Int J Adv Manuf Technol (2008) 37:564–575
DOI 10.1007/s00170-007-1004-1

ORIGINAL ARTICLE

# Combining multi-class queueing networks and inventory models for performance analysis of multi-product manufacturing logistics chains

**Yifan Wu · Ming Dong**

Received: 15 October 2006 / Accepted: 6 March 2007 / Published online: 31 March 2007

**Abstract** Manufacturing logistics chains consist of complex interconnections among several suppliers, manufacturing facilities, warehouses, retailers and logistics providers. Performance modeling and analysis become increasingly more important and difficult in the management of such complex manufacturing logistics networks. Many research studies have developed different methods to solve such problems. However, most of the research focuses on logistics systems with either a single stage or single type of product. In the real world, industries always involve multiple stages and produce multiple types of products at one stage. This paper is geared toward developing a new methodology by combining multi-class queueing networks and inventory models for the performance analysis of multi-product manufacturing logistic chains. A network of multi-class inventory queue models is presented for the performance analysis of a serial multi-stage manufacturing logistics chain in which multiple types of products are produced at each stage. A job queue decomposition strategy is employed to analyze the major performance measures and an approach for aggregating input streams and separating output streams is proposed to link all the sites or nodes in the logistics chain together. Numerical results show that the proposed method is effective for the application examples.

**Keywords** Multi-class queueing networks · Inventory models · Multi-stage manufacturing logistic chains · Aggregation · Separation

## 1 Introduction

A manufacturing logistic chain can be viewed as a network of suppliers, manufacturing sites, distribution centers, and customer locations, through which components and products flow. A node in a network can be a physical location, a sub-network, or just an operation process, while links represent material (components or products) flow. These networks find significant applications in manufacturing and logistics in many industries, such as the electronic and automobile industries [10]. Throughout these networks, there are different sources of uncertainties, including supply (availability and quality), process (machine breakdown, operator variation), and demand (arrival time and volume). Moreover, these variations will propagate from upstream stages to downstream stages. These uncertainties degrade the performances of a network such as longer cycle time and lower fill-rates. Inventories at different stages of a network can be used to buffer the uncertainties, but they also have varying costs and different impacts on the end-item service level. Their effective allocation and control becomes a great challenge to the managers of logistics chains. Performance modeling and analysis become increasingly more important and difficult in the management of such complex manufacturing logistics chains. Inventory including raw materials, components and finished goods usually represents from 20–60% of the total assets of

Y. Wu · M. Dong (✉)
Department of Industrial Engineering and Management,
School of Mechanical Engineering,
Shanghai Jiao Tong University,
800 Dongchuan Road Min-Hang District,
Shanghai 200240, People's Republic of China
e-mail: mdong@sjtu.edu.cn

M. Dong
Leaders for Manufacturing Program,
Massachusetts Institute of Technology,
77 Massachusetts Avenue, Suite E40-315,
Cambridge, MA, 02139, USA

manufacturing firms [2]. Therefore, a good inventory management system has always been important in the workings of an effective manufacturing logistic chain.

Motivated by this challenge, many researchers have devoted much work to this issue. However, most of the literature is focused on systems with single products only and literature on multi-stage logistics chains with multi-products is limited. The assumption that every stage or node of the network produces a single class of product does not characterize the real world very well since nearly all firms produce more than one kind of product with limited service capacity. In this paper, a model is developed to characterize the dynamics of complex manufacturing logistics chains with multi-product and finite capacity. An analytical method is proposed to obtain performance measures of such models. Numerical results show that the proposed method works well. Simulation techniques may generally be used to analyze the performance of a system, but to identify an optimal configuration of a logistics chain, many different system variants have to be evaluated. Simulation-based evaluation is usually very time-consuming. Analytical evaluation methods are therefore needed that can determine the key performance measures quickly, even if these methods only approximate the true performance of the logistics chain.

In order to evaluate the performance of a serial manufacturing logistics chain, a parametric decomposition approach is adopted, which has been widely used to analyze multi-stage systems or networks. The basic idea is to approximately analyze the individual queues separately after approximately characterizing the arrival processes to each queue by a few parameters (usually two, one to represent the rate and another to represent the variability). The goal is to approximately represent the network dependence through these arrival-process parameters. Once the congestion in each queue has been described, the total network performance can be approximated by acting as if all the queues are mutually independent, i.e., the rest of the approximation is performed as if the steady-state distribution of the numbers of customers at the queues had a product form [18]. In the proposed approach, the whole chain is decomposed into multiple single-stage multi-class inventory queues (an inventory-queue is a queueing model that incorporates certain inventory replenishment policies such as base stock). The inputs (raw materials or components arrival processes) of each single-stage multi-class inventory queue are used to capture the characteristics of input flows of the original chain.

The rest of the paper is organized as follows. Section 2 provides a review of the relevant literature. In Sect. 3, the operations and the principal characteristics of the developed model are described. A decomposition method that divides the whole logistics chain into multiple single-stage queuing networks is proposed and the performance measures by analyzing the single-stage queueing network are obtained in Sect. 4. Section 5 presents some numerical results. Section 6 summarizes this research and gives some future research directions.

## 2 Literature review

Significant literature exists on inventory management in logistics chains. In the following, some prior studies devoted to the issues which are similar to the above described problems are reviewed. First, some important work on single-product multi-stage systems is reviewed. Lee and Zipkin [11, 12] and Duri et al. [9] used the decomposition method to analyze the tandem queues and processing networks. They transformed the production system into a multi-echelon model with limited production capacities. Azaron et al. [3] developed an open queueing network for multi-stage assemblies in which each service station represents a manufacturing or assembly operation. In the proposed model, not only the manufacturing and assembly processing times are considered as the functions of the arrival and service rates of the various stages of the manufacturing process, but also the role of transport times between the service stations in the manufacturing lead time is considered. An assumption is that the arrival processes of the individual parts of the product are independent Poisson processes with equal rates. In each service station, there is a server with exponential distribution of processing time. The transport times between the service stations are assumed to be independent random variables with exponential distributions. By applying the longest path analysis in queueing networks, the distribution function of time spent by a product in the system or the manufacturing lead time can be obtained. The study in this paper is more similar to Liu et al. [13]. They developed a multi-stage inventory queue model and a job-queue decomposition approach that evaluates the performance of serial manufacturing and supply chain systems with inventory control at every stage. In this paper, the proposed method decomposes a queue at each stage into two components, a backlog queue and a material queue. Instead of the single type product queues in their model [13], the queues contain a multi-class of items in the proposed model.

Next, some work on multi-product systems is reviewed. Thonemann and Bradley [16] presented a model for analyzing the effect of product variety on supply-chain performance for a supply chain with a single manufacturer and multiple retailers. This model might be useful in analyzing some subsystems of the whole supply network, but it did not consider other structures in the complex supply networks such as assembly. Bitran and Tirupati [4] studied multi-product queueing networks with deterministic routing using the decomposition method. Their method did not consider the impacts of stock buffers; therefore, it may be proper for the system operating under make-to-order policy but not make-to-stock policy. Ettl et al. [10] developed a network of inventory queue models to analyze complex supply chains. Each stocking location is modeled as an $M^X/G/\infty$ inventory queue operating under a base-stock control policy. The authors classified the customers according to different product types and different service requirements. But they assumed the capacity of each store to be infinite which can hardly exist in practice. Especially for multi-product systems, the assumption with infinite capacity will transform the problem into single product systems with infinite capacity. Dong [8] modified the inventory queue model developed in Ettl et al. [10], they used $GI^X/G/1$ queue to model each store. However, they still concentrated on a single product problem. A multi-product assemble-to-order (ATO) system was studied by Lu et al. [15]. In the system, components are built to stock with inventory controlled by base-stock policies, but the final products are assembled to order. They developed an approximation of the expected number of backorders that uses marginal distribution of the number of outstanding orders of components only. Efficient algorithms are developed to solve these problems, and numerical examples illustrate the effectiveness of the approximations. An important assumption in the model is that customer orders of each product follow a batch Poisson process. Zhao and Simchi-Levi [19] considered both multi-product base-stock ATO systems and multi-product batch-ordering ATO systems where the replenishment lead times of the components are stochastic, sequential, and independent of the system state. For a base-stock ATO system with multiple end products and demand following independent Poisson processes, they characterized the dependence among the stockout delays of the components. They show that a multi-product ATO system can be decomposed into multiple single-product subsystems with each subsystem corresponding to one product. For a batch-ordering ATO system, efficient numerical methods for performance evaluation based on Monte Carlo simulation are developed. The authors also discussed the limits of the proposed approach such as Poisson demand processes and finite component production capacity. Buzacott and Shanthikumar [5] developed multi-class models using $M/M/1$ and $M/G/1$ queues, and they provided some good analytical solutions to the models, but their solutions work for single-stage problems only. The purpose of this paper is threefold: (1) To provide an integrated modeling framework for manufacturing logistics chains in which the interdependencies between model components are captured; (2) To develop a network of inventory-queue models for performance analysis of an integrated logistics chain with inventory control at all sites; and (3) To extend the previous work developed for a supply network model with base-stock control and service requirements. Instead of a single type product produced at each stage with infinite capacity, the problem of multiple types of products produced at each stage with finite capacity is considered.

## 3 An integrated modeling framework for logistics chains

Logistics chains may differ in the network structure (serial, parallel, assembly and arborescent distribution), product structure (levels of bill-of-materials), transportation modes, and degree of uncertainty that they face. However, they have some basic elements in common [8].

### 3.1 Sites and stores

A logistics chain can be viewed as a network of functional sites connected by different material flow paths. Generally, there are four types of sites: (1) *Supplier sites* which procure raw materials from outside suppliers; (2) *Fabrication sites* which transform raw materials into components; (3) *Assembly sites* which assemble the components into semi-finished products or finished goods; and (4) *Distribution sites* which deliver the finished products to warehouses or customers. All sites in the network are capable of building parts, subassemblies or finished goods in either make-to-stock or make-to-order mode. The sites can be treated as the building blocks for modeling the whole logistics chain. Figure 1 shows a physical model of a logistics chain.

Typically, there are two types of operations performed at a site in a logistics chain: *material receiving* and *production*. A material receiving operation is one that receives input materials from upstream sites and stocks them as a stockpile to be used for production. A production operation is one in which fabrication or assembly activities occur, transforming or assembling input materials into output materials. Correspondingly, each site in the logistics chain has two kinds of stores: *input stores* and *output stores*. Each store stocks a single SKU. The input stores model the stocking of different types of components received from upstream sites, and output stores model the stocking of finished-products at the site (in Fig. 2, a site is represented by the dashed box containing input and output stores).

### 3.2 Links

All stores in the logistics chains are connected together by links that represent supply and demand processes. Two types of links are defined: *internal link* and *external link*. Internal links are used to connect the stores within a site, i.e., they

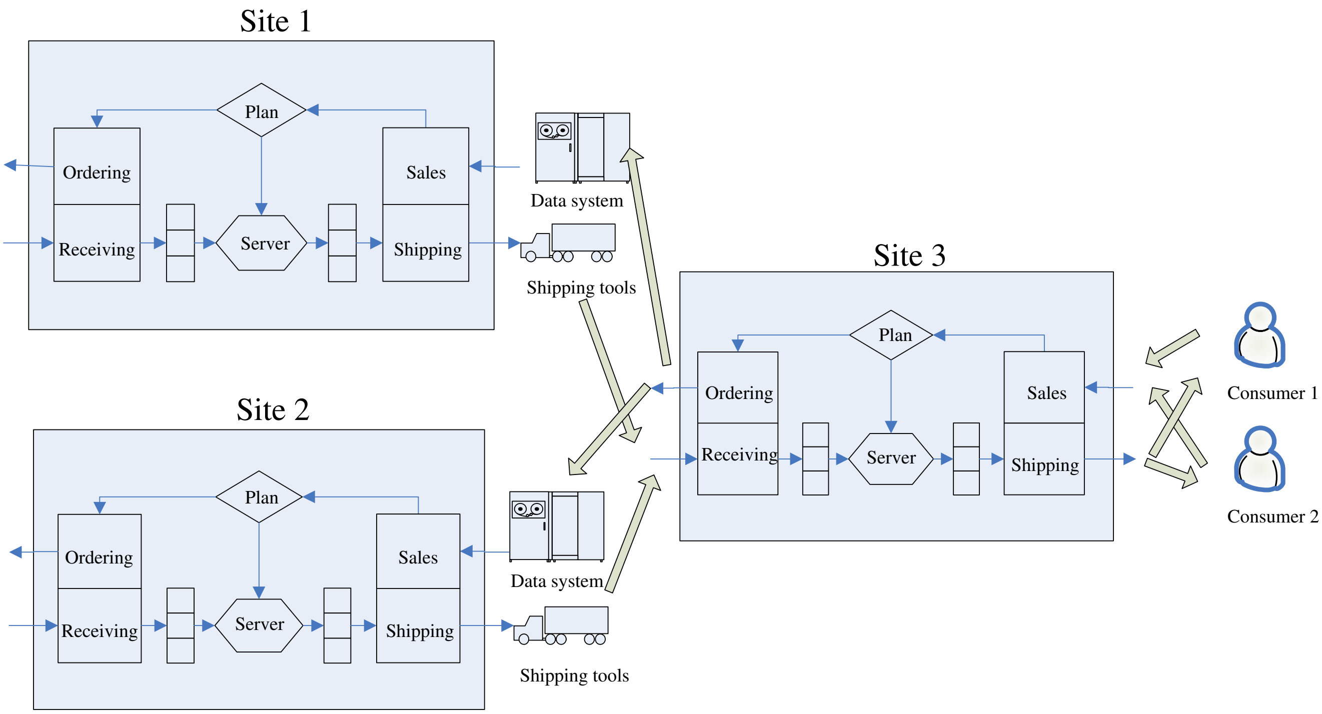


**Fig. 1** Physical model of a logistics chain

represent the material flow paths from input stores to output stores within a site. A link connecting an output store of one site to an input store of another site is called an external link. This kind of link represents that the output store provides replenishments to the specified downstream input store.

### 3.3 The relationships between stores

Let *ST* be the collection of stores in a logistics chain and *i* be a store in *ST*. The set of directly upstream supplying stores of store *i* is denoted as *UPST(i)*. The set of directly downstream receiving stores from store *i* is denoted as *DOWNST(i)*. If *i* is an input store, then *UPST(i)* is a singleton set, i.e., it contains only one upstream supplying store. That is, each input store can obtain replenishment from only one supplier. On the other hand, *DOWNST(i)* consists of one or more output stores at the same site. If *i* is an output store, then *UPST(i)* is either empty, in which case *i* is a *source* store (e.g., a supplier), or contains one or more stores, which are input stores at the same site. For *DOWNST(i)*, it is either empty, in which case *i* is an *end* store, or contains one or more input stores at its downstream site.

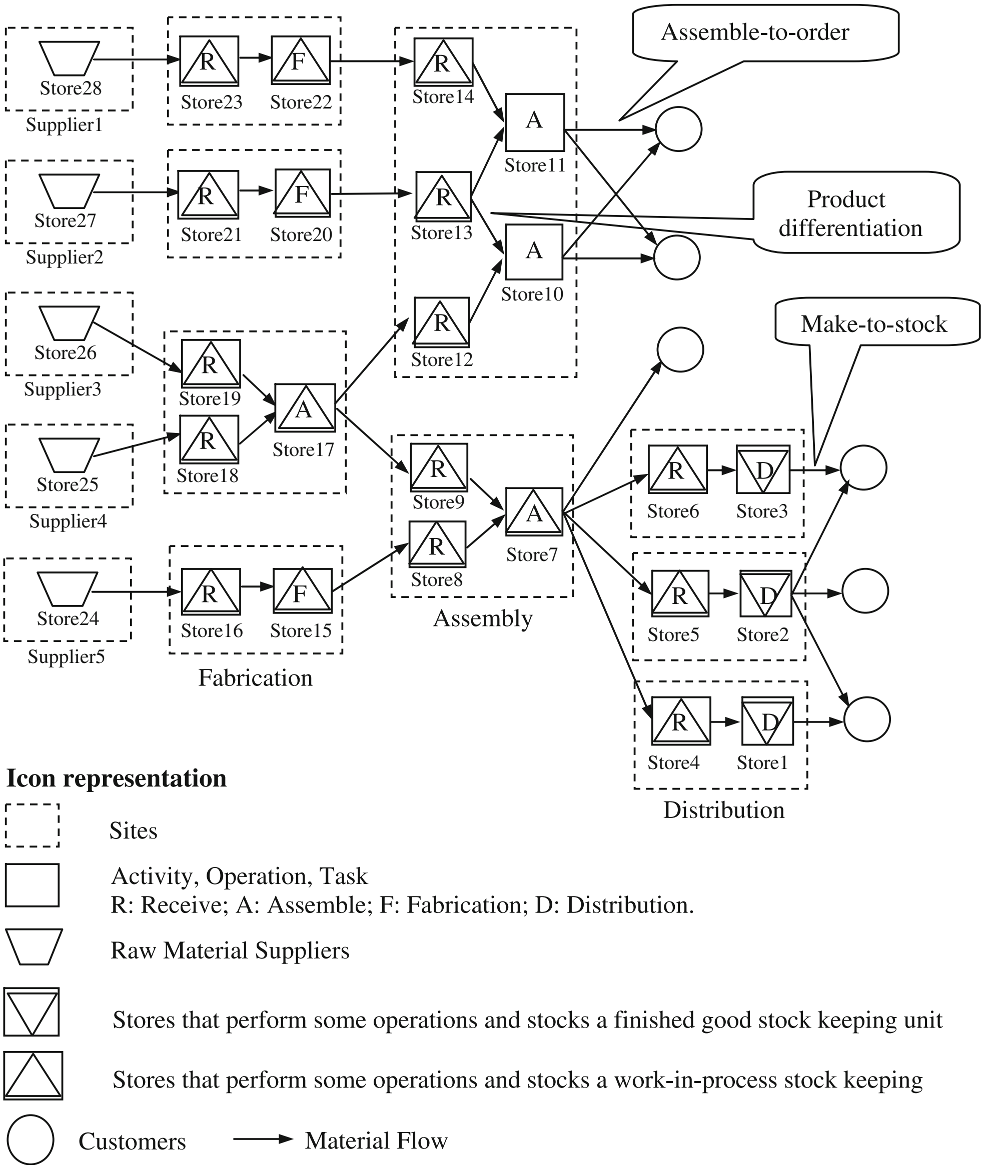
## 4 Multi-class inventory-queue model

A logistics chain with *n* sites (indexed as $i=1, ..., n$) in a complex network structure (serial, parallel and arborescent distribution) is considered. An example logistic chain ($n=5$) is given in Fig. 3. Each site *i* in the chain receives several kinds of incoming materials from upstream sites. With the incoming materials, site *i* produces different types of products. Each site is assumed to produce one type of product with just one kind of incoming material. This implies that assembly structure is not taken into consideration in this paper. Site *i* receives $r_i$ kinds of materials and produces a set of $r_i$ types of products to supply downstream sites or external customers. The set of direct upstream supplying sites of site *i* is denoted as *UPS (i)*. The set of direct downstream receiving sites from site *i* is denoted as *DOWNS (i)*. We further assume that one kind of material in site *i* comes from only one upstream supplying site and one type of product in site *i* goes to only one downstream site.

In the proposed model, each site *i* consists of two parts, a server with service rate $\mu_i$ and service time variation SCV (squared coefficient of variation) $C_{s,i}^2$ for all types of products, and an output buffer for semi-finished products or finished product. $I_{i,j}$ and $B_{i,j}$ are used to denote the on-hand inventory and backorder level of product *j* in site *i*, respectively. $N_i$ and $N_{i,j}$ are used to denote the job queue lengths of site *i* and product *j* in site *i* in steady-state, respectively.

The whole system operates in a make-to-stock mode using special policy similar to extended kanban control (EKC) proposed in Dallery and Liberopoulos [7]. Site *i* ($i=1, ...,n$) will maintain a safety stock of $S_{i,j}$ product *j* in its stock buffer and server *i* stops processing when there are

**Fig. 2** An integrated modeling framework for logistic networks



$S_{i,j}$ product $j$ in the buffer at site $i$; each product in the stock buffer has a tag attached to it. Whenever an order from downstream sites or external customers arrives, the product required will be delivered immediately if there is on-hand inventory of that type of product in the output buffer. In the meantime, the tag attached to the product will be transformed into a production authorization card and sent to the server for authorizing production in this site. Otherwise, it will be backordered, and another kind of tag will be sent to the server and join the job queue. So there are two types of tags in the output buffer, one is a production authorization card attached to semi-finished products or finished products while the other type is a backorder card that is not attached to any products within the output buffer. They are denoted $T^{p}_{i,j}$ and $T^{q}_{i,j}$, respectively. Thus, number of production authorization cards ($T^{p}_{i,j}$) in site $i$ for product $j$ is $S_{i,j}$ and we assume that there are sufficient backorder cards ($T^{q}_{i,j}$) in the output buffer. The above mechanism is illustrated in Fig. 4.

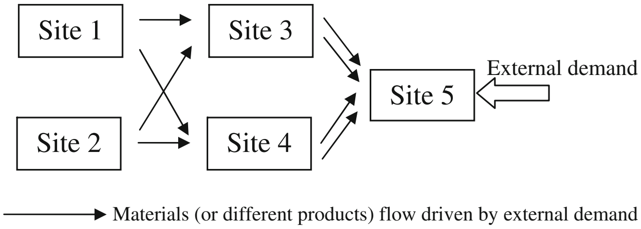

Site 1
Site 3
Site 5
External demand
Site 2
Site 4
Materials (or different products) flow driven by external demand

**Fig. 3** An example manufacturing logistics chain

There are two rules for the detailed operations of the proposed model.

First, after finishing processing in site $i$, product $P_{i,j}$ ($j$=1,..., $r_i$ ) with $T^{p}_{i,j}$ arrives at the output buffer, and will be released to the downstream site immediately if there are backorder cards ($T^{q}_{i,j}$) in the job queue. In this case,

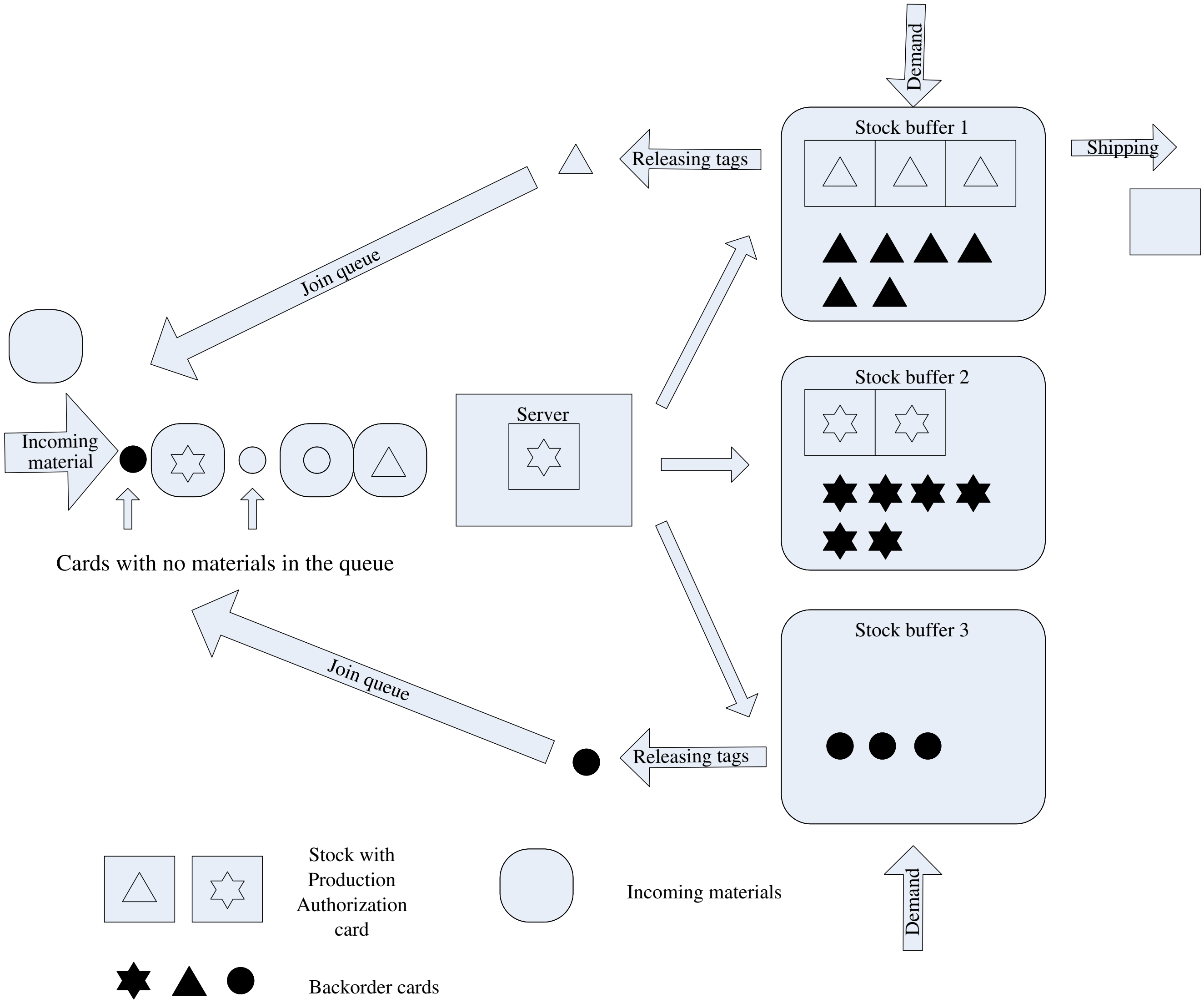


**Fig. 4** Illustration for a multi-class inventory-queue

production authorization card ($T^p_{i,j}$) will be unattached from the product and transmitted to the job queue to replace the first backorder card (if there are more than one backorder card in the queue). The backorder card, which is replaced, will join the output buffer and wait for downstream or external customer demand. Second, a product with a backorder card will be released to the downstream sites or external customers as soon as it arrives at the output buffer. The backorder card attached to the product will be removed and stay at the output buffer.

When an order from external customers arrives at the end site, a job is added to the job queue at the server of that site. At the same time, all related upstream supplying sites will have a job added to their job queues at the servers. Thus, the demand information becomes known or shared to all the related sites simultaneously. As shown in Fig. 5, when an external demand for a certain type of finished product arrives, an order will be added to sites 2, 3 and 5, respectively. Material flows through this path and will finally be turned into supply to the end site. The time and cost spent on setup and changeover are assumed to be insignificant and hence ignored. In this case, servers become busy when there are jobs in the job queues in which the FCFS (first come first serve) protocol is adopted.

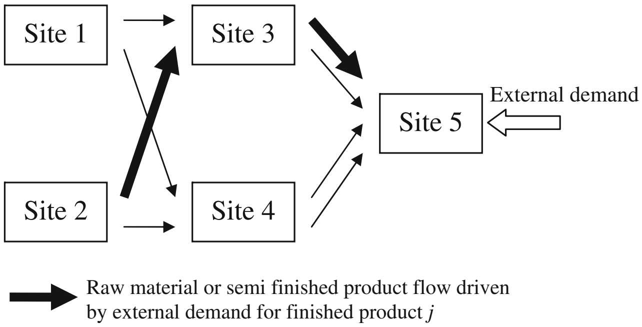

Site 1
Site 3
Site 5
External demand
Site 2
Site 4
Raw material or semi finished product flow driven by external demand for finished product j

**Fig. 5** Illustration of material and semi-finished product flows

From the above description, it is can be seen that the queue at the server of site $i$ contains multiple product types produced in site $i$, so we have

$$N_i = \sum_{j=1}^{r_i} N_{i,j} \tag{1}$$

Whenever an order arrives, it takes away a corresponding inventory item and a job joins the queue. So the on-hand inventory and backorder level have the following relationships:

$$\mathrm{I}_{i,j} = \left[S_{i,j} - N_{i,j}\right]^{+} \tag{2}$$

$$B_{i,j} = \left[N_{i,j} - S_{i,j}\right]^{+} \tag{3}$$

The realized fill rate of product $j$ (i.e., the fraction that on-hand inventory is greater than 0) in site $i$ is:

$$f_{i,j} = P\left\{N_{i,j} < S_{i,j}\right\} \tag{4}$$

## 5 Performance analysis

As mentioned in the literature review, exact computation of such models described above is very difficult. The main idea of the proposed approach is to divide the whole network into multiple single sites, aggregate the multi-class input streams at one site into a mixture stream. A decomposition method similar to the one developed by Liu et al. [13] is used to obtain the approximation of $N_i$ for all $i=1, ..., n$. Once the values of $N_{i,j}$ for all $i$ ($i=1, ..., n$) and $j$ ($j=1, ..., r_i$) become available, the major network performance measures such as $I_{i,j}$, $B_{i,j}$ and $f_{i,j}$ could be derived from $N_{i,j}$ (see Eqs. 2–4). Then, the departure process of each product in each site is analyzed. Suppose that the material transfer time between sites is deterministic, the arrival process of a downstream site is the summation of the departure processes of the corresponding upstream sites. Knowing the departure processes, we can link multiple single sites together and find out analytical solutions for every site (here, we assume that there is sufficient raw material supply at the source sites such as site 1 and site 2 in Fig. 3).

### 5.1 Aggregation of multi-class input streams and separation of output streams

With the assumption that sufficient materials are available at the source sites, the input processes of these sites can be viewed as the arrival processes of external demands for the corresponding finished products. However, the input processes of the non-source sites could not be viewed the same way because their input processes are related with not only the external customer demands but also the material supply processes from their upstream sites. In the following, first, an aggregation method used to approximate the mixture input stream with the multi-class input streams at every site is described. Then, departure process is analyzed by another approach called separation.

For an arbitrary site $i$ of the system, suppose site $i$ produces $r_i$ types of products. Therefore, there are $r_i$ input streams from upstream sites. Each input stream is approximated with mean arrival rate $\lambda_{i,j}$ (mean inter-arrival interval is $1/\lambda_{i,j}$) and SCV of inter-arrival time $C^2_{a,i,j}$ ($j$=1, ..., $r_i$). The service rate for all types of product at site $i$ is $\mu_i$ and SCV of service time is $C^2_{s,\,i}$.

With the idea of aggregation and separation in mind, we analyze the departure processes in the following (Fig. 6 illustrates the main idea of this method). $\lambda_i$ and $C^2_{a,i}$ are

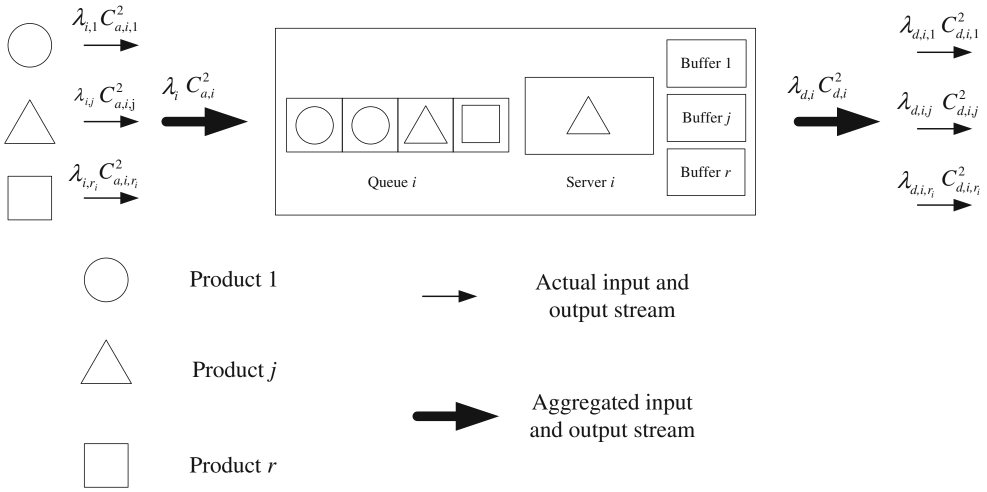

$\lambda_{i,1}\, C^2_{a,i,1}$
$\lambda_{i,j}\, C^2_{a,i,j}$
$\lambda_{i,r_i}\, C^2_{a,i,r_i}$
$\lambda_i\, C^2_{a,i}$
Buffer 1
Buffer $j$
Buffer $r$
Queue $i$
Server $i$
$\lambda_{d,i}\, C^2_{d,i}$
$\lambda_{d,i,1}\, C^2_{d,i,1}$
$\lambda_{d,i,j}\, C^2_{d,i,j}$
$\lambda_{d,i,r_i}\, C^2_{d,i,r_i}$
Product 1
Product $j$
Product $r$
Actual input and output stream
Aggregated input and output stream

**Fig. 6** Illustration of aggregation and separation

mean arrival rate and SCV of the aggregated input streams, respectively.

As the arrival processes of external demands for various products are assumed to be mutually independent, the input processes at all sites are independent. Then we have:

$$\lambda_i = \sum_{j=1}^{r_i} \lambda_{i,j} \tag{5}$$

$$C_{1,i}^2 = \frac{1}{\lambda_i}\sum_{j=1}^{r_i} \lambda_{i,j} C_{a,i,j}^2 \tag{6}$$

$$C_{2,i}^2 = I^2 \lambda_i^2 - 1 \tag{7}$$

$$C_{a,i}^2 = \omega C_{1,i}^2 + (1-\omega) C_{2,i}^2 \tag{8}$$

where $C_{1,i}^2$ and $C_{2,i}^2$ are SCV of inter-arrival time obtained through the asymptotic method and the stationary-interval method, respectively. $I$ is the inter-arrival time and $\omega$ is a function of $\rho$ (traffic density at a site) and $n^*$ (effective input stream number at a site) [1, 17].

The aggregated input streams can be characterized by their first two moments $\lambda_i$ and $C_{a,i}^2$, which are derived from the mean arrival rate and SCV of actual multiple input streams. In the following, we show how to derive the approximations of the first two moments of different departure streams with known $\lambda_{d,i}$ and $C_{d,i}^2$.

It can be seen that the mean departure rate is equal to the mean arrival rate. Then we have

$$\lambda_{d,i} = \lambda_i \tag{9}$$

$$\lambda_{d,i,j} = \lambda_{i,j} \quad j = 1, \ldots, r_i \tag{10}$$

$$C_{d,i,j}^2 = p_{i,j} C_{d,i}^2 + 1 - p_{i,j} \quad j = 1, \ ..., r_i \tag{11}$$

where $p_{i,j} = \lambda_{i,j}/\lambda_i$.

Equations 9 and 10 generate exact solutions. However, Eq. 11 does not fully incorporate the structure of the departure process which is strongly related to the arrival process. New approximation equations reflecting the impacts of arrival processes are given as follows:

$$C_{d,i,j}^2 = p_i C_{d,i}^2 + C_{W_{i,j}}^2 \quad j = 1, \ ..., r_i. \tag{12}$$

where $W_{i,\bar{j}} = Z_{i,\bar{j}} + 1$, and $Z_{i,\bar{j}}$ represents the number of products different from $j$ that depart within the interval of the time epoch in which two consecutive products $j$ depart.

Caldentey [6] analyzed four cases that are frequently used in the literature. The approximations of the SCV are provided in the following:

Case 1. Poisson arrival:

$$C_{d,i,j}^2 = p_{i,j} C_{d,\,i}^2 + (1-p_{i,j})\left[(1-p_{i,j}) C_{a,i,j}^2 + p_{i,j}\right] \quad j = 1, \ldots, r_i \tag{13}$$

Since SCV of inter-arrival time of the Poisson arrival process equals 1 ($C_{a,i,j}^2 = 1$), Eq. 13 is equivalent to Eq. 11.

Case 2. Asymptotic result:

$$C_{W_{i,j}}^2 = (1-p_{i,j})^2 C_{a,i}^2 + (1-p_{i,j}) p_{i,j} C_{i,\bar{j}}^2 \quad j = 1, \ldots, r_i \tag{14}$$

where $\bar{j}$ represents the aggregation of all products different from product $j$ at site $i$. This case is suitable in the situation in which each of the input streams has a small intensity when compared to the aggregated stream.

Based on the decomposition method, product streams are usually characterized by their first two moments. Depending on the value of the squared coefficient of variation, a sum or a mixture of two exponential random variables can be used to fit the first two moments [17]. The sum of two exponentials is used when the SCV is lower than 1, and the mixture of two exponentials is used when the SCV is greater than 1.

Case 3. Sum of two exponentials:

$$C_{W_{i,j}}^2 = (1-p_{i,j})^2 C_{a,i,j}^2 + p_{i,j}\sum_{k\neq j} p_{i,j} C_{a,i,k}^2 + \left(\frac{p_{i,j}^2}{2}\right)\sum_{k\neq j}\left(1 - C_{a,i,k}^2\right)^2\left[1 - \Phi_{i,j}\left(\frac{2\lambda_{i,k}}{1 - C_{a,i,k}^2}\right)\right] \quad j = 1, \ ..., r_i \tag{15}$$

where $\lambda_{i,j} = \frac{\mu_{i,j,1}\mu_{i,j,2}}{\mu_{i,j,1}+\mu_{i,j,2}}$, $\mu_{i,j,1} + \mu_{i,j,2} = \frac{2\lambda_{i,j}}{1-C_{a,i,j}^2}$ and $\Phi_{i,j}(s) = \left(\frac{\mu_{i,j,1}}{\mu_{i,j,1}+s}\right)\left(\frac{\mu_{i,j,2}}{\mu_{i,j,2}+s}\right)$.

Case 4. Mixture of two exponentials:

$$C_{W_{i,j}}^2 = (1-p_{i,j})^2 C_{a,i,j}^2 + p_{i,j}\sum_{k\neq j} p_{i,j} C_{a,i,k}^2 + 2p_{i,j}^2 \sum_{k\neq j}\left(C_{a,i,k}^2 - 1\right)\left[1 - \Phi_{i,j}\left(\frac{2\lambda_{i,k}}{1 + C_{a,i,k}^2}\right)\right] \quad j = 1, \ ..., r_i \tag{16}$$

where $\Phi_{i,j}(s) = \frac{\theta_{i,j,1}\mu_{i,j,1}}{\mu_{i,j,1}+s} + \frac{\theta_{i,j,2}\mu_{i,j,2}}{\mu_{i,j,2}+s}, \theta_{i,j,m} = \frac{\sqrt{C_{a,i,j}^2+1} \pm \sqrt{C_{a,i,j}^2-1}}{2\sqrt{C_{a,i,j}^2+1}}$ and $\mu_{i,j,m} = 2\lambda_{i,j}\theta_{i,j,m}$, $m$=1,2.

### 5.2 Job queue analysis

In the previous section, we showed how to derive individual departure processes with arrival processes and aggregated departure processes. However, the aggregated departure process is still unknown. The first two moments of inter-departure time are used to characterize the aggregated process. It can be seen that the mean departure time of a product is equal to the mean arrival time of the demand for it. Thus, deriving the second moment of the inter-departure time is the main task remaining. Furthermore, the job queue could be decomposed into two parts according to the proposed model.

Since the demand arrives in all related sites simultaneously (i.e., information is shared among these sites), an upstream site would deliver the materials or semi-finished products to the corresponding downstream site whenever the upstream site's job queue is increased by one. If there is sufficient supply at all sites, then each tag in the job queue should be attached to some raw materials or semi-finished products. However, this assumption could hardly exist. Hence, the job queue $N_{i,j}$ (job queue of product $j$ in site $i$) consists of two parts: a material queue $M_{i,j}$ of site $i$ and an empty card queue $E_{i,j}$. Thus, we have:

$$N_{i,j} = M_{i,j} + E_{i,j} \tag{17}$$

According to Eq. 1, we can obtain the following results:

$$M_i = \sum_{j=1}^{r_i} M_{i,j} \tag{18}$$

$$E_i = \sum_{j=1}^{r_i} E_{i,j} \tag{19}$$

$$N_i = M_i + E_i \tag{20}$$

$M_{i,j}$ and $E_{i,j}$ are assumed to be independent (though they are not). From the numerical results of Liu [14], this assumption is quite acceptable. It can be seen that $M_i$ and $E_i$ are also independent.

Because site $i$ demands materials for one type of product from just one upstream site, $E_{i,j}$ is equal to the backorder level of the material at its supply site. For example, from Fig. 3, we can see that $E_{3,2} = B_{2,1} = \left[N_{2,1} - S_{2,1}\right]^{+}$.

Next, we focus on the approximation of $C_{d,i}^2$ and $M_{i,j}$. Motivated by the approximation of SCV of inter-departure time by Buzacott and Shanthikumar [5] and Liu et al. [13], the following equation is used to approximate the SCV of aggregated inter-departure time:

$$C_{d,i}^2 = \left(1 - \rho_i^{2+S_i/2}\right) C_{a,i}^2 + \rho_i^{2+S_i/2} C_{s,i}^2 \quad i = 1, \ldots, n \tag{21}$$

where $S_i = \sum\limits_{j=1}^{r_i} S_{i,j}$, $\rho_i = \lambda_i / \mu_i$.

The queue length distribution of $M_{i,j}$ according to $M_i$ can be derived by the following relationship: $M_{i,j} = \frac{\lambda_{i,j}}{\lambda_i} M_i$; and

$$P\{M_i = l\} = \begin{cases} 1 - \rho_i, & l = 0 \\ \rho_i (1 - \widehat{\rho}_i) \widehat{\rho}_i^{\,l-1}, & l = 1 \end{cases} \tag{22}$$

where $\widehat{\rho}_i = \frac{\rho_i \left(C_{a,i}^2 + C_{s,i}^2\right)}{\rho_i \left(C_{a,i}^2 + C_{s,i}^2\right) + 2(1 - \rho_i)}$.

According to Eq. 22, we have:

$$E(M_i) = \frac{\rho_i}{1 - \widehat{\rho}_i}. \tag{23}$$

As we assume that there are sufficient materials at the source sites, $N_i$=$M_i$ ($i$ is a source site). With all the preparations above, the steady-state distribution of $N_i$ and the departure processes with known arrival processes could be derived. Thus, the performance measures of the complete logistics chain can be obtained as follows.

For any source site $i$, since there is sufficient material supply, we have $E_i$=0. According to Eqs. 17, 20 and $M_{i,j} = \frac{\lambda_{i,j}}{\lambda_i} M_i$, we can obtain $N_{i,j}$=$M_{i,j}$, $N_i$=$M_i$ and $M_{i,j} = \frac{\lambda_{i,j}}{\lambda_i} M_i$. Thus, $E\left(N_{i,j}\right) = \frac{\lambda_{i,j}}{\lambda_i} E(N_i) = \frac{\lambda_{i,j} \rho_i}{\lambda_i \left(1 - \widehat{\rho}_i\right)}$ and

$$P\left\{N_{i,j} = l\right\} = \begin{cases} 1 - \rho_{i,j}, & l = 0 \\ \rho_{i,j} \left(1 - \widehat{\rho}_{i,j}\right) \widehat{\rho}_{i,j}^{\,l-1}, & l \geq 1 \end{cases} \tag{24}$$

where $\rho_{i,j} = \frac{\lambda_{i,j}}{\lambda_i} \rho_i$ and $\widehat{\rho}_{i,j} = \frac{E\left(N_{i,j}\right) - \rho_{i,j}}{E\left(N_{i,j}\right)}$.

From Eqs. 2 and 24, we have:

$$P\left\{I_{i,j} = l\right\} = \begin{cases} \rho_{i,j} \widehat{\rho}_{i,j}^{\,S_{i,j}-1} & , l = 0 \\ \rho_{i,j} \left(1 - \rho_{i,j}\right) \widehat{\rho}_{i,j}^{\,S_{i,j}-l-1}, & l = 1, \ldots, S_{i,j} - 1 \\ 1 - \rho_{i,j}, & l = S_{i,j} \end{cases} \tag{25}$$

From Eqs. 3 and 24, we can obtain:

$$p\left\{B_{i,j} = l\right\} = \begin{cases} 1 - \rho_{i,j} \widehat{\rho}_{i,j}^{\,S_{i,j}}, & l = 0 \\ \rho_{i,j} \left(1 - \widehat{\rho}_{i,j}\right) \widehat{\rho}_{i,j}^{\,l-1+S_{i,j}}, & l \geq 1 \end{cases} \tag{26}$$

From Eqs. 4 and 24, we have:

$$f_{i,j} = P\left\{N_{i,j} < S_{i,j}\right\} = 1 - \rho_{i,j} \widehat{\rho}_{i,j}^{\,S_{i,j}-1} \tag{27}$$

Since $\widehat{\rho}_{i,j}$ is less than 1, $f_{i,j}$ increases with $S_{i,j}$, which means the fill rate increases when the stock level is set higher.

**Table 1** Numerical results of performance analysis

| $C^2_{s,1}, C^2_{s,2}, C^2_{s,3}, C^2_{s,4}, C^2_{s,5}$,<br>$\widehat{S}_1, \widehat{S}_2, \widehat{S}_3, \widehat{S}_4, \widehat{S}_5$ | $\rho_1$, $\rho_2$, $\rho_3$, $\rho_4$, $\rho_5$ | Methods | E($N_1$) | E($N_2$) | E($N_3$) | E($N_4$) | E($N_5$) |
|---|---|---|---|---|---|---|---|
| 1,1,1,1,1, 5,5,5,5,5 | 0.6,0.6,0.6,0.6,0.6 | Approx. | 1.500 | 1.500 | 1.500 | 1.500 | 1.500 |
| | | Simulation | 1.509 | 1.513 | 1.497 | 1.502 | 1.514 |
| | | Error (%) | −0.596 | −0.859 | 0.200 | −0.133 | 0.925 |
| | 0.9,0.9,0.9,0.9,0.9 | Approx. | 9.000 | 9.000 | 9.000 | 9.000 | 9.000 |
| | | Simulation | 9.206 | 8.632 | 9.220 | 8.552 | 8.544 |
| | | Error (%) | −2.238 | 4.263 | −2.386 | 5.238 | 5.337 |
| | 0.6,0.6,0.8,0.8,0.9 | Approx. | 1.500 | 1.500 | 4.000 | 4.000 | 9.000 |
| | | Simulation | 1.510 | 1.485 | 4.040 | 3.987 | 8.909 |
| | | Error (%) | −0.662 | 1.010 | 0.990 | 0.326 | 1.021 |
| | 0.8,0.8,0.9,0.9,0.6 | Approx. | 4.000 | 4.000 | 9.000 | 9.000 | 1.500 |
| | | Simulation | 4.012 | 4.011 | 8.916 | 8.797 | 1.468 |
| | | Error (%) | −0.299 | −0.274 | 0.942 | 2.308 | 2.180 |
| | 0.8,0.8,0.6,0.6,0.9 | Approx. | 4.000 | 4.000 | 1.500 | 1.500 | 9.000 |
| | | Simulation | 4.047 | 4.038 | 1.472 | 1.456 | 9.033 |
| | | Error (%) | −1.161 | −0.941 | 1.902 | 3.022 | −0.365 |
| 0.25,0.25,1,1,4, 5,5,8,8,10 | 0.6,0.6,0.6,0.6,0.6 | Approx. | 1.163 | 1.163 | 1.492 | 1.492 | 2.893 |
| | | Simulation | 1.165 | 1.164 | 1.521 | 1.505 | 2.832 |
| | | Error (%) | −0.172 | −0.086 | −1.907 | −0.864 | 2.154 |
| | 0.9,0.9,0.9,0.9,0.9 | Approx. | 5.962 | 5.962 | 8.421 | 8.421 | 21.132 |
| | | Simulation | 6.133 | 5.980 | 8.227 | 8.336 | 20.751 |
| | | Error (%) | −2.799 | −0.301 | 2.358 | 1.020 | 1.983 |
| | 0.6,0.6,0.8,0.8,0.9 | Approx. | 1.163 | 1.163 | 3.978 | 3.978 | 22.016 |
| | | Simulation | 1.167 | 1.165 | 4.069 | 3.995 | 21.575 |
| | | Error (%) | −0.343 | −0.172 | −2.236 | −0.426 | 2.044 |
| | 0.8,0.8,0.9,0.9,0.6 | Approx. | 2.800 | 2.800 | 8.661 | 8.661 | 2.721 |
| | | Simulation | 2.814 | 2.838 | 8.625 | 8.302 | 2.829 |
| | | Error (%) | −0.498 | −1.339 | 0.417 | 4.324 | −3.818 |
| | 0.8,0.8,0.6,0.6,0.9 | Approx. | 2.800 | 2.800 | 1.446 | 1.446 | 20.764 |
| | | Simulation | 2.802 | 2.864 | 1.473 | 1.455 | 21.331 |
| | | Error (%) | −0.071 | −2.235 | −1.833 | −0.618 | −2.658 |
| 1,1,4,4,0.25,8,8,5,5,10 | 0.6,0.6,0.6,0.6,0.6 | Approx. | 1.500 | 1.500 | 2.850 | 2.850 | 2.939 |
| | | Simulation | 1.507 | 1.494 | 2.893 | 2.892 | 2.891 |
| | | Error (%) | −0.464 | 0.402 | −1.486 | −1.452 | 1.660 |
| | 0.9,0.9,0.9,0.9,0.9 | Approx. | 9.000 | 9.000 | 21.150 | 21.150 | 13.076 |
| | | Simulation | 8.857 | 9.132 | 21.495 | 22.342 | 12.485 |
| | | Error (%) | 1.614 | −1.445 | −1.605 | −5.335 | 4.734 |
| | 0.6,0.6,0.8,0.8,0.9 | Approx. | 1.500 | 1.500 | 8.800 | 8.800 | 8.226 |
| | | Simulation | 1.479 | 1.520 | 9.099 | 8.468 | 8.445 |
| | | Error (%) | 1.420 | −1.316 | −3.286 | 3.921 | −2.593 |
| | 0.8,0.8,0.9,0.9,0.6 | Approx. | 4.000 | 4.000 | 21.150 | 21.150 | 2.077 |
| | | Simulation | 3.993 | 3.988 | 21.232 | 21.045 | 2.049 |
| | | Error (%) | 0.175 | 0.301 | −0.386 | 0.499 | 1.366 |
| | 0.8,0.8,0.6,0.6,0.9 | Approx. | 4.000 | 4.000 | 2.850 | 2.850 | 6.041 |
| | | Simulation | 4.064 | 4.006 | 2.814 | 2.912 | 6.205 |
| | | Error (%) | −1.575 | −0.150 | 1.279 | −2.129 | −2.643 |
| 4,4,0.25,0.25,1, 5,5,5,5,10 | 0.6,0.6,0.6,0.6,0.6 | Approx. | 2.850 | 2.850 | 1.176 | 1.176 | 1.569 |
| | | Simulation | 2.892 | 2.817 | 1.209 | 1.202 | 1.503 |
| | | Error (%) | −1.452 | 1.171 | −2.730 | −2.163 | 4.391 |
| | 0.9,0.9,0.9,0.9,0.9, | Approx. | 21.150 | 21.150 | 8.764 | 8.764 | 8.762 |
| | | Simulation | 24.452 | 21.840 | 8.950 | 8.640 | 9.032 |
| | | Error (%) | −13.504 | −3.159 | −2.078 | 1.435 | −2.989 |
| | 0.6,0.6,0.8,0.8,0.9 | Approx. | 2.850 | 2.850 | 2.858 | 2.858 | 8.577 |
| | | Simulation | 2.892 | 2.945 | 2.953 | 2.849 | 8.811 |
| | | Error (%) | −1.452 | 3.226 | −3.217 | 0.316 | −2.656 |

**Table 1** (continued)

| $C_{s,1}^2, C_{s,2}^2, C_{s,3}^2, C_{s,4}^2, C_{s,5}^2,$ $\widehat{S}_1, \widehat{S}_2, \widehat{S}_3, \widehat{S}_4, \widehat{S}_5$ | $\rho_1$, $\rho_2$, $\rho_3$, $\rho_4$, $\rho_5$ | Methods | E($N_1$) | E($N_2$) | E($N_3$) | E($N_4$) | E($N_5$) |
|---|---|---|---|---|---|---|---|
| 4,4,0.25,0.25,1, 8,8,10,10,10 | 0.8,0.8,0.9,0.9,0.6 | Approx. | 8.800 | 8.800 | 7.189 | 7.189 | 1.399 |
| | | Simulation | 8.697 | 9.073 | 7.495 | 6.777 | 1.348 |
| | | Error (%) | 1.184 | −3.009 | −4.083 | 6.079 | 3.783 |
| | 0.8,0.8,0.6,0.6,0.9 | Approx. | 8.800 | 8.800 | 1.260 | 1.260 | 9.767 |
| | | Simulation | 8.755 | 9.026 | 1.421 | 1.391 | 9.372 |
| | | Error (%) | 0.514 | −2.504 | −11.330 | −9.418 | 4.214 |
| | 0.8,0.6,0.6,0.8,0.9 | Approx. | 8.800 | 2.850 | 1.188 | 2.918 | 8.819 |
| | | Simulation | 8.904 | 2.859 | 1.248 | 2.907 | 9.004 |
| | | Error (%) | −1.168 | −0.315 | −4.808 | 0.378 | −2.055 |
| | 0.8,0.9,0.9,0.8,0.6 | Approx. | 8.800 | 21.150 | 7.296 | 3.272 | 1.539 |
| | | Simulation | 9.027 | 21.803 | 7.425 | 3.485 | 1.485 |
| | | Error (%) | −2.515 | −2.995 | −1.737 | −6.112 | 3.636 |

For any non-source site $i$, the empty card queue length for certain product $j$ equals the backorder level of the material at its supply site. Because the proposed algorithm starts at the source sites and computes the distributions of different measures stage by stage, the backorder level of the material at the supply site is known, which means the steady-state distribution of $E_{i,j}$ is available. For example, $E_{3,2} = B_{2,1} = \left[N_{2,1} - S_{2,1}\right]^+$ can be seen from Fig. 5. $N_{i,j}$ can be calculated from Eq. 17, so $B_{2,1}$ becomes known since it is the backorder level of product 1 at source site 2. This way, the performance measures such as on-hand inventory level and backorder level at non-source site $i$ can be obtained from the distribution of $N_{i,j}$.

Through this procedure, the whole network performance can be computed from the source sites to the end sites.

## 6 Application example and numerical results

In this section, the approximations derived from the proposed method are compared with the results obtained through simulation study. The example logistics chain is extracted and simplified from a real manufacturing logistics network (see Fig. 3). All the orders arrive with Poisson distribution. The overall accuracy of the proposed approximations is illustrated in terms of $N_i$, because all the other parameters such as inventory level $I_i$ and backorder level $B_i$ can be obtained through $N_i$.

The results including analytic approximation, simulation outputs and the relative errors between them are presented in Table 1. For simplicity, suppose that $\widehat{S}_i = S_{i,1} = S_{i,2} = \cdots S_{i,r_i}$ in this example, so we have $S_i = r_i\widehat{S}_i$.

From Table 1, it can be seen that the maximal error is 13.504% and the average error is 2.157%. This shows that the proposed analytical approximation approach is effective.

## 7 Conclusions

This paper proposes a serial manufacturing logistics chain model with multi-class inventory queue as the basic building block. Each site in the chain is modeled by a single server multi-class queue controlled by a pulling policy which is similar to extended kanban control policy. A job queue decomposition method is used to analyze the approximations of key performance measures. An aggregation and separation method is proposed to capture the main characteristics and then link all the sites together. Multi-class input stream is characterized by aggregation, and the output stream is characterized by separation. The numerical results of an application example show that the proposed model works well.

There are two main approaches in the literature that handle very similar problems to this paper. In order to provide more acceptable estimates of key parameters, one way is to include the interference among products in the decomposition method. With this modification, this approach is able to achieve significant improvement of performance (e.g., Bitran and Tirupati [4]). However, inventory is not considered and this makes the model appropriate for make-to-order environment only. While taking inventory into consideration, as in this paper, the departure processes become more complicated since the impact of interference among products and inventory on the departure processes has to be considered simultaneously. The second existing method treats the site in the networks quite differently from what this paper does. Stores instead of sites are the basic building blocks and each store is modeled with an $M^X/G/\infty$ inventory queue operating under a base-stock control policy (e.g., Ettl et al. [10]). This assumption ignores the capacity constraints which affect the departure processes of different kinds of products; however, the situation of unlimited capacity is not well common. This paper extends the work developed for the supply network model with base-stock control policy. Instead of a single type of product produced at each stage with infinite capacity, this paper studied the problem of multiple types of products produced at each stage with finite capacity consideration.

However, some important issues in manufacturing and supply chain systems (e.g., assembly structure) are ignored in this paper. A common general service time distribution for all kinds of products is also not a very reasonable assumption, it may be the case sometimes but it will not hold in other practices. The job queue length distribution of a specific product is obtained from the joint job queue length distribution through an approximation method. All these factors will be considered in future research. Once all the related performance measures are obtained, the optimization of inventory positions and allocations, which minimizes the total inventory cost, might be possible in the future study.

**Acknowledgements** The work presented in this paper has been supported by grants from the National Natural Science Foundation of China (70571050) and the Shanghai Pujiang Program (No. 05PJ14067).

## References

1. Albin SL (1984) Approximating a point process by a renewal process, II: superposition arrival processes to queues. Oper Res 32 (5):1133–1162
2. Arnold JRT (1998) Introduction to materials management. Prentice Hall, USA
3. Azaron A, Katagiri H, Kato K, Sakawa M (2006) Modelling complex assemblies as a queueing network for lead time control. Eur J Oper Res 174:150–168
4. Bitran GR, Tirupati D (1988) Multi-product queueing networks with deterministic routing: decomposition approach and the notion of interference. Manage Sci 34(1):75–100
5. Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice Hall, Englewood Cliffs, NJ
6. Caldentey R (2001) Approximations for multi-class departure processes. Queueing Syst 38:205–212
7. Dallery Y, Liberopoulos G (2000) Extended kanban control system: combining kanban and base stock. IIE Trans 32:369–386
8. Dong M (2003) Inventory planning of supply chains by linking production authorization strategy to queueing models. Prod Plan Control 14(6):533–541
9. Duri C, Frein Y, DiMascolo M (2000) Performance evaluation and design of base-stock systems. Eur J Oper Res 127:172–188
10. Ettl M, Feigin GE, Lin GY, Yao DD (2000) A supply network model with base-stock control and service requirements. Oper Res 48:216–232
11. Lee YJ, Zipkin PH (1992) Tandem queues with planned inventories. Oper Res 40:936–947
12. Lee YJ, Zipkin PH (1995) Processing networks with inventories: Sequential refinement systems. Oper Res 43(6):1025–1036
13. Liu LM, Liu XM, Yao DD (2004) Analysis and optimization of a multistage inventory-queue system. Manage Sci 50(3):365–380
14. Liu XM (1999) Performance analysis and optimization of supply networks. PhD Dissertation, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, People's Republic of China
15. Lu YD, Song JS, Yao DD (2005) Backorder minimization in multiproduct assemble-to-order systems. IIE Trans 37:763–774
16. Thonemann UW, Bradley JR (2002) The effect of product variety on supply-chain performance. Eur J Oper Res 143:548–569
17. Whitt W (1982) Approximating a point process by a renewal process: Two basic methods. Oper Res 30(1):125–147
18. Whitt W (1995) Variability functions for parametric-decomposition approximations of queueing networks. Manage Sci 41 (10):1704–1715
19. Zhao Y, Simchi-Levi D (2006) Performance analysis and evaluation of assemble-to-order systems with stochastic sequential lead times. Oper Res 54(4):706–724