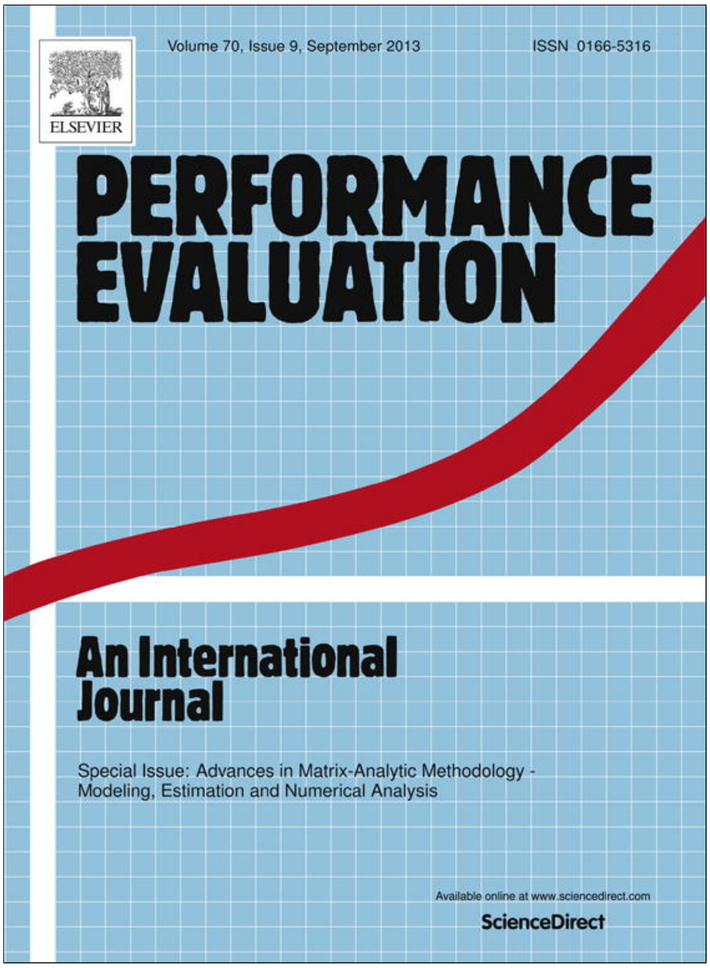
Volume 70, Issue 9, September 2013
ISSN 0166-5316
ELSEVIER
PERFORMANCE EVALUATION
An International Journal
Special Issue: Advances in Matrix-Analytic Methodology - Modeling, Estimation and Numerical Analysis
Available online at www.sciencedirect.com
ScienceDirect

Performance Evaluation 70 (2013) 623–638

Contents lists available at ScienceDirect

Performance Evaluation

journal homepage: www.elsevier.com/locate/peva

Performance analysis of an inventory–production system with shipment consolidation in the production facility

Qi-Ming He [a,*], Hanqin Zhang [b,c]

[a] University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1

[b] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100191, China

[c] NUS Business School, Singapore 119245, Singapore

ARTICLE INFO

Article history:
Available online 11 June 2013

Keywords:
Continuous time Markov chain
Matrix-analytic methods
Inventory–production system
Shipment consolidation

ABSTRACT

We consider an inventory–production system consisting of a warehouse and a production facility. The warehouse is used to store products to satisfy customer demands, and its inventory is controlled by an (r, Q) policy. Products ordered by the warehouse are processed in the production facility on a one-by-one basis, and finished products are consolidated into batches to be shipped from the production facility to the warehouse. Using the matrix-analytic methods, explicit solutions are obtained and computational methods are developed for analyzing system performance measures.

1. Introduction

The inventory–production system considered in this article consists of one warehouse and one production facility. The warehouse is used to store products to satisfy customer demands that arrive at the warehouse according to a non-renewal process, and the production facility processes products ordered by the warehouse on a one-by-one basis. Product processing times are assumed to be random, and are independent and identically distributed. The finished products at the production facility have to be consolidated into batches to be delivered to the warehouse. If there is no product available in the warehouse when a customer demand arrives, the customer will wait for one that will later on be delivered from the workshop. That is, the unsatisfied customer demand will be backlogged. The system cost includes the holding cost, the penalty cost and the fixed order cost incurred at the warehouse, and the holding cost of the finished products and the fixed delivery cost incurred at the workshop. Unlike the classical stochastic inventory systems where the inventory replenishment policy is determined mainly by inventory position/level, the inventory–production system takes into consideration information on the production facility in its inventory management. Here the production information includes the production capacity (processing time of each product) and the finished product delivery batch-size. Thus, inventory management is more sophisticated, if additional information on production is available and utilized. In this paper, based on a given finished product delivery batch-size and the capacity of the production facility, we derive explicit solutions and develop computational methods for analyzing system performance measures. We also develop a heuristic algorithm for finding an optimal inventory control policy at the warehouse so as to minimize the expected long-run average system-wide cost.

The (r, Q) policy, like the (s, S) policy for discrete time systems, is a popular type of inventory control policies (see [1–5]). It has been shown that the optimal policy for many inventory models is of the (r, Q) or (s, S) type (e.g., [6–13]). Algorithms

* Corresponding author.
E-mail addresses: q7he@uwaterloo.ca (Q.-M. He), bizzhq@nus.edu.sg (H. Zhang).

0166-5316/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.peva.2013.05.007

have been developed for computing the optimal (r, Q) policy for the inventory models. Thus, the (r, Q) policy can be a good choice for inventory control in inventory–production systems. In this paper, we will restrict ourselves among (r, Q)-type policies to find an optimal policy for the inventory–production system described above.

Based on the nature of the system, inventory–production systems can be basically categorized into two classes. One class consists of deterministic inventory–production systems: the product demands at the warehouse are deterministic and the product processing times at the workshop are also deterministic. The optimal replenishment policy for the warehouse is explored (e.g., see [14,15]). The other class includes stochastic inventory–production systems. Based on the system's operation flow, this class can be further classified into two types. One type is the assemble-to-order production–inventory system: the product demands to the warehouse are random and demand arrivals are usually formulated as a Poisson process, and the product processing times at the production facility are random (sometimes, the product may be, in parallel, processed at multiple production facilities with different processing times). When there is no fixed order/delivery cost at the warehouse/production facility, performance measures of such systems are analyzed (see [16]). The other type is the make-to-order production–inventory system. Compared with the assemble-to-order system, the operational flow for the make-to-order system is just reversed. The customer demands directly go to the production facility, process times of demands in the production facility are again random, while the warehouse just stores raw materials for production. Here the production facility makes raw material orders from the warehouse to be used for generating the products. When there is no fixed order cost for the production facility, He et al. [17–19] investigate the optimal inventory control policy at the warehouse. For the make-to-order system, furthermore, without considering the cost incurred at the warehouse, De Vericourt et al. [20], and Ha [21] give the optimal allocation policy when the production facility has several demand classes.

The inventory–production system considered in this paper is a stochastic assemble-to-order production–inventory system. Compared with those in the existing literature, our model has several special features. First, the shipment consolidation of finished products at the production facility is considered. This more general feature makes it possible to consider costs (e.g., inventory holding cost and transportation cost) associated with finished products. Furthermore, the shipment-size from the production facility to the warehouse may be different from the order-size determined by the warehouse. This relaxation on shipment-sizes would capture more about the workshop's transportation capability and ordering structure for raw materials to be used to generate products, and at the same time, the warehouse may use a different batch order size to reduce its orders' leadtime, or in other words, to improve the utilization of the production facility. Second, the fixed order cost incurred in the warehouse and fixed delivery cost incurred at the production facility are included. Thirdly, the demand process is modeled by a Markovian arrival process (*MAP*), which is a fairly general tool for modeling stochastic arrival processes (e.g., [22–24]). The *MAP*s can capture the possible correlation and burstiness in the demand process. Lastly, the production time is modeled by a phase-type distribution (e.g., [24–26]). As the phase-type distributions can approximate any probability distribution given by nonnegative random variables, our assumption on the production time is quite general.

Matrix-analytic methods are efficient methods for analyzing stochastic models (e.g., [25–27]). In the inventory management area, the matrix-analytic methods have been used successfully in analyzing system performance measures and determining the optimal inventory policies (e.g., [16–19,28–30]). By taking advantage of such methods, we develop efficient methods for computing performance measures for the inventory–production system of interest. The optimal (r, Q) policy of the system is characterized (partially) theoretically and numerically.

The remainder of the paper is organized as follows. In Section 2, the inventory–production system of interest is introduced. An irreducible $M/G/1$ type Markov chain is constructed for the system in Section 3. Based on matrix-analytic methods and Ramaswami's algorithm, a method for computing the stationary distribution of the Markov chain is presented. In Section 4, we derive analytic expressions for several key performance measures. Section 5 discusses some computational issues that may lead to improved algorithms for computing performance measures. A heuristic algorithm is proposed for computing the optimal (r, Q) policy. In Section 6, numerical examples are given and the sensitivity of system performance on system parameters is discussed. Section 7 concludes the paper.

2. The inventory–production system

The inventory–production system of interest consists of a warehouse and a production facility. Customer demands arrive at the warehouse. Demands are either satisfied immediately, if the warehouse has on-hand inventory, or backlogged, otherwise. The warehouse sends replenishment orders to the production facility. The production facility has infinite resource of raw materials for production and produces products ordered by the warehouse on a one-by-one basis. Finished products are stored in the production facility first. Once the total number of finished products reaches a threshold, all the cumulated products are consolidated into a batch and sent from the production facility to the warehouse. The transportation time between the production facility and the warehouse is negligible. The flows of demands, orders, and finished products in the inventory–production system are shown in Fig. 2.1.

More specifically, the inventory–production system is defined as follows.

1. Customer demands arrive according to a *Markovian arrival process (MAP)* with matrix representation (D_0, D_1), where D_0 and D_1 are $m_a \times m_a$ matrices with nonnegative elements, except for the diagonal elements of D_0, which are negative

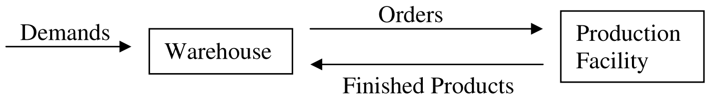

Fig. 2.1. Flows of demands, orders, and finished products.

numbers, and m_a is a positive integer. Let $D = D_0 + D_1$. Then D is the infinitesimal generator of the continuous time underlying Markov chain for the demand arrival process. We assume that D is irreducible. Denote by $I_a(t)$ the phase of the underlying Markov chain at time t. Denote by $\boldsymbol{\theta}_a$ the stationary distribution of D, i.e., $\boldsymbol{\theta}_a D = 0$, $\boldsymbol{\theta}_a \geq 0$, and $\boldsymbol{\theta}_a \mathbf{e} = 1$, where $\mathbf{e}$ is a column vector with all elements being one. (Note that, throughout this paper, the size of $\mathbf{e}$ depends on the context; Occasionally, the size of $\mathbf{e}$ is specified for clarity (e.g., $\mathbf{e}_{m_a}$ for $\mathbf{e}$ of size m_a).) The (average) arrival rate of demands is given by $\lambda = \boldsymbol{\theta}_a D_1 \mathbf{e}$. We refer to Lucantoni [23] and Neuts [24] for more about *MAP*s.

2. Inventory in the warehouse is reviewed continuously. All products that have been ordered but have not yet arrived in the warehouse are called *inventory on order*, which are in the production facility, waiting to be produced or shipped. Define *inventory position* as the on-hand inventory in the warehouse, plus the inventory on order, and minus the number of backlogged demands in the warehouse. The warehouse adopts an (r, q_1) policy for its inventory management (i.e., the usual (r, Q) policy for inventory systems with a continuous review scheme). That is: whenever the inventory position reaches r, an order of the amount q_1 is placed to the production facility. Consequently, the inventory position is brought up to $r + q_1$. The constant r is called the *reorder level* and q_1 is called the *order size*. Note that r can be any integer and q_1 has to be a positive integer.
3. The production facility always has enough resource for production. The production facility produces one product at a time. The production time of a product has a *phase-type distribution* with *PH*-representation $(\boldsymbol{\alpha}, T)$ of size m_s (a positive integer), where $\boldsymbol{\alpha}$ is a nonnegative row vector of size m_s and satisfies $\boldsymbol{\alpha}\mathbf{e} = 1$, T is an $m_s \times m_s$ *PH*-generator with negative diagonal elements and nonnegative off-diagonal elements. Denote by $I_s(t)$ the phase of the underlying Markov chain associated with $(\boldsymbol{\alpha}, T)$ at time t, if the production is on at time t; otherwise, set $I_s(t) = 0$. Then the state space of $I_s(t)$ is $\{0, 1, \ldots, m_s\}$. The mean production time is given by $\mu^{-1} = -\boldsymbol{\alpha} T^{-1}\mathbf{e}$, where μ is called the production rate. We assume that the *PH*-representation $(\boldsymbol{\alpha}, T)$ is irreducible, i.e., the infinitesimal generator $T + (-T\mathbf{e})\boldsymbol{\alpha}$ is irreducible. Let $\boldsymbol{\theta}_s$ be a row vector satisfying $\boldsymbol{\theta}_s(T + \mathbf{T}^0\boldsymbol{\alpha}) = 0$, $\boldsymbol{\theta}_s \geq 0$, and $\boldsymbol{\theta}_s \mathbf{e} = 1$. It is easy to verify that $\mu = \boldsymbol{\theta}_s \mathbf{T}^0$. We refer to Neuts [25] for more about *PH*-distributions.
4. Finished products are stored in the production facility first. A special shipment consolidation policy, called the *quantity policy*, is applied for the shipment of finished products. Namely, as soon as the number of finished products reaches q_2, where q_2 is a positive integer, all the cumulated finished products are shipped together to the warehouse.
5. The system costs include: the holding cost per product in the warehouse per unit time is h_w; the penalty cost per demand per unit time waiting in the warehouse is p_w; the ordering cost per order in the warehouse is K_w; the holding cost per finished product in the production facility per unit time is h_s; and the fixed delivery cost in the production facility is K_s.

To analyze the inventory–production system, we introduce the following variables to represent the system status.

(i) $IP(t)$: the inventory position at time t minus the reorder level r. Then $IP(t)$ takes integer values between 1 and q_1, and $r + IP(t)$ is the inventory position.
(ii) $q(t)$: the number of products being produced or waiting to be produced at time t, which form a queue in the production facility.
(iii) $w(t)$: the number of finished products in the production facility waiting to be shipped to the warehouse at time t.

With the above system variables, the inventory level, on-hand inventory and backlogs in the warehouse can be defined. Let $x^+ = \max\{0, x\}$.

(iv) $IP(t) + r - q(t) - w(t)$: the inventory level in the warehouse at time t.
(v) $(IP(t) + r - q(t) - w(t))^+$: the on-hand inventory in the warehouse at time t.
(vi) $(q(t) + w(t) - IP(t) - r)^+$: the number of backlogs (waiting demands) in the warehouse at time t.

All the above definitions are summarized in Fig. 2.2 for a complete view on the inventory–production system of interest.

For convenience, we also use the notation $\{IP(t), q(t), w(t)\}$ to denote the corresponding variables in steady state. Assume that the system can reach steady state. Then the expected total cost per unit time of the inventory–production system can be obtained as

$$C(r, q_1) = \frac{\lambda K_w}{q_1} + h_w E[(r + IP(t) - q(t) - w(t))^+] + p_w E[(q(t) + w(t) - r - IP(t))^+] + \frac{\lambda K_s}{q_2} + h_s E[w(t)]. \quad (2.1)$$

The objective of this paper is to develop methods for computing performance measures, such as $E[IP(t)]$, $E[q(t)]$, $E[w(t)]$, and $C(r, q_1)$, and for finding the optimal (r, q_1) policy that minimizes the function $C(r, q_1)$.

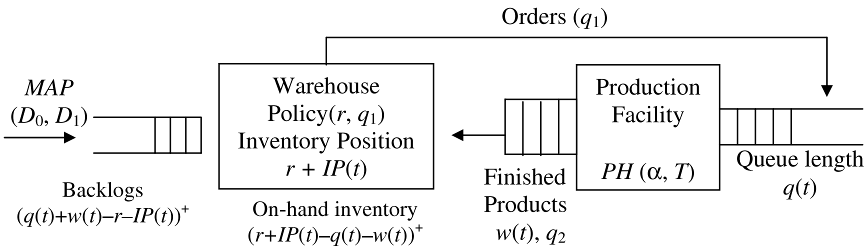

Fig. 2.2. Details of the inventory–production system.

3. Matrix-analytic solutions

In this section, we use matrix-analytic methods to study the continuous time Markov chain (CTMC) $\{(q(t), IP(t), I_a(t), I_s(t), w(t)), t \geq 0\}$ and find its stationary distribution. Our analysis consists of four parts. First, we begin with the simple CTMC $\{(IP(t), I_a(t)), t \geq 0\}$. The infinitesimal generator of the simple CTMC is constructed and an explicit solution is found for its stationary distribution. Second, we construct the infinitesimal generator for the CTMC $\{(q(t), IP(t), I_a(t), I_s(t)), t \geq 0\}$. Third, we construct an irreducible version $\{(q(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t \geq 0\}$ of the CTMC $\{(q(t), IP(t), I_a(t), I_s(t), w(t)), t \geq 0\}$ and its corresponding infinitesimal generator. Fourth, we re-block the infinitesimal generator of $\{(q(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t \geq 0\}$ to reveal its $M/G/1$ structure and to find its stationary distribution.

First, for a given (r, q_1) policy, the inventory position $r + IP(t)$ depends only on the Markovian arrival process (D_0, D_1) and q_1. Thus, $\{(IP(t), I_a(t)), t \geq 0\}$ is a CTMC with a state space $\{(i, j),\ i = 1, 2, \ldots, q_1,\ j = 1, 2, \ldots, m_a\}$. The infinitesimal generator of the process is

$$Q_{IP} = \begin{pmatrix} D_0 & & & & D_1 \\ D_1 & D_0 & & & \\ & D_1 & \ddots & & \\ & & \ddots & D_0 & \\ & & & D_1 & D_0 \end{pmatrix}_{(q_1 m_a)\times(q_1 m_a)} . \tag{3.1}$$

It is straightforward to obtain the following result.

Proposition 3.1. *The stationary distribution of* $\{(IP(t), I_a(t)), t \geq 0\}$ *is given by* $(\boldsymbol{\theta}_a, \boldsymbol{\theta}_a, \ldots, \boldsymbol{\theta}_a)/q_1$. *Consequently, in steady state, the distribution of the process* $\{IP(t), t \geq 0\}$ *is the discrete uniform distribution on* $\{1, 2, \ldots, q_1\}$.

Remark 3.1. Proposition 3.1 for $\{(IP(t), I_a(t)), t \geq 0\}$ can be useful in checking the computation accuracy for the matrix-analytic solutions for $\{(q(t), IP(t), I_a(t), I_s(t), w(t)), t \geq 0\}$.

Second, we have a look at the process $\{(q(t), IP(t), I_a(t), I_s(t)), t \geq 0\}$. By the irreducibility of D (the underlying process of the demand process) and $(\boldsymbol{\alpha}, T)$ (the production time), the process $\{(q(t), IP(t), I_a(t), I_s(t)), t \geq 0\}$ is an irreducible CTMC. The state space of the Markov chain is $\{(0, i, i_a),\ i = 1, 2, \ldots, q_1, i_a = 1, 2, \ldots, m_a\} \cup \{(q, i, i_a, i_s), q = 1, 2, \ldots, i = 1, 2, \ldots, q_1, i_a = 1, 2, \ldots, m_a, i_s = 1, 2, \ldots, m_s\}$. When the variable $q(t)$ changes its value, it either increases by q_1 or decreases by one. We call $q(t)$ the *level variable* and $(IP(t), I_a(t), I_s(t))$ the (vector) *phase variable*. Within each level, the states are ordered lexicographically. If $q(t) = 0$, i.e., the boundary level, there is no production. If an order is placed when $q(t) = 0$, $q(t)$ increases by q_1 and the production in the facility is initialized with distribution $\boldsymbol{\alpha}$. If an order is placed when $q(t) > 0$, $q(t)$ increases by q_1 and the phase of the production process remains the same. The infinitesimal generator Q_q of the Markov chain is given by

$$Q_q = \begin{pmatrix} Q_{IP,0} & 0 & \cdots & 0 & Q_{IP,1}\otimes\boldsymbol{\alpha} & & & \\ I\otimes\mathbf{T}^0 & Q_{IP,0}\oplus T & 0 & \cdots & 0 & Q_{IP,1}\otimes I & & \\ & I\otimes(\mathbf{T}^0\boldsymbol{\alpha}) & Q_{IP,0}\oplus T & 0 & \cdots & 0 & Q_{IP,1}\otimes I & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{3.2}$$

where I is the identity matrix (Note that, throughout this paper, the size of I depends on the context; Occasionally, the size of I is specified for clarity.), $\mathbf{T}^0 = -T\mathbf{e}$, "$\otimes$" is for the Kronecker product of matrices (i.e., for matrices $A = (a_{i,j})$ and B, their

Table 3.1
Relationship between $w(t)$, $\tilde{w}(t)$, and $q(t)$, if $q(0) = w(0) = 0$.

$\tilde{w}(t)$	1	2	3	…	$\tilde{q}_2$
$w(t)$ (if $q(t) = 0$)	0	g_{cd}	$2g_{cd}$	…	$(\tilde{q}_2 - 1)g_{cd}$
$w(t)$ (if $q(t) = ng_{cd} + k$)	$g_{cd} - k$	$2g_{cd} - k$	$3g_{cd} - k$	…	$\tilde{q}_2 g_{cd} - k$

Kronecker product is defined as $A \otimes B = (a_{i,j}B))$, and

$$\begin{aligned} &Q_{IP,0} = J(q_1) \otimes D_1 + I_{q_1\times q_1} \otimes D_0;\\ &Q_{IP,1} = L(q_1) \otimes D_1;\\ &Q_{IP,0} \oplus T \equiv Q_{IP,0} \otimes I_{m_s\times m_s} + I_{(q_1m_a)\times(q_1m_a)} \otimes T;\\ &J(q_1) = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}_{q_1\times q_1}, \quad L(q_1) = \begin{pmatrix} 0 & & & 1 \\ & \ddots & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}_{q_1\times q_1} = \mathbf{e}(1)\mathbf{v}(q_1), \end{aligned} \tag{3.3}$$

$\mathbf{e}(1)$ is a column vector with the first element being one and all others zero, and $\mathbf{v}(q_1)$ is a row vector with the last element (i.e., the q_1-th element) being one and all others zero. Note that $Q_{IP,0} + Q_{IP,1} = Q_{IP}$, where Q_{IP} is defined in Eq. (3.1). By regrouping the states, the Markov chain can be transformed into a quasi-birth-and-death process. Matrix geometric solutions can be found for its stationary distribution. Since the Markov chain is not used in the analysis of the inventory–production system, we shall not study it further.

Third, the process $\{(q(t), IP(t), I_a(t), I_s(t), w(t)), t \geq 0\}$ is a CTMC. Since variable $w(t)$ goes from 0 to $q_2 - 1$ in a cyclic manner (i.e., 0 to 1, 1 to 2, …, q_2-2 to q_2-1, and q_2-1 to 0), the infinitesimal generator of $\{(q(t), IP(t), I_a(t), I_s(t), w(t)), t \geq 0\}$ can be constructed from Q_q in a straightforward manner. Unfortunately, because of a close relationship between $w(t)$ and $q(t)$, the CTMC $\{(q(t), IP(t), I_a(t), I_s(t), w(t)), t \geq 0\}$ can be reducible. To study the steady-state behavior of the system, we need to explore the relationship between $w(t)$ and $q(t)$ and construct an irreducible Markov chain.

Suppose that q_1 and q_2 are co-prime integers, i.e., their greatest common divisor is one. For any integer k in $\{0, 1, \ldots, q_2 - 1\}$, there exist n and m such that $nq_1 - mq_2 = k$. Then $w(t)$ takes integer values in $\{0, 1, \ldots, q_2 - 1\}$ if $q(t) = 0$. Since $w(t)$ changes its values from 0 to $q_2 - 1$ cyclically, it is readily seen that $w(t)$ can take any values in $\{0, 1, \ldots, q_2 - 1\}$ for any $q(t)$. The Markov chain $\{(q(t), IP(t), I_a(t), I_s(t), w(t)), t \geq 0\}$ is irreducible.

For the general case, denote by g_{cd} the greatest common divisor of q_1 and q_2. A moment of refection leads to the fact that $q(t) + w(t) - w(0)$ must be a multiple of g_{cd} and can be any nonnegative multiple of g_{cd}. Thus, for any $q(t)$, $w(t)$ can take exactly q_2/g_{cd} values. Define

$$\tilde{q}_1 = \frac{q_1}{g_{cd}} \quad \text{and} \quad \tilde{q}_2 = \frac{q_2}{g_{cd}}. \tag{3.4}$$

The two integers $\tilde{q}_1$ and $\tilde{q}_2$ are co-prime.

Proposition 3.2. *Assume that $q(0) = w(0) = 0$. If $q(t) = 0$, then $w(t)$ can and can only take any number in $\{jg_{cd}, j = 1, 2, \ldots, \tilde{q}_2 - 1\}$. If $q(t) = ng_{cd} + k$, for $n \geq 0$ and $k = 1, 2, \ldots, g_{cd}$, then $w(t)$ can and can only take any number in $\{jg_{cd} - k, j = 1, 2, \ldots, \tilde{q}_2\}$.*

Proof. Under condition $q(0) = w(0) = 0$, $q(t) + w(t)$ is a multiple of g_{cd}. Since g_{cd} is a common divisor of $q(0)$, $w(0)$, q_1, and q_2, then $w(t)$ can be, and only be, one of $\{0, g_{cd}, 2g_{cd}, \ldots, (\tilde{q}_2 - 1)g_{cd}\}$, if $q(t)$ is a multiple number of g_{cd}. That proves the first part of the lemma. The second result follows from the fact that $w(t)$ changes its value cyclically. This completes the proof of Proposition 3.2. □

For convenience, we shall use $\{1, 2, \ldots, \tilde{q}_2\}$ to represent the $\tilde{q}_2$ states of $w(t)$ for each state of $(q(t), IP(t), I_a(t), I_s(t))$ as shown in Table 3.1. We use $\tilde{w}(t)$ to denote this new variable, which is defined as $\tilde{w}(t) = 1 + w(t)/g_{cd}$, if $q(t) = 0$; and $\tilde{w}(t) = (w(t) + k)/g_{cd}$, if $q(t) > 0$, where $k = q(t)$ modulo g_{cd} (i.e., k is the remainder of the division of $q(t)$ by g_{cd}). While $w(t)$ is the actual number of finished products in the production facility, which takes values from 0 to $q_2 - 1$, $\tilde{w}(t)$ is an artificial variable with $\tilde{q}_2$ states. The variable $w(t)$ is determined by $\tilde{w}(t)$ and $q(t)$ in a way specified in Table 3.1, if $q(0) = w(0) = 0$. For instance, if $q(t) = 0$ and $\tilde{w}(t) = 3$, then $w(t) = 2g_{cd}$. Note that if q_1 and q_2 are co-prime, i.e., $g_{cd} = 1$, then $w(t) = \tilde{w}(t) - 1$. Table 3.1 shall be used repeatedly throughout this paper for the relationship between $w(t)$, $q(t)$, and $\tilde{w}(t)$.

Remark 3.2. Without loss of generality, we can assume that $q(0) = 0$ and $w(0) = w_0$. Then $w(t)$ can be one of the integers in $\{(jg_{cd} + w_0 - k)$ modulo q_2, $j = 1, 2, \ldots, \tilde{q}_2\}$, if $q(t) = ng_{cd} + k$, for $n = 0, 1, 2, \ldots$, and $k = 1, 2, \ldots, g_{cd}$. Thus, the stationary distribution of the Markov chain can be found and the analysis can be carried out in a similar way. Yet the interpretation of the states of $\tilde{w}(t)$ may be slightly different from that of Table 3.1. For that reason, and since it reasonable to assume an empty system at $t = 0$, we assume that $q(0) = w(0) = 0$ throughout this paper.

According to Table 3.1 and Remark 3.2, there are exactly $\tilde{q}_2$ states of $\tilde{w}(t)$ associated with each state of $(q(t), IP(t), I_a(t), I_s(t))$, but the physical interpretations of the states can be different for different $q(t)$. Based on the physical interpretation of the states of $\tilde{w}(t)$ in Table 3.1, the transition of the state of $\tilde{w}(t)$ is given as follows.

(1) If $q(t)$ goes from $ng_{cd}+1$ to ng_{cd}, $\tilde{w}(t)$ (or corresponding $w(t)$) goes from j (or $jg_{cd}-1$) to $j+1$ (or jg_{cd}), if $j<\tilde{q}_2$; from $\tilde{q}_2$ to 1, if $j=\tilde{q}_2$. By Table 3.1, the corresponding transition of $\tilde{w}(t)$ is governed by the following transition probability matrix

$$U=\begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 1 & & & 0 \end{pmatrix}_{\tilde{q}_2\times\tilde{q}_2}, \tag{3.5}$$

which is a double stochastic matrix.

(2) For all the other cases, the transition of $\tilde{w}(t)$ is governed by the identity matrix I.

We call $q(t)$ the *level variable* and $(IP(t), I_a(t), I_s(t), \tilde{w}(t))$ the (vector) *phase variable*. The state space of the CTMC $\{(q(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t\ge 0\}$ is given as follows. If $q(t)=0$, then $I_s(t)=0$, since there is no production. We have $IP(t)=1,2,\ldots,q_1$, $I_a(t)=1,2,\ldots,m_a$, and $\tilde{w}(t)=1,2,\ldots,\tilde{q}_2$. Then level zero has $q_1m_a\tilde{q}_2$ states. If $q(t)\ge 1$, we have $IP(t)=1,2,\ldots,q_1$, $I_a(t)=1,2,\ldots,m_a$, $I_s(t)=1,2,\ldots,m_s$, and $\tilde{w}(t)=1,2,\ldots,\tilde{q}_2$. Such a level has $q_1m_am_s\tilde{q}_2$ states.

Utilizing matrices $J(q_1)$ and $L(q_1)$ defined in Eq. (3.3), the infinitesimal generator Q_w of $\{(q(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t\ge 0\}$ can be written as

$$Q_w=\begin{pmatrix} Q_{IP,0}\otimes I & 0 & \cdots & 0 & Q_{IP,1}\otimes\boldsymbol{\alpha}\otimes I & & & \\ I\otimes\mathbf{T}^0\otimes U & (Q_{IP,0}\oplus T)\otimes I & 0 & \cdots & 0 & Q_{IP,1}\otimes I\otimes I & & \\ & I\otimes(\mathbf{T}^0\boldsymbol{\alpha})\otimes I & (Q_{IP,0}\oplus T)\otimes I & 0 & \cdots & 0 & Q_{IP,1}\otimes I\otimes I & \\ & & \ddots & \ddots & \ddots & \cdots & \ddots & \ddots \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{3.6}$$

In Q_w, for $n=0,1,2,\ldots$, and $k=1,2,\ldots,g_{cd}$, the transition block from the level $q(t)=ng_{cd}+k$ to the level $q(t)=ng_{cd}+k-1$ is $I\otimes(\mathbf{T}^0\boldsymbol{\alpha})\otimes U$, if $k=1$; $I\otimes(\mathbf{T}^0\boldsymbol{\alpha})\otimes I$, otherwise. Note that the transition block from level 1 to level 0 is $I\otimes\mathbf{T}^0\otimes U$.

Eq. (3.6) indicates that Q_w is a level dependent $M/G/1$ type Markov chain, if $q(t)$ is considered as the level variable. To analyze $\{(q(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t\ge 0\}$ effectively, we need the level independent property. Fortunately, the level independent property can be obtained by re-blocking Q_w. We define a super level n as the set of all the states with $q(t)=(n-1)g_{cd}+1, (n-1)g_{cd}+2,\ldots,nq_{cd}$, for $n=1,2,\ldots$. That is: we re-block the matrix Q_w given in Eq. (3.6) such that each new block contains consecutive g_{cd} old blocks (except level zero) to obtain a level independent $M/G/1$ type Markov chain.

$$Q_w=\begin{pmatrix} A_{0,0} & 0 & \cdots & 0 & A_{0,\tilde{q}_1} & & & & \\ A_{1,0} & A_1 & 0 & \cdots & 0 & A_{\tilde{q}_1+1} & & & \\ & A_0 & A_1 & 0 & \cdots & 0 & A_{\tilde{q}_1+1} & & \\ & & A_0 & A_1 & 0 & \cdots & 0 & A_{\tilde{q}_1+1} & \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \tag{3.7}$$

with

$$\begin{aligned} A_{0,0}&=Q_{IP,0}\otimes I_{\tilde{q}_2\times\tilde{q}_2},\\ A_{0,\tilde{q}_1}&=\begin{pmatrix}0 & \cdots & 0 & Q_{IP,1}\otimes\boldsymbol{\alpha}\otimes I\end{pmatrix}_{(q_1m_a\tilde{q}_2)\times(g_{cd}q_1m_am_s\tilde{q}_2)}=\mathbf{v}(g_{cd})\otimes Q_{IP,1}\otimes\boldsymbol{\alpha}\otimes I_{\tilde{q}_2\times\tilde{q}_2},\\ A_{1,0}&=\begin{pmatrix} I\otimes\mathbf{T}^0\otimes U\\ 0\\ \vdots\\ 0\end{pmatrix}_{(g_{cd}q_1m_am_s\tilde{q}_2)\times(q_1m_a\tilde{q}_2)}=\mathbf{e}(1)\otimes I_{(q_1m_a)\times(q_1m_a)}\otimes\mathbf{T}^0\otimes U, \end{aligned} \tag{3.8}$$

$$A_{\tilde{q}_1+1} = \begin{pmatrix} Q_{IP,1} \otimes I_{m_s \times m_s} \otimes I_{\tilde{q}_2 \times \tilde{q}_2} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & Q_{IP,1} \otimes I_{m_s \times m_s} \otimes I_{\tilde{q}_2 \times \tilde{q}_2} \end{pmatrix}_{(g_{cd} q_1 m_a m_s \tilde{q}_2) \times (g_{cd} q_1 m_a m_s \tilde{q}_2)}$$
$$= I_{g_{cd} \times g_{cd}} \otimes Q_{IP,1} \otimes I_{m_s \times m_s} \otimes I_{\tilde{q}_2 \times \tilde{q}_2},$$
$$A_1 = I_{g_{cd} \times g_{cd}} \otimes \left(Q_{IP,0} \oplus T\right) \otimes I_{\tilde{q}_2 \times \tilde{q}_2} + J(g_{cd}) \otimes I_{(q_1 m_a) \times (q_1 m_a)} \otimes (\mathbf{T}^0 \boldsymbol{\alpha}) \otimes I_{\tilde{q}_2 \times \tilde{q}_2},$$
$$A_0 = \begin{pmatrix} 0 & & I \otimes (\mathbf{T}^0 \boldsymbol{\alpha}) \otimes U \\ & \ddots & \\ & & \ddots \\ & & 0 \end{pmatrix}_{(g_{cd} q_1 m_a m_s \tilde{q}_2) \times (g_{cd} q_1 m_a m_s \tilde{q}_2)}$$
$$= (\mathbf{e}(1)\mathbf{v}(g_{cd})) \otimes I_{(q_1 m_a) \times (q_1 m_a)} \otimes (\mathbf{T}^0 \boldsymbol{\alpha}) \otimes U,$$

where $\mathbf{v}(g_{cd})$ and $\mathbf{e}(1)$ are defined after Eq. (3.3), but the size of the vectors is g_{cd}.

It is clear that, after re-blocking, Q_w is now associated with an $M/G/1$ type Markov chain $\{(X(t), IS(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t \geq 0\}$ with a level independent transition structure. The relationship between $q(t)$, $X(t)$, and $IS(t)$ is given as

$$X(t) = \lfloor (q(t) - 1)/g_{cd} \rfloor + 1; \qquad (3.9)$$
$$IS(t) = q(t) - g_{cd} \max\{0,\ X(t) - 1\},$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

We call $X(t)$ the new *level variable* and $(IS(t), IP(t), I_a(t), I_s(t), \tilde{w}(t))$ the new (vector) *phase variable*. If $X(t) = 0$, $IS(t)$ and $I_s(t)$ are irrelevant, and we have $IP(t) = 1, 2, \ldots, q_1$, $I_a(t) = 1, 2, \ldots, m_a$, and $\tilde{w}(t) = 1, 2, \ldots, \tilde{q}_2$. The level zero has $q_1 m_a \tilde{q}_2$ states. If $X(t) \geq 1$, we have $IS(t) = 1, 2, \ldots, g_{cd}$, $IP(t) = 1, 2, \ldots, q_1$, $I_a(t) = 1, 2, \ldots, m_a$, and $I_s(t) = 1, 2, \ldots, m_s$, and $\tilde{w}(t) = 1, 2, \ldots, \tilde{q}_2$ values. Such a level has $g_{cd} q_1 m_a m_s \tilde{q}_2 = q_1 m_a m_s q_2$ states.

Taking advantage of the $M/G/1$ structure in the infinitesimal generator Q_w, a matrix-analytic solution can be obtained for the stationary distribution of $\{(X(t), (IS(t), IP(t), I_a(t), I_s(t), \tilde{w}(t))), t \geq 0\}$ by using Ramaswami's algorithm [31]. Denote by $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$, where $\boldsymbol{\pi}_0 = (\pi_{0,1,1,1}, \ldots, \pi_{0,q_1,m_a,\tilde{q}_2})$ and $\boldsymbol{\pi}_q = (\pi_{q,1,1,1,1}, \ldots, \pi_{q,q_1,m_a,m_s,\tilde{q}_2})$, $q = 1, 2, \ldots$, the stationary distribution of the CTMC $\{(q(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t \geq 0\}$. The elements in $\boldsymbol{\pi}_q$ are ordered lexicographically. It is well-known that $\boldsymbol{\pi}$ satisfies $\boldsymbol{\pi} Q_w = 0$ and $\boldsymbol{\pi}\mathbf{e} = 1$. To find the stationary distribution $\boldsymbol{\pi}$, we need to utilize the level independent structure given in Eq. (3.7). Similar to the re-blocking of Q_w, we re-block the vector $\boldsymbol{\pi}$ as follows: $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_{MG1,1}, \boldsymbol{\pi}_{MG1,2}, \ldots)$, where $\boldsymbol{\pi}_{MG1,n} = (\boldsymbol{\pi}_{(n-1)g_{cd}+1}, \boldsymbol{\pi}_{(n-1)g_{cd}+2}, \ldots, \boldsymbol{\pi}_{ng_{cd}})$, for $n = 1, 2, \ldots$. The stationary distribution $\boldsymbol{\pi}$ is partitioned into $(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$ for analysis and proofs and into $(\boldsymbol{\pi}_0, \boldsymbol{\pi}_{MG1,1}, \boldsymbol{\pi}_{MG1,2}, \ldots)$ for computation.

The stationary distribution $\boldsymbol{\pi}$ can be computed by Ramaswami's algorithm [31] as follows. Let matrix G be the minimal nonnegative solution to matrix equation

$$0 = A_0 + A_1 G + A_{\tilde{q}_1+1} G^{\tilde{q}_1+1}. \qquad (3.10)$$

Then we have, if $\boldsymbol{\pi}$ exists,

$$\boldsymbol{\pi}_{MG1,n} = \left(\boldsymbol{\pi}_0 \hat{B}_n + \sum_{k=2}^{\min\{n,\tilde{q}_1+1\}} \boldsymbol{\pi}_{MG1,n-k+1} B_k\right)(-B_1)^{-1}, \quad n = 1, 2, \ldots; \qquad (3.11)$$
$$\boldsymbol{\pi}_0 (A_{0,0} + \hat{B}_1 (-B_1)^{-1} A_{1,0}) = 0,$$
$$\boldsymbol{\pi}_0 \mathbf{e} = 1 - \lambda/\mu,$$

where $\hat{B}_n = A_{0,\tilde{q}_1} G^{\tilde{q}_1 - n}$, for $n = 1, 2, \ldots, \tilde{q}_1$; $\hat{B}_n = 0$, for $n > \tilde{q}_1$; and $B_1 = A_1 + A_{\tilde{q}_1+1} G^{\tilde{q}_1}$, $B_n = A_{\tilde{q}_1+1} G^{\tilde{q}_1+1-n}$, for $n = 2, 3, \ldots, \tilde{q}_1 + 1$; $B_n = 0$, for $n > \tilde{q}_1 + 1$. The existence of the stationary distribution and the normalization factor $\boldsymbol{\pi}_0 \mathbf{e} = 1 - \lambda/\mu$ are shown in the following proposition.

Proposition 3.3. *The continuous time Markov chain* $\{(q(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t \geq 0\}$ *(or* $\{(X(t), IS(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t \geq 0\}$*) defined by* Q_w *in Eq. (3.6) is irreducible. If* $\lambda/\mu < 1$*, the Markov chain is positive recurrent and its stationary distribution* $\boldsymbol{\pi}$ *is given in Eq. (3.11).*

Proof. First, the Markov chain $\{(q(t), IP(t), I_a(t), I_s(t)), t \geq 0\}$ is irreducible since the representations of demand arrival process and the *PH*-distribution are irreducible. The irreducibility of $\{(q(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t \geq 0\}$ is due to the chosen representation $\tilde{w}(t)$ for the states of $w(t)$ (see Proposition 3.2).

The ergodicity condition is obtained by using Neuts condition for $M/G/1$ type Markov chains in a straightforward manner [25,27]. First, we have

$$A_0 + A_1 + A_{\tilde{q}_1+1} = \begin{pmatrix} (Q_{IP} \oplus T) \otimes I & & & I \otimes (\mathbf{T}^0\boldsymbol{\alpha}) \otimes U \\ I \otimes (\mathbf{T}^0\boldsymbol{\alpha}) \otimes I & \ddots & & \\ & \ddots & \ddots & \\ & & I \otimes (\mathbf{T}^0\boldsymbol{\alpha}) \otimes I & (Q_{IP} \oplus T) \otimes I \end{pmatrix}. \tag{3.12}$$

It is easy to verify that $\boldsymbol{\theta}_A = (\mathbf{e}_{g_{cd}})' \otimes (\mathbf{e}_{q_1})' \otimes \boldsymbol{\theta}_a \otimes \boldsymbol{\theta}_s \otimes (\mathbf{e}_{\tilde{q}_2})'/(q_1q_2)$ is an invariant vector of $A_0 + A_1 + A_{\tilde{q}_1+1}$, where $\mathbf{e}'$ is the transpose of the column vector $\mathbf{e}$. By routine calculations, we obtain $\boldsymbol{\theta}_A(A_1 + (\tilde{q}_1 + 1)A_{\tilde{q}_1+1})\mathbf{e} = (\lambda - \mu)/g_{cd} < 0$. By Neuts [27], the CTMC is ergodic and its stationary distribution $\boldsymbol{\pi}$ exists. By Ramaswami's algorithm, Eq. (3.11) is obtained, except for the explicit expression for $\boldsymbol{\pi}_0\mathbf{e}$.

To find $\boldsymbol{\pi}_0\mathbf{e}$, we define $\boldsymbol{\varpi}_1^*(z) = \sum_{q=1}^{\infty} \boldsymbol{\pi}_q(I_{(q_1m_am_s)\times(q_1m_am_s)} \otimes \mathbf{e}_{\tilde{q}_2})z^q$, for $0 \le z \le 1$. By Eq. (3.6), the equation $\boldsymbol{\pi}Q_w = 0$ can be re-written as follows:

$$\begin{aligned}
&\boldsymbol{\pi}_0(Q_{IP,0} \otimes I) + \boldsymbol{\pi}_1(I_{(q_1m_a)\times(q_1m_a)} \otimes \mathbf{T}^0 \otimes U) = 0;\\
&1 \le ng_{cd} \le q_1 - 1: \ \boldsymbol{\pi}_{ng_{cd}}((Q_{IP,0} \oplus T) \otimes I) + \boldsymbol{\pi}_{ng_{cd}+1}(I_{(q_1m_a)\times(q_1m_a)} \otimes (\mathbf{T}^0\boldsymbol{\alpha}) \otimes U) = 0;\\
&\begin{matrix} 1 \le ng_{cd} + k \le q_1 - 1, \\ 1 \le k \le g_{cd} - 1: \end{matrix} \ \boldsymbol{\pi}_{ng_{cd}+k}((Q_{IP,0} \oplus T) \otimes I) + \boldsymbol{\pi}_{ng_{cd}+k+1}(I_{(q_1m_a)\times(q_1m_a)} \otimes (\mathbf{T}^0\boldsymbol{\alpha}) \otimes I) = 0;\\
&\boldsymbol{\pi}_0(Q_{IP,1} \otimes \boldsymbol{\alpha} \otimes I) + \boldsymbol{\pi}_{q_1}((Q_{IP,0} \oplus T) \otimes I) + \boldsymbol{\pi}_{q_1+1}(I_{(q_1m_a)\times(q_1m_a)} \otimes (\mathbf{T}^0\boldsymbol{\alpha}) \otimes U) = 0;\\
&q_1 + 1 \le ng_{cd}: \ \boldsymbol{\pi}_{ng_{cd}-q_1}(Q_{IP,1} \otimes I \otimes I) + \boldsymbol{\pi}_{ng_{cd}}((Q_{IP,0} \oplus T) \otimes I) + \boldsymbol{\pi}_{ng_{cd}+1}(I_{(q_1m_a)\times(q_1m_a)} \otimes (\mathbf{T}^0\boldsymbol{\alpha}) \otimes U) = 0;\\
&\begin{matrix} q_1 + 1 \le ng_{cd} + k, \\ 1 \le k \le g_{cd} - 1: \end{matrix} \ \boldsymbol{\pi}_{ng_{cd}+k-q_1}(Q_{IP,1} \otimes I \otimes I)\\
&\qquad + \boldsymbol{\pi}_{ng_{cd}+k}((Q_{IP,0} \oplus T) \otimes I) + \boldsymbol{\pi}_{ng_{cd}+k+1}(I_{(q_1m_a)\times(q_1m_a)} \otimes (\mathbf{T}^0\boldsymbol{\alpha}) \otimes I) = 0.
\end{aligned} \tag{3.13}$$

Using Eq. (3.13), by routine calculations, we obtain

$$\boldsymbol{\varpi}_1^*(z)\left((Q_{IP,0} + z^{q_1}Q_{IP,1}) \oplus \left(T + \frac{1}{z}\mathbf{T}^0\boldsymbol{\alpha}\right)\right) = -\boldsymbol{\pi}_0(I \otimes \mathbf{e}_{\tilde{q}_2})((Q_{IP,0} + z^{q_1}Q_{IP,1}) \otimes \boldsymbol{\alpha}). \tag{3.14}$$

To obtain Eq. (3.14), we multiply by $I_{(q_1m_a)\times(q_1m_a)} \otimes \boldsymbol{\alpha} \otimes I_{\tilde{q}_2\times\tilde{q}_2}$ on both sides of the first equation in Eq. (3.13) to ensure that all the vectors are of the same size in all equations in Eq. (3.13). Letting $z = 1$ in Eq. (3.14), we obtain

$$\boldsymbol{\pi}_0(I \otimes \mathbf{e}_{\tilde{q}_2})(Q_{IP} \otimes \boldsymbol{\alpha}) + \boldsymbol{\varpi}_1^*(1)(Q_{IP} \oplus (T + \mathbf{T}^0\boldsymbol{\alpha})) = 0. \tag{3.15}$$

Post-multiplying by $I_{(q_1m_a)\times(q_1m_a)} \otimes \mathbf{e}_{m_s}$ on both sides of Eq. (3.15) yields $\left(\boldsymbol{\pi}_0(I \otimes \mathbf{e}_{\tilde{q}_2}) + \boldsymbol{\varpi}_1^*(1)(I \otimes \mathbf{e}_{m_s})\right) Q_{IP} = 0$, which implies that $\left(\boldsymbol{\pi}_0(I \otimes \mathbf{e}) + \boldsymbol{\varpi}_1^*(1)(I \otimes \mathbf{e})\right) = (\boldsymbol{\theta}_a, \ldots, \boldsymbol{\theta}_a)/q_1$. Note that $\boldsymbol{\pi}_0(I \otimes \mathbf{e})\mathbf{e} + \boldsymbol{\varpi}_1^*(1)(I \otimes \mathbf{e})\mathbf{e} = 1$.

Post-multiplying by $\mathbf{e}_{q_1m_a} \otimes I_{m_s\times m_s}$ on both sides of Eq. (3.15) yields $\boldsymbol{\varpi}_1^*(1)(\mathbf{e} \otimes I)\left(T + \mathbf{T}^0\boldsymbol{\alpha}\right) = 0$, which implies that $\boldsymbol{\varpi}_1^*(1)(\mathbf{e} \otimes I) = c\boldsymbol{\theta}_s$, where $c = \boldsymbol{\varpi}_1^*(1)\mathbf{e}$. It is clear that $c = 1 - \boldsymbol{\pi}_0\mathbf{e}$ or $\boldsymbol{\pi}_0\mathbf{e} = 1 - c$. To find c, we take derivatives on both sides of Eq. (3.14), let $z = 1$, and multiply on both sides by $\mathbf{e}_{q_1m_am_s}$, to obtain

$$\begin{aligned}
0 &= \boldsymbol{\varpi}_1^*(1)(q_1(\mathbf{e}(1) \otimes (D_1\mathbf{e}) \otimes \mathbf{e}) - \mathbf{e} \otimes \mathbf{e} \otimes \mathbf{T}^0) + q_1\boldsymbol{\pi}_0(\mathbf{e}(1) \otimes (D_1\mathbf{e}) \otimes \mathbf{e})\\
&= q_1\boldsymbol{\varpi}_1^*(1)(\mathbf{e}(1) \otimes (D_1\mathbf{e}) \otimes \mathbf{e}) - \boldsymbol{\varpi}_1^*(1)(\mathbf{e} \otimes \mathbf{e} \otimes \mathbf{T}^0) + q_1\boldsymbol{\pi}_0(\mathbf{e}(1) \otimes (D_1\mathbf{e}) \otimes \mathbf{e})\\
&= q_1(\boldsymbol{\pi}_0(I \otimes \mathbf{e}) + \boldsymbol{\varpi}_1^*(1)(I \otimes \mathbf{e}))(\mathbf{e}(1) \otimes (D_1\mathbf{e})) - \boldsymbol{\varpi}_1^*(1)(\mathbf{e} \otimes \mathbf{e} \otimes I)\mathbf{T}^0\\
&= (\boldsymbol{\theta}_a, \ldots, \boldsymbol{\theta}_a)(\mathbf{e}(1) \otimes (D_1\mathbf{e})) - c\boldsymbol{\theta}_s\mathbf{T}^0\\
&= \lambda - c\mu,
\end{aligned} \tag{3.16}$$

which leads to the expected result. This completes the proof of Proposition 3.3. □

Remark 3.3. The re-blocking technique can be applied to Q_w given in Eq. (3.7) to generate a QBD structure. However, the space complexity of the QBD approach is significantly higher than that of the $M/G/1$ approach. Thus, we do not explore the QBD approach in this paper.

Remark 3.4. Since the underlying Markov chain of the demand arrival process is not affected by inventory management, we must have $\boldsymbol{\pi}_0(\mathbf{e}_{q_1} \otimes I_{m_a\times m_a} \otimes \mathbf{e}_{\tilde{q}_2}) + \sum_{q=1}^{\infty} \boldsymbol{\pi}_q(\mathbf{e}_{q_1} \otimes I_{m_a\times m_a} \otimes \mathbf{e}_{m_s} \otimes \mathbf{e}_{\tilde{q}_2}) = \boldsymbol{\theta}_a$, which can be used to check computation accuracy.

4. Performance measures

In this section, performance measures $\{E[IP(t)], E[w(t)], E[q(t)], E[(r + IP(t) - q(t) - w(t))^+], E[(q(t) + w(t) - r - IP(t))^+]\}$ are obtained either explicitly or in terms of the stationary distribution $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$. We begin with $IP(t)$ whose distribution can be found explicitly.

Proposition 4.1. *The inventory position $IP(t)$ has a uniform distribution on $\{1, 2, \ldots, q_1\}$. Consequently, we have $E[IP(t)] = (q_1 + 1)/2$.*

Proof. The result is obtained by Proposition 3.1. This completes the proof of Proposition 4.1. □

Remark 4.1. By definition, $E[IP(t)]$ can be expressed in terms of $\boldsymbol{\pi}$ as follows:

$$E[IP(t)] = \sum_{j=1}^{q_1} j \left(\boldsymbol{\pi}_0(\mathbf{e}(j) \otimes \mathbf{e}_{m_a\tilde{q}_2}) + \sum_{n=1}^{\infty} \boldsymbol{\pi}_{MG1,n}(\mathbf{e}_{g_{cd}} \otimes \mathbf{e}(j) \otimes \mathbf{e}_{m_a m_s \tilde{q}_2}) \right), \tag{4.1}$$

where $\mathbf{e}(j)$ is a column vector of size q_1 with the j-th element being one and all others zero. The above two expressions of $E[IP(t)]$ can be used for checking the computation accuracy of $\boldsymbol{\pi}$.

Let $\rho = \min\{1, \lambda/\mu\}$.

Proposition 4.2. *The distribution of $w(t)$ is given as, for $j = 0, 1, \ldots, \tilde{q}_2 - 1$,*

$$P\{w(t) = jg_{cd} + k\} = \begin{cases} \dfrac{1-\rho}{\tilde{q}_2} + \dfrac{\rho}{q_2}, & \text{if } k = 0; \\[2mm] \dfrac{\rho}{q_2}, & \text{if } k = 1, 2, \ldots, g_{cd} - 1. \end{cases} \tag{4.2}$$

Consequently, we have $E[w(t)] = (q_2 - \rho - g_{cd}(1-\rho))/2$. In particular, if $g_{cd} = 1$, then $w(t)$ has a discrete uniform distribution on $\{0, 1, \ldots, q_2 - 1\}$ and $E[w(t)] = (q_2 - 1)/2$; if $g_{cd} = q_2$, then $E[w(t)] = \rho(q_2 - 1)/2$.

Proof. Using Eq. (3.13), we first show that all elements of the vector $\boldsymbol{\pi}_q(\mathbf{e} \otimes I_{\tilde{q}_2\times\tilde{q}_2})$ are the same for $q = 0, 1, 2, \ldots$, i.e., $\boldsymbol{\pi}_q(\mathbf{e} \otimes I_{\tilde{q}_2\times\tilde{q}_2}) = (\boldsymbol{\pi}_q\mathbf{e})\mathbf{e}'/\tilde{q}_2$. Note that elements in the vector $\boldsymbol{\pi}_q(\mathbf{e} \otimes I)$ is the joint probability of the queue length $q(t)$ and $\tilde{w}(t)$, for $q = 0, 1, 2, \ldots$. Define $\boldsymbol{\psi}_0 = \boldsymbol{\pi}_0(I \otimes U)$, and $\boldsymbol{\psi}_q = \boldsymbol{\pi}_q(I \otimes U)$, for $q = 1, 2, \ldots$. Let $\boldsymbol{\psi} = (\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \ldots)$. It is easy to verify that $\boldsymbol{\psi}\mathbf{e} = \boldsymbol{\pi}\mathbf{e} = 1$, and $\boldsymbol{\psi}Q_w = \boldsymbol{\pi}Q_w(I \otimes U)$. Since $\boldsymbol{\pi}Q_w = 0$, we obtain $\boldsymbol{\psi}Q_w = 0$. Since the stationary distribution of the CTMC Q_w is unique, we must have $\boldsymbol{\psi} = \boldsymbol{\pi}$, i.e., $\boldsymbol{\psi}_q = \boldsymbol{\pi}_q = \boldsymbol{\pi}_q(I \otimes U)$ for $q = 0, 1, 2, \ldots$. Consequently, we must have $\boldsymbol{\pi}_q(\mathbf{e} \otimes I) = \boldsymbol{\pi}_q(\mathbf{e} \otimes U)$ for $q = 0, 1, 2, \ldots$, which implies that the $\tilde{q}_2$ elements in the vector $\boldsymbol{\pi}_q(\mathbf{e} \otimes I)$ are identical, i.e., $\boldsymbol{\pi}_q(\mathbf{e} \otimes I) = (\boldsymbol{\pi}_q\mathbf{e}, \ldots, \boldsymbol{\pi}_q\mathbf{e})/\tilde{q}_2$, for $q = 0, 1, 2, \ldots$.

Second, we show the vectors $\boldsymbol{\phi}_k = \sum_{n=0}^{\infty} \boldsymbol{\pi}_{ng_{cd}+k}(\mathbf{e} \otimes I_{(m_s\tilde{q}_2)\times(m_s\tilde{q}_2)})$ are identical vectors for $k = 1, 2, \ldots, g_{cd}$. Post-multiplying by $\mathbf{e} \otimes \boldsymbol{\alpha} \otimes I_{\tilde{q}_2\times\tilde{q}_2}$ on both sides of the first equation in Eq. (3.13), post-multiplying by $\mathbf{e} \otimes I_{m_s\times m_s} \otimes I_{\tilde{q}_2\times\tilde{q}_2}$ on both sides of all the equations corresponding to $q(t) = ng_{cd}$, for $n = 1, 2, \ldots$, and adding up the resulting equations, we obtain

$$\begin{aligned} 0 &= \boldsymbol{\pi}_0(Q_{IP}\mathbf{e} \otimes \boldsymbol{\alpha} \otimes I) + \sum_{n=0}^{\infty}(\boldsymbol{\pi}_{ng_{cd}+g_{cd}}(Q_{IP}\mathbf{e} \otimes I \otimes I + \mathbf{e} \otimes T \otimes I)) \\ &\quad + \sum_{n=0}^{\infty}(\boldsymbol{\pi}_{ng_{cd}+1}(Q_{IP}\mathbf{e} \otimes I \otimes I + \mathbf{e} \otimes (\mathbf{T}^0\boldsymbol{\alpha}) \otimes U)) \\ &= \left(\sum_{n=0}^{\infty} \boldsymbol{\pi}_{ng_{cd}+g_{cd}}(\mathbf{e} \otimes I \otimes I)\right)(T \otimes I) + \left(\sum_{n=0}^{\infty} \boldsymbol{\pi}_{ng_{cd}+1}(\mathbf{e} \otimes I \otimes I)\right)((\mathbf{T}^0\boldsymbol{\alpha}) \otimes U). \end{aligned} \tag{4.3}$$

Note that $Q_{IP}\mathbf{e} = 0$. Then Eq. (4.3) can be rewritten as

$$0 = \boldsymbol{\phi}_{g_{cd}}(T \otimes I_{\tilde{q}_2\times\tilde{q}_2}) + \boldsymbol{\phi}_1((\mathbf{T}^0\boldsymbol{\alpha}) \otimes U). \tag{4.4}$$

In a similar way, it can be shown that $0 = \boldsymbol{\phi}_k(T \otimes I_{\tilde{q}_2\times\tilde{q}_2}) + \boldsymbol{\phi}_{k+1}((\mathbf{T}^0\boldsymbol{\alpha}) \otimes I_{\tilde{q}_2\times\tilde{q}_2})$, for $k = 1, 2, \ldots, g_{cd} - 1$. Together, we have shown

$$0 = (\boldsymbol{\phi}_1, \ \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{g_{cd}}) \begin{pmatrix} T \otimes I & & & & (\mathbf{T}^0\boldsymbol{\alpha}) \otimes U \\ (\mathbf{T}^0\boldsymbol{\alpha}) \otimes I & T \otimes I & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & (\mathbf{T}^0\boldsymbol{\alpha}) \otimes I & T \otimes I \end{pmatrix}. \tag{4.5}$$

Eq. (4.5) indicates that $(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{g_{cd}})$ is the steady-state aggregation of the embedded process related to $A = A_0 + A_1 + A_{\tilde{q}_1}$ given in Eq. (3.12).

Recall that $\boldsymbol{\theta}_s$ is a row vector satisfying $\boldsymbol{\theta}_s(T + \mathbf{T}^0\boldsymbol{\alpha}) = 0$, $\boldsymbol{\theta}_s \geq 0$, and $\boldsymbol{\theta}_s\mathbf{e} = 1$. Since $\mathbf{e}'U = \mathbf{e}'$, $\boldsymbol{\phi}_k = c\boldsymbol{\theta}_s \otimes \mathbf{e}'$, for $k = 1, 2, \ldots, g_{cd}$, is a solution to Eq. (4.5), which is unique up to a positive constant. Consequently, we have $\boldsymbol{\phi}_k(\mathbf{e}_{m_s} \otimes I_{\tilde{q}_2\times\tilde{q}_2}) = c(\mathbf{e}_{\tilde{q}_2})'$, for $k = 1, 2, \ldots, g_{cd}$. Since $(\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2 + \cdots + \boldsymbol{\phi}_{g_{cd}})\mathbf{e} = 1 - \boldsymbol{\pi}_0\mathbf{e}$ and $q_2 = g_{cd}\tilde{q}_2$, we obtain $c = (1 - \boldsymbol{\pi}_0\mathbf{e})/q_2$. Hence, we have shown that

$$\sum_{n=0}^{\infty} \boldsymbol{\pi}_{ng_{cd}+k}(\mathbf{e} \otimes I_{\tilde{q}_2\times\tilde{q}_2}) = \frac{(1 - \boldsymbol{\pi}_0\mathbf{e})}{q_2}\mathbf{e}', \quad \text{for } k = 1, 2, \ldots, g_{cd}. \tag{4.6}$$

Eq. (4.6) indicates that, if $q(t) = k$ modulo g_{cd}, then $\tilde{w}(t)$ is uniformly distributed on $\{1, 2, \ldots, \tilde{q}_2\}$. By the interpretation given in Table 3.1, $w(t) = jg_{cd} + g_{cd} - k$ implies that $q(t) = k$ modulo g_{cd}, and $\tilde{w}(t) = j + 1$, for $j = 0, 1, \ldots, \tilde{q}_2 - 1$. Then we obtain, for $j = 0, 1, 2, \ldots, \tilde{q}_2 - 1$,

$$\begin{aligned} P\{w(t) = jg_{cd}\} &= (\boldsymbol{\pi}_0(\mathbf{e} \otimes I_{\tilde{q}_2\times\tilde{q}_2}))_j + \sum_{n=1}^{\infty}(\boldsymbol{\pi}_{ng_{cd}}(\mathbf{e} \otimes I_{\tilde{q}_2\times\tilde{q}_2}))_j \\ &= \boldsymbol{\pi}_0\mathbf{e}/\tilde{q}_2 + (1 - \boldsymbol{\pi}_0\mathbf{e})/q_2, \end{aligned} \tag{4.7}$$

and, for $k = 1, 2, \ldots, g_{cd} - 1$,

$$P\{w(t) = jg_{cd} + k\} = \sum_{n=0}^{\infty}(\boldsymbol{\pi}_{ng_{cd}+k}(\mathbf{e} \otimes I_{\tilde{q}_2\times\tilde{q}_2}))_j = (1 - \boldsymbol{\pi}_0\mathbf{e})/q_2. \tag{4.8}$$

The results are obtained since $\boldsymbol{\pi}_0\mathbf{e} = 1 - \lambda/\mu$. The expectation of $w(t)$ is obtained by routine calculations. This completes the proof of Proposition 4.2. □

This proposition can be interpreted intuitively as follows. If $g_{cd} = 1$, i.e., q_1 and q_2 are co-prime, $w(t)$ has the same probability of being any number between 0 and q_2 for any queue length. Therefore, $w(t)$ must have a discrete uniform distribution on $\{0, 1, 2, \ldots, q_2 - 1\}$. If $g_{cd} > 1$, (i) if the server is idle, $w(t)$ has the same probability of being $0, g_{cd}, 2g_{cd}, \ldots,$ and $(\tilde{q}_2 - 1)g_{cd}$; (ii) if the server is busy, $w(t)$ has the same (conditional) probability of being any number between 0 and $q_2 - 1$. Then the distribution of $w(t)$ is obtained by adding up the probabilities for individual values of $q(t)$.

It is interesting to see that the marginal distribution of $w(t)$ does not depend on the stationary distribution $\boldsymbol{\pi}$, but only on original system parameters. On the other hand, numerical results indicate that the conditional distributions of $w(t)$ and $\tilde{w}(t)$ are not independent of $q(t)$. Therefore, to find the expected total cost $C(r, q_1, q_2)$ (see Eq. (2.1)), we still need to consider the CTMC $\{(q(t), IP(t), I_a(t), I_s(t), \tilde{w}(t)), t \geq 0\}$, in steady of $\{(q(t), IP(t), I_a(t), I_s(t)), t \geq 0\}$.

Remark 4.2. By definition, $E[w(t)]$ can also be expressed in terms of $\boldsymbol{\pi}$ as follows:

$$E[w(t)] = \boldsymbol{\pi}_0\left(\mathbf{e} \otimes \begin{pmatrix} 0 \\ g_{cd} \\ \vdots \\ (\tilde{q}_2 - 1)g_{cd} \end{pmatrix}\right) + \sum_{n=1}^{\infty}\sum_{j=0}^{\tilde{q}_1-1}\sum_{k=1}^{g_{cd}} \boldsymbol{\pi}_{n,jg_{cd}+k}\left(\mathbf{e} \otimes \begin{pmatrix} g_{cd} - k \\ 2g_{cd} - k \\ \vdots \\ \tilde{q}_2 g_{cd} - k \end{pmatrix}\right). \tag{4.9}$$

In Eq. (4.9), $\boldsymbol{\pi}_n$ is partitioned according to the variable $IP(t)$ as $\boldsymbol{\pi}_n = (\boldsymbol{\pi}_{n,1}, \boldsymbol{\pi}_{n,2}, \ldots, \boldsymbol{\pi}_{n,q_1})$.

Proposition 4.3. *Assume that $\rho < 1$. For $q(t)$, we have the following results.*

(i) $E[q(t)] \geq \rho(q_1 + 1)/2$.
(ii)

$$E[q(t)] = \sum_{n=1}^{\infty} n\boldsymbol{\pi}_n\mathbf{e} = \frac{\mathbf{u}_1(q_1 Q_{IP,1} \otimes I - I \otimes \mathbf{T}^0\boldsymbol{\alpha})\mathbf{e} - \delta_1}{\mu - \lambda}, \tag{4.10}$$

where $\mathbf{u}_1$ and δ_1 are given in Eq. (4.13).

Proof. To prove part (i), we apply a sample path method. We first note that the queue length $q(t)$ increases in batches of size q_1. During a busy period, the server serves a number of batches of size q_1. The service of the batches cannot be overlapping, but one batch can be waiting while the server is serving (remaining) customers in another batch. Since service times are independent and identically distributed random variables, the mean queue length during the service of each batch has to be greater than or equal to $(q_1 + 1)/2$. Consequently, the mean queue length during a busy period is greater than or equal to $(q_1 + 1)/2$. By Proposition 3.3, the probability that the production facility is busy is ρ. Thus, by conditioning on the status of the production facility at (arbitrary) time t, we obtain part (i).

Part (ii) can be obtained by routine calculations for the first moment of the stationary distribution of $M/G/1$ type Markov chains (see Chapter 3 in [27]). Specifically, we take the first two derivatives of both sides in Eq. (3.14) to obtain

$$\begin{aligned}
&\boldsymbol{\varpi}_1^{*(1)}(z)\left((Q_{IP,0}+z^{q_1}Q_{IP,1})\oplus\left(T+\frac{1}{z}\mathbf{T}^0\boldsymbol{\alpha}\right)\right)+\boldsymbol{\varpi}_1^{*}(z)\left((q_1z^{q_1-1}Q_{IP,1})\oplus\left(-\frac{1}{z^2}\mathbf{T}^0\boldsymbol{\alpha}\right)\right)\\
&\quad=-\boldsymbol{\pi}_0(I\otimes\mathbf{e})((q_1z^{q_1-1}Q_{IP,1})\otimes\boldsymbol{\alpha});\\
&\boldsymbol{\varpi}_1^{*(2)}(z)\left((Q_{IP,0}+z^{q_1}Q_{IP,1})\oplus\left(T+\frac{1}{z}\mathbf{T}^0\boldsymbol{\alpha}\right)\right)+2\boldsymbol{\varpi}_1^{*(1)}(z)\left((q_1z^{q_1-1}Q_{IP,1})\oplus\left(-\frac{1}{z^2}\mathbf{T}^0\boldsymbol{\alpha}\right)\right)\\
&\quad+\boldsymbol{\varpi}_1^{*}(z)\left((q_1(q_1-1)z^{q_1-2}Q_{IP,1})\oplus\left(\frac{2}{z^3}\mathbf{T}^0\boldsymbol{\alpha}\right)\right)=-\boldsymbol{\pi}_0(I\otimes\mathbf{e})((q_1(q_1-1)z^{q_1-2}Q_{IP,1})\otimes\boldsymbol{\alpha}).
\end{aligned}\tag{4.11}$$

Let $Q_{as}=Q_{IP}\oplus(T+\mathbf{T}^0\boldsymbol{\alpha})=Q_{IP}\otimes I+I\otimes(T+\mathbf{T}^0\boldsymbol{\alpha})$ and $\boldsymbol{\theta}_{as}=\mathbf{e}'\otimes\boldsymbol{\theta}_a\otimes\boldsymbol{\theta}_s/q_1$. Then Q_{as} is an irreducible infinitesimal generator, $\boldsymbol{\theta}_{as}Q_{as}=0$ and $\boldsymbol{\theta}_{as}\mathbf{e}=1$. Then it can be shown that $Q_{as}+\mathbf{e}\boldsymbol{\theta}_{as}$ is invertible. By routine calculations, Eq. (4.11) leads to

$$\begin{aligned}
&\boldsymbol{\varpi}_1^{*(1)}(1)=(\boldsymbol{\varpi}_1^{*(1)}(1)\mathbf{e})\boldsymbol{\theta}_{as}+\mathbf{u}_1;\\
&\boldsymbol{\varpi}_1^{*(1)}(1)(q_1Q_{IP,1}\otimes I-I\otimes\mathbf{T}^0\boldsymbol{\alpha})\mathbf{e}=\delta_1,
\end{aligned}\tag{4.12}$$

where

$$\begin{aligned}
&\mathbf{u}_1=-(\boldsymbol{\varpi}_1^{*}(1)(q_1Q_{IP,1}\otimes I-I\otimes(\mathbf{T}^0\boldsymbol{\alpha}))+\boldsymbol{\pi}_0(I\otimes\mathbf{e})(q_1Q_{IP,1}\otimes\boldsymbol{\alpha}))(Q_{as}+\mathbf{e}\boldsymbol{\theta}_{as})^{-1};\\
&\delta_1=-\frac{1}{2}(\boldsymbol{\varpi}_1^{*}(1)(q_1(q_1-1)Q_{IP,1}\otimes I+2I\otimes(\mathbf{T}^0\boldsymbol{\alpha}))+\boldsymbol{\pi}_0(I\otimes\mathbf{e})((q_1(q_1-1)Q_{IP,1})\otimes\boldsymbol{\alpha}))\mathbf{e}.
\end{aligned}\tag{4.13}$$

Note that $\boldsymbol{\theta}_{as}(q_1Q_{IP,1}\otimes I-I\otimes\mathbf{T}^0\boldsymbol{\alpha})\mathbf{e}=\lambda-\mu$. Post-multiplying by $(q_1Q_{IP,1}\otimes I-I\otimes\mathbf{T}^0\boldsymbol{\alpha})\mathbf{e}$ on both sides of the first equation in Eq. (4.11), we obtain Eq. (4.10).

In Eq. (4.13), $\boldsymbol{\pi}_0$ can be obtained by solving Eq. (3.11). For $\boldsymbol{\varpi}_1^{*}(1)$, we use Eq. (3.14) in a similar way and the fact $\boldsymbol{\varpi}_1^{*}(1)\mathbf{e}=\rho$ to obtain

$$\boldsymbol{\varpi}_1^{*}(1)=\rho\boldsymbol{\theta}_{as}-\boldsymbol{\pi}_0(I\otimes\mathbf{e})(Q_{IP}\otimes\boldsymbol{\alpha})(Q_{as}+\mathbf{e}\boldsymbol{\theta}_{as})^{-1}.\tag{4.14}$$

This completes the proof of Proposition 4.3. □

Remark 4.3. The two expressions in Eq. (4.10) for $E[q(t)]$ can be used for checking the computation accuracy of $\boldsymbol{\pi}$.

Proposition 4.4. *Assume that $\rho<1$. Then we have*

$$\begin{aligned}
E[(r+IP(t)-q(t)-w(t))^+]&=\sum_{i=1}^{q_1}\sum_{i_a=1}^{m_a}\sum_{j=1}^{\tilde{q}_2}\pi_{0,i,i_a,w}(r+i-(j-1)g_{cd})^+\\
&\quad+\sum_{n=1}^{N_{\max}}\sum_{i=1}^{q_1}\sum_{j=1}^{\tilde{q}_2}\left(\sum_{k=1}^{g_{cd}}\sum_{i_a=1}^{m_a}\sum_{i_s=1}^{m_s}\pi_{(n-1)g_{cd}+k,i,i_a,i_s,j}\right)(r+i-(n-1+j)g_{cd})^+,
\end{aligned}\tag{4.15}$$

where $N_{\max}=1+\lceil(r+q_1)/g_{cd}\rceil$, where $\lceil x\rceil$ is the smallest integer greater than or equal to x. Then $E[(q(t)+w(t)-r-IP(t))^+]$ can be obtained similarly or from $E[(r+IP(t)-q(t)-w(t))^+]$, $E[IP(t)]$, $E[q(t)]$, and $E[w(t)]$.

Proof. Eq. (4.15) can be obtained by using Table 3.1. If $q(t)=(n-1)g_{cd}+k>0$ and $\tilde{w}(t)=j$, we have $w(t)=jg_{cd}-k$ and $q(t)+w(t)=(n-1+j)g_{cd}$. Note that $0\le IP(t)\le q_1$. If $r+q_1\le q(t)$, we have $r+IP(t)-(q(t)+w(t))\le 0$. In Eq. (4.15), $q(t)=(n-1)g_{cd}+k$. Then we have $r+IP(t)-(q(t)+w(t))\le 0$, if $n>N_{\max}$. The second result is obtained by applying $x^+=x+(-x)^+$ for any real number x. This completes the proof of Proposition 4.4. □

Propositions 4.1–4.4 indicate that the computation of the performance measures, including $C(r,q_1)$, can be done in finite steps and explicitly, except for the computation of matrix G. Efficient algorithms have been developed for computing G in the literature (see [26]). We would like to remark that the computation of the stationary distribution $\boldsymbol{\pi}$ involves large size matrices. The explicit results obtained in this section are useful not only for computing performance measures, but also useful for checking computation accuracy of $\boldsymbol{\pi}$, as indicated by Remarks 4.1–4.3.

5. Computational issue and heuristic algorithm

In this section, we refine the method for computing the stationary distribution $\boldsymbol{\pi}$ and introduce a heuristic algorithm for computing the optimal (r,q_1) policy.

In Ramaswami's algorithm for computing $\boldsymbol{\pi}$, the matrix G plays a key role. The computation of the matrix G can be made more efficient. Note that the matrix A_0 given in Eq. (3.8) has a special structure. Based on the special structure of A_0, it is easy to show that G has the following structure:

$$G = \begin{pmatrix} 0 & \cdots & 0 & G_1 \\ 0 & \cdots & 0 & G_2 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & G_{g_{cd}} \end{pmatrix}, \tag{5.1}$$

where $\{G_j,\ j = 1, 2, \ldots, g_{cd}\}$ are matrices of size $q_1 m_a m_s \tilde{q}_2$. By Eq. (3.10) and routine calculations, we obtain, for $j = 1, 2, \ldots, g_{cd}$,

$$0 = \delta_{(j=1)} I \otimes (\mathbf{T}^0 \boldsymbol{\alpha}) \otimes U + \delta_{(j\geq 2)} (I \otimes (\mathbf{T}^0 \boldsymbol{\alpha}) \otimes I) G_{j-1} + ((Q_{IP,0} \oplus T) \otimes I) G_j + \left(Q_{IP,1} \otimes I \otimes I\right) G_j G_{g_{cd}}^{\tilde{q}_1}, \tag{5.2}$$

where $\delta_{(\cdot)}$ is the indicator function. Eq. (5.2) can be used for computing $\{G_j,\ j = 1, 2, \ldots, g_{cd}\}$ iteratively. Eqs. (5.1) and (5.2) indicate that, if $g_{cd} > 1$, the computation of G can be more efficient. The special structure of G also leads to a more efficient way to compute $\boldsymbol{\pi}$ in Eq. (3.11):

$$\begin{aligned}
G^n &= \begin{pmatrix} 0 & \cdots & 0 & G_1 G_{g_{cd}}^{n-1} \\ 0 & \cdots & 0 & G_2 G_{g_{cd}}^{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & G_{g_{cd}} G_{g_{cd}}^{n-1} \end{pmatrix}, \quad \text{for } n = 2, 3, \ldots; \\
B_n &= \delta_{(n=1)} A_1 + \begin{pmatrix} 0 & \cdots & 0 & (Q_{IP,1} \otimes I) G_1 G_{g_{cd}}^{\tilde{q}_1 - n} \\ 0 & \cdots & 0 & (Q_{IP,1} \otimes I) G_2 G_{g_{cd}}^{\tilde{q}_1 - n} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & (Q_{IP,1} \otimes I) G_{g_{cd}} G_{g_{cd}}^{\tilde{q}_1 - n} \end{pmatrix}, \quad \text{for } n = 1, 2, \ldots, \tilde{q}_1 + 1; \\
\hat{B}_n &= \begin{pmatrix} 0 & \cdots & 0 & (Q_{IP,1} \otimes I) G_{g_{cd}}^{\tilde{q}_1 - n} \end{pmatrix}, \quad \text{for } n = 1, 2, \ldots, \tilde{q}_1.
\end{aligned} \tag{5.3}$$

Consequently, Eq. (3.11) becomes, for $n \geq 1$,

$$\begin{aligned}
&\boldsymbol{\pi}_0 \hat{B}_n + \sum_{k=1}^{n-1} \boldsymbol{\pi}_{MG1,k} B_{n-k+1} = \mathbf{v}(g_{cd}) \otimes \boldsymbol{\eta}_n + \delta_{(n\geq \tilde{q}_1+1)} \boldsymbol{\pi}_{MG1,n-\tilde{q}_1} B_{\tilde{q}_1+1}, \\
&\boldsymbol{\eta}_n = \delta_{(n\leq \tilde{q}_1)} \boldsymbol{\pi}_0 (Q_{IP,1} \otimes \boldsymbol{\alpha} \otimes I) G_{g_{cd}}^{\tilde{q}_1 - n} + \sum_{k=2}^{\min\{n,\tilde{q}_1\}} \sum_{j=1}^{g_{cd}} \boldsymbol{\pi}_{(n-k)g_{cd}+j} (Q_{IP,1} \otimes I) G_j G_{g_{cd}}^{\tilde{q}_1 - k}; \\
&\boldsymbol{\pi}_{MG1,n} = -\boldsymbol{\eta}_n ((B_1^{-1})_{g_{cd},1}, (B_1^{-1})_{g_{cd},2}, \ldots, (B_1^{-1})_{g_{cd},g_{cd}}) - \delta_{(n\geq \tilde{q}_1+1)} \boldsymbol{\pi}_{MG1,n-\tilde{q}_1} B_{\tilde{q}_1+1} B_1^{-1}; \\
&\boldsymbol{\pi}_0 (A_{0,0} - (Q_{IP,1} \otimes \boldsymbol{\alpha} \otimes I) G_{g_{cd}}^{\tilde{q}_1 - 1} (B_1^{-1})_{g_{cd},1} (I \otimes \mathbf{T}^0 \otimes U)) = 0,
\end{aligned} \tag{5.4}$$

where $((B_1^{-1})_{g_{cd},1},\ (B_1^{-1})_{g_{cd},2}, \ldots, (B_1^{-1})_{g_{cd},g_{cd}})$ is the last (block) row of B_1^{-1}, which can be obtained from $B_1^{-1} B_1 = I$ as follows:

$$\begin{aligned}
&(B_1^{-1})_{g_{cd},j} ((Q_{IP,0} \oplus T) \otimes I) + (B_1^{-1})_{g_{cd},j+1} (I \otimes (\mathbf{T}^0 \boldsymbol{\alpha}) \otimes I) = 0, \quad \text{for } j = 1, \ldots, g_{cd} - 1; \\
&(B_1^{-1})_{g_{cd},g_{cd}} ((Q_{IP,0} \oplus T) \otimes I) + \left(\sum_{j=1}^{g_{cd}} (B_1^{-1})_{g_{cd},j} (Q_{IP,1} \otimes I) G_j \right) G_{g_{cd}}^{\tilde{q}_1 - 1} = I.
\end{aligned} \tag{5.5}$$

Other blocks of B_1^{-1} can be found in a similar way. Details are omitted.

Eqs. (5.1)–(5.5) indicate that a large part of the computation of $\boldsymbol{\pi}$ can be done with matrix blocks of size $q_1 m_a m_s q_2 / g_{cd}$ or smaller. If $g_{cd} > 1$, Eqs. (5.1)–(5.5) lead to a reduction not only in the computation time of Ramaswami's algorithm, but also in the memory space necessary for the implementation of the algorithm.

Next, we develop a heuristic algorithm for computing (r, q_1) that minimizes the expected total cost defined in Eq. (2.1). We would like to point out again that we assume that $q(0) = w(0) = 0$ so that $q(t)$ and $w(t)$ satisfy the relationship shown in Table 3.1. If $q(0) = w(0) = 0$ does not hold, the analysis can carry through and the only difference is the interpretation of the states of $\tilde{w}(t)$.

First, Eq. (2.1) can be rewritten in the following form:

$$\begin{aligned}
C(r, q_1) = {} & \frac{\lambda K_w}{q_1} + p_w (E[q(t)] - E[IP(t)] - r) + (p_w + h_s) E[w(t)] \\
& + \frac{\lambda K_s}{q_2} + (h_w + p_w) E[(r + IP(t) - q(t) - w(t))^+].
\end{aligned} \tag{5.6}$$

For any given policy (r, q_1), formulas given in Section 4 can be used for computing $C(r, q_1)$. To find the optimal policy, we first derive some properties associated with the optimal policy. By now, it is evident that the CTMC $\{(q(t), IP(t), I_a(t), I_s(t), w(t)), t \geq 0\}$ is independent of the reorder point r. Therefore, some useful properties can be obtained.

For fixed q_1, define $r^*(q_1)$ and $\underline{r}^*(q_1)$ as

$$r^*(q_1) = \arg \min_{-\infty<r<\infty} \{C(r, q_1)\};$$
$$\underline{r}^*(q_1) = \arg \min_{-\infty<r<\infty} \left\{P\{r + IP(t) \geq q(t) + w(t)\} \leq \frac{p_w}{p_w + h_w}\right\}, \tag{5.7}$$

where

$$P\{r + IP(t) \geq q(t) + w(t)\} = \sum_{i=1}^{q_1}\sum_{i_a=1}^{m_a}\sum_{j=1}^{\tilde{q}_2} \pi_{0,i,i_a,j}\delta_{(r+i\geq(j-1)g_{cd})}$$
$$+ \sum_{n=1}^{\infty}\sum_{k=1}^{g_{cd}}\sum_{i=1}^{q_1}\sum_{i_a=1}^{m_a}\sum_{i_s=1}^{m_s}\sum_{j=1}^{\tilde{q}_2} \pi_{(n-1)g_{cd}+k,i,i_a,i_s,j}\delta_{(r+i\geq(n-1+j)g_{cd})}. \tag{5.8}$$

Proposition 5.1. *Assume that $\rho < 1$. Then we have*

(1) $r^*(q_1) + q_1 \geq 0$;
(2) *For fixed q_1, $C(r, q_1)$ is convex (discrete form) in r; and*
(3) $r^*(q_1) = \underline{r}^*(q_1)$ *or* $\underline{r}^*(q_1) + 1$.

Proof. If $r + q_1 < 0$, we must have $r + IP(t) < 0$. Then Eq. (5.6) becomes

$$C(r, q_1) = \frac{\lambda K_w}{q_1} + \frac{\lambda K_s}{q_2} + p_w(E[q(t)] - E[IP(t)]) + (p_w + h_s)E[w(t)] - p_w r, \tag{5.9}$$

which is decreasing in r. Therefore, the cost function is minimized at r such that $r + q_1 \geq 0$. This proves part (1).

To prove parts (2) and (3), we define $\Delta(r) = C(r + 1, q_1) - C(r, q_1)$. By Eq. (4.15), we obtain

$$\begin{aligned}
\Delta(r, q_1) &= (h_w + p_w)\sum_{i=1}^{q_1}\sum_{i_a=1}^{m_a}\sum_{j=1}^{\tilde{q}_2} \pi_{0,i,i_a,j}\delta_{(r+i\geq(j-1)g_{cd})} \\
&\quad + (h_w + p_w)\sum_{n=1}^{\infty}\sum_{k=1}^{g_{cd}}\sum_{i=1}^{q_1}\sum_{i_a=1}^{m_a}\sum_{i_s=1}^{m_s}\sum_{j=1}^{\tilde{q}_2} \pi_{(n-1)g_{cd}+k,i,i_a,i_s,j}\delta_{(r+i\geq(n-1+j)g_{cd})} - p_w \\
&= (h_w + p_w)P\{q(t) = 0,\ r + IP(t) \geq q(t) + w(t)\} + (h_w + p_w) \\
&\quad \times \sum_{n=1}^{\infty}\sum_{k=1}^{g_{cd}} P\{q(t) = (n-1)g_{cd} + k, r + IP(t) \geq q(t) + w(t)\} - p_w \\
&= (h_w + p_w)P\{r + IP(t) \geq q(t) + w(t)\} - p_w.
\end{aligned} \tag{5.10}$$

The function $\Delta(r, q_1)$ is clearly a nondecreasing function in r, which implies that $C(r, q_1)$ is convex (discrete form) in r. Further, we have $\Delta(r, q_1) \leq 0$ if $r \leq \underline{r}^*(q_1)$ and $\Delta(r, q_1) \geq 0$ if $r \geq \underline{r}^*(q_1) + 1$. Therefore, $C(r, q_1)$ is minimized at either $\underline{r}^*(q_1)$ or $\underline{r}^*(q_1) + 1$. This completes the proof of Proposition 5.1. □

Proposition 5.1 simplifies the search for the best reorder point r, for given order size q_1, significantly. Based on Proposition 5.1 and some observations on the optimal policies from a number of numerical examples, we propose the following heuristic algorithm for finding the optimal (r, q_1) policy for fixed q_2.

An heuristic algorithm for computing the optimal (r, q_1). Set $q_1 = q_1^* = 1$ and $C_{\min} = \infty$. Choose q^U as a big positive integer. Let $C^*(q_1) = C(r^*(q_1), q_1)$.

1. For q_1, find $\boldsymbol{\pi}$ by using Eq. (3.11).
2. Use Eqs. (5.7) and (5.8) to find $\underline{r}^*(q_1)$. Then calculate $C^*(q_1)$.
3. If $C^*(q_1) < C_{\min}$, set $q_1^* = q_1$ and $C_{\min} = C^*(q_1)$. Set $q_1 =: q_1 + 1$ and go to step 1.
4. If $C^*(q_1) \leq 2C_{\min}$ or $q_1 \leq q^U$, Set $q_1 =: q_1 + 1$ and go to step 1.
5. If $C^*(q_1) > 2C_{\min}$ and $q_1 > q^U$, stop.

The solution $(r^*(q_1^*), q_1^*)$ is likely to be the optimal solution. The selection of q^U is a key issue for the algorithm to find the optimal (r, q_1) policy successfully, which can be an interesting future research topic. It is clear that the optimal policy can be found if the upper-bound q^U is sufficiently large.

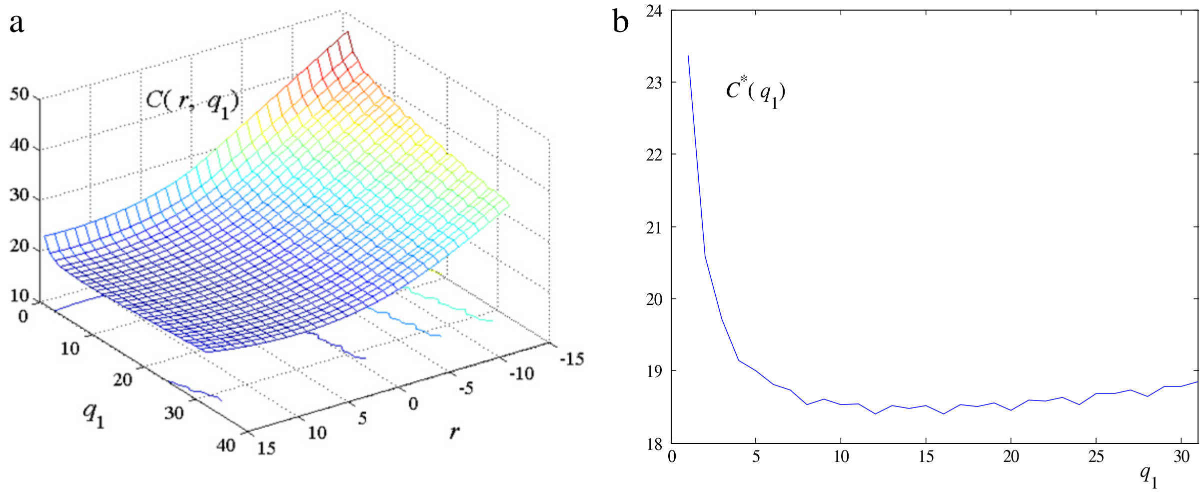

Fig. 6.1. Cost functions for Example 6.1.

6. Numerical examples and extensions

In this section, we use four cases to discuss issues related to the optimal (r, Q) policy (Examples 6.1 and 6.2), and model extension (Examples 6.3 and 6.4).

Example 6.1. Consider an inventory–production system with the following parameters:

$$MAP: \ m_a = 2, \qquad D_0 = \begin{pmatrix} -0.7 & 0.2 \\ 0 & -2 \end{pmatrix}, \qquad D_1 = \begin{pmatrix} 0.5 & 0 \\ 0.3 & 1.7 \end{pmatrix};$$

$$PH\text{-}distribution: \ m_s = 2, \qquad \boldsymbol{\alpha} = (0.9, 0.1), \qquad T = \begin{pmatrix} -8 & 1 \\ 0.4 & -0.4 \end{pmatrix};$$

$$\text{Costs}: \ h_w = 1, \qquad p_w = 1.2, \qquad h_s = 1.5, \qquad K_w = 5, \quad \text{and} \quad q_2 = 4.$$

The demand arrival rate is $\lambda = 1.1$ and the service rate is $\mu = 1.33333$. The coefficient of variation of the production time is 2.3938, which indicates that the production time is quite variable. Since the cost K_s does not affect the selection of (r, q_1), we set $K_s = 0$.

The cost functions $C(r, q_1)$ and $C^*(q_1)$ are plotted in Fig. 6.1(a) and (b), respectively. The optimal (r, Q) policy for the warehouse is $(r^*, q_1^*) = (9, 16)$ with an expected total cost per unit time $C^* = 18.4013$ per unit time. As shown in Fig. 6.1(a), the cost function $C(r, q_1)$ is not convex in (r, q_1). The function $C^*(q_1)$ is not convex in q_1. This makes it more challenging to develop an algorithm for finding the optimal (r, Q) policy. In addition, we find

(i) $r^*(q_1) = 13, 12, 12, 11, 11, 11, 11, 10, 10, 10, 10, 9, 9, 9, 9, 9, 8, 8, 8, 8, 8, 7, 7, 7, 7, 7, 7, 7, 6, 6, 6$, for $q_1 = 1, 2, \ldots, 31$.
(ii) $r^*(q_1) + q_1 = 14, 14, 15, 15, 16, 17, 18, 18, 19, 20, 21, 21, 22, 23, 24, 25, 25, 26, 27, 28, 29, 29, 30, 31, 32, 33, 34, 35, 35, 36, 37$, for $q_1 = 1, 2, \ldots, 31$.

The above results indicate that the reorder point $r^*(q_1)$ seems nonincreasing in q_1, and the order-up-to level $r^*(q_1) + q_1$ seems nondecreasing in q_1, which are consistent with intuition.

Example 6.2. Consider a model with Poisson demands with $\lambda = 1.1$ and exponential production times with $\mu = 1.33333$. All other parameters are the same as that of Example 6.1. Note that the demand arrival rate and the production rate are the same as that of Example 6.1 as well. Thus, the main difference between the models considered in Examples 6.1 and 6.2 are (i) the demand process in Example 6.1 is more bursty, and (ii) the production time in Example 6.1 is more variable (note that the coefficient of variation of an exponential random variable is 1).

For Example 6.2, the optimal (r, Q) policy is $(r^*, q_1^*) = (2, 12)$, which is quite different from (7, 16) for Example 6.1. The minimum expected total cost per unit time is 7.2237 for Example 6.2, which is also drastically different from 17.0835 for Example 6.1. The two models in Examples 6.1 and 6.2 have the same demand rates and the same production rates, but the performances of the two systems are significantly different.

Examples 6.1 and 6.2 show that the burstiness of the demand process and the variability of the production times have significant impact on the optimal inventory control in the warehouse. Thus, they should be considered in the design of such inventory–production systems. Examples 6.1 and 6.2 also indicate that the minimum expected total cost per unit time depends on not only the (average) demand rate and mean production time, but also the types of demand processes and production times. Thus, the utilization of *MAP*s for the demand process and *PH*-distributions for the production time becomes necessary for more accurate estimates of performance measures, in addition to inventory control.

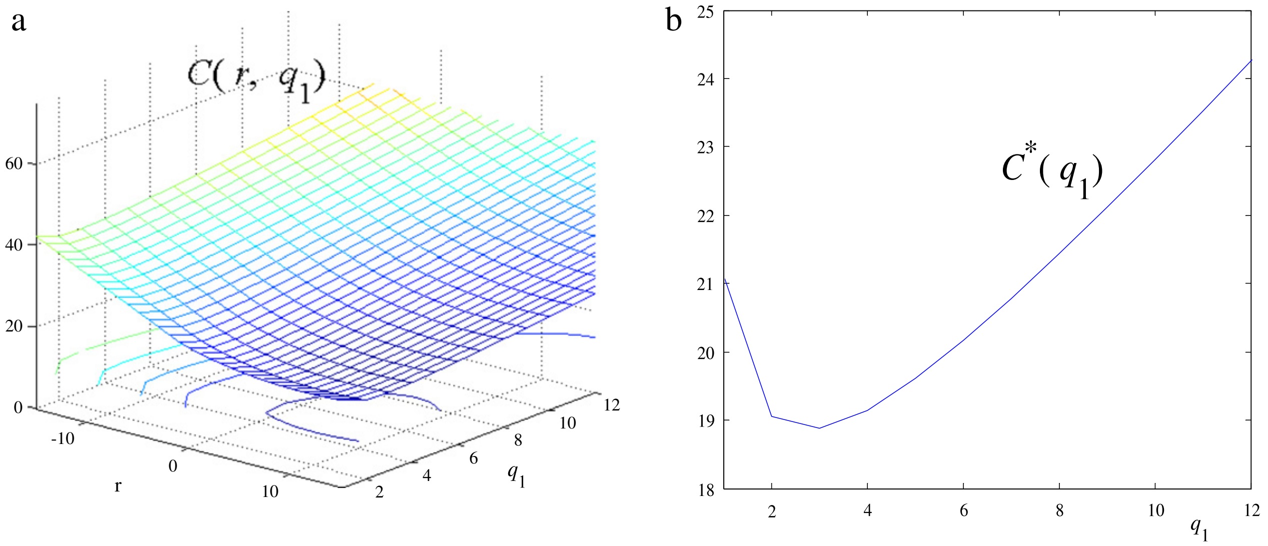

Fig. 6.2. Costs functions for Example 6.3.

Example 6.3. In this example, we consider the system with $q_2 = q_1$. In practice, the assumption implies that the finished products in an order must be transported together from the production facility to the warehouse. The immediate consequences of $q_2 = q_1$ are: (i) $g_{cd} = q_1$; and (ii) $q(t) + w(t) = \tilde{q}(t)q_1$, where $\tilde{q}(t)$ is the queue length of the $MAP(q_1)/PH(q_1)/1$ queue. The queue $MAP(q_1)/PH(q_1)/1$ can be defined from the $MAP/PH/1$ queue by grouping consecutive q_1 customers to form a super customer, where $\tilde{q}(t)$ counts the number of super customers in the queueing system. Then Eq. (2.1) becomes

$$C(r, q_1, q_1) = \frac{\lambda(K_w + K_s)}{q_1} + h_s\rho\frac{q_1 - 1}{2} + h_w E[(r + IP(t) - q_1\tilde{q}(t))^+] + p_w E[(q_1\tilde{q}(t) - r - IP(t))^+]. \quad (6.1)$$

The Markov chain to be analyzed is $\{(\tilde{q}(t),\ IP(t),\ I_a(t),\ I_s(t)), t \geq 0\}$. Although q_2 increases with q_1 in this case, the matrices involved in Ramaswami's algorithm (see Eq. (3.11)) are of size $q_1 m_a m_s$. Thus, numerical analysis of this model can be done efficiently.

Use all the parameters given in Example 6.1, except that of q_2. We obtain cost functions $C(r, q_1)$ and $C^*(q_1)$, which are plotted in Fig. 6.2(a) and (b), respectively. The optimal (r, Q) policy is $(r^*, q_1^*) = (11, 3)$ with $C^* = 18.8711$ per unit time. Note that the optimal solution $(11, 3)$ is quite different from the optimal solution $(9, 16)$ where q_2 is fixed at 4. On the other hand, the corresponding minimum costs are similar: 18.8711 and 18.4013.

An interesting observation is that the cost functions $C^*(r, q_1)$ and $C^*(q_1)$ seem to be convex, if $q_1 = q_2$. If it is true, the search for the optimal policy becomes feasible and can be efficient.

Example 6.4. In practice, demands may arrive in batches. In this example, we construct *MAP*s that can approximate batch arrival processes. The idea is to construct *MAP*s such that the arrival of one demand can be followed by several demands in a very short period of time. In general, a batch arrival process can be modeled by using *BMAP* [23], which has a matrix representation $(C_0, C_1, \ldots, C_K)$, where C_k is for the (matrix) arrival rate of batches of size k. We define

$$D_0 = \begin{pmatrix} C_0 & & & \\ & -\xi I & & \\ & & \ddots & \\ & & & -\xi I \end{pmatrix}, \quad D_1 = \begin{pmatrix} C_1 & C_2 & \cdots & C_K \\ \xi I & 0 & & \\ & \ddots & \ddots & \\ & & \xi I & 0 \end{pmatrix}. \quad (6.2)$$

If ξ is sufficiently large, then (D_0, D_1) is an *MAP* that approximates $(C_0, C_1, \ldots, C_K)$.

Remark 6.1. We do not consider *BMAP* directly in this paper due to the difficulty to construct an irreducible version Markov chain for $\{(q(t), IP(t), I_a(t), I_s(t), w(t)), t \geq 0\}$, if a *BMAP* is utilized.

7. Conclusions and discussion

This paper develops an efficient algorithm for the performance analysis of the inventory–production system of interest. Numerical results demonstrate the usefulness of *MAP*s and *PH*-distributions in getting accurate estimates of performance measures and in improving inventory management. Some issues, such as algorithms for computing the optimal (r, Q) policy and the analysis of an inventory–production system with a positive transportation time from the production facility to the warehouse, are worth further investigation.

References

[1] S. Axsäter, Inventory Control, Kluwer Academic Publishers, London, 2000.
[2] J.A. Buzacott, J.G. Shanthikumar, Stochastic Models of Manufacturing Systems, Prentice Hall, New York, 1993.
[3] G. Hadley, T. Whitin, Analysis of Inventory Systems, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.
[4] E. Porteus, Foundations of Stochastic Inventory Theory, Stanford University Press, Stanford, California, 2002.
[5] P.H. Zipkin, Foundations of Inventory Management, McGraw Hill, Boston, 2000.
[6] F. Chen, J.S. Song, Optimal policies for multi-echelon inventory problems with Markov modulated demand, Operations Research 49 (2001) 226–234.
[7] F. Chen, Y.S. Zheng, Evaluating echelon stock (R, nQ) policies in serial production/inventory systems with stochastic demand, Management Science 40 (1994) 1262–1275.
[8] D. Iglehart, Optimality of (s, S) policies in the infinite horizon dynamic inventory problem, Management Science 9 (1963) 259–267.
[9] H. Scarf, The optimality of (s, S) policies in dynamic inventory problems, in: K. Arrow, L. Karlin, P. Suppes (Eds.), Mathematical Methods in the Social Sciences, Stanford University Press, Stanford, California, 1960.
[10] S.P. Sethi, F. Cheng, Optimality of (s, S) policies in inventory models with Markovian demand, Operations Research 45 (1997) 931–939.
[11] J. Song, Inventory management in a fluctuating environment, Ph.D. Dissertation, Graduate School of Business, Columbia University, New York, 1991.
[12] Y.S. Zheng, A simple proof for optimality of (s, S) policies in infinite-horizon inventory systems, Journal of Applied Probability 28 (1991) 802–810.
[13] J.S. Song, P. Zipkin, Inventory control in a fluctuating demand environment, Operations Research 41 (1993) 351–370.
[14] O. Berman, E.H. Kaplan, D.G. Shimshak, Deterministic approximations for inventory management at service facilities, IIE Transactions 25 (1993) 98–104.
[15] S.K. Goyal, S.G. Deshmukh, Integrated procurement-production systems: a review, European Journal of Operational Research 62 (1992) 1–10.
[16] J.S. Song, S. Xu, B. Liu, Order-fulfillment performance measures in an assemble-to-order system with stochastic lead times, Operations Research 47 (1999) 131–149.
[17] Q.M. He, E.M. Jewkes, J. Buzacott, An efficient algorithm for computing the optimal replenishment policy for an inventory–production system, in: Advances in Matrix-Analytic Methods for Stochastic Models (Proceedings of the Second International Conference on Matrix-Analytic Methods, Winnipeg, Canada, 1998), Notable Publications, Inc., New Jersey, 1998, pp. 381–402.
[18] Q.M. He, E.M. Jewkes, J. Buzacott, Optimal and near-optimal inventory control policies for a make-to-order inventory–production system, European Journal of Operational Research 141 (2002) 113–132.
[19] Q.M. He, E.M. Jewkes, J. Buzacott, The value of information used in inventory control of a make-to-order inventory–production system, IIE Transactions 34 (2002) 999–1013.
[20] F. De Vericourt, F. Karaesmen, Y. Dallery, Optimal stock allocation for a capacitated supply chain, Management Science 48 (2002) 1486–1501.
[21] A.Y. Ha, Stock rationing in an $M/E_k/1$ make-to-stock queue, Management Science 46 (2000) 77–87.
[22] S. Asmussen, G. Koole, Marked point processes as limits of Markovian arrival streams, Journal of Applied Probability 30 (1993) 365–372.
[23] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Stochastic Models 7 (1991) 1–46.
[24] M.F. Neuts, A versatile Markovian point process, Journal of Applied Probability 16 (1979) 764–779.
[25] M.F. Neuts, Matrix-Geometric Solutions in Stochastic Models—An Algorithmic Approach, The Johns Hopkins University Press, Baltimore, 1981.
[26] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modelling, ASA & SIAM, Philadelphia, USA, 1999.
[27] M.F. Neuts, Structured Stochastic Matrices of M/G/1 Type and their Applications, Marcel Dekker, New York, 1989.
[28] R.N. Boute, S.M. Disney, M.R. Lambrecht, B. van Houdt, An integrated production and inventory model to dampen upstream demand variability in the supply chain, European Journal of Operational Research 178 (2007) 121–142.
[29] R.N. Boute, M.R. Lambrecht, B. van Houdt, Performance evaluation of a production/inventory system with periodic review and endogenous lead times, Naval Research Logistics 54 (2007) 462–473.
[30] V. Ramaswami, Algorithms for a continuous-review (s, S) inventory system, Journal of Applied Probability 18 (1981) 461–472.
[31] V. Ramaswami, Stable recursion for the steady state vector in Markov chains of $M/G/1$ type, Stochastic Models 4 (1988) 183–188.

Qi-Ming He is currently a Professor in the Department of Management Sciences at the University of Waterloo. 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. In investigating various stochastic models, his favorite methods are matrix analytic methods. Recently, he has been working on queueing systems with multiple types of customers, inventory systems with multiple types of demands, and representations of phase-type distributions and their applications.

Hanqin Zhang is a Professor in the Department of Decision Sciences, Business School, National University of Singapore, Singapore. He received his Ph.D. in Operations Research from Institute of Applied Mathematics, the Chinese Academy of Sciences, China, in 1991. His research interests are in queueing networks, stochastic manufacturing systems, inventory models, and supply chain management. He has published more than 60 papers in refereed journals such as *Advances in Applied Probability, INFORMS Journal on Computing, Journal of Applied Probability, Operations Research, Queueing Systems, and Stochastic Models*. He is a co-author of two monographs, *Average-Cost Control of Stochastic Manufacturing Systems* (with S. Sethi and Q. Zhang, Springer-Verlag, 2004), and *Inventory and Supply Chain Management with Forecast Updates* (with S. Sethi and H. Yan, Springer-Verlag, 2005).