EURANDOM PREPRINT SERIES
2010-036

Two Perishable Inventory Systems with One-way Substitution

Liqiang Liu, Ivo Adan, David Perry
ISSN 1389-2355

Two Perishable Inventory Systems with One-way Substitution

Liqiang Liu [†] , Ivo Adan [‡] , David Perry [§]

September 2010

Abstract: Motivated by the ABO issue of the blood banks system, in which the portions stored have constant shelf life, we consider two subsystems of perishable inventory. The two Perishable Inventory Subsystems – PIS A and PIS B, are correlated to each other through a so-called one-way substitution of demands. Specifically, the input streams and the demand streams applied to each subsystem are four Poisson processes which are independent of one another. However, if the shelf of PIS A (blood of type O) is empty of items an arriving demand of type A is unsatisfied, since demand of type A cannot be satisfied by an item of type B (blood portions of type AB), but if the shelf of PIS B is empty of items an arriving demand of type B is applied to PIS A, since demands of type B can be satisfied by both types. Such a one-way substitution of the issuing policy generates for PIS A a modulated Poisson demand process operating in a two-state non-Markovian environment. The performance analysis of PIS B is known from previous work. Hence, in this study we focus on the marginal performance analysis of PIS A. Based on a fluid formulation and a Markovian approximation for the one-way substitution demands process, we develop a unified approach to efficiently and accurately approximate the performance of PIS A. The effectiveness of the approach is investigated by extensive numerical experiments.

Keywords: PIS; One-way substitution; Approximation; Markovian arrival process; Performance evaluation.

[†]EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands (liu@eurandom.tue.nl)

[‡]Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands, and Department of Quantitative Economics, University of Amsterdam, P.O.Box 19268, 1000 GG Amsterdam, the Netherlands (iadan@win.tue.nl)

[§]Department of Statistics, University of Haifa, Haifa 31909 Israel (dperry@haifa.ac.il)

1. Introduction

Motivated by the operation of blood banks, we consider a stochastic input-output inventory system composed of two correlated Perishable Inventory Systems - PIS A and PIS B. Items of type A arrive in the shelf of PIS A according to a Poisson process with rate λ_A and items of type B arrive in the shelf of PIS B according to a Poisson process with rate λ_B. The shelf life of all items is a constant that without loss of generality is equal to 1. Demands of type A, that arrive according to a Poisson process with rate μ_A apply to PIS A and demands of type B, that arrive according to a Poisson process with rate μ_B, apply optionally, first to PIS B, but if PIS B is empty the application is routed to PIS A. The four Poisson streams are independent. However, since demands of type A can be satisfied only by items of type A, a demand of type A leaves unsatisfied if shelf A is empty. Asymmetrically, demands of type B can be satisfied by either items of type B or by items of type A. We assume that the issuing policy in both subsystems is first-in-first-out (FIFO).

We are interested in the marginal performance analysis of each subsystem. The performance measures are the long-run averages for the number of items on shelf s_A (s_B), the rate of item loss due to perishing ℓ_A (ℓ_B) and the rate of overall demand loss m. In fact, several versions of the marginal performance analysis of PIS B have already been carried out in previous work (for the analysis of PIS B as described above, see [13]). However, the performance analysis of PIS A appears to be new. To see the intricateness of a rigorous analysis, note that while the arrival process of items into PIS B is Poisson with rate λ_B and the demand process is Poisson with rate μ_B, the arrival process into PIS A is Poisson with rate λ_A, but the demand process applied to PIS A is a modulated Poisson process operating in a two-state random environment which is determined according to the environment status of PIS B. Namely, when shelf B is not empty, the demand process applied to PIS A is a Poisson process with rate μ_A, but when shelf B is empty, the demand process applied to PIS A is a Poisson process with rate $\mu_A+\mu_B$. Hence the demand process applied to PIS A is non-renewal, but tractable PISs in literature are mostly restricted to renewal arrival processes. Furthermore, while the time periods in which shelf B is empty are exponentially distributed (λ_B), the time periods in which shelf B is not empty are **not** exponential (but the law of these time periods can be computed), so that the random environment associated with the demand arrival process into PIS A is not Markovian. The latter fact makes a rigorous analysis of PIS A too complicated, if possible at all. In light of the intricateness of the demand process applied to PIS A, it seems that it is not likely to be possible to perform an exact analysis of the relevant performance measures of PIS A. This situation motivates us to develop a new approach to analyze a single PIS under a pure Markovian setting. Accordingly, we relax the renewal property and exploit the Markovian property by taking approximations when necessary, for tractability. The approach is a combination of two simple ideas of practical importance:

(a) Approximate the non-renewal arrival process by a Markovian arrival process (cf. [21]); specifically, a Markov-modulated Poisson process.
(b) Use a fluid formulation for the virtual waiting time process (e.g. [1], p. 308).

We apply an accurate approximation for the demand process applied to PIS A. To understand the rationale behind our approximation, let U_n be the nth non-emptiness period in PIS B, which we call ON period. Clearly $U_1, U_2, \ldots$ are independent and identically distributed (i.i.d.) random variables; let U be the generic random variable of the sequence. Since we know how to compute the law of U we also know how to compute its moments. Now take a random $\hat{U}$ with a known phase

type distribution such that $\mathbb{E}U^k = \mathbb{E}\hat{U}^k$ for all $k = 1, 2, \ldots, n$, for some predetermined n. Then intuitively, the model in which the original U is replaced by $\hat{U}$ will be a good approximation to the original model and the approximation will improve as n increases.

Our model is motivated by the ABO issue associated with blood banks. In practice, there are four types of blood – O, A, B and AB (in this study we assume a generic model of only two types). The blood portions arrive in accordance with four independent Poisson streams and are classified into the four categories (shelves) – O, A, B and AB. According to the formal statistics, more than 40% of the population belong to category O and less than 5% belong to category AB. Demands of type O can be satisfied by only blood portions of type O, but demands of type AB can be satisfied by either blood portions of type O or blood portions of type AB. It turns out that a shortage in blood portions of type O might have been a disaster, but a shortage in blood portions of type AB is important only from a managerial point of view. According to the formal standards and regulations, the maximum shelf life of all blood portions is 21 days. It means that after 21 days any blood portion cannot be used for transfusion. Then, after 1 unit of time (21 days), the blood portion is removed from the shelf. Except extreme urgency cases, it is natural to believe that the issuing policy of blood portions is FIFO. Namely, if the shelf is not empty, any arriving demand is satisfied by the oldest portion on the shelf. From a modeling point of view this means that the FIFO issuing policy can be used as a good approximation to reality.

The literature about perishable inventory systems is quite rich and comprises several directions of main streams. Over the last 30 years four review papers have been published on the general field (see [17] for the first review paper, [26], [12] and recently, a comprehensive updated monograph by Nahmias [18]). According to the review papers above it seems that most of the work in the field looks on the topic from an optimal control point of view. More specifically, in one main stream of models that are introduced in the literature there is a controller who faces the problem of 'when to place an order' and 'how much to order'. As a result, models of this type focus on optimization of policies. Our model belongs to a different main stream which is strongly related to queueing models with abandonments due to impatience of customers. In this model there is no controller, since the input is random. In addition, it is noted in [11, 17] that the finite shelf life of the stored items in the inventory system can be interpreted as the finite patience of the customers in a single server queueing system, also known as the finite dam model, which has been extensively studied (e.g. [4, 8, 9, 7, 3, 24]). However, Kaspi and Perry [13] notice a stronger connection between PIS and a queueing model with customers of waiting-line impatience. This connection lays the groundwork for further research on PIS from a queueing perspective (e.g. [14, 25, 23, 19]). As for the model considered in this paper, the management becomes much more intricate due to the one-sided interplay between the two types of items. Such an interplay between different types of items is commonly known as substitution or lateral transshipment in multi-product inventory systems (see, e.g., [5, 2, 27]).

The paper is organized as follows. In Section 2, we present a fluid formulation for the PIS. We apply the fluid formulation to PIS B and immediately obtain a fluid model driven by a continuous time Markov chain (CTMC). In Section 3, we address PIS B and illustrate the proposed approach in detail. We re-derive several known results. We introduce explicit formulas for quantities of our interest, in particular, for the ON period. A direct application of the fluid formulation to PIS A results in a non-Markovian model. In Section 4, we propose to approximate the ON period by certain phase type distributed random variables by using the same

moments of the ON period. We extend the state space of the driving CTMC. Then the approximate evaluation of PIS A is similar to that of PIS B. In the last section, we present numerical experiments to investigate approximation errors.

2. Fluid Formulation

Consider $A(t)$, the age of the oldest stock in a PIS at time $t \geq 0$. By convention, let $A(t) = 0$ if the system is empty (out-of-stock) at time t. Illustrated in upper Figure 1, the sample path of $A(t)$ has a linear growth of rate 1 when the inventory is in stock. The linear growth is interpreted as the aging of the oldest stock. A downward jump occurs when the oldest stock is removed from the system. A removal of the oldest stock happens due to either

(a) an arrival of demand (recall that in Section 1 we assume FIFO, namely the oldest stock is always used to satisfied a demand); or
(b) perishing.

Let $0 \leq S_1 \leq S_2 \leq \ldots$ be all epochs that an item is removed. We assume that the sequence $\{S_i\}_{i=1,2,\ldots}$ corresponds to a locally finite counting process, i.e. $\mathbb{P}\{\lim_{i\to\infty} S_i < \infty\} = 0$. Clearly the size of the ith jump, $i = 1, 2, \ldots$, is the inter-arrival time of the ith supply and the next, possibly truncated due to non-negativity of $A(t)$. We assume $A(t)$ is right continuous. We shall analyze the age process $\{A(t)\}_{t\geq 0}$ and show that the performance measures of our interest can be derived from the stationary distribution of the age process. Our approach is to couple the age process with a family of bivariate processes $\{(I_r, A_r)\}$ parameterized by a single positive constant r. The motivation is easy to see by thinking of vertical jumps in a sample path as segments of infinite slope ($r = \infty$). We construct A_r in such a way that the sample paths of A_r converge to the sample paths of A as $r \to \infty$. Under a Markovian setting, it turns out that the (I_r, A_r) process is a piecewise deterministic Markov process (PDMP, cf. [10]) with continuous (linear) sample paths, also known as fluid models. Thus we formulate and solve the PIS problem as a fluid model. A similar approach is also used in [16] to analyze the busy period of a $M/PH/1$ queueing model with impatient customers. Although the probability laws can be derived directly for the A process within the PDMP framework, we introduce such a construction because it is a unified approach, which is directly amenable to computations. The construction and our rationale to introduce it shall become clear in the following sections. The numerical experiments in Section 5 also serve as a convincing support of the approach.

The idea is explained on Figure 1. For any $r > 0$, let us define a process $\{A_r(t)\}_{t\geq 0}$ as the following transform of the A process.

(1)
$$
\begin{array}{lcll}
S_0 & = & 0, & \\
Z_i & = & S_{i+1} - S_i, & \\
\Delta_i^X & = & X_i \mathbb{1}_{\{A(S_i)+Z_i=1\}}/r, & \\
\Delta_i^Y & = & \left(A(S_i) + Z_i - A(S_{i+1})\right)/r, & \\
T_i & = & S_i + \sum_{j<i}(\Delta_j^X + \Delta_j^Y), & \\
A_r(t) & = & \left\{ \begin{array}{ll} A(t - T_i + S_i), & t \in [T_i,\ T_i + Z_i), \\ 1, & t \in [T_i + Z_i,\ T_{i+1} - \Delta_i^Y), \\ A(S_i) + Z_i - r(t + \Delta_i^Y - T_{i+1}), & t \in [T_{i+1} - \Delta_i^Y,\ T_{i+1}), \end{array} \right. & \\
i & = & 0, 1, \ldots, &
\end{array}
$$

where

(a) $X_0, X_1, \ldots$ are i.i.d. exponential random variables with mean 1;
(b) $\mathbb{1}_{\{x\}}$ denotes the indicator function that equals 1 if condition x is true and 0 otherwise.

Let

$$
\begin{aligned}
T_i' &= \inf\{t \in [T_i,\ T_i + Z_i) : A_r(t) > 0\}, \\
I_r(t) &= \begin{cases} -, & t \in [T_i,\ T_i'), \\ +, & t \in [T_i',\ T_i + Z_i), \\ +, & t \in [T_i + Z_i,\ T_{i+1} - \Delta_i^Y), \\ -, & t \in [T_{i+1} - \Delta_i^Y,\ T_{i+1}). \end{cases} \\
i &= 0, 1, \ldots
\end{aligned} \tag{2}
$$

Figure 1 is an illustration of the construction in (1) and (2), with $r = 1$.

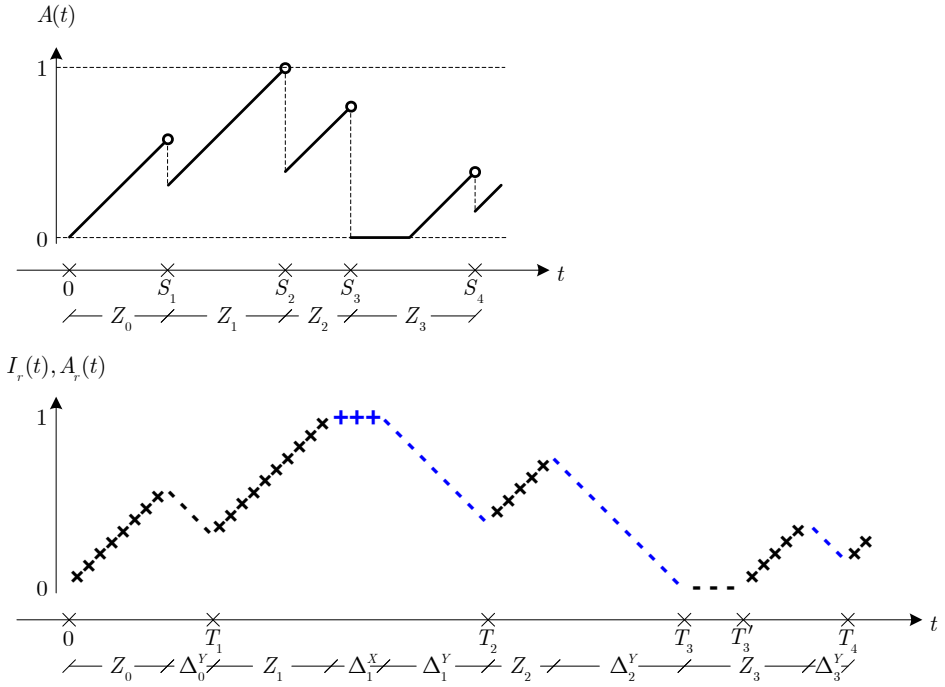

FIGURE 1. A sample path of the A process, with the corresponding sample path of the (I_r, A_r) process.

The construction, in the limit, gives an equivalent representation of the original A process. More precisely, we have the following proposition. Let τ be any stopping time of the A process, and τ_r be the corresponding one for the A_r process, $\tau_r := \tau + \sum_{i:S_i<\tau}(\Delta_i^X + \Delta_i^Y)$. Further let the random variable $A(\infty)$ and $A_r(\infty)$, respectively, be distributed as the limiting distribution of the A and A_r processes.

Proposition 1. *As $r \to \infty$:*

(a) *For all $t \geq 0$, $A_r(t) \to A(t)$ almost surely.*
(b) *$\tau_r \downarrow \tau$ almost surely.*
(c) *For all $n \geq 0$ and $t \geq 0$, $\mathbb{E}A_r^n(t) \to \mathbb{E}A^n(t)$.*
(d) *For all $n \geq 0$, $\mathbb{E}\tau_r^n \to \mathbb{E}\tau^n$.*
(e) *For all $n \geq 0$, $\mathbb{E}A_r^n(\infty) \to \mathbb{E}A^n(\infty)$.*

Proof. (a) and (b) immediately follow from the construction of the A_r process. (c) and (d) follow from dominated, respectively monotone convergence. To prove (e), note that A is regenerative. Let C be the first cycle in the A process, and C_r be the corresponding one in the A_r process. By (a) and (b),

$$\int_{C_r} A_r^n(t)\mathrm{d}t \to \int_C A^n(t)\mathrm{d}t,$$

with probability 1. Then, by dominated convergence

$$\mathbb{E}\int_{C_r} A_r^n(t)\mathrm{d}t \to \mathbb{E}\int_C A^n(t)\mathrm{d}t.$$

So

$$\mathbb{E}A_r^n(\infty) = \frac{\mathbb{E}\int_{C_r} A_r^n(t)\mathrm{d}t}{\mathbb{E}C_r} \to \frac{\mathbb{E}\int_C A^n(t)\mathrm{d}t}{\mathbb{E}C} = \mathbb{E}A^n(\infty).$$

□

The implication of Proposition 1 is that it is sufficient to perform our probabilistic computations, for both transient and limiting analysis, on the (I_r, A_r) process, then take the limit with respect to r. With this in mind we now turn our attention to the analysis of the (I_r, A_r) process. We treat the PIS B case in the next section before we go into a slightly more general setting for PIS A.

3. PIS B, Poisson Demands

In this section we apply the fluid formulation to the A process in PIS B, where both demand and supply are independent Poisson processes of rate μ_B and λ_B respectively. We give results for the stationary distributions of the (I_r, A_r) process (Proposition 2) and the A process (Theorem 3). We derive the performance measures from the stationary distribution (Section 3.2). We also study a first passage time of the age process in PIS B, which is useful when we proceed to PIS A. It is known from general PDMP studies that the stationary distribution, or the transform of a first passage time of the (I_r, A_r) process, is the solution to a certain boundary problem of linear ordinary differential equations (ODE). We exploit this result and give our solutions explicitly.

For conciseness, we omit the subscripts in λ_B and μ_B when handling solely PIS B. Since the inter-supply times, and the inter-demand times, for PIS B are i.i.d. exponential random variables, $\{(I_r(t), A_r(t))\}_{t\geq 0}$ is a time-homogeneous Markov process with state space $\{\texttt{+}, \texttt{-}\} \times [0, 1]$. The (I_r, A_r) process evolves as follows. When $I_r = \texttt{+}$ (resp. $I_r = \texttt{-}$), A_r increases (decreases) with rate 1 (r), unless $A_r = 1$ ($A_r = 0$). In the latter case A_r stays flat until I_r switches to the other state. When $I_r = \texttt{+}$ (resp. $I_r = \texttt{-}$), I_r stays for an exponentially distributed time with mean μ^{-1} ($r^{-1}\lambda^{-1}$) in that state, then switches to the other, unless $A_r = 1$ ($A_r = 0$). In the latter case I_r remains $\texttt{+}$ (resp. $\texttt{-}$) for an exponentially distributed time with mean r^{-1} (λ^{-1}) then switches to the other, which in consequence takes A_r away from the boundaries and the evolution again is driven according to the rules for $A_r \in (0, 1)$. The description above is exactly the so-called fluid model driven by a CTMC with a state space $\{\texttt{+}, \texttt{-}\}$ (cf. [15]), with special behavior on the boundaries. It is easy to see that the generator of I_r is as follows:

$$Q_t = \begin{cases} \begin{bmatrix} -r & r \\ 0 & 0 \end{bmatrix} \equiv \overline{Q}, & \text{if } A_r(t) = 1, \\ \begin{bmatrix} -\mu & \mu \\ r\lambda & -r\lambda \end{bmatrix} \equiv Q, & \text{if } 0 < A_r(t) < 1, \\ \begin{bmatrix} 0 & 0 \\ \lambda & -\lambda \end{bmatrix} \equiv \underline{Q}, & \text{if } A_r(t) = 0. \end{cases} \tag{3}$$

3.1. Stationary Distribution.

Let $\mathrm{diag}(\vec{v})$ denote the diagonal matrix of a vector $\vec{v}$ and

$$R = \mathrm{diag}([1 \ \ -r]) = \begin{bmatrix} 1 & 0 \\ 0 & -r \end{bmatrix}. \tag{4}$$

It can be proved rigorously within the PDMP framework that the stationary distribution of the (I_r, A_r) process has a density $f_r(i, x)$ at $(i, x) \in \{\texttt{+}, \texttt{-}\} \times (0, 1)$

and atoms $p_r(+)$, $p_r(-)$ at $(+,1)$ and $(-,0)$ respectively, which satisfy the following system of equations:

$$\vec{p}_r \underline{Q} - \vec{f}_r(0)R = \vec{0}, \tag{5a}$$

$$\vec{p}_r \underline{Q} + \int_0^x \vec{f}_r(u)\mathrm{d}u Q - \vec{f}_r(x)R = \vec{0}, \qquad x \in (0,1), \tag{5b}$$

$$\vec{p}_r \overline{Q} + \vec{f}_r(1)R = \vec{0}, \tag{5c}$$

where $\vec{p}_r$ and $\vec{f}_r(x)$ are row vectors as follows

$$\vec{p}_r = [p_r(+)\ p_r(-)], \quad \vec{f}_r(x) = [f_r(+,x)\ f_r(-,x)]. \tag{6}$$

Intuitively the ith equation of (5a)-(5c) are the global balance equations for the state sets $\{(i,0)\}$, $\{(i,a) : 0 \le a < x\}$ and $\{(i,1)\}$ respectively. Now write (5b) in derivative form as

$$\vec{f}_r(x)Q - \frac{\mathrm{d}\vec{f}_r(x)}{\mathrm{d}x}R = \vec{0}. \tag{7}$$

Thus (5) becomes a standard boundary problem of linear ODE. The explicit solution is given in the following proposition.

Proposition 2. *Let $\overline{Q}, Q$ and $\underline{Q}$ be as in (3). Let R be as in (4). The (I_r, A_r) process has a unique stationary distribution, which has a density f_r of the form*

$$\vec{f}_r(x) = \vec{p}_r \underline{Q} \exp(R^{-1}Qx)R^{-1}, \quad x \in (0,1), \tag{8}$$

and atoms $p_r(+)$, $p_r(-)$ at $(+,1)$ and $(-,0)$ respectively. The atoms are uniquely determined by

$$\vec{p}_r \left(\underline{Q} \exp(R^{-1}Q) + \overline{Q}\right) = \vec{0}, \tag{9a}$$

$$\left(\vec{p}_r + \int_{x=0}^1 \vec{f}_r(x)\mathrm{d}x\right)\vec{1} = 1, \tag{9b}$$

where $\vec{0}$ is the zero row vector and $\vec{1}$ is the column vector with all entries being 1.

Proof. From (5a) and (7) we get (8). Evaluating $\vec{f}_r(1)$ using (8) and substituting in (5c) we get (9a). Since $\underline{Q} + \overline{Q}$ is irreducible, the null space of $\left(\underline{Q} \exp(R^{-1}Q) + \overline{Q}\right)$ has exactly one dimension. Then $\vec{p}_r$ is completely determined by the normalization equation (9b). □

Remark 3.1. The matrix $R^{-1}Q$ in the exponent does not depend on r and is diagonalizable.

The stationary distribution of the A process can be obtained by taking the limit $r \to \infty$ (Proposition 1). Let

$$f(x) = \lim_{r\to\infty} f_r(+,x), \quad p = \lim_{r\to\infty} p_r(-). \tag{10}$$

Alternatively it is simply the conditional distribution given that $I_r = +$ and $A_r < 1$, or $I_r = -$ and $A_r = 0$ (e.g. [1], Proposition 1.12, p. 309). Hence, for any given $r > 0$,

$$f(x) = \frac{1}{\sigma_r} f_r(+,x), \quad p = \frac{1}{\sigma_r} p_r(-), \tag{11}$$

where

$$\begin{aligned} \sigma_r &= \lim_{t\to\infty} \mathbb{P}\left\{(I_r(t), A_r(t)) \in \{(+,a) : 0 \le a < 1\} \cup \{(-,0)\}\right\} \\ &= \int_{x=0}^1 f_r(+,x)\mathrm{d}x + p_r(-). \end{aligned} \tag{12}$$

Either way we give the explicit result in the following theorem and omit the proof.

Theorem 3 (Stationary Distribution)**.** *The age of the oldest stock in PIS B has a unique stationary distribution, which has a density*

$$f(x) = p\lambda e^{(\lambda-\mu)x}, \quad x \in (0,1)$$

and an atom

$$p = \begin{cases} \frac{\lambda-\mu}{\lambda e^{\lambda-\mu}-\mu}, & \text{if } \lambda \neq \mu \\ \frac{1}{\mu+1}, & \text{if } \lambda = \mu \end{cases}$$

at 0.

Remark 3.2. Theorem 3 is well known in the context of the finite dam model and coincides with Corollary 2.7 of [24].

We are interested in the distribution of the stock level in steady state, denoted by N. The generating function of N can be obtained by conditioning on the age of the oldest stock in steady state, namely $\mathbb{E}(z^N) = \mathbb{E}(\mathbb{E}(z^N|A))$, $|z| \leq 1$. Notice that if $A = a > 0$, then $N-1$, the number of supplies since the arrival of the oldest stock, has a Poisson distribution with mean λa, i.e., $\mathbb{E}(z^{N-1}|A = a > 0) = e^{-\lambda a(1-z)}$. The probability distribution can be obtained from the coefficients of the power series of the generating function. We give the result in the following corollary and omit the proof.

Corollary 4 (Stock Level)**.** *Let*

$$h(z) = \frac{\lambda z - \mu}{\lambda z e^{\lambda z - \mu} - \mu}.$$

The moment generating function of the stock level in steady state is

$$\mathbb{E}(z^N) = \frac{h(1)}{h(z)}, \quad |z| \leq 1,$$

and

$$\mathbb{P}(N = k) = p\left(\frac{\lambda}{\mu}\right)^k e^{-\mu} \sum_{i=k}^{\infty} \frac{\mu^i}{i!}.$$

Remark 3.3. A direct calculation of the long-run average stock level is as follows.

$$s_B = \mathbb{E}(N) = \mathbb{E}(\mathbb{E}(N|A)) = \int_{x=0}^{1} (1 + \lambda x) f(x) \mathrm{d}x. \tag{13}$$

3.2. **Performance Measures.** Recall that the performance measures we consider for PIS B are long-run average stock level (s_B) and perishing rate (ℓ_B). Also the substitution demand rate (m_B) is of interest, since the performance of PIS A depends on m_B. Clearly $m_B = \mu p$. Since every supply is either perished or issued to a demand, and every demand is either satisfied by a supply or lost, in the long run, we have $\lambda - \ell_B = \mu - m_B$. We use this conservation law to compute $\ell_B = \lambda - \mu + m_B$. The stock level can be computed from Corollary 4. Notice that all three performance measures can be expressed in terms of p. We list them as follows,

Corollary 5 (PIS B Performance)**.** *When* $\lambda \neq \mu$*,*

$$m_B = \frac{\mu(\lambda-\mu)}{\lambda e^{\lambda-\mu} - \mu}, \quad \ell_B = \lambda - \mu + m_B, \quad s_B = \frac{\lambda(\lambda+1)e^{\lambda}}{\lambda e^{\lambda} - \mu e^{\mu}} - \frac{\lambda}{\lambda - \mu}.$$

When $\lambda \to \mu$*,*

$$m_B \to \frac{\mu}{\mu+1}, \quad \ell_B \to \frac{\mu}{\mu+1}, \quad s_B \to \frac{\mu(\mu+2)}{2(\mu+1)}.$$

3.3. **First Passage Times.** We start by identifying the ON and OFF periods in PIS B, which are closely related to the performance measures of our interest. The ON (resp. OFF) period is the period during which $A(t) > 0$ ($A(t) = 0$) and demands are satisfied (unsatisfied and routed to PIS A). Let U (resp. D) be the generic random variables for the duration of the ON (OFF) period. The ON and OFF periods affect the performance of PIS A by modulating its demand process.

Obviously D is exponentially distributed with mean λ^{-1}. Now we define, for the (I_r, A_r) process, the first passage time, τ_r, and the Laplace-Stieltjes transform (LST), $\phi_{\alpha,r}$, which are related to U. Let

$$\tau_r = \inf\{t > 0 : A_r(t) = 0\}, \tag{14a}$$

$$\phi_{\alpha,r}(i,x) = \mathbb{E}(e^{-\alpha\tau_r}|I_r(0) = i, A_r(0) = x), \quad x \in [0,1], \quad Re(\alpha) > 0. \tag{14b}$$

Let $\phi(\alpha) = \mathbb{E}(e^{-\alpha U})$ be the LST of the ON period. From Proposition 1 we have

$$\phi(\alpha) = \lim_{r\to\infty} \phi_{\alpha,r}(\texttt{+}, 0). \tag{15}$$

An immediate result from the PDMP theory is that the column vector $\vec{\phi}_{\alpha,r}(x) = [\phi_{\alpha,r}(\texttt{+}, x)\ \phi_{\alpha,r}(\texttt{-}, x)]^{\mathrm{T}}$ satisfies the following differential equation:

$$\frac{\mathrm{d}\vec{\phi}_{\alpha,r}(x)}{\mathrm{d}x} = R^{-1}(\alpha I - Q)\vec{\phi}_{\alpha,r}(x), \quad x \in (0,1), \tag{16}$$

where I is the identity matrix. Next we specify the boundary conditions. Clearly

$$\phi_{\alpha,r}(\texttt{-}, 0) = 1. \tag{17}$$

Recall that the trajectory of A_r, given $I_r(0) = \texttt{+}$ and $A_r(0) = 1$, stays on the boundary for an exponentially distributed time with mean r^{-1}, and then leaves upon I_r switching from + to -. Therefore we have the following factorization

$$\phi_{\alpha,r}(\texttt{+}, 1) = \frac{r}{\alpha + r}\phi_{\alpha,r}(\texttt{-}, 1). \tag{18}$$

With these two boundary conditions, the solution is uniquely determined as given in the following proposition.

Proposition 6. *Let $\overline{Q}, Q$ and $\underline{Q}$ be as in* (3)*. Let R be as in* (4)*. The LST of the first passage time defined in* (14) *is given by*

$$\begin{bmatrix}\phi_{\alpha,r}(\texttt{+}, x)\\ \phi_{\alpha,r}(\texttt{-}, x)\end{bmatrix} = M(x)\begin{bmatrix}\phi_{\alpha,r}(\texttt{+}, 0)\\ 1\end{bmatrix}, \quad x \in [0,1], \tag{19}$$

where

$$M(x) = \exp(R^{-1}(\alpha I - Q)x), \quad \begin{bmatrix}m_+ & m_\downarrow\\ m_\uparrow & m_-\end{bmatrix} = M(1),$$

and

$$\phi_{\alpha,r}(\texttt{+}, 0) = \left(m_+ - \frac{r}{\alpha + r}m_\uparrow\right)^{-1}\left(\frac{r}{\alpha + r}m_- - m_\downarrow\right). \tag{20}$$

Proof. From (17) and (16) we get (19). Evaluate $\phi_{\alpha,r}(1)$ by (19), in which we substitute (18) and obtain

$$\begin{cases} \frac{r}{\alpha+r}\phi_{\alpha,r}(\texttt{-}, 1) &= m_+\phi_{\alpha,r}(\texttt{+}, 0) + m_\downarrow\\ \phi_{\alpha,r}(\texttt{-}, 1) &= m_\uparrow\phi_{\alpha,r}(\texttt{+}, 0) + m_- \end{cases}.$$

Solving the equation above yields (20). □

The following theorem is a direct result from Proposition 6 and (15).

Theorem 7 (ON Period, LST)**.** *The LST of the ON period of PIS B is given by*

$$\phi(\alpha) = \frac{\nu + (\alpha + \lambda - \mu) + e^{\nu}[\nu - (\alpha + \lambda - \mu)]}{\nu - (\alpha - \lambda + \mu) + e^{\nu}[\nu + (\alpha - \lambda + \mu)]},$$

where

$$\nu = \sqrt{(\alpha + \lambda - \mu)^2 + 4\mu\alpha} = \sqrt{(\alpha - \lambda + \mu)^2 + 4\lambda\alpha}.$$

Remark 3.4. Theorem 7 is in agreement with Perry and Asmussen [24], Corollary 3.1, Model II.

Remark 3.5. An alternative way to determine p, instead of using (9) is to use the mean of U and D. Notice that the process $\{\mathbb{1}_{\{A(t)>0\}}\}_{t\geq 0}$ is an alternating renewal process. Then $p = \mathbb{E}(D)/(\mathbb{E}(U) + \mathbb{E}(D))$.

The ith moment of the ON period can be obtained from Theorem 7 by

$$\mathbb{E}(U^i) = (-1)^i \left.\frac{\mathrm{d}\phi(\alpha)}{\mathrm{d}^i\alpha}\right|_{\alpha=0}.$$

We list the explicit formulas for the first three moments in

Corollary 8 (ON Period, Moments)**.** *When $\lambda \neq \mu$,*

$$\mathbb{E}(U^i) = \frac{i}{(\lambda - \mu)^{2i-1}} \sum_{j=0}^{i} c_{ij} e^{j(\lambda-\mu)}, \quad i = 1, 2, 3,$$

where

$$\begin{aligned}
c_{10} =& -1,\\
c_{11} =& 1,\\
c_{20} =& -\mu,\\
c_{21} =& -(\lambda(1+\lambda) - \mu(1+\mu)),\\
c_{22} =& \lambda,\\
c_{30} =& -2\mu(\lambda+\mu),\\
c_{31} =& 2\lambda^3 + \lambda^4 - 4\lambda\mu - 6\lambda^2\mu + 2\mu^2 + 2\lambda\mu^2 - 2\lambda^2\mu^2 + 2\mu^3 + \mu^4,\\
c_{32} =& -2\lambda\left(\lambda(1+2\lambda) - \mu(3+2\mu)\right),\\
c_{33} =& 2\lambda^2.
\end{aligned}$$

When $\lambda \to \mu$,

$$\mathbb{E}(U) \to 1, \quad \mathbb{E}(U^2) \to 1 + \frac{2}{3}\mu, \quad \mathbb{E}(U^3) \to 1 + 2\mu + \frac{4}{5}\mu^2.$$

These moments are useful when we deal with PIS A in the next section. The squared coefficient of variation (SCV) of the ON period, defined as

$$c_U^2(\lambda, \mu) = \mathbb{E}(U^2)/\mathbb{E}^2(U) - 1,$$

is given as follows.

$$c_U^2(\lambda, \mu) = \begin{cases} \frac{2}{3}\mu, & \text{if } \lambda = \mu, \\ \frac{\left(e^{2\lambda} - e^{2\mu} - 2e^{\lambda+\mu}(\lambda-\mu)\right)(\lambda+\mu)}{(e^{\lambda} - e^{\mu})^2(\lambda-\mu)}, & \text{if } \lambda \neq \mu. \end{cases}$$

This seems an interesting result for further comparative study with the diffusion approximation of $M/M/1/K$ queues (cf. Williams [28], Equation (6); Berger and Whitt [6], Equation (29)). Figure 2 shows a contour plot of $c_U^2(\lambda, \mu)$. Note $c_U^2(\lambda, \mu)$ is sensitive near the ridge $\lambda = \mu$, which is a symmetry axis of the function values as well, i.e., $c_U^2(\lambda, \mu) = c_U^2(\mu, \lambda)$. The sensitivity increases when λ $(= \mu)$ gets larger. The region $\{(\lambda, \mu) : c_U^2(\lambda, \mu) \geq c\}$ shrinks to a ray as $c \to \infty$. The observations

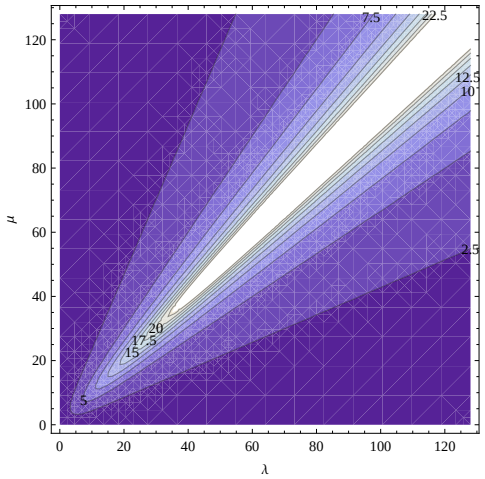

FIGURE 2. A contour plot of $c_U^2(\lambda, \mu)$.

above provide certain heuristic information to explore the parameter space in the numerical experiments (Section 5).

4. PIS A, Modulated Poisson Demands

As mentioned in Section 1, the demand process of PIS A is a modulated Poisson process. For PIS A, the resulting (I_r, A_r) process by the fluid formulation of (1) and (2) is not Markovian any more. We adopt an approximation as follows. First we use a phase type (PH) distribution (cf. [20]) with an irreducible representation $(\vec{\gamma}, T)$ to approximate the distribution of the ON period U, i.e.

$$U \approx \hat{U}, \quad \mathbb{P}\{\hat{U} > x\} = \vec{\gamma} e^{Tx} \vec{1}, \quad \text{for } x \geq 0.$$

To specify the representation $(\vec{\gamma}, T)$, a probably over-simplified solution is to take $\vec{\gamma} = 1$ and $T = -1/\mathbb{E}(U)$, i.e., approximate U by an exponential random variable with mean $\mathbb{E}(U)$. It is worth noting that closed form solutions are developed in [22] for mapping a general distribution (on the positive half-line) to a PH distribution, which matches the first three moments. In this section let us assume $(\vec{\gamma}, T)$ is given.

The PH approximation enables us to enlarge the state space of the I_r process in order to render the (I_r, A_r) process Markovian. Let $n - 1 \geq 1$ be the number of phases in the PH random variable that approximates the ON period. Let $J(t) = i, \quad i = 1, 2, \ldots, n-1$ if PIS B is in phase i of an ON period and $J(t) = n$ if PIS B is in an OFF period at time t. Then the process $\{J(t), \quad t \geq 0\}$ is a CTMC with a generator M (of size n) as follows.

$$M = \begin{bmatrix} T & -T\vec{1} \\ \lambda_B \vec{\gamma} & -\lambda_B \end{bmatrix}.$$

The steady-state analysis now proceeds as a straightforward extension of our treatment of PIS B. Recall that the fluid formulation approach is to construct a fluid model driven by a CTMC of finite state space, the I_r process. An additional modulating CTMC J introduced here is nothing but an extension of the state space of I_r from $\{+, -\}$ to $\{+1, +2, \ldots, +n, -1, -2, \ldots, -n\}$. The extended state now captures information about the J process as well, as the numeric part in our notation. For matrix notations, we index the states by 1 through $2n$ in the order they are listed above. The essence of the construction in (1) is to insert pieces (parameterized by $r > 0$) into the original A process so that, as $r \to \infty$, these pieces vanish

and the two processes coincide with probability 1. During the inserted periods the J process is not allowed to make a transition, i.e., its state will be suspended. This implies that the conditional (I_r, A_r) process is identical to the original (J, A) process, and hence, the same approach as (11) can be employed to obtain the stationary distribution of (J, A). Specifically we extend the matrices in (3) as follows.

$$Q_t = \begin{cases} \begin{bmatrix} -rI & rI \\ 0I & 0I \end{bmatrix} & \equiv \overline{Q}, \quad \text{if } A_r(t) = 1, \\ \begin{bmatrix} M - \mathrm{diag}(\vec{\mu}) & \mathrm{diag}(\vec{\mu}) \\ r\lambda_A I & -r\lambda_A I \end{bmatrix} & \equiv Q, \quad \text{if } 0 < A_r(t) < 1, \\ \begin{bmatrix} 0I & 0I \\ \lambda_A I & M - \lambda_A I \end{bmatrix} & \equiv \underline{Q}, \quad \text{if } A_r(t) = 0, \end{cases} \tag{21}$$

where I is the identity matrix of size n. Notice the diagonal matrices rI in $\overline{Q}$, and $r\lambda_A I$ in Q, effectively suspend the J process by disallowing phase transitions. The entries of the row vector $\vec{\mu}$ are the demand rates of PIS A modulated by J, i.e., all entries of $\vec{\mu}$ are μ_A, except the last one, which is $\mu_A + \mu_B$. The dimension of R is also extended as

$$R = \begin{bmatrix} I & 0 \\ 0 & -rI \end{bmatrix}. \tag{22}$$

For the stationary distribution, Proposition 2 is readily extendable. First we introduce the following vector notations.

$$\begin{aligned} \vec{f}_r(+, x) &= [f_r(+1, x)\ f_r(+2, x)\ \ldots\ f_r(+n, x)], \\ \vec{f}_r(-, x) &= [f_r(-1, x)\ f_r(-2, x)\ \ldots\ f_r(-n, x)], \\ \vec{p}_r(+) &= [p_r(+1)\ p_r(+2)\ \ldots\ p_r(+n)], \\ \vec{p}_r(-) &= [p_r(-1)\ p_r(-2)\ \ldots\ p_r(-n)]. \end{aligned}$$

Then we extend the notation introduced in (6) as

$$\vec{p}_r = [\vec{p}_r(+)\ \vec{p}_r(-)], \quad \vec{f}_r(x) = [\vec{f}_r(+, x)\ \vec{f}_r(-, x)].$$

Equipped with these notations, we extend Proposition 2 as follows.

Proposition 9. *Let $\overline{Q}, Q$ and $\underline{Q}$ be as in* (21)*. Let R be as in* (22)*. The (I_r, A_r) process has a unique stationary distribution, which has a density f_r of the form* (8) *and atoms $p_r(+j)$, $p_r(-j)$ at $(+j, 1)$ and $(-j, 0)$ respectively, for $j = 1, 2, \ldots, n$. The atoms are uniquely determined by* (9)*.*

For the stationary distribution of the (J, A) process, $\vec{f}(x)$ and $\vec{p}$, we extend (10)–(12) as follows.

$$\vec{f}(x) = \lim_{r\to\infty} \vec{f}_r(+, x), \quad \vec{p} = \lim_{r\to\infty} \vec{p}_r(-). \tag{23}$$

$$\vec{f}(x) = \frac{1}{\sigma_r}\vec{f}_r(+, x), \quad \vec{p} = \frac{1}{\sigma_r}\vec{p}_r(-), \tag{24}$$

$$\begin{aligned} \sigma_r &= \lim_{t\to\infty} \mathbb{P}\left\{ (I_r(t), A_r(t)) \in \bigcup_{j=1}^{n} \{(+j, a) : 0 \le a < 1\} \cup \{(-j, 0)\} \right\} \\ &= \left[\int_{x=0}^{1} \vec{f}_r(+, x)\mathrm{d}x + \vec{p}_r(-) \right] \vec{1}. \end{aligned} \tag{25}$$

However we can no longer get explicit expressions of such a simple form as in Theorem 3.

We can again compute the performance measures from the stationary distribution of the (J, A) process. Clearly $m = \vec{\mu}\vec{p}^{\mathrm{T}}$. By conservation we get $\ell_A = \lambda_A - (\mu_A + m_B) + m$. Extending (13), we get the average stock level

$$s_A = \int_{x=0}^{1} (1 + \lambda_A x)\vec{f}(x)\mathrm{d}x\vec{1}. \tag{26}$$

Although first passage times in PIS A are irrelevant to the performance measures in our current consideration, we note that Proposition 6 is readily extendable as well. Denote

$$\vec{\phi}_{\alpha,r}(+, x) = [\phi_{\alpha,r}(+1, x)\ \phi_{\alpha,r}(+2, x)\ \ldots\ \phi_{\alpha,r}(+n, x)]^{\mathrm{T}},$$
$$\vec{\phi}_{\alpha,r}(-, x) = [\phi_{\alpha,r}(-1, x)\ \phi_{\alpha,r}(-2, x)\ \ldots\ \phi_{\alpha,r}(-n, x)]^{\mathrm{T}}.$$

Then we have

Proposition 10. *Let $\overline{Q}, Q$ and $\underline{Q}$ be as in* (21)*. Let R be as in* (22)*. The LST of the first passage time defined in* (14) *is given by*

$$\begin{bmatrix} \vec{\phi}_{\alpha,r}(+, x) \\ \vec{\phi}_{\alpha,r}(-, x) \end{bmatrix} = M(x) \begin{bmatrix} \vec{\phi}_{\alpha,r}(+, 0) \\ \vec{1} \end{bmatrix}, \quad x \in [0, 1],$$

where

$$M(x) = \exp(R^{-1}(\alpha I - Q)x), \quad \begin{bmatrix} m_+ & m_\downarrow \\ m_\uparrow & m_- \end{bmatrix} = M(1),$$

$m_\bullet$ are $n \times n$ blocks and

$$\vec{\phi}_{\alpha,r}(+, 0) = \left(m_+ - \frac{r}{\alpha + r} m_\uparrow\right)^{-1} \left(\frac{r}{\alpha + r} m_- - m_\downarrow\right) \vec{1}.$$

To this end, the approach outlined above provides a unified treatment for similar models where the supply and demand are two independent Markovian arrival processes. Certainly it becomes delicate to construct the Q_t matrix and even more so to specify the Markovian arrival process of the unsatisfied demands. The size of the state space of the driving Markov chain may grow rapidly, which is a common limitation for approaches based on Markovianization with supplementary variables.

5. Numerical Experiments

To validate our approach, in this section we conduct experiments which focus on approximation errors in comparison to discrete event simulation of the inventory systems. Three approximations are in consideration:

- Poisson approximation (named PA): isolated PIS A has demand stream as Poisson process (PP) with rate $\mu_A + m_B$.
- Exponential approximation (named EA): isolated PIS A has demand stream as ON-OFF Markov-modulated PP with rates μ_A and $\mu_A + \mu_B$ respectively. The ON period is approximated by an exponential random variable whose mean coincides with the mean duration of the ON period of PIS B.
- Three-moments approximation (named M3A): Same as EA, except that the ON period is approximated by a PH random variable that matches the first three moments of the ON period, using the algorithm developed by Osogami and Harchol-Balter (cf. [22], Fig. 8).

Clearly PA is the most straightforward one to use and it is usually seen in the literature (e.g. [27]). The other two belong to the PH approximation discussed in Section 4. Compared to EA, M3A uses a more refined approximation for the ON period, thus better approximates the exact superposition demand process of PIS A. In principle we can continue to refine such an approximation to be as precise

as desired, attributed to the denseness of the class of PH distributions, which is a well-known fact stated as follows (eg. [29], p. 271). For any non-negative random variable, the ON period U in our case, there exists a sequence of PH random variables that converges to U in distribution. The main difficulty in practice is to find the PH approximation. A fruitful approach in this area is the so-called moment matching algorithm, for which we refer the reader to [22] and the references therein.

Although it would be of practical interest to bound the error for each approximation (so that one can choose an approximation of the lowest refinement level from all approximations that meet a given precision requirement), we are not going to pursue the matter here. However we are interested to see whether it is possible to draw qualitative conclusions at this stage. Intuitively, the significance of an exact description of the total demand process diminishes, if the substitution demands have a relatively negligible contribution. We may use a ratio as follows to roughly quantify the impact of PIS B on PIS A,

$$\eta = \frac{m_B}{\mu_A}.$$

Then one may think of η as an "amplifier" of the approximation error and expect that, for any of the three approximations, the accuracy decays as η increases, given other possible factors remain the same. Heuristically, another factor of importance seems the variability of the ON period of PIS B in the sense that the larger the normalized variance, the more significant an exact description of the total demand process will be. Hence the advantage of using a more refined approximation should be more prominent when the ON period is of higher variability. When we consider these factors simultaneously, it is plausible to conjecture that the normalized error is monotonically increasing in η, in c_U^2, and decreasing in the refinement level of the approximations. Therefore, at least, one would be somewhat assured to use PA when both η and c_U^2 are small, otherwise be alerted about the potential pitfall.

The main purpose of the experiments in this section is to evaluate the three approximations. Meanwhile we try to seek numerical evidence for the above discussion on the choice of an economical adequate approximation. The following experiments are carried out in three settings to generate test cases. We start with a setting to get a general review of the approximations for a fairly wide range of system parameters. This setting also conveniently serves as a cross validation for our implementations of the analytical computation and the simulation. Then we proceed to an interesting extreme setting for which we are able give a high contrast demonstration of the approximation quality. Finally we return to our motivating application and test the approximations for several series of realistic system parameters.

5.1. **The Wide Setting.** The test cases in this setting are generated as follows. We fix $\mu_A = 1$ (so that $\eta = m_B$) and $\lambda_A = \mu_A$ [1]. Then we vary both μ_B and the supply-demand ratio of PIS B $\rho_B = \lambda_B/\mu_B$ in $\{2^i; i = -2, \ldots, 2\}$. Thus we obtain 25 test cases in total.

The side-by-side comparison of M3A and simulation is reported in Table 1. We do not list the perishing rate (ℓ) since it is computed in terms of the demand lost rate (m).

We make the following observations from a close examination of Table 1. First, the wide coverage of these 25 test cases is evident by the ranges of η (from nearly 0 to 3.04), c_U^2 (from 0.1 to 2.67) and the size of the matrix T (from 2 to 20). Second, the precision is extraordinarily high. The absolute differences between the values by M3A and by simulation are in the scale of 10^{-4}. Hence M3A is remarkably

[1] The choice of an originally balanced PIS A is somewhat arbitrary. A simple reason is that in general a balanced system is more sensitive to perturbations than an imbalanced one.

Rate, Supply B	Rate, Demand B	Rate, Subs.	SCV (ON)	n (PH)	Stock Level A		Rate, Demand Lost		Sim. Time (Sec.)
					M3A	Sim.	M3A	Sim.	
0.0625	0.25	0.24	0.10	20	0.6974	0.6975	0.6543	0.6541	12.26
0.125	0.5	0.45	0.21	11	0.6523	0.6524	0.8065	0.8064	13.90
0.25	1	0.85	0.41	7	0.5782	0.5783	1.1116	1.1117	16.76
0.5	2	1.59	0.78	5	0.4695	0.4695	1.7424	1.7426	22.00
1	4	3.04	1.29	2	0.3336	0.3335	3.1001	3.0996	32.16
0.125	0.25	0.22	0.12	12	0.7002	0.7004	0.6458	0.6456	13.14
0.25	0.5	0.41	0.25	8	0.6616	0.6617	0.7763	0.7763	15.21
0.5	1	0.72	0.50	5	0.6042	0.6043	1.0119	1.0120	19.25
1	2	1.23	0.97	4	0.5288	0.5290	1.4414	1.4420	26.80
2	4	2.15	1.77	2	0.4360	0.4359	2.2871	2.2871	40.70
0.25	0.25	0.20	0.17	8	0.7055	0.7056	0.6301	0.6300	17.73
0.5	0.5	0.33	0.33	5	0.6778	0.6779	0.7238	0.7238	17.73
1	1	0.50	0.67	4	0.6474	0.6475	0.8516	0.8517	23.86
2	2	0.67	1.33	2	0.6256	0.6255	0.9970	0.9971	35.02
4	4	0.80	2.67	2	0.6204	0.6205	1.1334	1.1332	55.79
0.5	0.25	0.16	0.25	6	0.7145	0.7146	0.6034	0.6033	16.96
1	0.5	0.22	0.50	4	0.7027	0.7028	0.6448	0.6448	22.32
2	1	0.23	0.97	4	0.7030	0.7031	0.6549	0.6551	32.39
4	2	0.15	1.77	2	0.7216	0.7217	0.6035	0.6035	51.09
8	4	0.04	2.66	2	0.7432	0.7433	0.5272	0.5271	87.08
1	0.25	0.10	0.41	4	0.7276	0.7277	0.5648	0.5647	21.56
2	0.5	0.09	0.78	2	0.7306	0.7306	0.5582	0.5581	30.85
4	1	0.04	1.29	2	0.7420	0.7420	0.5253	0.5252	48.47
8	2	0.00	1.63	2	0.7492	0.7493	0.5025	0.5024	82.80
16	4	0.00	1.67	2	0.7500	0.7501	0.5000	0.4999	149.97

TABLE 1. M3A vs. simulation. The length of the 99% confidence intervals (CI-99) of all simulation estimates are in the scale of 10^{-4}. The M3A computation times are several milliseconds. ($\lambda_A = \mu_A = 1$)

accurate in this setting. Third, performance evaluation by M3A is efficient. For all cases, it takes merely several milliseconds on an ordinary desktop computer, which makes M3A accessible for evaluation-intensive optimization procedures.

On the other hand, EA (even PA) also performs reasonably well in this setting. Table 2 illustrates the relative errors. The outcome can be mostly explained by our intuitive rationale about the relation between η, c_U^2 and approximation error. For example, a comparison between Case $\lambda_B = 2, \mu_B = 4$ and Case $\lambda_B = 4, \mu_B = 2$ reveals the influence of η; a comparison between Case $\lambda_B = 0.25, \mu_B = 1$, Case $\lambda_B = 0.5, \mu_B = 1$ and Case $\lambda_B = 4, \mu_B = 4$ reveals the influence of c_U^2. The top three errors of PA indeed involve large η and/or c_U^2. For the remaining 22 cases, the error of PA is less than 4 percent. This observation motivates the next setting where we shall see the effort for a refined approximation is well paid off.

5.2. **The Extreme Setting.** Here we consider an interpretation of our PIS A/B model as follows. Let us think of A and B as two quality grades of a product with shelf life, say, one month. A customer who demands a grade B product will always be willing to accept a grade A (which is a higher grade) product but never the other way around. For certain reasons, we regard all customers equally important and decide to satisfy any demand whenever it is possible. Now suppose grade B product is a "fast mover", a product of high supply and demand rates, say, thousands of

Rate, Supply B	Rate, Demand B	Rate, Subs.	SCV (ON)	n (PH)	Error (%), Stock Level A		Error (%), Demand Lost	
					PA	EA	PA	EA
0.0625	0.25	0.24	0.10	20	-0.03	-0.02	0.00	0.03
0.125	0.5	0.45	0.21	11	-0.08	0.00	-0.15	0.02
0.25	1	0.85	0.41	7	-0.49	-0.01	-0.54	0.03
0.5	2	1.59	0.78	5	-2.41	0.01	-1.03	0.02
1	4	3.04	1.29	2	-8.51	-0.49	-0.81	-0.08
0.125	0.25	0.22	0.12	12	-0.05	-0.03	-0.04	0.03
0.25	0.5	0.41	0.25	8	-0.14	-0.01	-0.27	0.04
0.5	1	0.72	0.50	5	-0.72	0.00	-0.95	0.06
1	2	1.23	0.97	4	-3.20	-0.05	-2.10	-0.06
2	4	2.15	1.77	2	-10.80	-1.06	-2.58	-0.48
0.25	0.25	0.20	0.17	8	-0.05	-0.02	-0.09	0.03
0.5	0.5	0.33	0.33	5	-0.19	0.00	-0.45	0.05
1	1	0.50	0.67	4	-0.80	0.00	-1.44	0.05
2	2	0.67	1.33	2	-2.54	-0.11	-3.38	-0.28
4	4	0.80	2.67	2	-5.78	-0.84	-5.92	-1.31
0.5	0.25	0.16	0.25	6	-0.06	-0.02	-0.13	0.04
1	0.5	0.22	0.50	4	-0.20	-0.01	-0.54	0.04
2	1	0.23	0.97	4	-0.49	-0.02	-1.31	-0.02
4	2	0.15	1.77	2	-0.61	-0.06	-1.72	-0.23
8	4	0.04	2.66	2	-0.24	-0.05	-0.72	-0.15
1	0.25	0.10	0.41	4	-0.06	-0.02	-0.14	0.03
2	0.5	0.09	0.78	2	-0.10	-0.01	-0.33	0.02
4	1	0.04	1.29	2	-0.09	-0.01	-0.27	0.00
8	2	0.00	1.63	2	-0.02	-0.01	-0.02	0.02
16	4	0.00	1.67	2	-0.01	-0.01	0.02	0.02

TABLE 2. Relative errors for PA and EA. ($\lambda_A = \mu_A = 1$)

units per month; grade A product is a "slow mover", a product of low supply and demand rates, say, tens of units per month. An interesting case arises if PIS B is balanced, i.e., the supply rate equals the demand rate. In this case a shortage of grade B product can be quite unlikely. For example, if $\lambda_B = \mu_B = 1000$, then for a long term the probability of shortage is less than 0.1 percent. However, such an event is not too unlikely to be negligible for PIS A. For example, if $\mu_A = 9$, then the substitution demand amounts to almost 10 percent of the total demand of PIS A. Whenever a shortage of grade B product happens, PIS A bears a billowing surge of demands. This hints at a high variability in the superposition demand process of PIS A, for which we would expect a high contrast in accuracies of the three approximations.

We generate 10 test cases as follows. We fix $\lambda_A = \mu_A = 1$, balance supply and demand for PIS B, then vary μ_B in $\{2^i; i = -2, \ldots, 7\}$.

The side-by-side comparison of M3A and simulation is reported in Table 3. Figure 3 illustrates the relative errors for all three approximations. As we expect, the errors increase as η and c_U^2 increase simultaneously. We can clearly see that among the three approximations, M3A is the most accurate while PA is the least.

The experiment in this setting also reveals a limitation of our approach. Here we record an error slightly higher than 5 percent for the average stock level evaluated by M3A. If we keep increasing μ_B, then the accuracy of M3A may eventually become insufficient. This observation suggests a direction of further investigation in heavy-tailed traffic queueing systems for alternatives.

Another observation is that our approximations apparently tend to under-estimate the average stock level and the demand lost rate.

Rate, Supply B	Rate, Demand B	Rate, Subs.	SCV (ON)	n (PH)	Stock Level A		Rate, Demand Lost		Sim. Time (Sec.)
					M3A	Sim.	M3A	Sim.	
0.25	0.25	0.20	0.17	8	0.7055	0.7056	0.6301	0.6300	14.45
0.5	0.5	0.33	0.33	5	0.6778	0.6779	0.7238	0.7238	17.73
1	1	0.50	0.67	4	0.6474	0.6475	0.8516	0.8517	23.97
2	2	0.67	1.33	2	0.6256	0.6255	0.9970	0.9971	35.24
4	4	0.80	2.67	2	0.6204	0.6205	1.1334	1.1332	56.01
8	8	0.89	5.33	2	0.6294	0.6313	1.2423	1.2437	94.84
16	16	0.94	10.67	2	0.6432	0.6508	1.3181	1.3247	169.97
32	32	0.97	21.33	2	0.6549	0.6716	1.3652	1.3806	318.95
64	64	0.98	42.67	2	0.6626	0.6905	1.3919	1.4185	617.31
128	128	0.99	85.33	2	0.6670	0.7058	1.4063	1.4437	1210.82

TABLE 3. M3A vs. simulation. The length of CI-99 of all simulation estimates are in the scale of 10^{-4}, except for demand lost rate in the last 5 cases, where the lengths are in the scale of 10^{-3}. The M3A computation times are several milliseconds. ($\lambda_A = \mu_A = 1$)

5.3. **The Realistic Setting.** In this setting, we start from ($\lambda_A = 25, \mu_A = 20, \mu_B = 30, \lambda_B = 40$), which is supposedly close to the reality of our motivating blood bank application. We try to see the influence of substitution, as well as the relation between system parameters and performance. We generate five series of parameters as follows.

(a) Vary each parameter of the four, so that the supply-demand ratio for the corresponding system (PIS A or B) varies within the interval $[0.5, 1.5]$. This results in four series, named λ_A-series, μ_A-series, etc. For example, the λ_A-series is generated by varying λ_A from 10 to 30.

(b) Vary the shelf life parameter (b) within the interval $[1, 2]$. This results in one series, named b-series.

For λ_A-, μ_A- and b-series, we expect hardly any accuracy difference between the three approximations. All of them should be quite accurate, since the influence of PIS B on PIS A is negligible in all cases (the largest η being in the scale of 10^{-5}). Our expectation is indeed verified by the outcome. More interesting results from the rest series are shown in Figure 4. The performance of PIS A is sensitive to the varying parameter when the supply-demand ratio of PIS B is in the range of 0.5 to 1. Also in this region, accuracy difference among the approximations is observed. The accuracy of M3A is superior.

REFERENCES

[1] S. Asmussen. *Applied Probability and Queues*. Applications of Mathematics. Springer, New York, 2nd edition, 2003.

[2] S. Axsäter. Evaluation of unidirectional lateral transshipments and substitutions in inventory systems. *European Journal of Operational Research*, 149(2):438 – 447, 2003. Sequencing and Scheduling.

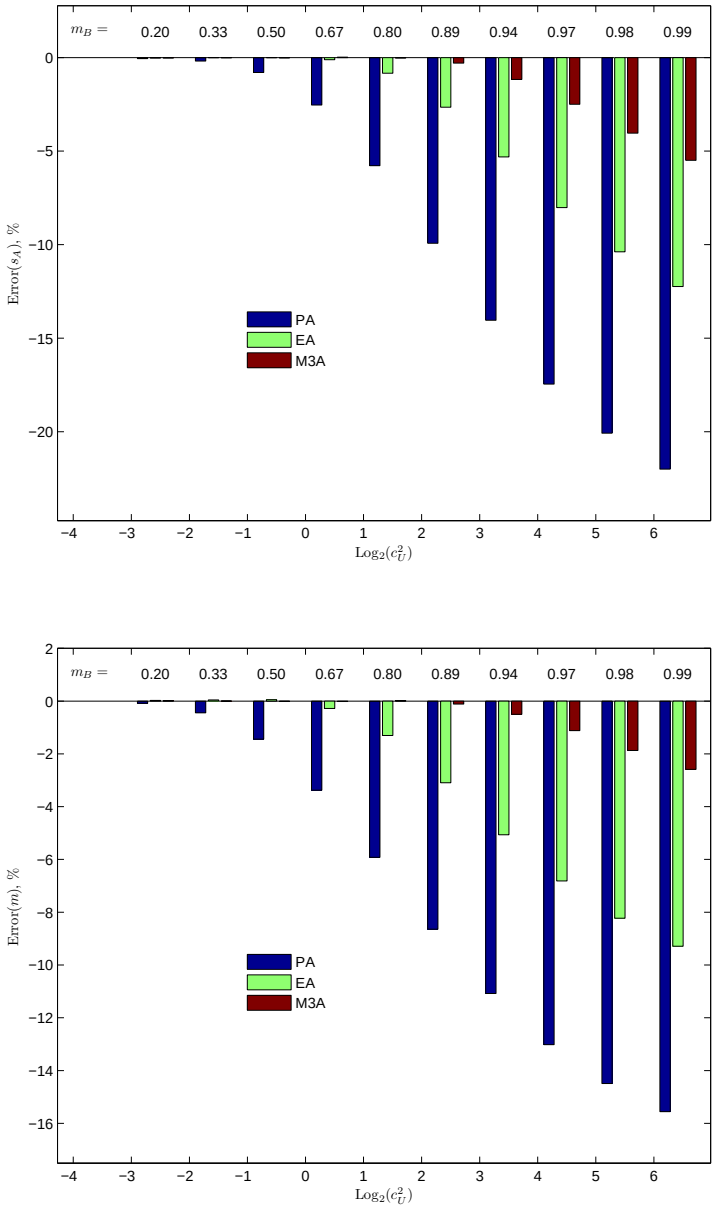

FIGURE 3. Relative errors, $\lambda_A = \mu_A = 1$.

[3] F. Baccelli, P. Boyer, and G. Hebuterne. Single-server queues with impatient customers. *Advances in Applied Probability*, 16(4):887–905, 1984.

[4] D. Y. Barrer. Queuing with impatient customers and ordered service. *Operations Research*, 5(5):650–656, 1957.

[5] Y. Bassok, R. Anupindi, and R. Akella. Single-period multiproduct inventory models with substitution. *Operations Research*, 47(4):632–642, 1999.

[6] A. Berger and W. Whitt. The Brownian approximation for rate-control throttles and the $G/G/1/C$ queue. *Discrete Event Dynamic Systems*, 2(1):7–60, July 1992.

[7] J. W. Cohen. Single server queue with uniformly bounded virtual waiting time. *Journal of Applied Probability*, 5(1):93–122, 1968.

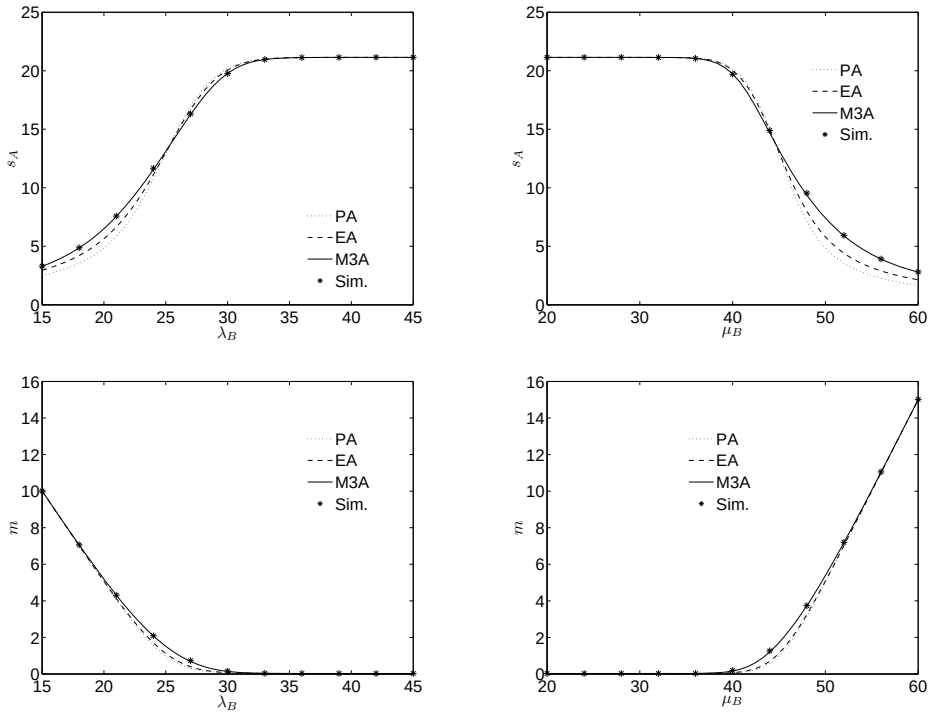

FIGURE 4. Approximations vs. simulation, λ_B-series (μ_B=30) and μ_B-series (λ_B=40), $\lambda_A = 25$, $\mu_A = 20$.

[8] D. J. Daley. Single-server queueing systems with uniformly limited queueing time. *Journal of the Australian Mathematical Society*, 4(04):489–505, 1964.

[9] D. J. Daley. General customer impatience in the queue $GI/G/1$. *Journal of Applied Probability*, 2(1):186–205, 1965.

[10] M. H. A. Davis. Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society Series B-Methodological*, 46(3):353–388, 1984.

[11] S. C. Graves. The application of queueing theory to continuous perishable inventory systems. *Management Science*, 28(4):400–406, 1982.

[12] I. Karaesmen, A. Scheller-Wolf, and B. Deniz. Managing perishable and aging invetories: review and future research directions. In K. Kempf, P. Keskinocak, and P. Uzsoy, editors, *Handbook of Production Planning*, Kluwer International Series in Operations Research and Management Science, Advancing the State-of-the-Art Subseries. Kluwer, 2009. To appear.

[13] H. Kaspi and D. Perry. Inventory systems of perishable commodities. *Advances in Applied Probability*, 15(3):674–685, 1983.

[14] H. Kaspi and D. Perry. Inventory systems for perishable commodities with renewal input and Poisson output. *Advances in Applied Probability*, 16(2):402–421, 1984.

[15] V. G. Kulkarni. Fluid models for single buffer systems. In J. H. Dshalalow, editor, *Frontiers in Queueing: Models and Applications in Science and Engineering*, chapter 11, pages 321–338. CRC Press, Boca Raton, FL, 1997.

[16] L. Liu and V. G. Kulkarni. Busy period analysis for $M/PH/1$ queues with workload dependent balking. *Queueing Systems*, 59(1):37–51, May 2008.

[17] S. Nahmias. Perishable inventory theory: A review. *Operations Research*, 30(4):680–708, 1982.

[18] S. Nahmias. *Perishable Inventory Systems*. Springer-Verlag, 2011. To appear.

[19] S. Nahmias, D. Perry, and W. Stadje. Perishable inventory systems with variable input and demand rates. *Mathematical Methods of Operations Research*, 60(1):155–162, 2004.

[20] M. Neuts. Probability distributions of phase type. *Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206, 1975.

[21] M. F. Neuts. A versatile Markovian point process. *Journal of Applied Probability*, 16(4):764–779, 1979.

[22] T. Osogami and M. Harchol-Balter. Closed form solutions for mapping general distributions to quasi-minimal PH distributions. *Performance Evaluation*, 63(6):524 – 552, 2006. Modelling Techniques and Tools for Computer Performance Evaluation.

[23] D. Perry. Analysis of a sampling control scheme for a perishable inventory system. *Operations Research*, 47(6):966–973, 1999.

[24] D. Perry and S. Asmussen. Rejection rules in the $M/G/1$ queue. *Queueing Systems*, 19(1-2):105–130, 1995.

[25] D. Perry and M. J. M. Posner. An $(S-1, S)$ inventory system with fixed shelf life and constant lead times. *Operations Research*, 46(3):S65–S71, 1998.

[26] F. Raafat. Survey of literature on continuously deteriorating inventory models. *The Journal of the Operational Research Society*, 42(1):27–37, Jan., 1991.

[27] I. C. Reijnen, T. Tan, and G. J. van Houtum. Inventory planning for spare parts networks with delivery time requirements. Technical report, School of Industrial Engineering, Eindhoven University of Technology, the Netherlands, 2009.

[28] R. J. Williams. Asymptotic variance parameters for the boundary local times of reflected Brownian motion on a compact interval. *Journal of Applied Probability*, 29(4):996–1002, 1992.

[29] R. W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, 1989.