CHAPTER 5

An Integrated Framework for Multi-Stage Production-Inventory-Distribution Supply Chain Networks

5.1 Network Topology

The arborescent structure of the networks treated by multi-echelon inventory networks and related models are appropriate for distribution networks. However, in a general supply chain network, the model involves not only distribution but also assembly, in which multiple components are required for the production of one part. Extending multi-echelon inventory networks into supply chain networks is not straightforward. See Ernst and Pyke (1993) and Cohen and Lee (1988) for research in this direction.

Supply 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.

5.1.1 Sites and Stores

A supply chain network 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*: they procure raw materials from outside suppliers;

(2) *Fabrication sites*: they transform raw materials into components;

(3) *Assembly sites*: they assemble the components into semi-finished products or finished goods;

(4) *Distribution sites*: they delivery 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 part that a site produces is a single-level BOM.

All sites in a supply chain typically perform two types of operations: *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 supply chain has two kinds of stores: *input stores* and *output stores*. Each store stocks a single stock keeping unit (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 Figure 5.1, a site is represented by the dashed box containing input and output stores).

There are two special types of stores in this network topology: *source stores* and *end stores*. Source stores are those output stores that do not have any upstream input stores. They represent the upstream boundaries of the supply chain model. End stores are those output stores that have at least one customer demand stream associated with them. An output store could be both a supplier to downstream input stores and a supplier to external customer demand streams. The external customer demand streams constitute the downstream boundaries of the supply chain model.

The sites can be treated as the building blocks for modeling the whole supply chain. Therefore, for the performance analysis of a supply chain model, the performance of each store is analyzed first, then, the whole supply chain performance is analyzed. It also can be seen that a supply chain network structure is closely related to the product structure (BOM) and process structures (serial, parallel, assembly and arborescent structures).

5.1.2 Links

All stores in the supply chain are connected together by links that represent supply and demand processes. Two types of links are defined: *internal link* and *external link*.

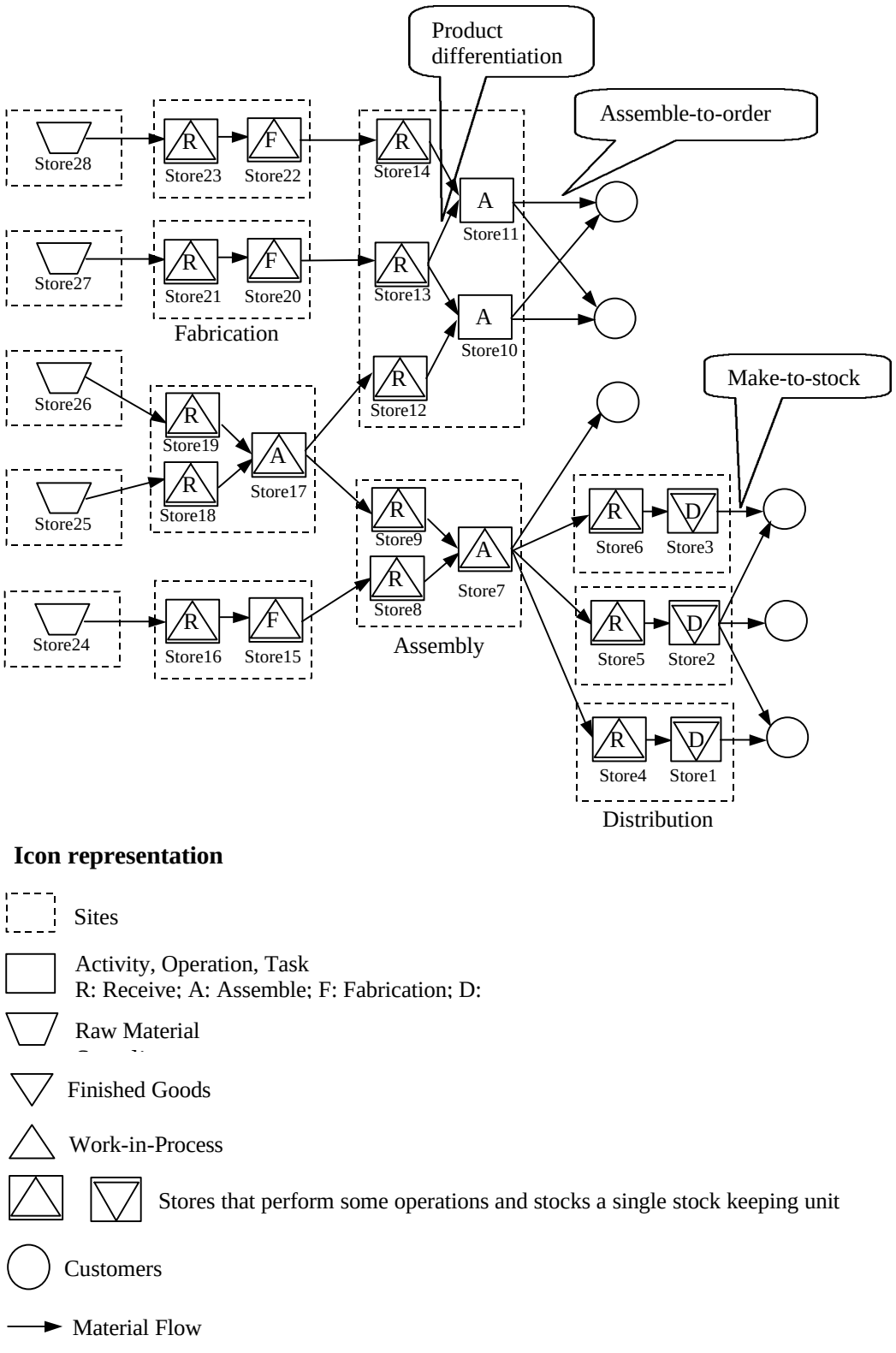

Figure 5.1 An integrated framework for multi-stage supply chain networks

Internal links are used to connect the stores within a site, i.e., they represent the material flow paths from input stores to output stores within a site. Associated with an internal link connecting an input store *i* to an output store *j* is a usage count, u_{ij}, which indicates the number of SKUs in the input store *i* required to produce a SKU in the output store *j*. Along with the usage counts, the internal links connecting input stores and output stores constitute the single-level BOM for that output store.

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. In the network topology, we define that a downstream input store has only one link between it and its upstream output store (see Figure 5.1).

The demand placed on SKUs at a downstream site is translated into a demand for components at the current site via the bill of materials, or equivalently, the usage count. The downstream demand, in turn, creates demand at the supplying site. Hence, the whole supply chain network behaves as a "pull" system in terms of material requirements. This is also called the *demand transmission process* (Garg 1999).

5.2 The Relationships between Stores

Let ***ST*** be the collection of stores in a supply chain network 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.

5.2.1 Inventory Positions within Stores

At each store, there are three types of inventory quantities: *on-hand inventory quantity*, *on-order quantity* and *back-order quantity*, which constitute the inventory position at that store.

At a store, on-hand inventory refers to the number of SKUs that are currently in inventory. A store may have some outstanding orders (the total replenishment stock ordered from upstream suppliers but have not been physically delivered to the store. These outstanding orders are called on-order quantity. Back-order quantity is the quantity of SKUs from current store that have been ordered by downstream customers but not yet filled. The inventory position at a store is then defined as the total of on-hand inventory plus on-order quantity minus the backorders (see Figure 5.2).

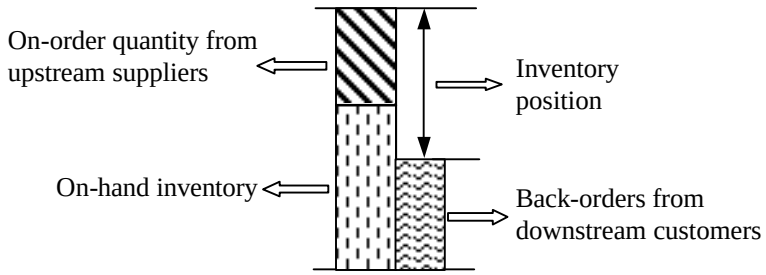

Figure 5.2 Inventory position at a store

5.2.2 Nominal and Actual Replenishment Lead Times of Stores

The lead time of a store is defined as the time required to obtain a replenishment for a SKU in that store. In terms of store types, two cases are discussed in the following.

(1) Case 1: *i* is an input store

In this case, according to our network topology, *i* has a unique supplier, i.e., *UPST*(*i*).

If *i* places an order and its upstream supplier *UPST*(*i*) has on-hand inventory, then the order is filled immediately and shipped to store *i*. The shipping time or transportation time constitutes the nominal lead time of store *i*, denoted by $L_n(i)$.

If *i* places an order and its upstream supplier, *UPST*(*i*), has a stock-out, then the order joins a backorder queue at this supplier store, i.e., the order has to wait to get filled. Therefore, the order has an additional delay time due to non-availability of components from upstream suppliers. When considering the possibility of a stock-out at the supply stores, the corresponding lead time is called as actual lead time, denoted by $L_a(i)$. Or more general, the actual lead time associated with a store is defined as the difference between the time an order is placed by this store and the time the filled order arrives at the store.

(2) Case 2: *i* is an output store

In this case, the replenishment order becomes essentially a production order. If all input stores of *i* have enough on-hand inventory, the production starts immediately. The production lead time constitutes the nominal lead time of store *i*, also denoted by $L_n(i)$.

If any one of the input stores of *i* has a stock-out, the production order joins a backorder queue and is not started until all components required for the production are available. The corresponding actual lead time is also denoted by $L_a(i)$.

For a distribution site, both input stores and output stores represent the material stockpiles within the same site (i.e., no physical material transformation occurs). The nominal lead time would be used to model the time required to retrieve parts from bins and prepare orders for shipment.

In summary, if an order is placed inside a site, then the lead time is either nominal production lead time or actual production lead time. If an order is placed outside a site, i.e., from upstream sites, then the lead time is either nominal shipping lead time or actual shipping lead

time. Or equivalently, the material flow time associated with an external link is called material (shipping) lead time, the flow time associated with an internal link is called production lead time.

5.2.3 Nominal and Actual Replenishment Lead Times of Sites

In the case of there is a unique output store within a site, e.g., a non-assembly site, the nominal lead time (NLT) of this site is defined as its actual replenishment lead time (ALT) under the conditions that its input store has a stock-out, but the supplier of this input store has available components. Thus, NLT of a site equals to the summation of input store NLT and output store NLT.

The actual lead time of a non-assembly site is the actual lead time of its output store when its input store has a stock-out and the supplier of this input store also has a stock-out. Thus, ALT of a site equals to the summation of input store ALT and output store NLT.

When a site has more than one output stores, the corresponding NLT and ALT of this site would be the ones with the maximal values among several output stores' NLTs and ALTs, respectively, under above conditions.

For assembly sites, the situation is somewhat complex. If an order is triggered by an output store which has a BOM consisting of more than one type of SKU, the assembly operation will only proceed when sufficient quantities of all input SKUs are available. These input materials have different nominal shipping lead time, and delays resulting from the availability of materials at the supplying sites differ.

In terms of above definitions, for assembly sites, we have:

$$NLT = \max_{j \in OS} \{ \max_{i \in UPST(j)} [NLT(i)] + NLT(j) \} \quad (5\text{-}1)$$

and

$$ALT = \max_{j \in OS} \{ \max_{i \in UPST(j)} [ALT(i)] + NLT(j) \} \quad (5\text{-}2)$$

where OS is the set of output stores in a site, $UPST(j)$ is the set of directly upstream supplying stores j.

CHAPTER 6

Performance Analysis and Optimization of Multi-Stage Production-Inventory-Distribution Supply Chain Networks

An important issue in a supply chain and the primary purpose of the supply chain model is controlling the inventory at different sites or stores while meeting customer service level requirements, therefore quantifying the trade-off between inventory investment and customer service levels. Since the trade-off between inventory investment and service levels may change over time, this will request that the supply chain performance to be evaluated continuously so that the supply chain managers be able to make timely and right decisions.

For each SKU at every store in a supply chain, either a target service level (fill rate) or a target inventory stock level is specified. Therefore, through the performance analysis, one can determine the required inventory stock level to support the target service level or the achieved service performance given the inventory stock level.

6.1 Information Flows in Supply Chain Networks

Nodes (sites or stores) in the supply chain networks are connected by three types of flows: the information flow, the material flow and control flow. The information flow consists of two subtypes: demand information and service requirement information. The direction of information flow is from the downstream end-product stages to the upstream raw material stages. The material flow originates at the upstream raw material stages and terminates at the downstream end-product stages. The control flow can go either upstream stages or downstream stages.

The demand and service level requirement information will be specified when customers place orders for end products. Then, these demand and service information flow from the end-product stages to the upstream stages in supply chain networks. From the information on end-product stages, the demand and service level information for other stages can be derived

according to the bill-of-materials and routings in supply chain networks. The determination procedure of demands for non-end-product stages is called *demand transmission* and the determination procedure of service levels for non-end-product stages is called *service level transmission*.

6.1.1 Demand Transmission Process in Supply Chain Networks

The demand transmission is used to determine the mean and variance of the demands for SKUs of upstream stages of end-product stages. The demand placed on SKUs at a downstream site is translated into a demand for components at the current site via the bill of materials. The downstream demand, in turn, creates demand at the supplying site. Hence, the whole supply chain network behaves as a "pull" system in terms of material requirements. This is called the *demand transmission process* (Garg 1999).

Garg (1999), Graves (1988) and Lee and Billington (1993) assume that the demand for finished goods is a stationary, uncorrelated and normally distributed random variable. Apparently, the derived demand for each SKU at upstream sites is not stationary, uncorrelated and normally distributed since the commonality of components among products will result in positively correlated demands for common components. However, based on the simulation results on an assembly system, Yao (1994) compares some performance measures (such as lead times, backorder levels, etc.) of correlated demand and uncorrelated demand situations. He finds that the difference between them is insignificant when service level requirements for finished goods are high. Therefore, we can simplify the positively correlated demands by approximating them into uncorrelated demands. That is, the derived demand for each SKU at upstream sites is also assumed to be stationary, uncorrelated and normally distributed.

Through the bill-of-materials, the *demand transmission process* can be represented mathematically as follows:

Let

u_{ij} = the number of component *i* needed for each SKU *j*;

δ_{ij} = the indicator variable = 1 if component *i* is used in SKU *j*

= 0 otherwise;

$\mu(j, h)$ = mean demand per unit time period for SKU *j* at store *h*;

$\nu(j, h)$ = variance of the demand per unit time period for SKU *j* at store *h*;

$p_i(g, h)$ = the proportion of the requirement of component *i* from store *g* at its downstream store *h*;

DOWNST(*g*, *i*) = downstream Stores of store *g* which supplies components *i*;

DOWNST(*h*, *j*) = downstream Stores of store *h* which supplies components *j*.

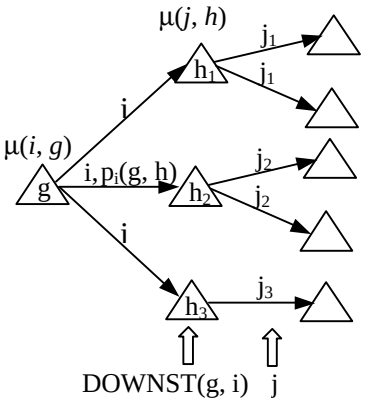

Figure 6.1 The illustration of an demand transmission process

From Figure 6.1, we have following equations:

$$m(i,g) = \sum_{h \in DOWNST(g,i)} \sum_{j} d_{ij} m_{ij} p_i(g,h) m(j,h) \qquad (6\text{-}1)$$

$$n(i,g) = \sum_{h \in DOWNST(g,i)} \sum_{j} d_{ij} m_{ij}^2 p_i^2(g,h) n(j,h) \qquad (6\text{-}2)$$

6.1.2 Service Level Transmission in Supply Chain Networks

Consider an output store *h* that requires input component *i* for its production of *j* from input store *g* at the same site. The component *i* is sourced from an upstream output store *f*. Then, the *immediate availability level* of component *i* is defined as the fill rate of the material at store *f*.

When store *f* has a stock-out, delays due to material shortage at store *f* are called the response time of material at store *f*.

The availability of a SKU at a site depends upon response times and fill rates of its input components from the upstream sites. Therefore, service-level requirements on end products drive the service-level requirements for each upstream SKU. This is called the *service transmission process*. Or equivalently, Lee and Billington (1993) define the transformation of output product service performance measures at a supplying site to input material availability measures at a downstream site as the *availability transfer process*.

The service-level requirement at a site can be defined as a function of this site's *target response time* and the SKU *fill rate* at this site. For instance, a service-level requirement that requires at least 98% demand should be met within 2 days can be specified as a response time 2 days and fill rate of 98%.

The expression for response time requirements of components is a result of *response time transmission*.

Let the actual response time for SKU *j* at site *h* be $ALT_j(h)$, i.e., the actual replenishment lead time. Let the mean and variance of the nominal replenishment lead time for SKU *j* at site *h* be $\mu_j(h)$ and $\nu_j(h)$, respectively. If actual response time for SKU *j* is greater than its replenishment lead time, then the SKU is make-to-order or assemble-to-order. Suppose site *g* supplies component *i* for the production of component *j*. Thus, the start of component *i*'s production could be delayed by $\{ALT_j(h)-\mu_j(h)-\kappa[\nu_j(h)]^{1/2}\}^+$, where κ is the safety factor. If component *i* is used by multiple SKUs downstream, the maximal delay time is $min_{j\in DSi}\{ALT_j(h)-\mu_j(h)-\kappa[\nu_j(h)]^{1/2}\}^+$, where DS_i is the set of SKUs that use component *i*.

6.1.3 Product Demands and Target Service Levels

In our supply chain model, the product demands are represented as independent streams of orders for SKUs stored at any of the output stores in the sites. Each stream models the demands of a customer or a group of customers over time. Multiple demand streams could exist

for the same output store, representing the fact that multiple customers may order products from the same location in a supply chain.

Let M denote the set of all customer demand streams. Associated with each customer demand stream m∈ M are: (1) the end (output) store that supplies the demand, *ENDST*(*m*); (2) a forecast or real demand stream *D*(*m, t*); (3) the delivery time, T_m, which models the time to delivery the product from the end store to the customer; and (4) target or desired service-level requirements for that customer demand stream. The target service-level requirements are specified through two given parameters: due date b_m (could be a random number) for demand stream *m*, and fill rate a_m that is the fraction of class *m* orders filled before the due date.

Let W_m be the waiting time to receive an order from customer demand stream *m*, then the service-level requirement for demand stream *m* is given as follows:

$$P[W_m \leq b_m] \geq a_m \qquad (6\text{-}3)$$

6.2 Performance Analysis of a Single-Product-Type Store

6.2.1 Make-to-Stock Stores

The principle motivation for operating a manufacturing system on a make-to-stock basis is the desire to improve the level of service to customers and gain the competitive advantages that result from improved service. Make-to-stock operation can reduce the delay in filling customer orders, and availability of products on stock may lead to increased sales. Make-to-stock production may also result in savings in the costs associated with setting up or starting up the machines or processes.

However, on the other hand, make-to-stock operation usually implies that there is an inventory of products available to fill customer orders as they arrive. This inventory will require an investment of money that could be used elsewhere in the firm. Thus, in deciding whether to produce to stock, the firm has to weigh the advantages of improved service, and perhaps lower manufacturing costs against the costs and other problems associated with keeping inventories.

For a make-to-stock store, the following system characteristics have to be considered.

(1) Product Variety

If a store will produce several different products, then the interrelationship between the demand and manufacturing of the products is a critical issue to consider. In a common situation in which only one of the products can be produced at a time on a machine, and there is a change-over cost between products. Demands for the products would be related if a customer orders one each of all products produced on the facility.

(2) Pattern of Demand

Two aspects of the demand pattern are significant: the customer arrival pattern and demand pattern for products given by a customer. Customer arrivals may be non-stationary (e.g., they may exhibit trends with time and seasonal variations) as well as vary with such external factors as the level of economic activity. And, arrival patterns may be influenced by the pricing and promotion policies.

(3) Manufacturing Capacity

All manufacturing processes are to some extent unreliable or uncertain. For example, machines can fail, tools can break, or operators can be absent. And some finished products may fail to meet the specification and have to be scrapped or reworked. This yield loss can be quite significant in some industries such as very large scale integrated (VLSI) chip fabrication.

(4) Change-over and Set-up

In multiple product facilities change-over between products can require a significant time. Also there may be costs associated with starting up or shutting down the facility, or there may be a need to shut down the facility for cleaning, maintenance, or to replace worn tools after some given total quantity of items has been produced. This may lead to production in batches, where the batch size can be fixed and constant, or it can vary between successive runs of a given product.

6.2.2 Production Authorization (PA) Cards Control Mechanism

A store is considered as a single-stage manufacturing facility, that is, one or a group of c identical machines, any of which is capable (perhaps after appropriate set-up) of making any one of the SKUs produced by the facility. Finished SKUs of each component are kept in an inventory area. As demands arrive their orders are met by the SKUs from the inventory area. If all orders cannot be met immediately then they would join the back-order queue.

Production authorization (PA) mechanism works as follows: when each SKU is produced by the manufacturing facility, a tag is attached to this SKU and thus for every SKU in inventory area there is a tag. When a SKU is shipped to a customer the tag is removed, and this then becomes the production authorization or PA card for that SKU.

Some control rules can be applied for determining when the PA cards would be transmitted to the manufacturing facility (Buzacott 1989).

(1) Immediate transmittal to the facility as soon as a PA card is generated.

(2) Transmittal of batches of fixed size q as soon as at least q cards have been accumulated.

(3) Transmittal of a batch of all available PA cards once at least q cards have been accumulated.

Observe that if customer demands are of unit size there is essentially no difference between rules 2 and 3.

Usually, the presence of a PA card at the manufacturing facility authorizes the production of one SKU. If there is a yield loss in manufacturing, however, then the PA card could authorize production of more than one SKU; alternatively, when a SKU is found to be defective it will immediately create a PA card.

For the single-product-type manufacturing facility, if set-up requires negligible time or cost then the simplest rule is to produce whenever PA cards are available. Otherwise the following rules could be used: (1) Start up once there are at least q PA cards present, continue

producing as long as there are any PA cards, then shut down once there are no PA cards and the facility becomes idle. (2) If PA cards arrive in batches, set up for each batch.

Once the rules for PA card generation and the control rules for manufacturing facility set-up have been decided, the next task is to determine the parameters of rules. The major parameters are the desired inventory level at a store and the lot sizes for transmittal of PA cards. The corresponding optimal values can be found using cost models analogous to classical inventory, where the cost components are inventory holding costs, set-up costs and backlog costs. To establish such an optimization model, the following performance measures are needed: (1) the average finished good inventory, (2) the average backlogged demand, (3) the average lead time to fill an order, (4) the probability a demand becomes backlogged, and (5) the frequency of set-up.

PA cards convey the information on the occurrence of demands to the manufacturing facility. The PA mechanism enables the manufacturing facility to implement strategies for releasing jobs for processing based on the physical inventory level and work in process at the facility. Using such information significantly enhances the performance of the supply chain with respect to both inventory and service levels. From a modeling perspective, the PA card mechanism allows one to see the insights for the dynamics of the system and how information can be used to control material flow. This facilitates a systematic approach to the modeling of the make-to-stock systems.

Compared with other material control mechanisms such as KANBAN, CONWIP, starvation avoidance, MRP and etc., the PA card mechanism offers more flexibility. There are a wide number of possibilities for setting lead time delays, PA card limits, and inventory targets at the store.

6.2.3 Performance Analysis with $GI^x/G/1$ Queue

Ettl et al. (2000) modeled each store as an infinite-server queue operating under a base-stock control rule. In particular, they use the $M^x/G/\infty$ queue, where demand arrivals follow a

Poisson process with rate λ, and each arrival brings a batch of *X* orders. The use of one-for-one base-stock policy avoids the determination of lot size at each store. This maybe suitable for the situation in which the production set-up cost is negligible. Since the one-for-one replenishment policy requires the machines to reset-up for each production of an order, this policy will result in a higher cost if the set-up cost is not negligible. And in reality, most time this set-up cost cannot be neglected. Therefore, instead of using (*S-1*, *S*) base-stock policy, we adopt (*s*, *S*) inventory policy. This policy works as follows: when the inventory at a store is less than *s*, the production will start to produce the items until the up-to-order level *S* is achieved. In addition, the assumption of uncapacitated production is often not true in practice. In the following, we use M^X/G/1 queue operating under (*s*, *S*) inventory control rule to analyze the performance of a single-product make-to-stock store.

To analyze the performance of a make-to-stock store under (*s*, *S*) inventory policy, we adopted the target-level PA mechanism with fixed lot size. This mechanism works as follows: let *S* be the order-up-to point of a store inventory, each item at the store is attached with a tag. When a demand stream arrives and there are some available items at the store, this demand is met and the tag with this demand is removed from the item. Then this tag is activated into a PA card. Whenever *q* or more PA cards are accumulated at the store, *q* PA cards are transmitted to the manufacturing facility. In other words, *q* is the lot size. Each of these *q* PA cards authorizes the machine to produce one item. And the overlapping production initiation policy is adopted: as soon as the manufacturing of all *q* items of a lot of PA cards is initiated, the production of items for the subsequent lot of PA cards can be initiated (whenever machines become available). This is so called (*q*, *S*) fixed-batch target-level PA mechanism (Buzacott 1989). When the production of *q* items is completed and shipped to the store inventory area, the PA cards with these items will be converted into tags. Observe that the maximum number of items in the store will always be less than or equal to *S*. Therefore, this mechanism is essentially the same as the traditional reorder point/order quantity inventory control policy. Unlike in inventory modeling, the fixed-lot

target-level PA mechanism will explicitly model the process that determines the lead time for replenishment of items at the store.

Consider a queuing system consisting of a waiting (dispatch) area and a service facility. Customers arrive at the queueing system according to the arrival process $\{A_n, n = 1, 2, \ldots\}$. The number of orders brought in by the *n*th customer is $X_n = q$, n = 1, 2, …. That is, the lot size is fixed and equals to *q*. Here, we assume that each order corresponds to one SKU. Each customer on its arrival enters the waiting area. If the number of orders in the service facility is less than *S*, orders from the waiting area are sent into the service facility one at a time until there are *S* customers in the service facility.

Suppose the store is initially full at time zero. The *n*th customer arrives at time $\tilde{A}_n = A_{nq}$ and brings *q* orders. The service time of the *n*th order is T_n, n = 1, 2, …Let D_n be the departure time of the *n*th order from this queueing system. Then $\tilde{D}_n = D_{nq}$ is the time epoch where the *n*th lot of *q* PA cards order is filled. In other words, every *q*th order departure corresponds to a time at which a batch of *q* PA card orders is filled.

Let $N(t)$ be the number of orders in the batch arrival $\tilde{G}I^q/G/1$ queueing system with arrival process $\{\tilde{A}_n, n = 1, 2, \ldots\}$, service times $\{T_n, n = 1, 2, \ldots\}$ and batch size $\{X_n = q, n = 1, 2, \ldots.\}$. For the batch of PA cards currently being manufactured, define $C(t)$ be the number of finished SKUs at the manufacturing facility waiting to be shipped. Since for each batch there are at most *q* items that are in manufacturing, we have

$$C(t) = \left\lceil \frac{N(t)}{q} \right\rceil \cdot q - N(t), \qquad t \geq 0 \tag{6-4}$$

At each arrival epoch of $\tilde{A}_n = A_{nq}$, we create a lot tag (i.e., one tag for every *q* orders) and destroy one whenever a lot of *q* PA cards is returned to inventory area. Let $R(t)$ be the number of orders arriving at or before time t, but after the last lot tag was created. Then

$$R(t) = A(t) - \left\lfloor \frac{A(t)}{q} \right\rfloor \cdot q, \qquad t \geq 0 \tag{6-5}$$

where $A(t)$ is the number of orders that arrived during (0, t].

Since $C(t)+N(t)$ is the number of orders within the manufacturing facility and $R(t)$ is the number of orders waiting at outside the manufacturing facility, the total number of orders during (0, t] is $C(t)+N(t)+R(t)$. Therefore the number of orders backlogged at time t is

$$B(t) = [C(t) + N(t) + R(t) - S]^+ = \left(\left\lceil \frac{N(t)}{q} \right\rceil \cdot q + R(t) - S \right)^+, \qquad t \geq 0 \qquad (6\text{-}6)$$

Similarly, the inventory of finished SKUs at inventory area is

$$I(t) = \left(S - \left\lceil \frac{N(t)}{q} \right\rceil \cdot q - R(t) \right)^+, \qquad t \geq 0 \qquad (6\text{-}7)$$

From equation (6-5), it can be seen that as t increases, $R(t)$ will uniformly increase from *0* to q-1, drop to *0*, increase from *0* to q-1, and so on. Then we have

$$P\{R = n\} = \frac{1}{q}, \qquad n = 0,1,...,q-1 \qquad (6\text{-}8)$$

Assume that $N(t)$ and $R(t)$ are independent, and let $p_N(n)$ be the distribution of $N(t)$. From equations (6-4), (6-6) and (6-7), it can be shown

$$P\{C = n\} = p_N\left(\left\lceil \frac{n}{q} \right\rceil \cdot q - n \right), \qquad n = 1,2,... \qquad (6\text{-}9)$$

$$P\{B = n\} = \frac{1}{q} p_N\left(\left\lceil \frac{n}{q} \right\rceil \cdot q + S \right), \qquad n = 1,2,... \qquad (6\text{-}10)$$

$$P\{I = n\} = \frac{1}{q} p_N\left(S - \left\lceil \frac{n}{q} \right\rceil \cdot q \right), \qquad n = 1,2,...,S \qquad (6\text{-}11)$$

Next, some approximations are used to derive the distribution of $N(t)$, i.e., $p_N(n)$.

An approach to solve the M^X/G/1 model is to analyze the $M/\tilde{G}/1$ queue with arrival rate λ and service time equal in distribution to $\sum_{n=1}^{X} T_n$. Similarly, for the renewal customer arrival processes (not necessarily Poisson), the $GI/\tilde{G}/1$ queue can be used to model the dynamics of

the jobs in $GI^X/G/1$ problem, where each job is the aggregate of orders brought in by each customer, and has an arrival process {A_n, n = 1, 2, …} and service times $\tilde{T}_n = \sum_{j=1}^{X_n} T_{nj}$, n = 1, 2, …. The mean of the service times $\tilde{T}$ is $E[X]E[T]$ (refer to Wolff 1989 for general background materials for queues). The squared coefficient of variation the service times $\tilde{T}$, which is defined as $C_{\tilde{T}}^2 = Var(\tilde{T})/E[\tilde{T}]^2$, is then $C_X^2 + (1/E[X])C_T^2$.

Let $\tilde{W}_{GI/G/1}(l, m, C_a^2, C_T^2)$ be one of the approximations for the average waiting time in queue for a GI/G/1 queue with mean interarrival time 1/λ, mean service time 1/μ, squared coefficient of variation of interarrival time C_a^2, and squared coefficient of variation of service time C_T^2. Then the average waiting time of a customer in the $GI/\tilde{G}/1$ queue can be approximated by

$$w_c = \tilde{W}_{GI/G/1}(l, 1/E[X]E[T], C_a^2, C_X^2 + (1/E[X])C_T^2) \tag{6-12}$$

Let $r = l/m = l/E[X]E[T]$ and $C_{\tilde{T}}^2 = C_X^2 + (1/E[X])C_T^2$, three approximations for w_c are w_{c1}, w_{c2} and w_{c3} (Wolff 1989):

$$w_{c1} = \left[\frac{r^2(1+C_{\tilde{T}}^2)}{1+r^2C_{\tilde{T}}^2}\right]\cdot\left[\frac{C_a^2 + r^2C_{\tilde{T}}^2}{2l\cdot(1-r)}\right] + E[\tilde{T}] \tag{6-13}$$

$$w_{c2} = \left[\frac{r(1+C_{\tilde{T}}^2)}{2-r+rC_{\tilde{T}}^2}\right]\cdot\left[\frac{r(2-r)C_a^2 + r^2C_{\tilde{T}}^2}{2l\cdot(1-r)}\right] + E[\tilde{T}] \tag{6-14}$$

$$w_{c3} = \frac{rC_a^2(1-(1-r)C_a^2) + r^2C_{\tilde{T}}^2}{2l\cdot(1-r)} + E[\tilde{T}] \tag{6-15}$$

Extensive empirical testing shows that approximations for w_{c1} and w_{c2} work very well when $C_a^2 \le 2$. When the squared coefficient of variation of interarrival time is very large the w_{c1} can be very poor. This is to be expected because the range of the exact mean flow time for different distributions of interarrival times with a fixed mean and large squared coefficient of variation can be very large. If $C_a^2 \le 1$, w_{c3} will provide a better approximation.

Observe that when $C_a^2 = 1$, the approximation agrees with the results for Poisson arrival process, provided that the approximation $\tilde{W}_{GI/G/1}$ selected is exact for M/G/1 queues.

The average number of customers waiting in the queue can be obtained from Little's formula:

$$N_c = l \cdot w_c \tag{6-16}$$

Since each customer in queue consists of an average of $E[X]$ (or q) orders, the average number of orders in the system corresponding to those customers waiting in the queue (i.e., orders in waiting) is

$$N_{ow} = l \cdot E[X] \cdot w_c = l \cdot q \cdot w_c \tag{6-17}$$

The average number of orders in the system corresponding to the customers in service (i.e., orders in service), if any, is

$$N_{os} = \frac{E[X^2] + E[X]}{2E[X]} \tag{6-18}$$

Because the probability that a customer is in service is $\rho = \lambda E[X]E[T]$, the average number of orders, $E[N_o]$, in the system can be approximated by

$$E[N_o] = N_{ow} + N_{os} \cdot r = l \cdot E[X] \cdot w_c + \frac{l \cdot (E[X^2] + E[X])E[T]}{2}$$

$$= l \cdot q \cdot w_c + \frac{l \cdot (q^2 + q)E[T]}{2} \tag{6-19}$$

Therefore, the distribution of the number of orders in this $GI^X/G/1$ is approximated by

$$P\{N_o = n\} \approx \begin{cases} (1-r), & n = 0 \\ r(1-s)s^{n-1}, & n = 1,2,... \end{cases} \tag{6-20}$$

where

$$s = \frac{E[N_o] - r}{E[N_o]} \tag{6-21}$$

is chosen such that the average of the approximated distribution is equal to $E[N_o]$.

Substitute equation (6-20) into equations (6-10) and (6-11), we have

$$P\{B=n\}=\frac{r}{q}(1-s\ )s^{\left\lceil\frac{n}{q}\right\rceil\cdot q+S-1},\qquad n=1,2,... \tag{6-22}$$

and

$$P\{I=n\}=\frac{r}{q}(1-s\ )s^{S-\left\lceil\frac{n}{q}\right\rceil\cdot q-1},\qquad n=1,2,...,S \tag{6-23}$$

Our next target is to find the means of B(t) and I(t), therefore, we have

$$E[B]=\sum_{n=1}^{\infty}n\cdot\frac{r}{q}(1-s\ )s^{\left\lceil\frac{n}{q}\right\rceil q+S-1}=\frac{r}{q}(1-s\ )s^{\ S-1}\sum_{n=1}^{\infty}n\cdot s^{\left\lceil\frac{n}{q}\right\rceil q} \tag{6-24}$$

The last summation in equation (5-24) can be obtained as follows

$$\begin{aligned}\sum_{n=1}^{\infty}n\cdot s^{\left\lceil\frac{n}{q}\right\rceil\cdot q}&=[1+2+\cdots+(q-1)+q]s^{\ q}+[(q+1)+(q+2)+\cdots+2q]s^{\ 2q}+\cdots\\&\quad+[(mq+1)+(mq+2)+\cdots+(mq+q)]s^{\ (m+1)q}+\cdots\\&=\frac{(1+q)q}{2}s^{\ q}+\frac{(1+3q)q}{2}s^{\ 2q}+\frac{(1+5q)q}{2}s^{\ 3q}+\cdots+\frac{[1+(2m+1)q]q}{2}s^{\ (m+1)q}+\cdots\\&=\frac{q}{2}s^{\ q}\{(1+q)+(1+3q)s^{\ q}+(1+5q)s^{\ 2q}+\cdots+[1+(2m+1)q]s^{\ mq}+\cdots\}\end{aligned}$$

Let

$$Z=\{(1+q)+(1+3q)s^{\ q}+(1+5q)s^{\ 2q}+\cdots+[1+(2m+1)q]s^{\ mq}\} \tag{6-25}$$

and multiply σ^q to both sides of equation (6-25), then

$$s^{\ q}Z=\{(1+q)s^{\ q}+(1+3q)s^{\ 2q}+(1+5q)s^{\ 3q}+\cdots+[1+(2m+1)q]s^{\ (m+1)q}\} \tag{6-26}$$

From (6-25)-(6-26), we have

$$\begin{aligned}(1-s^{\ q})Z&=(1+q)-[1+(2m+1)q]s^{\ (m+1)q}+2qs^{\ q}[1+s^{\ q}+s^{\ 2q}+\cdots+s^{\ (m-1)q}]\\&=(1+q)-[1+(2m+1)q]s^{\ (m+1)q}+2qs^{\ q}\left[\frac{1-s^{\ mq}}{1-s^{\ q}}\right]\end{aligned} \tag{6-27}$$

thus

$$Z=\frac{1+q}{1-s^{\ q}}-\frac{[1+(2m+1)q]s^{\ (m+1)q}}{1-s^{\ q}}+\frac{2qs^{\ q}}{1-s^{\ q}}\cdot\left[\frac{1-s^{\ mq}}{1-s^{\ q}}\right] \tag{6-28}$$

From equation (6-21), we know $|s|<1$. So the limit of Z is

$$\lim_{m\to\infty} Z = \frac{1+q}{1-s^{\,q}} + \frac{2qs^{\,q}}{1-s^{\,q}} \cdot \left[\frac{1}{1-s^{\,q}}\right] \tag{6-29}$$

Therefore

$$E[B] = \frac{r}{q}(1-s\,)s^{\,S-1}\sum_{n=1}^{\infty} n \cdot s^{\left\lceil \frac{n}{q} \right\rceil \cdot q} = \frac{(1+q)(1-s\,)rs^{\,S+q-1}}{2(1-s^{\,q})} + \frac{qs^{\,2q+S-1}r(1-s\,)}{(1-s^{\,q})^2} \tag{6-30}$$

Similarly,

$$E[I] = \sum_{n=1}^{S} n \cdot \frac{r}{q}(1-s\,)s^{\,S-\left\lceil \frac{n}{q} \right\rceil \cdot q-1} = \frac{r}{q}(1-s\,)s^{\,S-1}\sum_{n=1}^{S} n \cdot s^{\,-\left\lceil \frac{n}{q} \right\rceil \cdot q} \tag{6-31}$$

and define $\left\lceil \frac{S}{q} \right\rceil = m+1$, then

$$\begin{aligned}
\sum_{n=1}^{S} n \cdot s^{\,-\left\lceil \frac{n}{q} \right\rceil \cdot q} &= [1+2+\cdots+(q-1)+q]s^{\,-q} + [(q+1)+(q+2)+\cdots+2q]s^{\,-2q} + \cdots \\
&\quad + \{[(m-1)q+1]+[(m-1)q+2]+\cdots+mq\}s^{\,-mq} + [(mq+1)+(mq+2)+\cdots+S]s^{\,-(m+1)q} \\
&= \frac{(1+q)q}{2}s^{\,-q} + \frac{(1+3q)q}{2}s^{\,-2q} + \cdots + \frac{[1+(2m-1)q]q}{2}s^{\,-mq} + \frac{(1+mq+S)(S-mq)}{2}s^{\,-(m+1)q} \\
&= \frac{q}{2}s^{\,-q}\{(1+q)+(1+3q)s^{\,-q}+\cdots+[1+(2m-1)q]s^{\,-mq}\} + \frac{(1+mq+S)(S-mq)}{2}s^{\,-(m+1)q}
\end{aligned}$$

Let

$$Z' = \{(1+q)+(1+3q)s^{\,-q}+\cdots+[1+(2m-1)q]s^{\,-mq}\} \tag{6-32}$$

and multiply σ^q to both sides of equation (6-25), then

$$s^{\,-q}Z' = \{(1+q)s^{\,-q}+(1+3q)s^{\,-2q}+\cdots+[1+(2m-3)q]s^{\,-mq}+[1+(2m-1)q]s^{\,-(m+1)q}\} \tag{6-33}$$

From (6-32)-(6-33), we have

$$\begin{aligned}
(1-s^{\,-q})Z' &= (1+q)-[1+(2m-1)q]s^{\,-(m+1)q} + 2qs^{\,-q}[1+s^{\,-q}+s^{\,-2q}+\cdots+s^{\,-(m-1)q}] \\
&= (1+q)-[1+(2m-1)q]s^{\,-(m+1)q} + 2qs^{\,-q}\left[\frac{1-s^{\,-mq}}{1-s^{\,-q}}\right]
\end{aligned}$$

that is

$$Z' = \frac{(1+q)}{(1-s^{\,-q})} - \frac{[1+(2m-1)q]s^{\,-(m+1)q}}{(1-s^{\,-q})} + 2qs^{\,-q}\left[\frac{1-s^{\,-mq}}{(1-s^{\,-q})^2}\right] \tag{6-34}$$

Therefore

$$E[I] = \frac{r}{q}(1-s)s^{S-1}\sum_{n=1}^{S} n \cdot s^{-\left\lceil \frac{n}{q} \right\rceil \cdot q}$$

$$= \frac{r}{2}(1-s)s^{S-q-1}Z' + \frac{r}{q}(1-s)s^{-(m+1)q+S-1}\frac{(1+mq+S)(S-mq)}{2} \qquad (6\text{-}35)$$

6.2.4 Optimization of Inventory Positions

Since each arrival customer will bring in $E[X]$ (or q) orders, the effective arrival rate is $lE[X]$. If the setup cost is k_1 per batch, the inventory carrying cost is k_2 per unit time, and the backlogging cost is k_3 per unit time, then the total cost rate $TC(q, S)$ for this target level PA mechanism with fixed-lot size is

$$TC(q,S) = k_1\left\lceil \frac{l \cdot E[X]}{q} \right\rceil + k_2 E[I] + k_3 E[B]$$

$$= k_1 l + k_2\left[\frac{r}{2}(1-s)s^{S-q-1}Z' + \frac{r}{q}(1-s)s^{-(m+1)q+S-1}\frac{(1+mq+S)(S-mq)}{2}\right]$$

$$+ k_3\left[\frac{(1+q)(1-s)rs^{S+q-1}}{2(1-s^q)} + \frac{qs^{2q+S-1}r(1-s)}{(1-s^q)^2}\right] \qquad (6\text{-}36)$$

This result can be used to obtain the optimal values of q and S that minimize the total cost of inventory, backlogging as well as machine set-up costs if the machines are set up for each batch of PA cards.

6.2.5 Performance Analysis of a Store

6.2.5.1 Stock-out Probability of a Store

The stock-out probability at a store or the probability of a customer is backlogged, denoted by p, is the fraction of time that the on-hand inventory at the store is zero:

$$p = P[I = 0] = P\left[S \le \left\lceil \frac{N(t)}{q} \right\rceil \cdot q + R(t)\right] \qquad (6\text{-}37)$$

Since $R(t)$ uniformly increases from *0* to q-1, drops to *0*, increases from *0* to q-1, and so on. The distribution of $R(t)$ then is $1/q$. Thus, we have the following expressions:

When $R(t) = 0$,

$$I(t) = \left\{ S - \left\lceil \frac{N(t)}{q} \right\rceil q \right\}, \qquad N(t) = \left\lfloor \frac{S}{q} \right\rfloor q + 1, \left\lfloor \frac{S}{q} \right\rfloor q + 2, \cdots \tag{6-38}$$

When $R(t) = 1$,

$$I(t) = \left\{ (S-1) - \left\lceil \frac{N(t)}{q} \right\rceil q \right\}, \qquad N(t) = \left\lfloor \frac{S-1}{q} \right\rfloor q + 1, \left\lfloor \frac{S-1}{q} \right\rfloor q + 2, \cdots \tag{6-39}$$

When $R(t) = 2$,

$$I(t) = \left\{ (S-2) - \left\lceil \frac{N(t)}{q} \right\rceil q \right\}, \qquad N(t) = \left\lfloor \frac{S-2}{q} \right\rfloor q + 1, \left\lfloor \frac{S-2}{q} \right\rfloor q + 2, \cdots \tag{6-40}$$

……

When $R(t) = q$-1,

$$I(t) = \left\{ (S-q+1) - \left\lceil \frac{N(t)}{q} \right\rceil q \right\}, \quad N(t) = \left\lfloor \frac{S-q+1}{q} \right\rfloor q + 1, \left\lfloor \frac{S-q+1}{q} \right\rfloor q + 2, \cdots \tag{6-41}$$

Then, by the distribution of $N(t)$ in equation (6-20), we have

$$\begin{aligned} p &= P[I = 0] \\ &= \frac{1}{q} r(1-s)s^{q\left\lfloor \frac{S}{q} \right\rfloor} \sum_{n=0}^{\infty} s^n + \frac{1}{q} r(1-s)s^{q\left\lfloor \frac{S-1}{q} \right\rfloor} \sum_{n=0}^{\infty} s^n + \cdots + \frac{1}{q} r(1-s)s^{q\left\lfloor \frac{S-q+1}{q} \right\rfloor} \sum_{n=0}^{\infty} s^n \\ &= \frac{r}{q} s^q \{ s^{\left\lfloor \frac{S}{q} \right\rfloor} + s^{\left\lfloor \frac{S-1}{q} \right\rfloor} + \cdots + s^{\left\lfloor \frac{S-q+1}{q} \right\rfloor} \} \end{aligned} \tag{6-42}$$

Let $S = mq + r$, where $0 \le r < q$, from equation (6-41) we have the stock-out probability

$$p = P[I = 0] = \frac{r}{q} s^q [(r+1)s^m + (q-r-1)s^{m-1}] \tag{6-43}$$

6.2.5.2 Fill Rate of a Store

The fill rate at a store, denoted by *f*, is the fraction of customer orders that is filled by on-hand inventory. The fill rate is also the fraction of time that the on-hand inventory at the store is greater than zero:

$$f = P[I > 0] = P\left[S > \left\lceil \frac{N(t)}{q} \right\rceil \cdot q + R(t) \right] = 1 - p$$

$$= 1 - \frac{r}{q} s^{q}[(r+1)s^{m} + (q-r-1)s^{m-1}] \tag{6-44}$$

Let $\tilde{f} = 1 - f$, we have $\tilde{f} = p$.

6.2.5.3 Performance Analysis of a Non-Source, Non-Assembly Store

Compared with a source store, the performance analysis of a non-source, non-assembly store has to consider the reliability of its unique upstream store. Since the upstream store may not be 100% reliable, the actual lead time of the store may differ from the nominal lead time. To account for this probability, we use the same basic model as above but rather than using the nominal lead time as the service time for $\tilde{G}I^{q}/G/1$ model we use the actual lead time, which is defined as the nominal lead time plus possibly some additional time that corresponds to the delay experienced by an order when the upstream store has a stock-out.

Let *i* be the store we are analyzing and let *j* be the single upstream store of store *i*. Then the actual lead time of store *i* is:

$$\tilde{L}_i = \begin{cases} L_i & w.p. \quad f_j \\ L_i + t_j & w.p. \quad \tilde{f}_j \end{cases} \tag{6-45}$$

In words, the actual lead time of a store is the nominal lead time if the order is filled right away by the upstream store; otherwise it is the nominal lead time augmented by an additional delay term, t_j, which is the typical delay experienced by a back order at store *j*.

6.2.5.4 Performance Analysis of a Non-Source Assembly Store

Non-source assembly stores refer to the output stores that require components from more than one upstream store (for example, an assembly operation).

When an order arrives at store *i* and on-hand inventory at store *i* is empty, store *i* will send demands to all its upstream stores. If these stores all have on-hand inventory, then the probability of the actual lead time equaling to store *i*'s own nominal lead time (i.e., $\tilde{L}_i = L_i$) is:

$$q_{i\_nom} = \prod_{j \in UPST(i)} f_j \tag{6-46}$$

where f_j is the fill rate, the fraction of orders supplied by on-hand inventory at store j. In other words, q_{i_nom} represents the fraction of orders at store i that experiences the store's nominal lead time.

On the other hand, if some of the upstream stores are empty or stock-out, denoted by the set $E \subseteq UPST(i)$, then the actual lead time of store i is:

$$\tilde{L}_i = L_i + \max_{j \in E}(t_j) \tag{6-47}$$

where t_j denotes the additional delay at store $j \in E$.

This will happen with probability

$$q_{i\_delay} = \prod_{j \in UPST(i) \setminus E} f_j \cdot \prod_{j \in E} \tilde{f}_j \tag{6-48}$$

In words, q_{i_delay} represents the fraction of orders at store i that experiences the store's actual lead time.

As stated by Ettl and et al. (2000), there is an assumption for above arguments: the product-form assumes independence among the set of stores that supply i. But they are not really independent since these stores always receive the same replenishment request simultaneously from store i. Even with this approximation, there exists a combinatorial explosion associated with the lead time calculation: $\tilde{L}_i$ is a mixture of $2^{UPST(i)}$ cases of orders not filled on arrival. To simplify the problem: ignore those cases in which an order is not filled simultaneously at two or more stores and renormalize the probabilities, we have (Ettl et al. 2000):

$$\tilde{L}_i = \begin{cases} L_i & w.p.\ \ p_{0i} \\ L_i + t_j & w.p.\ \ p_{ji}, \quad j \in UPST(i) \end{cases} \tag{6-49}$$

where $p_{0i} = \left(1 + \sum_{h \in UPST(i)} \tilde{f}_h / f_h\right)^{-1}, \qquad p_{ji} = \left(\tilde{f}_h / f_h\right) p_{0i}.$ (6-50)

In general, the quantity t_j is quite intractable. Here, we use an approximation given by Ettl and et al. (2000).

At store j, suppose there is zero on-hand inventory, and R_j orders are being processed. Then t_j can be approximated as:

$$t_j = \tilde{L}_j r_j, \qquad where \quad r_j = \frac{E(B_j)}{p_j(R_j+1)}. \tag{6-51}$$

p_j can be obtained from (6-43).

6.3 The Optimization Model

Similar to Ettl and et al. (2000), the objective of the optimization model is to minimize the total expected inventory capital throughout the network while satisfying customer service-level requirements.

From (6-3), we can rewrite the service-requirement as:

$$P[W_i \le b_i] = f_i P[L_i \le b_i] + (1 - f_i) P[L_i + t_i \le b_i] \ge a_i \tag{6-52}$$

and

$$f_i = \frac{a_i - P[L_i + t_i \le b_i]}{P[L_i \le b_i] - P[L_i + t_i \le b_i]} \tag{6-53}$$

There are two types of inventory at each store in the network: finished goods inventory and WIP inventory. From the performance analysis of above, we know that the expected finished goods inventory at a store is $E[I]$, and the expected WIP is $E[N]$.

For end stores that provide finished goods, let c_i denote the inventory capital per SKU at store i. For non-end stores that provide components, the average between the inventory value at store i and the value of all the components that make up the SKU at store i is used, i.e.,

$$\tilde{c}_i = \frac{1}{2}\left(c_i + \sum_{j \in UPST(i)} c_j u_{ji} \right) \tag{6-54}$$

Therefore, the objective function is:

$$\begin{aligned} C(q,S) = \sum_i \{ \tilde{c}_i \cdot \left[l \cdot q \cdot w_c + \frac{l \cdot (q^2+q) E[T]}{2} \right] + c_i \cdot \\ \left[\frac{r}{2}(1-s)s^{S-q-1}Z' + \frac{r}{q}(1-s)s^{-(m+1)q+S-1}\frac{(1+mq+S)(S-mq)}{2} \right] \} \end{aligned} \tag{6-55}$$

Our purpose is to minimize this objective function, subject to meeting the fill-rate constraints for end stores as specified in (6-53).

6.4 Numerical Examples

In this section, several numerical examples are used to illustrate the validity of the proposed model. In particular, the numerical comparison between results from the analytical performance evaluation model and those obtained from simulation study is made.

6.4.1 Performance Analysis Results

Figure 6.2 gives a supply chain network with fourteen stores: two supplier stores, four manufacturing stores, three assembly stores, two warehouse stores and three retailer stores.

6.4.1.1 Example 1

In terms of the analytical results from previous sections, the performance of the supply chain network can be calculated as follows:

Example 1

Retailer1 Fill Rate

E[X] = q	S	E[T]	μ=E[X]E[T]	λ	ρ=λ/μ	C_a^2	C_T^2	$C_{T\sim}^2$	w_c	N_{ow}	N_{os}
8	26	3.00	24	5	0.20833	1	2.4	9.3	3.0565	122.259	4.5
$E[N_o]$	σ	E[B]	m	SumE[I]	MultiplierE[I]		E[I]	r	p	f (R1 Fill Rate)	
123.196	0.9983	14.65	3	363.984	0.0250		9.09	2	0.20471	0.79529	

Example 1

Retailer2 Fill Rate

E[X] = q	S	E[T]	μ=E[X]E[T]	λ	ρ=λ/μ	C_a^2	C_T^2	$C_{T\sim}^2$	w_c	N_{ow}	N_{os}
8	20	3.20	25.6	3	0.11719	1	2.4	9.3	3.2267	77.441	4.5
$E[N_o]$	σ	E[B]	m	SumE[I]	MultiplierE[I]		E[I]	r	p	f (R1 Fill Rate)	
77.968	0.9985	9.41	2	215.592	0.0142		3.07	4	0.11550	0.88450	

Example 1

Retailer3 Fill Rate

E[X] = q	S	E[T]	μ=E[X]E[T]	λ	ρ=λ/μ	C_a^2	C_T^2	$C_{T\sim}^2$	w_c	N_{ow}	N_{os}
8	32	4.20	33.6	2	0.05952	1	2.4	9.3	4.2097	67.355	4.5
$E[N_o]$	σ	E[B]	m	SumE[I]	MultiplierE[I]		E[I]	r	p	f (R1 Fill Rate)	
67.623	0.9991	8.19	3	538.169	0.0072		3.90	8	0.05894	0.94106	

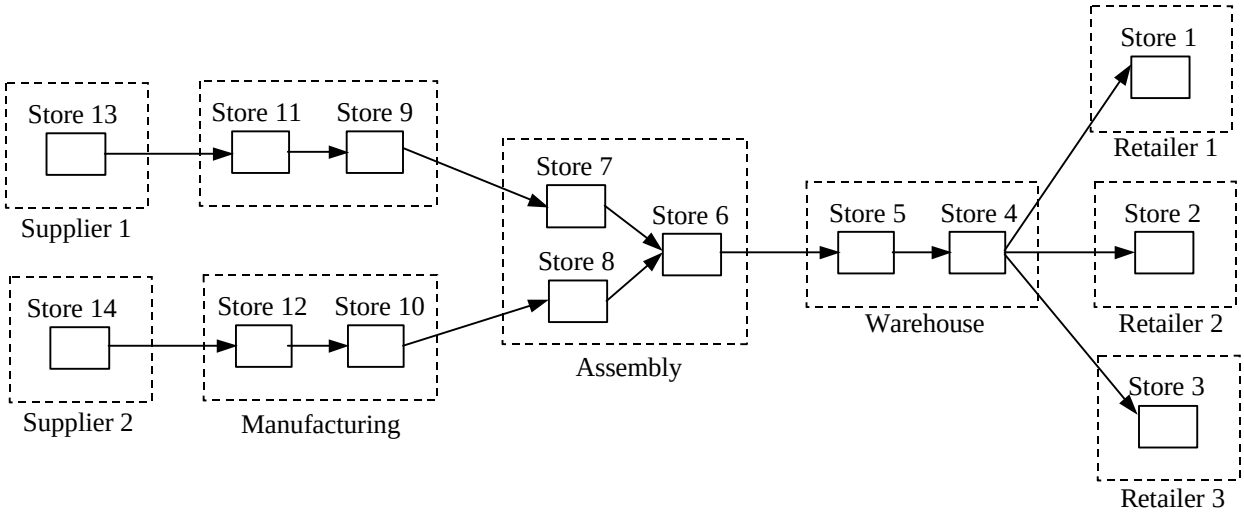

Figure 6.2 An example of an integrated supply chain network

The simulation model is written in STROBOSCOPE, which is a general-purpose discrete-event simulation language based on activity scanning and activity cycle diagrams (Martinez 1998). The input parameters for the simulation study are given as follows:

P1C1OrderAmount = 78; P1C2OrderAmount = 93; P1C3OrderAmount = 70;

BatchOfComp1 = 5; BatchOfComp2 = 5;

ReOrderPointOfRE11P1 = 6; ReOrderPointOfRE12P1 = 8; ReOrderPointOfRE13P1 = 10;

ROPW1P1 = 32; ROPP1Inv = 26;

ROPCM1Inv = 20; ROPCM2Inv = 8;

EOQW1RE11 = 20; EOQW1RE12 = 25; EOQW1RE13 = 28; EOQP1W1 = 60;

EPQP1Inv = 42; EOQCM1Inv = 25; EOQCM2Inv = 10;

BatchOfP1C1Trans = 9; BatchOfP1C2Trans = 15; BatchOfP1C3Trans = 20;

Each product consists of three type 1 components and two type 2 components. The transshipment lead times are assumed to be normal distributions.

The corresponding simulation results are given as follows (the confidence intervals are constructed from 50 independent replications):

Example 1

R1 Fill Rate

Ave	SD	90% Confidence Interval Low	90% Confidence Interval High
0.778	0.09	0.743	0.813

Example 1

R2 Fill Rate

Ave	SD	90% Confidence Interval Low	90% Confidence Interval High
0.928	0.057	0.907	0.950

Example 1

R3 Fill Rate

Ave	SD	90% Confidence Interval Low	90% Confidence Interval High
0.914	0.05	0.895	0.934

6.4.1.2 Example 2

By changing the reorder points of retailers and the number of orders per unit time, the performance of the supply chain network of example 2 can be calculated as follows:

Example 2

Retailer1 Fill Rate

E[X] = q	S	E[T]	μ=E[X]E[T]	λ	ρ=λ/μ	C_a^2	C_T^2	$C_{T\sim}^2$	w_c	N_{ow}	N_{os}
11	24	3.50	38.5	3	0.07792	1	2.4	9.21818	3.5112	115.870	6
$E[N_o]$	σ	E[B]	m	SumE[I]	MultiplierE[I]		E[I]	r	p	f (R1 Fill Rate)	
116.338	0.9993	10.37	2	305.872	0.0070		2.13	2	0.07728	0.92272	

Example 2

Retailer2 Fill Rate

E[X] = q	S	E[T]	μ=E[X]E[T]	λ	ρ=λ/μ	C_a^2	C_T^2	$C_{T\sim}^2$	w_c	N_{ow}	N_{os}
9	26	3.50	31.5	3	0.09524	1	2.4	9.26667	3.5172	94.963	5
$E[N_o]$	σ	E[B]	m	SumE[I]	MultiplierE[I]		E[I]	r	p	f (R1 Fill Rate)	
95.439	0.9990	10.29	2	356.284	0.0103		3.68	8	0.09420	0.90580	

Example 2

Retailer3 Fill Rate

E[X] = q	S	E[T]	μ=E[X]E[T]	λ	ρ=λ/μ	C_a^2	C_T^2	$C_{T\sim}^2$	w_c	N_{ow}	N_{os}
14	23	4.50	63	3	0.04762	1	2.4	9.17143	4.504	189.170	7.5
$E[N_o]$	σ	E[B]	m	SumE[I]	MultiplierE[I]		E[I]	r	p	f (R1 Fill Rate)	
189.527	0.9997	13.44	1	277.577	0.0034		0.94	9	0.04744	0.95256	

Similarly, the corresponding simulation results are given as follows:

Example 2

R1 Fill Rate

Ave	SD	90% Confidence Interval Low	90% Confidence Interval High
0.89	0.114	0.845	0.934

Example 2

R2 Fill Rate

Ave	SD	90% Confidence Interval Low	90% Confidence Interval High
0.934	0.042	0.918	0.951

Example 2

R3 Fill Rate

Ave	SD	90% Confidence Interval Low	90% Confidence Interval High
0.932	0.053	0.911	0.953

6.4.1.3 Comparisons

From the comparisons between results from the analytical performance evaluation model and those obtained from the simulation study, it can be seen that these results are close to each other. The differences between results are less than 5% (relative error). Therefore, the simulation results show that the proposed methodology is effective to supply chain network performance analysis.

Example 1

S	λ	Analytical Results	Simulation Results	Difference	Percentage
20	12	Retailer 1 Fill Rate 0.7953	Simulation for Retailer 1 0.778±0.09	Retailer 1 Fill Rate 0.0173	Retailer 1 Fill Rate 2.224%
25	7	Retailer 2 Fill Rate 0.8845	Simulation for Retailer 2 0.928±0.057	Retailer 2 Fill Rate 0.0435	Retailer 2 Fill Rate 4.688%
28	9	Retailer 3 Fill Rate 0.9411	Simulation for Retailer 3 0.914±0.05	Retailer 3 Fill Rate 0.0271	Retailer 3 Fill Rate 2.965%

Example 2

S	λ	Analytical Results	Simulation Results	Difference	Percentage
24	10	Retailer 1 Fill Rate 0.9227	Simulation for Retailer 1 0.89±0.114	Retailer 1 Fill Rate 0.0327	Retailer 1 Fill Rate 3.674%
26	9	Retailer 2 Fill Rate 0.9058	Simulation for Retailer 2 0.934±0.042	Retailer 2 Fill Rate 0.0282	Retailer 2 Fill Rate 3.019%
23	8	Retailer 3 Fill Rate 0.9526	Simulation for Retailer 3 0.932±0.053	Retailer 3 Fill Rate 0.0206	Retailer 3 Fill Rate 2.210%

6.4.2 Optimization Results

Consider the distribution system of the previous example, the fill rates of retailers are specified and used as constraints. In the following, we specify retailer 1's fill rate as 90%, retailer 2's fill rate as 90% and retailer 3's fill rate as 92%. In addition, assume that the inventory capitals per SKU at all stores are same, which is $10 per SKU. The following Tables give the optimization results:

Optimization Example
Retailer1 Fill Rate

E[X] = q	S	E[T]	μ=E[X]E[T]	λ	ρ=λ/μ	C_a^2	C_T^2	$C_{T\sim}^2$	w_c	N_{ow}	N_{os}
9.8539616	*40*	3.00	29.561885	3	0.10148	1	2.4	9.244	3.0196	89.264	5.42698
$E[N_o]$	σ	E[B]	m	SumE[I]	MultiplierE[I]		E[I]	r	p	f (R1 Fill Rate)	
89.815	0.99887	8.67	4	850.1101	0.0099		8.38	0.584	0.10000	0.90000	

Optimization Example
Retailer2 Fill Rate

E[X] = q	S	E[T]	μ=E[X]E[T]	λ	ρ=λ/μ	C_a^2	C_T^2	$C_{T\sim}^2$	w_c	N_{ow}	N_{os}
9.2462621	*36*	3.20	29.588039	3	0.10139	1	2.4	9.26	3.2196	89.307	5.12313
$E[N_o]$	σ	E[B]	m	SumE[I]	MultiplierE[I]		E[I]	r	p	f (R1 Fill Rate)	
89.826	0.99887	9.28	3	688.004	0.0105		7.25	8.261	0.10000	0.90000	

Optimization Example
Retailer3 Fill Rate

E[X] = q	S	E[T]	μ=E[X]E[T]	λ	ρ=λ/μ	C_a^2	C_T^2	$C_{T\sim}^2$	w_c	N_{ow}	N_{os}
8.2690953	*28*	4.50	37.210929	3	0.08062	1	2.4	9.29	4.5121	111.934	4.63455
$E[N_o]$	σ	E[B]	m	SumE[I]	MultiplierE[I]		E[I]	r	p	f (R1 Fill Rate)	
112.307	0.99928	13.28	3	412.9049	0.0096		3.95	3.193	0.08000	0.92000	

Rounding to the integers, the following optimal solutions can be obtained:

Retailer 1: q = 10 and S = 40;

Retailer 2: q = 9 and S = 36;

Retailer 3: q = 8 and S = 28; and

Total inventory capital = $3185.