Proposition 2. *For fixed* S, $\beta_g(S, S_g)$ *is nondecreasing in* S_g.

Proof. Consider two systems with identical event sequences $\{(n, T_n, E_n); n = 1, 2, 3, \ldots\}$. In one system, the policy parameters are (S, S_g) and the resulting states are given by $\{(R_n, B_{s,n}); n = 1, 2, 3, \ldots\}$. In the second system, the policy parameters are (S, S_g') with $S_g' > S_g$ and the resulting states are given by $\{(R_n', B_{s,n}'); n = 1, 2, 3, \ldots\}$. We claim that $R_n' = R_n$ and $B_{s,n}' \geq B_{s,n}$. That $R_n' = R_n$ has already been established. To show $B_{s,n}' \geq B_{s,n}$ by induction, we first assume that it is true for some value of n and then establish the result for $n + 1$. Since the backorders change by at most one unit per transition, it suffices to assume $B_{s,n}' = B_{s,n}$ and then show that $B_{s,n+1}' < B_{s,n+1}$ is not possible. Thus, suppose $B_{s,n+1} = B_{s,n} + 1$. This can happen only if a *silver* demand occurs and $R_n \geq S - S_g$. Since $S_g' > S_g$, we have $R_n' = R_n \geq S - S_g'$ and in this situation we will have $B_{s,n+1}' = B_{s,n}' + 1 = B_{s,n} + 1 = B_{s,n+1}$. Now suppose $B_{s,n+1}' = B_{s,n}' - 1$. This can happen only if a delivery occurs and $B_{s,n}' = R_n' - (S - S_g')$. But this would imply

$$B_{s,n} = B_{s,n}' = R_n' - (S - S_g') > R_n - (S - S_g),$$

which is not possible. Under all sample paths, therefore, we have that $B_{s,n}' \geq B_{s,n}$. By Equation (2), the on-hand inventory for the second system will be no less than the on-hand inventory for the first system. Consequently, the *gold* fill rate for the second system must be at least as high as for the first system. □

4. Optimization algorithm

In this section we use the previous results to develop an algorithm to solve the two-demand-class fill rate optimization problem. Two important features of the algorithm are that it requires only one computation of the stationary probability distribution and it relies on simple line searches and recursive calculations for the remaining steps.

Let Δ^* be the smallest value of $\Delta = S - S_g$ that satisfies the required *silver* fill rate

$$\Delta^* = \min_{\Delta \in \{1,2,\ldots\}} \left\{ \Delta : \sum_{k=0}^{\Delta-1} \frac{(\lambda T)^k e^{-\lambda T}}{k!} \geq c_s \right\} \tag{7}$$

and let S^* be the smallest value of S that satisfies the required *gold* fill rate under the condition that $S_g = S - \Delta^*$:

$$S^* = \min_{S \in \{1,2,\ldots\}} \left\{ S : \beta_g \left(S, S - \Delta^* \right) \geq c_g \right\}. \tag{8}$$

Proposition 3. *The parameters* $(S, S_g) = (S^*, S^* - \Delta^*)$ *are optimal for the fill rate optimization problem.*

Proof. Suppose there exists another solution (S', S_g') that is feasible but for which $S' < S^*$. For this solution to be feasible with respect to the *silver* fill rate constraint we must have $S' - S_g' \geq \Delta^*$. Consider the solution $(S', S' - \Delta^*)$. By construction, this solution satisfies the *silver* fill rate constraint. Since the *gold* fill rate is nondecreasing in S_g for fixed S and since $S' - \Delta^* \geq S_g'$, we must have $\beta_g(S', S' - \Delta^*) \geq c_g$. However, this implies $S' \geq S^*$ by the definition of S^*, a contradiction. Consequently, there is no other feasible solution with a smaller value of S. □

Let $\overline{S}$ denote an upper bound on the optimal target inventory level. A natural choice is to set

$$\overline{S} = \min_{S \in \{1,2,\ldots\}} \left\{ S : \sum_{k=0}^{S-1} \frac{(\lambda T)^k e^{-\lambda T}}{k!} \geq c_g \right\}. \tag{9}$$

In this case, we can set $S_g = 0$ and then $\beta_g(\overline{S}, 0) = \beta_s(\overline{S}, 0) \geq c_g > c_s$ and so both service level constraints are satisfied.

We are now in a position to sketch the optimization algorithm (Table 1). We do not address implementation issues such as when to truncate infinite series.

The algorithm uses simple line searches to find Δ^*, $\overline{S}$, and the smallest value of S satisfying $\beta_g \geq c_g$. Each of these can be implemented using recursive forms. The validity of the calculation for β_g is a consequence of Corollary 4. The optimality of the solution is ensured by Proposition 3. The only challenging calculation is the determination of the stationary probabilities, $\pi_{(r,b_s)}(\Delta^*)$, for all $(r, b_s) \in Z_0 \times Z_0$. This is the focus of the remainder of this article. Note, however, that these probabilities are computed exactly once in the algorithm. This is an important consequence of the invariance results established above.

Table 1. Optimization algorithm for the two-demand-class problem.

```
1. Compute $\Delta^*$ and $\beta_s(\Delta^*)$ using (7);
2. Compute $\overline{S}$ using (9);
3. Compute $\pi_{(i,j)}(\Delta^*)$, for all $(i, j) \in Z_0 \times Z_0$, by some method, to be
   determined;
4. Compute $\pi_k(\overline{S}, \overline{S} - \Delta^*)$ for $k = 1, 2, \ldots \overline{S} - \Delta^*$ using (6);
5. Set $\beta_g = \beta_s(\Delta^*)$, $i = \Delta^*$;
6. If $\beta_g \geq c_g$, go to 7.
   Otherwise, repeat until $\beta_g \geq c_g$ :
   a) Set $i = i + 1$;
   b) Set $\beta_g = \beta_g + \pi_{\overline{S}-i+1}(\overline{S}, \overline{S} - \Delta^*)$;
7. Set $S^* = i$; Set $S_g^* = S^* - \Delta^*$;
8. Return $(S, S_g) = (S^*, S_g^*)$.
```

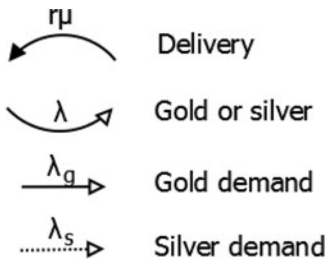

Figure 2. Flow rates.

5. Special case: Exponentially distributed lead times

In general, exact determination of the stationary probabilities, $\pi_{(r,b_s)}(\Delta^*)$, for all $(r, b_s) \in Z_0 \times Z_0$, for arbitrary lead time distributions is an unsolved problem. An excellent approximation for the case of constant lead times is provided by Fadiloglu and Bulut (2010b). In this article, we show that exact expressions for the stationary probabilities are available in the special case of exponentially distributed lead times and that these probabilities can be computed using a straightforward recursive scheme. In subsequent sections we present analytical and numerical evidence suggesting that this method provides fill rate estimates that are good approximations for systems with other lead time distributions, provided that the coefficient of variation of these distributions is not too high.

For the remainder of this section, we assume that lead times are exponentially distributed with rate $\mu = 1/T$. It is easy to see that in this case the resulting process $\{(R(t), B_s(t)) : t \geq 0\}$ is a CTMC. Let the state space be denoted by $D(\Delta) = \{(r, b_s) \in Z_0 \times Z_0 : b_s \leq r - \Delta\}$. Let the matrix A denote the infinitesimal generator of the process. The balance equations $\pi A = 0$ can be developed easily with reference to the state transition diagram (Fig. 1) and the key to transition rates (Fig. 2). Table 2 lists the resulting balance equations.

We are unable to obtain a closed-form solution to these balance equations. However, a rearrangement of terms in these equations suggests a computational scheme. Table 3 presents this suggested scheme.

Note that by Palm's Theorem, $\pi_{(0,0)} = e^{-\lambda/\mu}$, so the recursive scheme has a valid starting point. However, a study of the table reveals that there is no formula for determining $\pi_{(\Delta+k,k-1)}$ for $k = 1, 2, \ldots,$ in terms of quantities obtainable recursively through other formulas. We refer to these states $\{(\Delta + k, k - 1) : k = 1, 2, \ldots\}$ as *bridge states*: they lie at the interface between the states where deliveries are restricted to *gold* and states where deliveries are either unrestricted or are used to fill *silver* backorders. In Fig. 1, they are depicted as rounded rectangles. The challenge, then, is to determine the stationary probabilities of these bridge states. We present a solution to this challenge in the next sections and then integrate the results with Table 3 to present an algorithm for solving the CTMC balance equations.

5.1. *Determining the stationary probabilities of the bridge states*

5.1.1. *The bridge probabilities and the bridge theorem*

We first review basic CTMC results as applied to this system. Let θ_s denote the parameter of the exponential distribution of a

Table 2. Balance equations for the two-demand-class CTMC.

State classification	*Balance equation*
$r = 0, b_s = 0$	$\lambda\pi_{(0,0)} = \mu\pi_{(1,0)}$
$0 < r < \Delta, b_s = 0$	$(r\mu + \lambda)\pi_{(r,0)} = \lambda\pi_{(r-1,0)} + (r+1)\mu\pi_{(r+1,0)}$
$r = \Delta, b_s = 0$	$(r\mu + \lambda)\pi_{(r,0)} = \lambda\pi_{(r-1,0)} + (r+1)\mu\pi_{(r+1,0)} + (r+1)\mu\pi_{(r+1,1)}$
$r > \Delta, b_s = 0$	$(r\mu + \lambda)\pi_{(r,0)} = \lambda_g\pi_{(r-1,0)} + (r+1)\mu\pi_{(r+1,0)}$
$r > \Delta, b_s = r - \Delta$	$(r\mu + \lambda)\pi_{(r,b_s)} = \lambda_s\pi_{(r-1,b_s-1)} + (r+1)\mu\pi_{(r+1,b_s)} + (r+1)\mu\pi_{(r+1,b_s+1)}$
$r > \Delta, 0 < b_s < r - \Delta$	$(r\mu + \lambda)\pi_{(r,b_s)} = \lambda_s\pi_{(r-1,b_s-1)} + \lambda_g\pi_{(r-1,b_s)} + (r+1)\mu\pi_{(r+1,b_s)}$

sojourn time in state $s \in D(\Delta)$. If $s = (r, b_s)$, then, since r represents $R(t)$, the number of units in resupply, the rate θ_s is given by

$$\theta_s = r\mu + \lambda.$$

Let the matrix Q denote the probability transition matrix of the underlying Markov chain. If $s, s' \in D(\Delta)$ and $s \neq s'$, then the elements of Q can be deduced from Table 2 and the relation

$$Q_{s,s'} = A_{s,s'}/\theta_s.$$

The Markov chain on the set $D(\Delta)$ given by Q is irreducible. Consequently, since state $(0, 0)$ is recurrent (Palm's Theorem), all states in $D(\Delta)$ are recurrent. Let $\widetilde{\pi} = (\widetilde{\pi}_s)$ denote the stationary probability distribution of the imbedded Markov chain. Then $\widetilde{\pi}$ is an invariant measure

$$\widetilde{\pi}Q = \widetilde{\pi}.$$

Furthermore

$$\pi_s = \frac{\widetilde{\pi}_s/\theta_s}{\sum_{s'} \widetilde{\pi}_{s'}/\theta_{s'}}. \tag{10}$$

We are thus led to focus on $\widetilde{\pi}_s$, the stationary probabilities of the bridge states in the underlying Markov chain.

The fundamental relationship we exploit is the following result for discrete-time Markov chains:

Proposition 4 *(Resnick, 1992, p. 118). Let ξ_n be the system state at time n, $\tau_{s'}(1)$ be the first hitting time to state s', and $\widetilde{\pi}_s$ be the steady-state probability of being in state s. Also, let $s' \in S$ be recurrent, and define for $s \in S$:*

$$\begin{aligned} \nu_s &= E_{s'} \sum_{0 \le n \le \tau_{s'}(1)-1} 1_{\{\xi_n = s\}} \\ &= \sum_{n=0}^{\infty} P\left\{\xi_n = s, \tau_{s'}(1) > n | \xi_0 = s'\right\}. \end{aligned}$$

Then ν is an invariant measure. If state s' is positive recurrent so that $E_{s'}\tau_{s'}(1) < \infty$, then

$$\widetilde{\pi}_s = \frac{\nu_s}{E_{s'}\tau_{s'}(1)}.$$

The proposition states that the stationary probability of any state s in a discrete-time Markov chain is proportional to the expected number of times the state is visited in one cycle of consecutive visits to another recurrent state s'.

We relate this result to the Markov chain on $D(\Delta)$ as follows. Let s' be any recurrent state in $D(\Delta)$. Let ν_s denote the expected number of visits to state $s \in D(\Delta)$ between two consecutive visits to s'. Let τ' denote the first hitting time of state s' and let $E_s[\tau']$ denote the expected first hitting time of this state starting from state $s \in D(\Delta)$.

Corollary 5. *For all $k = 1, 2, \ldots,$ and any pair of states $s, s'' \in D(\Delta)$:*

$$\frac{\pi_s\theta_s}{\pi_{s''}\theta_{s''}} = \frac{\widetilde{\pi}_s}{\widetilde{\pi}_{s''}} = \frac{\nu_s}{\nu_{s''}}.$$

Proof. By Proposition 4 we have

$$\widetilde{\pi}_s = \frac{\upsilon_s}{E_{s'}[\tau']}, \text{ and } \widetilde{\pi}_{s''} = \frac{\upsilon_{s''}}{E_{s'}[\tau']}.$$

The expected cycle length cancels when computing the ratio. The extension to the CTMC stationary probabilities, π, follows from Equation (10). □

In what follows, the role of the recurrent state s' in the proposition is played by the deliver-to-*silver* recurrent states: $\{(\Delta + k, k) : k = 1, 2, \ldots\}$. For any value of k, to simplify notation, we reference nodes of interest relative to the node $s' = (\Delta + k, k)$ and suppress the dependence on k. Figure 3 illustrates. The deliver-to-*gold* states of interest are those with $k - 1$ *silver* backorders ($B_s = k - 1$). These are labeled $u_1, u_2, u_3, \ldots$ Thus, in general, $u_i, i = 1, 2, \ldots,$ refers to state $(\Delta + k - 1 + i, k - 1)$.

Table 3. Computational scheme for the two-demand-class CTMC.

State classification	*Suggested computation*
$r = 0, b_s = 0$	$\pi_{(1,0)} = \frac{\lambda}{\mu}\pi_{(0,0)}$
$0 < r < \Delta, b_s = 0$	$\pi_{(r+1,0)} = \frac{(r\mu+\lambda)}{(r+1)\mu}\pi_{(r,0)} - \frac{\lambda}{(r+1)\mu}\pi_{(r-1,0)}$
$r = \Delta, b_s = 0$	$\pi_{(r+1,1)} = \frac{(r\mu+\lambda)}{(r+1)\mu}\pi_{(r,0)} - \frac{\lambda}{(r+1)\mu}\pi_{(r-1,0)} - \pi_{(r+1,0)}$
$r > \Delta, b_s = 0$	$\pi_{(r+1,0)} = \frac{(r\mu+\lambda)}{(r+1)\mu}\pi_{(r,0)} - \frac{\lambda_g}{(r+1)\mu}\pi_{(r-1,0)}$
$r > \Delta, b_s = r - \Delta$	$\pi_{(r+1,b_s+1)} = \frac{(r\mu+\lambda)}{(r+1)\mu}\pi_{(r,b_s)} - \frac{\lambda_s}{(r+1)\mu}\pi_{(r-1,b_s-1)} - \pi_{(r+1,b_s)}$
$r > \Delta, 0 < b_s < r - \Delta$	$\pi_{(r+1,b_s)} = \frac{(r\mu+\lambda)}{(r+1)\mu}\pi_{(r,b_s)} - \frac{\lambda_s}{(r+1)\mu}\pi_{(r-1,b_s-1)} - \frac{\lambda_g}{(r+1)\mu}\pi_{(r-1,b_s)}$

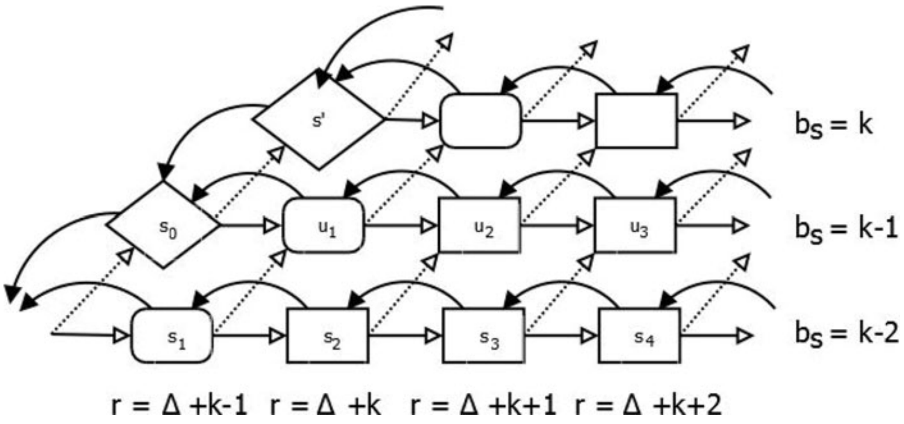

Figure 3. General bridge node notation.

We refer to these states as *gated states*, $G = \{u_1, u_2, u_3, \dots\}$. The bridge state of interest—that is, the state whose steady-state probability we seek to calculate—is the leading gated state: u_1. We also need to reference non-gated states that can reach states in G by means of a single demand arrival. We label these states by $s_0, s_1, s_2, s_3, \dots$ Thus, for $k \geq 2$, s_0 refers to state $(\Delta + k - 1, k - 1)$, which can reach G with a single *gold* arrival and, in general, for $i = 1, 2, \dots,$ state s_i refers to state $(\Delta + k - 2 + i, k - 2)$, which can reach G with a single *silver* arrival. We refer to these states as *feeder states*, $F = \{s_0, s_1, s_2, \dots\}$. The case $k = 1$ is special: In that case, there is only one feeder state: $F = \{s_0\} = \{(\Delta, 0)\}$. Let f denote the set of possible feeder indices: $f = \{0\}$ when $k = 1$ and $f = \{0, 1, 2, \dots\}$, otherwise.

State s' communicates with the gated states G only through state s_0. Between two consecutive visits to s', the Markov process may visit any of the gated states multiple times, up until a *silver* demand occurs in one of these states. Once a *silver* demand occurs in a gated state the process must visit state s' before any of these states can be revisited. Similarly, the gated states communicate with the feeder states only through state s_0. These are the facts we exploit to develop a solution.

Let τ' denote the first hitting time of state s' in the Markov chain on $D(\Delta)$. We continue to suppress the dependence on k. Let τ_0 denote the first hitting time of state s_0. Let p_i denote the probability that the process will reach state s_0 before it reaches state s', starting from gated state u_i:

$$p_i = P\left\{\tau_0 < \tau' | \xi_0 = u_i\right\}.$$

We refer to these probabilities p_i as *bridge probabilities*. A method for computing these probabilities is described in the next section. In this section, we establish their relationship to the *bridge state probability*, π_{u_1}. The major result of this section is the following.

Theorem 2 *(Bridge Theorem). For a given deliver-to-silver state $s' = (\Delta + k, k)$, gated states G, and feeder states F, defined relative to index k, the stationary probabilities, π, of the CTMC satisfy the following relationships: For $k = 1$:*

$$(\Delta + 1)\,\mu\pi_{u_1} = \lambda_g p_1 \pi_{s_0}.$$

For $k > 1$:

$$(\Delta + k)\,\mu\pi_{u_1} = \lambda_g p_1 \pi_{s_0} + \sum_{i=1}^{\infty} \lambda_s p_i \pi_{s_i}.$$

Proof. Provided in the online supplement. □

Interpreting the result, we imagine flows across a bridge from the bridge state u_1 to state s_0. In steady state, this flow occurs at rate $(\Delta + k)\mu\pi_{u_1}$, which is the rate of deliveries multiplied by the steady-state probability of the bridge state. Due to the unusual form of the Markov chain, the only flow that is possible in this direction must have first come from one of the feeder states, s_i for $i \in f$. The rate of flow in this direction (from s_i to the matching gated state) is $\lambda_g \pi_{s_0}$ if $i = 0$ and $\lambda_s \pi_{s_i}$ otherwise. However, only a portion of this flow returns across the bridge. The rest of the flow will return to the feeder states through s', the deliver-to-*silver* state, which is our reference state. The fraction that returns across the bridge is given by p_i, the probability that the underlying Markov chain will make that transition before visiting state s', given that it starts in the bridge state.

5.1.2. Computing the bridge probabilities

In this section, we develop simple recursive formulas for calculating the bridge probabilities p_i introduced in the previous section. In one sense, we continue the analysis of the previous section but in another sense, we establish general results for a simplified discrete-time Markov chain and apply these results to the imbedded Markov chain of the previous section.

The simplified Markov chain is depicted in Fig. 4. It has two absorbing states, s_0 and s', and an infinite number of transient states, $u_1, u_2, u_3, \dots$ Each transient state u_i can make a transition directly to absorbing state s' with probability γ_i and to state u_{i-1}, provided $i > 0$, with probability α_i. State u_1 can make a transition directly to absorbing state s_0 with probability α_1. Each transient state u_i can also make a transition directly to state u_{i+1} with probability β_i. No other transitions are possible. We assume that all transition probabilities are positive: $\alpha_i, \beta_i, \gamma_i > 0$. We further assume that $\{\alpha_i\}$ (respectively, $\{\beta_i\}$) is a monotonically increasing (respectively, decreasing) series with limit 1 (respectively, 0) as $i \to \infty$. The annotations for r and b_s in the figure can be ignored for now; they will be useful later in this section. Let τ' be the hitting time of absorbing state s' and let τ_0 denote the hitting time of absorbing state s_0. Let ξ_n denote the state of the Markov chain at step n. Let p_i denote the probability that the process reaches state s_0 before it reaches state s', given that it starts in state u_i:

$$p_i = P\left\{\tau_0 < \tau' | \xi_0 = u_i\right\}.$$

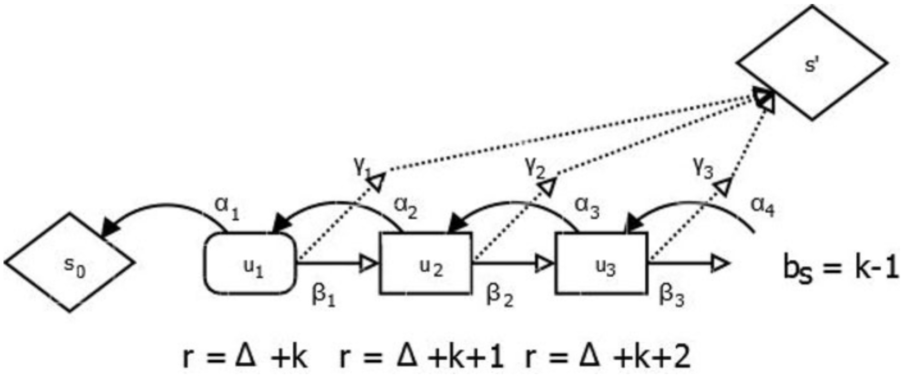

Figure 4. Simplified Markov chain: Bridge transitions.

Let $p_0 = 1$. It is easy to see the following relation must hold

$$p_i = \alpha_i p_{i-1} + \beta_i p_{i+1} \tag{11}$$

for $i = 1, 2, \ldots$ This gives rise to a general recursive formula for these probabilities:

$$p_{i+1} = \beta_i^{-1} \left(p_i - \alpha_i p_{i-1} \right). \tag{12}$$

The only difficulty is in finding the value of p_1 to initiate the calculation. We begin by relating each p_i to p_1. The first few formulas in the series are

$$\begin{aligned} p_2 &= \beta_1^{-1} p_1 - \alpha_1 \beta_1^{-1}, \\ p_3 &= \frac{1}{\beta_2}\left(\frac{1}{\beta_1} - \alpha_2\right) p_1 - \frac{1}{\beta_2}\frac{\alpha_1}{\beta_1}, \\ p_4 &= \frac{1}{\beta_3}\left(\frac{1}{\beta_2}\left(\frac{1}{\beta_1} - \alpha_2\right) - \alpha_3 \frac{1}{\beta_1}\right) p_1 \\ &\quad - \frac{1}{\beta_3}\left(\frac{1}{\beta_2}\frac{\alpha_1}{\beta_1} - \alpha_3 \frac{\alpha_1}{\beta_1}\right). \end{aligned}$$

The pattern should be apparent: let $a_i, b_i,\ i = 2, 3, \ldots,$ be defined recursively

$$\begin{aligned} a_{i+1} &= \beta_i^{-1} \left(a_i - \alpha_i a_{i-1} \right) \\ b_{i+1} &= \beta_i^{-1} \left(b_i - \alpha_i b_{i-1} \right) \end{aligned} \tag{13}$$

with initiating values

$$a_0 = -1\ , a_1 = 0,\ b_0 = 0,\ b_1 = 1. \tag{14}$$

The following proposition states the pattern.

Proposition 5. *For $i = 1, 2, \ldots$:*

$$p_i = b_i p_1 - a_i. \tag{15}$$

Proof. It is easily verified for $i = 1, 2, 3$, and 4. Assume it is true for i. Then for $i + 1$ we have

$$\begin{aligned} p_{i+1} &= \beta_i^{-1} \left(p_i - \alpha_i p_{i-1} \right) \\ &= \beta_i^{-1} \left(b_i p_1 - a_i - \alpha_i \left[b_{i-1} p_1 - a_{i-1} \right] \right) \\ &= \beta_i^{-1} \left(b_i - \alpha_i b_{i-1} \right) p_1 - \beta_i^{-1} \left(a_i - \alpha_i a_{i-1} \right) \\ &= b_{i+1} p_1 - a_{i+1}. \end{aligned}$$

The result, therefore, holds by induction. □

We refer to the quantities a_i and b_i as *bridge factors*. Observe that the same linear recursion generates each series: $\{p_i\}$, $\{a_i\}$, and $\{b_i\}$. The series differ only in their initial values: $\{p_0, p_1\}$, $\{a_0, a_1\}$, and $\{b_0, b_1\}$.

Proposition 6. *Under the condition that $\alpha_i \nearrow 1$ and $\beta_i \searrow 0$ as $i \to \infty$:*

$$\lim_{i \to \infty} \frac{p_i}{b_i} = 0.$$

Proof. Provided in the online supplement. □

Corollary 6. *Under the conditions of Proposition 6:*

$$p_1 = \lim_{i \to \infty} \frac{a_i}{b_i}. \tag{16}$$

To relate this result to the previous section, we consider Fig. 4 again but, this time, with attention to the annotations for r and b_s. Focusing on state u_1, we have

$$\begin{aligned} \alpha_1 &= \frac{(\Delta + k)\,\mu}{(\Delta + k)\,\mu + \lambda}, \\ \beta_1 &= \frac{\lambda_g}{(\Delta + k)\,\mu + \lambda}, \\ \gamma_1 &= \frac{\lambda_s}{(\Delta + k)\,\mu + \lambda}, \end{aligned} \tag{17}$$

and it follows that $\alpha_1 + \beta_1 + \gamma_1 = 1$. In general, for any state u_i, $i = 1, 2, \ldots,$

$$\begin{aligned} \alpha_i &= \frac{(\Delta + k + i - 1)\,\mu}{(\Delta + k + i - 1)\,\mu + \lambda}, \\ \beta_i &= \frac{\lambda_g}{(\Delta + k + i - 1)\,\mu + \lambda}, \\ \gamma_i &= \frac{\lambda_s}{(\Delta + k + i - 1)\,\mu + \lambda}. \end{aligned} \tag{18}$$

Observe that $\alpha_i \nearrow 1$ and $\beta_i \searrow 0$ as $i \to \infty$, so the conditions of Proposition 6 are satisfied. Therefore, Equation (16) shows that the bridge probability p_1 for state $u_1 = (r + \Delta, k - 1)$ can be computed as the limit of a ratio of bridge factors, which can be recursively computed using Equation (13) from initial values (14). Successive bridge probabilities for states with $b_s = k - 1$ can then be computed using Equation (12). Let $\varepsilon > 0$ denote a tolerance factor and let K denote an upper limit on the number of *silver* demand backorders to compute. Table 4 defines a function $BP(\Delta, k, \mu, \lambda_s, \lambda_g, K, \varepsilon)$ that returns a vector $p = (p_1, p_2, \ldots, p_K)$ by applying these recursions.

Table 4. Bridge probability function, $BP()$.

```
Step  Inputs: $\Delta, k, \mu, \lambda_s, \lambda_g, K, \varepsilon$
1.    Initialize: $\alpha_1, \beta_1$ using (17), $a_0, a_1, b_0, b_1$ using (14);
2.    Compute: $\alpha_2, \alpha_3, \beta_2, \beta_3$ using (18); Compute $a_2, a_3, b_2, b_3$ using (13);
3.    $n \leftarrow 3$;
4.    While ($\frac{a_n}{b_n} - \frac{a_{n-1}}{b_{n-1}} > \varepsilon$ and $n < K$) :
          Compute: $\alpha_{n+1}, \beta_{n+1}$ using (18); Compute $a_{n+1}, b_{n+1}$ using (13);
          $n \leftarrow n + 1$; End while loop;
5.    $p_1 \leftarrow \frac{a_n}{b_n}$; $n \leftarrow 1$;
6.    While ($n < K$) : Compute $p_{n+1}$ using (12);
          $n \leftarrow n + 1$; End while loop;
          Output: $(p_1, p_2, \ldots, p_K)$.
```

5.2. The Bridge Algorithm

In this section, we assemble the previous results into an algorithm for computing the stationary distribution, π, of the two-demand-class CTMC defined on the state space $D(\Delta)$, for given $\Delta = S - S_g$. Control parameters K, ε, and the bridge probability function, $BP(\Delta, k, \mu, \lambda_s, \lambda_g, K, \varepsilon)$, are as described in the previous section. Table 5 summarizes the algorithm, called here the *Bridge Algorithm*, for computing the stationary distribution of the two-demand-class CTMC. What is noteworthy about the algorithm is that it is composed entirely of recursive calculations. No matrix inversion is required.

Computational complexity: Examining Table 4, we observe that computing the bridge probabilities requires $O(K)$ computations for each combination of Δ and k considered (steps 4 and 6 iterate over n and $n < K$). Similarly, in Table 5, steps 1 to 3 require $O(\Delta)$ steps and we can take Δ to be less than K. Steps 4 to 7 are each at most $O(K)$. Step 8 is $O(K^2)$. Consequently, the Bridge Algorithm has computational complexity at most $O(K^2)$. We earlier showed that when optimizing stock levels, the Bridge Algorithm needs to be run for only one value of Δ, the value Δ^* that optimizes Equation (7). This calculation and the other line searches to find $\bar{S}$ and S^* are no more than $O(K)$, so it follows that optimization of stock levels using this approach is at most $O(K^2)$. This is in contrast with the embedded Markov chain approach of Fadiloglu and Bulut (2010b) that requires $O(D^6_{max})$ computations where $D_{max} \approx K$. Hence, this simple recursive algorithm is several orders of magnitude less complex than the embedded Markov chain approach.

6. The case of generally distributed lead times

The previous section detailed an exact analysis leading to a recursive algorithm to determine the stationary probabilities for the (R, B_s) process in the special case of exponentially distributed lead times. For the case of generally distributed lead times, two examples shall suffice to demonstrate the difficulty of exact analysis for this problem. Both examples assume that lead times, L, are constant.

Example 1: The sequence of arrivals matters

Given the state (R, B_s) at any point in time, the state of the system a lead time from now depends not only on the total numbers of *gold* and *silver* demand arrivals during the next L time periods but also on the sequence of these arrivals. To see this, suppose $S = 5$, $S_g = 2$, $R = 0$, and $B_s = 0$. Suppose also that during the next L time periods there are exactly three *silver* demands followed by two *gold* demands. The resulting state after L time periods will be $R = 5$ and $B_s = 0$. On the other hand, if the sequence of arrivals had been reversed (two *gold* demands followed by three *silver* demands), the resulting state would have been $R = 5$ and $B_s = 2$.

Example 2: The sequence of deliveries matters

Given the state (R, B_s) at any point in time, the state of the system a lead time from now depends not only on the total numbers of *gold* and *silver* demand arrivals during the next L time periods but also on the delivery times of the units currently in resupply. To see this, suppose $S = 5$ and $S_g = 2$ as before, but this time $R = 1$ and $B_s = 0$. Suppose also that during the next L

Table 5. The Bridge Algorithm for computing the stationary distribution of the two-demand-class CTMC.

```
Step
1.  $\pi_{(0,0)} \leftarrow e^{-\lambda/\mu}$;
2.  $\pi_{(1,0)} \leftarrow \frac{\lambda}{\mu}\pi_{(0,0)}$;
3.  For each $r$ in $1, 2, \ldots, \Delta - 1$ :
        $\pi_{(r+1,0)} \leftarrow \frac{(r\mu+\lambda)}{(r+1)\mu}\pi_{(r,0)} - \frac{\lambda}{(r+1)\mu}\pi_{(r-1,0)}$; Next $r$;
4.  $(p_1, p_2, \ldots, p_K) \leftarrow BP(\Delta, 1, \mu, \lambda_s, \lambda_g, K, \varepsilon)$;
5.  $\pi_{(\Delta+1,0)} \leftarrow \frac{\lambda_g}{(\Delta+1)\mu} p_1 \pi_{(\Delta,0)}$;
6.  $\pi_{(\Delta+1,1)} \leftarrow \frac{(\Delta\mu+\lambda)}{(\Delta+1)\mu}\pi_{(\Delta,0)} - \frac{\lambda}{(\Delta+1)\mu}\pi_{(\Delta-1,0)} - \pi_{(\Delta+1,0)}$;
7.  For each $r$ in $\Delta + 1, \Delta + 2, \ldots, \Delta + K$ :
        $\pi_{(r+1,0)} \leftarrow \frac{(r\mu+\lambda)}{(r+1)\mu}\pi_{(r,0)} - \frac{\lambda_g}{(r+1)\mu}\pi_{(r-1,0)}$; Next $r$;
8.  For each $k$ in $2, 3, \ldots, K$ :
    a) $(p_1, p_2, \ldots, p_K) \leftarrow BP(\Delta, k, \mu, \lambda_s, \lambda_g, K, \varepsilon)$;
    b) $\pi_{(\Delta+k,k-1)} \leftarrow \frac{\lambda_g}{(\Delta+k)\mu} p_1 \pi_{(\Delta+k-1,k-1)} + \frac{\lambda_s}{(\Delta+k)\mu} \sum_{i=1}^{K} p_i \pi_{(\Delta+k-2+i,k-2)}$;
    c) For each rin $\Delta + k, \Delta + k + 1, \ldots, \Delta + K$ :
        $\pi_{(r+1,k-1)} \leftarrow \frac{(r\mu+\lambda)}{(r+1)\mu}\pi_{(r,k-1)} - \frac{\lambda_s}{(r+1)\mu}\pi_{(r-1,k-2)} - \frac{\lambda_g}{(r+1)\mu}\pi_{(r-1,k-1)}$; Next $r$;
    d) $\pi_{(\Delta+k,k)} \leftarrow \frac{((\Delta+k-1)\mu+\lambda)}{(\Delta+k)\mu}\pi_{(\Delta+k-1,k-1)} - \frac{\lambda_s}{(\Delta+k)\mu}\pi_{(\Delta+k-2,k-2)} - \pi_{(\Delta+k,k-1)}$;
    e) Next $k$;
```

time periods there are three *silver* demands followed by two *gold* demands. The state of the system a lead time from now depends on when the unit in resupply is delivered. If it is received before any of the demands occur then the resulting state after L periods will be $R = 5$ and $B_s = 0$. On the other hand, if it is received after all of the demands occur, then the resulting state will be $R = 5$ and $B_s = 1$.

As suggested by these examples, there is no known exact solution for this rationing policy, except for the special case of exponentially distributed lead times, as detailed above. Several papers in the literature consider the constant lead time case and propose approximation methods to solve for the stationary distribution of (R, B_s). Dekker *et al.* (1998) base their approximation on the assumption that at an arbitrary point in time, t, the on-hand inventory at time $t - L$, a lead time ago, was equal to the order-up-to-level S. Deshpande *et al.* (2003) use a different reasoning for their approach and allow for order quantities greater than one. However, when applied to an $(S - 1, S)$ policy, the resulting formulas are identical to the approach of Dekker *et al.* (1998). We refer to this approach as the *single-cycle approach*. More recently, Fadiloglu and Bulut (2010b) consider the (R, B_s) process sampled at multiples of the fixed lead time as a Markov chain. In determining the transition probabilities of the Markov chain, they assume, for the purpose of approximation, that the delivery times are unaffected by the level of the *silver* backorders, B_s. They then develop a scheme for computing the stationary probabilities of the Markov chain using recursive calculations. Rapidly converging upper and lower bounds on the stationary probabilities are then computed using a sophisticated technique from the field of computational linear algebra. The accuracy of the resulting fill rate estimates compared with simulation runs is excellent. We refer to this as the *embedded Markov chain approach* and make use of a similar assumption in our approach.

As noted, Palm's Theorem implies that the stationary distribution of $R(t)$ for general lead time distributions is identical to that obtained when the lead time is exponentially distributed, with the same mean. A similar result obtains for the stationary distribution of $(R(t), B_s(t))$ if the following condition holds.

Definition 1. The Independence Condition is said to hold if, whenever the state of the system $(R, B_s) = (r, b_s)$ at an arbitrary point in time t, the probability of a unit delivery in the interval $(t, t + h)$ for an infinitesimally small $h > 0$ does not depend on the value of b_s.

Observe that this condition is very similar to that used in the embedded Markov chain approach for constant lead times. The *Independence Condition* holds in the case of exponentially disributed lead times due to the memoryless property of the exponential distribution. To show the importance of this condition, we offer the following theorem.

Theorem 3. *Assuming a general, positively valued lead time distribution having finite mean, T, with no probability mass at zero, then, if the Independence Condition is true, the steady-state distribution of (R, B_s) satisfies the same balance equations as a system with an exponential lead time distribution with the same mean.*

Proof. See Appendix A. □

Although the proof of Theorem 3 uses several concepts from the classic proof of Palm's Theorem, it employs a new approach to describe the limiting behavior of state transitions under the threshold rationing policy. The classic proof of Palm's Theorem does not consider the state transition probabilities and their limiting behaviors (Muckstadt and Sapra, 2010). The proof in Appendix A also identifies the critical point where the *Independence Condition* is required. This highlights the essential difficulty of exact analysis for this problem: dependence of the probability distribution of delivery times of units in resupply on B_s, the number of *silver* backorders. The theorem holds for the case of exponentially distributed lead times but, as suggested by the examples above, it is unlikely to hold in general. On the other hand, if the dependence is weak, the theorem suggests that the stationary distribution under exponentially distributed lead times might lead to a very good approximation for general lead time distributions. It is this conjecture that motivates the experimental studies of this article. We refer to our approach as the CTMC approach, which uses the results from exponential lead time distributions to approximate general lead time distribution situations. Furthermore, if the *Independence Condition* were true, then we would expect that CTMC approach would lead to exactly the same result as the embedded Markov chain approach in the case of constant lead times. Differences in numerical results must therefore trace either to numerical issues or to a failure of the *Independence Condition*.

As the *Independence Condition* is central to both the embedded Markov chain approach for constant lead times and the CTMC approach for general lead time distributions, we investigate it in some detail. It is well known that if we condition on the total number of Poisson arrivals in the interval $(t - L, t]$, say, r_0, then the unordered demand arrival times would be distributed as r_0 independent random variables, each uniformly distributed on $(t - L, t]$. Under the $(S - 1, S)$ policy, each demand arrival triggers a replenishment order that is to be received L periods later. Consequently, the replenishment order delivery times in $(t, t + L]$ would be distributed as r_0 uniform random variables on $(t, t + L]$. As Fadiloglu and Bulut (2010b) note, this property is no longer guaranteed to hold when one conditions also on the value of B_s, the *silver* backorders. Vicil and Jackson (2015) report on simulation experiments that demonstrate, indeed, that the distribution of replenishment order delivery times in $(t, t + L]$ is not uniformly distributed, when the value of $B_s(t)$ is known. Nevertheless, Fadiloglu and Bulut (2010b) report that the embedded Markov chain approach works quite well for constant lead times. The purpose of this article is to show how well the CTMC approach works.

7. Performance analysis using numerical simulation

For the remainder of this article, we concentrate on using numerical simulation to evaluate the quality of the CTMC approach for the two-demand-class model under a variety of lead time probability distributions. Unless otherwise stated, the duration of each simulation is 200 000 time periods and 10 independent simulations are performed for each parameter scenario. We use the observed *gold* fill rate, β_g, from each of the 10 simulations to construct confidence intervals around the performance metric. The confidence intervals are constructed based on the *t-distribution*, as the sample size is small. In each scenario, the *silver* fill rate, β_s, can be determined analytically.

There are currently three heuristics in the literature for constant lead times: Dekker *et al.* (1998), Deshpande *et al.* (2003), and Fadiloglu and Bulut (2010b). For zero setup costs, the model of Deshpande *et al.* (2003) is identical to the single-cycle approach of Dekker *et al.* (1998).

Our numerical study is divided into two major sections. First we compare the CTMC approach with the single-cycle approach of Dekker *et al.* (1998). We also summarize other simulation results evaluating the quality of the CTMC approach for constant lead times that are included in Vicil and Jackson (2015). Then, we compare the CTMC approach with the embedded Markov chain approach, which is the most recent heuristic.

7.1. A comparison of the CTMC approach with the single-cycle approach

To compare the CTMC approach with the single-cycle approach of Dekker *et al.* (1998), we construct a series of experiments for which λ_s and λ_g values vary and we assume order lead times are constant. The parameters are chosen in such a way that $\beta_s \geq 60\%$ and $\beta_g \geq 85\%$, levels, which are no less than what we would anticipate in practice and capture the situation where *gold* customers contract for substantially higher service levels than *silver* customers.

In Table 6, 30 different cases are presented in order to compare the accuracy of approximations with respect to various system parameters. From these results, we conclude that several factors affect the performance of the Dekker *et al.* heuristic. First, it is clear that as long as the expected lead time demand is sufficiently low, the Dekker *et al.* heuristic provides a good approximation. However, as soon as the expected lead time demand exceeds some threshold (e.g., 15 units) in these experiments, we start observing significant deviations from the simulated fill rate figures (cases (19) to (24) are good examples of this pattern). Second, it is also apparent that the accuracy of the Dekker *et al.* heuristic improves for high *gold* fill rates (i.e., 95%). Third, we also observe that in addition to *gold* fill rates, *silver* fill rates are also driving factors in the quality of the approximation of the Dekker *et al.* heuristic. For example, cases (11) and (12) both correspond to high *gold* fill rates, 98.84% and 97.23%, respectively. However, the former has a 82.17% *silver* fill rate, whereas the latter has a 65.32% fill rate. Although both cases correspond to high *gold* fill rates, the quality of the approximation in the Dekker *et al.* (1998) heuristic is lower for the lower *silver* fill rate (compare cases (17) and (18)).

On the other hand, it can be concluded that the *Independence Condition* holds well for these system parameters and the CTMC approach works well for all cases. In fact, the CTMC approach provides a very high-quality approximation across all the scenarios considered. The predicted *gold* fill rate differs from the center of the confidence interval by no more than 0.5%. However, it is apparent that the CTMC approach consistently but slightly overestimates the simulated *gold* fill rate, in contrast with the single-cycle approach, which underestimates the *gold* fill rate, often by a substantial amount.

7.1.1. Summary of the additional numerical studies

In Vicil and Jackson (2015), we explore a wide range of system parameters and, where possible, compare the results of

Table 6. Comparison of the CTMC approximation to the single-cycle heuristic.

Case	S	S_g	$\lambda_g/(\lambda_s+\lambda_g)$	λL	β_s (%)	$\beta_{g\,(Simulation)}$ (%)	$\beta_{g\,(CTMC)}$ (%)	$\beta_{g\,(single-cycle)}$ (%)
(1)	5	2	1/2	1.5	80.88	99.53 ± 0.02	99.57	97.40
(2)	7	2	1/2	3	81.52	99.17 ± 0.03	99.23	98.78
(3)	10	2	1/2	6	74.40	97.90 ± 0.04	98.08	96.31
(4)	19	2	1/2	15	66.41	95.38 ± 0.07	95.80	89.76
(5)	29	3	1/2	24	63.19	97.78 ± 0.05	98.01	91.46
(6)	37	4	1/2	30	68.34	99.26 ± 0.03	99.35	95.50
(7)	5	1	1/3	2.25	80.94	97.41 ± 0.04	97.51	96.64
(8)	7	1	1/3	4.50	70.29	94.32 ± 0.08	94.63	91.62
(9)	13	2	1/3	9	70.60	98.60 ± 0.03	98.75	96.43
(10)	27	1	1/3	22.5	74.33	93.42 ± 0.13	93.59	87.29
(11)	44	2	1/3	36	82.17	98.84 ± 0.04	98.85	95.33
(12)	50	2	1/3	45	65.32	97.23 ± 0.08	97.37	87.20
(13)	6	2	2/3	2.25	80.94	98.79 ± 0.02	98.86	98.50
(14)	8	2	2/3	4.5	70.29	96.16 ± 0.06	96.44	94.65
(15)	13	2	2/3	9	70.60	94.50 ± 0.09	94.83	91.49
(16)	27	1	2/3	22.5	74.33	86.84 ± 0.21	87.10	82.63
(17)	44	2	2/3	36	82.17	95.36 ± 0.14	95.34	91.46
(18)	50	2	2/3	45	65.32	89.01 ± 0.16	89.37	79.25
(19)	8	2	1/5	3.75	82.29	99.86 ± 0.01	99.87	99.69
(20)	11	2	1/5	7.5	66.20	99.47 ± 0.03	99.51	98.29
(21)	19	2	1/5	15	66.41	99.26 ± 0.05	99.34	96.83
(22)	42	2	1/5	37.5	63.71	98.97 ± 0.05	99.04	92.86
(23)	65	2	1/5	60	63.38	98.89 ± 0.05	98.93	90.32
(24)	81	2	1/5	75	66.28	98.94 ± 0.05	98.99	90.28
(25)	9	3	4/5	3.75	82.29	99.22 ± 0.03	99.30	99.05
(26)	12	2	4/5	7.5	77.64	94.98 ± 0.09	95.14	93.65
(27)	20	2	4/5	15	74.89	92.09 ± 0.11	92.31	89.40
(28)	43	2	4/5	37.5	69.52	87.07 ± 0.18	87.26	81.51
(29)	66	3	4/5	60	63.38	88.02 ± 0.31	88.43	79.06
(30)	82	3	4/5	75	66.28	88.59 ± 0.31	88.93	79.90

the CTMC approach with competing heuristics. We briefly summarize some of the results here.

The impact of total workload changes: Our aim in this part is to analyze the effect of total workload on the performance of approximations, while keeping all other system parameters fixed. We fix $\lambda_g/(\lambda_s+\lambda_g)=0.5,\ S=5$, and $S_g=2$ and vary the total workload, λL. The results are presented in Appendix B, Table B1. As the total workload increases, we observe that the absolute error of the approximation increases up to some point and then starts to decrease. It is also interesting to observe that as workload increases, with the rest of the system parameters kept fixed, the *gold* customer fill rate is not significantly affected after $\lambda L=15$ in these experiments. This might be counter-intuitive. One explanation for this phenomenon is that for $\lambda L \geq 15$, *silver* customers do not get any service at all despite the existence of *silver* customer demands. On the other hand, all of the replenishment orders due to *silver* customer demands are used to satisfy *gold* customers. Hence, this offsets the negative effect of an increase in workload on *gold* customer fill rate. However, the degree of such an offset would vary depending on the ratio $\lambda_g/(\lambda_s+\lambda_g)$. We investigate the impact of that ratio next.

Varying the demand rate for gold service: Our aim in this series of experiments is to analyze the performance of approximations under a fixed workload while varying the ratio $\lambda_g/(\lambda_s+\lambda_g)$. We set $S=8,\ S_g=2$, and $\lambda L=5$. The results are presented in Appendix B, Table B2. Based on the numerical results, we see that the CTMC approximation provides a higher-quality approximation in all cases than the single-cycle heuristic. We also observe that as the ratio $\lambda_g/(\lambda_s+\lambda_g)$ increases up to 2/3, the performance of both the CTMC approximation and the single-cycle heuristic are negatively affected. As the ratio increases beyond this point, the quality of both approximations increases. One explanation for this behavior is that as the ratio approaches zero, the system behaves more like a single-customer system with *silver* demands, whereas as the ratio approaches one, the system moves toward a single-customer system with *gold* demands. Hence, the effect of rationing decreases and therefore both approximations provide higher-quality results at the extremes.

7.2. A comparison of the CTMC approach with the embedded Markov chain approach

In this part of the the study, we compare the performance of the CTMC approach with respect to the embedded Markov chain approach under lognormal, geometric, and Erlang lead time distributions, as well as under constant lead times. As suggested by Theorem 3, the form of the lead time distribution will have no effect on the stationary distribution and, hence, no effect on customer service levels provided the *Independence Condition* holds. The extensive experimentation reported in Vicil and Jackson (2015) reveals that, in fact, the achieved *gold* service level is relatively insensitive to the form of the lead time distribution. Therefore, another important contribution of this article is to note that any good approximation algorithm for the constant lead time case or any other lead time distribution can be used to approximate general lead time distributions. Hence, our analytical and experimental results suggest that both the embedded Markov chain and our CTMC approach should work well across a variety of lead time distributions.

On the other hand, as shown earlier, the CTMC approach is several orders of magnitude less complex than the embedded Markov chain approach.

In the following series of experiments, we refer to the same numerical examples considered in Fadiloglu and Bulut (2010b). The expected lead time is the same in each example. Note that the embedded Markov chain approach assumes that the *Independence Condition* holds for constant lead times, whereas the CTMC approach assumes the same condition holds for general lead time distributions.

In Table 7, we report simulation studies that consider a constant lead time and Erlang-distributed lead times with shape parameters 16, 4, and 2. (For the Erlang distribution, the coefficient of variation $CV=\sqrt{1/k}$, where k is the shape parameter. Hence, the CV of the Erlang distribution varies between zero and one for $k\geq 1$.) For the constant lead time case, we observe that the *Independence Condition* appears to hold, as long as the *silver* fill rate is not too low. In particular, for $\beta_s \geq 90\%$, we observe that the absolute error for the estimated *gold* fill rate under the CTMC approach is zero, whereas for $\beta_s \geq 42.32\%$, the absolute error is still less than 1.15%. On the other hand, for β_s as low as 6.20%, the absolute error increases up to 3.22%. However, for the cases considered, the embedded Markov chain approach estimates the *gold* fill rate extremely well, even when the *silver* fill rate is small.

On the other hand, for Erlang-distributed lead times, it is interesting to observe that as the CV increases, the quality of the CTMC approach increases, whereas the quality of embedded Markov chain approach decreases. For the cases with $CV=0.707$ and $\beta_s=42.32\%$, the maximum absolute error for the CTMC approach drops to 0.63% and for β_s it reaches as low

Table 7. Comparison of CTMC approximation versus the results in Fadiloglu and Bulut (2010b), $S=4$, $S_g=1$.

				Erlang				
λT	λ_g/λ	β_s (%)	$\beta_{g\,(constant)}$ (%)	$\beta_{g(CV=0.25)}$ (%)	$\beta_{g(CV=0.50)}$ (%)	$\beta_{g(CV=0.707)}$ (%)	$\beta_{g\,(CTMC)}$ (%)	$\beta_{g\,(Fadiloglu\ and\ Bulut)}$ (%)
1	1/4	91.97	99.54 ± 0.01	99.54 ± 0.02	99.53 ± 0.02	99.51 ± 0.03	99.54	99.5
	1/2		99.07 ± 0.02	99.06 ± 0.02	99.04 ± 0.03	99.07 ± 0.04	99.07	99.1
	3/4		98.59 ± 0.04	98.58 ± 0.02	98.59 ± 0.02	98.57 ± 0.02	98.59	98.6
3	1/4	42.32	91.13 ± 0.10	91.15 ± 0.13	91.30 ± 0.12	91.48 ± 0.09	91.87	91.2
	1/2		82.38 ± 0.13	82.38 ± 0.06	82.69 ± 0.18	82.84 ± 0.12	83.47	82.4
	3/4		73.67 ± 0.10	73.62 ± 0.11	73.74 ± 0.14	74.04 ± 0.14	74.59	73.7
6	1/4	6.20	78.89 ± 0.09	78.93 ± 0.17	79.19 ± 0.12	79.74 ± 0.21	80.90	78.7
	1/2		58.05 ± 0.06	57.98 ± 0.08	58.62 ± 0.16	59.43 ± 0.15	61.27	58.1
	3/4		37.60 ± 0.11	37.70 ± 0.09	38.00 ± 0.14	38.89 ± 0.19	40.49	38.1

Table 8. Comparison of CTMC approximation versus Fadiloglu and Bulut (2010b) approximation, $S = 4$, $S_g = 1$.

			Lognormal					
λT	λ_g/λ	β_s (%)	$\beta_{g(CV=3.00)}$ (%)	$\beta_{g(CV=2.00)}$ (%)	$\beta_{g(CV=1.50)}$ (%)	$\beta_{g\,(geometric)}$ (%)	$\beta_{g\,(CTMC)}$ (%)	$\beta_{g\,(Fadiloglu\ and\ Bulut)}$ (%)
1	1/4	91.97	99.58 ± 0.03	99.55 ± 0.02	99.55 ± 0.02	99.52 ± 0.02	99.54	99.5
	1/2		99.12 ± 0.02	99.07 ± 0.03	99.06 ± 0.02	99.06 ± 0.03	99.07	99.1
	3/4		98.66 ± 0.03	98.62 ± 0.04	98.57 ± 0.02	98.58 ± 0.03	98.59	98.6
3	1/4	42.32	92.87 ± 0.06	92.30 ± 0.09	91.79 ± 0.13	91.39 ± 0.12	91.87	91.2
	1/2		85.06 ± 0.10	84.08 ± 0.14	83.23 ± 0.16	82.82 ± 0.13	83.47	82.4
	3/4		76.34 ± 0.20	75.29 ± 0.21	74.50 ± 0.11	74.11 ± 0.22	74.59	73.7
6	1/4	6.20	83.25 ± 0.13	81.75 ± 0.12	80.40 ± 0.12	80.02 ± 0.11	80.90	78.7
	1/2		65.07 ± 0.18	62.73 ± 0.16	60.62 ± 0.16	60.08 ± 0.15	61.27	58.1
	3/4		44.72 ± 0.30	42.21 ± 0.21	40.08 ± 0.11	39.49 ± 0.22	40.49	38.1

as 6.20% and the maximum absolute error drops to 1.84%. On the other hand, the maximum absolute error for the embedded Markov chain approach can be as high as 1.33%.

These results drive our motivation to study other cases to observe how the quality of approximation changes as the *CV* increases. To do so, we use the same setting as before but this time with lognormal and geometric lead time distributions. We study the cases with $CV = 1.50$, 2.00, and 3.00. The results are presented in Table 8. For all of the cases with lognormal lead time distributions, the CTMC approach either matches or dominates the embedded Markov chain approach. Furthermore, we also see that the embedded Markov chain approach underestimates the simulated *gold* fill rates. However, this situation varies for the CTMC approach depending on the CV values and other system parameters. For $CV \leq 2.00$ and β_s as low as 42.32%, the maximum absolute error for the CTMC approach is 0.7%, wherever the error can be as high as 1.6% for the embedded Markov chain approach. For the geometric lead time distribution cases, for $\beta_s = 91.97\%$, both methods provide excellent approximations. For the cases with $\beta_s = 42.32\%$, the two methods are in a tie in terms of approximation performance. On the other hand, for β_s as low as 6.20%, the CTMC approach provides better approximations than the embedded Markov chain approach. It is surprising that as the CV increases, the CTMC approach outperforms the embedded Markov chain approach given that the *Independence Condition* is the basis for both approaches. It is noteworthy that the differences are most pronounced in scenarios where the *Independence Condition* is least likely to hold.

8. Conclusions

In this article, we consider a model where there are two priority demand classes exhibiting mutually independent, stationary, Poisson demand processes with non-zero order lead times that are independent and identically distributed. We assume an (S-1, S) ordering policy and a threshold-level-based allocation and backorder clearing policy.

Currently, there is no exact solution for this rationing policy in the literature, except for what we have provided in the special case of exponentially distributed lead times. We pinpoint the difficulty for exact steady-state analysis and then show why a CTMC approach might provide a good approximation to the calculation of stationary probabilities under general lead time distributions. We also present a procedure to solve the CTMC by exploiting the special structure of the transition matrix in a novel way. This results in an efficient recursive procedure. For the generally distributed lead times setting, we develop an efficient algorithm in which the optimal parameters can be found by computing stationary probabilities only once. The algorithm relies on a simple line search. We are the first to provide an optimization scheme for this model subject to demand class specific fill rate constraints.

We compare our results with the the single-cycle approach of Dekker *et al.* (1998). We report that for constant lead times, the resulting solution outperforms their approach. Based on the simulation studies, for realistic scenarios that we expect to see in real-life situations (such as $\beta_s \geq 60\%$ and $\beta_g \geq 85\%$), the absolute error for the CTMC approximation is less than 0.5%.

We also compare the performance of the CTMC approximation with respect to the most recent approximation provided by Fadiloglu and Bulut (2010b). Although both approaches share an *Independence Condition*, their method is customized to the constant lead time case. For the numerical examples considered with constant lead times, although our method provides a reasonably good approximation, their method is clearly more accurate. We also show that as the form of the lead time distribution changes, the *gold* service levels do not vary by much. Therefore, another important contribution of our article is that, as Theorem 3 establishes the theoretical foundation, any valid approximation algorithm for constant lead time case or any other lead time distributions might be used to approximate general lead time distributions. Comparing the performance of the CTMC approach with the embedded Markov chain approach, for lognormal and Erlang lead time distributions, we demonstrate that as the *CV* of the lead time increases, the quality of the embedded Markov chain approach diminishes, whereas the quality of the CTMC approach increases.

For practical applications, it is important to provide simple and accurate approximations and to investigate their behavior under different system settings. Therefore, our proposed method, which requires only knowledge on the mean value of the lead time distributions, performs well over a wide range of parameter settings for general lead time distributions, provided that the *silver* fill rate is maintained in excess of 60%. Also, our simple recursive algorithm is several orders of magnitude less computationally complex than the embedded Markov chain approach. Hence, it may be worth exploring this approach with different rationing models under general lead time distributions.

It is straightforward, but tedious, to extend the model to consider three demand classes: adding a *platinum* demand class to the previously described *gold* and *silver* demand classes. Let λ_p denote the arrival rate for *platinum* customers. *Platinum* customers are assumed to require a higher level of service than both *gold* and *silver* customers. We extend the rationing policy to include a threshold $S_p \leq S_g$ at and below which only *platinum* customers are served. The state space must be expanded to include *gold* backorders: (R, B_s, B_g) but, in the case of exponentially distributed lead times, it is not difficult to derive the balance equations that can be solved for the steady-state probabilities. The balance equations and the numerical results for the three-demand-class model are included in Vicil and Jackson (2015). For the cases considered there, we observe a similar pattern as in the two priority demand classes setting: the CTMC approximation overestimates the true *gold* and *platinum* fill rates. However, for sufficiently high *silver* fill rates (i.e., $\beta_s \geq 50\%$), the absolute errors for CTMC approximation with respect to (mean) simulated *gold* and *platinum* fill rates are less than 0.75%. We also conclude from those experiments that the two-step rationing provides even larger protection from being backordered for the highest-priority demand class than the single-step rationing policy.

As a suggestion for future research, since the CTMC approximation provides quite satisfactory results under a static rationing policy for general lead time distributions, it may be interesting to explore the performance of this approach under dynamic replenishment policies.

Notes on contributors

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

Peter Jackson is a Professor in the School of Operations Research and Industrial Engineering (ORIE), Cornell University. He received his Ph.D. in Operations Research from Stanford University in 1980. He has served at Cornell since 1980. He is the Director of Graduate Studies for, and a former Director of, the Systems Engineering Program within the College of Engineering. He also serves as the Director of Undergraduate Studies for ORIE. His research interests include planning and scheduling for integrated production, transportation and inventory management systems, supply chain management, and business modeling and data analysis. He has consulted with several companies in these areas, including Agco, PTC-Servigistics, General Motors, Cleveland Clinic, Xelus, Clopay Building Products, General Electric, Aeroquip, and Quaker Oats. He was the recipient of a General Motors Research and Development Innovation award in 2011 for a business process to optimize retail inventories. He is also active in educational curriculum development for operations research and systems engineering. He is the recipient of several awards for curriculum innovation in addition to numerous student-voted awards for teaching excellence. He is the author of an introductory textbook to systems engineering, *Getting Design Right: A Systems Approach* (CRC Press, 2009).

References

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

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.

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.

Deshpande, V. and Cohen, M.A. (2005) A nested threshold inventory rationing policy for multiple demand classes in inventory systems with replenishment. Working paper, Krannert School of Management, Purdue University, West Lafayette, IN.

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

Ding, Q., Kouvelis, P. and Milner, J.M. (2006) Dynamic pricing through discounts for optimizing multiple-class demand fulfillment. *Operations Research*, **54**(1)169–183.

Ding, Q., Kouvelis, P. and Milner, J.M. (2007) Dynamic pricing for multiple class deterministic demand fulfillment. *IIE Transactions*, **39**(11) 997–1013.

Fadiloglu, M.M. and Bulut, O. (2010a) A dynamic rationing policy for continuous-review inventory Systems. *European Journal of Operational Research*, **202**, 675–685.

Fadiloglu, M.M. and Bulut, O. (2010b) An embedded Markov chain approach to stock rationing. *Operations Research Letters*, **38**(6), 510–515.

Kleijn, M.J. and Dekker, R. (1999) An overview of inventory systems with several demand classes. *Lecture Notes in Economics and Mathematical Systems*, **480**, 253–265.

Kocaga, Y.L. and Sen, A. (2007) Spare parts inventory management with demand lead times and rationing. *IIE Transactions*, **39**(9), 879–898.

Muckstadt, J.A. and Sapra, A. (2010) *Principles of Inventory Management: When You Are Down to Four, Order More*, Springer, New York, NY.

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

Resnick, S.I. (1992) *Adventures in Stochastic Processes*, Birkhäuser, Boston, MA.

Teunter, R.H. and Haneveld, W.K.K. (2008) Dynamic inventory rationing strategies for inventory systems with two demand classes, Poisson demand and backordering. *European Journal of Operational Research*, **190**(1), 156–178.

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

Vicil, O. and Jackson, P. (2015) Numerical validation of fill rate estimation methods for two and three demand-class rationing policies with one-for-one replenishment and general lead time distributions. Technical Report No. 1484, School of ORIE, Cornell University, Ithaca, NY.

Appendices

Appendix A

The proof of Theorem 3 is modeled after the proof of Palm's Theorem. We focus initially on the aggregate number of units in resupply, whether they are from *gold* or *silver* demands. We analyze the limiting behavior of transition probabilities for this process during an infinitesimal time interval h. This is accomplished in two lemmas, the first of which provides limits related to $\tilde{p}(h)$, the probability that a unit in resupply will not be delivered in the next h time units, and the second of which applies these limits to provide formulas for derivatives of the transition probabilities. We are then in a position to prove the main result. For a general system state $(R, B_s) = (i, j)$ with $i > S - S_g$ and $j \geq 1$, we provide the limiting behavior of the transition probabilities during an infinitesimal time interval h by conditioning on the state of the system at time t. Under the *Independence Condition*, the resulting system of equations is identical to the balance equations in the CTMC. Consequently, we show that under the *Independence Condition*, the steady-state distribution of $\left(R, B_s\right)$ satisfies the balance equations of a system with an exponential lead time distribution having the same mean.

Our first task is to establish the stochastic behavior of a general unit in resupply. Let $G(\cdot)$ denote the probability distribution of the lead time. We assume that lead times are positively valued random variables, thus $G(\cdot)$ has no atom at zero. Any probability distribution can be written as the sum of a discrete distribution and an absolutely continuous distribution. That is, there exist constants $\{(w_k, y_k) : k = 1, 2, \ldots\}$ and a non-negative, continuous function $g()$ such that

$$G(t) = \sum_k w_k 1_{\{t \geq y_k\}} + \int_0^t g(u)du.$$

Let $G_c(t) = \int_0^t g(u)du$ denote the absolutely continuous portion of the distribution. Initially, we assume the existence of a constant, $\bar{y}$, that bounds the discrete portion of the distribution $y_k < \bar{y}$ for all k. Later, dependence on this assumption will be omitted.

Let p be the common probability that any demand that arrives during $[0, t)$ remains in the resupply system at time t. Recall that for a Poisson arrival, given that an arrival occurs during $[0, t)$, the time of arrival is uniformly distributed over this interval. Consequently, conditioning on the time of arrival, we have

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

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

$$\begin{aligned}
&\tilde{p}(h)\\
&= \frac{P\left\{\text{a unit arrives in } [0, t) \text{ and is in resupply at time } t + h\right\}}{P\left\{\text{a unit arrives in } [0, t) \text{ and is in resupply at time } t\right\}}\\
&= \frac{\int_0^t [1 - G(t + h - s)] \frac{1}{t} ds}{\int_0^t [1 - G(t - s)] \frac{1}{t} ds}\\
&= \frac{\int_0^t [1 - G(t + h - s)] ds}{\int_0^t [1 - G(t - s)] ds}.
\end{aligned}$$

Trivially, $\lim_{h \to 0} \tilde{p}(h) = 1$.

Lemma A1. *For* $t > \bar{y}$:

$$\lim_{h \to 0} \frac{1}{h} [\tilde{p}(h)^i - \tilde{p}(h)^{i+1}] = \frac{G(t)}{\int_0^t [1 - G(u)] du}$$

and

$$\lim_{h \to 0} \frac{1 - e^{-\lambda h} \tilde{p}(h)^i}{h} = \lambda + i \frac{G(t)}{\int_0^t [1 - G(u)] du}.$$

Proof. Assume initially that there exists a constant $\underline{y}$ such that $0 < \underline{y} < y_k < \bar{y}$ for all k and that we consider only values of h and t such that $h < \underline{y}$ and $t > \bar{y}$. Under this assumption:

$$\begin{aligned}
&\int_0^t [1 - G(t + h - s)] ds\\
&= \int_0^t [1 - G_c(t + h - s)] ds - \sum_k w_k \int_0^t 1_{\{t+h-s \geq y_k\}} ds\\
&= \int_0^t [1 - G_c(t + h - s)] ds - \sum_k w_k (t + h - y_k).
\end{aligned}$$

and hence $\tilde{p}(h)$ is differentiable with

$$\begin{aligned}
\tilde{p}'(h) &= \frac{-\int_0^t g(t + h - s) ds - \sum_k w_k}{\int_0^t [1 - G(t - s)] ds}\\
&= \frac{-\int_0^t g(t + h - s) ds - \sum_k w_k 1_{\{t+h \geq y_k\}}}{\int_0^t [1 - G(t - s)] ds}
\end{aligned}$$

since $1_{\{t \geq y_k\}} = 1$ for all k. After a change of variable, this leads to

$$\begin{aligned}
\tilde{p}'(h) &= \frac{-\int_h^{t+h} g(u) du - \sum_k w_k 1_{\{t+h \geq y_k\}}}{\int_0^t [1 - G(u)] du}\\
&= \frac{-(G_c(t + h) - G_c(h)) - \sum_k w_k 1_{\{t+h \geq y_k\}}}{\int_0^t [1 - G(u)] du}\\
&= -\frac{G(t + h) - G_c(h)}{\int_0^t [1 - G(u)] du}.
\end{aligned}$$

Observe that t is not a point of discontinuity of $G()$. Consequently,

$$\lim_{h \to 0} \tilde{p}'(h) = -\frac{G(t)}{\int_0^t [1 - G(u)] du}.$$

This result enables us to apply L'Hopital's rule:

$$\begin{aligned}
\lim_{h \to 0} \frac{1}{h} [\tilde{p}(h)^i - \tilde{p}(h)^{i+1}] &= \lim_{h \to 0} \left[i \tilde{p}(h)^{i-1} - (i + 1) \tilde{p}(h)^i\right] \tilde{p}'(h)\\
&= \frac{G(t)}{\int_0^t [1 - G(u)] du}.
\end{aligned}$$

Also

$$\begin{aligned}
\lim_{h \to 0} \frac{1 - e^{-\lambda h} \tilde{p}(h)^i}{h} &= \lim_{h \to 0} \left[\lambda e^{-\lambda h} \tilde{p}(h)^i - i e^{-\lambda h} \tilde{p}(h)^{i-1} \tilde{p}'(h)\right]\\
&= \lambda + i \frac{G(t)}{\int_0^t [1 - G(u)] du}.
\end{aligned}$$

Since the result is true for all $\underline{y} > 0$, it will hold in the limit as $\underline{y} \to 0$. □

Suppose there are i replenishment orders outstanding. Let $\mathrm{u}_{[k]}$ denote the age of the kth oldest replenishment order and let $\mathrm{u} = (\mathrm{u}_{[1]}, \mathrm{u}_{[2]}, \ldots, \mathrm{u}_{[i]})$ denote the age-of-pipeline vector.

Denote the state of the system at time t by $\xi_t = (i, j, \mathrm{u})$ where i is the number of replenishment orders outstanding, j is the number of *silver* backorders, and u is the age-of-pipeline vector. We assume $\xi_0 = (0, 0, \emptyset)$. That is, the process begins with nothing on order.

It is easily seen that, for general lead time distributions, the process $\xi = \{\xi_t,\ t \geq 0\}$ is a Markov process.

With an abuse of notation, we write $\xi_t = (i, j)$ to denote all possible states with i replenishment orders outstanding and j *silver* backorders. Similarly, we write $\xi_t = i$ to denote all possible states with i replenishment orders outstanding.

We define

$$\begin{aligned}\bar{Q}_{i,j}(t,t+h) \equiv P\big[&\text{number of units in resupply at time } t+h \text{ is}\\ &j \mid \text{number of units in resupply at time } t \text{ is}\\ &i \text{ and number of units in}\\ &\text{resupply at time zero is } 0\big]\\ = P\big[&\xi_{t+h} = j \mid \xi_t = i, \xi_0 = 0\big].\end{aligned}$$

In the case where $i < j$, we further qualify this quantity by $\zeta \in \{s, g\}$, the last type of demand (*silver* or *gold*) to arrive:

$$\begin{aligned}\bar{Q}^{(\zeta)}_{i,j}(t,t+h) \equiv P\big[&\text{number of units in resupply at time } t+h \text{ is } j,\\ &\text{the most recent demand in } [0, t+h) \text{ is of}\\ &\text{type } \zeta, \text{ and all units in resupply at time } t \text{ are}\\ &\text{still in resupply at time } t+h \mid \text{the number}\\ &\text{of units in resupply at time } t \text{ is } i, \text{ and the}\\ &\text{number of units in resupply at time zero is } 0\big].\end{aligned}$$

We have the following expressions for the derivatives of these quantities.

Lemma A2. *For* $t > \bar{y}$,

$$\begin{aligned}\lim_{h\to 0}\frac{1-\bar{Q}_{i,i}(t,t+h)}{h} &= \lambda + i\frac{G(t)}{\int_0^t[1-G(u)]du};\\ \lim_{h\to 0}\frac{\bar{Q}_{i+1,i}(t,t+h)}{h} &= (i+1)\,\frac{G(t)}{\int_0^t[1-G(u)]du};\\ \lim_{h\to 0}\frac{\bar{Q}^{(s)}_{i-1,i}(t,t+h)}{h} &= \lambda_s; \textit{ and}\\ \lim_{h\to 0}\frac{\bar{Q}^{(g)}_{i-1,i}(t,t+h)}{h} &= \lambda_g.\end{aligned}$$

Proof. The state of the system changes if an arrival of either type of demand occurs or a unit is received from the resupply system. Since demand is a Poisson process and orders from suppliers are triggered whenever a demand occurs, during an infinitesimal time interval h, the probability of more than one event to occur is $o(h)$ due to the Poisson nature of the process. In addition, keep in mind that demands are independent and identically distributed so the resupply process is independent of the subsequent demand process. We next define a set of probabilities that will be used later in the proof.

$$\begin{aligned}\bar{Q}_{i,i}(t,t+h) = P\big(&\text{no demand occurs during } (t,t+h]\\ &\text{and all } i \text{ units in resupply at time } t \text{ are still}\\ &\text{in resupply after } h \text{ time units}\big) + o(h)\\ = e^{-\lambda h}&\binom{i}{i}\tilde{p}(h)^i\left(1-\tilde{p}(h)\right)^0 + o(h)\\ = e^{-\lambda h}&\tilde{p}(h)^i + o(h);\end{aligned}$$

$$\begin{aligned}\bar{Q}_{i+1,i}(t,t+h) = P\big(&\text{no demand occurs during } (t,t+h]\\ &\text{and among the } i+1 \text{ units in resupply}\\ &\text{at time } t, \text{ only one of them is received}\\ &\text{in } (t,t+h]\big) + o(h)\\ = e^{-\lambda h}&\binom{i+1}{i}\tilde{p}(h)^i\left(1-\tilde{p}(h)\right) + o(h)\\ = e^{-\lambda h}&(i+1)\left[\tilde{p}(h)^i - \tilde{p}(h)^{i+1}\right] + o(h);\end{aligned}$$

and

$$\begin{aligned}\bar{Q}_{i-1,i}(t,t+h) = P\big(&\text{a demand occurs during } (t,t+h] \text{ and}\\ &\text{all } i-1 \text{ units in resupply at time } t \text{ are still}\\ &\text{in resupply after } h \text{ time units}\big) + o(h)\\ = \lambda h\, e^{-\lambda h}&\binom{i-1}{i-1}\tilde{p}(h)^{i-1}\left(1-\tilde{p}(h)\right)^0 + o(h)\\ = \lambda h\, e^{-\lambda h}&\tilde{p}(h)^{i-1} + o(h).\end{aligned}$$

Let

$$\begin{aligned}\bar{Q}^{(\zeta)}_{i-1,i}(t,t+h) = P\big(&\text{a demand of type } \zeta \text{ occurs during}\\ &(t,t+h] \text{ and all } i-1 \text{ units in resupply}\\ &\text{at time } t \text{ are still in resupply after}\\ &h \text{ time units where } \zeta \in \{s,g\}\big) + o(h).\end{aligned}$$

Given that a customer demand occurs, the probability of it being a *silver* or *gold* customer demand are λ_s/λ and λ_g/λ, respectively. It follows that

$$\begin{aligned}\bar{Q}^{(s)}_{i-1,i}(t,t+h) &= \lambda_s h\, e^{-\lambda h}\tilde{p}(h)^{i-1} + o(h)\\ \bar{Q}^{(g)}_{i-1,i}(t,t+h) &= \lambda_g h\, e^{-\lambda h}\tilde{p}(h)^{i-1} + o(h).\end{aligned}$$

Next, we derive the limits of the above expressions as $h \to 0$, which we will use for the steady-state analysis of system behavior.

Limits as $h \to 0$:

(a)

$$\begin{aligned}&\lim_{h\to 0}\frac{1-\bar{Q}_{i,i}(t,t+h)}{h}\\ &\quad= \lim_{h\to 0}\frac{1-e^{-\lambda\, h}\tilde{p}(h)^i}{h} - \lim_{h\to 0}\frac{o(h)}{h}\\ &\quad= \lambda + i\frac{G(t)}{\int_0^t[1-G(u)]\,du}, \qquad \text{due to Lemma A1.}\end{aligned}$$

(b)

$$\begin{aligned}&\lim_{h\to 0}\frac{\bar{Q}_{i+1,i}(t,t+h)}{h}\\ &\quad= \lim_{h\to 0}\frac{e^{-\lambda h}(i+1)\left[\tilde{p}(h)^i - \tilde{p}(h)^{i+1}\right]}{h} + \lim_{h\to 0}\frac{o(h)}{h},\\ &\quad= \lim_{h\to 0}\left[e^{-\lambda h}(i+1)\right]\lim_{h\to 0}\frac{\left[\tilde{p}(h)^i - \tilde{p}(h)^{i+1}\right]}{h}\\ &\quad= (i+1)\frac{G(t)}{\int_0^t[1-G(u)]\,du}, \text{ due to Lemma A1.}\end{aligned}$$

(c)

$$\lim_{h\to 0}\frac{\bar{Q}^{(s)}_{i-1,i}(t,t+h)}{h} = \lim_{h\to 0}\frac{1}{h}\lambda_s h\, e^{-\lambda h}\tilde{p}(h)^{i-1} + \lim_{h\to 0}\frac{o(h)}{h}$$

$$= \lim_{h\to 0}\lambda_s\, e^{-\lambda h}\tilde{p}(h)^{i-1}$$

$$= \lambda_s \lim_{h\to 0}\tilde{p}(h)^{i-1}$$

$$= \lambda_s,$$

$$\lim_{h\to 0}\frac{\bar{Q}^{(g)}_{i-1,i}(t,t+h)}{h} = \lim_{h\to 0}\frac{1}{h}\lambda_g h\, e^{-\lambda h}\tilde{p}(h)^{i-1} + \lim_{h\to 0}\frac{o(h)}{h}$$

$$= \lim_{h\to 0}\lambda_g\, e^{-\lambda h}\tilde{p}(h)^{i-1}$$

$$= \lambda_g \lim_{h\to 0}\tilde{p}(h)^{i-1}$$

$$= \lambda_g.$$

□

Theorem 3. *Assuming a general, positively valued lead time distribution having finite mean, T, with no probability mass at zero, then, if the Independence Condition is true, the steady-state distribution of (R, B_s) satisfies the same balance equations as a system with an exponential lead time distribution with the same mean.*

Proof. Now, let us develop the ideas for the original problem. The proof for the balance equations (which holds under the *Independence Condition*) will be given for a more general general state (i, j) with $i > S - S_g$ and $j \geq 1$. The other cases can be proven in a similar way. In Fig. A1, the balance equation for $i > S - S_g$ and $j \geq 1$ is given by

$$\pi_{(i,j)}\left[\lambda_s + \lambda_g + \frac{i}{T}\right] = \pi_{(i-1,j)}\lambda_g + \pi_{(i+1,j)}\frac{i+1}{T} + \pi_{(i-1,j-1)}\lambda_s.$$

Recall that at time 0 there are no orders outstanding by assumption; hence the system state is (0, 0). Let

$$P_{(k,l),(i,j)}(t,t') \equiv P\big[\xi_{t'} = (i,j) \mid \xi_t = (k,l)\big],$$
$$\bar{P}_{(k,l),(i,j)}(t,t') \equiv P\big[\xi_{t'} = (i,j) \mid \xi_0 = (0,0),\, \xi_t = (k,l)\big].$$

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

$$P_{(0,0),(i,j)}(0,t+h) = \sum_{k,l} P_{(0,0),(k,l)}(0,t)\bar{P}_{(k,l),(i,j)}(t,t+h).$$

Assuming h is an infinitesimal time unit, the probability of more than one event to happen is $o(h)$, and we have

$$\begin{aligned} P_{(0,0),(i,j)}(0,t+h) = \Big\{ & P_{(0,0),(i-1,j)}(0,t)\bar{P}_{(i-1,j),(i,j)}(t,t+h) \\ & + P_{(0,0),(i+1,j)}(0,t)\bar{P}_{(i+1,j),(i,j)}(t,t+h) \\ & + P_{(0,0),(i-1,j-1)}(0,t)\bar{P}_{(i-1,j-1),(i,j)}(t,t+h) \\ & + P_{(0,0),(i,j)}(0,t)\bar{P}_{(i,j),(i,j)}(t,t+h) \\ & + o(h) \Big\}. \end{aligned}$$

Subtracting $P_{(0,0),(i,j)}(0,t)$ from both sides and then taking the limits as $h \to 0$:

$$\begin{aligned} \lim_{h\to 0}\frac{P_{(0,0),(i,j)}(0,t+h) - P_{(0,0),(i,j)}(0,t)}{h} = \Big\{ & P_{(0,0),(i-1,j)}(0,t)\lim_{h\to 0}\frac{\bar{P}_{(i-1,j),(i,j)}(t,t+h)}{h} \\ & + P_{(0,0),(i+1,j)}(0,t)\lim_{h\to 0}\frac{\bar{P}_{(i+1,j),(i,j)}(t,t+h)}{h} \\ & + P_{(0,0),(i-1,j-1)}(0,t)\lim_{h\to 0}\frac{\bar{P}_{(i-1,j-1),(i,j)}(t,t+h)}{h} \\ & - \lim_{h\to 0}\frac{\Big(1 - \bar{P}_{(i,j),(i,j)}(t,t+h)\Big)}{h}P_{(0,0),(i,j)}(0,t) \\ & + \lim_{h\to 0}\frac{o(h)}{h}\Big\}. \end{aligned} \tag{A1}$$

For the Right-Hand Side (RHS) of the above equation, the transitions, $\bar{P}(\cdot)$, expressed in the limit terms arise because a *gold* demand occurs, a unit is received from resupply, a *silver* demand occurs, and nothing happens, respectively. Now it is time to use the concept that was developed earlier in this section. Examining each of these terms:

$$\begin{aligned} \bar{P}_{(i-1,j),(i,j)}(t,t+h) &= P\Big(\text{a } \textit{gold} \text{ demand occurs during } (t,t+h] \text{ and all } i-1 \\ &\quad \text{units in resupply at time } t \text{ are still in resupply after } h \text{ time} \\ &\quad \text{units} \mid \xi_t = (i-1,j),\ \xi_0 = (0,0)\Big) + o(h) \\ &= P\Big(\text{a } \textit{gold} \text{ demand occurs during } (t,t+h] \text{ and all } i-1 \\ &\quad \text{units in resupply at time } t \text{ are still in resupply after } h \text{ time} \\ &\quad \text{units} \mid \xi_t = i-1,\ \xi_0 = 0\Big) + o(h), \end{aligned}$$

by the *Independence Assumption*. Hence,

$$\bar{P}_{(i-1,j),(i,j)}(t,t+h) = \bar{Q}^{(g)}_{i-1,i}(t,t+h).$$

In like manner,

$$\begin{aligned} \bar{P}_{(i+1,j),(i,j)}(t,t+h) &= P\Big(\text{no demand occurs during } (t,t+h] \text{ and among the } i+1 \\ &\quad \text{units in resupply at time } t \text{, only one of them is} \\ &\quad \text{received in } (t,t+h] \mid \xi_t = i+1,\ \xi_0 = 0\Big) + o(h) \\ &= \bar{Q}_{i+1,i}(t,t+h). \end{aligned}$$

$$\begin{aligned} \bar{P}_{(i-1,j-1),(i,j)}(t,t+h) &= P\Big(\text{a } \textit{silver} \text{ demand occurs during } (t,t+h] \text{ and all } i-1 \\ &\quad \text{units in resupply at time } t \text{ are still in resupply after } h \text{ time} \\ &\quad \text{units} \mid \xi_t = i-1,\ \xi_0 = 0\Big) + o(h) \\ &= \bar{Q}^{(s)}_{i-1,i}(t,t+h). \end{aligned}$$

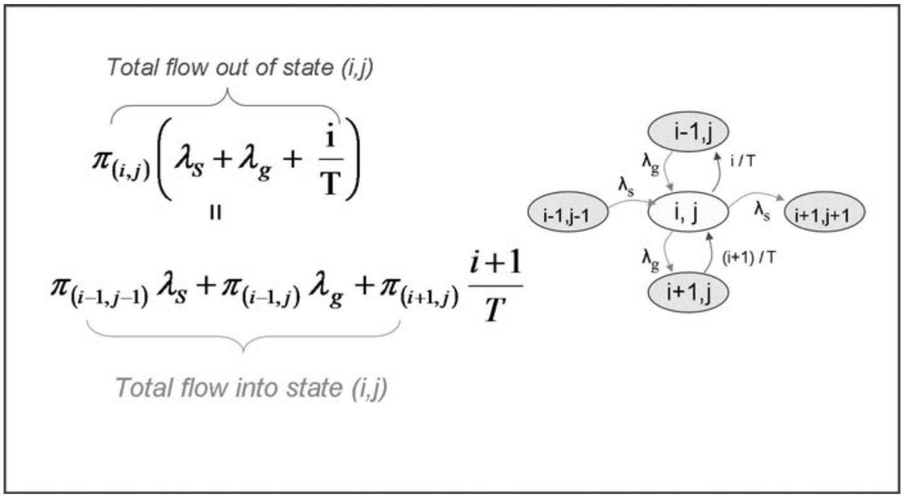
Total flow out of state (i,j)
$\pi_{(i,j)}\left(\lambda_s+\lambda_g+\frac{i}{T}\right)$
=
$\pi_{(i-1,j-1)}\lambda_s+\pi_{(i-1,j)}\lambda_g+\pi_{(i+1,j)}\frac{i+1}{T}$
Total flow into state (i,j)
i-1,j
i-1,j-1
i, j
i+1,j+1
i+1,j
λ_g
i / T
λ_s
λ_s
λ_g
(i+1) / T

Figure A1. Under the *Independence Condition*, the balance equation for a generic system state (*i, j*) for $i > S - S_g$ and $j \geq 1$.

$$\bar{P}_{(i,j),(i,j)}(t,t+h) = P\big(\text{no demand occurs during } (t,t+h] \text{ and all } i \text{ units in resupply at time } t \text{ are still in resupply after } h \text{ time units} \mid \xi_t = i,\ \xi_0 = 0\big) + o(h)$$

$$= \bar{Q}_{i,i}(t,t+h).$$

As a result, we have

$$\lim_{h\to 0}\frac{\bar{P}_{(i-1,j),(i,j)}(t,t+h)}{h} = \lim_{h\to 0}\frac{\bar{Q}^{(g)}_{i-1,i}(t,t+h)}{h} = \lambda_g;$$

$$\lim_{h\to 0}\frac{\bar{P}_{(i+1,j),(i,j)}(t,t+h)}{h} = \lim_{h\to 0}\frac{\bar{Q}_{i+1,i}(t,t+h)}{h} = (i+1)\cdot\frac{G(t)}{\int_0^t[1-G(u)]\,du};$$

$$\lim_{h\to 0}\frac{\bar{P}_{(i-1,j-1),(i,j)}(t,t+h)}{h} = \lim_{h\to 0}\frac{\bar{Q}^{(s)}_{i-1,i}(t,t+h)}{h} = \lambda_s;$$

$$\lim_{h\to 0}\frac{1-\bar{P}_{(i,j),(i,j)}(t,t+h)}{h} = \lim_{h\to 0}\frac{1-\bar{Q}_{i,i}(t,t+h)}{h} = \lambda_s+\lambda_g + i\cdot\frac{G(t)}{\int_0^t[1-G(u)]\,du}.$$

Using the above results in Equation (A1), we get

$$\lim_{h\to 0}\frac{P_{(0,0),(i,j)}(0,t+h)-P_{(0,0),(i,j)}(0,t)}{h} = P_{(0,0),(i-1,j)}(0,t)\cdot\lambda_g + P_{(0,0),(i+1,j)}(0,t)\cdot(i+1)\cdot\frac{G(t)}{\int_0^t[1-G(u)]\,du} + P_{(0,0),(i-1,j-1)}(0,t)\cdot\lambda_s - \left(\lambda_s+\lambda_g+i\cdot\frac{G(t)}{\int_0^t[1-G(u)]\,du}\right)\cdot P_{(0,0),(i,j)}(0,t).$$

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

$$\lim_{t\to\infty}\lim_{h\to 0}\frac{P_{(0,0),(i,j)}(0,t+h)-P_{(0,0),(i,j)}(0,t)}{h} = \lim_{t\to\infty}\Bigg\{P_{(0,0),(i-1,j)}(0,t)\cdot\lambda_g + P_{(0,0),(i+1,j)}(0,t)\cdot(i+1)\cdot\frac{G(t)}{\int_0^t[1-G(u)]\,du} + P_{(0,0),(i-1,j-1)}(0,t)\cdot\lambda_s - \left(\lambda_s+\lambda_g+i\cdot\frac{G(t)}{\int_0^t[1-G(u)]\,du}\right)\cdot P_{(0,0),(i,j)}(0,t)\Bigg\}. \tag{A2}$$

Since

$$\lim_{t\to\infty}\frac{G(t)}{\int_0^t[1-G(u)]\,du} = \frac{1}{T}$$

and due to Theorem 1, the limit exists for the RHS of Equation (A2). After reordering the terms, the RHS becomes

$$= \left\{\pi_{(i-1,j)}\cdot\lambda_g + \pi_{(i+1,j)}\cdot\frac{i+1}{T} + \pi_{(i-1,j-1)}\cdot\lambda_s - \pi_{(i,j)}\cdot\left(\lambda_s+\lambda_g+\frac{i}{T}\right)\right\}.$$

The left-hand side of Equation (A2) is given by

$$= \lim_{t\to\infty} P'_{(0,0),(i,j)}(t).$$

Note that $P_{(0,0),(i,j)}(t)$ is bounded by zero and one for all t. Therefore, if $\lim_{t\to\infty} P'_{(0,0),(i,j)}(t)$ converges, then it must converge to zero. However, we have already shown that it is convergent by the RHS of Equation (A2).

As a result, by rearranging the terms in Equation (A2), we get

$$\pi_{(i,j)}\cdot\left(\lambda_s+\lambda_g+\frac{i}{T}\right) = \pi_{(i-1,j)}\cdot\lambda_g + \pi_{(i+1,j)}\cdot\frac{i+1}{T} + \pi_{(i,j-1)}\cdot\lambda_s.$$

This is the same balance equation as in CTMC with $\mu = 1/T$. The other balance equations for special cases $i = S - S_g$ as well as $j = 0$ follow from similar analysis and match the corresponding balance equations in the CTMC. □

Appendix B

Table B1. Performance of approximations with respect to an increase in workload.

Case	*S*	S_g	λL	β_s (%)	$\beta_{g\,(Simulation)}$ (%)	$\beta_{g\,(CTMC)}$ (%)	AE $_{CTMC}$ (%)	$\beta_{g\,(sgl-cycle)}$ (%)	AE$_{(sgl-cycle)}$ (%)
(I)	5	2	1.5	80.88	99.53 ± 0.02	99.57	0.04	99.40	0.13
(II)	5	2	3	42.41	95.42 ± 0.04	96.05	0.63	92.47	2.95
(III)	5	2	6	6.33	82.59 ± 0.13	85.92	3.33	55.94	26.65
(IV)	5	2	15	~ 0	75.45 ± 0.23	78.93	3.48	2.44	73.01
(V)	5	2	24	~ 0	75.12 ± 0.13	77.64	2.52	~ 0	75.12
(VI)	5	2	30	~ 0	75.26 ± 0.34	77.17	1.91	~ 0	75.26

Table B2. Performance of approximations with respect to an increase in ratio $\lambda_g/(\lambda_s + \lambda_g)$ under fixed workload.

Case	$\lambda_g/(\lambda_s + \lambda_g)$	β_s (%)	$\beta_{g\,(Simulation)}$ (%)	$\beta_{g\,(CTMC)}$ (%)	AE $_{CTMC}$ (%)	$\beta_{g\,(sgl-cycle)}$ (%)	AE$_{(sgl-cycle)}$ (%)
(I)	1/10	61.60	99.86 ± 0.02	99.89	0.03	99.60	0.26
(II)	1/5	61.60	99.47 ± 0.03	99.54	0.07	98.61	0.86
(III)	1/3	61.60	98.54 ± 0.04	98.70	0.16	96.77	1.77
(IV)	1/2	61.60	96.66 ± 0.07	97.02	0.36	94.14	2.52
(V)	2/3	61.60	94.10 ± 0.08	94.57	0.47	91.48	2.62
(VI)	4/5	61.60	91.50 ± 0.12	91.96	0.46	89.45	2.05
(VII)	9/10	61.60	89.21 ± 0.17	89.56	0.35	88.01	1.20