Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 176 (2007) 275–294
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Production, Manufacturing and Logistics

Modeling and analysis of a supply–assembly–store chain

Zhao Xiaobo [a,*], Deju Xu [b], Hanqin Zhang [c], Qi-Ming He [d]

[a] *Department of Industrial Engineering, Tsinghua University, Beijing 100084, China*
[b] *Department of Mathematics, Capital Normal University, Beijing 100037, China*
[c] *Institute of Applied Mathematics, Academy of Mathematics and Systems Science, The Chinese Academy of Sciences, Beijing 100080, China*
[d] *Department of Industrial Engineering, Dalhousie University, Halifax, Canada B3J 2X4*

Received 24 March 2004; accepted 27 June 2005
Available online 11 October 2005

Abstract

We consider a supply–assembly–store chain with produce-to-stock strategy, which comprises a set of component suppliers, a mixed-model assembly line with a constantly moving conveyor linking a set of workstations in series, and a set of product storehouses. Each supplier provides components of a specified family, which are assembled at a corresponding workstation. Units belonging to different models of products are sequentially fed onto the conveyor, and pass through the workstations to generate finished products. Each storehouse stores finished products belonging to a specific model for satisfying customer demands. The suppliers deliver components according to a just-in-time supply policy with stochastic leadtimes. Customer demands for a particular model of products arrive at the corresponding storehouse according to a Poisson stream. The paper conducts a modeling and performance analysis in the design stage of the system in the sense of "long-term-behavior". A rolling technique is constructed for analyzing stationary probability distributions of the numbers of components. A two-dimensional Markov chain with infinite states is introduced for analyzing stationary probability distributions of inventories of finished products. Based on these distributions, performance measures of the system, such as work-in-process of components, inventory amounts of finished products, as well as service levels for customers, can be easily obtained. Managerial insights are obtained from both analytical and numerical results.

Keywords: Supply chain; Mixed model assembly line; Markov chain; Work-in-process; Inventory; Service level

* Corresponding author. Tel.: +86 10 6278 4897/4898; fax: +86 10 6279 4399.
E-mail address: xbzhao@mail.tsinghua.edu.cn (Z. Xiaobo).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2005.06.053

1. Introduction

Supply chain management addresses the management of materials and information across the entire chain from suppliers to producers, distributors, retailers, and customers. Traditionally, each company performs purchasing, production and marketing activities independently, so that it is difficult to make an optimal plan for the whole chain. In recent years, it has been realized that actions taken by one member of the chain can influence all others in the chain (see, for example, Chopra and Meindl, 2001; Silver et al., 1998). More and more companies have gradually recognized that each of them serves as part of a supply chain against other supply chains in terms of competition, rather than as a single firm against other individual firms. Since 1990, as the information technology has continuously developed, it is possible to coordinate all organizations and all functions involved in the whole chain. Consequently, supply chain management has been increasingly receiving attention from both academic researchers and practitioners. Roughly speaking, research on supply chain management has been mainly focused on three major issues. One is the behavior of information flow through a supply chain (see, for example, Lee et al., 1997). The second issue deals with inventory management, which regards a supply chain as a multi-echelon inventory system (see, for example, Axsater, 2000a; Zipkin, 2000, and the cited references in them). The third issue is orientated to planning and operations management of a supply chain based on queueing systems, which has not been addressed enough in the literature yet (see the most related works, for example, Parlar, 2000; Raghavan and Viswanadham, 2001; Song and Yao, 2002).

Since a supply chain deals with material flows and information flows across the entire chain, from suppliers of original components to final customers, it comprises at least two major domains: the physical transformation domain (mining, smelting, casting, alloying, machining, assembling; etc.), and the goods distribution domain (conveyance, storage, and transportation).

The physical transformation domain is formed by several manufacturing enterprises that generate goods through a series of processes provided by different firms. In recent years, in this domain, the "just-in-time" (JIT) principle has been adopted as a supply mechanism in many firms in actual supply chains (see, for example, Aigbedo, 2004; Grout and Christy, 1999; Kelle and Miller, 2001; Olhager, 2002; Pan and Yang, 2002). Originally, the JIT philosophy was developed by the Toyota Motor Corporation through the kanban control for the objective of minimizing inventories. Since the mid-1980's it has become one of the principal methods used as an *internal* production management system within a single manufacturer. Therefore a great deal of the research in JIT production systems treats leadtimes for internal (to the firm) supplies as controllable; usually they are assumed to be constant or even zero (see, for example, Miyazaki et al., 1988; Muckstadt and Tayur, 1995; Sarker and Balan, 1999; Spearman and Zazanis, 1992). If we extend the JIT principle as an intra-firm supply mechanism in a supply chain, the leadtimes become a major factor of concern, and the assumption of constant or zero supply leadtimes becomes no longer tenable. Therefore we contend that leadtimes should be treated as random variables. So far, to the authors' knowledge, few papers have treated stochastic leadtimes for JIT supply mechanisms; see, for example, Yanagawa et al. (1994) who assume a discrete probability distribution, and Grout and Christy (1999) who consider a uniform distribution.

In the goods distribution domain, goods are moved from warehouses to distributors, from distributors to retailers, and finally from retailers to customers. The focus of research in this area is inventory management, which is somewhat different from production management. For example, in the physical transformation domain the JIT supply mechanism among different companies can be implemented by exterior supply kanbans (Monden, 1998). It can be difficult to introduce such a supply mechanism in the distribution area. So far, the popular inventory management approaches in this area remain the (s, S) policy under periodic review (see, form example, Cetinkaya and Parlar, 2002; Sobel and Zhang, 2001), the (R, Q) policy under continuous review (see, for example, Axsater, 2000b), etc.

Mixed model assembly lines (MMALs) are widely developed in modern industries. In the last few decades MMALs have received considerable attention in the literature; see, for example, the survey articles of

Gagnon and Ghosh (1991), Lima-Fernandes and Groover (1995) and Yano and Bolat (1989). In a supply chain that includes an MMAL, the MMAL plays a central role as a bridge that connects the transformation domain and the distribution domain. Components supplies are pulled from the transformation domain to the MMAL, and the finished products are replenished from the MMAL to the distribution domain. Therefore, the planning and operations management for the MMAL directly impact the performance of the entire chain. For such MMAL-centered supply chains, goods replenishment processes from the MMAL to the distribution area are not controlled by the conventional policies such as the (s, S) policy or the (R, Q) policy. They are also different from conventional ‘‘one producer–multiple retailers with capacitated resource’’ systems in the literature (Evans, 1967; Perez and Zipkin, 1997; Rajagopalan, 2002; Tayur, 1993). We shall provide modeling and analysis of such supply chains in the subsequent sections.

On the other hand, for production strategies, basically speaking, there are two popular types: ‘‘produce-to-stock’’ and ‘‘produce-to-order’’ (see Buzacott and Shanthikumar, 1993). Production planning with produce-to-stock strategy is based on the market forecasting information, which is widely used in appliance industry, IT industry, common-type autos industry, etc. Consequently, for these production systems, lost sales may occur naturally. Production strategy of produce-to-order is adopted by industries such as aircraft manufacturing, ship manufacturing, special-type autos manufacturing, etc. The production planning for such industries is made only after receiving customers' demands. Hence, a lost sale may not take place in such systems.

In this paper, we consider a supply–assembly–store chain with produce-to-stock strategy. The chain comprises a set of component suppliers, a mixed model assembly line (MMAL) with a conveyor linking a set of workstations in series, and a set of finished product storehouses. Each supplier provides components of a specified component family to a specified workstation in the MMAL. All suppliers deliver components according to the JIT mechanism with respective stochastic leadtimes. Each workstation assembles components of some specified families. Units belonging to different models of products are sequentially fed onto the conveyor of the MMAL, and are moved by the conveyor at a constant speed to pass through the workstations to generate products. Finished products enter the storehouses, each of which stores finished products of a specified model. Customers arrive at the storehouses to get their desired products. This system is close to some real supply chains in such as electronic appliance industries and auto manufacturing (see, for example, Monden, 1998). Thus, its study is of considerable value to practitioners in these industries.

We model the above system and analyze stationary probability distributions of the numbers of components and stationary probability distributions of finished products in storehouses, in the sense of long-term-behavior. These analyses are useful and valuable to obtain performance measures of the system, such as averages and variances of work-in-processes (WIPs) of components, averages and variances of inventories of finished products in storehouses, as well as service levels for customers. They are also useful and valuable in risk and sensitivity analysis for the system. Such studies are usually essential at the design stage of the supply chain to ensure satisfactory performance. For example, designers are concerned with the system performance with respect to the system parameters, the system configuration, etc., to make an appropriate decision. Similarly, they also expect to obtain performance measures to evaluate a given design plan. The methods developed in this paper can be used to compare different system designs based on their stationary probability distributions, i.e., their ‘‘long-term-behavior’’. Therefore, the main contribution of this paper is threefold: formulation of stochastic models for the number of components and the inventory of finished products, performance analysis conducted by a ‘‘rolling method’’ and by using matrix analytic methods, and the provision of some interesting insights into the behavior of such systems.

The rest of the paper is organized as follows. In Section 2, the model formulation of the supply chain is given. In Section 3, a rolling technique is proposed to analyze stationary probability distributions of the numbers of components, and managerial insights are obtained from analytical results. In Section 4, the stationary probability distributions of inventories in the storehouses are discussed based on the matrix analytic method. Section 5 provides numerical results to investigate the impact on performance with respect to several managerial measures. Future work is pointed out in Section 6 to conclude the paper.

2. Model formulation

The supply–assembly–store chain considered in this paper comprises three parts: a set of N component suppliers, a mixed model assembly line (MMAL), and a set of M product storehouses. Components come from the N suppliers and are held in N component shelves near the MMAL before used in assembly. Assemblies occur in K workstations (work zones) along the MMAL, where components are taken from shelves and put on units to form M models of products. Finished products off the MMAL are stored in the M product storehouses waiting to be sold to customers. In the rest of this section, we give a detailed description of the three parts. (See an example in Fig. 1 with $M = 3$ product models, $N = 5$ component suppliers, and $K = 3$ workstations.)

2.1. Storehouses of finished products

Products off the MMAL are stored in the M storehouses that are numbered from 1 to M with model m products ($1 \leqslant m \leqslant M$) stored in storehouse m. The demands from customers for model m products follow a stationary Poisson process with parameter λ_m with each demand requiring a single product. Since the system adopts the produce-to-stock strategy, it is reasonable, as described in the previous section, that during any period of no products in storehouse m, customers for model m products are lost.

2.2. Deliveries of components from suppliers

Components used in assembling on the MMAL are supplied by N suppliers that are numbered 1 to N. Each supplier is responsible for the delivery of a family of (possibly different) components. For instance, in a car assembly, all kinds of engines are from the same supplier but with different horsepower, which are used in different model cars. We assume that supplier n is responsible for the delivery of the components of family n, $1 \leqslant n \leqslant N$. There is a shelf near the MMAL for each family to hold the received components.

When an order for components of family n is issued, it takes a random time for supplier n to deliver the ordered components. Such leadtimes of delivering components by supplier n are independently identically distributed random variables (i.i.d.r.v.s) with cumulative distribution function $F_n(t)$. Let ξ_n be the generic

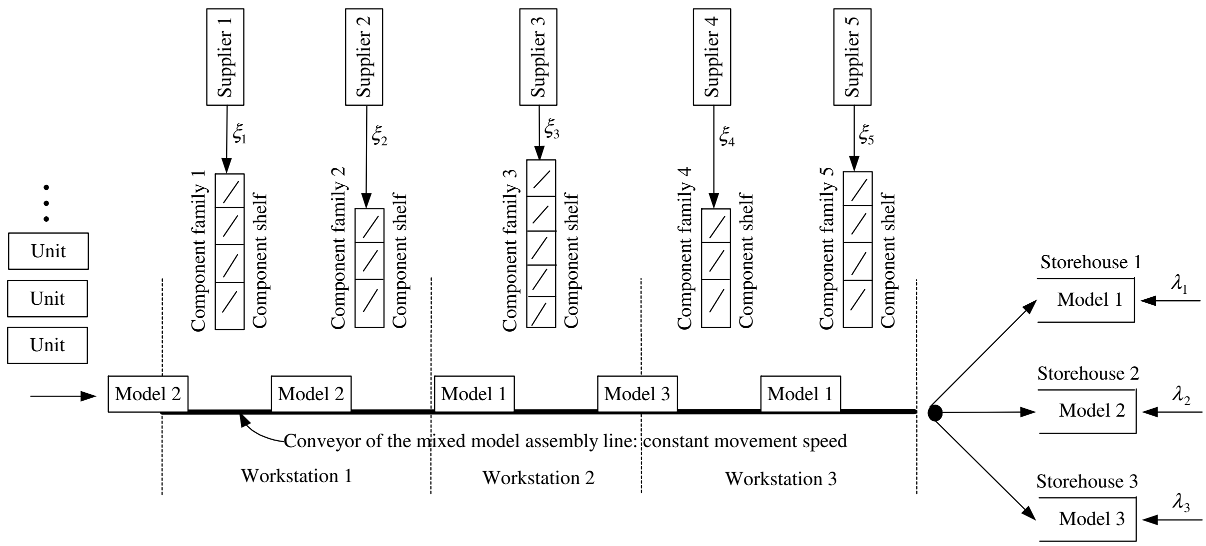

Fig. 1. A supply–production–store chain.

random variable corresponding to $F_n(t)$. We assume that ξ_n takes nonnegative real values and is bounded, i.e., there is a finite positive constant T_n such that

$$F_n(T_n) = \Pr\{\xi_n \leqslant T_n\} = 1. \tag{2.1}$$

For later use and without loss of generality, we assume that T_n is the smallest integer for which Eq. (2.1) holds. The above constraint on ξ_n is not restrictive since any real delivery can be completed within a finite time interval.

An actual MMAL is usually operated under the condition that a shortage of any component never occurs. To satisfy that condition and to keep the inventory of components to a minimum, the following supply policy is adopted. For components of family n, an order is placed at time t for components to be used at time $t + T_n$. This supply policy is called *just-in-time* (JIT) delivery with stochastic leadtimes. It is easy to see that, under the JIT supply policy, no shortage of components of family n will occur since the leadtime ξ_n is less than or equal to T_n with probability one. It will be made clear that the production schedule at the MMAL is deterministic. Thus, future requirements of any family of components can be entirely known. Therefore, the JIT supply policy can be applied. We point out that the replenishment processes of different families of components depend only on the production schedule (to be specified next) of the MMAL and are independent of each other.

2.3. The MMAL

The MMAL consists of a constantly moving conveyor and K workstations each of which represents a work zone. There are M different models of products produced on the MMAL. A model m ($1 \leqslant m \leqslant M$) product begins with an initial unit that is fed onto the front of the conveyor and moved by the conveyor from workstations 1 to $2, 2$ to $3, \ldots,$ and $K-1$ to K. Initial units fed onto the conveyor can be either common or specific to different models of products produced by the MMAL. For convenience, we call a unit for model m product a model m unit. When a model m unit passes through a workstation, components required for the model are assembled onto the unit. After passing all workstations (at the end of the conveyor), a model m product is finished and is stored in storehouse m. The numbers of components from different component families required for different models of products are given in the following $M \times N$ matrix:

$$A = \begin{pmatrix} A_{1,1} & \cdots & A_{1,N} \\ \vdots & \vdots & \vdots \\ A_{M,1} & \cdots & A_{M,N} \end{pmatrix}, \tag{2.2}$$

where the (m,n)th element $A_{m,n}$ is the number of components of family n required by a unit of model m product. The fact that a component family comprises a set of similar but different components and different models of products may require different numbers of components makes the MMAL different from single model assembly lines (SMALs).

Without loss of generality, we assume that workstation k assembles components from suppliers $\{n_{k-1}+1, n_{k-1}+2, \ldots, n_k\}$, where $n_0 = 0 < n_1 < n_2 < \cdots < n_K = N$. (Otherwise, we can simply re-number the component suppliers so that the assumption holds). A workstation has a group of operators to complete tasks of assembling the specified components. As soon as a unit is brought into workstation k by the conveyor, all its required components from suppliers $\{n_{k-1}+1, n_{k-1}+2, \ldots, n_k\}$ are simultaneously taken from their shelves. As the unit on the conveyor moves through the work zone, the operators of the workstation assemble these components onto the unit.

Fig. 1 illustrates a system with five suppliers, an MMAL with three workstations, and three storehouses. Workstation 1 is used to assemble components of families 1 and 2, workstation 2 is to assemble components of family 3, and workstation 3 is to assemble components of families 4 and 5. For this case, $N = 5$, $K = 3$,

$M = 3$, and $n_0 = 0$, $n_1 = 2$, $n_2 = 3$, and $n_3 = 5$, i.e., the first workstation is for families $\{1, 2\}$, the second is for family $\{3\}$ and the third is for families $\{4, 5\}$.

We note that the assembling process of the MMAL is affected by the physical characteristics of the MMAL, such as the speed of the conveyor, the lengths of the workstations, and the time duration between feeding units onto the conveyor. For our model, it is not needed to consider the lengths of the workstations nor the speed of the conveyor as long as the operators can complete their assembling tasks during the sojourn time of a unit in the workstations. The MMAL operates in such a manner that: (1) one unit is fed onto the conveyor every T_c units of time; and (2) the conveyor is moving at a constant speed. Without loss of generality, we assume that one unit is fed onto the conveyor per unit time (i.e., $T_c = 1$) so that the conveyor moves one unit to each workstation per unit time. A moment's reflection indicates that regardless of the speed of the conveyor and the lengths of the workstations, if one unit is fed onto the conveyor per unit time, then one unit moves into each workstation per unit time and one finished product goes off the MMAL per unit time.

A production schedule is determined by the feeding order of units for producing different models of products, i.e., the so-called *sequence* (see, for example, Scholl, 1999). For actual MMALs in the operational stage, such a sequence represents the production schedule of one working day or one working shift. As was indicated, we focus on the design stage of the supply chain for modeling and analysis based on the "long-term-behavior", so in our model the production sequence is the same for each day (see Remark 2.1 for more discussions on this assumption). Let $\Sigma = \{\sigma(0), \sigma(1), \ldots, \sigma(D-1)\}$ be the sequence, where $\sigma(j)$ identifies the model of the unit fed onto the conveyor at time j and D is the total number of products produced per day, $1 \leqslant \sigma(j) \leqslant M$ and $0 \leqslant j \leqslant D-1$. By stationarity, it follows that $\sigma(j)$ identifies the model of the unit fed onto the conveyor at time $j + iD$ in the ith day, for $0 \leqslant j \leqslant D-1$, and i is any nonnegative integer. Therefore, production is scheduled in cycles with a cycle length D and Σ represents a production schedule for time periods $[0, D-1]$, $[D, 2D-1]$, and so on.

Let d_m represent the number of model m products in the sequence Σ, which is also the number of model m products produced per day.

Remark 2.1. In the operation of an MMAL, an important element is the daily production schedule, i.e., the sequence. It needs to be dynamically adjusted if demands by customers are nonstationary or the manufacturer takes a produce-to-order strategy for which a schedule is determined according to the received order bills. This paper discusses long-term-behavior in design stage under conditions such as stationary customers' demands and produce-to-stock strategy. In doing so, a fixed sequence Σ is used for daily production for the analysis of the system in the sense of long-term-behavior.

Remark 2.2. In this paper, we assume that all leadtimes incurred by the same supplier are independent and we do not make any further assumption on the delivery discipline such as "first-supply-first-arrive".

Remark 2.3. The JIT supply policy and T_c imply that the duration of the interval between two orders to a supplier is just T_c. Nevertheless, if T_c is small, then the supply frequency may be too high. In such cases, we can consider to take the supply duration as multiples of T_c, which means that the amount of components in a single delivery will be used by multiple units. Then the analysis principle for the number of components in the subsequent section is the same for such cases.

Remark 2.4. The existing literature mainly focuses on MMALs under one of three objectives: (1) leveling the usage rate of components consumption without supply leadtimes; (2) balancing the assembling load in the workstations; and (3) smoothing the product output over all models under produce-to-order strategy. In many actual systems, supply leadtimes of the components are essential, and inventories of finished products as well as service levels for customers are crucial under produce-to-stock strategy. Our model is more appropriate for such systems from practical perspective.

3. Analysis of components on shelves

The number of components of family n on the corresponding shelf (we call it the work-in-process (WIP) thereafter), at integer time epoch s is denoted by $W_n(s)$, $1 \leqslant n \leqslant N$. In this section, we develop a method to analyze the probability distribution of $W_n(s)$. Since no shortage is allowed and no coordination among suppliers exists, the replenishment processes of the n families of components are independent. Thus, our analysis can be focused on a single family of components. We shall first introduce a "rolling" technique for tracking the status of the placed orders, and then find its steady state distribution. After that, we obtain the probability moments of the WIP, and finally present managerial insights gained from the analytical results.

3.1. Rolling technique

Consider the components of family n. Suppose that the corresponding workstation of component family n is k, i.e., components of family n are used in workstation k. Without loss of generality, we assume that the first unit enters workstation k at time 0. Then subsequent units enter the workstation only at positive integer epochs since a unit is fed onto the conveyor each unit of time. In order to study $W_n(s)$, we first analyze the status of the placed orders at integer epochs (i.e., unit-entry epochs).

The method is based on a "rolling" technique that is characterized by two variables: a pointer and a string.

The pointer, $X(s)$, tracks the product model of the unit just entered workstation k at time s in the following manner. Let $X(s) = \mathrm{Rem}(s, D)$ be the remainder of the nonnegative integer s divided by D. (Note that, throughout this paper, we use t for continuous time and s for integer time epochs.) The variable $X(s)$ is actually deterministic in a cyclic fashion and takes nonnegative integer values $\{0, 1, \dots, D-1\}$. By the definition of the production sequence $\Sigma = \{\sigma(0), \sigma(1), \dots, \sigma(D-1)\}$, $\sigma(X(s))$ represents the product model of the unit just entered workstation k at time s.

Next, the string, a vector of random variables $\mathbf{J}(s) = (j_1(s), j_2(s), \dots, j_{T_n-1}(s))$, is introduced to track the status of the placed orders. Since the leadtimes of supplies are less than or equal to T_n, there can be at most $T_n - 1$ outstanding orders whose status is still uncertain at any time epoch s. In fact, these are the most recent $T_n - 1$ orders. Let, for $1 \leqslant i \leqslant T_n - 1$,

$$j_i(s) = \begin{cases} 1, & \text{if the components of family } n \text{ ordered at } s - T_n + i \text{ have arrived,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

The random variable $j_i(s)$ provides information about the order placed at time $s - T_n + i$; the components placed at that time will be used at time $s + i$. In other words, the components associated with the ith element in vector $\mathbf{J}(s)$ will be used by the unit $\sigma(\mathrm{Rem}(X(s) + i, D))$ in the sequence Σ.

It is clear that the vector $(X(s), \mathbf{J}(s))$ provides all the information needed about the replenishment process of components of family n.

Fig. 2 shows an example for which $D = 5$ and $T_n = 7$. Suppose at time epoch 17, $(X(17), \mathbf{J}(17)) = (2, (1, 0, 1, 0, 1, 1))$. Then the pointer is 2, two orders placed at times 12 and 14 have not arrived yet and other orders placed at times 11, 13, 15, and 16 have arrived.

It is obvious that the pointer is deterministic which takes values cyclically and sequentially from 0 to $D - 1$ whereas the string is stochastic characterized by the vector $\mathbf{J}(s) = (j_1(s), j_2(s), \dots, j_{T_n-1}(s))$. The state space of $X(s)$ is

$$\Omega_p = \{0, 1, \dots, D-1\}. \tag{3.2}$$

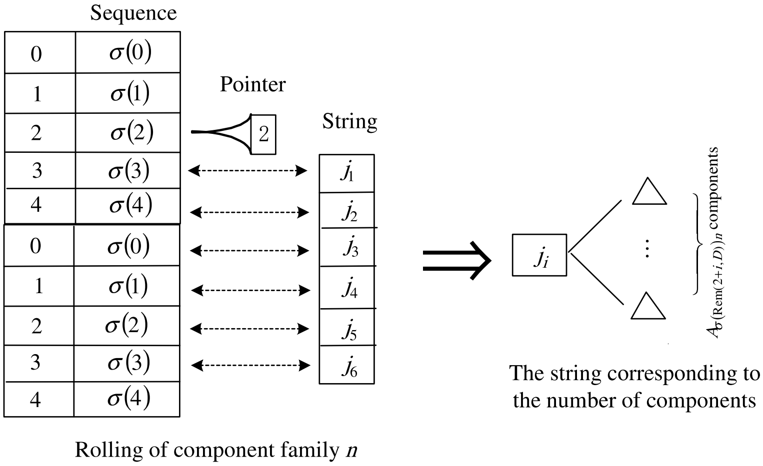

Fig. 2. The rolling approach.

That of $\mathbf{J}(s)$ is given by the cross product of $T_n - 1$ number of the set $\{0, 1\}$, i.e,

$$\Omega_s = \{0, 1\}^{T_n - 1}. \tag{3.3}$$

Since any leadtime incurred by the supplier is independent of the pointer, the following result is straightforward.

Proposition 3.1. *The pointer $X(s)$ and the string $\mathbf{J}(s)$ are independent of each other.*

At an arbitrary integer observation epoch s, $j_i(s)$ in $\mathbf{J}(s)$ characterizes whether or not the order placed at time $s - T_n + i$ has arrived by taking values 1 and 0, with probability $F_n(T_n - i)$ being 1 and probability $1 - F_n(T_n - i)$ being 0. Note that these probabilities are independent of the observation epochs. In other words, they are the same at all observation epochs, hence the stationarity of the distribution. Moreover, due to the independence among all the leadtimes by the supplier, the stationary distribution must possess a product form over $T_n - 1$ dimensions for tracking the $T_n - 1$ placed orders.

Let $\pi_{\boldsymbol{J}}$ denote the stationary probability of $\mathbf{J}(s)$ at state $\boldsymbol{J} = (j_1, \ldots, j_{T_n - 1}) \in \Omega_s$. Then we have the following result.

Proposition 3.2. *The stationary distribution $\pi = (\pi_{\boldsymbol{J}})_{\boldsymbol{J} \in \Omega_s}$ is given by*

$$\pi_{\boldsymbol{J}} = \prod_{i=1}^{T_n - 1} p_i(j_i), \tag{3.4}$$

where

$$p_i(j_i) = \begin{cases} 1 - F_n(T_n - i), & \text{if } j_i = 0, \\ F_n(T_n - i), & \text{if } j_i = 1. \end{cases} \tag{3.5}$$

3.2. Moments of the WIP

Since the vector $(X(s), \mathbf{J}(s))$ provides all the information about the replenishment process of components of family n, we can analyze the WIP from the production sequence $\Sigma = \{\sigma(0), \sigma(1), \ldots, \sigma(D - 1)\}$ and the stationary distribution π.

The manner can best be understood by referring to Fig. 2. Suppose the pointer is 2. Then the first element in $\mathbf{J}(s)$ is related to the third unit in the sequence, the second element is related to the fourth unit, the third one is related to the zeroth unit, and so forth. In general, if the pointer is at j, then the ith element of $\mathbf{J}(s)$ is related to the unit $\sigma(\mathrm{Rem}(j+i, D))$ in the sequence Σ. Thus, from the requirement matrix A given in Eq. (2.2), the corresponding requirement for components of family n for that unit is $A_{\sigma(\mathrm{Rem}(j+i,D)),n}$.

Suppose the pointer is zero. Then the WIP corresponding to the state $\boldsymbol{J} = (j_1, \ldots, j_{T_n-1})$ is given by

$$\sum_{i=1}^{T_n-1} j_i \cdot A_{\sigma(\mathrm{Rem}(i,D)),n}. \tag{3.6}$$

From Proposition 3.2, the expectation of the above is then

$$\pi_{\boldsymbol{J}} \cdot \sum_{i=1}^{T_n-1} j_i \cdot A_{\sigma(\mathrm{Rem}(i,D)),n} = \left(\prod_{i=1}^{T_n-1} p_i(j_i)\right)\left(\sum_{i=1}^{T_n-1} j_i \cdot A_{\sigma(\mathrm{Rem}(i,D)),n}\right).$$

The expectation of the WIP at the pointer being zero is given by taking the summation over the state space of $\mathbf{J}(s)$, i.e.,

$$\sum_{\boldsymbol{J}\in\Omega_s}\left(\pi_{\boldsymbol{J}} \cdot \sum_{i=1}^{T_n-1} j_i \cdot A_{\sigma(\mathrm{Rem}(i,D)),n}\right) = \sum_{\boldsymbol{J}\in\Omega_s}\left(\left(\prod_{i=1}^{T_n-1} p_i(j_i)\right)\left(\sum_{i=1}^{T_n-1} j_i \cdot A_{\sigma(\mathrm{Rem}(i,D)),n}\right)\right).$$

Substituting (3.5) into the above with some algebra leads to

$$A_{\sigma(\mathrm{Rem}(1,D)),n} \cdot F_n(T_n-1) + A_{\sigma(\mathrm{Rem}(2,D)),n} \cdot F_n(T_n-2) + \cdots + A_{\sigma(\mathrm{Rem}(T_n-1,D)),n} \cdot F_n(1)$$

In general, the expectation of the WIP at pointer x, $0 \leqslant x \leqslant D-1$, is given by

$$A_{\sigma(\mathrm{Rem}(x+1,D)),n} \cdot F_n(T_n-1) + A_{\sigma(\mathrm{Rem}(x+2,D)),n} \cdot F_n(T_n-2) + \cdots + A_{\sigma(\mathrm{Rem}(x+T_n-1,D)),n} \cdot F_n(1)$$

Summation over the state space of the pointer and taking the average, we obtain the first moment of the WIP as follows

$$\begin{aligned} E[W_n] &= \frac{1}{D}\sum_{x\in\Omega_p}[A_{\sigma(\mathrm{Rem}(x+1,D)),n} \cdot F_n(T_n-1) + \cdots + A_{\sigma(\mathrm{Rem}(x+T_n-1,D)),n} \cdot F_n(1)] \\ &= \frac{1}{D}\left[(F_n(1) + F_n(2) + \cdots + F_n(T_n-1)) \cdot \sum_{x=0}^{D-1} A_{\sigma(x+1,D),n}\right]. \end{aligned} \tag{3.7}$$

For the second moment of the WIP, we may not give an explicit expression form as writ the first moment above. Nevertheless, we can easily calculate its value according to the following

$$E[W_n]^2 = \frac{1}{D}\sum_{x\in\Omega_p}\sum_{\boldsymbol{J}\in\Omega_s}\left(\pi_{\boldsymbol{J}} \cdot \left(\sum_{i=1}^{T_n-1} j_i \cdot A_{\sigma(\mathrm{Rem}(x+i,D)),n}\right)^2\right). \tag{3.8}$$

In general, the kth moment of the WIP is calculated in accordance with

$$E[W_n]^k = \frac{1}{D}\sum_{x\in\Omega_p}\sum_{\boldsymbol{J}\in\Omega_s}\left(\pi_{\boldsymbol{J}} \cdot \left(\sum_{i=1}^{T_n-1} j_i \cdot A_{\sigma(\mathrm{Rem}(x+i,D)),n}\right)^k\right). \tag{3.9}$$

Remark 3.1. The expected WIP given in Eq. (3.7) is for the WIP at integer time epochs. However, supplied components may arrive at any time between integers. For instance, j_1 may be zero, but the corresponding components may arrive before the next integer time epoch. For such a case, we must consider the possibility in which supplied components arrive between two successive observations in order to find the expected WIP at an arbitrary time. Define a conditional distribution function as

$$H_i(t) = \Pr\{\xi_n \leqslant t | \xi_n > i\}, \quad i = 0, 1, \ldots, T_n - 1.$$

If $j_1 = 0$ for state $\boldsymbol{J}(s)$ observed at an arbitrary integer time epoch s, according to the "time proportion" defined for a WIP, the expected WIP in the interval $(s, s+1]$ is given by

$$A_{\sigma(\mathrm{Rem}(x+1,D)),n} \cdot \int_{T_n-1}^{T_n} (T_n - t) H_{T_n-1}(\mathrm{d}t).$$

For $j_i = 0$, $i = 2, \ldots, T_n - 1$, the expected WIP in the interval $(s, s+1]$ is given by

$$A_{\sigma(\mathrm{Rem}(x+i,D)),n} \cdot \int_{T_n-i}^{T_n-i+1} [T_n - i + 1 - t] H_{T_n-i}(\mathrm{d}t).$$

Then adding the above to (3.6) forms the expected WIP at any time epoch.

Remark 3.2. For an actual MMAL, some models of products may not require a particular family of components, i.e., $A_{m,n} = 0$ for some m and n. Such a case can still be treated by our method by introducing a dummy supply with the same leadtime ξ_n. In doing so, the stationary probability distribution does not change. This dummy supply does not influence the real value of WIP since $A_{m,n} = 0$.

Remark 3.3. Once the WIPs of individual component families are found, the total WIP associated with workstation k can be found easily by summing up all WIPs of the component families $\{n_{k-1}+1, n_{k-1}+2, \ldots, n_k\}$. In practice, such a result is useful for the configuration of the workstations.

3.3. Some insights

In the design stage of the system, one important issue is to plan resources, such as space or the sizes of the shelves for holding arrived components. The first and second moments of the WIPs provide helpful and valuable information for such work.

Consider the WIP of family n. Recall that d_m is the number of model m products in the sequence Σ. Then it holds that

$$\sum_{x=0}^{D-1} A_{\sigma(x+1,D),n} = \sum_{m=1}^{M} A_{m,n} \cdot d_m.$$

The above implies that given the d_m's, the number of model m products produced per day, the expectation of the WIP (3.7) is independent of the sequence Σ. This is reasonable because a particular requirement for components only changes the relative usage time in different sequences. Their contributions to the expected WIP are the same at any using time under the JIT supply policy. Nevertheless, the second moment of the WIP may depend on the sequence, which leads to different variances of the WIP. We shall investigate such behavior numerically in Section 5.

From (3.7), on the other hand, the expected WIP depends on the distribution function $F_n(t)$ of the supply times ξ_n's. It indicates that the smaller the summation term $F_n(1) + F_n(2) + \cdots + F_n(T_n-1)$, the smaller the

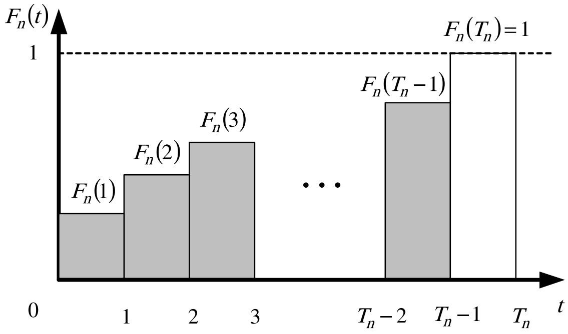

Fig. 3. Relationship of WIP and $F_n(t)$.

expected WIP. This summation is equivalent to the shaded area in Fig. 3. To reduce the expected WIP, the endeavors may take into consideration of shortening the upper bound of the leadtimes T_n first (see, for example, Pan and Yang, 2002; Grout and Christy, 1999). However, it is not a unique approach for lowering the expected WIP. In fact, even if T_n is large, we can still maintain the expected WIP at a low level through enhancing the stability of ξ_n's. Since $F_n(t)$ is a probability distribution function, it possesses the nondecreasing property. This implies that to reduce the shaded area shown in Fig. 3, reduction of $F_n(i)$ is more important than reduction of $F_n(j)$ for $i < j$. Moreover, we can consider the shaded area at a limiting form as follows

$$\int_0^{T_n} F_n(t)\,\mathrm{d}t.$$

On the other hand, if we take the expectation of ξ_n, we have

$$E[\xi_n] = \int_0^{T_n} \bar{F}_n(t)\,\mathrm{d}t = \int_0^{T_n} (1 - F_n(t))\,\mathrm{d}t = T_n - \int_0^{T_n} F_n(t)\,\mathrm{d}t,$$

or equivalently,

$$\int_0^{T_n} F_n(t)\,\mathrm{d}t = T_n - E[\xi_n]. \tag{3.10}$$

The above gives us a insight such that the shaded area or the expected WIP $E[W_n]$ can be reduced, no matter how long the upper bound of the leadtimes T_n is, provided we can make the average leadtime $E[\xi_n]$ as close as to the upper bound as possible.

Remark 3.4. Components are delivered with leadtimes according to the T_n's to avoid component shortages. The maximum resource needed by a particular component family for holding arrived components is relevant to its T_n. If the resources for all component families are set according to these maximum requirements, components can be admitted whenever they arrive, but the utilization ratio of the resources may be low. Other than such safe mode, to improve resource utilization, we can determine the resource capacities based on the first and second moments of the WIPs. Then risks of resource shortages may happen. Nevertheless, a special common resource can be spared to absorb risks of resource shortages over all component families. In such a case, the total resources required can be much less than in the safe mode.

4. Analysis of finished products in storehouses

In this section, we study the inventories of finished products in storehouses, and the service levels for customers. It is clear that the different models of the finished products are independent. Thus we focus on just one model of products, say model m ($1 \leqslant m \leqslant M$), and we call customers who require model m products "type m customers". First, we introduce an embedded Markov chain associated with the number of model m products at integer time epochs. Then we analyze the Markov chain in the steady state. Based on the stationary distribution of the Markov chain, moments of the inventory and service level are obtained. Recall that $T_c = 1$ and the production sequence is $\Sigma = \{\sigma(0), \sigma(1), \ldots, \sigma(D-1)\}$.

4.1. The embedded Markov chain

Without loss of generality, we assume that the first finished product leaves the MMAL at time 0. Then subsequent finished products move off the MMAL at integer time epochs; each of them enters the corresponding storehouse for satisfying customer demand. We construct a Markov chain by observing the inventory of model m products at the epochs at which a finished product (regardless of its model) leaves the MMAL. Define a variable $Y(s) = \mathrm{Rem}(s, D)$. Then $Y(s)$, together with the sequence Σ, provide information about the model of the product moving off the MMAL at (integer) time s. The variable $Y(s)$ takes values $\{0, 1, 2, \ldots, D-1\}$ cyclically. (Although $Y(s)$ takes the same values as $X(s)$ defined in the previous section, their physical meanings are somewhat different. Therefore we introduce $Y(s)$ as a new variable here.) Let $I_m(s)$ be the number of model m products in storehouse m right after time s (i.e., at time s^+). The random variable $I_m(s)$ takes nonnegative integer values.

Putting $Y(s)$ and $I_m(s)$ together, we obtain a process ($I_m(s)$, $Y(s)$) with information about the inventory status in storehouse m at integer time epochs. It is easy to see that ($I_m(s)$, $Y(s)$) is a Markov chain. We call $I_m(s)$ the *level variable* and $Y(s)$ the *phase variable*.

Let $\delta_{m,\sigma(y)} = 1$, if $\sigma(y) = m$; 0, otherwise, for $0 \leqslant y \leqslant D-1$. Then we have

$$d_m = \sum_{y=0}^{D-1} \delta_{m,\sigma(y)} \tag{4.1}$$

being the number of model m products in the sequence Σ, or the number of model m products produced per day. Note that if $\sigma(Y(s)) = m$, we must have $I_m(s) > 0$. Thus, the infinite state space of the Markov chain ($I_m(s)$, $Y(s)$) is:

$$\Omega_I = \{(0, y) : 0 \leqslant y \leqslant D-1, \sigma(y) \neq m\} \cup \{(i, y) : i \geqslant 1 \text{ and } 0 \leqslant y \leqslant D-1\}.$$

It is easy to see that every level has D states except level 0, which has $D - d_m$ states.

Between two consecutive observation epochs (i.e., during the period of a unit time), $I_m(s)$ can increase at most by one. On the other hand, since there can be any number of demands arriving during a unit time, $I_m(s)$ may drop to zero. Therefore, the Markov chain ($I_m(s)$, $Y(s)$) is a typical $GI/M/1$ type Markov chain (see Neuts, 1981), where GI refers to the entry process of finished products into the storehouse and M refers to the arrival process of type m customers. Next, we find the one-step transition probabilities of ($I_m(s)$, $Y(s)$).

During a unit time, the probability that i type m customers arrived at storehouse m is given by

$$u_i = \frac{(\lambda_m)^i}{i!} \mathrm{e}^{-\lambda_m}, \quad i \geqslant 0. \tag{4.2}$$

Define a diagonal matrix Γ as follows

$$\Gamma = \begin{pmatrix} \delta_{m,\sigma(0)} & & & \\ & \delta_{m,\sigma(1)} & & \\ & & \ddots & \\ & & & \delta_{m,\sigma(D-1)} \end{pmatrix}. \tag{4.3}$$

The matrix Γ is used to find out whether or not a finished product is of model m.

Let

$$P = \begin{pmatrix} 0 & I_{D-1} \\ 1 & 0 \end{pmatrix},$$

where I_{D-1} is an identity matrix of size $D-1$.

Define

$$\hat{U}_{0,0} = (I_D - \Gamma)P(I_D - \Gamma) + \Gamma P \quad \text{and} \quad \hat{U}_{0,1} = (I_D - \Gamma)P\Gamma, \tag{4.4}$$

and for $k \geqslant 1$

$$\begin{aligned} \hat{U}_{k,0} &= \left(1 - \sum_{i=0}^{k-1} u_i\right) P(I_D - \Gamma), \\ U_{k,1} &= \left(1 - \sum_{i=0}^{k-1} u_i\right) P\Gamma + u_{k-1}P(I_D - \Gamma), \\ U_{k,j} &= U_{k+1-j} \equiv u_{k-j}P(I_D - \Gamma) + u_{k+1-j}P\Gamma, \quad 2 \leqslant j \leqslant k, \\ U_{k,k+1} &= U_0 \equiv u_0 P\Gamma, \\ U_{k,j} &= 0, \quad j \geqslant k+2. \end{aligned} \tag{4.5}$$

Let $U_{0,0}$ be a $(D-d_m)\times(D-d_m)$ matrix obtained by deleting all rows and columns in the matrix $\hat{U}_{0,0}$ that satisfy $\sigma(i)=m$, where i is the row number or the column number. Let $U_{0,1}$ be a $(D-d_m)\times D$ matrix obtained by deleting all rows in matrix $\hat{U}_{0,1}$ that satisfy $\sigma(i)=m$, where i is the row number. Moreover, let $U_{k,0}$ be a $D\times(D-d_m)$ matrix obtained by deleting all columns in matrix $\hat{U}_{k,0}$ that satisfy $\sigma(i)=m$, where i is the column number. The probability transition blocks from level 0 to level 0 and from level 0 to level 1 are $U_{0,0}$ and $U_{0,1}$, respectively. The probability transition blocks from level k to other levels are $\{U_{k,j}, j \geqslant 0\}$ for $k \geqslant 1$. Note that matrix P plays the role of shifting the production sequence by one unit of time, and the matrix Γ is for tracking whether or not the model of the product just finished is of model m. The above transition blocks can be obtained easily from their intuitive interpretations. For instance, the transition from k to j for $2 \leqslant j \leqslant k$ can take place in two ways: one is when $k-j$ demands arrived and the finished product is not of model m; the other is when $k+1-j$ demands arrived and the finished product is of model m.

With the above background, we are ready to give the one step transition matrix of the Markov chain $(I_m(s), Y(s))$.

Proposition 4.1. *The one-step transition probability matrix of the Markov chain* $(I_m(s), Y(s))$ *is given by*

$$P_F = \begin{pmatrix} U_{0,0} & U_{0,1} & & & & \\ U_{1,0} & U_{1,1} & U_0 & & & \\ U_{2,0} & U_{2,1} & U_1 & U_0 & & \\ U_{3,0} & U_{3,1} & U_2 & U_1 & U_0 & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \tag{4.6}$$

where the matrix blocks in P_F *are defined in Eq.* (4.5) *or following the discussion immediately after Eq.* (4.5).

Let $U = \sum_{k=0}^{\infty} U_k$. It can be verified that

$$U = P\Gamma + P(I_D - \Gamma) = P, \tag{4.7}$$

which indicates that U is also irreducible. Therefore, there exists a strictly positive left-eigenvector $\mathbf{\theta}$ of U, such that $\mathbf{\theta}U = 1$ and $\mathbf{\theta e} = 1$. In fact, it is easy to see that $\mathbf{\theta} = (1/D, \ldots, 1/D)$. Furthermore, define Λ^* as

$$\Lambda^* = \sum_{k=1}^{\infty} kU_k\mathbf{e}. \tag{4.8}$$

A condition for the Markov chain ($I_m(s)$, $Y(s)$) (with infinite state space) to be positive recurrent is given in the following important proposition.

Proposition 4.2. *The irreducible Markov chain* ($I_m(s)$, $Y(s)$) *is positive recurrent if and only if*

$$\rho_m = \frac{d_m}{\lambda_m D} < 1. \tag{4.9}$$

Proof. According to Neuts (1981), since the Markov chain ($I_m(s)$, $Y(s)$) is irreducible, it is positive recurrent if and only if $\mathbf{\theta}\Lambda^* > 1$. By using expressions in Eq. (4.5), routine calculations lead to

$$\mathbf{\theta}\Lambda^* = \mathbf{\theta}\left(\sum_{k=1}^{\infty} k[u_{k-1}P(I_D - \Gamma) + u_k P\Gamma]\right)e = \mathbf{\theta}((1+\lambda_m)P(I_D - \Gamma) + \lambda_m P\Gamma)e = 1 + \lambda_m - \frac{d_m}{D}. \tag{4.10}$$

Thus, $\mathbf{\theta}\Lambda^* > 1$ is equivalent to Eq. (4.9). This completes the proof of the proposition. □

Remark 4.1. For a standard $GI/G/1$ queueing system, a condition for system stability is $\rho = \lambda/\mu < 1$, i.e., the arrival rate λ is less than the service rate μ (see, for example, Wolff, 1989). For the Markov chain ($I_m(s)$, $Y(s)$), ρ_m defined in Eq. (4.9) gives a similar system stability condition. In fact, λ_m is the demand rate of model m products (equivalent to the service rate μ in a standard $GI/G/1$ queueing system) and d_m/D is the production rate of model m products (equivalent to the arrival rate λ in a standard $GI/G/1$ queueing system). Thus, ρ_m is the ratio of demand rate to the production rate.

We assume that $\rho_m < 1$ so that the stationary distribution of the Markov chain ($I_m(s)$, $Y(s)$) exists. Denote the stationary distribution by

$$\tilde{\Pi} = (\tilde{\pi}(0), \tilde{\pi}(1), \tilde{\pi}(2), \ldots), \tag{4.11}$$

where $\tilde{\pi}(0) = (\tilde{\pi}_0(0), \tilde{\pi}_1(0), \ldots, \tilde{\pi}_{D-d_m-1}(0))$ and $\tilde{\pi}(k) = (\tilde{\pi}_0(k), \tilde{\pi}_1(k), \ldots, \tilde{\pi}_{D-1}(k))$, $k \geqslant 1$. Let R be the minimal nonnegative solution to the equation:

$$R = \sum_{k=0}^{\infty} R^k U_k. \tag{4.12}$$

Using results in Neuts (1981), the stationary distribution of the Markov chain ($I_m(s)$, $Y(s)$) can be obtained.

Proposition 4.3. *If condition* (4.9) *is satisfied, then the spectrum of the matrix R is less than one.* $(\tilde{\pi}(0), \tilde{\pi}(1))$ *is the unique positive solution of the following equation*

$$\begin{cases} \tilde{\pi}(0) = \tilde{\pi}(0)U_{0,0} + \tilde{\pi}(1)\sum\limits_{k=1}^{\infty} R^{k-1}U_{k,0}, \\ \tilde{\pi}(1) = \tilde{\pi}(0)U_{0,1} + \tilde{\pi}(1)\sum\limits_{k=1}^{\infty} R^{k-1}U_{k,1}, \\ \tilde{\pi}(0)e + \tilde{\pi}(1)(I_D - R)^{-1}e = 1, \end{cases} \tag{4.13}$$

and for $k = 2, 3, \ldots,$ *we have*

$$\tilde{\pi}(k) = \tilde{\pi}(1)R^{k-1}. \tag{4.14}$$

Proof. The results can be obtained from Theorem 1.3.2 of Neuts (1981). □

According to Neuts (1981), the procedure to compute the stationary distribution can be roughly stated as follows. First, given an initial matrix $R = 0$, calculate an approximate solution of Eq. (4.12) by an iterative method. Then calculate $\tilde{\pi}(0)$ and $\tilde{\pi}(1)$ according to equations in (4.13). Finally, calculate other $\{\tilde{\pi}(k), k \geqslant 2\}$ by using Eq. (4.14).

4.2. Moments of inventories

Based on the stationary distribution of the Markov chain, we can easily calculate the expected inventory of model m finished products in the corresponding storehouse, $E[I_m]$, as follows

$$E[I_m] = \sum_{k=1}^{\infty} \sum_{y=0}^{D-1} \tilde{\pi}_y(k) \cdot k. \tag{4.15}$$

The second moment, $E[I_m]^2$, is given by

$$E[I_m]^2 = \sum_{k=1}^{\infty} \sum_{y=0}^{D-1} \tilde{\pi}_y(k) \cdot k^2. \tag{4.16}$$

In general, its nth moment, $E[I_m]^n$, is then

$$E[I_m]^n = \sum_{k=1}^{\infty} \sum_{y=0}^{D-1} \tilde{\pi}_y(k) \cdot k^n. \tag{4.17}$$

In the design stage, one important consideration is to determine spaces (or the sizes of the storehouses) for storing finished products. The first and second moments of the inventories provide helpful and valuable information for such work. Moreover, the moments may depend on the sequence Σ. We shall investigate behavior in terms of the sequence in the next section for some managerial insights.

Remark 4.2. Since the observations occur at integer time epochs, the above stationary distribution characterizes the inventories at these time epochs. If one wants to find the inventory at an arbitrary time, we need to find the inventory between integer time epochs (or between observation epochs), since demands may arrive at any time and take away products upon their arrivals. Let τ_i be the arrival time of the ith demand from a Poisson stream, $i \geqslant 1$. According to Theorem 2.3.1 in Ross (1983), given that there are exactly i demands arrived in $[0, 1]$, $\{\tau_1, \ldots, \tau_i\}$ possess the same joint distribution as the order statistics corresponding to i independent random variables uniformly distributed on the interval $[0, 1]$. Following a general definition of inventory on the time proportion (see, for example, Section 3.2.2 in Buzacott and Shanthikumar, 1993), given that the inventory is k at the beginning of the interval $[0, 1]$, the expected inventory in that interval is given by, for $i \leqslant k$,

$$\begin{aligned} &E[k\tau_1 + (k-1)(\tau_2 - \tau_1) + \cdots + (k-i)(1 - \tau_1 - \cdots - \tau_i) | \tau_i \leqslant 1 < \tau_{i+1}] \\ &\quad = E[k - i + \tau_1 + \tau_2 + \cdots + \tau_i | \tau_i \leqslant 1 < \tau_{i+1}] = k - \frac{i}{2}, \end{aligned} \tag{4.18}$$

and, for $i > k$,

$$E[k\tau_1 + (k-1)(\tau_2 - \tau_1) + \cdots + (\tau_k - \tau_{k-1}) | \tau_i \leqslant 1 < \tau_{i+1}] = E\left[\sum_{j=1}^{k} \tau_j \middle| \tau_i \leqslant 1 < \tau_{i+1}\right] = \sum_{j=1}^{k} \frac{j}{i+1} = \frac{k(k+1)}{2(i+1)}. \tag{4.19}$$

Note that if $i > k$, the inventory will become zero after the kth demand arrival epoch. Hence we only consider the first k arrival epochs. By using Eqs. (4.18) and (4.19), given that the initial state is (k, y), the expected inventory between two consecutive integer time epochs is given by

$$\begin{aligned}
&\left(k - \frac{0}{2}\right) \cdot \frac{(\lambda_m)^0}{0!} \mathrm{e}^{-\lambda_m} + \left(k - \frac{1}{2}\right) \cdot \frac{(\lambda_m)^1}{1!} \mathrm{e}^{-\lambda_m} + \cdots + \left(k - \frac{k}{2}\right) \cdot \frac{(\lambda_m)^k}{k!} \mathrm{e}^{-\lambda_m} \\
&\quad + \frac{k(k+1)}{2(k+2)} \cdot \frac{(\lambda_m)^{k+1}}{(k+1)!} \mathrm{e}^{-\lambda_m} + \frac{k(k+1)}{2(k+3)} \cdot \frac{(\lambda_m)^{k+2}}{(k+2)!} \mathrm{e}^{-\lambda_m} + \cdots \\
&= \sum_{i=0}^{k} \left(k - \frac{i}{2}\right) \cdot \frac{(\lambda_m)^i}{i!} \mathrm{e}^{-\lambda_m} + \sum_{i=k+1}^{\infty} \frac{k(k+1)}{2(i+1)} \cdot \frac{(\lambda_m)^i}{i!} \mathrm{e}^{-\lambda_m} \\
&= \mathrm{e}^{-\lambda_m} \left\{ \sum_{i=0}^{k} \left(k - \frac{i}{2} - \frac{k(k+1)}{2(i+1)}\right) \frac{(\lambda_m)^i}{i!} \right\} + \frac{k(k+1)}{2\lambda_m} (1 - \mathrm{e}^{-\lambda_m}).
\end{aligned} \tag{4.20}$$

From the above analysis, the expected inventory of finished products in storehouse m at an arbitrary time (in the steady state) is given by

$$\sum_{k=1}^{\infty} \sum_{y=0}^{D-1} \tilde{\pi}_y(k) \left(\mathrm{e}^{-\lambda_m} \left\{ \sum_{i=0}^{k} \left(k - \frac{i}{2} - \frac{k(k+1)}{2(i+1)}\right) \frac{(\lambda_m)^i}{i!} \right\} + \frac{k(k+1)}{2\lambda_m} (1 - \mathrm{e}^{-\lambda_m}) \right). \tag{4.21}$$

4.3. Service levels for customers

The service level, S_m, for type m customers is defined as the proportion of the number of satisfied customers in steady state. Similar to (4.21), the expected number of satisfied customers per time unit is

$$\sum_{k=1}^{\infty} \sum_{y=0}^{D-1} \tilde{\pi}_y(k) \left(\sum_{i=0}^{k-1} (\lambda_m - k) \frac{(\lambda_m)^i}{i!} \mathrm{e}^{-\lambda_m} - \frac{(\lambda_m)^k}{(k-1)!} \mathrm{e}^{-\lambda_m} + k \right). \tag{4.22}$$

According to the Poisson property, the expected number of customer arrivals per time unit is λ_m. By definition, the service level for type m customers is consequently given by

$$S_m = \frac{\sum_{k=1}^{\infty} \sum_{y=0}^{D-1} \tilde{\pi}_y(k) \left(\sum_{i=0}^{k-1} (\lambda_m - k) \frac{(\lambda_m)^i}{i!} \mathrm{e}^{-\lambda_m} - \frac{(\lambda_m)^k}{(k-1)!} \mathrm{e}^{-\lambda_m} + k \right)}{\lambda_m}. \tag{4.23}$$

Remark 4.3. Since the arrival rate of type m customers is λ_m, the production rate of model m products from the MMAL is d_m/D, and model m products are taken by type m customers only, we must have $S_m = \rho_m = (d_m/D)/\lambda_m = d_m/(\lambda_m\ D)$. Consequently, we have the following convergence conclusion of the series in the numerator of (4.23)

$$\sum_{k=1}^{\infty} \sum_{y=0}^{D-1} \tilde{\pi}_y(k) \left(\sum_{i=0}^{k-1} (\lambda_m - k) \frac{(\lambda_m)^i}{i!} \mathrm{e}^{-\lambda_m} - \frac{(\lambda_m)^k}{(k-1)!} \mathrm{e}^{-\lambda_m} + k \right) = \frac{d_m}{D}. \tag{4.24}$$

5. Numerical results

The sequence Σ plays a central role in the whole system. In this section, we investigate, through numerical results, the influences on WIPs of components, inventories of finished products and service levels for customers by different sequences.

5.1. WIP of components

In Section 3, we have provided managerial insights in terms of the leadtimes for the expected WIP. We have shown that the first moment of the WIP is independent of the sequence Σ. Here we investigate the variances of the WIP for different sequences.

Consider the system with the following parameters. (As pointed out in Section 3, without loss of generality, we only focus on a single family of components.)

The number of product models: $M = 10$.

The component requirements for the family by the product models: $[A_{1,n}, \ldots, A_{10,n}] = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$.

The bound of the leadtimes: $T_n = 10$.

The distribution function of the leadtimes: $[F(1), \ldots, F(T_n - 1)] = [F(1), F(2), F(3), F(4), F(5), F(6), F(7), F(8), F(9)] = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$.

The total number of products per day: $D = 100$.

The number of each product model per day: $d_1 = d_2 = \cdots = d_{10} = 10$.

For the above system, we consider the following three typical sequences:

(1) Batch-sequence: $\Sigma_1 = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10]$;

(2) Uniform-sequence: $\Sigma_2 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$;

(3) Even-sequence: $\Sigma_3 = [1, 10, 2, 9, 3, 8, 4, 7, 5, 6, 1, 10, 2, 9, 3, 8, 4, 7, 5, 6, 1, 10, 2, 9, 3, 8, 4, 7, 5, 6, 1, 10, 2, 9, 3, 8, 4, 7, 5, 6, 1, 10, 2, 9, 3, 8, 4, 7, 5, 6, 1, 10, 2, 9, 3, 8, 4, 7, 5, 6, 1, 10, 2, 9, 3, 8, 4, 7, 5, 6, 1, 10, 2, 9, 3, 8, 4, 7, 5, 6, 1, 10, 2, 9, 3, 8, 4, 7, 5, 6, 1, 10, 2, 9, 3, 8, 4, 7, 5, 6]$.

The "batch-sequence" continuously feeds the same model onto the conveyor till exhaustion of the products to the model in a production cycle. The "uniform-sequence" repeats from model 1 to model 10 in the sequence. Referring to A, the component requirements by the product models, the "even-sequence" means that units with less component requirement and more component requirement or units with closer component requirement to the average requirement alternatively appear in the sequence. That is, a unit with less component requirement is followed by a unit with more component requirement, and a unit with the component requirement closer to the average requirement is followed by one with the component requirement also closer to the average requirement. The computational results are as follows:

Expected WIP	Variances of WIP		
24.75	Batch-sequence: Σ_1 208.02	Uniform-sequence: Σ_2 78.79	Even-sequence: Σ_3 66.09

Although the expected WIP is not large, the variances can be very different under different sequences. The batch-sequence is the worst sequencing method, whereas the even-sequence is the best one. Therefore, the manager should consider adopting an even-sequence as far as possible for saving shelf resources or for reducing the risk of space shortage for capacitated shelf resources.

5.2. Inventories of finished products

Suppose the system is the same as in the previous subsection. Customers' demands for all product models follow Poisson streams with rates: $[\lambda_1,\ldots,\lambda_{10}] = [0.105, 0.110, 0.114, 0.117, 0.119, 0.120, 0.125, 0.130, 0.135, 0.140]$.

For the uniform-sequence Σ_2 and the even-sequence Σ_3, any product model appears evenly in the both sequences. It is clear, from the long-term-behavior, that the uniform-sequence Σ_2 possesses the same efficiency as the even-sequence Σ_3 as far as the behavior of finished products in the storehouses is concerned. Therefore, we only need to consider two typical sequences: Batch-sequence, and Even-sequence.

We investigate the first moments and the variances of inventories of finished products. They provide important information for planning the storehouses, such as the space of the storehouses. The following are the computational results:

The first moments										
Product model	1	2	3	4	5	6	7	8	9	10
Batch-sequence	12.99	7.92	6.43	5.76	5.42	5.28	4.71	4.31	4.00	3.76
Even-sequence	10.21	5.21	3.78	3.15	2.84	2.71	2.21	1.87	1.63	1.45
The variances										
Product model	1	2	3	4	5	6	7	8	9	10
Batch-sequence	111.83	35.34	22.73	18.49	16.71	16.01	13.67	12.40	11.61	11.09
Even-sequence	103.47	26.81	14.09	9.77	7.94	7.22	4.81	3.48	2.66	2.12

The results show that the batch-sequence generates larger average (the first moment) inventories and larger variances. The relative differences of the average values or the variance values indicate that product model 1 is the most nonsensitive one whereas product model 10 is the most sensitive one with respect to the sequence. Referring to the stability condition (the positive recurrent condition (4.9)) by Proposition 4.2, we know that $[\rho_1,\ldots,\rho_{10}] = [0.952, 0.909, 0.877, 0.855, 0.840, 0.833, 0.800, 0.770, 0.742, 0.714]$; product model 1 with the value ρ_1 closest to 1, and the larger the index number m, the smaller the ρ_m values.

The above analysis gives us insight to determine the number of each product model per day, d_m's, for $\rho_m = d_m/\lambda_m D$ from (4.9). To attain relatively stable behavior of inventories with respect to different sequences, the manager should determine the d_m's in such a manner as to let the ρ_m's be as close to 1 as possible. Of course, this will lead the corresponding inventories to increase.

Remark 5.1. A storehouse is regarded as a $GI/M/1$ type queueing system with GI referring to the output process of finished products related to the sequence. In a $GI/M/1$ type queueing system, the average number of customers should be minimized if the variance of arrivals is minimized. Therefore, a pattern of the optimal sequence for minimizing the expected inventories may be such that the output over all product models is smoothed as far as possible. In other words, different product models positioned in a cycle length D should be smooth as far as possible.

5.3. Service levels for customers

As expected from (4.23) and (4.24), the service level for type m customers is given by

$$S_m = \frac{d_m}{\lambda_m D},$$

which is independent of sequences. For the system with the parameters in the previous subsections, calculations generate the service levels being $[S_1, \ldots, S_{10}] = [0.952, 0.909, 0.877, 0.855, 0.840, 0.833, \ 0.800, 0.770, 0.742, 0.714]$ for any sequence.

An important concern in the design stage is to determine the production capacity, or equivalently to determine D that is formed by $d_1, \ldots, d_M$. Consequently, assigning $d_1, \ldots, d_M$ for a given D is an interesting task. One approach for determining the appropriate d_m's for any stationary demand processes is to make $d_m / \sum_{i=1}^{M} d_i (= d_m/D)$ as close to λ_m as possible. Raising customer service levels by letting d_m/D be as close to λ_m as possible will cause ρ_m to be close to 1, hence increases inventories of finished products in the corresponding storehouse. Therefore, the manager must make a balance from the trade-off between inventories and service levels.

6. Future extensions

This paper addresses an important issue in the design stage of the supply chain, performance analysis, for investigating the "long-term-behavior" of the system. One of the future research tasks is the daily control for the system, which is a crucial issue involved in its operational stage. The most challenging issue is to dynamically determine the production sequence every day. That is, given the information on a particular day such as the WIP on the shelves at the beginning of the day, the distributions of customers' demands during the day, the inventories in the storehouses at the beginning of the day, what is the optimal sequence for that day? This problem is well discussed in the literature as a typical issue in an MMAL in deterministic model (see, for example, Aigbedo, 2004; Scholl, 1999 and the cited references in it). For the supply chain in the stochastic environment considered in this paper, the objective of the optimal sequence can be set as minimizing the variances of WIP, minimizing the inventories of finished products in the storehouses, and maximizing the service levels for customers. Particularly, minimizing the variances of WIP is just fitting Toyota's goal: smoothing the usage rate of every component family (Monden, 1998).

Acknowledgements

The research is supported by NSF of China under grant 70325004 for the first author, grant 10071051 for the second author, the grants of the Distinguished Young Investigator and the Hundred Talents Program of the Chinese Academy of Sciences for the third author, and by NSERC for the fourth author. The authors would like to thank Professor Salah E. Elmaghraby (from Department of Industrial Engineering, North Carolina State University) for his valuable comments and kind proofreading that lead to the improvement of the paper.

References

Aigbedo, H., 2004. Analysis of parts requirements variance for a JIT supply chain. International Journal of Production Research 42, 417–430.

Axsater, S., 2000a. Inventory Control. Kluwer Academic Publisher.

Axsater, S., 2000b. Exact analysis of continuous review (R, Q) policies in two-echelon inventory systems with compound Poisson demand. Operations Research 48, 686–696.

Buzacott, J.A., Shanthikumar, J.G., 1993. Stochastic Models of Manufacturing Systems. Prentice Hall, New Jersey.

Chopra, S., Meindl, P., 2001. Supply Chain Management. Prentice Hall, New Jersey.

Cetinkaya, S., Parlar, M., 2002. Note: Optimality conditions for an (s, S) policy with proportional and lump-sum penalty costs. Management Science 48, 1635–1639.

Evans, R., 1967. Inventory control of a multiproduct system with a limited production resource. Laval Research Logistics Quarterly 14, 173–184.

Gagnon, R., Ghosh, S., 1991. Assembly line research: Historical roots, research life cycles and future directions. Omega 19 (5), 381–399.

Grout, J.R., Christy, D.P., 1999. A model of supply responses to just-in-time delivery requirements. Group Decision and Negotiation 8, 139–156.

Kelle, P., Miller, P.A., 2001. Stockout risk and order splitting. International Journal of Production Economics 71, 407–415.

Lee, H.L., Padmanabhan, V., Whang, S., 1997. Information distortion in a supply chain: The bullwhip effect. Management Science 43, 546–558.

Lima-Fernandes, C.J., Groover, M.P., 1995. Mixed-model assembly line balancing and sequencing: A survey. Engineering Design and Automation 1 (1), 33–42.

Miyazaki, S., Ohta, H., Nishiyama, N., 1988. The optimal operation planning of kanban to minimize the total operation cost. International Journal of Production Research 26, 1605–1611.

Monden, Y., 1998. Toyota Production System, third ed. Chapman and Hall.

Muckstadt, J.A., Tayur, S.R., 1995. A comparison of alternative kanban control mechanisms. I: Background and structural results. IIE Transactions 27, 140–150.

Neuts, M.F., 1981. Matrix-Geometric Solutions in Stochastic Models, an Algorithm Approach. The Johns Hopkins University Press, Baltimore and London.

Olhager, J., 2002. Supply chain management: A just-in-time perspective. Production Planning and Control 13, 681–687.

Pan, J.C.H., Yang, J.S., 2002. A study of an integrated inventory with controllable lead time. International Journal of Production Research 40, 1263–1273.

Parlar, M., 2000. Probabilistic analysis of renewal cycles: An application to a non-Markovian inventory problem with multiple objectives. Operations Research 48, 243–255.

Perez, A.P., Zipkin, P., 1997. Dynamic scheduling rules for a multiproduct make-to-stock queue. Operations Research 45, 919–930.

Raghavan, N.R.S., Viswanadham, N., 2001. Generalized queueing network analysis of integrated supply chains. International Journal of Production Research 39, 205–224.

Rajagopalan, S., 2002. Make to order or make to stock: Model and application. Management Science 48, 241–256.

Ross, S.M., 1983. Stochastic Processes. John Wiley & Sons, Inc.

Sarker, B.R., Balan, V.C., 1999. Operations planning for a multi-stage kanban system. European Journal of Operational Research 112, 284–303.

Scholl, A., 1999. Balancing and Sequencing of Assembly Lines, second ed. Physica-Verlag, A Springer-Verlag Company.

Silver, E.A., Pyke, D.F., Peterson, R., 1998. Inventory Management and Production Planning and Scheduling, third ed. John Wiley & Sons, Inc.

Sobel, M.J., Zhang, R.Q., 2001. Inventory policies for systems with stochastic and deterministic demand. Operations Research 49, 157–162.

Song, J., Yao, D., 2002. Performance analysis and optimization of assemble-to-order systems with random leadtimes. Operations Research 50, 889–903.

Spearman, M.L., Zazanis, M.A., 1992. Push and pull production systems: Issues and comparisons. Operations Research 40 (3), 521–532.

Tayur, S.R., 1993. Computing the optimal policy for capacitated inventory models. Stochastic Models 9, 585–598.

Wolff, R.W., 1989. Stochastic Modeling and the Theory of Queues. Prentice Hall, New Jersey.

Yanagawa, Y., Miyazaki, S., Ohta, H., 1994. The optimal operation planning of a kanban system with variable lead times. Production Planning and Control 5, 21–29.

Yano, C.A., Bolat, A., 1989. Survey, development and application of algorithms for sequencing paced assembly lines. Journal of Manufacturing and Operations Management 2 (3), 172–198.

Zipkin, P., 2000. Foundations of Inventory Management. McGraw-Hill.