Hindawi Publishing Corporation
Advances in Operations Research
Volume 2012, Article ID 126254, 34 pages
doi:10.1155/2012/126254

Review Article

Performance Evaluation of Stochastic Multi-Echelon Inventory Systems: A Survey

David Simchi-Levi[1] and Yao Zhao[2]

[1] *Engineering Systems Division, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2] *Department of Supply Chain Management and Marketing Sciences, Rutgers, The State University of New Jersy, Newark, NJ 07102, USA*

Correspondence should be addressed to Yao Zhao, yaozhao@andromeda.rutgers.edu

Received 31 August 2011; Accepted 3 November 2011

Academic Editor: Shangyao Yan

Globalization, product proliferation, and fast product innovation have significantly increased the complexities of supply chains in many industries. One of the most important advancements of supply chain management in recent years is the development of models and methodologies for controlling inventory in general supply networks under uncertainty and their widefspread applications to industry. These developments are based on three generic methods: the queueing-inventory method, the lead-time demand method and the flow-unit method. In this paper, we compare and contrast these methods by discussing their strengths and weaknesses, their differences and connections, and showing how to apply them systematically to characterize and evaluate various supply networks with different supply processes, inventory policies, and demand processes. Our objective is to forge links among research strands on different methods and various network topologies so as to develop unified methodologies.

1. Introduction

Many real-world supply chains, such as those found in automotive, electronics, and consumer packaged goods industries, consist of large-scale assembly and distribution operations with geographically dispersed facilities. Clearly, many of these supply chains support the production and distribution of multiple end-products which are assembled from hundreds or thousands of subsystems and components with widely varying lead times and costs.

One challenge in all these supply chains is the efficient management of inventory in a complex network of facilities and products with stochastic demand, random supply and high inventory and transportation costs. This requires one to specify the inventory policy for each

Table 1: Classification of the literature.

1	Single-period models or models with zero lead times	Models with positive lead times
2	Supply chains with capacity limits	Uncapacitated supply chains
3	Optimal policy characterization	Policy evaluation and optimization
4	Guaranteed service time models	Stochastic service time models

product at each facility so as to minimize the system-wide inventory cost subject to customer service requirements. For many years, both practitioners and academicians have recognized the potential benefit of effective inventory control in such networks. In fact, the literature on multi-echelon inventory control can be dated back to the 1950s. However, it is only in the last few years that some of these benefits have been realized, see, for example, Lee and Billington [1], Graves and Willems [2], and Lin et al. [3]. Three reasons have contributed to this trend:

(1) the availability of data, not only on network structure and bill of materials (BOMs), but also on demand processes, transportation lead times and manufacturing cycle times, and so forth;

(2) industry that is searching for scientific methods for inventory management that help to cope with long lead times and the increase in customer service expectations;

(3) recent developments in modeling and algorithms for the control of general structure multi-echelon inventory systems.

These developments are built on three generic methods: the queueing-inventory method, the lead-time demand method, and the flow-unit method. While the first two methods take a snapshot of the system and focus on quantities (e.g., backorders and on-hand inventory), the third method follows the movement of each flow unit and focuses on times (e.g., stockout delays and inventory holding times). This paper discusses the strengths and weaknesses of these methods, differences and connections among these methods, and demonstrates their abilities in handling various network topologies, inventory policies, and demand processes.

1.1. Classification of Literature

To position our survey in perspective, we classify the related literature by several dimensions (see Table 1).

Models with Zero Lead Times versus Models with Positive Lead Times

Models with zero lead times can be used to analyze strategic issues as well as tactical or operational issues when the lead times can be ignored, see, for example, the celebrated Newsvendor model [4] and some mathematical-programming-based models [5]. Models with positive lead times, such as the multi-echelon inventory models, explicitly consider lead times and even uncertain lead times.

Capacitated Supply Chains versus Uncapacitated Supply Chains

Supply chain models with limited production capacity received significant attention in the literature. We refer to Kapuscinski and Tayur [6] for a review of multistage single-product supply chains, to Sox et al. [7] for single stage multiproduct systems, and to Shapiro [5] for

mathematical programming models of production-inventory systems. In uncapacitated supply chains, we typically assume a positive exogenous "transit time" for processing a job, where the "transit time" is defined as the total time it takes from job inception to job completion. This transit time may represent manufacturing cycle time, transportation lead time, or warehouse receiving and processing times. The literature on uncapacitated supply chains can be further classified into two categories: i.i.d. or sequential transit time. In the former, the transit times are i.i.d. random variables; while, in the latter, the transit times are sequential in the sense that jobs are completed in the same sequence as they are released.

Optimal Policy Characterization versus Policy Evaluation and Optimization

The focus of the former is on identification and characterization of the structure of the optimal inventory policy. We refer to Federgruen [8], Zipkin [9], and Porteus [10] for excellent reviews. Unfortunately, the optimal policy is not known for general supply chains except for some special cases. When the optimal policy is unknown or known but too complex to implement, an alternative approach is to evaluate and optimize simple heuristic policies which are optimal in special cases but not in general.

Guaranteed Service Time Model versus Stochastic Service Time Model

In the former, it is assumed that in case of stockout, each stage has resources other than the on-hand inventory (such as slack capacity and expediting) to satisfy demand so that the committed service times can always be guaranteed. In the latter, it is assumed that in case of stockout, each stage fully backorders the unsatisfied demand and fills the demand until on-hand inventory becomes available. Thus, the delay due to stockout (i.e., the stockout delay) is random, and the committed service times cannot be 100% guaranteed. A recent comparison between the two models is provided by Graves and Willems [11].

1.2. The Scope and Objective of the Survey

This survey focuses on the stochastic service time model for uncapacitated supply chains. Because we are interested in general supply networks, we focus on policy evaluation and optimization. Given a certain class of simple but effective inventory policies, the specific problem that we address in this survey is how to characterize and evaluate system performance in general structure supply chains. The challenge arises from the fact that the inventory policy controlling one product at one facility may have an impact on all other products/facilities in the network either directly or indirectly.

For guaranteed service time models, Graves and Willems [11] summarize recent development and demonstrate its potential applications in industry-size problems. These developments are based on the lead-time demand method. For the stochastic service time model, Hadley and Whitin [12] provide the first comprehensive review for single-stage systems. Chen [13] reviews the lead-time demand method in serial supply chains, and de Kok and Fransoo [14] discuss some of its applications in more general supply chains. Song and Zipkin [15] provide an in-depth review of the literature on assembly systems, while Axsater [16] presents an excellent survey for serial and distribution systems. Zipkin [9] presents an excellent and comprehensive review for the queueing-inventory method in single-stage systems and the lead-time demand method in single-stage, serial, pure distribution, and pure assembly systems.

The objective of this paper is to compare the effectiveness of queueing-inventory method, the lead-time demand method, and the flow-unit method in supply chains along the following dimensions: network topology, inventory policy, and demand process. Specifically, we discuss how to apply each method systematically to evaluate various network topologies with either i.i.d. or sequential transit times, either base stock or batch ordering inventory policy, and either unit or batch demand process. The network topology considered includes single-stage (see Section 2), serial (Section 3.1), pure distribution (Section 3.2), pure assembly and 2-level general networks (Section 3.3), and tree and more general networks (Section 3.4). For each network topology, we discuss the three methods side by side and address questions such as, how are different stages connected and dependent? How does each method work? How are the results/methods connected to those of single-stage systems and systems of other topologies? What are the weakness and strength of each method? And what are the differences and connections among the methods? Some open questions are summarized in Section 4.

While some of the materials covered here appeared in previous reviews, we present these materials (together with recent results) in a coherent way by building connections among different methods and establishing uniform treatment of each method across different network topologies. We also shed some lights on the strengths and limitations of each method.

2. Single-Stage Systems

In this section, we consider single-stage systems and review the key assumptions and results of the three generic methods. We show how each method can handle different inventory policies, transit times, and demand processes. Following convention, we define a stage (a node, equivalently) to be a unique combination of a facility and a product, where the facility refers to a processor plus a storage where the latter carries inventory processed by the former.

Inventory Policies

In this paper, we focus on either continuous-review or periodic-review base-stock and batch ordering policies. For any stage in a supply chain, we define inventory position to be the sum of its on-hand inventory and outstanding orders subtracting backorders. Under continuous review, a base-stock policy with base-stock level s works as follows: whenever inventory position drops below s, order up to s. A batch-ordering policy with reorder point r and batch size Q works as follows: whenever the inventory position drops to or below the reorder point r, an order of size nQ is placed to raise the inventory position up to the smallest integer above r. Clearly, a base-stock policy is a special case of the batch-ordering policy with a batch size $Q = 1$. Continuous-review base-stock policies are often used for expensive products facing low-volume but highly uncertain demand (e.g., service parts). Batch-ordering policies are often used where economies of scale in production and transportation cannot be ignored (commodities).

Under periodic review, the base-stock and batch ordering policies work in similar ways as their continuous-review counterparts except that inventory is reviewed only once in one period. The sequence of events is as follows [12]. At the beginning of a review period, the replenishment is received, the inventory is reviewed, and then an order decision is made. Demand arrives during the period. At the end of the period, costs are calculated. Some work in the literature assumes that all demands arrive at the end of the period; see, for example,

Zipkin [9, Chapter 9]. Under this assumption, a single-stage periodic-review inventory system can be viewed as a special case of its continuous-review counterpart with constant demand interarrival times and batch demand sizes. In this survey, we assume demand arrives during the period unless otherwise mentioned.

Transit Times

If the transit times (Section 1.1) are sequential and stochastic, namely, "stochastic sequential transit times," then they must be dependent over consecutive orders. Kaplan [17] presents a discrete-time model for the stochastic sequential transit time in a periodic-review single-stage system, where the evolution of the outstanding order vector is modeled by a Markov chain. See Song and Zipkin [18] for a generalization of the model. For continuous-review single-stage systems, Zipkin [19] presents a continuous-time model for stochastic and sequential transit times.

Definition 2.1. The exogenous, stochastic, and sequential transit times are defined as follows: there exits an exogenous continuous-time stochastic process $\{U(t)\}$ that is stationary and ergodic with finite limiting moments, such that the sample path of $\{U(t)\}$ is left-continuous, the transit time at t, $L(t) = U(t)$, and $t + L(t)$ is nondecreasing.

Svoronos and Zipkin [20] apply this model to multistage supply chain with two additional assumptions: (1) the transit times are independent of the system state, for example, demand and order placement and (2) the transit times are independent across stages.

In practice, the transit times can be either parallel or sequential or somewhere in between. Many production and transportation processes in the real world are subject to random exogenous events. Indeed, the orders placed by the systems under consideration may be a negligible portion of their total workload. Thus, the transit times are *exogenous* and should be estimated from data. While in some practical cases, the sequential transit time model may be more realistic than the i.i.d. transit time model [20], in cases such as repairing and maintenance, the i.i.d. transit time model may be a better approximation [21].

Demand Processes

Both unit demand and batch demand processes are studied in the literature. On arrival of a batch demand, one shall address questions such as: should all units of the demand be satisfied together (unsplit demand)? Or should each demand unit be satisfied separately (split demand)? For a supply system (either production or transportation) processing a job of multiple units, one needs to address questions like: is the job processed and replenished as an individable entity (unsplit supply)? Or is each unit processed and replenished separately (split supply)? If the former is true, does the transit time depend on job size? See Zipkin [9] for more discussions on these questions. While the case of split demand is easier to handle and thus widely studied in the literature, the case of unsplit demand is much more difficult; see Section 2.1 for more details.

The Basic Assumption

For the ease of exposition, we make the following assumption throughout the survey unless otherwise mentioned.

Assumption 2.2. The system is under continuous review; unsatisfied demands are fully backordered; outside suppliers have ample stock; the transit times are exogenous either i.i.d. or sequential; demand is satisfied on a first-come first-serve (FCFS) basis; demand can be split; supply cannot be split; transit times do not depend on job sizes.

Throughout the survey, we use the following notations: $a^+ = \max\{a, 0\}$, $a^- = \max\{-a, 0\}$. $E(\cdot), V(\cdot)$ are the mean and variance of a random variable, respectively. If random variables X and Y are independent, we denote $X \perp Y$. We consider base-stock policies with $s \geq 0$ and batch-ordering policies with $r \geq 0$ unless otherwise mentioned.

We define the basic model for single-stage systems as follows: inventory is controlled by a base-stock policy, demand follows Poisson process with rate λ, and the transit time (i.e., lead time) L is constant. In the following subsections, we first discuss the methods in the basic model and then extend the results to more general demand process, inventory policies, and supply process.

2.1. The Queueing-Inventory Method

Let $\{IO(t), t \geq 0\}$ be the outstanding order process, $\{IP(t), t \geq 0\}$ the inventory position process, and $\{IL(t), t \geq 0\}$ the process of net inventory (on-hand minus backorder). Define $\{I(t), t \geq 0\}$ ($\{B(t), t \geq 0\}$) to be the process of on-hand inventory (backorder, resp.). For appropriate initial conditions, the following equations hold under Assumption 2.2,

$$IO(t) + IL(t) = IP(t), \quad t \geq 0, \tag{2.1}$$

$$I(t) = IL(t)^+, \tag{2.2}$$

$$B(t) = IL(t)^-. \tag{2.3}$$

For unsplit demand, (2.2)-(2.3) do not hold since $I(t) > 0$ and $B(t) > 0$ can hold simultaneously.

Note that $IO(t)$ is the number of jobs in the supply process. The queueing-inventory method characterizes the probability distribution of $IO(t)$ by identifying the appropriate queueing analogue. One can follow a 3-step procedure to characterize the system performance: (1) the distribution of $IP(t)$, (2) the distribution of $IO(t)$, and (3) the dependence of $IO(t)$ and $IP(t)$. We focus on steady-state analysis and define $IO = \lim_{t\to\infty} IO(t)$. The same notational rule applies to IL, IP, and I and B.

Clearly, $IP = s$ for base-stock policies. For batch-ordering policies, the distribution of IP only depends on the demand process. IP is uniformly distributed in $\{r+1, r+2, \ldots, r+Q\}$ for renewal batch demand under mild regularity assumptions [22]. See Zipkin [19] for a discussion of more general demand processes. The distribution of IO depends on the demand process, the inventory policy, and the supply system (see discussions below). For batch-ordering policies, IP depends on IO. Intuitively, the lower the IP, the longer the time since the last order, and therefore the lower the IO.

i.i.d. Transit Time

Consider first the basic model with constant L, the queueing analogue is a $M/D/\infty$ queue. By Palm theorem [23], IO follows Poisson (λL) distribution. If L is stochastic, then the queueing

analogue is a $M/G/\infty$ queue and IO follows Poisson $(\lambda E(L))$ distribution. Because demand is satisfied on a FCFS basis, the stockout delay differs from L even at $s = 0$; see Muckstadt [24, page 96] for an exact analysis. For renewal unit demand, the queueing analogue is a $G/G/\infty$ queue. For compound Poisson demand, then the queueing analogue is a $M^Y/G/\infty$ queue where $\{Y_n\}$ is the demand size process. The distribution of IO is compound Poisson under Assumption 2.2.

Consider the basic model but with a batch ordering policy, the queueing analogue is a $Er^Q/D/\infty$ queue where Er stands for Erlang interarrival times. See Galliher et al. [25] for an exact analysis. For batch demand processes, tractable approximations become appealing. One can first assume $IP \perp IO$ and then approximate the distribution of IO by results from systems with base-stock policy and batch-demand processes [9, Section 7.2.4].

Sequential Transit Time

Consider the basic model with sequential transit times (Definition 2.1). Let $D(t_1, t_2]$ be the demand during time interval $(t_1, t_2]$, where $t_1 \leq t_2$, and let $D(\infty \mid L] = \lim_{t\to\infty} D(t - L, t]$. By Svoronos and Zipkin [20]:

Proposition 2.3. *IO has the same distribution as $D(\infty \mid L]$.*

Proof. See the appendix for a proof. □

D$(t - L, t]$ $(D(\infty \mid L])$ is called the lead-time demand. If demand follows compound Poisson process, Proposition 2.3 also holds under Assumption 2.2.

For the basic model with constant transit time, one can obtain Proposition 2.3 by an alternative approach [9]. At time t, because all orders placed on or before $t-L$ are replenished while all orders placed after $t-L$ are still in transit, $IO(t)$ equals to the number of orders placed during $(t-L, t]$. Due to the Poisson demand and the continuous-review base-stock policy, one must have

$$IO(t) = D(t - L, t]. \tag{2.4}$$

Consider the batch ordering policy in the basic model with sequential transit times (Definition 2.1). Equation (2.4) does not hold because $IO(t)$ is clearly not the demand during $(t - L, t]$. In addition, $IO(t)$ depends on $IP(t)$. Exact analysis of these systems using the queueing-inventory method is rare. Fortunately, such systems can be easily handled by the lead-time demand method and the flow-unit method.

2.2. The Lead-Time Demand Method

Consider the basic model. Observe that at time t, the system receives all orders placed on or before $t - L$ but none of the orders placed after $t - L$, then

$$IL(t) = IP(t - L) - D(t - L, t]. \tag{2.5}$$

Equations (2.2)-(2.3) remain true here. Although (2.5) looks quite similar to (2.1) and (2.4), they follow completely different logic. Indeed, IL and IP are measured at different

times (t or $t-L$) in the lead-time demand method rather than the same time (t) in the queueing-inventory method.

Let $\{IP(t_n)\}$ be the embedded discrete time Markov chain (DTMC) formed by observing $IP(t)$ right after each ordering decision (at t_n). Zipkin [19] shows the following.

Proposition 2.4. *Consider a single-stage system. If* (i) *the inventory policy depends only on inventory position,* (ii) *the demand sizes are i.i.d. random variables independent of the arrival epochs,* (iii) $\{IP(t_n), n \geq 0\}$ *is irreducible, aperiodic, and positive recurrent,* (iv) *the arrival epochs form a counting process which is either stationary or converges to a stationary process in distribution as* $t \to \infty$, *and* (v) *the transit times are sequential and exogenous (Definition 2.1), then*

(1) IP *has the same distribution as* $\{IP(t_n)\}$ *as* $n \to \infty$,

(2) $IL = IP - D(\infty \mid L]$,

(3) $IP \perp D(\infty \mid L)$.

The inventory policy includes the batch-ordering policy and the (s, S) policy, and the demand process includes renewal batch process and the superposition of independent renewal batch processes [19]. We point out that for (2.5) and Proposition 2.4 to hold, the assumptions of sequential transit time, FCFS rule, and split demand are necessary.

In the basic model, the stockout delay, X, for a demand at t, satisfies [26]

$$\Pr\{X \leq x\} = \Pr\{D(t-L+x, t) < s\}, \quad \text{for } 0 \leq x \leq L. \tag{2.6}$$

To see this, note that, at $t+x$, all orders triggered by demand on or prior to $t+x-L$ are replenished. Because the demand at t has priority over demand after t, the demand at t is satisfied on or before $t+x$ if and only if the orders triggered by demand during $(t+x-L, t)$ are less than s. By the same logic, for compound Poisson demand, the stockout delay for the kth unit of a demand, $X(k)$, is given by

$$\Pr\{X(k) \leq x\} = \Pr\{D(t-L+x, t) \leq s-k\}, \quad \text{for } 0 \leq x \leq L. \tag{2.7}$$

Consider now the basic model under periodic review. Let $IP(n)$ be the inventory position at the beginning of period n after order decision is made and $IL(n)$ ($I(n)$ and $B(n)$) the net inventory (inventory on-hand and backorder) at the end of period n after demand is realized. Let L here be an integer multiple of a review period and $D[n, m]$ the demand from period n to m inclusive. According to the sequence of events (see beginning of Section 2), (2.5) and (2.2)-(2.3) become $IL(n) = IP(n-L) - D[n-L, n]$, $I(n) = IL(n)^+$, and $B(n) = IL(n)^-$, respectively. By Hausman et al. [27], for $x \leq L$,

$$\Pr\{\text{all demand in period } n \text{ is satisfied within } x \text{ periods}\} = \Pr\{D[n-L+x, n] \leq s\}. \tag{2.8}$$

2.3. The Flow-Unit Method

For the basic model, suppose a demand arrives at time t, then the order triggered by this demand will satisfy the sth demand after t [28, 29]. Alternatively, the corresponding order

that satisfies the demand at time t is placed at $t - T(s)$, where $T(s)$ is determined by starting at time t, counting backwards until the number of demand arrivals reaches s [30]. We call the former the "forward method" because, for each order, it looks forward to identify the corresponding demand. We call the latter the "backward method" because, for each demand, it looks backward to identify the corresponding order.

Both methods yield the same result for single-stage systems. For general networks, the two methods may take different angles, and thus one can be more convenient than the other (Section 3). We focus on the backward method unless otherwise mentioned. The stockout delay, X, for the demand at time t and the holding time, W, for the product that satisfies this demand are given by

$$X = (L - T(s))^+, \tag{2.9}$$

$$W = (T(s) - L)^+. \tag{2.10}$$

Unlike the queueing-inventory method and the lead-time demand method, the flow-unit method focuses on the stockout delay (the inventory holding time) associated with each demand (product) rather than the on-hand inventory and backorders at a certain time. Equations (2.9)-(2.10) hold also for stochastic sequential lead times (Definition 2.1) and for any point unit-demand process [31]. We should point out that the assumptions of sequential lead time and FCFS rule are necessary for (2.9)-(2.10). By (2.9), the distribution of the stockout delay, X, is given by,

$$\Pr\{X \le x\} = \Pr\{L - T(s) \le x\}, \quad \text{for } 0 \le x \le L. \tag{2.11}$$

For compound Poisson demand, different units in one demand face statistically different stockout delays [29]. Consider the kth unit of a demand at t, the backorder delay, $X(k)$, and the inventory holding time, $W(k)$, for the corresponding item that satisfies this unit are

$$X(k) = [L - T(J(k))]^+, \tag{2.12}$$

$$W(k) = [T(J(k)) - L]^+, \tag{2.13}$$

where $J(k)$ is obtained by starting at time t, counting backwards demand arrivals until the cumulative demand becomes greater than $s - k$ in the first time. See Forsberg [32] and Zhao [33] for extended discussions.

A comparison between (2.6)-(2.7) and (2.11)-(2.12) demonstrates the connections between the lead-time demand method and the flow-unit method. Because $D(t - L, t)$ is the cumulative demand and $T(s)$ is the sum of interarrival times, the event $\{T(s) \ge L - x\}$ is equivalent to the event $\{D(t - L + x, t) < s\}$ for unit demand [34, page 406]. Similarly, the event $\{T(J(k)) \ge L - x\}$ is equivalent to the event $\{D(t - L + x, t) \le s - k\}$ for batch demand.

For the basic model under periodic review, if demand arrives at the end of each period, then the system is a special case of its continuous-review counterpart [33]. If demand arrives

during a period, the flow-unit method also applies, see, for example, Axsater [35]. For the basic model with batch ordering policy, by Axsater [36],

$$\begin{aligned} X &= (L - T(S))^+, \\ W &= (T(S) - L)^+, \end{aligned} \tag{2.14}$$

where S is a random integer uniformly distributed in $\{r+1, r+2, \ldots, r+Q\}$. See also Zhao and Simchi-Levi [30]. For the basic model with both batch ordering policy and compound Poisson demand, the analysis is more involved but still tractable, see Axsater [37].

3. Multistage Supply Chains

Multistage supply chains differ from single-stage systems because the lead time at one stage depends on other stages' stock levels. For a stage, the lead time is the total time needed from order placement to order delivery. Clearly, lead times include but are not limited to the "transit times."

Notation 1. Consider a supply chain under Assumption 2.2 with node set $\mathcal{N}$ and arc set $\mathcal{A}$. An arc refers to a pair of nodes with direct supply-demand relationship. We define the following.

(i) $\{IO_j(t), t \geq 0\}$: the outstanding order process at node $j \in \mathcal{N}$.

(ii) $\{IP_j(t), t \geq 0\}$: the inventory position process at node j.

(iii) $\{IL_j(t), t \geq 0\}$: the net inventory (on-hand minus backorder) process at node j.

(iv) $\{I_j(t), t \geq 0\}$ ($\{B_j(t), t \geq 0\}$): the process of on-hand inventory (backorder) at node j.

(v) L_j ($L_{i,j}$): the processing cycle time at node j (transportation lead time over arc $(i,j) \in \mathcal{A}$).

(vi) IT_j ($IT_{i,j}$): the inventory in-transit during L_j (during $L_{i,j}$).

(vii) $\tilde{L}_j$: the total replenishment lead time at node j.

(viii) X_j (W_j): the stockout delay (inventory holding time) at node j.

(ix) τ_j (α_j, β_j): the committed service time (target type 1, 2 service) at node j.

(x) $a_{i,j}$: the BOM structure, that is, one unit at node j requires $a_{i,j}$ unit(s) from node i.

(xi) h_j (π_j): the inventory holding cost (penalty cost) per unit item per unit time at node j.

(xii) s_j (r_j, Q_j): base-stock level (reorder point, batch size) at node j.

3.1. Serial Systems

In this section, we extend the methodologies and results of the single-stage systems to a serial supply chain where nodes $j \in \mathcal{J}$ are numbered by $1, 2, \ldots, |\mathcal{J}|$. Node $|\mathcal{J}|$ receives external supply, node $j+1$ supplies node j, and node 1 supplies external demand. The transit time of node $|\mathcal{J}|$ is $L_{|\mathcal{J}|}$, and the transit time between stage $j+1$ and j is L_j. This system can be controlled either by an installation policy or an echelon policy. For an installation policy, the notation is defined as above. For an echelon policy, we need the following notation.

(i) IP_j^e: the echelon inventory position at stage j, which is the sum of inventory on-hand and on-order at stage j plus inventory on-hand and in-transit at all downstream stages of j subtracting B_1.

(ii) $IL_j^e = IP_j^e - IO_j$: the echelon net inventory at stage j.

(iii) $I_j^e = IL_j^e + B_1$: the echelon on-hand inventory.

(iv) $IT_j^e = IT_j + IL_j^e$: the echelon inventory in-transit.

(v) s_j^e (r_j^e): the echelon base-stock level (reorder point).

An echelon batch-ordering policy works as follows: whenever IP_j^e drops to or below r_j^e, an order of size nQ_j is placed to raise the echelon inventory position up to the smallest integer above r_j^e. According to convention, we assume that Q_{j+1} and r_{j+1} are integer multiples of Q_j for all j.

We define the basic model for serial systems as follows: each stage controls its inventory by an installation base-stock policy; external demand follows Poisson process; the transit times are constant, and $a_{j+1,j} = 1$, for all j. We focus on the penalty cost model and refer to Boyaci and Gallego [38] and Shang and Song [39] for discussions on the service constraint model.

Echelon Policies versus Installation Policies

The echelon policies (base-stock or batch ordering) are equivalent to their installation counterparts under certain conditions. According to Axsater and Rosling [40], two policies are equivalent if given identical initial conditions, the two policies share the same sample path for their inventory positions at all stages of the supply chain for any external demand sequence.

For serial systems under either continuous review or periodic review with identical periods, one can construct an equivalent echelon batch-ordering policy for each installation batch-ordering policy by setting $r_1^e = r_1$; $r_{j+1}^e = r_j^e + Q_j + r_{j+1}$, $j = 1, 2, \ldots, |\mathcal{J}| - 1$. The initial conditions are $r_j < I_j(0) \leq r_j + Q_j$, and $I_j(0) - r_j$ is an integer multiple of Q_{j-1}.

For an echelon policy, one may not always find an equivalent installation policy unless the echelon policy is nested: stage $j + 1$ orders only when stage j orders for each j. The initial condition is $r_j^e < I_j^e(0) \leq r_j^e + Q_j$. The result on batch ordering policies remain valid in pure assembly systems but not in distribution systems. Indeed, Axsater and Juntti [41] compare numerically the performance of echelon and installation batch ordering policies in a pure distribution system with Poisson demand and show that either policies can outperform the other and the difference is up to 5%.

Joint Distribution of Inventory Positions

Consider a continuous-review serial system with installation batch-ordering policies and compound Poisson demand, the inventory position vector $\overline{IP}(t) = (IP_j(t), j \in \mathcal{J})$ forms a continuous-time Markov chain (CTMC) with state space $\mathcal{S} = \otimes_{j\in\mathcal{J}}\{r_j + Q_{j-1}, r_j + 2Q_{j-1}, \ldots, r_j + Q_j\}$ where $Q_0 = 1$. We focus on 3 questions: (1) what is the marginal distribution of IP at each stage? (2) When are the IPs independent across stages? (3) What is the distribution of IP seen by an order placed by a downstream stage?

Proposition 3.1. *If the CTMC of $\overline{IP}(t)$ is irreducible and aperiodic, then as $t \to \infty$*

(1) *$\overline{IP}(t)$ is uniformly distributed in $\mathcal{S}$.*

(2) *The inventory positions are independent across different stages.*

(3) *Each order of stage j sees IP_{j+1} in its time averages.*

Proof. See the appendix for a proof. □

A sufficient condition for $\overline{IP}$ to be irreducible and aperiodic is that the external demand can equal 1. For a serial supply chain with echelon batch-ordering policies, the inventory position vector has a state space $\mathcal{S}^e = \otimes_{j\in\mathcal{J}}\{r_j^e + 1, r_j^e + 2, \ldots, r_j^e + Q_j\}$. Because inventory positions at different stages are driven by a common demand process, they may not be independent. Proposition 3.1 does not hold here because the CTMC of $\overline{IP^e}(t)$ may be reducible and depends on initial conditions, see Axsater [42]. Fortunately, if one assumes randomized initial conditions, then $\overline{IP^e}$ is uniformly distributed in $\mathcal{S}^e$ [43]. So far, the only result on non-Markovian demand process is that Proposition 3.1 holds for renewal unit external demand. See Section 3.2 for more discussions.

3.1.1. The Queueing-Inventory Method

Consider the basic model. Applying (2.1) to each stage, $IP_j(t) = IO_j(t) + IL_j(t)$, $j \in \mathcal{J}$. Define $B_{|\mathcal{J}|+1}(t) \equiv 0$. Because $IO_j(t) = B_{j+1}(t) + IT_j(t)$, for all j, we must have

$$IP_j(t) \;=\; B_{j+1}(t) + IT_j(t) + IL_j(t), \quad j \in \mathcal{J}. \tag{3.1}$$

That is, the inventory position at stage j consists of three elements: backorders at stage $j+1$, inventory in-transit from stage $j+1$ to j, and net inventory at stage j. By (3.1) and (2.3),

$$B_j(t) \;=\; \left(B_{j+1}(t) + IT_j(t) - IP_j(t)\right)^+, \quad j \in \mathcal{J}. \tag{3.2}$$

Note that $IP_j(t)$ is not independent of $B_{j+1}(t)$ in general. Equations (3.1)-(3.2) hold for any serial system under Assumption 2.2 and extend to periodic-review systems [44]. The queueing-inventory method focuses on characterizing IO_j and IT_j for each stage.

i.i.d. Transit Time

Consider the basic model with i.i.d. transit times. Other than the special case of $s_j = 0$, for all $j \neq 1$, where the system forms a Jackson network with mutually independent $IT_j(t)$, the serial system poses a substantial challenge for exact analysis under the queueing-inventory method because IT_j depends on B_{j+1}. An exact analysis is unknown [9]. Various approximations are proposed, see discussions of the distribution systems (Section 3.2.1).

Sequential Transit Time

For the basic model with stochastic sequential transit times (Definition 2.1), the analysis here is a special case of those of pure distribution systems. We postpone the discussion to Section 3.2.1. For batch ordering systems, the exact analysis by the queueing-inventory

method is difficult because B_{j+1} (and thus IO_j) is not independent of IP_j. Fortunately, such systems can be easily handled by the lead-time demand method and the flow-unit method.

3.1.2. The Lead-Time Demand Method

Consider the basic model with sequential transit times (Definition 2.1). We discuss both installation and echelon policies. Extensions to compound Poisson demand is straightforward.

Installation Policies

By (3.1), $IP_j(t-L_j) = B_{j+1}(t-L_j) + IT_j(t-L_j) + IL_j(t-L_j)$, for all j. By the lead-time demand method, at time t, all outstanding orders except $B_{j+1}(t-L_j)$ will be available at stage j. Therefore,

$$IL_j(t) = IP_j(t-L_j) - B_{j+1}(t-L_j) - D(t-L_j,t], \quad j \in \mathcal{J}. \tag{3.3}$$

Equation (3.3) is similar to (2.5) in single-stage systems. The difference is that here only part of $IO_j(t-L_j)$, that is, $IT_j(t-L_j)$, is available at t. For base-stock policies, $IP_j(t) \equiv s_j$. Equation (3.3) implies the following recursive equations for B_j in steady-state:

$$B_j = \left(B_{j+1} + D(\infty \mid L_j] - s_j\right)^+, \quad j \in \mathcal{J}, \tag{3.4}$$

where $D(\infty \mid L_j]$s are mutually independent. We refer the reader to Van Houtum and Zijm [45, 46], Chen and Zheng [47], and Gallego and Zipkin [48] for extended discussions. By (3.4), a serial supply chain can be decomposed into $|\mathcal{J}|$ single-stage systems where one can characterize B_j from $j = |\mathcal{J}|$ to $j = 1$ consecutively.

Extension to batch-ordering policy is not straightforward because IP_j depends on B_{j+1}. See Badinelli [49] for an exact analysis of systems with Poisson demand and constant lead times. Indeed, echelon policies are easier to handle using the lead-time demand method.

Echelon Policies

First consider echelon base-stock policies. By (2.5),

$$\begin{aligned} IL^e_{|\mathcal{J}|}(t) &= s^e_{|\mathcal{J}|} - D(t-L_{|\mathcal{J}|},t], \\ IL^e_j(t) &= IT^e_j(t-L_j) - D(t-L_j,t] \\ &= \min\left\{IL^e_{j+1}(t-L_j), s^e_j\right\} - D(t-L_j,t], \quad j = 1,2,\ldots,|\mathcal{J}|-1. \end{aligned} \tag{3.5}$$

In steady-state, $IL^e_{|\mathcal{J}|} = s^e_{|\mathcal{J}|} - D(\infty \mid L_{|\mathcal{J}|}]$, $IL^e_j = \min\{IL^e_{j+1}, s^e_j\} - D(\infty \mid L_j]$, $j = 1,2,\ldots,|\mathcal{J}|-1$, where the $D(\infty \mid L_j]$s are mutually independent. Equation (3.5) can be extended to periodic-review systems [44, 47].

Next, we consider batch-ordering policies. For the most upstream stage, $IL^e_{|\mathcal{J}|}(t) = IP^e_{|\mathcal{J}|}(t-L_{|\mathcal{J}|}) - D(t-L_{|\mathcal{J}|},t]$. Given $IL^e_{j+1}(t)$, $IT^e_j(t)$ is uniquely determined as follows [44]: if $IL^e_{j+1}(t) \le r^e_j$, then $IT^e_j(t) = IL^e_{j+1}(t)$; otherwise, $IT^e_j(t) > r^e_j$. Because $IT^e_j(t) \le r^e_j + Q_j$ and

$IL^e_{j+1}(t) - IT^e_j(t)$ must be an integer multiple of Q_j, $IT^e_j(t) = IL^e_{j+1}(t) - mQ_j$, where m is the largest integer so that $IL^e_{j+1}(t) - mQ_j > r^e_j$. Define $IT^e_j(\mathrm{t}) = \varpi(IL^e_{j+1}(t), Q_j)$. In steady-state,

$$\begin{aligned} IL^e_{|\mathcal{J}|} &= IP^e_{|\mathcal{J}|} - D(\infty \mid L_{|\mathcal{J}|}], \\ IL^e_j &= \varpi\left(IL^e_{j+1}, Q_j\right) - D(\infty \mid L_j], \quad j = 1, 2, \ldots, |\mathcal{J}| - 1. \end{aligned} \tag{3.6}$$

Here, $IP^e_{|\mathcal{J}|}$ is uniformly distributed in $\{r^e_{|\mathcal{J}|} + 1, \ldots, r^e_{|\mathcal{J}|} + Q_{|\mathcal{J}|}\}$, $D(\infty \mid L_j]$s are mutually independent and independent of IT^e_js. See Chen and Zheng [44] and Chen [50] for more discussions.

Approximations and Bounds

Policy evaluation based on the exact analysis can be time consuming. One can compute the system performance approximately but fast using two-moment approximations. For instance, one can compute (3.4) by fitting a negative binomial or Gamma distribution to the lead-time demand utilizing the first two moments [20, 51]. Equation (3.4) can also be regarded as incomplete convolutions of the form $(X_1 - a)^+ + X_2$. Van Houtum and Zijm [45, 46] fit the incomplete convolutions by mixed Erlang or hyperexponential distributions.

An alternative approach is to develop bounds. The "Restriction-Decomposition" heuristic [48] is based on the observation that by (3.3)-(3.4), $I_j \le (s_j - D(\infty \mid L_j))^+$ and $B_1 \le B_2 + (D(\infty \mid L_1 - s_1)^+ \le \cdots \le \sum_{j\in\mathcal{J}}(D(\infty \mid L_j) - s_j)^+$. Thus, the system total cost $TC = \sum_{j\in\mathcal{J}} h_j I_j + \pi_1 B_1 \le \sum_{j\in\mathcal{J}}[h_j(s_j - D(\infty \mid L_j))^+ + \pi_1(D(\infty \mid L_j) - s_j)^+]$. The latter is the sum of single-stage cost functions. One can then choose the base-stock levels that optimize the bound.

Shang and Song [52] develop Newsvendor types of close-form bounds and approximations for the optimal base-stock levels. The key idea is to construct a subsystem for each stage that includes itself and its downstream stages then replace the installation holding costs at all stages of the subsystem by either a upper or a lower bound. Such a subsystem effectively collapses into a single-stage system, for which one can use the newsboy model. For batch-ordering policies, Chen and Zheng [53] develop lower and upper bounds for the total cost by either under- or overcharging a penalty cost for each stage. The resulting bounds are sums of $|\mathcal{J}|$ many single-stage cost functions.

Finally, we mention that the performance gap between echelon and installation policies may be minor. Chen [50] compares the best echelon policy with the best installation policy in serial systems. For different number of stages, lead times, batch sizes, demand variabilities, and holding/penalty costs, it is shown, in a numerical study, that the % difference of their performance (based on the optimal cost of echelon policies) range from 0% to 9% with an average 1.75%.

3.1.3. The Flow-Unit Method

The flow-unit method provides an exact analysis for the basic model with either Poisson or compound Poisson demand. Because the analysis here is a special case of that of pure distribution systems, we postpone the discussion to Section 3.2.3. In the basic model with installation batch ordering policy, applying (2.14) to each $j \in \mathcal{J}$ yields, $X_j = (X_{j+1} + L_j - T_j(S_j))^+$

and $W_j = (T_j(S_j) - X_{j+1} - L_j)^+$, where S_j is uniformly distributed in $\{r_j + Q_{j-1}, r_j + 2Q_{j-1}, \ldots, r_j + Q_j\}$ and $S_j, j \in \mathcal{J}$ are independent (Proposition 3.1). Furthermore, $T_j(\cdot)$s are not overlapping, and therefore $T_j(S_j),\ j \in \mathcal{J}$ are mutually independent. Consequently, a serial system can be decomposed into multiple single-stage systems as in Section 3.1.2.

The flow-unit method can also be applied to serial systems with echelon batch ordering policy [42] or base-stock policy under periodic review [32, 33, 41]. We postpone the discussion to distribution systems (Section 3.2.3).

3.2. Pure Distribution

In this section, we focus on 2-level pure distribution systems (distribution systems, for brevity), where node 0, the distribution center (DC), is the unique supplier for nodes $j \in \mathcal{J}$ (the retailers) that face external demand. The transit time of node 0 is L_0, and the transit time between stage 0 and j is L_j. Distribution systems are more complex than serial systems because (i) the demand process faced by the DC is a superposition of the order processes of all retailers and (ii) DC needs to allocate inventory among retailers in case of shortages. In this section, we focus on installation policies and FCFS rule unless otherwise mentioned.

Redefine $\mathcal{S} = \otimes_{j\in\{0\}\bigcup\mathcal{J}}\{r_j + \delta_j, r_j + 2\delta_j, \ldots, r_j + Q_j\}$, where $\delta_j = 1$, for all $j \in \mathcal{J}$, and δ_0 is the maximum common factor of $Q_j,\ j \in \mathcal{J}$, by the proof of Proposition 3.1, see also [54].

Corollary 3.2. *Proposition 3.1 holds for the inventory position vector of the DC and all retailers.*

For demand under non-Markovian assumptions, Cheung and Hausman [55] show that if external demand follows independent renewal unit processes, then the first two statements of Proposition 3.1 hold for the inventory position vector of the DC and all retailers.

We define the basic model for distribution systems as follows: each stage utilizes an installation base-stock policy, external demand follow independent Poisson processes with rates $\lambda_j,\ j \in \mathcal{J},\ L_j,\ j \in \{0\}\bigcup\mathcal{J}$ are constant, and $a_{0,j} = 1$, for all j. No lateral transshipment is allowed.

3.2.1. The Queueing-Inventory Method

By (2.1), $IO_j(t) + IL_j(t) = IP_j(t)$ holds for $j \in \{0\}\bigcup\mathcal{J}$ under Assumption 2.2. Because $IO_j(t) = B_{0,j}(t) + IT_j(t)$, for all $j \in \mathcal{J}$, and $B_{0,j}(t)$ is the orders placed by stage j backlogged at stage 0,

$$B_0 = \sum_{j\in\mathcal{J}} B_{0,j}(t), \tag{3.7}$$

$$IP_j(t) = B_{0,j}(t) + IT_j(t) + IL_j(t). \tag{3.8}$$

For the basic model, conditioning on $B_0 = b$, $B_{0,j}$ follows a binomial distribution with b number of trials and a successful rate of $\lambda_j / \sum_{l\in\mathcal{J}}\lambda_l$ per trial (the "binomial decomposition," [51, 56]). This is true because the probability that an order received by the DC is placed by retailer j is $\lambda_j \setminus \sum_{l\in\mathcal{J}}\lambda_l$, and each order is independent of the others. This result holds as long as external demand follows independent Poisson processes, retailers utilize continuous-review base-stock policy, and DC serves retailers' orders on a FCFS basis. For compound Poisson demand or batch ordering policy, it is much more involved to decompose B_0 into $B_{0,j}$, see Shanker [57] and Chen and Zheng [43].

i.i.d. Transit Time

Consider the basic model. Similar to serial systems (Section 3.1.1), such a system is difficult for exact analysis unless $s_0 = 0$. Various approximations are proposed where the basic idea is to decompose the system into multiple single-stage systems with the input parameters depending on other stages.

A simple approximation (METRIC, [21]) works as follows: first, apply the single-stage results (Section 2.1) to the DC by noting that IO_0 is a Poisson random variable with parameter $\sum_{j\in\mathcal{J}}\lambda_j \cdot E(L_0)$. By (2.1)–(2.3), one can characterize IL_0, I_0, and B_0. By Little's law, the expected stockout delay at DC is $E(X_0) = E(B_0)/\sum_{j\in\mathcal{J}}\lambda_j$. Second, for each retailer j, regarding its supply system as an infinite server queue with a mean service time $E(X_0) + E(L_j)$, one can again apply the single-stage results to obtain the distribution of IO_j, I_j, and B_j. Clearly, the second step is an approximation because the orders placed by the retailers are satisfied by the DC on a FCFS basis.

Muckstadt [58] generalizes METRIC to include a hierarchical or indentured product structure (MOD-METRIC): when an assembly needs repair, then exactly one of its subassemblies (modules) needs repair. To illustrate the idea, let us consider a single-stage system with a single assembly and its modules $k \in \mathcal{K}$. Let s_0 (s_k) be the stock-level of the assembly (module k) and R_0 (R_k) its repair time. Assume the assembly failure rate is λ with probability p_k that module k needs repair, then the expected total repair time for an assembly is $E(\widetilde{R}_0) = E(R_0) + \sum_{k\in\mathcal{K}} p_k E(X_k)$, where $E(X_k) = E(B_k)/(p_k\lambda)$ is the expected delay due to stockout of module k. $E(B_k)$ is the expected backorders of module k which can be computed by (2.1) and (2.3) and the fact that IO_k follows Poisson $(E(R_k)p_k\lambda)$ (Section 2.1). Once $E(\widetilde{R}_0)$ is known, one can use METRIC to compute the performance measure at the assembly.

Sherbrooke [59] considers a similar model as Muckstadt [58] but utilizes a different approximation (VARI-METRIC). The key difference is to compute the first 2 moments (rather than the first moment) of the backorders at the depot and the outstanding orders at each base then fit their distributions by negative binomial distributions. Numerical study shows that VARI-METRIC improves the accuracy of METRIC. For a thorough literature review on inventory control in supply chains with repairable items, see Muckstadt [24].

Sequential Transit Time

Consider again the basic model. Note that each order placed by the retailers faces statistically the same stockout delay at the DC (by the independent Poisson demand and the FCFS rule), the exact analysis works as follows: first, compute the distribution of IO_0 by L_0 and the demand process at DC by Proposition 2.3. Then, determine the distribution of B_0 (by (2.3)). The distribution of X_0 can be determined by the fact that demand during X_0 (from all retailers) has the same probability distribution as B_0 (by the proof of Proposition 2.3). For any retailer j, the total replenishment lead time $\widetilde{L}_j = X_0 + L_j$. Given the demand process at retailer j, one can compute the distribution of IO_j and then B_j and X_j in a similar way. Svoronos and Zipkin [20] develop exact expressions of system performance for phrase-type transit times and present a two-moment approximation based on negative binomial distributions.

For compound Poisson demand, although the probability distribution of backorders may differ from that of the demand during stockout delay [29], the latter serves as a good approximation to the former. Zipkin [29] generalizes the 2-moment approximation of Svoronos and Zipkin [20] to distribution systems and presents an exact analysis based on the flow-unit method for phrase-type transit times and demand sizes (see also Section 3.2.3).

3.2.2. The Lead-Time Demand Method

Consider the basic model with sequential lead times (Definition 2.1). Applying (2.5) to DC yields $IL_0(t) = IP_0(t - L_0) - D_0(t - L_0, t]$, where $D_0(t - L_0, t]$ is the lead time demand for DC. By Proposition 2.4 and Corollary 3.2, we can determine the distribution of $D_0(\infty \mid L_0]$, IL_0, B_0, and I_0. For the retailers, we consider two cases.

Base-Stock Policy

By (3.8), $B_{0,j}(t-L_j)+IT_j(t-L_j)+IL_j(t-L_j) = IP_j(t-L_j) \equiv s_j,\ j \in \mathcal{J}$. By the lead-time demand method, at time t, all outstanding orders except $B_{0,j}(t-L_j)$ will be delivered to stage j, yielding

$$IL_j(t) = s_j - B_{0,j}\left(t - L_j\right) - D_j\left(t - L_j, t\right], \tag{3.9}$$

where $D_j(t-L_j, t]$ is the lead-time demand for retailer j. Since the distribution of $B_{0,j}$ is known ("binomial decomposition", Section 3.2.1), one can exactly characterize the distribution of I_j and B_j for all j [51, 56]. For fast computation, a two-moment approximation is proposed that fits $B_{0,j} + D(\infty \mid L_j)$ by a negative binomial distribution. In a numerical study, Graves [51] shows that the 2-moment approximation is more accurate than "METRIC" which only utilizes the first moment.

Exact analysis is feasible for distribution systems where each retailer has multiple supply modes, for example, upon arrival of a demand, a retailer can order a unit either from the DC (mode 1) or from mode 2 with constant lead time L'' [56]. The decision for each order is independent of others, so the total demand at stage j can be split into two independent Poisson processes each is served by a supply mode. Let $D'_j(t-L_j, t]$ $(D''_j(t-L''_j, t])$ be the lead-time demand served by mode 1 (2.2), then $IL_j(t) = s_j - B_{0,j}(t-L_j) - D'_j(t-L_j, t] - D''_j(t-L''_j, t]$, where all random variables on the right-hand side are independent.

Consider the basic model but assume that each stage utilizes a periodic-review base-stock policy. An important issue here is how to allocate DC's on-hand inventory to the retailers when the total demand exceeds the supply. The optimal allocation rule does not have a simple form, see, for example, Clark and Scarf [60] and Federgruen and Zipkin [61]. Therefore, most work so far focuses on heuristic rules, such as the "myopic" allocation rule [61], the random allocation rule [62, section 3.2.3], and the "virtual allocation" rule [63]. The "virtual allocation" rule works as follows: the DC observes external demand at all retailers and commits its stock in the sequence of external demand arrivals rather than the sequence of retailers' orders. An exact procedure is developed to characterize the inventory levels at all stages. Numerical study shows that virtual allocation has good performance although it is not optimal.

Batch Ordering Policy

As we mentioned at the beginning of Section 3.2, one of the challenges in distribution system is that the DC's demand process is a superposition of the retailers' order processes. This demand process becomes difficult to characterize when the retailers' use batch-ordering policies. Even for a simple system with identical retailers, the DC's demand process is a superposition of $|\mathcal{J}|$ many independent Erlang processes (by Corollary 3.2), thus it is nonrenewal [64]. Inspired by the "METRIC" approach, Deuermeyer and Schwarz [64], Lee and Moinzadeh [65, 66], and Svoronos and Zipkin [67] decompose the distribution system

into single-stage systems and propose various approximations for the retailers' lead-time demand. The key idea here is to characterize the moments of the DC backorders and then approximately determine either the delay due to stock at DC or the retailer j's share of the DC backorder. Finally, utilize either (2.5) or (3.9) to determine the moments of the lead-time demand at each retailer. See Axsater [16] for an extended discussion.

Chen and Zheng [43] consider the basic model with echelon batch ordering policies where the retailers may not be identical. The paper presents an exact analysis for Poisson demand and approximations for compound Poisson demand. To illustrate the idea, let IP_j^e (or IL_j^e) be the echelon inventory position (echelon inventory level) at stage $j \in \{0\} \bigcup \mathcal{J}$ where $IP_0^e = IO_0 + I_0 + \sum_{j\in\mathcal{J}}[IT_j + IL_j]$ and $IL_0^e = IP_0^e - IO_0$. First, one has $IL_0^e(t) = IP_0^e(t - L_0) - D_0(t - L_0, t]$ and $B_0(t) = [\sum_{j\in\mathcal{J}} IP_j^e(t) - IL_0^e(t)]^+$. The distribution of B_0 can be determined by the fact that $IP_j^e,\ j \in \{0\} \bigcup \mathcal{J}$ are independent (due to randomized initial conditions). Then, decompose the DC's backorders to each retailer to obtain $B_{0,j}, j \in \mathcal{J}$. Finally, $IT_j^e = IP_j^e - B_{0,j}$ and $IL_j^e = IT_j^e - D_j(\infty \mid L_j]$, see (3.6).

3.2.3. The Flow-Unit Method

The flow-unit method enables exact analysis for a wide range of distribution systems. Consider first the basic model with the sequential lead time (Definition 2.1). Suppose a demand arrives at retailer $j \in \mathcal{J}$ at time t, the stockout delay for this demand and the inventory holding time for the product that satisfies this demand are given by (2.9)-(2.10), $X_j = (\widetilde{L}_j - T_j(s_j))^+$ and $W_j = (T_j(s_j) - \widetilde{L}_j)^+$, where $\widetilde{L}_j$ is the total replenishment lead time for the order placed by stage j at time $t - T_j(s_j)$. For this order, the stockout delay and the inventory holding time for the corresponding item at the DC are $X_0 = (L_0 - T_0(s_0))^+$ and $W_0 = (T_0(s_0) - L_0)^+$. Therefore, $\widetilde{L}_j = X_0 + L_j$. Note that $T_j(s_j)$ is based on the demand of retailer j while $T_0(s_0)$ is based on the demand at DC. Because of Poisson demand, $T_0(s_0)$ (and thus X_0) is statistically the same for all retailer orders. Because $T_j(s_j), j \in \mathcal{J}$ are not overlapping with $T_0(s_0)$, $T_j(s_j) \perp T_0(s_0)$. This implies that the distribution system can be decomposed into single-stage systems where one can first evaluate the performance of the DC and then the performance of each retailer, see, for example, Axsater [28], Zipkin [29], and Simchi-Levi and Zhao [31].

For compound Poisson demand, let us consider the kth unit of a demand at node j. One needs to identify not only the corresponding order placed by stage j but also the corresponding unit in that order that satisfies this demand unit. By Zhao [33], $X_j(k) = (X_0(M_j(k)) + L_j - T_j(J_j(k)))^+$ and $W_j(k) = (T_j(J_j(k)) - L_j - X_0(M_j(k)))^+$, where $X_0(m) = (L_0 - T_0(J_0(m)))^+$. Here, $J_j(k)$ is the index of the corresponding order defined in Section 2.3, and $M_j(k)$ is the index of the unit in the corresponding order that satisfies the kth demand unit at node j. The analysis extends to a periodic-review systems with base-stock policy and virtual allocation rule (see Axsater [35] for Poisson demand and Forsberg [32] for compound Poisson demand).

We point out that for the special case of serial systems, the lead-time demand method handles Poisson demand and compound Poisson demand in the same way (3.4) but the flow-unit method becomes considerably more complex. On the other hand, for compound Poisson demand, the flow-unit method handles the serial and distribution systems in the same way but the lead-time demand method becomes much involved (the "Binomial decomposition" fails) as one moves from serial to distribution systems [57].

Batch-ordering policy complicates the analysis considerably due to the complex demand process faced by the DC. To see this, let us consider the basic model with identical

retailers and installation batch-ordering policy. The number of system demand (i.e., the demand of all retailers) between two consecutive retailers' orders is now random (versus a constant in the case of a single retailer). Forsberg [68] provides an exact analysis for distribution systems with batch ordering policy and Poisson demand. Axsater [36, 54] provides various approximations.

For distribution systems with both batch ordering policy and compound Poisson demand, Axsater [37] presents an exact analysis for installation policies and Axsater [42] considers echelon policies. The exact evaluation is, however, time consuming. Let m be a multiplier of the batch sizes. The computational effort is $O(|\mathcal{J}|^5)$ and $O(m^2)$ [68], $O(|\mathcal{J}|^2)$ and $O(m^4)$ [37], and $O(|\mathcal{J}|^{5/2})$ and $O(m^2)$ [42]. Cachon [62] provides an exact analysis for a periodic-review system with installation batch ordering policy, identical retailers, and i.i.d. demand, where the DC randomly allocates stock to orders received in the same period but follows the FCFS rule to serve orders in consecutive periods.

Because the flow-unit method requires the FCFS rule and the assumptions that orders are replenished in the sequence as they are placed, it is not clear how to apply this method to problems where these assumptions fail, for example, systems with multiple supply modes (Section 3.2.2), systems with reverse material flows [69], and systems with rationing rules [70]. For these systems, the lead-time demand method still applies.

3.3. *Assembly Systems*

In this section, we consider both pure assembly systems where each stage has at most one customer and two-level general networks where each stage can have multiple customers or suppliers.

In a two-level general network, stages in $\mathcal{O}$ are suppliers and stages in $\mathcal{J}$ are customers. Supply-demand relationship exists only between sets $\mathcal{O}$ and $\mathcal{J}$. It is convenient to call the set $\mathcal{O}$ components and the set $\mathcal{J}$ products. Let $\mathcal{O}_j = \{j \in \mathcal{J} \mid a_{i,j} > 0\}$ be the component set for product j, and $\mathcal{J}_i = \{j \in \mathcal{J} \mid a_{i,j} > 0\}$ the product set served by component i. Let L_i (L_j) the transit time at stage $i \in \mathcal{O}$ ($j \in \mathcal{J}$), and $L_{i,j}$ be the transit time (e.g., transportation lead time) from stage $i \in \mathcal{O}$ to stage $j \in \mathcal{J}$. Note that each stage $i \in \mathcal{O}$ is performing a distribution operation and each stage $j \in \mathcal{J}$ is performing an assembly operation. We assume that a product can be assembled only when all necessary components are available.

The two-level general network includes the following important special cases: (i) pure assembly systems where $|\mathcal{J}| = 1$. Here, we index the unique stage in $\mathcal{J}$ by 0. (ii) Assemble-to-order (ATO) systems where $L_{i,j} = 0$ for all i, and j, $L_j = 0$ for all j and all stages in $\mathcal{J}$ carry zero inventory. This model can be applied to CTO (configure-to-order) systems, repairable items with multiple failure [71], and the "pick and ship" systems in B2C e-commerce.

The optimal policies on ordering or allocation in such a network are either not known or state-dependent and thus too complex to implement [72]. In practice, only suboptimal but simple ordering policies (e.g., installation policies) and simple allocation rules (e.g., FCFS) are implemented. Here, we focus on installation policies and FCFS rule unless otherwise mentioned.

Assembly systems pose a significant challenge for policy evaluation because of the common demand processes shared by different components. One has to address the question of how to characterize the dependence among components? And what is the impact of the dependence on system performance?

We define the basic model for assembly systems as follows: each stage utilizes an installation base-stock policy, external demand follows independent Poisson processes with

rates $\lambda_j, j \in \mathcal{J}$, all transit times are constant. Let $a_{i,j}$ be either zero or one unless otherwise mentioned. When a stage $j \in \mathcal{J}$ places an order and some of its suppliers have on-hand inventory but others do not, we assume that the available stocks are shipped to stage j immediately. Clearly, each stage $j \in \mathcal{J}$ may hold inventory for components $i \in \mathcal{O}_j$ which is not yet processed due to shortages of other components. We call this inventory the "committed stock" [15].

3.3.1. The Queueing-Inventory Method

Consider the two-level general network under Assumption 2.2, by (2.1) and (2.3), $IO_l(t) + IL_l(t) = IP_l(t)$ and $B_l(t) = IL_l(t)^-$, $l \in \mathcal{O} \bigcup \mathcal{J}$. Let $B_{i,j}$ be the orders placed by stage j backlogged at stage i. Similar to (3.7),

$$B_i(t) = \sum_{j \in \mathcal{J}_i} B_{i,j}(t). \tag{3.10}$$

For each product $j \in \mathcal{J}$, let $IT_{i,j}$ be the inventory in-transit from stage i to j during time $L_{i,j}$, IT_j the inventory in-transit during L_j, and $I_{i,j}$ the committed stock of component i at stage j. Then,

$$\begin{aligned} IO_j(t) &= \max_{i \in \mathcal{O}_j} \{B_{i,j}(t) + IT_{i,j}(t)\} + IT_j(t), \quad i \in \mathcal{O}, \\ I_{i,j}(t) &= \max_{l \in \mathcal{O}_j} \{B_{l,j}(t) + IT_{l,j}(t)\} - B_{i,j}(t) - IT_{i,j}(t). \end{aligned} \tag{3.11}$$

In the special case of ATO systems, the backorders at stage j, $B_j(t)$, and the on-hand plus committed inventory of component i, $\widetilde{I}_i(t)$, are given by

$$B_j(t) = \max_{i \in \mathcal{O}_j} \{B_{i,j}(t)\}, \tag{3.12}$$

$$\widetilde{I}_i(t) = IL_i(t) + \sum_{j \in \mathcal{J}_i} [B_j(t) - B_{i,j}(t)] = IP_i(t) - IO_i(t) + \sum_{j \in \mathcal{J}_i} B_j(t). \tag{3.13}$$

If $|\mathcal{J}| = 1$ in the ATO systems, then (3.12)-(3.13) reduce to

$$B_0(t) = \max_{i \in \mathcal{O}} \{B_i(t)\}, \tag{3.14}$$

$$\widetilde{I}_i(t) = IP_i(t) - IO_i(t) + B_0(t). \tag{3.15}$$

Because the ATO systems capture the dependence among the components in the two-level general networks, we focus on ATO systems for the rest of Section 3.3.

Consider first the basic model for ATO systems with i.i.d. transit times and $|\mathcal{J}| = 1$. The stages $i \in \mathcal{O}$ form $|\mathcal{O}|$ parallel $M/G/\infty$ queues with common demand arrivals. The objective of the queue-inventory method is to characterize the joint distribution of the outstanding orders (i.e., job in queues): $\overline{IO} = (IO_i,\ i \in \mathcal{O})$. Once $\overline{IO}$ is known, B_0 is given by (3.14), $\widetilde{I}_i$ is given by (3.15), and the order-based fill rate $f_0 = \Pr\{s_i - IO_i >,\ \forall i \in \mathcal{O}\}$.

The analysis of $\overline{IO}$ is based on the following observation (see, e.g., [73]). For simplicity, let $\mathcal{O} = 2$. Define $\varphi_i(\cdot)$ (or $\Psi_i(\cdot)$) to be the pdf (cdf) function of L_i, for all i. Let $\Psi_i^c(u) = 1-\Psi_i(u)$. Consider an arbitrary demand arrival in $[0,t]$. Due to Poisson demand, the arrival time of this demand is uniformly distributed in $[0,t]$. Conditioning on the arrival time $0 \le u \le t$, the probability that both queues ($i = 1,2$) are still processing the job triggered by this demand at t is $p_{1,2}(u) = \Pr\{L_1 > t-u\}\Pr\{L_2 > t-u\}$ which equals to $\Psi_1^c(t-u)\Psi_2^c(t-u)$. Similarly, the probability that only queue 1 (or 2) is still processing the job at t is $p_1(u) = \Psi_1^c(t-u)\Psi_2(t-u)$ ($p_2(u) = \Psi_1(t-u)\Psi_2^c(t-u)$, resp.). Finally, the probability that both queues finish the job at t is $p_0(u) = \Psi_1(t-u)\Psi_2(t-u)$. Unconditioning on u, $p_{1,2} = (1/t)\int_0^t \Psi_1^c(t-u)\Psi_2^c(t-u)du$, same logic applies to p_1, p_2, and p_0.

Let $\widetilde{N}(t)$ be the total jobs up to time t. Among these jobs, let $N_{1,2}(t)$ be those in process in both queues $j = 1$ and 2, $N_1(t)$ (or $N_2(t)$) those in process only in queue 1 (or 2, resp.), and $N_0(t)$ those left both queues. Because all arrivals are independent, conditioning on $\widetilde{N}(t) = n$, $(N_{1,2}(t), N_1(t), N_2(t), N_0(t))$ follows multinomial distribution with parameters $n, p_{1,2}, p_1, p_2, p_0$. Clearly, $IO_i(t) = N_{1,2}(t) + N_i(t), i = 1,2$, and IO_is are dependent due to the common element, $N_{1,2}$. Applying the logic to ATO systems with any $|\mathcal{O}|$ and let $t \to \infty$,

$$IO_i = \sum_{\forall \Omega \subseteq \mathcal{O} | i \in \Omega} N(\lambda_0 \theta_\Omega). \tag{3.16}$$

Here, $N(\cdot)$s are independent Poisson random variables and $\theta_\Omega = \int_0^\infty [\prod_{i\in\Omega}\Psi_i^c(u)] \cdot [\prod_{i\in\mathcal{O}\setminus\Omega}\Psi_i(u)]du$. Note that there are $2^{|\mathcal{O}|} - 1$ Poisson random variables.

Lu et al. [74] generalize the result to ATO systems with multiple products and provide the generating function for $\overline{IO}$ and bounds for the order-based fill rates. Lu et al. [75] present bounds for the order-based backorders. Interestingly, the lower bound on $E(B_j)$ is related to the "binomial decomposition" in distribution systems (Section 3.2.1). Due to independent Poisson demand and FCFS rule, $B_{i,j}$ in (3.10) follows a binomial distribution for any given B_i. By (3.12), $E(B_j) \ge \max_{i\in\mathcal{O}_j}\{E(B_{i,j})\} = \max_{i\in\mathcal{O}_j}\{E(B_i)\lambda_j/\sum_{l\in\mathcal{O}_i}\lambda_l\}$.

Lu and Song [76] formulate a nonconstrained cost-minimization problem for the model, where the total cost includes backorder cost and holding cost for both on-hand and committed stock. It is shown that the total cost is submodular in $s_i, i \in \mathcal{O}$. For other types of ATO systems, Gallien and Wein [77], Cheung and Hausman [71], and Dayanik et al. [78] characterize the distribution of $\overline{IO}$ which leads to either exact analysis or bounds on the key performance measure. See Song and Zipkin [15] for an extended discussion.

To date, it is not clear how to use the queueing-inventory method to characterize ATO systems with either stochastic sequential lead times or batch-ordering policies because the joint distribution of the outstanding orders is difficult to characterize and $(IO_i, i \in \mathcal{O})$ depends on $(IP_i, i \in \mathcal{O})$. Fortunately, some of these systems can be handled by the lead-time demand method and the flow-unit method.

3.3.2. The Lead-Time Demand Method

We first consider the basic model for the ATO systems with $|\mathcal{O}| = 1$. Let us index the components $i \in \mathcal{O}$ in a nondecreasing order of their lead times, that is, $L_1 \le L_2 \le \cdots \le L_{|\mathcal{O}|}$. By (2.5), $IL_i(t) = s_i - D(t-L_i,t]$, $i \in \mathcal{O}$. Since all components face identical demand process, by Zipkin [9, Section 8.4.5],

$$D(t-L_i,t] = D(t-L_i,t-L_{i-1}] + D(t-L_{i-1},t], \quad i = 2,3,\ldots,|\mathcal{O}|. \tag{3.17}$$

By (3.14) and (3.17), $B_0(t) = \max_{i\in\mathcal{O}}\{[D(t-L_1,t] + \sum_{l=2}^{i} D(t-L_l, t-L_{l-1}] - s_i]^+\}$. For component i, the on-hand inventory is IL_i^+ and the committed inventory is $B_0 - B_i$ where $B_i = IL_i^-$. Analogous to (3.15), the total on-hand plus committed inventory of component i is $\widetilde{I}_i(t) = s_i - D(t-L_1,t] - \sum_{l=2}^{i} D(t-L_l, t-L_{l-1}] + B_0(t)$. Because $D(t-L_1,t]$ and $D(t-L_i,t-L_{i-1}]$ are independent, exact analysis is feasible. The key idea here is to identify the common lead-time demand shared by different components.

This approach can be generalized to multiproduct ATO systems with constant lead times. Consider the basic model with $|\mathcal{J}| > 1$. Because the demand processes for different components may not be completely identical, (3.17) no longer holds. Consider two components, i and ι. There are 4 cases.

(1) $\mathcal{J}_i \bigcap \mathcal{J}_\iota = \emptyset$. Then, $D(t-L_i,t] \perp D(t-L_\iota,t]$.

(2) $\mathcal{J}_i = \mathcal{J}_\iota$. This case can be handled by (3.17).

(3) $\mathcal{J}_i \subset \mathcal{J}_\iota$. Consider the following two subcases.

(i) If $L_i < L_\iota$, $D_\iota(t-L_\iota,t] = D_\iota(t-L_\iota,t-L_i] + D_i(t-L_i,t] + D_{\mathcal{J}_\iota\setminus\mathcal{J}_i}(t-L_i,t]$ where $D_{\mathcal{J}_\iota\setminus\mathcal{J}_i}(t-L_i,t]$ is total demand of products in set $\mathcal{J}_\iota \setminus \mathcal{J}_i$ during $(t-L_i,t]$.

(ii) If $L_i \geq L_\iota$, $D_\iota(t-L_\iota,t] = D_i(t-L_\iota,t] + D_{\mathcal{J}_\iota\setminus\mathcal{J}_i}(t-L_\iota,t]$ and $D_i(t-L_i,t] = D_i(t-L_i,t-L_\iota] + D_i(t-L_\iota,t]$.

All lead-time demand on the right-hand side of the equations are independent.

(4) $\mathcal{J}_i \bigcap \mathcal{J}_\iota \neq \emptyset$ but $\mathcal{J}_i \not\subset \mathcal{J}_\iota$ and $\mathcal{J}_\iota \not\subset \mathcal{J}_i$. This case is more complex but still tractable see, for example, [79]. The key idea is to identify the common lead-time demand for both components.

Using convolution, Song [79] presents exact expressions for the order-based fill rates $f_j = \Pr\{s_i - D_i(t-L_i,t] > 0, i \in \mathcal{O}_j\}$. It is also shown that $f_j \geq \prod_{i\in\mathcal{O}_j}\Pr\{s_i - D_i(t-L_i,t] > 0\}$. This inequality implies that ignoring the correlation among components results in underestimating the fill rates.

To determine the expected order-based backorders, Song [80] utilizes the relation between the fill rate and the stockout delay. Let X_j be the stockout delay for product j. Clearly, $0 \leq X_j \leq \max\{L_i, i \in \mathcal{O}_j\}$, and, by (2.6), $\Pr\{X_j \leq x\} = \Pr\{D_i(t-L_i+x,t) < s_i, i \in \mathcal{O}_j \text{ and } x < L_i\}$. By Little's law, one can translate the problem of the expected backorders to the problem of the expected stockout delays. See Song [80] for a detailed discussion.

Consider the basic model for the ATO systems except that stage $i \in \mathcal{O}$ utilizes a batch ordering policy (r_i, Q_i). If external demand follows a compound multivariate Poisson process, Song [81] shows that the inventory position vector of all components, $(IP_i,\ i \in \mathcal{O})$, is uniformly distributed in $\otimes_{i\in\mathcal{O}}\{r_i+1, r_i+2, \ldots, r_i+Q_i\}$ if the CTMC of $(IP_i,\ i \in \mathcal{O})$ is irreducible and aperiodic. Therefore, the expected order-based backorders and fill rates, of a batch-ordering ATO system, can be expressed as the average of the counterparts of multiple base-stock systems.

For ATO systems under periodic review, the idea is similar: identify common lead-time demand shared by components. However, the allocation rule for common components becomes an important issue. Hausman et al. [27] consider a multi-item system where $D_i(n)$, the demand in nth period, follows multivariate normal random distribution. Assuming constant lead times, FCFS rule, and independent demand across periods, the probability of satisfying all demand in period n within τ periods of time is $\Pr\{D_i[n-L_i+\tau, n] \leq s_i,\ i \in \mathcal{O}\}$ by (2.8). Zhang [82] considers a different allocation rule, the "fixed-priority" rule: while demands in consecutive periods are served on a FCFS basis, demands in the same period are

served based on their priority. Let $j \succeq j$ denote that demand j has higher priority over demand $\imath$. The fill rate for customer type j is given by $\Pr\{D_i[n-L_i, n-1\text{s}] + \sum_{l\in\mathcal{J}, l\succeq j} a_{i,l} D_l(n) \le s_i, i \in \mathcal{I}_j\}$, where $D_l(n)$ is the demand of product l at period n. Since high dimensional multivariate normal distributions are computationally intensive, bounds on the fill rates are developed. Agrawal and Cohen [83] study the "fair-share" rule for demand in the same period: if component i has a shortage in period n, then the fraction of component i's available stock allocated to product j equals to $D_j(n)/D_i(n)$. The resulting order-based fill rate is identical to that of Hausman et al. [27].

de Kok [84] imposes an "ideal" product structure on the model of Hausmans et al. [27]: if $L_\imath \le L_i$, then either $\mathcal{J}_\imath \bigcap \mathcal{J}_i = \emptyset$ or $\mathcal{J}_\imath \subseteq \mathcal{J}_i$. An ATO system is "strongly ideal" if it has an idea product structure and satisfies the condition that for any product $j \in \mathcal{J}_\imath \bigcap \mathcal{J}_i$, $a_{i,j} = a_{\imath,j}$. Further assume a linear allocation rule and demand occuring at the end of each period, it is shown that the order-based fill rates satisfy $\Pr\{\sum_{l=0}^{L_i-1} \sum_{k\in\mathcal{J}} a_{i,k} D_k(n-l) \le s_i, i \in \mathcal{I}_j\}$, for all j. If the ATO system is strongly ideal, then fill rates have the form of $\Pr\{\sum_{l=1}^{m} Z_l \le c_m,\ m = 1, 2, \ldots, M\}$ which is a generalized finite horizon nonruin probability studied extensively in the actuarial literature.

Unlike serial and distribution systems (Sections 3.1-3.2), extensions from constant lead times to stochastic sequential lead times (by the lead-time demand method) is not straightforward because it is difficult to determine the common lead time demand. The flow-unit method, which separates demand from the lead time, provides a simpler and cleaner analysis.

3.3.3. The Flow-Unit Method

Consider the basic model for ATO systems with stochastic sequential lead times where the component inventory is managed by either continuous-time base-stock policies or batch-ordering policies. We refer to the latter as a *batch-ordering* system and the former as a base-stock system. The following discussion is based on Zhao and Simchi-Levi [30].

Single-Product Base-Stock Systems

Let $|\mathcal{J}| = 1$. Consider components i and $\imath$. Without loss of generality, let $s_i \le s_\imath$. Suppose a demand arrives at time t, then the corresponding orders of the components i and $\imath$ that satisfy this demand are placed at time $t - T(s_i)$ and $t - T(s_\imath)$, respectively (the "backward method", see Section 2.3). It is easily seen that $T(s_i)$ overlaps with $T(s_\imath)$ over the time period $[t - T(s_i), t]$, and therefore $T(s_\imath) = T(s_i) + T(s_\imath - s_i)$. The dependence among the arrival times $t - T(s_i) + L_i$, $i \in \mathcal{I}$, is quite intuitive: if the interarrival times are short for recent demands, and as a result $T(s_i)$ is small for all $i \in \mathcal{I}$, then all components are likely to be out of stock.

Indexing the components in the nondecreasing order of their base-stock levels, for any sequence of $t_1 \le t_2 \le \cdots \le t_{|\mathcal{I}|}$, the joint probability density function of $T(s_i)$, $i \in \mathcal{I}$, is given by

$$\begin{aligned} &\Pr\{T(s_1) = t_1, T(s_2) = t_2, \ldots, T(s_{|\mathcal{I}|}) = t_{|\mathcal{I}|}\} \\ &\quad = \Pr\{T(s_1) = t_1\} \\ &\qquad \times \Pr\{T(s_2 - s_1) = t_2 - t_1\} \cdots \Pr\{T(s_{|\mathcal{I}|} - s_{|\mathcal{I}|-1}) = t_{|\mathcal{I}|} - t_{|\mathcal{I}|-1}\}. \end{aligned} \tag{3.18}$$

For other sequences of $t_1, t_2, \ldots, t_{|\mathcal{I}|}$, $\Pr\{T(s_1) = t_1, T(s_2) = t_2, \ldots, T(s_{|\mathcal{I}|}) = t_{|\mathcal{I}|}\} = 0$.

By (3.18), we can derive the probability distribution for the product stockout delay, $X_0 = [\max_{i\in\mathcal{O}}\{L_i - T(s_i)\}]^+$. For any service time $\tau(\geq 0)$, conditioning on $\overline{L} = \overline{l} = (l_1, l_2, \ldots, l_{|\mathcal{O}|})$ yields

$$\begin{aligned}\Pr\{X_0 \leq \tau\} = \Pr\Big\{&T(s_1) \geq (l_1 - \tau)^+, T(s_1) + T(s_2 - s_1) \geq (l_2 - \tau)^+, \ldots, \\ &T(s_1) + T(s_2 - s_1) + \cdots + T\left(s_{|\mathcal{O}|} - s_{|\mathcal{O}|-1}\right) \geq \left(l_{|\mathcal{O}|} - \tau\right)^+\Big\}.\end{aligned} \tag{3.19}$$

The waiting time of component $i, i \in \mathcal{O}$, is determined by $W_i = X - L_i + T(s_i)$.

The backward method may work better for assembly systems than the forward method because, in the latter, the orders (of components) triggered by a demand will satisfy different demand in the future; while in the former, we focus on a demand and identify all the orders placed beforehand that satisfy this demand. The flow-unit method separates the demand process from the lead times rather than putting them together as lead time demand. Thus, the demand process determines $T(s_i)$, $i \in \mathcal{O}$, whose joint distribution can be easily characterized, and the supply system determines $L_i, i \in \mathcal{O}$, which need not be independent.

Multiproduct Base-Stock Systems

Let $|\mathcal{O}| > 1$. Assuming that a demand of product type $j \in \mathcal{J}$ arrives at time t, then the corresponding order of component $i \in \mathcal{O}_j$ that satisfies this demand is placed at time $t - T_{i,j}(s_i)$, where $T_{i,j}(s_i)$ is determined by starting at time t, counting backward demand arrivals of all products that require component i until the total number of arrivals reaches s_i. Because of the lead time, an order placed at time $t - T_{i,j}(s_i)$ will arrive at time $t - T_{i,j}(s_i) + L_i$.

For each product $j \in \mathcal{J}$, the stockout delay is $X_j = [\max_{i\in\mathcal{O}_j}\{L_i - T_{i,j}(s_i)\}]^+$, and component i's waiting time, when it is committed to product j, is $W_{i,j} = X_j - L_i + T_{i,j}(s_i)$. Thus, the multiproduct ATO system can be decomposed into $|\mathcal{J}|$ single-product subsystems with each subsystem corresponding to a product $j \in \mathcal{J}$ and its component set $\mathcal{O}_j$. It is important to note that these single-product subsystems are not identical to the single-product assembly systems because $T_{i,j}(s_i)$ is associated with the superposition of the demand processes of all products that require component i. Close-form expressions are derived for the covariance matrix of $T_{i,j}(s_i)$, $i \in \mathcal{O}_j$. Zhao [33] characterizes their joint probability distribution.

Zhao and Simchi-Levi [30] proposes two numerical methods to evaluate system performance. The first method is based on Monte Carlo simulation while the second method is based on a two-moment approximation. A numerical study of an example inspired by a real world problem, the *Dimension 2400 Pentium of* **Dell**, shows that the simulation-based method is scalable and can evaluate large size, real world ATO systems; while the method based on the 2-moment approximation can handle up to medium size ATO systems with multiple products.

Multiproduct Batch-Ordering Systems

Now assume that inventory of each component is controlled by a continuous-time batch-ordering policy. Let $\mathcal{S}_j = \otimes_{i\in\mathcal{O}_j}\{r_i + 1, r_i + 2, \ldots, r_i + Q_i\}$. Based on Song [81], Zhao and Simchi-Levi [30] prove the following proposition.

Proposition 3.3. *Assume that the Markov chain of the inventory position vector of the components is irreducible and aperiodic. Suppose that a demand for product $j \in \mathcal{J}$ arrives at time t, then the*

corresponding order of component i, $i \in \mathcal{O}_j$, that satisfies this demand is placed at time $t - T_{i,j}(S_i)$, where the random vector $(S_i, i \in \mathcal{O}_j)$ is uniformly distributed in $\mathcal{S}_j$.

Based on Proposition 3.3, the order-based fill rates and the expected stockout delays can be expressed as the averages of their counterparts in the base-stock systems. However, the number of the corresponding base-stock systems is exponential in the number of components. By exploring the problem structure, Zhao and Simchi-Levi [30] develop efficient numerical methods based on Monte Carlo simulation. Given the sample size, the number of products, and the reorder points, the computational complexity of the methods is no more than that of sorting a set of real numbers, where the set size equals to the sum of the batch sizes of all components.

3.4. General Supply Networks

In this section, we discuss extensions of the three generic methods to general supply chains.

Supply Chain Characteristics

A supply chain consists of facilities and products. To specify a network, we need to know the processing cycle time for each product at each facility and the transportation lead time between every two facilities. We also need to know the BOM structure, external demand processes, target service levels (e.g., the committed service times and the target fill rates), and the value added at each facility for each product.

Network Classification

A node (or a stage) refers to a unique combination of facility and product, and an arc refers to a pair of nodes with direct supply-demand relationship. A tree network is the one where breaking any arc results in two separate subnetworks. A tree network includes serial distribution and pure assembly as special cases. Networks with at most one directed path between every two nodes include tree as a special case but are not limited to tree, for example, the two-level general networks (Section 3.3). An acyclic network is more general which allows multiple directed paths between two nodes. Finally, supply chains may have feedback loops or reverse flows which form into close loop networks.

Unless otherwise mentioned, we assume that Assumption 2.2 holds. In addition, each node utilizes an installation base-stock policy, $a_{i,j}$ equals either zero or one, and external demand follows independent Poisson processes in case of continuous review or is i.i.d. random variables in case of periodic-review.

3.4.1. The Lead-Time Demand Method

We follow the development of the literature by first considering the lead-time demand method. The idea here is the same as "METRIC": breaking a network into multiple single-stage systems with the input parameters depending on each other.

Lee and Billington [1] analyze the Hewlett-Packard DeskJet printer supply chain with the objective of providing tools for managers to evaluate various stock positioning strategies. Each stage in the supply chain utilizes a periodic-review base-stock policy, the transit times in manufacturing and transportation processes are stochastic and sequential. Demand

process at each stage can be obtained by aggregation of the BOM. For each node j, the total replenishment lead time $\widetilde{L}_j$ consists of three parts: the processing time at node j, transportation lead times, and stockout delays from immediate suppliers. For assembly systems, it is assumed that at most one supplier can be out of stock in each period. Let f_i be the fill rate at stage i. Hence,

$$E\left(\widetilde{L}_j\right) \approx \sum_{(i,j)\in\mathcal{A}} a_{i,j}E(L_{i,j}) \Big/ \sum_{(i,j)\in\mathcal{A}} a_{i,j} + \sum_{(i,j)\in\mathcal{A}} (1-f_i)E(X_i) + E(L_j). \tag{3.20}$$

Similarly, $V(\widetilde{L}_j)$ can be determined by the first two moments of X_i, L_j, and $L_{i,j}$.

Let RP be the length of one review period. One can compute the first 2 moments of the lead-time demand at node j by $[E(\widetilde{L}_j) + RP_j]\mu_j$ and $[E(\widetilde{L}_j) + RP_j]\sigma_j^2 + \mu_j^2 V(\widetilde{L}_j)$, respectively. Here, μ_j (or σ_j) is the mean (standard deviation) of demand in one period. Approximating the lead time demand by a normal random variable, then the on-hand inventory is determined by Proposition 2.4; $E(X_j)$ and $V(X_j)$ are computed based on (2.8) where L is replaced by $E(\widetilde{L}_j)$.

3.4.2. The Queueing-Inventory Method

Ettl et al. [85] applies the queueing-inventory method to supply chains where each stage utilizes a continuous-time base-stock policy, all transit times are i.i.d. random variables, and the external demand follows compound Poisson process. For each node j, $\widetilde{L}_j$ is given by

$$\widetilde{L}_j \;=\; \max_{(i,j)\in\mathcal{A}} \{X_i\} + L_j. \tag{3.21}$$

To compute the moments of $\widetilde{L}_j$, it is assumed that at most one supplier can be out of stock at any time [1]. Then, the supply process at node j is approximated by a $M^Y/G/\infty$ queue with $\widetilde{L}_j$ being the service time. By queueing theory, one can derive expressions for the moments of IO_j, which in turn yields the statistics of I_j, B_j (see (2.1)–(2.3)), and customer service levels. Since it is a challenge to determine $E(X_j)$, an upper bound based on $M/M/\infty$ queue is utilized.

In addition to performance evaluation, Ettl et al. [85] optimize the total inventory investment, that is, the sum of expected work-in-process and finished goods inventory, in the supply chain subject to meeting certain service requirements of the external customers. Using the safety factors (service levels) as decision variables, the authors developed analytic expressions for the gradients, and therefore the constrained nonlinear optimization problem can be solved by the conjugate gradient method. Numerical studies show that this problem has many local optimal solutions, and the strategy of setting high fill rates at all stages can perform poorly relative to the optimal solution (an average of roughly 20% gap is recorded).

3.4.3. The Flow-Unit Method

Applying the flow-unit (backward) method, one can provide exact analysis of supply chains with exogenous, stochastic, and sequential transit times (Definition 2.1).

Simchi-Levi and Zhao [31] consider tree networks with independent Poisson demand and continuous-review base-stock policies and develop exact recursive equations for the stockout delays at all stages of the supply chain. At node j, we must have, (see (2.9)),

$$\begin{aligned} X_j &= \left(\widetilde{L}_j - T_j(s_j)\right)^+, \\ \widetilde{L}_j &= \max_{(i,j)\in\mathcal{A}} \left\{X_i + L_{i,j}\right\} + L_j. \end{aligned} \tag{3.22}$$

Clearly, $X_i = (\widetilde{L}_i - T_i(s_i))^+$ where $T_i(s_i)$ may be dependent (see, e.g., (3.18)). The key idea here is that, for each external demand, we look backward in time to identify the corresponding order placed by each stage in the supply chain that eventually satisfies the demand. Thus, the recursive equations hold not only for systems in steady state, but also for systems in transient states with time varying and/or temporally correlated demand.

If supplier i is in turn supplied by other node(s) in the system, then $\widetilde{L}_i$ may be correlated across nodes i where $(i, j) \in \mathcal{A}$, and $\widetilde{L}_i$ may also be correlated with $T_\iota(s_\iota)$ for $i \neq \iota$. Indeed, the stockout delays of parallel branches in a multistage assembly system can be correlated. See Simchi-Levi and Zhao [31] for an indepth analysis of the correlations. The following proposition characterizes the impact of the correlations on system performance.

Proposition 3.4. *Consider a tree-structure supply chain. If external demand follows independent Poisson processes, then any assembly node in the system has stochastically shorter backorder delay and longer inventory holding time than their counterparts in an analogous system with independent lead times.*

Based on the recursive equations, Simchi-Levi and Zhao [31] prove the following properties.

Theorem 3.5. *Given two serially linked nodes, node 2 (supplier) and node 1 (customer), in a tree supply network. Let $s_2 > 0$. Then, moving one unit of inventory from node 2 downstream to node 1 yields* (i) *stochastically shorter backorder delay (equivalently, stockout delay) at node 1 and* (ii) *stochastically shorter inventory holding time for any item traveling through both nodes.*

This theorem holds for any tree-structure supply chain facing point demand processes under the assumption that demand and supply can be split. One application of this theorem is that moving inventory from all upstream stages to the most downstream stage reduces (stochastically) the total inventory holding time for any item in the system as well as the backorder delays to the external customers.

Proposition 3.6. *Under Definition 2.1 and the assumption of independent Poisson demand processes, $\widetilde{L}_j$ is independent of $T_j(s_j)$ at every node $j \in \mathcal{N}$.*

Guided by the exact analysis, Simchi-Levi and Zhao [31] present two-moment approximations and tractable decompositions that lead to an efficient evaluation and optimization algorithm for general tree-structure supply chains. The algorithm computes the first two moments of the stockout delay at each stage of the network according to Proposition 3.6 and (3.22). To identify the optimal or near optimal stock levels in the supply chain that minimize system-wide inventory cost subject to service level constraints, the algorithm employs a dynamic programming routine to evaluate all stages sequentially.

The algorithm is tested in various supply chains including a 22-stage and 21-arc assembly network inspired by a real world problem, the Bulldozer supply chain, see Graves and Willems [11]. Comparing to simulation results, the approximations are sufficiently accurate for a wide range of system parameters, and the algorithm computes the optimal or near optimal stock levels efficiently. It is shown that the lead time uncertainties have significant impact on the stock levels and stock positions, and ignoring lead time uncertainties can lead to substantial errors.

Zhao [33] extends the analysis and approximation to compound Poisson demand and networks with at most one directed path between every two nodes. Shi and Zhao [86] consider acyclic supply chain and discover some simple yet unique properties.

4. Conclusion

We conclude the paper by pointing out some extensions of the models and methodologies and some of the remaining challenges.

General Supply Network with Batch Ordering Policy

In practice, economies of scale in production or transportation costs may drive batch ordering policies across the supply chain. General supply networks, for example, tree, with batch ordering policies and lead times have not been studied in the literature. Indeed, Ettl et al. [85] and Muckstadt [24] call for models and algorithms to handle these systems.

Supply Chains with Multiple Products: Design of Network and BOM

For exact analysis of general structure supply chains with multiple products, two challenges remain: (1) the mapped network may be acyclic and (2) $a_{i,j}$ may be any nonnegative integer. Resolving these challenges requires an extension of the stochastic, sequential lead time model (Definition 2.1) to include joint probability distributions for the transit times [33].

Despite these challenges, inventory positioning in multiproduct supply chains with common components deserves attentions as it holds the promises of jointly optimizing BOM, network, and inventory. Without doubt, the design of network such as selection of suppliers, transportation modes, manufacturing capabilities, and locations of facilities greatly affect the inventory costs and service levels. Moreover, the implementation of strategies such as component commonality, modular design, and postponement has made significant impact on real-world supply chains, see, for example, Feitzinger and Lee [87]. Given recent developments in the inventory positioning literature (in particular, in assembly systems and general networks), we see huge opportunities in this direction.

Supply Chains with Reverse Material Flows

The reverse material flows can be caused by returns, recycling, or feedbacks. Supply chains with returns are different from those handling repairable items because, in the latter, a returned defected item is always accompanied by a demand for a workable item and the defected item cannot be reused immediately. So far, researchers have applied the lead-time demand method to supply chains with returns, see, for example, Fleischmann et al. [69] for single-stage systems, DeCroix et al. [88] for serial systems and DeCroix and Zipkin [89] for assembly systems. It is not clear, though, how the other two methods can be applied here.

Supply Chains with Processing Capacity Constraints

Positioning inventory in supply chains with processing (e.g., production) capacity constraints poses a substantial challenge. To see this, let us consider the basic model for serial systems with an exponential server at each stage. The transit time at stage j depends on the departure process at stage $j + 1$ which is not even renewal [90]. For ATO system, introducing capacitated suppliers significantly complicates the way components interact, see, for example, Song et al. [91] and Zhao and Simchi-Levi [30]. Exact evaluation and optimization of capacitated supply chains are difficult, we refer to Buzacott et al. [90], Lee and Zipkin [92], Glasserman and Tayur [93] and Liu et al. [94] for various approximations, and to Glasserman and Tayur [95] for a simulation-based optimization algorithm.

Supply Chains with Nonstationary and/or Correlated Demand

So far, we assume that demand processes are stationary and uncorrelated. In practice, demand can be nonstationary (due to seasonality or short product life-cycle) and correlated, and demand forecast can be updated. Supply networks with nonstationary and/or correlated demand pose significant analytical and numerical challenges. For instance, in a periodic-review system with nonstationary demand, one has to determine stock levels not only across facilities but also across time. See Ettl et al. [85] for a rolling horizon approach and Graves and Willems [96] for an extension of the guaranteed service model.

Supply chains with correlated demand is difficult because they generally cannot be decomposed into single-stage systems. Erkip et al. [97] provide a decomposition result under the "balanced assumption" [98]. Dong and Lee [99] considers a serial system and provide a low bound for the optimal base-stock levels. Truong et al. [100] provide a simple heuristic policy (by tracing flow-units) with a constant worst performance guarantee.

Other Inventory Control Policies

Clearly, the base-stock and batch ordering policies are not the only ones studied in literature and used in practice. For instance, the periodic-review (s, S) policy, known as the min-max policy in practice, has received lots of attention in the literature (see, e.g., Zipkin [9]). In practice, the inventory control policies can be far more complex than these simple policies because one has to take the batch sizes and the minimum and maximum order quantities into account. In addition, many allocation rules other than FCFS are used in practice and studied in the literature, such as priority rules, fair-share rules, and the FCFS rule without the "committed stock" (Section 3.3). It would be interesting to apply the existing methods or develop new method to characterize and evaluate supply chains with these policies and allocation rules.

Appendix

Proof of Proposition 2.3. We can regard the supply system as a queue with $IO(t)$ being the number of jobs in system at time t. Due to the Poisson demand, IO has the same distribution as the number of jobs in the queue seen by a departing job in steady state because all departures see the system in its time averages [34, Theorems 7.1–7.2]. Indeed, the latter is the demand during the transit time of the departing job which has a distribution identical to $D(\infty \mid L]$ (again due to Poisson demand). □

Proof of Proposition 3.1. To prove the first statement, note that the CTMC of $\overline{IP}$ is ergodic (due to finite state space) and it has a unique steady-state distribution. Because the uniform distribution is stationary for the inventory position vector [101, page 64], $\overline{IP}$ is uniformly distributed in $\mathcal{S}$. The uniform distribution of $\overline{IP}$ immediately implies the independence of the inventory positions.

Clearly, the interarrival times between two orders placed by stage j may not follow exponential distribution. Thus, to prove the third statement, we utilize "ASTA" by Melamed and Whitt [102]. By the second statement, IP_{j+1} is independent of the order placement at stage j. Therefore, the weak lack of anticipation assumption (see [102, Definition 1]) holds for IP_{j+1} and the order process by stage j. Because $E(IP_{j+1})$ does not depend on t, it follows from Theorem 2 of Melamed and Whitt [102] that, in steady state, each order placed by stage j sees IP_{j+1} in its time averages. □

Acknowledgments

The authors are grateful to the comments made by the review team that have allowed us to improve the paper. The research of the first author is partially supported by NSF Contract CMMI-0758069, Masdar Institute of Science and Technology (MIST), Bayer Business Services, and SAP. The second author was supported in part by the National Science Foundation Career Award CMMI-0747779.

References

[1] H. L. Lee and C. Billington, "Material management in decentralized supply chains," *Operations Research*, vol. 41, no. 5, pp. 835–847, 1993.
[2] S. C. Graves and S. P. Willems, "Optimizing strategic safety stock placement in supply chains," *Manufacturing and Service Operations Management*, vol. 2, no. 1, pp. 68–83, 2000.
[3] G. Lin, M. Ettl, S. Buckley et al., "Extended-enterprise supply-chain management at IBM personal systems group and other divisions," *Interfaces*, vol. 30, no. 1, pp. 7–25, 2000.
[4] S. Nahmias, *Production and Operations Analysis*, McGraw-Hill/Irwin, Boston, Mass, USA, 4th edition, 2000.
[5] J. F. Shapiro, "Mathematical programming models and methods for production planningand scheduling," in *Handbooks in Operations Research and Management Science*, vol. 4, North-Holland, Amsterdam, The Netherlands, 1993.
[6] R. Kapuscinski and S. Tayur, "Optimal policies and simulation-based optimization for capacitated production inventory systems," in *Quantitative Models for Supply Chain Management*, S. Tayur, R. Ganeshan, and M. J. Magazine, Eds., Kluwer Academic Publishers, Boston, Mass, USA, 1998.
[7] C. R. Sox, P. L. Jackson, A. Bowman, and J. A. Muckstadt, "Review of the stochastic lot scheduling problem," *International Journal of Production Economics*, vol. 62, no. 3, pp. 181–200, 1999.
[8] A. Federgruen, "Centralized planning models for multi-echelon inventory systems underuncertainty," in *Handbooks in OR & MS*, S. C. Graves et al., Ed., vol. 4, North Holland, Amsterdam, The Netherlands, 1993.
[9] P. Zipkin, *Foundations of Inventory Management*, McGraw Hill, Boston, Mass, USA, 2000.
[10] E. L. Porteus, *Foundations of Stochastic Inventory Theory*, Stanford University Press, Stanford, Calif, USA, 2002.
[11] S. C. Graves and S. P. Willems, "Supply chain design: safety stock placement and supplychain configuration," in *Handbooks in Operations Research and Management Science Vol. 11, Supply Chain Management: Design, Coordination and Operation*, A. G. de Kok and S. C. Graves, Eds., North-Holland, Amsterdam, The Netherlands, 2003.
[12] G. Hadley and T. M. Whitin, *Analysis of Inventory Systems*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1963.
[13] F. Chen, "On (R, NQ) policies in serial inventory systems," in *Quantitative Model for Supply Chain Management*, S. Tayur, R. Ganeshan, and M. J. Magazine, Eds., Kluwer Academic Publishers, Boston, Mass, USA, 1998.

[14] T. G. de Kok and J. C. Fransoo, "Planning supply chain operations: definition and comparisonof planning concepts," in *Handbooks in Operations Researchand Management Science, Vol. 11: Supply Chain Management*, A. G. de Kok and S. C. Graves, Eds., Elsevier, Amsterdam, The Netherlands, 2003.

[15] J. S. Song and P. Zipkin, "Supply chain operations: assemble-to-order systems," in *Handbooks in Operations Research and Management Science, Vol. 11: Supply Chain Management*, A. G. de Kok and S. C. Graves, Eds., Elsevier, Amsterdam, The Netherlands, 2003.

[16] S. Axsater, "Supply chain operations: serial and distribution inventory systems," in *Handbooks in Operations Research and Management Science, Vol. 11: Supply Chain Management: Design, Coordination and Operation*, A. G. de Kok and S. C. Graves, Eds., North-Holland, Amsterdam, The Netherlands, 2003.

[17] R. Kaplan, "A dynamic inventory model with stochastic lead times," *Management Science*, vol. 16, pp. 491–507, 1970.

[18] J. S. Song and P. H. Zipkin, "Inventory control with information about supply conditions," *Management Science*, vol. 42, no. 10, pp. 1409–1419, 1996.

[19] P. Zipkin, "Stochastic lead-times in continuous-time inventory models," *Naval Research Logistics Quarterly*, vol. 33, pp. 763–774, 1986.

[20] A. Svoronos and P. Zipkin, "Evaluation of one-for-one replenishment policies for multiechelon inventory systems," *Management Science*, vol. 37, no. 1, pp. 68–83, 1991.

[21] C. Sherbrooke, "METRIC: a multi-echelon technique for recoverable item control," *Operations Research*, vol. 16, pp. 122–141, 1968.

[22] I. Sahin, "On the stationary analysis of continuous review (*s, S*) inventory systems with constant lead times," *Operations Research*, vol. 27, no. 4, pp. 717–729, 1979.

[23] C. Palm, "Analysis of the Erlang traffic formula for busy signal arrangements," *EricssonTechnics*, vol. 5, pp. 39–58, 1938.

[24] J. A. Muckstadt, *Analysis and Algorithms for Service Parts Supply Chains*, Springer, NewYork, NY, USA, 2000.

[25] H. P. Galliher, P. M. Morse, and M. Simond, "Dynamics of two classes of continuous reviewinventory systems," *Operations Research*, vol. 7, pp. 362–384, 1959.

[26] W. K. Kruse, "Waiting time in a continuous review (*s, S*) inventory system with constant lead times," *Operations Research*, vol. 29, no. 1, pp. 202–207, 1981.

[27] W. H. Hausman, H. L. Lee, and A. X. Zhang, "Joint demand fulfillment probability in a multi-item inventory system with independent order-up-to policies," *European Journal of Operational Research*, vol. 109, no. 3, pp. 646–659, 1998.

[28] S. Axsater, "Simple solution procedures for a class of two-echelon inventory problems," *Operations Research*, vol. 38, no. 1, pp. 64–69, 1990.

[29] P. Zipkin, "Evaluation of base-stock policies in multiechelon inventory systems with compound-Poisson demands," *Naval Research Logistics*,, vol. 38, pp. 397–412, 1991.

[30] Y. Zhao and D. Simchi-Levi, "Performance analysis and evaluation of assemble-to-order systems with stochastic sequential lead times," *Operations Research*, vol. 54, no. 4, pp. 706–724, 2006.

[31] D. Simchi-Levi and Y. Zhao, "Safety stock positioning in supply chains with stochastic lead times," *Manufacturing and Service Operations Management*, vol. 7, no. 4, pp. 295–318, 2005.

[32] R. Forsberg, "Optimization of order-up-to-*S* policies for two-level inventory systems with compound Poisson demand," *European Journal of Operational Research*, vol. 81, no. 1, pp. 143–153, 1995.

[33] Y. Zhao, "Evaluation and optimization of installation base-stock policies in supply chainswith compound Poisson processes," *Operations Research*, vol. 56, pp. 437–452, 2008.

[34] V. G. Kulkarni, *Modeling and Analysis of Stochastic Systems*, Chapman & Hall, New York, NY, USA, 1995.

[35] S. Axsater, "Optimization of order-up-to-*S* policies in two-echelon inventory systems with periodic review," *Naval Research Logistics*, vol. 40, pp. 245–253, 1993.

[36] S. Axsater, "Continuous review policies for multi-level inventory systems with stochastic demand," in *Logistics of Production and Inventory*, S. Graves, A. Rinnooy Kan, and P. Zipkin, Eds., Elsevier; North-Holland, Amsterdam, The Netherlands, 1993.

[37] S. Axsater, "Exact analysis of continuous review (*R, Q*) policies in two-echelon inventory systems with compound Poisson demand," *Operations Research*, vol. 48, no. 5, pp. 686–696, 2000.

[38] T. Boyaci and G. Gallego, "Serial production/distribution systems under service constraints," *Manufacturing and Service Operations Management*, vol. 3, no. 1, pp. 43–50, 2001.

[39] K. H. Shang and J. S. Song, "A closed-form approximation for serial inventory systems and its application to system design," *Manufacturing and Service Operations Management*, vol. 8, no. 4, pp. 394–406, 2006.
[40] S. Axsater and K. Rosling, "Notes: installation vs. Echelon stock policies for multilevelinventory control," *Management Science*, vol. 39, pp. 1274–1280, 1993.
[41] S. Axsater and L. Juntti, "Comparison of echelon stock and installation stock policies for two-level inventory systems," *International Journal of Production Economics*, vol. 45, no. 1–3, pp. 303–310, 1996.
[42] S. Axsater, "Simple evaluation of echelon stock (*R, Q*) policies for two-level inventory systems," *IIE Transactions*, vol. 29, no. 8, pp. 661–669, 1997.
[43] F. Chen and Y. S. Zheng, "One-warehouse multiretailer systems with centralized stock information," *Operations Research*, vol. 45, no. 2, pp. 275–287, 1997.
[44] F. Chen and Y. S. Zheng, "Evaluating echelon stock (*R, nQ*) policies in serial production/inventory systems with stochastic demand," *Management Science*, vol. 40, no. 10, pp. 1262–1275, 1994.
[45] G. J. van Houtum and W. H. M. Zijm, "Computational procedures for stochastic multi-echelon production systems," *International Journal of Production Economics*, vol. 23, no. 1–3, pp. 223–237, 1991.
[46] G. J. van Houtum and W. H. M. Zijm, "Incomplete convolutions in production and inventory models," *OR Spectrum*, vol. 19, no. 2, pp. 97–107, 1997.
[47] F. Chen and Y. S. Zheng, "Lower bounds for multi-echelon stochastic inventory systems," *Management Science*, vol. 40, pp. 1426–1443, 1994.
[48] G. Gallego and P. Zipkin, "Stock positioning and performance estimation in serial production transportation systems," *Manufacturing and Service Operations Management*, vol. 1, pp. 77–88, 1999.
[49] R. Badinelli, "A model for continuous-review pull policies in serial inventory systems," *Operations Research*, vol. 40, pp. 142–156, 1992.
[50] F. Chen, "Echelon reorder points, installation reorder points, and the value of centralizeddemand information," *Management Science*, vol. 44, pp. 0221–0234, 1998.
[51] S. C. Graves, "Multi-echelon inventory model for a repairable item with one-for-one replenishment," *Management Science*, vol. 31, no. 10, pp. 1247–1256, 1985.
[52] K. H. Shang and J. S. Song, "News vendor bounds and heuristic for optimal policies in serial supply chains," *Management Science*, vol. 49, no. 5, pp. 618–638, 2003.
[53] F. Chen and Y. S. Zheng, "Near-optimal echelon-stock (*R, nQ*) policies in multistage serial systems," *Operations Research*, vol. 46, no. 4, pp. 592–602, 1998.
[54] S. Axaster, "Evaluation of installation stock based (*R, Q*)-policies for two-level inventory systems with poisson demand," *Operations Research*, vol. 46, no. 3, pp. S135–S145, 1998.
[55] K. L. Cheung and W. H. Hausman, "Exact performance evaluation for the supplier in a two-echelon inventory system," *Operations Research*, vol. 48, no. 4, pp. 646–653, 2000.
[56] R. M. Simon, "Stationary properties of a two- echelon inventory model for low demand items," *Operations Research*, vol. 19, no. 3, pp. 761–773, 1971.
[57] K. Shanker, "Exact analysis of a two-echelon inventory system for recoverable items under batch inspection policy," *Naval Research Logistics Quarterly*, vol. 28, no. 4, pp. 579–601, 1981.
[58] J. A. Muckstadt, "A model for a multi-item, multi-echelon, multi-indenture inventory system," *Management Science*, vol. 20, no. 4, pp. 472–481, 1973.
[59] C. C. Sherbrooke, "VARI-METRIC: improved approximations for multi-indenture, multiechelon availability models," *Operations Research*, vol. 34, no. 2, pp. 311–319, 1986.
[60] A. J. Clark and H. Scarf, "Optimal policies for a multi-echelon inventory problem," *Management Science*, vol. 50, no. 12, pp. 1782–1795, 2004.
[61] A. Federgruen and P. Zipkin, "Approximation of dynamic, multi-location production and inventory problems," *Management Science*, vol. 30, no. 1, pp. 69–84, 1984.
[62] G. P. Cachon, "Exact evaluation of batch-ordering inventory policies in two-echelon supply chains with periodic review," *Operations Research*, vol. 49, no. 1, pp. 79–98, 2001.
[63] S. C. Graves, "A multiechelon inventory model with fixed replenishment intervals," *Management Science*, vol. 42, no. 1, pp. 1–18, 1996.
[64] B. Deuermeyer and L. B. Schwarz, "A model for the analysis of system service level in warehouse/retailer distribution systems: the identical retailer case," in *Multilevel Production/Inventory Control Systems: Theory and Practice*, L. Schwarz, Ed., Elsevier; North-Holland, Amsterdam, The Netherlands, 1981.
[65] H. L. Lee and K. Moinzadeh, "Two-parameter approximations for multi-echelon repairable inventory models with batch ordering policy," *IIE Transactions*, vol. 19, no. 2, pp. 140–149, 1987.

[66] H. L. Lee and K. Moinzadeh, "Operating characteristics of a two-echelon inventory systemfor repairable and consumable items under batch ordering and shipment policy," *Naval Research Logistics Quarterly*, vol. 34, pp. 365–380, 1987.
[67] A. Svoronos and P. Zipkin, "Estimating the performance of multi-level inventory system," *Operations Research*, vol. 36, no. 1, pp. 57–72, 1988.
[68] R. Forsberg, "Exact evaluation of (*R, Q*)-policies for two-level inventory systems with Poisson demand," *European Journal of Operational Research*, vol. 96, no. 1, pp. 130–138, 1997.
[69] M. Fleischmann, R. Kuik, and R. Dekker, "Controlling inventories with stochastic item returns: a basic model," *European Journal of Operational Research*, vol. 138, no. 1, pp. 63–75, 2002.
[70] V. Deshpande, M. A. Cohen, and K. Donohue, "A threshold inventory rationing policy for service-differentiated demand classes," *Management Science*, vol. 49, no. 6, pp. 683–703, 2003.
[71] K. L. Cheung and W. H. Hausman, "Multiple failures in a multi-item spares inventory model," *IIE Transactions*, vol. 27, no. 2, pp. 171–180, 1995.
[72] Y. Akçay and S. H. Xu, "Joint inventory replenishment and component allocation optimization in an assemble-to-order system," *Management Science*, vol. 50, no. 1, pp. 99–116, 2004.
[73] J. S. Song and D. D. Yao, "Performance analysis and optimization of assemble-to-order systems with random lead times," *Operations Research*, vol. 50, no. 5, pp. 889–903, 2002.
[74] Y. Lu, J. S. Song, and D. D. Yao, "Order fill rate, leadtime variability, and advance demand information in an assemble-to-order system," *Operations Research*, vol. 51, no. 2, pp. 292–308, 2003.
[75] Y. Lu, J. S. Song, and D. D. Yao, "Backorder minimization in multiproduct assemble-to-order systems," *IIE Transactions*, vol. 37, no. 8, pp. 763–774, 2005.
[76] Y. Lu and J. S. Song, "Order-based cost optimization in assemble-to-order systems," *Operations Research*, vol. 53, no. 1, pp. 151–169, 2005.
[77] J. Gallien and L. M. Wein, "A simple and effective component procurement policy for stochastic assembly systems," *Queueing Systems*, vol. 38, no. 2, pp. 221–248, 2001.
[78] S. Dayanik, J. S. Song, and S. H. Xu, "The effectiveness of several performance bounds for capacitated assemble-to-order systems," *Manufacturing and Service Operations Management*, vol. 5, no. 3, pp. 230–251, 2003.
[79] J. S. Song, "On the order fill rate in a multi-item, base-stock inventory system," *Operations Research*, vol. 46, no. 6, pp. 831–845, 1998.
[80] J. S. Song, "Order-based backorders and their implications in multi-item inventory systems," *Management Science*, vol. 48, no. 4, pp. 499–516, 2002.
[81] J. S. Song, "Note on assemble-to-order systems with batch ordering," *Management Science*, vol. 46, no. 5, pp. 739–743, 2000.
[82] A. X. Zhang, "Demand fulfillment rates in an assemble-to-order system with multiple products and dependent demands," *Production and Operations Management*, vol. 6, no. 3, pp. 309–323, 1997.
[83] N. Agrawal and M. A. Cohen, "Optimal material control in an assembly system with component commonality," *Naval Research Logistics*, vol. 48, no. 5, pp. 409–429, 2001.
[84] T. G. de Kok, "Evaluation and optimization of strongly ideal Assemble-To-Order systems," Tech. Rep., Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 2003.
[85] M. Ettl, G. E. Feigin, G. Y. Lin, and D. D. Yao, "Supply network model with base-stock control and service requirements," *Operations Research*, vol. 48, no. 2, pp. 216–232, 2000.
[86] J. Shi and Y. Zhao, "Technical note: some structural results on acyclic supply chains," *Naval Research Logistics*, vol. 57, no. 6, pp. 605–613, 2010.
[87] E. Feitzinger and H. L. Lee, "Mass customization through postponement," *Harvard Business Review*, vol. 75, pp. 116–121, 1997.
[88] G. DeCroix, J. S. Song, and P. Zipkin, "A series system with returns: stationary analysis," *Operations Research*, vol. 53, no. 2, pp. 350–362, 2005.
[89] G. A. DeCroix and P. H. Zipkin, "Inventory management for an assembly system with product or component returns," *Management Science*, vol. 51, no. 8, pp. 1250–1265, 2005.
[90] J. Buzacott, S. Price, and J. Shanthikumar, "Service level in multi-stage MRP and base-stock controlled production systems," in *New Directions for Operations Research in Manufacturing*, G. Fandel, T. Gulledge, and A. Hones, Eds., Springer, Berlin, Germany, 1991.
[91] J. S. Song, S. H. Xu, and B. Liu, "Order-fulfillment performance measures in an assemble-to-order system with stochastic leadtimes," *Operations Research*, vol. 47, no. 1, pp. 131–149, 1999.
[92] Y. J. Lee and P. Zipkin, "Tandem queues with planned inventories," *Operations Research*, vol. 40, no. 5, pp. 936–947, 1992.

[93] P. Glasserman and S. Tayur, "A simple approximation for a multistage capacitated production-inventory system," *Naval Research Logistics*, vol. 43, no. 1, pp. 41–58, 1996.
[94] L. Liu, X. Liu, and D. D. Yao, "Analysis and optimization of a multistage inventory-queue system," *Management Science*, vol. 50, no. 3, pp. 365–380, 2004.
[95] P. Glasserman and S. Tayur, "Sensitivity analysis for base-stock levels in multiechelon production-inventory systems," *Management Science*, vol. 41, no. 2, pp. 263–281, 1995.
[96] S. C. Graves and S. P. Willems, "Strategic inventory placement in supply chains: non stationary demand," Tech. Rep., Sloan School of Management, MIT, Cambridge, Mass, USA, 2005.
[97] N. Erkip, W. H. Hausman, and S. Nahmias, "Optimal centralized ordering policies in multi-echelon inventory systems with correlated demands," *Management Science*, vol. 36, no. 3, pp. 381–392, 1990.
[98] G. D. L. Schrage, "Centralized ordering policies in a multi-warehouse system with lead time and random demand," in *Multi-level Production/Inventory Systems: Theory and Practice*, L. B. Schwarz, Ed., pp. 51–67, North-Holland, Amsterdam, The Netherlands, 1981.
[99] L. Dong and H. L. Lee, "Optimal policies and approximations for a serial multiechelon inventory system with time-correlated demand," *Operations Research*, vol. 51, no. 6, pp. 969–980, 2003.
[100] V. A. Truong, R. Levi, and R. O. Roundy, "Provably nearly optimal balancing policies formulti-echelon stochastic inventory control models," Tech. Rep., Cornell University, New York, NY, USA, 2006.
[101] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 2, John Wiley & Sons, New York, NY, USA, 2nd edition, 1971.
[102] B. Melamed and W. Whitt, "On arrivals that see time averages," *Operations Research*, vol. 38, no. 1, pp. 156–172, 1990.