Queueing Syst (2012) 70:207–231
DOI 10.1007/s11134-011-9272-8

Production-inventory systems in stochastic environment and stochastic lead times

Vidyadhar Kulkarni · Keqi Yan

Received: 16 February 2010 / Published online: 28 December 2011

Abstract We consider a production-inventory system where the production and demand rates are modulated by a finite state Continuous Time Markov Chain (CTMC). When the inventory position (inventory on hand – backorders + inventory on order) falls to a reorder point r, we place an order of size q from an external supplier. We consider the case of stochastic leadtimes, where the leadtimes are i.i.d. exponential(μ) random variables, and orders may or may not be allowed to cross. We derive the distribution of the inventory level, and analyze the long run holding, backlogging, and ordering cost rate per unit time. We use simulation to study the sensitivity of the system to the distribution of the lead times.

Keywords Inventory systems · Lead times · Fluid queues with jumps

Mathematics Subject Classification (2000) 60K15 · 60K25 · 60K37

1 Introduction

In this paper we study a production-inventory system that evolves in a random environment. Let $I(t)$ be the inventory level at time t. If $I(t) > 0$, it denotes the actual inventory on hand at time t, and if $I(t) < 0$, $-I(t)$ denotes the backorders at time t. Let $Z(t)$ be the state of the environment at time t. We assume that $\{Z(t), t \geq 0\}$ is an irreducible CTMC on state space $\Omega = \{1, 2, \ldots, n\}$ with infinitesimal generator

V. Kulkarni (✉)
Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599, USA
e-mail: vkulkarn@email.unc.edu

K. Yan
SAS Institute, Inc., Cary, NC 27511, USA

$Q = [q_{ij}]$. The inventory level process $\{I(t), t \geq 0\}$ is modulated by the environment as follows: when the environment is in state i, the production occurs continuously at a constant rate r_i, and demand occurs at rate d_i. Thus, while the environment is in state i, the inventory level changes at rate $R_i = r_i - d_i$. Let R be the diagonal matrix defined as $R = \text{diag}(R_1, R_2, \ldots, R_n)$. The demand that cannot be immediately satisfied is backlogged.

We can place orders from outside suppliers to supplement the production and satisfy the demand. We assume the lead times are nonnegative and stochastic. Thus, an order placed at time zero is delivered after a random amount of time.

We define *inventory position* as the inventory on hand minus backorders plus the inventory on order (which is the amount that has been ordered but not yet delivered). Let $P(t)$ be the inventory position at time t. We follow the standard (r, q) ordering policy that operates as follows: when the inventory position decreases to a prespecified reorder point r, we place an order of size q, which arrives after a random period of time. Note that the order size q is fixed and is independent of the environmental state at the time of order placement. We may place a new order before a previous order arrives if the inventory position reduces to r before the order is delivered. Let $O(t)$ denote the number of outstanding orders at time t. The inventory position at time t is given by

$$P(t) = I(t) + qO(t).$$

Now define

$$X(t) = P(t) - r. \tag{1}$$

Since the holding and back-order costs depend on $I(t)$, our main interest is in computing the limiting distribution of the inventory level process $\{I(t), t \geq 0\}$. Since

$$I(t) = X(t) + r - qO(t), \tag{2}$$

we need the limiting joint distribution of $(X(t), O(t))$ to compute the limiting distribution of $I(t)$. In order to compute this we study a trivariate process $\{(X(t), O(t), Z(t)), t \geq 0\}$. We consider two cases.

Serial case This case arises if the orders do not cross, that is, they are delivered in the same sequence in which they are placed. We assume that if there is at least one outstanding order, the next order will be delivered after an $\exp(\nu)$ amount of time.

Parallel case This case arises if the orders can cross, that is, orders may be delivered out of sequence. We assume that the lead times for the individual orders are i.i.d. $\exp(\nu)$. Thus, if there are m outstanding orders, the next order will be delivered after an $\exp(m\nu)$ amount of time.

Both the above cases can be covered by a general model: when there are m outstanding orders, the next order is delivered at rate ν_m. When $\nu_m = \nu$ for $m \geq 1$, we get the serial case, and when $\nu_m = m\nu$, we get the parallel case.

One can think of the trivariate process $\{(X(t), O(t), Z(t)), t \geq 0\}$ as a fluid process $\{X(t), t \geq 0\}$ modulated by the extended environment process $\{(Z(t), O(t)), t \geq 0\}$. However, one cannot use the methods of Yan [28] to compute the limiting distribution of this trivariate process since the state space of the extended environment

process is $\{(j,k) : j \in \Omega, k \geq 0\}$, which is infinite. Yan [27] circumvents this issue by truncating the state space by assuming that there is a finite upper limit on the number of outstanding orders. In this paper we present an analysis under the assumption that there is no upper limit on the number of outstanding orders.

In the analysis we shall need results from the transient and limiting behavior of the standard fluid model (with no jumps), as well as the first passage times. One method of deriving these results is the spectral method: derive a system of linear differential equations and solve them by using standard techniques. This is the classical approach taken in the seminal papers Anick et al. [2], and Mitra [14], and followed by many others. More recently, Asmussen [4] has analyzed the same model (with and without a Brownian motion component) using martingale methods, and by Ahn and Ramaswami [1, 3] and Ramaswami [21] using matrix analytic methods. None of these papers include the analysis of the model with jumps, which is studied by Kulkarni and Yan [11], using the spectral method.

We shall follow the spectral method in this paper. Clearly, the same analysis can be done by these other methods as well. The solutions by the spectral method become particularly concise when the rates R_i are nonzero in each state, as reported by Tanaka et al. [25]. It is tempting to argue that the states with zero rates do not matter, since we can always "skip over" them without affecting the (X, Z) process. However, this argument works only in the study of the limiting behavior of the process. In our case, however, we need to study the transient behavior of the process, and hence we will need to explicitly account for the states with zero rate. The added generality of allowing zero rate states does not shed any new light on the behavior of the system, and is more distracting than useful. Hence, strictly for simplicity, we shall assume in this paper that there are no zero-rate states. This allows us to present many of the results in a closed form.

Stochastic lead times are a well studied aspect of inventory systems. There are many ways to model stochastic lead times that lead to tractable analysis. See Kaplan [10], Nahmias [15], and Zipkin [29]. In general, one assumes that the stochastic lead times are i.i.d. non-negative random variables, and researchers have looked at two cases: orders can cross (as if they are coming from different sources), or that they cannot cross (as if they are coming from a single source that processes the orders sequentially). In this paper, we call them the parallel case and the serial case. Zipkin [29] contains a good description of how these two lead time models can arise. In the inventory literature, the (r, q) policy was first introduced in the classical paper by Galliher et al. [8] when lead times are present. Since then, it has become standard policy to analyze, since an optimal policy can often be found with that form. Zipkin [29] shows that this policy can be optimal under quite general settings. Another stream of inventory literature investigates different demand models. There is a large literature where demand rate is assumed to be constant (as in the standard EOQ model, see Hadley and Whitin [9] and Zipkin [30]), or Poisson (see Galliher et al. [8]), or possess independent increments (see Stidham [24] and Whitt [26]). More recently, researchers have studied more complicated demand processes. For example, Browne and Zipkin [7] have considered a demand process modulated by a continuous time stochastic process, and Song and Zipkin [23] consider a Markov modulated discrete demand process. The recent book by Beyer et al. [6] discusses Markovian

demand models in discrete time and provides a good reference for the history of such demand models. Yan and Kulkarni [28] and Berman and Perry [5] consider a Markov modulated fluid model of demands and productions, but without the lead times.

The key contribution of this work is to combine Markov modulated fluid models for demands, and stochastic lead times. As far as we know, this is the first time a fluid model of a production-inventory system with stochastic lead times and random environment has been considered in the literature. The main insight is to identify the process of outstanding orders as an $SM/M/1$ queue in the serial case and an $SM/M/\infty$ queue in the parallel case, where the notation SM represents an arrival process that generates a single arrival whenever a prespecified semi-Markov process undergoes a transition. Then we use the well-known results by Neuts [17, 18] and Ramaswami [20] to compute the limiting distributions. It is useful to point out here that if we "skip over" the zero-rate states and construct an environmental state with reduced state space, the process of outstanding orders can no longer be analyzed as an $SM/M/1$ or $SM/M/\infty$ queue.

The paper is organized as follows: In the next section we collect all the relevant results about the standard fluid model in one place for ready reference. We mention which are known and which are new at appropriate places. In Sect. 3 we study the process of outstanding orders as an $SM/M/1$ or an $SM/M/\infty$ queue. We use these results to compute the limiting distribution of the $\{(X(t), O(t), Z(t)), t \geq 0\}$ process in Sect. 4. We compute the optimal reorder point in Sect. 5. All the material developed here is then illustrated with a numerical example in Sect. 6. Finally, in Sect. 7, we summarize our conclusions and discuss several possible extensions.

2 The process $\{(X(t), Z(t)), t \geq 0\}$

We can see that the $\{X(t), t \geq 0\}$ process is the same as the inventory process studied in Kulkarni and Yan [11] and Yan and Kulkarni [28], modulated by the environment process $\{Z(t), t \geq 0\}$, and with upward jumps of size q whenever the inventory process hits zero. We need to study this process starting in state $(X(0), Z(0)) = (q, i)$ over the interval $[0, T)$, where

$$T = \min\bigl\{t > 0 : X(t) = 0\bigr\}. \tag{3}$$

Thus, the upward jump behavior at T plays no role in this analysis. Hence, in this section, we assume that $\{(X(t), Z(t)), t \geq 0\}$ is the standard fluid process with no jumps. Here, we state some relevant results for completeness. Some are known previously, while some are new. We clearly state it when the results are new.

Stability The process $\{(X(t), Z(t)), t \geq 0\}$ is stable (i.e., it has a nondefective limiting distribution) if

$$\Delta = -\sum_{i \in \Omega} p_i R_i > 0, \tag{4}$$

where

$$p_i = \lim_{t \to \infty} \mathsf{P}\bigl\{Z(t) = i\bigr\}, \quad i \in \Omega. \tag{5}$$

As discussed before, we shall assume that $R_i \neq 0$ for all $i \in \Omega$, that is, the diagonal matrix R is invertible. Note that the differential equations (6), (12), and (25) that appear below are valid for all R, but the closed form solution is valid only when R is invertible.

Transient analysis Here, we state the main result about the transient distribution of $(X(t), Z(t))$ from Tanaka et al. [25]. Suppose the bivariate process starts in state $(X(0), Z(0)) = (q, i)$ where $q > 0$. Let

$$\pi_{ij}(t,x) = \mathsf{P}\big(X(t) \le x, Z(t) = j | X(0) = q, Z(0) = i\big),$$

and $\pi(t,x) = [\pi_{ij}(t,x)]$. Define the Laplace Transform (LT) as

$$\pi^*_{ij}(s,x) = \int_0^\infty e^{-st}\pi_{ij}(t,x)\,dt.$$

Then, for a given s with $\mathrm{Re}(s) > 0$, the matrix $\pi^*(s,x) = [\pi^*_{ij}(s,x)]$ satisfies the equation

$$\frac{d\pi^*(s,x)}{dx}R = -\pi^*(s,x)(sI - Q) + \pi(0,x), \tag{6}$$

with boundary conditions

$$\pi_{ij}(0,x) = \begin{cases} \delta_{ij} & \text{if } x \ge q, \\ 0 & \text{if } x < q, \end{cases} \tag{7}$$

and

$$\pi_{ij}(t,0) = 0 \quad \text{if } R_j > 0, \ \ t > 0. \tag{8}$$

Here, $\delta_{ij} = 1$ if $i = j$, and 0 otherwise. We need the following notation to write the solution of (6). Assume that $(sI - Q)R^{-1}$ is diagonalizable, and write

$$(sI - Q)R^{-1} = V\Theta V^{-1}, \tag{9}$$

where the right eigenvectors V, and the eigenvalues $\Theta = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_n)$ of $(sI - Q)R^{-1}$ depend on s, but we suppress the dependence to simplify the notation. We assume, without loss of generality, that the states and the eigenvalues θ are numbered so that R_i's and $\mathrm{Re}(\theta_i)$'s both increase in $i \in \Omega$. Let $n^+(n^-)$ be the number of states i with $R_i > 0$ ($R_i < 0$). For $\mathrm{Re}(s) > 0$, we know that (See Anick et al. [2]) $n^+(n^-)$ θ_j's have positive (negative) real part. The solution $\pi^*(s,x)$ can now be written as

$$\pi^*(s,x) = \begin{cases} (sI - Q)^{-1}[I - \exp(-(sI - Q)R^{-1}(x - q))] & \\ \quad + \phi^*(s)\exp(-(sI - Q)R^{-1}x) & \text{if } x \ge q \\ \phi^*(s)\exp(-(sI - Q)R^{-1}x) & \text{if } x < q, \end{cases} \tag{10}$$

where

$$\phi^*(s) = R^{-1}VI^-\Theta^{-1}\exp(\Theta q)\big(I^-VI^- + I^+\big)^{-1}I^-, \tag{11}$$

where I^+ is a diagonal matrix with 1 in the (i,i)th place if $R_i > 0$ and zero otherwise, and $I^- = I - I^+$.

Remark 1 When can we expect $(sI - Q)R^{-1}$ to be diagonlizable? One simple but useful example is the birth and death processes. We can use the results in Ledermann [13] to show that this is the case (when $\mathrm{Re}(s) \geq 0$) if Q is the infinitesimal generator of a birth and death process. In production-inventory applications, a birth and death process can represent an environment that is slowly improving or deteriorating, such as a set of production facilities that come on and off one at a time. We shall use such a system to do numerical computations in Sect. 6.

Remark 2 When $(sI - Q)R^{-1}$ is not diagonalizable, the analysis can still be done using Jordan form. However, the concise representation of the solution disappears. This case does not shed any new light on the system behavior, and hence we shall not treat such a case here.

Special case Suppose $R_i < 0$ for all $1 \leq i \leq n$. Then $X(0) = q$ implies that $0 \leq X(t) \leq q$ for all $t \geq 0$. In this case, the above result simplifies to

$$\pi^*(s,x) = R^{-1}V\Theta^{-1}\exp\bigl(\Theta(q-x)\bigr)V^{-1}, \quad 0 \leq x \leq q.$$

The absorbing case In the sequel we will need the extension of the above transient analysis to the case where the first component of the bivariate stochastic process $\{(X(t), Z(t)), t \geq 0\}$ stays zero once it reaches zero, while the second component continues to evolve as before. We call this the absorbing case and denote the corresponding stochastic process by $\{(X^a(t), Z(t)), t \geq 0\}$. Thus, $X^a(t) = 0 \Rightarrow X^a(t') = 0$ for all $t' \geq t$. Let

$$\pi^a_{ij}(t,x) = \mathsf{P}\bigl(X^a(t) \leq x, Z(t) = j | X^a(0) = q, Z(0) = i\bigr),$$

and $\pi^{a*}_{ij}(s,x)$ be its LT. Then the matrix $\pi^{a*}(s,x) = [\pi^*_{ij}(s,x)]$ satisfies the equation

$$\frac{d\pi^{a*}(s,x)}{dx}R = -\pi^{a*}(s,x)(sI - Q) + \pi^a(0,x), \tag{12}$$

which is same as (6). The boundary conditions are given by

$$\pi^a_{ij}(0,x) = \begin{cases} \delta_{ij} & \text{if } x \geq q, \\ 0 & \text{if } x < q, \end{cases} \tag{13}$$

which is same as (7), and

$$\left.\frac{d\pi^a_{ij}(t,x)}{dx}\right|_{\{x=0\}} = 0 \quad \text{if } R_j > 0,\ t \geq 0, \tag{14}$$

which differs from (8), and reflects the absorbing nature of the X^a process. The solution $\pi^{a*}(s,x)$ can now be written as

$$\pi^{a*}(s,x) = \begin{cases} (sI-Q)^{-1}[I-\exp(-(sI-Q)R^{-1}(x-q))] & \\ \quad +\phi^{a*}(s)\exp(-(sI-Q)R^{-1}x) & \text{if } x \geq q \\ \phi^{a*}(s)\exp(-(sI-Q)R^{-1}x) & \text{if } x < q, \end{cases} \tag{15}$$

where

$$\phi^{a*}(s) = R^{-1}VI^{-}(\Theta)^{-1}\exp(\Theta q)\big[V\Theta V^{-1}I^{+}+VI^{-}\big]^{-1}I^{-}. \tag{16}$$

Equation (15) is identical to (10), but (16) differs from (11), reflecting the changed absorbing behavior. The results in (15) and (16) are new.

Special case Suppose $R_i < 0$ for all $1 \leq i \leq n$. Then $X(0) = q$ implies that $0 \leq X(t) \leq q$ for all $t \geq 0$. In this case the bivariate process $\{(X^a(t), Z(t)), t \geq 0\}$ is identical to $\{(X(t), Z(t)), t \geq 0\}$ and the above result simplifies to

$$\pi^{a*}(s,x) = \pi^{*}(s,x), \quad 0 \leq x \leq q,$$

as expected.

First passage times Let T be the first passage time into the state 0, as defined by (3). For a fixed $q > 0$, define

$$H_{ij}(t,x) = \mathsf{P}\big(T > t, X(t) \leq x, Z(t) = j | X(0) = q, Z(0) = i\big),$$
$$t \geq 0, \quad x \geq 0, \ i,j \in \Omega,$$

and let

$$H(t,x) = \big[H_{ij}(t,x)\big]. \tag{17}$$

Since $X^a(t) > 0$ is equivalent to $T > t$, we get

$$H_{ij}(t,x) = \mathsf{P}\big(0 < X^a(t) \leq x, Z(t) = j | X^a(0) = q, Z(0) = i\big) = \pi^a_{ij}(t,x) - \pi^a_{ij}(t,0). \tag{18}$$

Thus, we can use the analysis of the absorbing case to derive an expression for the LT $H^*(s,x)$ of $H(t,x)$ as follows:

$$H^*(s,x) = \begin{cases} (sI-Q)^{-1}[I-\exp(-(sI-Q)R^{-1}(x-q))] & \\ \quad +\phi^{a*}(s)[\exp(-(sI-Q)R^{-1}x)-I] & \text{if } x \geq q \\ \phi^{a*}(s)[\exp(-(sI-Q)R^{-1}x)-I] & \text{if } x < q, \end{cases} \tag{19}$$

where $\phi^{a*}(s)$ is given by (16). These results are new.

An embedded Markov renewal sequence Next, define

$$D_{ij}(t,x) = \mathsf{P}\big(T \le t, Z(T) = j | X(0) = x, Z(0) = i\big), \quad x \ge 0, \tag{20}$$

$$\begin{aligned}\tilde{D}_{ij}(s,x) &= \mathsf{E}\big(e^{-sT} 1_{\{Z(T)=j\}} \big| X(0) = q, Z(0) = i\big) \\ &= \int_0^\infty e^{-st}\, dD_{ij}(t), \quad x \ge 0. \end{aligned} \tag{21}$$

Also define

$$D_{ij}(t) = D_{ij}(t,q), \tag{22}$$

$$\tilde{D}_{ij}(s) = \tilde{D}_{ij}(s,q). \tag{23}$$

Let $D(t) = [D_{ij}(t)]$ and $\tilde{D}(s) = [\tilde{D}_{ij}(s)]$. Clearly $D(\infty) = \tilde{D}(0)$ is a transition probability matrix of an irreducible Discrete Time Markov Chain (DTMC). Let $\hat{\pi} = [\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n]$ be the unique solution to

$$\hat{\pi} = \hat{\pi}\, \tilde{D}(0), \qquad \sum_{i=1}^{n} \hat{\pi}_i = 1. \tag{24}$$

These quantities have been analyzed by Ramaswami [21] and Narayanan [16]. We recapitulate the main results of Narayanan [16]. The matrix $\tilde{D}(s,x) = [\tilde{D}_{ij}(t,x)]$ satisfies the following differential equation:

$$R \frac{d\tilde{D}(s,x)}{dx} = (sI - Q)\tilde{D}(s,x), \tag{25}$$

with initial conditions

$$\tilde{D}_{ij}(s,x) = 0 \quad \text{if } R_j > 0,\ x > 0,$$

$$\tilde{D}_{ij}(s,0) = \delta_{ij} \quad \text{if } R_i < 0.$$

From (9), it follows that

$$R^{-1}(sI - Q) = R^{-1} V * \Theta * V^{-1} R \tag{26}$$

The solution to (25) can now be written as

$$\tilde{D}(s,x) = R^{-1} V I^- e^{\Theta x} I^- V^{-1} R \big(I^+ V^{-1} R + I^-\big)^{-1} I^-. \tag{27}$$

In particular, we get

$$\tilde{D}(s) = R^{-1} V I^- e^{\Theta q} I^- V^{-1} R \big(I^+ V^{-1} R + I^-\big)^{-1} I^-. \tag{28}$$

Now, suppose $X(0) = q$, and S_k be the kth time the X process hits zero ($S_0 = 0$). Define $Z_k = Z(S_k)$. Due to the Markov nature of the $\{(X(t), Z(t)), t \ge 0\}$ process, and since $X(S_k+) = q$ for all $k \ge 0$, it follows that $\{(S_k, Z_k), k \ge 0\}$ is a Markov

renewal sequence (MRS, see Kulkarni [12] for definition) with kernel $D(\cdot)$ whose LST is given by $\tilde{D}(s)$. These results are new.

Limiting distribution, no jumps Suppose the stability condition of (4) is satisfied. Then the limiting distribution defined by

$$\pi_j(x) = \lim_{t\to\infty} \pi_{ij}(t,x)$$

exists and is unique. Let $\pi(x) = [\pi_1(x), \pi_2(x), \ldots, \pi_n(x)]$. Let Θ and V be as given by (9) with $s = 0$. Again, when there are no jumps, the results of Tanaka et al. [25] can be written in matrix forms as follows:

$$\pi(x) = p - pI^+\left(I^+V^{-1}I^+ + I^-\right)^{-1}\exp(\Theta x)I^+V^{-1}$$

where $p = [p_1, p_2, \ldots, p_n]$ is the limiting distribution of the CTMC $\{Z(t), t \geq 0\}$.

Limiting distribution, with jumps Now consider the case when the X process jumps to q whenever it hits zero. The jumps have no effect on the Z process. Let the limiting distribution be defined by

$$\pi_j^q(x) = \lim_{t\to\infty} \mathsf{P}\big(X(t) \leq x, Z(t) = j\big).$$

Let $\pi^q(x) = [\pi_1^q(x), \pi_2^q(x), \ldots, \pi_n^q(x)]$. Let Θ and V be as given by (9) with $s = 0$. The result of Kulkarni and Yan [11] can be written in a matrix form as follows:

$$\pi^q(x) = p\min(x/q, 1) - pI^+\left(I^+V^{-1}I^+ + I^-\right)^{-1}E(x)I^+V^{-1} \tag{29}$$

where $E(x)$ is a diagonal matrix with

$$E_{ii}(x) = \begin{cases} (\exp(-\theta_i \max(x-q,0)) - \exp(-\theta_i x))/(\theta_i q) & \text{if } \theta_i > 0 \\ 0 & \text{if } \theta_i \leq 0. \end{cases}$$

With these results in place we are ready to analyze the $\{O(t), t \geq 0\}$ process in the next section.

3 The process $\{O(t), t \geq 0\}$

In this section we consider the process $\{O(t), t \geq 0\}$. The (r, q) policy implies that an order of size q is placed with the supplier at times S_k, $k \geq 0$. Thus, $O(t)$ increases by one at every S_k, and if $O(t) = m$ it decreases by one at an exponential rate ν_m. Since $\{(S_k, Z_k), k \geq 0\}$ is an MRS, we can think of $\{O(t), t \geq 0\}$ as the queue length process in a queue with arrivals modulated by a semi-Markov process (SM). In the serial order-processing case, we have $\nu_m = \nu$ for all $m \geq 1$, and hence $\{O(t), t \geq 0\}$ is the queue-length process in an $SM|M|1$ queue. Similarly, in the parallel order-processing case, we have $\nu_m = m\nu$ for all $m \geq 0$, and hence $\{O(t), t \geq 0\}$ is the

queue-length process in an $SM|M|\infty$ queue. We use this structure to compute the limiting distribution of the outstanding orders in the next two subsections.

Define $O_k = O(S_k-)$. Assuming the limits exist, define

$$g_{m,i} = \lim_{k\to\infty} \mathsf{P}(O_k = m, Z_k = i), \quad i \in \Omega,\ m \geq 0,$$

and

$$g_m = [g_{m,1}, g_{m,2}, \ldots, g_{m,n}].$$

3.1 The serial case

In this case, $\{(O_k, Z_k), k \geq 0\}$ is the embedded chain in an $SM|M|1$ queue and its limiting distribution can be computed using the results of Neuts [18], or from Theorem 4.2.1 of Neuts [19]. First, we study the stability condition.

Let τ be the mean time between two consecutive orders in steady state. Since the steady state rate of demands is Δ of (4), and the size of each order is q, a simple conservation of inventory argument implies that

$$\tau\Delta = q. \tag{30}$$

An $SM|M|1$ queue is stable (i.e., it has a non-defective limiting distribution) if $\nu\tau > 1$, hence the condition of stability for the $\{O(t), t \geq 0\}$ process can be written as

$$\Delta > 0 \quad \text{and} \quad \frac{\Delta}{\nu q} < 1. \tag{31}$$

The next theorem gives the result about the limiting distribution of the bivariate process $\{(O_k, Z_k), k \geq 0\}$.

Theorem 3.1 *Suppose the stability condition* (31) *is satisfied. Let M be the smallest non-negative solution to*

$$M = \int_0^\infty \exp\bigl(-\nu t(I - M)\bigr)\, dD(t), \tag{32}$$

where D is from (22)*. Then the limiting distribution is given by*

$$g_m = \hat{\pi}(I - M)M^m, \quad m \geq 0, \tag{33}$$

where $\hat{\pi}$ is from (24)*.*

Proof Follows from Neuts [18], or from Theorem 4.2.1 of Neuts [19]. We omit the details. □

The simplest method of obtaining M in (32) is to use the recursion: $M(0) = 0$ (an n by n matrix of zeroes),

$$M(k+1) = \int_0^\infty \exp\bigl(-\nu t\bigl(I - M(k)\bigr)\bigr)\, dD(t), \quad k \geq 0. \tag{34}$$

Then it is known that (see Neuts [19]) $M(k)$ (rapidly) converges to M as $k \to \infty$. The simplest method of evaluating the above integral is to diagonalize $M(k)$ and write

$$M(k) = V(k)\Theta(k)U(k),$$

where $\Theta(k) = \operatorname{diag}(\theta_1(k), \theta_2(k), \ldots, \theta_n(k))$, and $U(k) = (V(k))^{-1}$. Now let $V_i(k)$ be the ith column of $V(k)$ and $U_i(k)$ be the ith row of $U(k)$. Then, assuming the eigenvalues are distinct, we have

$$\exp\bigl(-\nu t\bigl(I - M(k)\bigr)\bigr) = \sum_{i=1}^{n} e^{-\nu(1-\theta_i(k))t} V_i(k)U_i(k).$$

Substituting in (34) we get

$$M(k+1) = \sum_{i=1}^{n} V_i(k)U_i(k)\tilde{D}\bigl(\nu\bigl(1-\theta_i(k)\bigr)\bigr), \quad k \geq 0.$$

Thus, $M(k+1)$ can be easily computed from $M(k)$ using the LT $\tilde{D}(s)$ given in (28). If the eigenvalues are not distinct, the above method can still be used, but we have to use Jordan normal forms, and we need to use derivatives of $\tilde{D}(s)$ matrix. This makes the method more complicated to implement, and numerically delicate. In our examples, the eigenvalues are observed to be distinct.

3.2 The parallel case

In this case, $\{(O_k, Z_k), k \geq 0\}$ is the embedded chain in an $SM|M|\infty$ queue, and its limiting distribution can be computed using the results of Neuts [17]. This queue is stable (i.e., has nondefective limiting distribution) if the condition in (4) is satisfied. (Note that, we need this to make the semi-Markov process positive recurrent.) We also need $\nu > 0$. We find it useful to define the binomial moments of O_k in steady state as follows:

$$b_{m,j} = \lim_{k\to\infty} \mathsf{E}\left(\binom{O_k}{m} 1_{\{Z_k=j\}}\right).$$

Then

$$b_m = [b_{m,1},\ b_{m,2}, \ldots, b_{m,n}] = \sum_{r=m}^{\infty} \binom{r}{m} g_r.$$

We give the main result below:

Theorem 3.2 *The limiting distribution vectors g_m exist if the stability condition in (4) is satisfied and $\nu > 0$, and their binomial moments are given by*

$$b_m = \hat{\pi} \prod_{k=1}^{m} \tilde{D}(k\nu)\bigl(I - \tilde{D}(k\nu)\bigr)^{-1}, \quad m \geq 0. \tag{35}$$

where $\tilde{D}$ *is from* (23) *and* $\hat{\pi}$ *is from* (24). *Also*,

$$g_m = \hat{\pi} \sum_{k=0}^{\infty} (-1)^k \binom{m+k}{m} b_{m+k}. \tag{36}$$

Armed with Theorems 3.1 and 3.2, we can now analyze the limiting distribution of the trivariate process $\{(X(t), O(t), Z(t)), t \geq 0\}$ in the next section.

4 The process $\{(X(t), O(t), Z(t)), t \geq 0\}$

In this section we derive the limiting distribution of the trivariate process $\{(X(t), O(t), Z(t)), t \geq 0\}$ by utilizing the fact that it is a Markov regenerative process (MRGP) with embedded Markov renewal sequence (MRS) $\{(S_k, (O_k, Z_k)), k \geq 0\}$. (See Kulkarni [12] for the relevant definitions.) This is similar to the procedure followed by Ramaswami in [20].

Let

$$p(x, m, j) = \lim_{t \to \infty} \mathsf{P}\big(X(t) \leq x, O(t) = m, Z(t) = j\big),$$
$$x \geq 0, \quad m = 0, 1, 2, \ldots, \ j \in \Omega. \tag{37}$$

We state the results for the serial and the parallel cases separately.

Theorem 4.1 (The serial case) *Suppose the stability conditions in* (31) *holds. Then the vectors*

$$p_m(x) = \big[p(x, m, 1), \ p(x, m, 2), \ldots, p(x, m, n)\big], \quad m \geq 0$$

are given by

$$p_0(x) = \pi^q(x) - \hat{\pi}\Psi(x),$$
$$p_m(x) = g_{m-1}\Psi(x), \quad m \geq 1,$$

where

$$\Psi(x) = \frac{\Delta}{q} \int_0^{\infty} \exp\big\{-\nu t(I - M)\big\} H(t, x)\, dt,$$

where M *is as in* (32), $H(t, x)$ *is as defined in* (17), *and* $\pi^q(x)$ *is as given by* (29).

Proof We use the key renewal theorem for the Markov regenerative processes from Kulkarni [12]. Let

$$S(x, m, j) = \big\{(y, m, j) : 0 \leq y \leq x\big\},$$

and let $\alpha_{q,k,i}(x, m, j)$ be the expected time spent by the trivariate process in the set $S(x, m, j)$ over $[0, T)$ starting with $X(0) = q$, $O(0) = k$, and $Z(0) = i$, where T is

as defined in (3). Now, over the interval $[0, T)$, the $O(t)$ process is independent of the bivariate process $(X(t), Z(t))$. Using this fact we get, for $k \geq m \geq 1$,

$$\alpha_{q,k,i}(x, m, j) = \int_0^\infty e^{-\nu t} \frac{(\nu t)^{k-m}}{(k-m)!} H_{ij}(t, x)\, dt.$$

Using τ from (30) in the key renewal theorem for Markov regenerative processes, we get

$$p(x, m, j) = \frac{\Delta}{q} \sum_{r=m-1}^{\infty} \sum_{i=1}^{n} g_{r,i} \alpha_{q,r+1,i}. \tag{38}$$

In matrix form, we get

$$\begin{aligned} p_m(x) &= \frac{\Delta}{q} \sum_{r=m-1}^{\infty} \int_0^\infty e^{-\nu t} \frac{(\nu t)^{r+1-m}}{(r+1-m)!} g_r H(t, x)\, dt \\ &= \frac{\Delta}{q} \int_0^\infty e^{-\nu t} \sum_{r=m-1}^{\infty} \frac{(\nu t)^{r+1-m}}{(r+1-m)!} \hat{\pi}(I - M) M^r H(t, x)\, dt \\ &\quad \text{(From Theorem 3.1)} \\ &= \hat{\pi}(I - M) M^{m-1} \frac{\Delta}{q} \int_0^\infty e^{-\nu t} \sum_{r=m-1}^{\infty} \frac{(\nu t M)^{r+1-m}}{(r+1-m)!} H(t, x)\, dt \\ &= g_{m-1} \frac{\Delta}{q} \int_0^\infty e^{-\nu t(I-M)} H(t, x)\, dt. \end{aligned}$$

This yields the theorem. □

Next, we give a computationally useful method of computing $\Psi(x)$ when M is diagonalizable. Assume this is the case and write

$$M = ABA^{-1},$$

where $B = \text{diag}(b_1, b_2, \ldots, b_n)$ and $A = [A_1, A_2, \ldots, A_n]$, where b_i is an eigenvalue of M and A_i is the right eigenvector of M. Let C_i be the ith row of $C = A^{-1}$. Then

Theorem 4.2

$$\Psi(x) = \frac{\Delta}{q} \sum_{i=1}^{n} A_i C_i H^*\big(\nu(1 - b_i), x\big). \tag{39}$$

Proof Using

$$\exp(M) = \sum_{i=1}^{n} e^{b_i} A_i C_i$$

we get

$$\begin{aligned}\Psi(x) &= \frac{\Delta}{q}\int_0^\infty e^{-\nu t(I-M)}H(t,x)\,dt\\ &= \frac{\Delta}{q}\int_0^\infty e^{-\nu t}\sum_{i=1}^n e^{\nu b_i t}A_iC_iH(t,x)\,dt\\ &= \frac{\Delta}{q}\sum_{i=1}^n\int_0^\infty e^{-\nu(1-b_i)t}A_iC_iH(t,x)\,dt\\ &= \frac{\Delta}{q}\sum_{i=1}^n\int_0^\infty e^{-\nu(1-b_i)t}A_iC_iH(t,x)\,dt\\ &= \frac{\Delta}{q}\sum_{i=1}^n A_iC_iH^*\big(\nu(1-b_i),x\big)\end{aligned}$$

as desired. □

Since H^* is explicitly given by (19), the above theorem yields a simple method of computing $\Psi(x)$. If M is not diagonalizable (has repeated eigenvalues) the computation of $\Psi(x)$ requires the use of Jordan forms, and derivatives of the H^* matrix. This makes computation more involved and numerically delicate.

Next, we study the parallel case. Again, we denote the binomial moments as follows:

$$\beta_m(x) = \sum_{r=m}^\infty \binom{r}{m} p_r(x).$$

The main result is given in the following theorem.

Theorem 4.3 (Parallel case) *Suppose the stability condition in* (4) *holds and* $\nu > 0$. *Then the binomial moment vectors are given by*

$$\beta_m(x) = (b_m + b_{m-1})H^*\big((m+1)\nu, x\big), \quad m \geq 0,$$

where $b_{-1} = 0$ *and* b_m, $m \geq 0$, *are as given in* (35) *and* H^* *is as in* (19). *Also*,

$$p_m(x) = \hat{\pi}\sum_{k=0}^\infty (-1)^k \binom{m+k}{m}\beta_{m+k}(x), \quad m \geq 0. \tag{40}$$

Proof We use the key renewal theorem for the Markov regenerative processes from Kulkarni [12]. Let $S(x,m,j)$ be as in the proof of Theorem 4.1 and let

$$\begin{aligned}&\alpha^*_{q,k,i}(x,m,j)\\ &\quad = \mathsf{E}\left(\int_0^T \binom{O(t)}{m} 1_{\{X(t)\leq x\}} 1_{\{Z(t)=j\}} | X(0)=q, O(0)=k, Z(0)=i\right)\end{aligned}$$

where T is as defined in (3). Now, over the interval $[0, T)$, the $O(t)$ process is independent of the bivariate process $(X(t), Z(t))$. Also, given $O(0) = k$, and $T > t$, $O(t)$ is a $\text{Bin}(k, e^{-\nu t})$ random variable. It is easy to show that

$$\mathsf{E}\left(\binom{\text{Bin}(k,\, p)}{m}\right) = \binom{k}{m} p^m.$$

Using these facts we get, for $k \geq m \geq 1$,

$$\alpha^*_{q,k,i}(x, m, j) = \int_0^\infty \binom{k}{m} e^{-m\nu t} H_{ij}(t, x)\, dt = \binom{k}{m} H^*_{ij}(m\nu, x).$$

Using τ from (30),

$$\begin{aligned}
\beta_m(x) &= \frac{\Delta}{q} \sum_{r=m-1}^{\infty} \int_0^\infty \binom{r+1}{m} e^{-m\nu t} g_r H(t, x)\, dt \\
&= \frac{\Delta}{q} \sum_{r=m-1}^{\infty} \binom{r+1}{m} g_r H^*(m\nu, x) \\
&= \frac{\Delta}{q} \sum_{r=m-1}^{\infty} \left[\binom{r}{m} + \binom{r}{m-1}\right] g_r H^*(m\nu, x) \\
&= \frac{\Delta}{q} [b_m + b_{m-1}] H^*(m\nu, x).
\end{aligned}$$

This yields the theorem. The relation holds even at $m = 0$ if we define $b_{-1} = 0$. □

Since H^* is explicitly given by (19), the above theorem yields a simple method of computing the limiting joint distribution of $(X(t), O(t), Z(t))$. Using this limiting distribution and (2), we can compute the limiting distribution of $I(t)$ as given in the next theorem.

Theorem 4.4 *Let*

$$G(x) = \sum_{m=0}^{\infty} p_m(x + mq) \cdot e, \quad -\infty < x < \infty, \tag{41}$$

where e is an n by 1 vector of ones. Then

$$\lim_{t\to\infty} P\big(I(t) \leq x\big) = G(x - r). \tag{42}$$

Proof Using (37), we see that

$$\begin{aligned}
&\lim_{t\to\infty} P\big(X(t) - qO(t) \leq x\big) \\
&\quad = \sum_{m=0}^{\infty} \sum_{j=1}^{n} P\big(X(t) \leq x + qm,\, O(t) = m,\, Z(t) = j\big)
\end{aligned}$$

$$= \sum_{m=0}^{\infty} \sum_{j=1}^{n} p(x, m, j) = G(x).$$

The theorem then follows by using (2). Note that we use the results of Theorem 4.1 if we are in the serial case and those of Theorem 4.3 in the parallel case to compute $p_m(x)$ for $x \geq 0$. For $x < 0$, we set $p_m(x) = 0$. □

Using this limiting distribution of $I(t)$ and an appropriate cost model, we can compute the optimal r and q that minimize the long run average cost per unit time. The details are given in the next section.

5 The cost model

Using the limiting distribution of $I(t)$ in Theorem 4.4, we see that the limiting expected amount of inventory on hand is given by

$$\lim_{t\to\infty} E\big(I(t)^+\big) = \int_0^{\infty} \big(1 - G(x - r)\big)\, dx,$$

and the limiting expected amount of inventory on back-order is given by

$$\lim_{t\to\infty} E\big(I(t)^-\big) = \int_0^{\infty} G(r - x)\, dx.$$

Now, let α be the rate which orders are placed from the outside suppliers in steady state. Since all orders are of size q, the rate at which items are received from the outside suppliers is αq in steady state. Also, in steady state the in-house production takes place at rate $\sum p_i r_i$, and demand occurs at rate $\sum p_i d_i$, where p_i is as given in (5). The conservation of inventory implies that

$$\alpha q + \sum_{i=1}^{n} p_i r_i = \sum_{i=1}^{n} p_i d_i.$$

This implies

$$\alpha = \frac{\Delta}{q},$$

where Δ is as defined in (4). Note that, in steady state, the in-house production rate or the amount ordered from the external suppliers per unit time do not depend on the parameters r and q of the inventory management policy.

Now, assume that h is the cost to hold one item in the inventory for one unit of time; b is cost to backlog one unit of demand for one unit of time; and k is the fixed cost to place an order from an outside supplier. We do not consider production costs, or the cost of externally ordered items, because as discussed above, they are not affected by the parameters r and q. The long run cost rate is thus given by

$$c(r, q) = h \int_{-r}^{\infty} \big(1 - G(x)\big)\, dx + b \int_{-\infty}^{-r} G(x)\, dx + \frac{k\Delta}{q}.$$

Following the standard analysis in inventory models, we see that for a given q, $c(r,q)$ is a convex function of r, and it is minimized at

$$r^*(q) = -G^{-1}\left(\frac{h}{h+b}\right). \tag{43}$$

The problem of minimizing the function $c(r,q)$ thus reduces to that of minimizing the function of a single variable $c(r^*(q),q)$ to obtain the optimal order quantity q^*. Then the optimal reorder point is given by $r^* = r^*(q^*)$. This needs to be done numerically.

6 A numerical example

Consider a workshop with n independent and identical machines. Each machine stays up for an exponential amount of time with parameter μ and then fails. The repairs take an exponential amount time with rate λ. While a machine is up, it produces items continuously at rate θ, and when it is down it produces nothing. The demands occur continuously at rate nd, where $d > 0$, is a fixed parameter. Let $Z(t)$ be the number of working machines at time t. Then $\{Z(t), t \geq 0\}$ is a birth and death CTMC on $\{0, 1, \ldots, n\}$, with birth rates

$$\lambda_i = (n-i)\lambda, \quad 0 \leq i \leq n,$$

and death rates

$$\mu_i = i\mu, \quad 0 \leq i \leq n.$$

The expected number of up machines is $n\lambda/(\lambda+\mu)$ in steady state, and hence

$$\Delta = n\left(d - \theta\frac{\lambda}{\lambda+\mu}\right).$$

We assume the following values:

$$\lambda = 1, \quad \mu = 2, \quad d = 1, \quad q = 1, \quad h = 1, \quad b = 2, \quad k = 1, \quad \nu = 2.$$

We study the series and parallel cases for the following cases:

$$(n,\theta) = (1, 1.2), (2, 1.2), (3, 1.2), (1, 0.8), (2, 0.8).$$

In the first three cases, R_i is positive in some states, and negative in the others; while in the last two cases $R_i < 0$ in all states. Note that the parallel system (n,θ) is stable for all n if $0 \leq \theta < 3$. However, $(n, 1.2)$ is unstable for the series system if $n \geq 4$, and $(n, 0.8)$ is unstable for the series case for $n \geq 3$. Hence, we do not include them in our computational experiment. We also do not find the optimal q, since that is a simple numerical optimization in one variable. We assume that $q = 1$, and compute

1. r: the optimal reorder point,
2. $\mathsf{E}(O)$: the expected number of outstanding orders in steady state,

3. $\mathsf{E}(I^+)$: the expected inventory on hand in steady state,
4. $\mathsf{E}(I^-)$: the expected inventory on back order in steady state.

In order to gauge the sensitivity of the results to the assumption that the lead times are exponentially distributed with mean $1/\nu$, we use simulation to estimate the four performance measures mentioned above for three other distributions for the lead times:

1. Deterministic with mean $1/\nu$ and variance zero,
2. Erlang$(2, 2\nu)$ with mean $1/\nu$ and variance $1/2\nu^2$,
3. Hyper-exponential, a mixture of Exp(2ν) with probability $2/3$, and Exp$(\nu/2)$ with probability $1/3$. This has mean $1/\nu$ and variance $2/\nu^2$.

Note that all these distributions are chosen to have the same mean, and their variances are ordered as Var(Det) < Var(Erl) < Var(Exp) < Var(Hyp). Table 1 summarizes the above performance measures for the series system, and Table 2 for the parallel system. In each table, the performance measure for the exponential case is from the numerical calculations, while the others are from simulation. For the exponential, we do not report the simulation results since they are more or less the same as the numerical (as expected) results.

We see that each of the performance measures increases as the number of machines increases for both the parallel and the series case. This is as expected. Also, the parallel system performs better than the series system, also as expected.

Table 1 Series system

Lead-time	n	θ	r	$\mathsf{E}(O)$	$\mathsf{E}(I^+)$	$\mathsf{E}(I^-)$
Det	1	1.2	−0.06	0.299	0.266	0.085
Erl	1	1.2	−0.06	0.310	0.282	0.114
Exp	1	1.2	−0.06	0.332	0.289	0.145
Hyp	1	1.2	−0.06	0.392	0.299	0.211
Det	2	1.2	0.29	0.601	0.296	0.094
Erl	2	1.2	0.49	0.836	0.422	0.256
Exp	2	1.2	0.67	1.051	0.532	0.401
Hyp	2	1.2	0.95	1.483	0.703	0.722
Det	3	1.2	1.74	2.013	0.714	0.485
Erl	3	1.2	4.20	4.270	1.808	1.372
Exp	3	1.2	6.06	5.811	2.673	1.899
Hyp	3	1.2	11.47	11.316	5.128	2.787
Det	1	0.8	0.03	0.366	0.228	0.070
Erl	1	0.8	0.03	0.382	0.254	0.112
Exp	1	0.8	0.04	0.393	0.273	0.154
Hyp	1	0.8	0.04	0.509	0.286	0.260
Det	2	0.8	0.40	0.733	0.242	0.079
Erl	2	0.8	0.85	1.179	0.495	0.329
Exp	2	0.8	1.34	1.657	0.756	0.578
Hyp	2	0.8	2.29	2.637	1.269	1.119

Table 2 Parallel system

Lead-time	n	θ	r	$\mathsf{E}(O)$	$\mathsf{E}(I^+)$	$\mathsf{E}(I^-)$
Det	1	1.2	−0.06	0.299	0.261	0.085
Erl	1	1.2	−0.07	0.300	0.275	0.110
Exp	1	1.2	−0.08	0.300	0.281	0.126
Hyp	1	1.2	−0.09	0.300	0.283	0.138
Det	2	1.2	0.29	0.601	0.295	0.094
Erl	2	1.2	0.28	0.598	0.344	0.151
Exp	2	1.2	0.29	0.600	0.372	0.170
Hyp	2	1.2	0.28	0.602	0.378	0.190
Det	3	1.2	0.62	0.898	0.327	0.104
Erl	3	1.2	0.64	0.896	0.422	0.176
Exp	3	1.2	0.66	0.900	0.458	0.197
Hyp	3	1.2	0.66	0.897	0.477	0.213
Det	1	0.8	0.03	0.366	0.228	0.070
Erl	1	0.8	0.02	0.368	0.252	0.105
Exp	1	0.8	0.01	0.367	0.264	0.126
Hyp	1	0.8	0.00	0.365	0.273	0.143
Det	2	0.8	0.40	0.733	0.242	0.079
Erl	2	0.8	0.43	0.734	0.338	0.148
Exp	2	0.8	0.45	0.733	0.382	0.170
Hyp	2	0.8	0.45	0.730	0.405	0.190

The simulation results point out that the series system is more sensitive to the lead time distribution than the parallel system. In general, all performance measures increase as the variance of the lead time increases. Thus, having a deterministic lead time seems most beneficial. Also, the influence of the lead time distribution increases as the number of machines increases. In the parallel case, the reorder point and the expected number of outstanding orders seem to be highly insensitive. The insensitivity of the outstanding orders may be a result of the fact that it forms an infinite server queue. However, the arrival process is not Poisson, and the $SM/G/\infty$ queue is not known to be insensitive to the G. In fact, Neuts and Ramaswami [22] report that $PH/G/\infty$ queue is highly sensitive to the variability of the interarrival times. Our simulation seems to indicate that the expected queue length in the $SM/G/\infty$ queue seems insensitive to the service time distribution. In general, we can recommend that the exponential lead time analysis can be used safely in the parallel case, but in the series case one needs to be more careful about the distribution of the lead time.

Next, we plot the limiting marginal cdf of the inventory level. Figure 1 shows this for the serial case when $n = 1, 2, 3$ when $\theta = 1.2$, and Fig. 2 shows the same for the parallel case. Figures 3 and 4 show similar results for $n = 1, 2$ and $\theta = 0.8$. Note that we use the appropriate r value from Tables 1 and 2 in plotting $G(x - r)$ for the different cases. Since r is chosen to be equal to $-G^{-1}(1/3)$, all the graphs pass through the point $(0, 1/3)$ in these four figures.

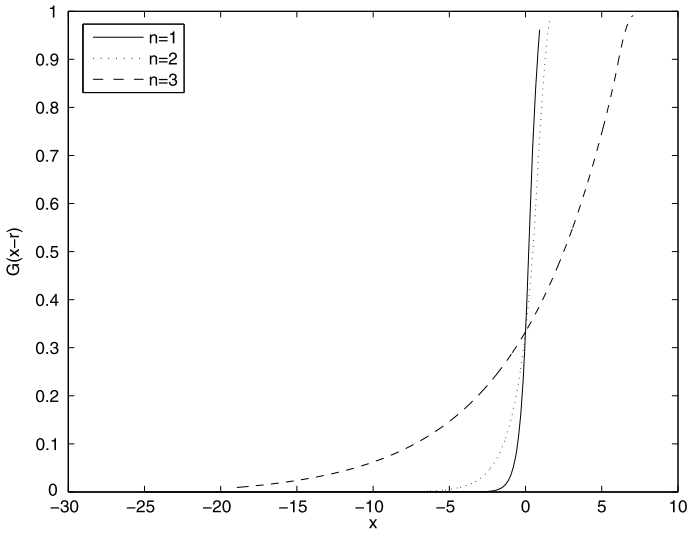

Fig. 1 Limiting cdf of the inventory level for the series case, with $\theta = 1.2$

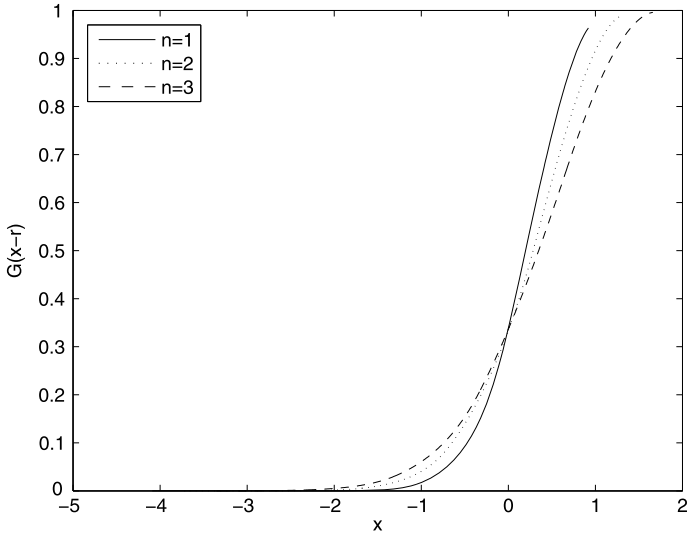

Fig. 2 Limiting cdf of the inventory level for the parallel case, with $\theta = 1.2$

Finally, we show the limiting marginal pmf of the number of outstanding orders. Figure 5 shows this for the serial case for $n = 1, 2, 3,$ and $\theta = 1.2$, and Fig. 6 shows the same for the parallel case. Figures 7 and 8 show similar results for $n = 1, 2$ and $\theta = 0.8$.

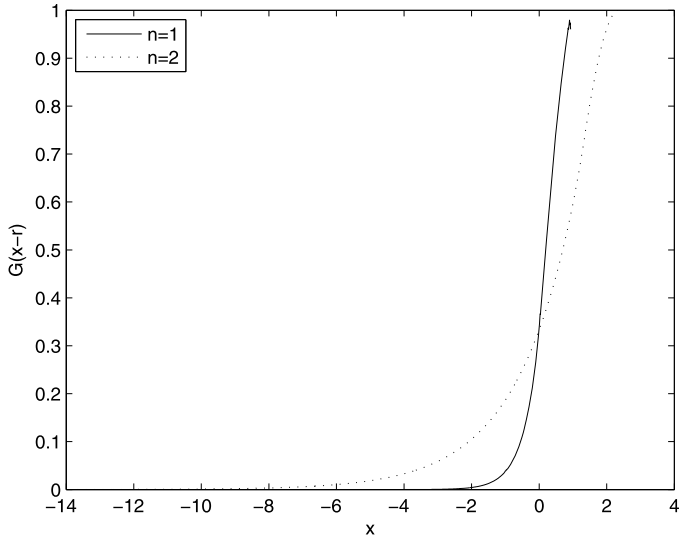

Fig. 3 Limiting cdf of the inventory level for the series case, with $\theta = 0.8$

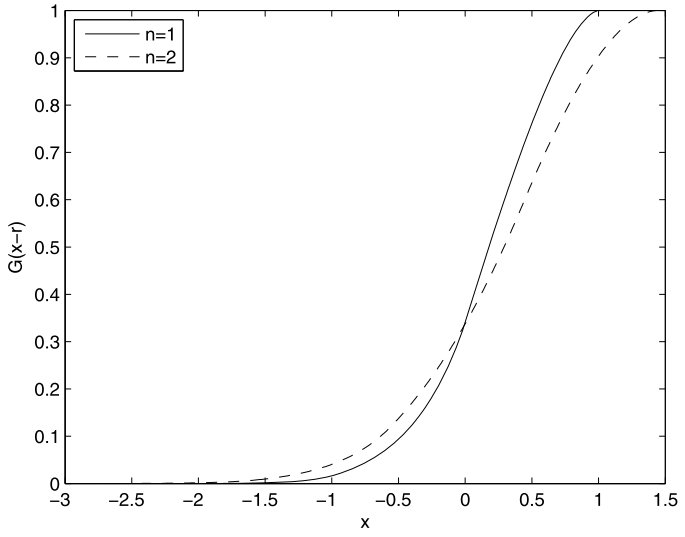

Fig. 4 Limiting cdf of the inventory level for the parallel case, with $\theta = 0.8$

7 Conclusions and extensions

In this paper we have developed procedures to compute the limiting joint distribution of the inventory level and the number of outstanding orders for a fluid inventory system when the production and demands are modulated by an external stochastic environment and the lead times are stochastic. The analysis uses the $SM|M|1$ and

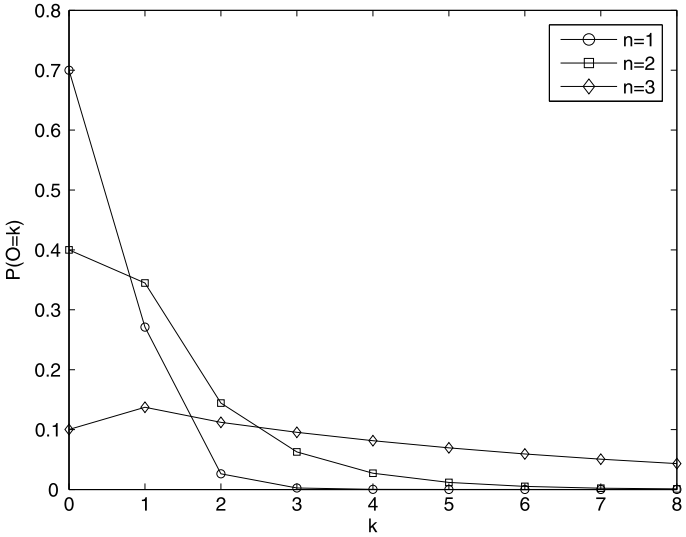

Fig. 5 Limiting pmf of the outstanding orders for the series case, with $\theta = 1.2$

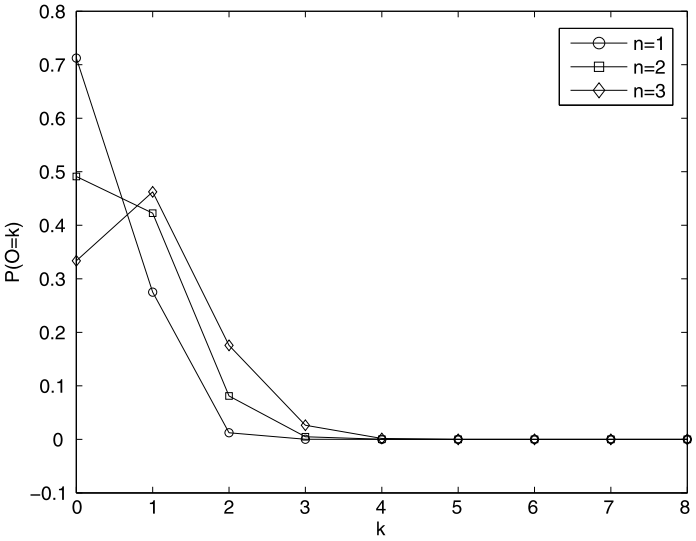

Fig. 6 Limiting pmf of the outstanding orders for the parallel case, with $\theta = 1.2$

$SM|M|\infty$ queues as the building blocks. Our simulation results indicate that the serial system is much more sensitive to the assumption of exponentially distributed lead times than the parallel system. We discuss several possible extensions below.

State dependent order sizes One of the main restrictions of our analysis is that the order size is always q. It would be interesting to extend this analysis to the case when the order size is allowed to depend on the state of the environment at the time of

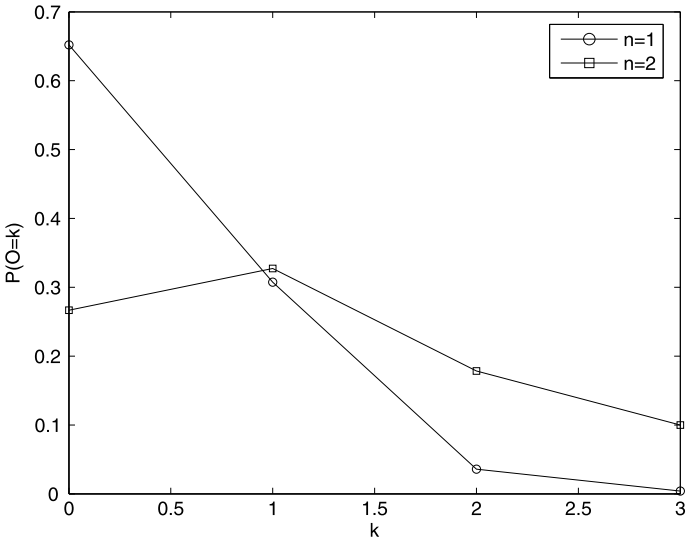

Fig. 7 Limiting pmf of the outstanding orders for the series case, with $\theta = 0.8$

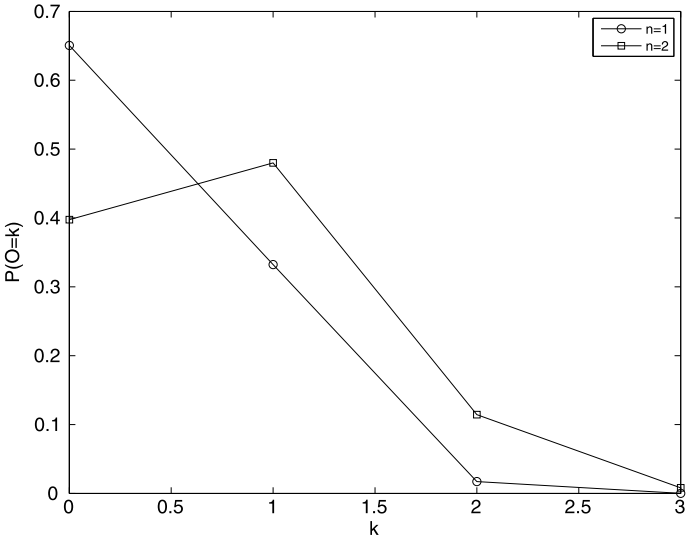

Fig. 8 Limiting pmf of the outstanding orders for the parallel case, with $\theta = 0.8$

order placement. Although it is possible to formulate this as a fluid process modulated by a multidimensional environment process that keeps track of number of orders of different sizes, the analysis promises to be challenging.

Continuous order filling Our model assumes that when we place an order of size q, it is delivered in its entirety after a random amount of time. What if the order delivery

is continuous, rather than lumpsum? There are several ways to model this situation. We shall consider the following. Suppose that we have a standby production facility that can be used to add to the inventory at a fixed rate, say θ_0. When the inventory level hits a preset level r, we turn on this standby production facility, and keep it running until the inventory level reaches $r+q$, where $q>0$ is a fixed constant. Then we turn the standby facility off until the inventory level falls to r. This (r,q) policy will produce a stable system if $\theta_0 > \Delta$, where Δ is given by (4). Note that there are no lead times in this model. The steady state distribution of the (X,Z) process can be computed by using the Markov regenerative process analysis by considering the system as going though two alternating phases: in phase one the standby facility is on, and in phase two it is off. All the required building blocks for this analysis are already developed in Sect. 2, and the analysis promises to be straight forward.

Lost sales Our analysis assumes backlogging. It will be useful to do a similar analysis assuming lost sales. However, under the assumption of lost sales, the (X,Z) process studied in this paper is no longer a fluid process with jumps, and we are forced to study the trivariate process (X,O,Z) directly. This promises to be a challenging extension.

Acknowledgements We thank the two referees whose thorough reading and valuable suggestions have improved this paper substantially.

References

1. Ahn, S., Ramaswami, V.: Fluid flow models and queues—a connection by stochastic coupling. Commun. Stat., Stoch. Models **19**, 325–348 (2003)
2. Anick, D., Mitra, D., Sondhi, M.M.: Stochastic theory of a datahandling system with multiple sources. Bell Syst. Tech. J. **61**, 1871–1894 (1982)
3. Ahn, S., Ramaswami, V.: Efficient algorithms for transient analysis of stochastic fluid flow models. J. Appl. Probab. **42**, 531–549 (2005)
4. Asmussen, S.: Stationary distributions for fluid flow models with or without Brownian noise. Commun. Stat., Stoch. Models **11**, 21–49 (1995)
5. Berman, O., Perry, D.: An EOQ model with state dependent demand rate. Eur. J. Oper. Res. **176**, 255–272 (2006)
6. Beyer, D., Cheng, F., Sethi, S.P., Taksar, M.: Markovian Demand Inventory Models. Springer, Berlin (2010)
7. Browne, S., Zipkin, P.: Inventory models with continuous, stochastic demands. Ann. Appl. Probab. **1**, 419–435 (1991)
8. Galliher, H.P., Morse, P.M., Simond, M.: Dynamics of two classes of continuous-review inventory systems. Oper. Res. **7**, 362–384 (1959)
9. Hadley, G., Whitin, T.: Analysis of Inventory Systems. Prentice-Hall, Englewood Cliffs (1963)
10. Kaplan, R.S.: A dynamic inventory model with stochastic lead times. Manag. Sci., Theor. Ser. **16**, 491–507 (1970)
11. Kulkarni, V.G., Yan, K.: A fluid model with upward jumps at the boundary. Questa **56**, 103–117 (2007)
12. Kulkarni, V.G.: Modeling and Analysis of Stochastic Systems, 2nd edn. CRC Press, Boca Raton (2009)
13. Ledermann, W., Reuter, G.E.H.: In: Spectral Theory for the Differential Equations of Simple Birth and Death Processes. Philosophical Transactions of the Royal Society: Mathematical, Physical, and Engineering Sciences, vol. 246, pp. 312–369 (1954)
14. Mitra, D.: Stochastic theory of fluid models of multiple failure-susceptible producers and consumers coupled by a buffer. Adv. Appl. Probab. **20**, 646–676 (1988)

15. Nahmias, S.: Simple approximations for a variety of dynamic lead-time lost-sales inventory models. Oper. Res. **27**, 904–924 (1979)
16. Narayanan, A., Kulkarni, V.G.: First Passage times in fluid models with applications to two-priority fluid systems. In: Proceedings of the IPDS'96, Urbana-Champaign, pp. 166–175 (1996)
17. Neuts, M.F., Chen, S.Z.: The infinite server queue with semi-Markovian arrivals and negative exponential services. J. Appl. Probab. **9**, 178–184 (1972)
18. Neuts, M.F.: Markov chains with applications in queueing theory, which have a matrix geometric invariant vector. Adv. Appl. Probab. **10**, 185–212 (1978)
19. Neuts, M.F.: Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach. Johns Hopkins University Press, Baltimore (1981)
20. Ramaswami, V.: Algorithms for a continuous-review (s, S) inventory system. J. Appl. Probab. **18**, 461–472 (1981)
21. Ramaswami, V.: Passage times in fluid models with application to risk processes. Methodol. Comput. Appl. Probab. **8**, 497–515 (2006)
22. Ramaswami, V., Neuts, M.F.: Some explicit formulas and computational methods for infinite-server queues with phase- type arrivals. J. Appl. Probab. **17**, 498–514 (1980)
23. Song, J., Zipkin, P.: Inventory control in a fluctuating demand environment. Oper. Res. **41**, 351–370 (1993)
24. Stidham, S., Jr.: Cost models for stochastic clearing systems. Oper. Res. **21**, 100–127 (1979)
25. Tanaka, T., Hashida, O., Takahashi, Y.: Transient analysis of fluid model for ATM statistical multiplexer. Perform. Eval. **23**, 145–162 (1995)
26. Whitt, W.: The stationary distribution of a stochastic clearing process ward Whitt. Oper. Res. **29**, 294–308 (1981)
27. Yan, K.: Fluid models for productio-inventory systems, Ph.D. Thesis, Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599 (2007)
28. Yan, K., Kulkarni, V.G.: Optimal inventory policies under stochastic production and demand rates. Stoch. Models **24**, 173–190 (2008)
29. Zipkin, P.: Stochastic leadtimes in continuous-time inventory models. Nav. Res. Logist. Q. **33**, 763–774 (1986)
30. Zipkin, P.: Foundations of Inventory Management. McGraw-Hill, Boston (2000)