IIE Transactions (2000) **32**, 409–419

Performance measures of a make-to-order inventory-production system

QI-MING HE[1] and E.M. JEWKES[2]

[1]*Department of Industrial Engineering, DalTech, Dalhousie University, Halifax, Nova Scotia, Canada B3J 2X4*
E-mail: Qi-Ming.He@dal.ca
[2]*Department of Management Sciences, University of Waterloo, Waterloo, Canada N2L 3G1*
E-mail: emjewkes@mansci.uwaterloo.ca

Received November 1997 and accepted May 1999

This paper develops two algorithms for computing the average total cost per product and other performance measures for a make-to-order inventory-production system. The two algorithms are developed by using matrix analytic methods. The first algorithm is based on the matrix-geometric solution of the Quasi-Birth-and-Death (QBD) Markov process. The second algorithm is based on the fundamental period of the QBD Markov process. The advantages and disadvantages of the two algorithms are discussed.

1. Introduction

This paper deals with a simple supply chain that is modeled as an *inventory-production system* consisting of a warehouse and a workshop (Fig. 1). Demands from customers are accepted at the workshop and products are produced in the workshop to satisfy customer demands. Raw materials used in production are supplied by an outside supplier through the warehouse to the workshop.

One issue in this make-to-order inventory-production system is how to reduce the inventory cost of the raw materials in the system. This is related to two types of problems: (i) how to find the optimal replenishment policy with respect to the average total cost per product; and (ii) how to compare the efficiency of different replenishment policies or how to analyze a system with a particular replenishment policy. For the first problem, algorithms based on Markov decision processes were developed in He *et al.* [1], that allowed the computation of the optimal replenishment policy and some nearly optimal policies. For the second problem two algorithms are developed in this paper that allow the computation of the average total cost per product and other performance measures for any feasible replenishment policy.

The inventory-production model of interest is a special two-echelon model. Typical inventory control policies used in continuous review two-echelon models are "order up to", $(S-1, S)$, or (r, Q). Performance analysis of such models has been carried out by Axsater [2,3], Buzacott and Shanthikumar [4], Federgruen [5], Mitra and Mitrani [6], and Zipkin [7]. This paper considers a set of replenishment policies that utilize information about the inventory status as well as information about production status (Equation (1)). The results are useful in analyzing the value of information used in inventory control [8].

The inventory-production model is also related to queueing models and inventory models since it is a combined queueing and inventory model. Some results from the $M/M/1$ queueing system [9] and the stochastic EOQ model [10] are used in this paper. Although the $M/M/1$ queue and the stochastic EOQ model are simple and well studied, their combined model is complicated and is unexplored.

The main mathematical tools used in this paper are matrix analytic methods. One of the main features of matrix analytic methods is that numerically tractable solutions can be obtained for problems where solutions would be otherwise unavailable or much more difficult to obtain. This explains why matrix analytic methods are utilized in this paper. Readers are referred to Neuts [11, 12] for more details on matrix analytic methods and Chakravarthy and Alfa [13] for details of recent developments in matrix analytic methods.

This paper gives detailed explanations of the methods used and the results obtained. The reason for doing so is to ensure that readers with a limited knowledge of matrix analytic methods can follow the solution process. This paper illustrates how useful matrix analytic methods can be in analyzing certain stochastic models.

The rest of the paper is organized as follows. In Section 2, the inventory-production model with zero leadtimes is introduced in more detail. The feasible replenishment policy set is also defined. Section 3 constructs the QBD Markov process for each feasible replenishment policy.

0740-817X © 2000 "IIE"

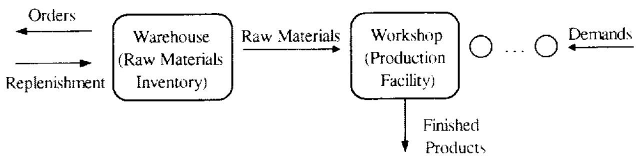

Fig. 1. The make-to-order inventory-production system.

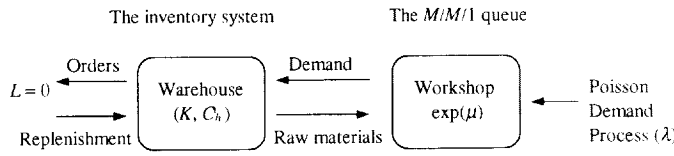

Fig. 2. The inventory-production system with zero leadtimes.

Section 4 then studies the stationary distribution of the inventory-production system. An algorithm for computing the average total cost per product is developed based on the matrix-geometric solution of the QBD Markov process. Sections 5 and 6 study the total costs incurred in a busy cycle and an algorithm for computing the average total cost incurred in a busy cycle is developed. Finally in Section 7, the results obtained in this paper are summarized and a brief discussion is given to some extensions and future research directions.

2. Model definition

The inventory-production system of interest in this paper is defined as follows. The workshop manufactures make-to-order products based on customer demands. Customer demands arrive one at a time to the workshop according to a Poisson process with parameter λ. All demands are processed in the workshop by a single machine in batch sizes of one. Production (or processing) times have a common exponential distribution with parameter μ. When the machine is ready to process a customer order, a call for a unit of raw materials is sent to the warehouse. A unit of raw materials is immediately sent to the workshop and production on that unit begins. The transportation time between the warehouse and the workshop is assumed negligible. Raw materials in the warehouse are replenished from a supplier according to a continuous review replenishment policy. Replenishment leadtimes of raw materials are assumed to be zero so that production in the workshop can occur whenever there are demands. The replenishment policy for raw materials is such that it does not allow raw material shortages, i.e., production will occur whenever there are demands in the workshop.

The cost components of the inventory-production system are as follows: there is a fixed ordering cost K associated with each order from the warehouse to the supplier, regardless of the order size, and the holding cost is C_h per unit of raw materials held in the system per unit time.

According to the above definition, leadtimes at both the supplier and the warehouse are zero. No shortage of raw materials is allowed at the supplier, the warehouse, or the workshop. All demands will be processed. Thus, the inventory-production system can be decomposed into two subsystems: an $M/M/1$ queue (the workshop) and an inventory system (the warehouse) (Fig. 2). The workshop is modeled as an $M/M/1$ queue [9] since no shortage of raw materials is allowed and the queueing process is not influenced by raw material replenishment. Without loss of generality, it is assumed that all demands are processed on a first-come-first-served basis. The warehouse is modeled as an inventory system with zero replenishment leadtimes and demands that come from the workshop when the workshop begins to produce new products.

The status of the $M/M/1$ queue at time t is represented by the number of customers (the number of unfilled demands or the queue length) in the workshop, denoted by $q(t)$. It is assumed that $\rho = \lambda/\mu < 1$ so that the $M/M/1$ queue is ergodic. Let $I(t)$ be the total number of units of raw materials in the inventory-production system at time t, i.e., the number of units of raw materials in the warehouse plus the one unit of raw materials in the workshop, if the workshop is working. Then the status of the inventory-production system can be represented by $(q(t), I(t))$ at time t.

Raw material replenishment in the warehouse is controlled according to a specific replenishment policy. This replenishment policy determines when and how much of the raw materials to order from the outside supplier. In this paper, only replenishment policies based on system status $(q(t), I(t))$ are considered. This implies that both the queue length $q(t)$ and the inventory level $I(t)$ are reviewed continuously. Since raw material leadtimes are zero, it makes no sense to order raw materials when $I(t)$ is positive. Thus, a replenishment policy is a function of the number of customers $q(t)$ only and can be represented as a vector $\boldsymbol{\pi} = (\pi(0), \pi(1), \pi(2), \ldots)$, where $\pi(n)$ is the order size when *the inventory level is zero* and the number of customers is n. At time t, if $I(t) = 0$ and $\pi(q(t)) > 0$, an order of size $\pi(q(t))$ is issued and filled; otherwise, no action takes place. If an order of the size $\pi(q(t))$ is filled at time t, the inventory level becomes $\pi(q(t))$, i.e., $I(t+) = \pi(q(t))$.

He *et al.* [1] have shown that the optimal replenishment policy (which minimizes the average total cost per product) has the following structural property: $\pi^*(q) = \pi^*_\infty$ for all $q \geq q^*$, where π^*_∞ and q^* are finite positive integers. This property says that the optimal order sizes are finite and equal to a finite number when the queue length is large enough. Thus, only replenishment policies that have a finite tail order size, such as the optimal replenishment policy, are considered. The set of feasible replenishment policies under consideration in this paper is defined as

$$\Pi = \{\pi : \pi(0) \geq 0, \pi(q) \geq 1, 1 \leq q < \infty, \pi(q) \equiv \pi_\infty > 0, \text{ for all } q \geq q^* > 0, \pi_\infty \text{ and } q^* \text{ are finite}\}. \quad (1)$$

Despite many restrictions, set Π is still rather general and includes many of the replenishment policies of special interest to the inventory-production system [8]. The next example shows what a typical replenishment policy $\pi \in \Pi$ looks like.

Example 1. Consider an inventory-production system with system parameters $\lambda = 0.3$, $\mu = 1$, $K = 30$, and $C_h = 1$. The optimal replenishment policy π^* (where the average total cost per product is minimized) is shown in Table 1.

For instance, when the optimal replenishment policy π^* is applied, if the queue length is three when the inventory level of the system becomes zero, an order of size six is issued and filled.

Based on system assumptions, it is easy to see that, for any replenishment policy $\pi \in \Pi$, the corresponding stochastic process $(q(t), I(t))$ is a Markov process. The objective of this paper is to analyze performance measures (especially the average total cost per product) by exploring the Markov process $(q(t), I(t))$ for any $\pi \in \Pi$.

3. The quasi-birth-and-death Markov process

Consider an inventory-production system with a replenishment policy $\pi \in \Pi$. This section defines the Markov process $(q(t), I(t))$ explicitly.

It is easy to see that the stochastic process $(q(t), I(t))$ is a two-dimensional Markov process. Since the inventory level $I(t)$ is uniformly finite, the process $(q(t), I(t))$ is essentially a one-dimensional Markov process. To construct the state space and the infinitesimal generator of the Markov process $(q(t), I(t))$, a function π^e, which is called the envelope function of π, is introduced first. Let

Table 1. The optimal replenishment policy of Example 1

q (queue length)	$\pi^*(q)$
0	0
1	4
2	5
3	6
4	7
5	8
6	8
7	9
8	8
9	7
10	7
(≥)11	8

$$\pi_{\max} = \max_{\{q \geq 0\}}\{\pi(q)\} \quad \text{and} \quad q_{\max} = \min_{\{q \geq 0\}}\{q : \pi(q) = \pi_{\max}\}. \quad (2)$$

$\pi_{\max}$ is the maximum order size of π and $q_{\max}$ is the least queue length where the order size is $\pi_{\max}$. The envelope π^e is defined recursively as

$$\pi^e(q) = \begin{cases} \pi_{\max}, & q \geq q_{\max}; \\ \max\{\pi(0), \pi(1), \ldots, \pi(q), \pi^e(q+1) - 1\}, & 0 \leq q < q_{\max}. \end{cases} \quad (3)$$

While π may fluctuate dramatically, π^e is a nondecreasing function. Note that, while the envelope function π^e is introduced for a better presentation, it is also useful in numerical computation. An example of π and π^e is shown in Fig. 3 for the optimal replenishment policy of Example 1. It can be proved that π^e is the minimum nondecreasing function which is no less than π pointwise.

For $(q(t), I(t))$, intuitively, when the state (q, i) is reachable, the state $(q + j, i)$, for $j > 0$, is also reachable since there could be any number of arrivals before the completion of a product. Since any number of products could be completed before an arrival occurs, the state $(q - j, i - j)$ is reachable for $0 < j \leq \min\{q, i\}$. The state space of $(q(t), I(t))$ is then given by $S = \cup_{q=0}^{\infty} S_q$, where

$$S_0 = \{(0, I\{\pi(0) \neq 0\}), \ldots, (0, \pi^e(0))\},$$
$$S_q = \{(q, 1), \ldots, (q, \pi^e(q))\}, \quad q \geq 1,$$

and $I\{\cdot\}$ is the indicator function. The subset S_q in which all states have a queue length q is called *level* q. The number of states in each level is nondecreasing since the envelope π^e is nondecreasing. Any state in the state space S can reach any other states in S. It is clear that the queue length $q(t)$ can increase or decrease at most by one at a transition so $(q(t), I(t))$ is a quasi-birth-and-death (QBD) Markov process. This leads to the following theorem.

Theorem 1. *$(q(t), I(t))$ is a QBD Markov process and is irreducible for any feasible policy π in Π defined in Equation (1).*

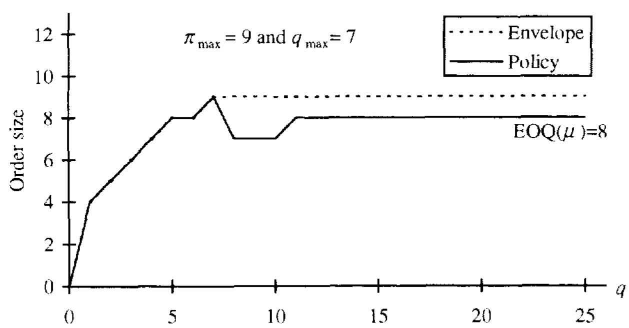

Fig. 3. The envelope of an optimal replenishment policy.

Transitions of $(q(t), I(t))$ can be categorized into the following three types:

(i) A transition from state (q, i) to $(q+1, i)$ occurs when a demand arrives before the completion of the product in service, if any.
(ii) A transition from (q, i) to $(q-1, i-1)$, for $q > 0$ and $i > 0$, occurs when a product is completed before the next arrival.
(iii) A transition from $(q, 1)$ to $(q-1, \pi(q-1))$, for $q > 0$, takes place when a product is completed before the next arrival and an order is issued and filled.

The case $q = 0$ is slightly different and has only transitions of type (i). A replenishment may occur when the queue length goes from zero to one if $\pi(0) = 0$. Then the infinitesimal generator of $(q(t), I(t))$ is given by

$$\mathbf{Q} = \begin{pmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,1} & & & & & \\ \mathbf{A}_{1,0} & \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \mathbf{A}_{q^*,q^*-1} & \mathbf{A}_{q^*,q^*} & \mathbf{A}_{q^*,q^*+1} & & \\ & & & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \\ & & & & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 \\ & & & & & \ddots & \ddots & \ddots \\ & & & & & & \ddots & \ddots \end{pmatrix}, \tag{4}$$

where all the matrix blocks are given explicitly in Appendix A and all other elements are zero. The constant $q^* = \min\{q : \pi(n) = \pi_\infty, \text{ for all } n \geq q\}$.

Let $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots)$ be the stationary distribution of $(q(t), I(t))$, where $\mathbf{x}_q = (x_{q,1}, x_{q,2}, \ldots, x\_\{q, \pi^e(q)\})$ (Notice that a_b denotes a subscript b for typographical clarity) and

$$x_{q,i} = \lim_{t\to\infty} \mathrm{P}\{q(t) = q, I(t) = i\}, \quad \text{for} \quad 1 \leq i \leq \pi^e(q), \ q \geq 0. \tag{5}$$

Notice that $x_{0,0}$ is defined in a similar way when $\pi(0) = 0$. The limit exists when the Markov process is positive recurrent. The stationary distribution vector $\mathbf{X}$ satisfies $\mathbf{XQ} = 0$, $\mathbf{Xe} = 1$ and $\mathbf{X} > 0$, where $\mathbf{e}$ is a vector with all components 1.

Let $\boldsymbol{\alpha}$ be the marginal stationary distribution of the inventory level. Then $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha\_\pi_{\max})$, where

$$\alpha_i = \lim_{t\to\infty} \mathrm{P}\{I(t) = i\}, \quad 0 \leq i \leq \pi_{\max}, \tag{6}$$

i.e., the probability that at an arbitrary epoch the inventory level is i.

The literature on stationary distributions of QBD Markov processes is extensive (see for example Neuts [11], Ramaswami and Taylor [14], and references therein). However no general explicit solution is available. For the QBD Markov process $(q(t), I(t))$ of interest, two algorithms will be developed to allow the computation of the stationary distribution and other performance measures. To end this section, a special case where explicit solutions can be found is presented.

Example 2. Consider a feasible replenishment policy $\pi = (l, l, \ldots)$ where l is a positive integer. This replenishment policy is special since the order size is always l (independent of the queue length at the order epochs). This replenishment policy is used when no information about the queue length is available for inventory control in the warehouse. (Strictly speaking, when no information about the queue length is available, an order should be placed when $I(t)$ becomes one since it implies zero inventory in the warehouse). This explains why replenishment policies with $\pi(0) > 0$ are included in set $\mathbf{\Pi}$ (Equation (1)). It can be verified that the stationary distribution is given by $\mathbf{x}_q = \rho^q(1-\rho)(1/l, \ldots, 1/l)$, $q \geq 0$, when $\rho = \lambda/\mu < 1$. The distribution of the inventory level at an arbitrary time is given as

$$(\alpha_1, \alpha_2, \ldots, \alpha_l) = \sum_{q=0}^{\infty} \mathbf{x}_q = \left(\frac{1}{l}, \frac{1}{l}, \ldots, \frac{1}{l}\right). \tag{7}$$

Therefore, the inventory level is uniformly distributed over $\{1, 2, \ldots, l\}$. Let $g_{-1}(l)$ be the average total cost per product with respect to order size l. By Equation (7) and direct calculations,

$$g_{-1}(l) = \frac{K}{l} + \frac{(l+1)C_h}{2\lambda}. \tag{8}$$

This special case is important not only because explicit solutions can be found for some performance measures, but also because the replenishment policy uses the least information about the queue length in inventory replenishment. It is the base case in analyzing the value of information about the queue length [8].

4. The matrix-geometric solution

This section presents the matrix-geometric solution of the QBD Markov process $(q(t), I(t))$. The stationary joint distribution of the inventory level and the queue length is found first. Then the stationary distribution of the inventory level is derived. A probabilistic argument leads to expressions for the average holding cost, average ordering cost, and average total cost per product (or per unit time). Some of the mathematical details or proofs will not be shown since much of the theory has been established in Neuts [11,12], and Ramaswami and Taylor [14].

First, the ergodicity condition of the Markov process $(q(t), I(t))$ is $\lambda < \mu$. This is obvious since the Markov process is irreducible. $I(t)$ is uniformly finite and $q(t)$ is ergodic when $\lambda < \mu$. The matrix-geometric solution of $\mathbf{X}$ is given by [11]:

$$\mathbf{x}_{q+1} = \mathbf{x}_q \mathbf{R}_q, \quad 0 \le q < q^*,$$
$$\mathbf{x}_{q+1} = \mathbf{x}_q \mathbf{R}, \quad q \ge q^*, \tag{9}$$

where $\mathbf{x}_0$ is the unique solution to equations

$$\mathbf{x}_0(\mathbf{A}_{0,0} + \mathbf{R}_0\mathbf{A}_{1,0}) = \mathbf{x}_0(-\lambda\mathbf{I} + \mathbf{R}_0\mathbf{A}_{1,0}) = 0,$$
$$\mathbf{x}_0\left[\mathbf{I} + \sum_{n=0}^{q^*-1}\left(\prod_{q=0}^{n}\mathbf{R}_q\right) + \sum_{n=q^*+1}^{\infty}\left(\prod_{q=0}^{q^*-1}\mathbf{R}_q\right)\mathbf{R}^{n-q^*}\right]\mathbf{e} = 1$$
$$\text{or} \quad \mathbf{x}_0\mathbf{e} = 1 - \rho, \tag{10}$$

matrix $\mathbf{R}$ is the minimal nonnegative solution to the quadratic matrix equation

$$\mathbf{R}^2\mathbf{A}_2 + \mathbf{R}\mathbf{A}_1 + \mathbf{A}_0 = 0, \tag{11}$$

and for $0 \le q < q^*$, matrix $\mathbf{R}_q$ is the minimal nonnegative solution to the equation

$$\mathbf{R}_q\mathbf{R}_{q+1}\mathbf{A}_{q+2,q+1} + \mathbf{R}_q\mathbf{A}_{q+1,q+1} + \mathbf{A}_{q,q+1} = 0. \tag{12}$$

$\mathbf{R}_q$ is a $\pi^e(q) \times \pi^e(q+1)$ matrix. $\mathbf{I}$ is the identity matrix. It is easy to see that

$$\mathbf{R}_q = -\mathbf{A}_{q,q+1}(\mathbf{A}_{q+1,q+1} + \mathbf{R}_{q+1}\mathbf{A}_{q+2,q+1})^{-1}. \tag{13}$$

Since $\mathbf{R}_{q^*} = \mathbf{R}$, then all $\mathbf{R}_q, 0 \le q < q^*$, can be obtained when $\mathbf{R}$ is known. The invertibility of the matrix in the above equation can be justified when the Markov process $(q(t), I(t))$ is positive recurrent, i.e., $\lambda < \mu$. The eigenvalue with the largest real part of $\mathbf{R}$, denoted as $sp(\mathbf{R})$, is less than 1 when the Markov process is positive recurrent.

In terms of $\{\mathbf{x}_q, q \ge 0\}$, the stationary distribution of the inventory level is given by

$$\boldsymbol{\alpha} = \left(\alpha_0, \hat{\mathbf{x}}_0 + \sum_{q=1}^{q^*-1}\mathbf{x}_q + \mathbf{x}_{q^*}(\mathbf{I} - \mathbf{R})^{-1}\right), \tag{14}$$

where $\alpha_0 = x_{0,0}$ and $\hat{\mathbf{x}}_0 = (x_{0,1}, \ldots, x_{\{0,\pi^e(0)\}})$ when $\pi(0) = 0$; $\alpha_0 = 0$ and $\hat{\mathbf{x}}_0 = \mathbf{x}_0$ when $\pi(0) > 0$. The vectors $\{\mathbf{x}_q\}$ may be of different dimensions. The summation in Equation (14) means that elements with the same second index are added together. The average inventory at an arbitrary time can be calculated using $\Sigma i\alpha_i$. Unfortunately, a closed form formula for the average inventory level is difficult to derive, except for a few special cases (e.g., Example 2).

Now, the focus turns to cost related measures. Assuming that the system is in a steady state, consider the costs incurred in the interval $(t, t+\delta t)$. An ordering cost K is incurred when the inventory level hits zero and the queue length is q with $\pi(q) > 0$ (i.e., a product is completed and it leaves a system with no inventory and a queue length q with $\pi(q) > 0$ behind). For cases with $\pi(0) = 0$, an ordering cost K is incurred when an arrival finds no other demand and no inventory of raw materials in the system. The probability that a product is completed in $(t, t+\delta t)$ is given by $\delta t\mu + o(\delta t^2)$ [15], given that production is going on at time t. The probability that a demand arrives in $(t, t+\delta t)$ is given by $\delta t\lambda + o(\delta t^2)$. Then the ordering cost in $(t, t+\delta t)$ is given by

$$K\delta t\mu\left[\sum_{q=2}^{\infty}x_{q,1} + I\{\pi(0) > 0\}x_{1,1}\right]$$
$$+ K\delta t\lambda I\{\pi(0) = 0\}x_{0,0} + o(\delta t^2). \tag{15}$$

If the inventory level at t is i, then the holding cost in $(t, t+\delta t)$ is approximately $i\delta tC_h$. Therefore, the holding cost in $(t, t+\delta t)$ is given by

$$C_h\delta t\sum_{i=0}^{\pi_{\max}} i\alpha_i + o(\delta t^2). \tag{16}$$

Theorem 2. *The average total cost per product for replenishment policy* $\boldsymbol{\pi}(\in \boldsymbol{\Pi})$ *is given by*

$$g(\boldsymbol{\pi}) = \frac{1}{\lambda}\left[C_h\sum_{i=0}^{\pi_{\max}} i\alpha_i + K\mu[\alpha_1 - I\{\pi(0) = 0\}x_{1,1} - x_{0,1}]\right.$$
$$\left.+ K\lambda I\{\pi(0) = 0\}x_{0,0}\right]. \tag{17}$$

Proof. Adding the ordering and the holding costs given by Equations (15) and (16) together and dividing the summation by δt, the average total cost per unit time is obtained. Since the average number of arrivals per unit time is λ, the average total cost per product is obtained by dividing the average total cost per unit time by λ. ■

In summary, the computation of the stationary distribution $\mathbf{X}$ and related performance measures can be done in the following six steps.

A R-related algorithm

Step 1. Input system parameters $\{\lambda, \mu, K, C_h\}$ and the replenishment policy $\boldsymbol{\pi}$.

Step 2. Construct the envelope function $\boldsymbol{\pi}^e$ of $\boldsymbol{\pi}$ (Equation (3)).

Step 3. Construct the infinitesimal generator $\mathbf{Q}$ (Appendix A).

Step 4. Compute the matrix $\mathbf{R}$ and $\mathbf{R}_q$, $0 \le q \le q^*$ (Equations (11) and (13)).

Step 5. Compute vector $\mathbf{x}_0$ and the stationary distribution $\mathbf{X}$ (Equations (9) and (10)).

Step 6. Compute other performance measures (Equations (14) to (17)).

Steps 1, 2, 3, and 6 can be easily implemented since only simple algebra is involved. Step 4 involves finding the minimal nonnegative solution to a matrix quadratic equation. Existing algorithms for computing $\mathbf{R}$ can be found in Neuts [11] and Latouche and Ramaswami [16]. Step 5 requires solving a linear equation. Elsner's method [11] can be used to solve this equation.

The above algorithm for computing the performance measures is based on the stationary distribution of the system. The algorithm is efficient when the maximal order size of a replenishment policy is not "too large" (depending on the computer in use), but may not be so for an arbitrary replenishment policy. The reason is that all the matrices $\mathbf{R}$ and $\{\mathbf{R}_q, 0 \leq q \leq q_{\max}\}$ have to be stored for computing the stationary distribution. Another concern with this approach is that it produces no result about the costs incurred in busy periods. In the next two sections, the costs incurred in a busy cycle are studied.

5. Costs Incurred in a busy cycle

The second part of the paper derives the distribution (or the Laplace Stieltjes Transform) of the total inventory costs incurred in a fundamental period. As a result, the total costs incurred in a busy cycle can be obtained. This leads to expressions for the average total (or holding, or ordering) cost per product (or per unit time). To study the costs incurred in a busy cycle, the concept of fundamental period is used.

Neuts [11] has defined *a fundamental period* as the first passage time in which the Markov process $(q(t), I(t))$ reaches level q, given that the Markov process starts in level $q+1$. It is easy to see that the *busy period* is a special type of the fundamental period where the Markov process goes from level 1 to level 0 for the first time. A *busy cycle* is the time period between two consecutive returns to level zero.

Let u and v be random variables with exponential distributions with parameters λ and μ respectively. Random variables u and v are the generic random variables for the interarrival time and the production time respectively. Let $\tau(q,i)$ be the time the Markov process $(q(t), I(t))$ stays in state (q,i) before the next transition. Let $C(q,i)$ be the total cost incurred during the time $\tau(q,i)$. Then

$$\tau(q,i) = \min\{u, v\}, \ \mathrm{E}\exp\{-s\tau(q,i)\} = \frac{s}{s+\lambda+\mu}, \ q, i \geq 1. \tag{18}$$

$$C(q,i) = \begin{cases} iC_h\tau(q,i), & \text{if } i > 1, \ q \geq 1, \\ C_h\tau(q,1), & \text{if } u < v, i = 1, \ q \geq 2, \\ C_h\tau(q,1) + K, & \text{if } u \geq v, i = 1, \ q \geq 2. \end{cases} \tag{19}$$

For $q = 0$, $\tau(0,i) = u$. Cases with $q = 1$ and $q = 0$ are special.

$$\begin{aligned} \text{For } \pi(0) > 0, \ & C(1,1) = C_h\tau(1,1) + K, \ \text{if } u \geq v, \\ & C(1,1) = C_h\tau(1,1), \ \text{if } u < v, \\ & C(0,i) = iC_h\tau(0,i). \\ \text{For } \pi(0) = 0, \ & C(1,1) = C_h\tau(1,1), \ C(0,0) = K, \ i = 0, \\ & C(0,i) = iC_h\tau(0,i), \ i \geq 1. \end{aligned}$$

The cases with $q > q^*$ are dealt with first. Assuming $q > q^*$, the transition matrix blocks are $\{\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2\}$. The sojourn time in the state (q,i) is exponentially distributed with parameter $-(\mathbf{A}_1)_{i,i} = \lambda + \mu$ for all i. The Laplace Stieltjes Transform of the sojourn time is $(-\mathbf{A}_1)_{i,i}/(s - (\mathbf{A}_1)_{i,i})$. The probability of the transition from (q,i) to $(q+1,j)$ is $(\mathbf{A}_0)_{i,j}/(-(\mathbf{A}_1)_{i,i})$. The probability of the transition from state (q,i) to $(q-1,j)$ is $(\mathbf{A}_2)_{i,j}/(-(\mathbf{A}_1)_{i,i})$. The values of $(\mathbf{A}_0)_{i,j}$ and $(\mathbf{A}_2)_{i,j}$ are either $0, \lambda$, or μ. The notation $\{(\mathbf{A}_2)_{i,j}, (\mathbf{A}_1)_{i,j}, (\mathbf{A}_0)_{i,j}\}$ are used in the following derivation since it is more informative and they show the actual transition states.

Conditioning on the state of the Markov process, the Laplace Stieltjes Transform (LST) of the cost function $C(q,i)$ is given as (for $q > q^*$)

$$\begin{aligned} &\mathrm{E}[\exp\{-sC(q,i)\} : q(\tau(q,i)+) = q+1, I(\tau(q,i)+) = j \\ &\quad |q(0) = q, I(0) = i] \\ &= \mathrm{E}[\exp\{-iC_h s\tau(q,i)\}]\mathrm{P}\{q(\tau(q,i)+) = q+1, \\ &\quad I(\tau(q,i)+) = j|q(0) = q, I(0) = i\} \\ &= \frac{(-\mathbf{A}_1)_{i,i}}{[iC_h s - (\mathbf{A}_1)_{i,i}]} \frac{(\mathbf{A}_0)_{i,j}}{(-\mathbf{A}_1)_{i,i}}, \end{aligned} \tag{20}$$

and

$$\begin{aligned} &\mathrm{E}[\exp\{-sC(q,i)\} : q(\tau(q,i)+) = q-1, I(\tau(q,i)+) = j|q(0) = q, I(0) = i] \\ &= \begin{cases} \mathrm{E}[\exp\{-siC_h\tau(q,i)\}]\mathrm{P}\{v < u, I(\tau(q,i)+) = j|q(0) = q, I(0) = i\}, & i \geq 2, \\ \mathrm{E}[\exp\{-s(C_h\tau(q,1) + K)\}]\mathrm{P}\{v < u, I(\tau(q,i)+) = j|q(0) = q, I(0) = i\}, & i = 1, \end{cases} \\ &= \begin{cases} \dfrac{(-\mathbf{A}_1)_{i,i}}{[iC_h s - (\mathbf{A}_1)_{i,i}]} \dfrac{(\mathbf{A}_2)_{i,j}}{(-\mathbf{A}_1)_{i,i}}, & i \geq 2; \\ e^{-sK} \dfrac{(-\mathbf{A}_1)_{1,1}}{[C_h s - (\mathbf{A}_1)_{1,1}]} \dfrac{(\mathbf{A}_2)_{1,j}}{(-\mathbf{A}_1)_{1,1}}, & i = 1. \end{cases} \end{aligned} \tag{21}$$

Equations (20) and (21) give the LSTs of the cost incurred between two consecutive transitions. Next, the LST of the total cost incurred during the fundamental period from level q to level $q-1$ is derived. Define the following function associated with the costs incurred in a fundamental period, for $q > q^*$,

$\Phi_{i,j}(x)$ = The probability that the total cost incurred in a fundamental period is less than x, the Markov process reaches level $q-1$ for the first time in state $(q-1, j)$, given that it starts in state (q, i).

The LST of this function is defined as

$$\Phi^*_{i,j}(s) = \int_0^\infty e^{-sx} d\mathbf{\Phi}_{i,j}(x), \quad \mathrm{Re}(s) > 0. \tag{22}$$

Let $\mathbf{\Phi}^*(s)$ be a $\pi_{\max} \times \pi_{\max}$ matrix with elements $\mathbf{\Phi}^*_{i,j}(S)$.

The Markov process $(q(t), I(t))$ stays in (q, i) for $\tau(q, i)$ and then transits to another state. Conditioned on the next state reached from (q, i), which is either in level $q+1$ or $q-1$, then

$$\begin{aligned}\mathbf{\Phi}^*_{i,j}(s) &= \mathrm{E}[\exp\{-sC(q,i)\} : q(\tau(q,i)+) = q-1, I(\tau(q,i)+) \\ &= j | q(0) = q, I(0) = i] + \sum_{k=1}^{\pi_{\max}} \mathrm{E}[\exp\{-sC(q,i)\} : \\ &\quad q(\tau(q,i)+) = q+1, I(\tau(q,i)+) = k | q(0) \\ &= q, I(0) = i] \mathbf{\Phi}^*_{k,j}(2,s),\end{aligned} \tag{23}$$

where the one period cost $C(q, i)$ is apparently independent of the cost incurred after the first transition. $\mathbf{\Phi}^*_{kj}(2, s)$ is the LST of the total cost incurred during the first passage time from level $q+1$ to level $q-1$. Notice that the transition from level q to itself is ignored since such a transition does not exist for the $(q(t), I(t))$ defined in Section 3. Combining Equations (21) and (23), for $0 \le i, j \le \pi_{\max}$, yields

$$\mathbf{\Phi}^*_{i,j}(s) = \frac{(-\mathbf{A}_1)_{i,i}}{[iC_h s - (\mathbf{A}_1)_{i,i}]} \left[\frac{\exp\{-sKI\{i=1\}\}(\mathbf{A}_2)_{i,j}}{(-\mathbf{A}_1)_{i,i}} + \sum_{k=1}^{\pi_{\max}} \frac{(\mathbf{A}_0)_{i,k}}{(-\mathbf{A}_1)_{i,i}} \mathbf{\Phi}^*_{k,j}(2,s) \right]. \tag{24}$$

In matrix form, after some algebra, Equation (24) becomes

$$\begin{aligned}&[sC_h \mathrm{diag}(1, \ldots, \pi_{\max}) - \mathbf{A}_1]\mathbf{\Phi}^*(s) \\ &\quad = \mathrm{diag}(e^{-sK}, 1, \ldots, 1)\mathbf{A}_2 + \mathbf{A}_0 \mathbf{\Phi}^*(2,s).\end{aligned} \tag{25}$$

The function $\mathbf{\Phi}^*(s)$ can be derived if the relationship between $\mathbf{\Phi}^*(s)$ and $\mathbf{\Phi}^*(2, s)$ can be found. The relationship between $\mathbf{\Phi}^*(2, s)$ and $\mathbf{\Phi}^*(s)$ is given in the following lemma.

Lemma 1. *For Re(s) > 0,* $\mathbf{\Phi}^*(2, s) = [\mathbf{\Phi}^*(s)]^2$.

Proof. The total cost during the first transition from level $q-1$ to level $q-1$ consists of the total cost from level $q+1$ to level q and the total cost from level q to level $q-1$. The two parts are independent by the Markovian property possessed by the process $(q(t), I(t))$ and have a common LST. Let $C_{(q+1,q-1)}$ be the total cost incurred during the fundamental period from level $q+1$ to level $q-1$. Similarly $C_{(q+1,q)}$ and $C_{(q,q-1)}$ are defined for levels $q+1$, q and $q-1$ respectively. Let $I_{q+1,q-1}$ be the inventory level when the Markov process first reaches level $q-1$ from level $q+1$. $I_{q+1,q}$ and $I_{q,q-1}$ are defined for levels $q+1, q$, and $q-1$ respectively. Then the following equality holds.

$$\begin{aligned}&[C_{(q+1,q-1)} : I_{q+1,q-1} = j | I(0) = i] \\ &= \sum_{k=1}^{\pi_{\max}} \{ [C_{(q+1,q)} : I_{q+1,q} = k | I(0) = i] \\ &\quad + [C_{(q,q-1)} : I_{q,q-1} = j | I_{q+1,q} = k] \}.\end{aligned} \tag{26}$$

Taking LST on both sides and using the independence of the random variables in the braces on the second line of (26) gives:

$$\mathbf{\Phi}^*_{i,j}(2,s) = \sum_{k=1}^{\pi_{\max}} \mathbf{\Phi}^*_{i,k}(s) \mathbf{\Phi}^*_{k,j}(s). \tag{27}$$

The matrix form of this equality gives the expected result. ■

With the above lemma, the LST of the total cost incurred in a fundamental period $\mathbf{\Phi}^*(s)$ is obtained. This is the key result of this section.

Theorem 3. *For Re(s) > 0 and $q > q^*$, $\mathbf{\Phi}^*(s)$ satisfies*

$$\begin{aligned}&[sC_h \mathrm{diag}(1, \ldots, \pi_{\max}) - \mathbf{A}_1]\mathbf{\Phi}^*(s) \\ &\quad = \mathrm{diag}(e^{-sK}, 1, \ldots, 1)\mathbf{A}_2 + \mathbf{A}_0[\mathbf{\Phi}^*(s)]^2.\end{aligned} \tag{28}$$

Proof. The result is obtained by using Equation (25) and Lemma 1. ■

Now, an equation is established for the LST of the total cost in a fundamental period for $q > q^*$. For boundary levels ($q \le q^*$), similar results can be derived. Equations for $\{\mathbf{\Phi}^*_q(s)\}$ are given without detailed derivation.

$\mathbf{\Phi}^*_q(s)$ is defined for the LST of the total cost incurred during the fundamental period from level q to level $q-1$, $q > 0$. $\mathbf{\Phi}^*_0(s)$ is defined as the LST of the total cost incurred during a busy cycle, i.e., from level 0 to level 0. For $q > 1$,

$$\begin{aligned}\mathbf{\Phi}^*_q(s) &= [sC_h \mathrm{diag}(1, \ldots, \pi^e(q)) - \mathbf{A}_{q,q} - \mathbf{A}_{q,q-1}\mathbf{\Phi}^*_{q+1}(s)]^{-1} \\ &\quad \times \mathrm{diag}(e^{-sK}, 1, \ldots, 1)\mathbf{A}_{q,q-1}.\end{aligned} \tag{29}$$

When $\pi(0) > 0$, the above equation holds for $q = 1$, since an order is issued when $(q(t), I(t))$ transits from (1, 1) to (0, 1). When $\pi(0) = 0$,

$$\Phi_1^*(s) = [sC_h\mathrm{diag}(1,\dots,\pi^e(1)) - \mathbf{A}_{1,1} - \mathbf{A}_{1,2}\Phi_2^*(s)]^{-1}\mathbf{A}_{1,0}. \quad (30)$$

When $\pi(0) > 0$, no order is placed when $(q(t), I(t))$ transits from (0, 1) to (1, 1), then

$$\Phi_0^*(s) = [sC_h\mathrm{diag}(1,\dots,\pi^e(0)) - \mathbf{A}_{0,0}]^{-1}\mathbf{A}_{0,1}\Phi_1^*(s). \quad (31)$$

When $\pi(0) = 0$,

$$\Phi_0^*(s) = [sC_h\mathrm{diag}(1,\dots,\pi^e(0)) - \mathbf{A}_{0,0}]^{-1} \times \mathrm{diag}(\mathrm{e}^{-sK}, 1,\dots,1)\mathbf{A}_{0,1}\Phi_1^*(s). \quad (32)$$

Since $\Phi_{q^*-1}^*(s) = \Phi^*(s)$, it is clear that $\{\Phi_q^*(s)\}$ can be obtained recursively once $\Phi^*(s)$ is known.

Now that the LSTs of the total costs incurred during fundamental periods have been found, some performance measures such as the distributions and moments of these costs can be determined.

6. The average total cost incurred in a fundamental period

This section derives formulas for the average total cost incurred in a fundamental period. As a result, the average total cost incurred in a busy period as well as the average total cost per product is obtained. Again, the analysis starts with level $q > q^*$.

Differentiating both sides of (28) with respect to s, yields

$$\begin{aligned} &C_h\mathrm{diag}(1,\dots,\pi_{\max})\Phi^*(s) \\ &\quad + [sC_h\mathrm{diag}(1,\dots,\pi_{\max}) - \mathbf{A}_1]\Phi^{*(1)}(s), \\ &= \mathrm{diag}(-K\mathrm{e}^{-sK}, 0,\dots,0)\mathbf{A}_2 \\ &\quad + \mathbf{A}_0[\Phi^*(s)\Phi^{*(1)}(s) + \Phi^{*(1)}(s)\Phi^*(s)]. \end{aligned} \quad (33)$$

Setting $s = 0$ and denoting by $\Phi^{(1)} = \Phi^{*(1)}(0)$, after some algebra,

$$\begin{aligned} \mathbf{A}_1\Phi^{(1)} &= C_h\mathrm{diag}(1,\dots,\pi_{\max})\mathbf{G} + \mathrm{diag}(K, 0,\dots,0)\mathbf{A}_2 \\ &\quad - \mathbf{A}_0[\mathbf{G}\Phi^{(1)} + \Phi^{(1)}\mathbf{G}]. \end{aligned} \quad (34)$$

where the equalities $\mathbf{G} = \Phi^*(0) = \Phi(\infty)$ are used. Matrix $\mathbf{G}$ is the minimal nonnegative solution to the equation

$$\mathbf{A}_2 + \mathbf{A}_1\mathbf{G} + \mathbf{A}_0\mathbf{G}^2 = 0. \quad (35)$$

Theorem 4. *The expected conditional costs incurred in a fundamental period satisfy*

$$\Phi^{(1)} = \Phi_0^{(1)} + (-\mathbf{A}_1^{-1}\mathbf{A}_0)(\mathbf{G}\Phi^{(1)} + \Phi^{(1)}\mathbf{G}), \quad (36)$$

where

$$\Phi_0^{(1)} = \mathbf{A}_1^{-1}[C_h\mathrm{diag}(1,\dots,\pi_{\max})\mathbf{G} + \mathrm{diag}(K, 0,\dots,0)\mathbf{A}_2]. \quad (37)$$

Clearly, elements of the matrix $-\Phi_0^{(1)}$ are the average costs incurred before the next transition. The second part on the right hand side of Equation (36) is the average cost (with a negative sign) after the first transition. The (i, j)th element of $-\Phi^{(1)}$ is the average total cost incurred in a fundamental period when the Markov process first arrives in level $q - 1$ at $(q - 1, j)$ from level q, given that the initial inventory is i. To further simplify Equation (36), let

$$\mathbf{u}_c = -\frac{\mathrm{d}\Phi^*(s)}{\mathrm{d}s}\mathbf{e}\bigg|_{s=0} = -\Phi^{(1)}\mathbf{e}. \quad (38)$$

The element $(\mathbf{u}_c)_i$ is the average total cost incurred during a fundamental period started in the state (q, i).

Theorem 5.

$$\mathbf{u}_c = -[\mathbf{I} - (-\mathbf{A}_1^{-1}\mathbf{A}_0)(\mathbf{I} + \mathbf{G})]^{-1}\Phi_0^{(1)}\mathbf{e}. \quad (39)$$

Proof. Postmultiplying both sides of Equation (36) by $\mathbf{e}$, Equation (39) is obtained after simple algebra. The invertibility of matrix $\mathbf{A}_1 + \mathbf{A}_0(\mathbf{I} + \mathbf{G})$ has been proved in Neuts [11]. ■

Compared to equation (36), expression (39) is computationally straightforward. No recursion is needed when matrix $\mathbf{G}$ is known.

Equations (36) and (39) can be extended to the boundary levels. For instance, let

$$\mathbf{u}_{c,q} = -\frac{\mathrm{d}\Phi_q^*(s)}{\mathrm{d}s}\mathbf{e}\bigg|_{s=0}. \quad (40)$$

Then, for $2 \le q \le q_{\max}$,

$$\begin{aligned} \mathbf{u}_{c,q} &= [\mathbf{I} - (-\mathbf{A}_{q,q}^{-1}\mathbf{A}_{q,q+1})\mathbf{G}_{q+1}]^{-1} \\ &\quad \times [-\Phi_{q,0}^{(1)}\mathbf{e} + (-\mathbf{A}_{q,q}^{-1}\mathbf{A}_{q,q+1})\mathbf{u}_{c,q+1}], \end{aligned} \quad (41)$$

where matrices $\mathbf{G}_q = \Phi_q^*(0)$ (and $\mathbf{G}_{q-1} = \Phi_{q+1}^*(0)$) and $\Phi_{q,0}^{(1)}$ can be calculated using equations

$$\mathbf{G}_q = -(\mathbf{A}_{q,q} + \mathbf{A}_{q,q-1}\mathbf{G}_{q+1})^{-1}\mathbf{A}_{q,q-1}, \quad q \ge 1. \quad (42)$$

$$\begin{aligned} \Phi_{q,0}^{(1)} &= -\mathbf{A}_{q,q}^{-1}[C_h\mathrm{diag}(1, 2,\dots,\pi^e(q))\mathbf{G}_q \\ &\quad + \mathrm{diag}(K, 0,\dots,0)\mathbf{A}_{q,q-1}]. \end{aligned} \quad (43)$$

For $q = 1$ and $\pi(0) > 0$, $\mathbf{u}_{c,1}$ and $\Phi_{1,0}^{(1)}$ are given by Equations (41) and (43). For $q = 1$ and $\pi(0) = 0$, $\Phi_{1,0}^{(1)}$ has no ordering cost, which is obtained by removing the second term in the right hand side of Equation (43).

When $q = 0$ and $\pi(0) = 0$,

$$\mathbf{u}_{c,0} = -\Phi_{0,0}^{(1)}\mathbf{e} + (-\mathbf{A}_{0,0}^{-1}\mathbf{A}_{0,1})\mathbf{u}_{c,1}. \quad (44)$$

where

$$\begin{aligned} \Phi_{0,0}^{(1)} &= \mathbf{A}_{0,0}^{-1}[C_h\mathrm{diag}(0, 1,\dots,\pi^e(0))\mathbf{G}_0 \\ &\quad + \mathrm{diag}(K, 0,\dots,0)\mathbf{A}_{0,1}]. \end{aligned} \quad (45)$$

and

$$\mathbf{G}_0 = (-\mathbf{A}_{00}^{-1}\mathbf{A}_{01})\mathbf{G}_1. \quad (46)$$

When $q = 0$ and $\pi(0) > 0$, $\mathbf{\Phi}_{0,0}^{(1)}$ is obtained by removing the second term on the right hand side of Equation (45).

To find the average total cost per product, consider the average total cost incurred in a busy cycle. The idea is to consider the embedded Markov chain at the end points of busy periods. By definition, it is known that this embedded Markov chain has a transition matrix $\mathbf{G}_0$. It is easy to see that $\mathbf{G}_0$ is an irreducible stochastic matrix when the Markov process is positive recurrent. Let $\boldsymbol{\beta}_0$ be the left invariant vector of $\mathbf{G}_0$, i.e., $\boldsymbol{\beta}_0\mathbf{G}_0 = \boldsymbol{\beta}_0$, $\boldsymbol{\beta}_0 \geq \mathbf{0}$, and $\boldsymbol{\beta}_0\mathbf{e} = 1$. Conditioning on the initial state of a busy cycle, the average total cost incurred in a busy cycle can be obtained as

$$\sum_{i=I\{\pi(0)\neq 0\}}^{\pi^*(0)} (\boldsymbol{\beta}_0)_i(\mathbf{u}_{c,0})_i, \quad (47)$$

i.e., the product of vectors $\boldsymbol{\beta}_0$ and $\mathbf{u}_{c,0}$: $\boldsymbol{\beta}_0\mathbf{u}_{c,0}$. Since the total number of products produced in a busy cycle is the total number of customers served in a busy cycle for an $M/M/1$ queue, the average number of products produced in a busy cycle is $\mu/(\mu - \lambda)$ [9]. Then the average total cost per product is given by, when the policy π is applied,

$$g(\pi) = \boldsymbol{\beta}_0\mathbf{u}_{c,0}(\mu - \lambda)/\mu. \quad (48)$$

In summary, an algorithm for computing performance measures related to busy cycles can be developed as follows.

A G-related Algorithm

Steps 1, 2, and 3 are identical to the **R**-related algorithm given in Section 4.

Step 4. Compute the matrix **G** (Equation (35)).

Step 5. Compute $\mathbf{u}_c$, $\mathbf{G}_q$, $\mathbf{u}_{c,q}$, for $0 \leq q \leq q^*$, and the vector $\boldsymbol{\beta}_0$ (Equations (39), (41), (42), (44), and (46)).

Step 6. Compute other performance measures.

An algorithm for computing matrices **G** and $\{\mathbf{G}_q, 0 \leq q \leq q^*\}$ can be found in Neuts [11].

It is clear that the **G**-related algorithm only requires a space complexity $O((\pi_{\max})^2)$, while the **R**-related algorithm presented in Section 4 requires a space complexity $O(q^*(\pi_{\max})^2)$. When q^* is large, i.e., the replenishment policy fluctuates with the queue length until the queue length is fairly large, the **R**-related algorithm will require significantly more computing resources than the **G**-related algorithm which is simply doing basic matrix operations. The shortcoming of the fundamental period approach is that it only deals with the costs incurred in the fundamental periods and cannot be used to deal with the distributions of the queue length and the inventory level.

Example 3. (Example 1 continued). The average total costs per product for the optimal replenishment policy given in Example 1 and the best EOQ policy of this system (Example 2) are $g^* = 15.640$ and $g_{-1}(4) = 15.833$, respectively. Clearly, the use of information about the queue length in inventory control reduces total inventory costs. The reason is that using information about the queue length to adjust the order size leads to a reduction in the inventory cost incurred during idle periods. More discussion on the value of information in inventory control can be found in He *et al.* [8].

7. Summary, extensions, and generalizations

To develop algorithms for computing the average total cost per product of any feasible replenishment policy $\pi \in \mathbf{\Pi}$, this paper accomplished the following.

(i) The quasi-birth-and-death Markov process associated with a feasible policy was constructed and its transition matrix was given explicitly.

(ii) Based on the matrix-geometric solution, the **R**-related algorithm was developed for computing various performance measures such as the distribution of the inventory level, the mean inventory level, and the average total cost per product.

(iii) Using the fundamental period approach, the LST of the total costs incurred in a fundamental period was derived. The **G**-related algorithm was developed for computing performance measures such as the average total cost incurred in a busy cycle and the average total cost per product.

The main focus of this paper has been in developing two algorithms that can be used for computing various performance measures for the inventory-production system. The underlying methods used in the algorithms have been well tested by other researchers, thus this paper does not focus on a comparison of the computational aspects of the resulting algorithms. The contribution of this paper has been in showing how matrix analytic methods can be applied to a new problem setting so that performance measures for a particular replenishment policy can be easily computed, and to illustrate how some results can be explicitly derived. Once developed, the two algorithms given in this paper could be used to evaluate a variety of replenishment policies and to gain insights into the behaviour of the inventory-production system. Such results are reported in He *et al.* [8].

The analysis of this paper can be extended to include higher moments of inventory costs. For instance, the second moments of the costs incurred during the fundamental periods can be derived from equations given in Section 5. The results for the second moments are useful in analyzing the variability of the total cost incurred in fundamental periods.

The analysis carried out in this paper can also be extended to inventory-production models with exponential or phase-type leadtimes as well (assuming a finite shortage cost and at most one outstanding order). For such cases, the workshop can no longer be modelled as an $M/M/1$ queue and new random variables must be added to $(q(t), I(t))$ to represent the inventory-production model. Thus, the Markov process associated with the inventory-production model becomes $(q(t), I(t), J(t), k(t))$, where $q(t), I(t), J(t)$, and $k(t)$ are the queue length, the inventory level, the size of the outstanding order (if any), and the state of the leadtime process of the outstanding order at time t, respectively. For replenishment policies similar to that of $\mathbf{\Pi}$ defined in Section 2, $(q(t), I(t), J(t), k(t))$ is a QBD Markov process where $(I(t), J(t), k(t))$ is the auxiliary random variable. This QBD Markov process is more complicated than the one defined in Section 3. Nonetheless, the modeling approach used in this paper facilitates such generalizations so that the theory developed of Neuts [11] can be applied. Most of the work laid out in this paper is still valid. Some work in this direction can be found in He *et al.* [8]

Acknowledgements

The authors would like to thank the referees and the editor for valuable comments and suggestions.

References

[1] He, Q.-M., Jewkes, E.M. and Buzacott, J. (1997). Inventory control policies for an inventory-production system. (submitted).

[2] Axsäter, S. (1993) Continuous review policies for multi-level inventory systems with stochastic demand, *Logistics of Production and Inventory*, Nemhauser, G.L. and Rinnooy Kan, A.H.G. (eds), North-Holland, New York, pp. 195–198.

[3] Axsäter, S. (1993) Exact and approximate evaluation of batch-ordering policies for two-level inventory systems. *Operations Research*, **41**, 777–785.

[4] Buzacott, J.A. and Shanthikumar, J.G. (1993) *Stochastic Models of Manufacturing Systems*, Prentice Hall, Englewood Cliffs, NJ.

[5] Federgruen, A. (1993) Centralized planning models for multi-echelon inventory systems under uncertainty, in *Logistics of Production and Inventory*, Nemhauser, G.L. and Rinnooy Kan, A.H.G. (eds), North-Holland, New York, pp. 133–174.

[6] Mitra, D. and Mitrani, I. (1990) Analysis of a kanban discipline for cell coordination in production lines I. *Management Sciences*, **36**, 1548–1566.

[7] Zipkin, P. (1989) A kanban-like production control system: analysis of simple models. Working paper 89-1, Graduate School of Business, Columbia University, NY.

[8] He, Q.-M., Jewkes, E.M. and Buzacott, J. (1997) The value of information used in inventory control of an inventory-production system. (submitted).

[9] Cohen, J.W. (1982) *The Single Server Queue*, North-Holland, New York.

[10] Bartmann, D. and Beckmann, M.J. (1992) *Inventory Control: Models and Methods*, Springer-Verlag, Berlin, New York.

[11] Neuts, M.F. (1981) *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore, London. Chps. 1 and 3.

[12] Neuts, M.F. (1989) *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York.

[13] Chakravarthy, S. and Alfa, A.S. (1996) *Matrix-Analytic Methods in Stochastic Models*, Marcel Dekker, Inc, New York.

[14] Ramaswami, V. and Taylor, P.G. (1996) Some properties of the rate matrices in level dependent quasi-birth-and-death processes with a countable number of phases. *Stochastic Models*, **12**, 143–164.

[15] Ross, S.M. (1970) *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, CA.

[16] Latouche, G. and Ramaswami, V. (1993) A logarithmic reduction algorithm for quasi-birth-and-death processes, *Journal of Applied Probability*, **30**, 650–674.

Appendix A

Transition blocks of Markov process $(q(t), I(t))$

First, the dimensions of the transition blocks are given as follows. $\{\mathbf{A}_{q,q-1}\}$ are $\pi^e(q) \times \pi^e(q-1)$ matrices since level q (i.e., S_q) has $\pi^e(q)$ states and level $q-1$ (S_{q-1}) has $\pi^e(q-1)$ states; Similarly, $\{\mathbf{A}_{q,q}\}$ are $\pi^e(q) \times \pi^e(q)$ matrices; $\{\mathbf{A}_{q,q+1}\}$ are $\pi^e(q) \times \pi^e(q+1)$ matrices; $\mathbf{A}_0$, $\mathbf{A}_1$ and $\mathbf{A}_2$ are $\pi_{\max} \times \pi_{\max}$ matrices. Note that the envelope function $\pi^e(q)$ represents the number of states in level q (except level zero). Based on the discussion given in Section 3, the transition blocks can be written explicitly as follows.

$\mathbf{A}_{0,0} = -\lambda\mathbf{I}$: no transitions among the states of level zero; the next transition epoch is the arrival epoch of a demand.

$\mathbf{A}_{0,1} = \lambda(\mathbf{I}, 0)$, for $\pi(0) > 0, \pi^e(0) = \pi^e(1) - 1$;

$\mathbf{A}_{0,1} = \lambda\mathbf{I}$, for $\pi(0) > 0, \pi^e(0) = \pi^e(1)$;

$$\mathbf{A}_{0,1} = \begin{pmatrix} 0 & 0 & \cdots & \lambda & \cdots & 0 & 0 \\ & \lambda & & & & & \\ & & \ddots & & & & \\ & & & \ddots & & & \\ & & & & \ddots & & \\ & & & & & \lambda & 0 \end{pmatrix},$$

for $\pi(0) = 0$, $\pi^e(0) = \pi^e(1) - 1$.

When $\pi(0) > 0$, no order is issued at the next arrival epoch when the queue length becomes one. When $\pi(0) = 0$, an order of the size $\pi(1)$ is issued and filled when the Markov process moves from (0, 0) to $(1, \pi(1))$, which brings the inventory level to $\pi(1)$. For matrix $\mathbf{A}_{0,1}$ defined for the $\pi(0) = 0$ case, the right hand side column (the zero column) is removed when $\pi^e(0) = \pi^e(1)$.

$$\text{For } q > 0,\ \mathbf{A}_{q,q-1} = \begin{pmatrix} 0 & \cdots & \mu & \cdots & 0 \\ \mu & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \mu & 0 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}.$$

for $\pi^e(q) > \pi^e(q-1)$.

When $\pi^e(q) = \pi^e(q-1)$, the last row is removed from $\mathbf{A}_{q,q-1}$. When $\pi(0) = 0$, the first row in $\mathbf{A}_{1,0}$ becomes zeros. When $(q(t), I(t))$ makes the transition from $(q, 1)$ to level $q-1$, an order of the size $\pi(q-1)$ is issued and filled.

$\mathbf{A}_{q,q} = -(\lambda + \mu)\mathbf{I}$. There is no transition among the states in the same level.
$\mathbf{A}_{q,q-1} = (\lambda\mathbf{I}, 0)$, when $\pi(q) > \pi(q+1), q \geq 1$, and
$\mathbf{A}_{q,q-1} = \lambda\mathbf{I}$, when $\pi(q) = \pi(q+1)$
$\mathbf{A}_0 = \lambda\mathbf{I}, \mathbf{A}_1 = -(\lambda + \mu)\mathbf{I}$,

and

$$\mathbf{A}_2 = \begin{pmatrix} 0 & \cdots & \mu & \cdots & 0 \\ \mu & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \mu & 0 \end{pmatrix}.$$

In $\mathbf{A}_2$, when the Markov process goes down one level from the state $(q, 1)$, an order of size π_∞ is issued and filled.

Biographies

Qi-Ming He is currently an Assistant Professor in the Industrial Engineering Department, Dalhousie University. He received a Ph.D. from the Institute of Applied Mathematics, Chinese Academy of Sciences in 1989 and a Ph.D. from the Department of Management Science at the University of Waterloo in 1996. His main research areas are algorithmic methods in applied probability, queueing theory, inventory control, and production management. He is currently working on queueing and inventory systems with multiple types of customers (demands).

Elizabeth Jewkes is currently an Associate Professor in the Department of Management Science at the University of Waterloo. Her research is in stochastic models of manufacturing systems. Recent work has involved optimal investment in reducing setup time and improving quality and in the value of information in supply chain management.

Contributed by the Production Planning Department