Chapter 11*
Supply Chain Operations: Assemble-to-Order Systems

Jing-Sheng Song
Graduate School of Management
University of California, Irvine, CA 92697
Tel. (949) 824-2482, e-mail: jssong@uci.edu

Paul Zipkin
Fuqua School of Business
Duke University, Durham, NC 27708
Tel. (919) 660-7853, e-mail: paul.zipkin@duke.edu

September 2001
Revised November 2002, January 2003

Abstract

This chapter reviews the research to date on assemble-to-order systems. It covers modeling issues and analytical methods, and also summarizes managerial insights gained from the research. At the end, it identifies some directions for future research.

*Forthcoming in *Handbooks in Operations Research and Management Science*, Vol. XXX: *Supply Chain Management*, T. de Kok and S. Graves (eds.), North-Holland.

1 Introduction

An *assemble-to-order* (or *ATO*) *system* includes several *components* and several *products*. Demands occur only for products, but the system keeps inventory only of components. To make each product requires a particular selection of components, comprising only a subset of them, but possibly several units of certain ones. Some or all components are shared by several products. The time to assemble a product from its components is negligible. The time to acquire or produce a component, however, is substantial. A product is assembled only in response to demand. See Figure 1.

[INSERT FIGURE 1 HERE.]

A *configure-to-order* (or *CTO*) *system* is a special case. The components are partitioned into subsets, and the customer *selects* components from those subsets. A computer, for example, is configured by selecting a processor from several options, a monitor from several options, etc. The difference between a CTO system and an ATO system is important at the demand-elicitation level. At the operational level, however, the differences are minor. Our discussion focuses on general ATO systems.

Such systems have been employed for some time in various industries, but lately their popularity has soared. An ATO system is an efficient way to deliver a high level of product variety to customers, while maintaining reasonable response times and costs.

One well-known ATO system (actually, a CTO system) is Dell Computer's. Dell lets the customer select among several processors, monitors, disk drives, etc. – these are the components. Thus, the number of products (combinations of options) is huge. This approach has been so successful that most other makers of personal computers are adopting similar systems. Indeed, the ATO approach has become widespread throughout the electronics industry. The major U.S. automobile companies are studying ambitious ATO systems for the assembly of cars (Kerwin (2000)). (In a real system, product assembly may take some time, though not more than customers are willing to wait.)

Certain other types of systems have the same structure. Consider a mail-order or e-commerce retailer, which maintains inventories of the items in its catalogue. The items correspond to components, and a product is any combination of them. The assembly of a product entails picking out the items in the customer's order and packaging them. Also, consider the problem of stocking spare parts for the repair of equipment. The parts are the components, and a product is a particular type

of repair job, requiring particular parts. The parts may be located at a central point, where equipment needing service arrives (e.g., vehicles), or the parts may travel to stationary equipment (as in field service of computers, copiers, and factory machines). In either case, the part requirements of a job are usually unknown in advance.

This chapter reviews the research to date on ATO systems. It covers modeling issues and analytical methods, and also summarizes managerial insights gained from the research. At the end, it identifies some directions for future research. (See Song and Yao (2001) for other related articles.)

Two special cases are worth identifying. An *assembly system* has just one product, and a *distribution system* has just one component. The key issue in an assembly system is the *coordination* of the components, while the key issue in a distribution system is the *allocation* of the component among the products. (This assembly system is a special one, due to the negligible assembly time, which implies that there is no reason to assemble the product in advance of demand. The distribution system is special in the same way.) An ATO system combines the elements of assembly and distribution, and so must resolve both coordination and allocation issues. This is what makes ATO systems difficult to analyze, design and manage. See Figure 2.

[INSERT FIGURE 2 HERE.]

Research in this area has two major goals. One is efficient operations. *Given* a particular system design, i.e., a line of products and a set of components, this work aims to evaluate its performance under various conditions, including inventory levels of the components. It also seeks to find good operating policies, including inventory levels that balance cost and customer service. The second research goal is to understand the impacts of alternative system designs, for example, the effects of designing several products to share common components.

In Section 2, we discuss one-period models. Section 3 focuses on multi-period, discrete-time models, while Section 4 presents continuous-time models. Section 5 summarizes research on system design. Section 6 points out some future directions.

2 One-Period Models

This section focuses on one-period models. Some systems can be understood fruitfully in a static framework. Either each time period really can be treated in isolation, or the model can serve as a

myopic heuristic for more complex scenarios. As in the classic news-vendor model (one component and one product), there is no need to distinguish between backorders and lost sales, and we suppress procurement leadtimes. We present a fairly general formulation of the problem and then discuss specific works in the literature.

The sequence of events within the period is as follows: (1) Components are produced or acquired. (2) Demands for the products are realized. (3) Components are allocated to products, and costs are assessed on the ending situation. The basic approach is stochastic linear programming with simple recourse. We consider these events in reverse order.

The problem for stage (3) can be formulated as follows: Define

m = total number of components
n = total number of products
i = index for components (subscript)
j = index for products (superscript)
a_i^j = units of component i required to make one unit of product j, $A = (a_i^j)$ (matrix)
d^j = demand for product j, $\mathbf{d} = (d^j)$ (vector)
y_i = supply of component i, $\mathbf{y} = (y_i)$ (vector)
p^j = penalty cost for unit shortage of product j, $\mathbf{p} = (p^j)$ (vector)
h_i = cost for unit excess of component i, $\mathbf{h} = (h_i)$ (vector)
z^j = production of product j, $\mathbf{z} = (z^j)$ (vector)
w^j = shortage of product j, $\mathbf{w} = (w^j)$ (vector)
x_i = excess of component i, $\mathbf{x} = (x_i)$ (vector)

(The parameters a_i^j, d^j, and y_i are nonnegative real numbers, the p^j are positive real numbers, and the h_i are any real numbers. A negative h_i represents a salvage value.) The problem is then

$$\begin{aligned} \text{(P3)} \qquad \hat{G}(\mathbf{y},\mathbf{d}) = \text{minimize} \quad & \mathbf{hx} + \mathbf{pw} \\ \text{subject to} \quad & \\ & A\mathbf{z} + \mathbf{x} = \mathbf{y} \\ & \mathbf{z} + \mathbf{w} = \mathbf{d} \\ & \mathbf{w}, \mathbf{x}, \mathbf{z} \geq \mathbf{0} \end{aligned}$$

This is a linear program. (It can be quite large if there is a large number of products.) The minimal cost function $\hat{G}(\mathbf{y},\mathbf{d})$ is convex.

Several modeling issues are worth noting: This formulation treats demands, supplies, etc. as continuous. In some situations these quantities may actually be discrete. If the model is revised accordingly, it becomes an integer linear program. This model is of course much harder to solve than the original.

Also, this formulation assumes that the stockout penalty cost for each product is linear in the shortage. An alternative formulation replaces the term $\mathbf{pw}$ in the objective with $\mathbf{p1}(\mathbf{w})$, where $\mathbf{1}(\mathbf{w}) = (\mathbf{1}(w^j))$ is the vector of 0-1 variables indicating which of the w^j are positive. Here, a cost p^j is incurred if there is *any* shortage of product j, regardless of how much. This model too is harder to solve than the original. Yet another formulation has a term $p \max_j[\mathbf{1}(\mathbf{w})]$ in the objective. Here, a cost p is incurred if there is *any* shortage of *any* product. Alternatively, any of these service measures can be constrained instead of penalized. (In our opinion, the original linear formulation better represents service to customers, in addition to being more tractable computationally.)

Lastly, the formulation assumes that a demand for a product must be filled entirely or not at all. This makes sense when the product is really a product, i.e., incomplete with any of its components missing. In a retailing setting, however, the situation is less clear. The customer gets *some* benefit from partial fulfillment.

Now let us examine two special cases. First, consider a distribution system (one component, so we can omit the index i). Here, the model reduces to a continuous knapsack problem, which is easy to solve: For each product compute the ratio p^j/a^j, the shortage cost per unit of the component. Select the product with the largest ratio, and satisfy its demand as much as possible, i.e., $z^j = \min\{y/a^j, d^j\}$. If there is any of the component left, satisfy as much as possible the demand for the product with the *next* largest ratio. Continue in this manner until the component is exhausted, or all demands are filled. (This assumes $h \geq 0$. If $h < 0$, i.e., there is a positive salvage value $-h$, omit any product with $p^j/a^j < -h$. Here, the solution can have both unfilled demand and remaining component.) Thus, in this case, the best allocation of the component is determined by a priority ranking of the products, which depends on the cost and usage data only, not demand conditions.

Next, consider an assembly system (one product, so we can omit the index j). Here too, the model is easy to solve: Set the production level z to fill all demand, if possible, or else to use up the most limiting component. i.e., $z = \min\{\min\{y_i/a_i\}, d\}$.

Now let us move to stage (1). Here, $\mathbf{d}$ is a random variable. Let $\mathbf{x}_0$ be the initial component inventory vector. The expected cost at stage (3) given $\mathbf{y}$ is $G(\mathbf{y}) = E_{\mathbf{d}}[\hat{G}(\mathbf{y}, \mathbf{d})]$. Let $c(\mathbf{y} - \mathbf{x}_0)$

denote the cost of acquiring the components. The problem, then, is

$$\begin{array}{lll} \text{(P1)} & \text{minimize} & c(\mathbf{y}-\mathbf{x}_0)+G(\mathbf{y}) \\ & \text{subject to} & \mathbf{y}\geq\mathbf{x}_0 \end{array}$$

The function $G(\mathbf{y})$ is convex. If $c(\cdot)$ too is convex (e.g., linear), then the overall objective function is convex, and so the model is relatively easy to solve.

In particular, assume $c(\cdot)$ is linear with cost-coefficient vector $\mathbf{c}$. Let $\mathbf{y}^*$ be the global minimizer of $\mathbf{cy}+G(\mathbf{y})$. Then, the optimal ordering policy is a base-stock policy with base-stock level $\mathbf{y}^*$. That is, if $\mathbf{x}_0\leq\mathbf{y}^*$, then order up to $\mathbf{y}^*$. Standard stochastic linear programming techniques can be employed to compute G and $\mathbf{y}^*$.

In the alternative formulations mentioned above, where the objective of (P3) includes $\mathbf{p1}(\mathbf{w})$ or something similar, the corresponding G includes terms representing stockout probabilities. Except in special cases, the function G is not convex.

Some elements of this formulation are due to Gerchak and Henig (1986). Assuming $\mathbf{x}_0=\mathbf{0}$, they compare the ATO system with a make-to-stock (MTS) system in which the assembly is performed before demands are realized. In the latter, there is no need to solve problem (P3), and $\mathbf{y}=A\mathbf{z}$. The optimal value of $\mathbf{z}$ is the solution of problem (P1) with $A\mathbf{z}$ replacing $\mathbf{y}$. Problem (P1) now reduces to n separate news-vendor problems. Let $\mathbf{z}^{\mathrm{S}}$, $\mathbf{y}^{\mathrm{S}}$, and C^{S} denote the optimal solution and cost for this MTS setting, and let $\mathbf{z}^{\mathrm{O}}$, $\mathbf{y}^{\mathrm{O}}$ and C^{O} be their counterparts in the original ATO setting (by solving both (P3) and (P1)). Then,

a) $C^{\mathrm{S}}\geq C^{\mathrm{O}}$. (The cost is lower when assembly is postponed until demand is realized.)

b) $\mathbf{z}^{\mathrm{O}}\geq\mathbf{z}^{\mathrm{S}}$. (More demand is fulfilled in the ATO system.)

c) If component i is product-specific (used by just one product), then $y_i^{\mathrm{O}}\geq y_i^{\mathrm{S}}$.

d) On the other hand, one cannot predict the ordering of y_i^{O} and y_i^{S} if i is a common component (used by more than one product).

The optimal stock of a common component may be higher or lower than the combined optimal stocks of the specialized components it replaces. Thus, moving to an ATO system need not reduce inventories. It does improve overall performance, but the improvement may show up instead in reduced stockouts.

The study of Gerchak and Henig (1986) builds on earlier work comparing simple systems with and without common components among the products. See Baker, Magazine and Nuttle (1986), Gerchak, Magazine and Gamble (1988), and references therein. They assume a single constraint on the overall stockout probability. In our notation, the problem is

$$\begin{aligned} &\text{minimize} && \mathbf{cy} \\ &\text{subject to} && \mathsf{P}(A\mathbf{d} \leq \mathbf{y}) \geq \beta \end{aligned}$$

Here, A is a $0-1$ matrix, and β is a prespecified service level. For the system without common components, each column of A has only one non-zero entry. In the system with common components, A has fewer columns, and some columns have several non-zero entries. This is obtained by adding some columns of the original A, i.e., combining some components into common components.

It is shown that the total inventory investment required to meet the constraint is lower with a common component than without it. Certain additional qualitative properties of the solution derived in the first paper are, however, shown in the second paper to hold only in special cases. Thus, while commonality is certainly a good thing, all else being equal, its detailed effects are hard to predict. See also Bagchi and Gutierrez (1992), Eynan (1996) and Eynan and Rosenblatt (1996) for variations of these studies and similar conclusions. Eynan (1996) provides a summary of this line of research and additional references.

The repair-kit problem also is related. The repair kit carried by a repairman is a multi-item inventory. A demand is a repair job, which may require several different items in the kit. Typically, each job requires either one or zero unit of each item, so the matrix A contains only entries 0 and 1. The problem is to determine an optimal kit that minimizes either the expected inventory and penalty cost, or the inventory cost subject to a constraint on the job-completion rate. Some models assume that the kit can be restocked after each job, therefore the decision variables are binary; we only need to decide whether to carry an item or not. See, e.g., Smith et al. (1980) and Graves (1982). Other models consider multiple units of each item in stock; the performance measure is the expected number of jobs which can be completed before stock out or the probability of serving all jobs arriving in a fixed period of time. There is no inventory replenishment during the period. The difference between this model and the one-period model (P1) and (P3) is that, within the period, the allocation is dynamic as demand occurs, and it follows the FCFS rule. Network-flow and combinatorial techniques are developed to solve the optimization problem. See, e.g., Mamer and Smith (1982, 1985) and Brumelle and Granot (1993). Mamer and Smith (2001) review this

literature.

Swaminathan and Tayur (1999) consider an extension of the basic model above. Between components and products is another layer of *sub-assemblies*. A sub-assembly can be made from components in stage (1), before demand is realized. The model explicitly includes a production resource (e.g., capacity or time) that is used in the creation of products from components and/or sub-assemblies in stage (3). To make a product from sub-assemblies consumes less of this resource than making it directly from components. See Chapter 8 of this handbook for a detailed discussion.

3 Multi-Period, Discrete-Time Models

This section focuses on discrete-time, multi-period models. Within a single period, the decisions and the sequence of events are the same as in the one-period problem. However, additional complications arise when we link different periods, as the ending state in one period becomes the beginning state in the next. One complication is due to leadtimes for component replenishments. When the leadtimes for different components are different, the replenishment decision in one period will affect the inventory levels in different periods in the future. The problem becomes even more complex if the leadtimes are uncertain. Another complication is how to deal with shortages – whether the unsatisfied demand in one period is backlogged or lost. In the backlogging case, there is also a partial-shipment or inventory-commitment issue. That is, if we have only part of the components a demand requests, should we ship or put aside the available components as committed inventory to this customer while waiting for those unavailable components to come? If we keep committed inventory, then we pay both the backorder cost for the unsatisfied demand and the holding cost for the committed inventory. Yet another complication is the relative priority of backlogged demand and current demand. This complication does not arise in the two special cases – one product or one component. It is problematic only when there are different products with overlapping component sets.

We first summarize results on the characterization of optimal policies and then results on performance evaluation and optimization techniques for a given type of policy.

3.1 Characterization of Optimal Policies

With all the complications mentioned above, the state space becomes quite large: We need to track not only the component inventory vector ($\mathbf{x}$), but also the backordered products ($\mathbf{w}$) and

the outstanding-order vector for each component. As a result, the literature includes only partial results for special cases: zero leadtimes, or positive leadtimes but only one product.

Consider the case with no component replenishment leadtimes and no committed inventory. Backlogged demand is merged with current demand, i.e., no priority is given to the backordered demand. For simplicity, assume stationary cost factors and component usages. Let T be the time horizon and t the time index, $t = 0, ...T.$. Use an additional subscript t to index the variables. The sequence of events within each period is the same as in the one-period model above.

Then the problem is

$$\begin{array}{lll} \text{(P)} & \text{minimize} & \mathsf{E}\{\sum_{t=0}^{T}[c(\mathbf{y}_t - \mathbf{x}_t) + \mathbf{h}\mathbf{x}_{t+1} + \mathbf{p}\mathbf{w}_{t+1}] \ \} \\ & \text{subject to} & \\ & & \mathbf{x}_{t+1} = \mathbf{y}_t - A\mathbf{z}_t \\ & & \mathbf{w}_{t+1} = \mathbf{w}_t + \mathbf{d}_t - \mathbf{z}_t \\ & & \mathbf{w}_t, \mathbf{x}_t, \mathbf{z}_t \geq \mathbf{0}, \quad \mathbf{y}_t \geq \mathbf{x}_t, \quad t = 0, ...T. \end{array}$$

For linear $c(\cdot)$ the optimal-cost function for each period is convex, and again a base-stock policy is optimal. That is, there are vectors $\mathbf{y}_t^*(\mathbf{w})$ such that, if $\mathbf{x}_t \leq \mathbf{y}_t^*(\mathbf{w}_t)$, then order up to $\mathbf{y}_t^*(\mathbf{w}_t)$. Otherwise, we know of no results characterizing the optimal policy for the general case. (However, there are some results for variations of the model. See Veinott (1965).)

Because exact solution of (P) is difficult, various computational and heuristic approaches have been developed. One approach is myopic: In each period, solve the embedded allocation problem, after the demand for that period is realized (i.e, solve for $\mathbf{z}_t$ after observing $\mathbf{d}_t$, ignoring future periods), as in (P3). Stochastic programming techniques are then used to determine the optimal order quantities for the entire horizon, as in (P1). The problem can be reformulated as a nested stochastic program. See Swaminathan and Tayur (1999).

Gerchak and Henig (1989) study a lost-sales model with stationary data and linear order cost. (The paper states that the results hold for the backlog case too, but the formulation does not completely cover that case. There are no state variables or costs for backlogs.) In particular, $\mathbf{p} = \mathbf{0}$, and $\mathbf{d}_t$ has the same distribution across t. Let r^j be the unit revenue of product j, and denote $\mathbf{r} = (r^j)$. Then, the problem can be expressed as follows:

$$\begin{array}{lll} \text{(Lost Sales)} & \text{minimize} & \mathsf{E}\{\sum_{t=0}^{T}[c(\mathbf{y}_t - \mathbf{x}_t) + \mathbf{h}\mathbf{x}_{t+1} - \mathbf{r}\mathbf{z}_t \ \} \\ & \text{subject to} & \\ & & \mathbf{x}_{t+1} = \mathbf{y}_t - A\mathbf{z}_t \end{array}$$

$$\mathbf{z}_t \leq \mathbf{d}_t$$
$$\mathbf{x}_t, \mathbf{z}_t \geq \mathbf{0}, \quad \mathbf{y}_t \geq \mathbf{x}_t, \quad t = 0, ...T.$$

They too adopt the myopic-allocation policy for this problem. That is, in each period, after the demand $\mathbf{d}_t$ is realized, solve a linear program to find the $\mathbf{z}_t$ that maximizes $\mathbf{rz}_t$ while satisfying the constraints. They show that, because of the stationary data, this myopic policy is optimal, so the multi-period solution is identical to the single-period solution. In other words, a base-stock order policy and the myopic-allocation policy are optimal. Van Mieghem and Rudi (2001) obtain some related results and explain in detail why the myopic result does not extend to the backlog case.

Hillier (2000) studies a model in which a common component is shared by all the products, and each product has a unique, product-specific component. Let $n+1$ index the common component. Thus, the matrix A has $n+1$ rows and n columns, with $a_i^i = 1$, $a_i^j = 0$ for $j \neq i$ and $i, j = 1, ..., n$, and $a_{n+1}^j = 1$ for $j = 1, ..., n$. The myopic-allocation policy above is employed. Under certain special assumptions about purchase, holding and backorder costs, as well as zero leadtimes, the paper concludes that commonality may not be beneficial if the common component is more expensive than the components it would replace.

Distribution Systems

Next, consider the special case of a distribution system, that is, a single component and multiple products (or demand classes). Topkis (1968) analyzes a model in which the order decisions can be made only at certain fixed times. The number of periods between such decisions, called a stocking cycle, equals the leadtime. Thus, at any time there is only one outstanding order, and the order is received only at the end of the stocking cycle. Also, all previous backlogs are cleared at each reorder point. Thus, the problem is essentially a single-cycle problem. It is shown that, under certain conditions, a base-stock policy is optimal for ordering, and a rationing policy (described below) is optimal for allocation in each period within a cycle.

Let us take a closer look at a cycle. Suppose the cycle length is $L+1$ periods. The problem within the cycle is a special case of (P), in which $T = L$, y_0 is fixed, and $y_t = 0$, $t = 1, ..., L$. We can formulate this problem as a dynamic program. There are two state variables. One is x_t, the inventory level at the beginning of period t. The other is $\mathbf{u}_t = \mathbf{w}_t + \mathbf{d}_t$, the vector of outstanding product demands (previous backorders plus demand in the period). Define $V_t(x, \mathbf{u})$ to be the minimal expected cost in periods $t, t+1, ..., L$, assuming period t begins with inventory level $x_t = x$ and outstanding demand $\mathbf{u}_t = \mathbf{u}$. Then,

$V_t(x, \mathbf{u}) = \min\{h_t x_{t+1} + \mathbf{p}_t \mathbf{u}_{t+1} + \mathsf{E}[V_{t+1}(x_{t+1}, \mathbf{u}_{t+1})]\}$

subject to

$$x_{t+1} = x - A\mathbf{z}_t$$
$$\mathbf{u}_{t+1} = \mathbf{u} + \mathbf{d}_{t+1} - \mathbf{z}_t$$
$$\mathbf{u}_{t+1}, x_{t+1}, \mathbf{z}_t \geq \mathbf{0}, \quad t = 0, ...T.$$

Assume $a^j = 1$ for all j. Renumber the products so that $p_t^1 \leq p_t^2 \leq \cdots \leq p_t^n$. Topkis shows that the optimal allocation policy is determined by nonnegative rationing limits $\{\tilde{z}_t^j, j = 1, ...n\}$ with $\tilde{z}_t^1 \geq \tilde{z}_t^2 \geq \cdots \geq \tilde{z}_t^n$. The rule works as follows: Start with product n. Allocate as much as possible to it, as long as the stock level does not drop below the rationing limit $\tilde{z}_t^n$. If any demand remains unfilled, stop. Otherwise, apply the same rule to product $n-1$, then $n-2$, and so forth. This rule is easy to implement, but the computation of the rationing levels is difficult. (Note that the solution to the single-period problem discussed in the last section corresponds to $\tilde{z}^j = 0$ for all j. This is because, in that case, we do not need to consider future demands.)

Sobel and Zhang (2001) study a model with no leadtime and two demand sources, one deterministic ($\bar{d}_t$) and the other stochastic ($\hat{d}_t$). Think of these as two distinct products. The deterministic demand in each period must be satisfied immediately, and the stochastic demand can be backordered. So, the allocation policy in each period is fixed: Satisfy all the deterministic demand, and then satisfy as much of the stochastic demand as possible, including backorders carried from previous periods. The order cost in each period is a fixed cost k plus a linear cost with rate c. Problem (P) now becomes

$$\text{minimize} \quad E\{\textstyle\sum_{t=0}^{T}[k\delta(y_t - x_t) + c(y_t - x_t) + hx_{t+1} + pw_{t+1}]\ \}$$

subject to

$$x_{t+1} = y_t - \bar{d}_t - z_t$$
$$w_{t+1} = w_t + \hat{d}_t - z_t$$
$$x_t, w_t \geq 0, \quad y_t \geq \max\{x_t, \bar{d}_t\}, \quad t = 0, ...T,$$

where $\delta(x) = 1$ if $x > 0$ and $\delta(x) = 0$ if $x = 0$.

Assuming positive h and p, clearly, in the optimal solution, $z_t = \min\{y_t - \bar{d}_t, w_t + \hat{d}_t\}$. Also, x_{t+1} and w_{t+1} cannot be both positive. Let $\tilde{x}_t$ be the net inventory at beginning of period t, i.e., $\tilde{x}_t = x_t - w_t$, and $\tilde{y}_t$ the net inventory after ordering. The problem can be rewritten as

$$\text{minimize} \quad E\{\textstyle\sum_{t=0}^{T}[k\delta(\tilde{y}_t - \tilde{x}_t) + c(\tilde{y}_t - \tilde{x}_t) + h(\tilde{y}_t - d_t)^+ + p(d_t - \tilde{y}_t)^+]\}$$

subject to

$$\tilde{y}_t \geq \tilde{x}_t + \bar{d}_t$$
$$\tilde{x}_{t+1} = \tilde{y}_t - d_t, \quad t = 0, ...T,$$

where $d_t = \bar{d}_t + \hat{d}_t$. Except for the special constraint, the model is the standard one. Indeed, a modified (s, S) policy is optimal. The parameters s_t and S_t are defined as usual. The optimal policy for period t is to order up to S_t (set $y_t = S_t$) if $x_t < \max\{s_t, \bar{d}_t\}$. Otherwise, do not order. (When the leadtime is positive, the analysis breaks down.)

Assembly Systems

Now, return to the general model (P), and consider the special case of an assembly system (one product). The optimal allocation policy is simple: Just satisfy as much backorders and demand as possible. It remains to determine the component replenishment policy. When all components have the same leadtime, then the inventories of all components (adjusted for usage) should be equal at all times, and the problem reduces to a single-item model. Different leadtimes, however, are challenging.

Building on Schmidt and Nahmias (1985), Rosling (1989) studies a multi-stage assembly system with deterministic leadtimes. He shows that, under some mild conditions on initial inventories, the assembly system is equivalent to a series system. So, following Clark and Scarf (1960), an echelon base-stock policy is optimal. Applying this result to the ATO system considered here (i.e., no subassemblies and no final assembly time), the optimal policy is a *balanced base-stock policy*. This is like a base-stock policy, but the components are coordinated as follows: Let L_i be the leadtime for component i. Without loss of generality, assume $L_m > \cdots > L_1$. Redefine units if necessary so that $a_i = 1$ for all i. Orders for component m follow a standard base-stock policy, as if component m were the only one. For component $i < m$, order precisely the quantity of component m ordered $(L_m - L_i)$ periods ago. Thus, the same amounts of all components arrive at each time. See Zhang (1995).

In the following, we relate this problem (with different leadtimes) to the formulation (P) and provide a proof of the optimal policy. Note that, with positive leadtimes, the order decision for component i at t does not influence the system cost until $t + L_i$. Thus, the variable y_{it} in (P) can no longer serve as a decision variable. Instead, a directly controllable variable is the inventory position, the sum of net inventory and inventory on order. Let

$\hat{x}_{it}$ = inventory position of component i at the beginning of period t before ordering, $\hat{\mathbf{x}}_t = (\hat{x}_{it})$ (vector)

$\hat{y}_{it}$ = inventory position of component i at the beginning of period t after ordering, $\hat{\mathbf{y}}_t = (\hat{y}_{it})$ (vector)

$\tilde{x}_{it}$ = net inventory of component i at the beginning of period t

$d[t, s)$ = cumulative demand in periods $t, t+1, ..., s-1$, for $s > t$

$d[t, s]$ = cumulative demand in periods $t, t+1, ..., s$, for $s > t$.

Then $x_{it} = [\tilde{x}_{it}]^+$ and $w_t = \max_i[\tilde{x}_{it}]^-$. So, the inventory-backorder cost can be expressed in terms of the $\tilde{x}_{it}$. It is well known that

$$\tilde{x}_{i,t+1} = \hat{y}_{i,t-L_i} - d[t - L_i, t]. \tag{1}$$

Since there is only one product, an optimal policy must guarantee that, after some initial periods, the net inventories of all components at the end of each period are equal. Applying (1), we have

$$\hat{y}_{i,t-L_i} - d[t - L_i, t] = \hat{y}_{m,t-L_m} - d[t - L_m, t] \tag{2}$$

for all $i < m$. This is equivalent to

$$\hat{y}_{i,t-L_i} = \hat{y}_{m,t-L_m} - d[t - L_m, t - L_i) \tag{3}$$

for all $i < m$. Thus, the optimal policy is the balanced one described above, and the problem reduces to a single-item problem with decision variable $\hat{y}_{m,t}$. In particular, the optimal policy for component m is base-stock with base-stock level s_m^*, the solution to

$$F_m(s) = p/(p + h_m),$$

where $F_m(\cdot)$ is the cumulative distribution function of $d[1, L_m]$. Each time we order component m, we order the same amount for component i, $i < m$, but at $(L_m - L_i)$ periods later. (Zhang (1995) achieves the same result by interpreting Rosling's result in the ATO setting.)

When the leadtimes are stochastic, the above approach does not work. It is no longer possible to define m so that the difference $L_m - L_i$ is always nonnegative. Song et al. (2000) consider the special case of a one-time stochastic demand. The problem is to determine when and how much of each component to order, to minimize the total expected cost. The optimal order quantity for each component is generally less than in the standard newsvendor model with no assembly structure and

no leadtime uncertainty. Several simple but reliable heuristic procedures are developed. Numerical studies indicate that, in this setting, leadtime variability often has a larger impact than demand variability. Moreover, it is better to approximate leadtime uncertainty than to ignore it.

3.2 Performance Evaluation

The optimal policy for the general system with positive leadtimes is unknown. One attractive heuristic is a base-stock order policy along with some allocation rule. Several authors focus on performance evaluation and optimization techniques for such policies, assuming that demand in each period has a multivariate normal distribution. The biggest challenge here is the numerical evaluation of such distributions.

Hausman *et al.* (1998) assume the FCFS allocation policy, so that backlogged demands are filled in order of arrival. This implies that all available inventory is committed to the earliest backlogged demands, even if those units can be used to fill later demands. The service measure of interest is the fill rate with time-window τ, the probability of filling a demand within time τ, where τ is a given nonnegative integer. This measure is hard to compute exactly. The paper focuses on a lower bound, namely, the probability of filling *all* demand in a period within time τ, denoted R_τ.

For each component i, let s_i be the base-stock level and $d_i(L_i-\tau+1)$ the demand over $L_i-\tau+1$ time periods. It is shown that

$$R_\tau = \mathsf{P}(d_i(L_i - \tau + 1) \leq s_i, \forall i). \tag{4}$$

It turns out that $R_\tau = \tilde{R}_0$, where $\tilde{R}_0$ indicates R_0 in a revised system with truncated leadtimes $\tilde{L}_i = [L_i - \tau]^+$. The paper examines the problem of maximizing the R_0, which is a multivariate normal probability, subject to a linear budget constraint. Even this objective is hard to evaluate in general, and so the paper develops heuristic methods. The best of these seems to be an *equal fractile heuristic*, which selects the base-stock levels to equalize the components' fill rates.

The model of Agrawal and Cohen (2001) minimizes the total expected component inventory costs, subject to constraints on the order fill rates. Without the constraints, the problem separates by components. The allocation policy is the following:

1. Partial FCFS: Assign the available stock of components to specific finished-product orders, and release units for delivery, even when the entire subset of components is not available. (However, the product order is considered complete only when the entire kit has been delivered.)

2. Fair-share allocation: In case of a component shortage, the available stock is allocated to the orders, based on the actual demand in the period. Specifically, for each component i, product j receives a fraction of the stock, the ratio of its demand for the component to the total component demand.

The paper develops an expression for the order fill rate under this policy. Again, this requires the evaluation of multivariate normal distributions. The paper shows that the objective function is convex and the constraints are quasi-convex, so the globally optimum solution for the optimization problem exists and is characterized by the Kuhn-Tucker conditions. This approach leads to quite different solutions from the equal-fractile heuristic of Hausman et al.

Zhang (1997) studies a similar problem using a different allocation rule. He assumes that demands in different periods are filled according to the FCFS rule. However, for demands within the same period, a product priority rule is followed. Also, the following stock-commitment policy is used: Once component units are allocated to a product as above, these units remain committed to the product, even if the demand cannot be filled due to inadequate stock of other components. Two easy-to-compute lower bounds on the order fill rate are proposed, based on properties of the multivariate normal distribution, and their performances compared through numerical experiments. The results indicate that neither bound dominates the other.

Cheng et al. (2002) assume i.i.d. replenishment leadtimes and FCFS allocation rule. They study the problem of minimizing average component inventory holding cost subject to product-family dependent fill rate constraints. They use an approximation for the fill rate in each product family, so that the constraint functions are linear functions of the item fill rates. An exact algorithm and a greedy heuristic algorithm are developed. Using the solution techniques and real data from applications in IBM, the paper further conducted numerical experiment to highlight several key benefits of the ATO operation in terms of risk pooling. Their numerical examples also show that, compared to the Make-to-Stock model, ATO can lead to substantial inventory savings and improved demand forecast accuracy.

de Kok and Visschers (1999) propose a modified base-stock policy that adapts Rosling's approach for pure assembly systems to more general ATO systems. This approach essentially fixes the allocation of components among products, before the components actually arrive in inventory. The result, instead of a series system, is a multi-level distribution system. See Chapter 13 of this handbook for details.

To summarize, the research to date has developed several plausible, reasonably effective heuristic methods. These methods are quite different, however, in spirit as well as in detail. One cannot yet draw broad conclusions about which approaches are most promising in practice.

4 Continuous-Time Models

This section reviews continuous-time models. Again, we discuss first the characterization of optimal policies and then results on performance evaluation and optimization of a given type of policy.

4.1 Characterization of Optimal Policies

Again, little is known about optimal polices in the general case, and so our discussion focuses on special cases.

For assembly systems, Chen and Zheng (1994) extend the results of Rosling (1989) discussed above in several directions. In particular, they show that the results hold for continuous as well as discrete time. So, a continuous-time, single-product ATO system again reduces to a single-item one.

Ha (1997) studies a single-item $M/M/1$ make-to-stock queue with several demand classes and lost sales. (In our terms, this is a type of distribution system.) His results have the same flavor as those of Topkis (1968). Each demand class has a rationing limit. When the inventory is at or below the limit, it is optimal to reject demands of this class in anticipation of higher-priority demands. de Vricourt et al. (1999) study a similar problem with backlogs. The optimal policy has the same form.

4.2 Performance Evaluation

For continuous-review, multi-product systems, all research to date assumes independent compound-Poisson processes for product demands. This implies that demand for each component too is compound-Poisson. Otherwise, the works differ in the modeling of the component supply and in analytical approaches.

Before summarizing individual works, we first introduce a formulation of the demand process, the stock allocation policy, and the inventory control policy, which is largely due to Song (1998).

Let $\mathcal{I} = \{1, 2, \cdots, m\}$ denote the set of component indices. Product-demand epochs form a stationary Poisson process, denoted $\{A(t), t \geq 0\}$, with rate λ. Each demand may require several components in different amounts simultaneously. For any subset of components $K \subseteq \mathcal{I}$, we say a

demand is of type K if it requests $Z_i^K > 0$ units of component $i \in K$, and 0 units in $\mathcal{I} \setminus K$. The random variable $Z^K = (Z_i^K, i \in K)$ has a known discrete probability distribution ψ^K. Assume that each order's type is independent of the other orders' types and of all other events. Also, there is a fixed probability q^K that an order is of type K, $\sum_K q^K = 1$. Thus, the type-K order stream forms a Poisson process with rate $\lambda^K = q^K \lambda$. (An order type is thus a bit different from a product, as the word is used above. A product corresponds to a fixed recipe of components. An order type has a fixed set of components, but the quantities are random.)

Let $\mathcal{K}$ be the set of all demand types, that is, $\mathcal{K} = \{K \subset \mathcal{I} : q^K > 0\}$. Note that $\mathcal{K}$ is not necessarily the set of all possible subsets of $\mathcal{I}$. For each component i, let $\mathcal{K}_i$ denote the family of subsets of $\mathcal{K}$ that contain i. The demand process for component i forms a compound Poisson process with rate $\lambda_i = \sum_{K \in \mathcal{K}_i} \lambda^K = q_i \lambda$ and batch size Z_i, the mixture of Z_i^K for all $K \in \mathcal{K}_i$.

In general, the demand model does not impose any restrictions on the batches Z_i^K among $i \in K$. It is, however, worth mentioning three important special cases:

1) *Unit Demand*: $Z_i^K = 1$ for all $i \in K$. That is, a demand of type-K requires one and only one unit of each component in K. Such a demand process is especially common when the items are relatively expensive or durable. For instance, a customer of a bookstore may buy several books but only one copy each. In the consumer market of the mail-order personal computer business, a demand typically requires one motherboard, one keyboard, one monitor, and at most one video card. Here, ψ^K concentrates on a single point.

2) *Assembly of Multiple Products*: $Z_i^K = a_i^K \xi^K$, where the a_i^K are constant positive integers and ξ^K is a positive-integer random variable. In this case, a type-K demand requests a random number ξ^K of units of product K, which has the fixed bill-of-material $a_i^K, i \in K$. Here, ψ^K is in effect a one-dimensional distribution – that of ξ^K.

3) *Pick and Pack*: Z_i^K are independent across $i \in K$. This is a reasonable approximation for demands in distribution systems, such as in mail-order retailing, especially when the items in K are not too closely related, e.g., women's sweaters and men's slacks. Here, ψ^K is the product of the marginal distributions ψ_i^K of Z_i^K.

Demands are filled on a first-come-first-served (FCFS) basis. Demands that cannot be filled immediately are backlogged. When a demand arrives and some of its required components are in

stock but others are not, we either ship the in-stock components or put them aside as committed inventory. However, a demand is considered backlogged until it is satisfied completely.

The inventory of each component is controlled by a base-stock policy, with

$$s_i := \text{ the base-stock level for component } i.$$

Let $t \geq 0$ be the continuous time variable, and for each t denote

$$\begin{aligned}
IN_i(t) &= \text{net inventory of item } i, \\
A^K(t) &= \text{number of type-}K\text{ demands by time } t, \\
D_i(t) &= \text{cumulative demand for } i \text{ by time } t \\
B^K(t) &= \text{type-}K\text{ backorder at } t \\
&= \text{number of type-}K\text{ orders that are not yet completely satisfied by } t, \\
B_i(t) &= \text{number of backorders for item } i \text{ at } t.
\end{aligned}$$

Let D_i stand for the steady-state limit of $D_i(t - L_i, t) = D_i(t) - D_i(t - L_i)$, the lead-time demand of item i. Let IN_i be the steady-state limit of $IN_i(t)$, and define B^K and B_i similarly. Also, define

$$W^K = \text{steady-state waiting time of a type-}K\text{ backorder.}$$

The performance measures of interest are, for any demand type K,

$$\begin{aligned}
f^{K,w} &= \text{type-}K\text{ order fill rate with time window } w \\
&= \text{probability of satisfying a type-}K\text{ order within a time window } w \\
&= \mathsf{P}[W^K \leq w] \\
f^K &= \text{fill rate of type-}K\text{ demand} = f^{K,0} \\
\mathsf{E}[B^K] &= \text{average number of type-}K\text{ backorders.}
\end{aligned}$$

With these order-based performance measures, one can easily obtain the following system performance measures:

$$\begin{aligned}
f &= \text{average (over all demand types) off-shelf fill rate} = \sum_K q^K f^K. \\
\mathsf{E}[B] &= \text{total average order-based backorders} = \sum_K \mathsf{E}[B^K].
\end{aligned}$$

It is also interesting to relate the order-based performance measures to the component-based ones:

$$\begin{aligned} f_i &= \text{off-shelf fill rate of component } i, \\ \mathsf{E}[B_i] &= \text{average number of backorders of component } i. \end{aligned}$$

4.3 Constant Leadtimes

Let L_i be the leadtime for component i, a constant. Then,

$$IN_i = s_i - D_i.$$

Performance evaluation thus involves the joint distribution of the leadtime demands $(D_1, ..., D_m)$. For example,

$$f^K = \mathsf{P}(D_i + Z_i^K \leq s_i,\, i \in K).$$

Let $f^{K,w}(\mathbf{s}|\mathbf{L})$ be $f^{K,w}$ in a system with base-stock levels $\mathbf{s} = (s_i)_i$ and leadtimes $\mathbf{L} = (L_i)_i$. When $L_i = L$ for all i, we write $f^K(\mathbf{s}|\mathbf{L})$ as $f^K(\mathbf{s}|L)$. It is shown in Song [42] that, for any fixed K and $0 \leq w < \max_{i \in K}\{L_i\}$,

$$f^{K,w}(\mathbf{s}|\mathbf{L}) = f^{K,0}(\mathbf{s}|(L_1 - w)^+, ..., (L_n - w)^+). \tag{5}$$

Thus, $f^{K,w}$ equals the immediate fill rate f^K in a transformed system, where the leadtimes are truncated by w. So, we only need to focus on f^K henceforth. (This result is similar to the one for discrete-time systems discussed above.)

Song (1998) observes that there are several independent random variables shared by the elements of the leadtime-demand vector $(D_i)_i$, each of which is a univariate Poisson random variable. Thus, the dimension of the distribution of the vector can be reduced by conditioning on these common elements. As a result, the order fill rates can be obtained through convolutions of one-dimensional distributions. The paper develops an algorithm that sequences the conditioning steps. This procedure makes the calculation much simpler and faster than the direct approach using the joint distribution of the net inventories.

To illustrate the result, consider a 2-component, unit-demand system. Here, there are three types of demand: A type-1 customer requires one unit of component 1 only; type-2 requires one unit of component 2 only; and type-12 (short notation for type-$\{1, 2\}$) asks for one unit of each component. In this unit-demand case, D_i, which has the same distribution as $D_i(L_i)$, has a Poisson

distribution with parameter $\lambda_i L_i$. For convenience, let $p(\cdot|a)$, $P(\cdot|a)$, and $P^c(\cdot|a)$ denote the probability mass function, cdf and complementary cdf, respectively, of the Poisson distribution with parameter a.

The type-i fill rate is exactly component i's fill rate

$$f_i = \mathsf{P}(IN_i > 0) = \mathsf{P}(D_i < s_i) = P(s_i - 1|\lambda_i L_i). \quad i = 1, 2.$$

The type-12 fill rate is

$$f^{12} = \mathsf{P}(IN_1 > 0, IN_2 > 0) = \mathsf{P}(D_1 < s_1, D_2 < s_2).$$

Assume $L_1 = L_2 = L$. Then,

$$D_i = D_i(L) = D^i(L) + D^{12}(L), \qquad i = 1, 2.$$

Here, $D^K(L)$ has the Poisson distribution with parameter $\lambda^K L$. Moreover, the $D^K(L)$ are independent. By conditioning on $D^{12}(L)$ and then deconditioning, we obtain

$$f^{12}(\mathbf{s}|L) = \sum_{k=0}^{\min\{s_1,s_2\}-1} p(k|\lambda^{12}L)P(s_1 - k - 1|\lambda^1 L)P(s_2 - k - 1|\lambda^2 L).$$

In the case $L_1 \neq L_2$, a few extra calculations yield a similar result.

Now, let us return to the general problem. Song (1998) also develops simpler bounds on the order fill rate, which require lower-dimensional joint distributions or merely the marginal distributions. More specifically, it is shown that, for any K,

$$f^K \geq \prod_{\ell=1}^{k} f^{\mathcal{S}_\ell}, \tag{6}$$

where $\{\mathcal{S}_1, ..., \mathcal{S}_k\}$ is any partition of K. In particular,

$$f^K \geq \prod_{i \in K} f_i. \tag{7}$$

Also,

$$f^K \leq \min_{i \in K} f_i.$$

It turns out that f^{ij}, the fill rate for type-$\{i, j\}$ demand, is quite easy to obtain for any i and j; it has the same formula as f^{12} given above. So, (6) yields a better lower bound on f^K than $\prod_{i \in K} f_i$ without too much computational effort. Finding the best *pair partition* so that this lower bound

is maximized is equivalent to the *nonbipartite weighted matching problem*, which can be solved by existing algorithms in the combinatorial-optimization literature.

Song (2002) studies the evaluation of order-based backorders for the same model. Let $B_i^K(t)$ be the number of backorders for item i at time t that are due to demand type K, where $K \in \mathcal{K}_i$. Let B_i^K be the steady-state limit of $B_i^K(t)$. Then

$$\mathsf{E}[B^K] = \mathsf{E}[\max_{i \in K} B_i^K]. \tag{8}$$

Given $B_i = n$, B_i^K is a binomial random variable with n trials and success probability λ^K/λ_i in each trial. However, since the B_i are correlated random variables, computing its joint distribution alone is difficult, not to mention the conditional binomial distributions and the max operation within the expectation. Song presents a much simpler approach. To illustrate the results, consider the 2-item, unit-demand system with equal leadtimes discussed above. Let $\mathsf{E}[B^K(\mathbf{s}|L)]$ denote the expected type-K backorders with base-stock levels s_i and common leadtime L. First, notice that a request for item i is due to a type-i order with probability λ^i/λ_i, so the average type-i backorders equals

$$\bar{B}^i(s_i|L) = \frac{\lambda^i}{\lambda_i}\mathsf{E}[B_i(s_i|L)],$$

where

$$\mathsf{E}[B_i(s_i|L)] = \lambda_i L - \sum_{k=0}^{s_i-1} P^c(k|\lambda_i L).$$

Also,

$$\mathsf{E}[B^{12}(\mathbf{s}|L)] = \lambda^{12}L - q^{12}\sum_{\ell=0}^{\min\{s_1,s_2\}-1}\sum_{m=0}^{s_1-\ell-1}\sum_{j=0}^{s_2-\ell-1}\frac{(\ell+m+j)!}{\ell!\,m!\,j!}(q^{12})^\ell(q^1)^m(q^2)^j P^c(\ell+m+j|\lambda L).$$

Using the formulas, the paper discusses a few examples to gain managerial insights. It is shown, for example, that for a given level of inventory investment, using common components or fewer product configurations may *not* reduce backorders. This is in contrast to conclusions drawn from more restrictive single-period models in the literature.

Although the exact result enjoys a tremendous computational advantage over simulation, it can still be computationally demanding for large systems. The paper also develops easy-to-compute bounds. In particular,

$$LB^K \;:=\; \lambda^K \max_{i \in K}\frac{\mathsf{E}[B_i]}{\lambda_i} \quad\leq\quad \mathsf{E}[B^K] \quad\leq\quad \lambda^K \sum_{i \in K}\frac{\mathsf{E}[B_i]}{\lambda_i} \;:=\; UB^K.$$

Summing these inequalities yields bounds on the total average order-based backorders:

$$LB \;:=\; \sum_K LB^K \;\le\; \mathsf{E}[B] \;\le\; \sum_K UB^K \;:=\; UB.$$

It can be verified that

$$UB = \sum_{i=1}^{J} \mathsf{E}[B_i] = \text{ the total average item-based backorders} := \mathsf{E}[B_I].$$

So, the total item-based backorders always dominates the total order-based backorders.

A natural approximation for $\mathsf{E}[B^K]$ is the simple average of LB^K and UB^K. Numerical results indicate that the approximation performs extremely well.

Song (2000) extends the above results to more general policies, the (R, nQ) policies. (Here, for each component i, there is a base lot-size Q_i, a positive integer. When the inventory position of item i falls to or below the reorder point R_i, an order of size nQ_i is placed, where n is the integer such that the inventory position after ordering is between $R_i + 1$ and $R_i + Q$. When all $Q_i = 1$, the policy reduces to a base-stock policy.) Under reasonable conditions, the service measures can be computed as simple averages of their counterparts under base-stock policies.

4.4 Uncapacitated Stochastic Leadtimes

Next, consider the case where the supply system of each component consists of many parallel processors, so that its leadtimes are i.i.d. random variables. The constant-leadtime model is a special case. Let

$$X_i(t) = \text{number of outstanding orders of component } i \text{ at time } t.$$

Then (returning to base-stock policies),

$$IN_i(t) = s_i - X_i(t).$$

Thus, performance evaluation involves the distribution of the outstanding-order vector $X(t) = (X_1(t), ..., X_m(t))$ and its steady-state limit $X = (X_1, ..., X_m)$.

I.i.d. leadtimes were assumed in the earliest studies of dynamic product-based performance, in the literature on multi-indenture models of multi-echelon inventory systems. Here, an end item (product) consists of several repairable modules (components). A failure of a product is due to the failure of one component. The performance measure of interest is the number of products backordered. Cannibalization is allowed, that is, a good component in a failed product can be used

to replace a failed component in another failed product. Thus, the number of products backordered is the maximum of the component backorders, i.e.,

$$B = \max_i B_i = \max_i [(X_i - s_i)^+]. \tag{9}$$

Suppose the product fails according to a Poisson process. Then, one can calculate the cumulative distribution of B. For other arrival processes, one can obtain the expected value of B. See Nahmias (1981) for a review.

Cheung and Hausman (1995) consider multivariate Poisson demand, so there can be simultaneous failures of several components. Again, they assume complete cannibalization, so (9) holds. The authors propose the following disaggregation approach for the joint distribution of X_i: Define Y^K to be the number of jobs of type-K that have one or more components outstanding. Then, according to Palm's result for $M/G/\infty$ queues, Y^K has a Poisson distribution with mean $\lambda^K \mathsf{E}[\max_{i \in K} L_i]$, and the Y^K are independent over K. Let X_i^K be the number of outstanding orders of component i that originated from demand type-K, so $X_i = \sum_{i \in K} X_i^K$. Then, $\mathsf{E}[B]$ can be evaluated by conditioning on $(\mathbf{Y} = \mathbf{y}) = (Y^K = y^K, \forall K)$. The conditional probability $\mathsf{P}[X_1 \leq x_1, ..., X_m \leq x_m | \mathbf{Y} = \mathbf{y}]$ must be obtained through computation. This probability is easy when all $y^K = 0$ or 1. For larger y^K, it becomes complicated.

Two approximations are proposed to simplify the computation of $\mathsf{E}[B]$ for large m. However, the approximations still employ the conditional probability mentioned above. Also, Jensen's inequality yields a simple lower bound:

$$\mathsf{E}[B] \geq \max_i \{(\lambda_i \mathsf{E}[L_i] - s_i)^+\}.$$

Unfortunately, as the paper shows, this bound works poorly as an approximation.

Gallien and Wein (2001) assume i.i.d. component leadtimes in a single-product assembly system. Demand is Poisson with rate λ. Inspired by Rosling's result for deterministic leadtimes, the following class of policies is considered: Start with stocks of all components at a common base-stock level, s. Every demand triggers an order for each component after a component-dependent delay $\delta_i \geq 0$. Numerical experiments show that this policy achieves nearly identical performance to the standard base-stock policy with base-stock levels $s_i = s - \lambda \delta_i$.

To keep the analysis tractable, the paper imposes a synchronization assumption, namely, that components are assembled in the same sequence they are ordered in. Thus, the time needed to replenish a complete set of components is $\max_i(L_i + \delta_i)$. Let Y be the steady-state number of replenishment orders for complete sets of components, for which at least one of the m individual

component orders has not yet arrived. Then, applying Palm's result, Y has a Poisson distribution with mean $\rho = \lambda \mathsf{E}[\max_i(L_i + \delta_i)]$. Also, $I = (s - Y)^+$ and $B = (Y - s)^+$. Using a tandem queueing network analogous, it is shown that

$$\mathsf{E}[I_i] = \lambda(\mathsf{E}[\max_i(L_i + \delta_i) - \mathsf{E}[L_i] - \delta_i).$$

Now, assume L_i has the Gumbel distribution with cdf $exp(-\alpha_i e^{-mx})$, $m > 0$. This implies that all the L_i have the same variance σ^2. Then, $\mathsf{E}[\max_i(L_i + \delta_i)]$ can be written in closed form for any δ_i. This permits the closed-form determination of optimal values of δ_i and s that minimizes the long-run average cost

$$\sum_{i=1}^{m} h_i \mathsf{E}[I_i] + (\sum_{i=1}^{m} h_i)\mathsf{E}[I] + b[B],$$

where h_i and b are unit holding-cost rate and backorder cost rate, respectively. That is,

$$\delta_i^* = \max_j \left(\mathsf{E}[L_j] - \frac{\sqrt{6}}{\pi}\sigma \ln h_j \right) - \left(\mathsf{E}[L_i] - \frac{\sqrt{6}}{\pi}\sigma \ln h_i \right), \quad i = 1, ..., m$$

$$\rho^* = \frac{\lambda\sqrt{6}\sigma}{\pi} \ln \left[\sum_{i=1}^{m} \exp(\frac{\pi(\mathsf{E}[L_i] + \delta_i^*)}{\sqrt{6}\sigma} \right]$$

s^* is the smallest integer that satisfies $P(s^*|\rho^*) \geq \dfrac{b}{b+h}$.

Numerical results for an 11-component system indicate that this approximate solution is within 2% of the best among this class of policies (found by simulation), in all cases with a common leadtime variance. It also significantly outperforms policies that ignore either component dependence or leadtime variance. In addition, it is reasonably robust with respect to various modeling assumptions. However, the order-synchronization approximation does affect the results; it overestimates the amount of inventory required.

Song and Yao (2000) also study the single-product system with Poisson demand and i.i.d. leadtimes. They adopt the standard base-stock policy, without order synchronization. Under any base-stock policy, the outstanding-order vector $X(t)$ is precisely the numbers of jobs in m M/G/∞ queues with a common arrival stream. This $X(t)$ has a steady-state limit, X. Let $G_i(\cdot)$ be the cdf of L_i and $G_i^c = 1 - G_i$. Let $N(a)$ denote a Poisson random variable with mean a. Then, X can be expressed as partial sums of 2^{m-1} independent Poisson random variables as follows:

$$X_i = \sum_{\mathcal{S}: i \in \mathcal{S}} N(\lambda\theta_{\mathcal{S}}), \quad \text{with} \quad \theta_{\mathcal{S}} = \int_0^\infty [\prod_{k \in \mathcal{S}} G_k^c(x)][\prod_{j \in \mathcal{I} \setminus \mathcal{S}} G_j(x)]dx. \tag{10}$$

Here, for any subset $\mathcal{S}$ of $\mathcal{I}$, $N(\lambda\theta_{\mathcal{S}})$ is the number of jobs (in steady state) still in process in the queues $k \in \mathcal{S}$, but completed by the other queues.

In principle, all the performance measures can be evaluated exactly using (10). However, there are $2^m - 1$ independent Poisson random variables. This exponential growth in the number of components means that the method is impractical for large systems.

The paper investigates the effect of leadtime variability by comparing two systems, the original system with leadtimes L_i, and another system with leadtimes $\tilde{L}_i$. Assume $\mathsf{E}[L_i] = \mathsf{E}[\tilde{L}_i] = \ell_i$, and that L_i is more variable than $\tilde{L}_i$ in the sense of the "increasing convex ordering", denoted $L_i \geq_{icx} \tilde{L}_i$, i.e.,

$$\int_x^\infty G_i^c(u)du \geq \int_x^\infty \tilde{G}_i^c(u)du$$

for $x \geq 0$. (Here, $\tilde{G}_i^c$ is the complementary cdf of $\tilde{L}_i$.) Note that the above implies $\mathsf{Var}[L_i] \geq \mathsf{Var}[\tilde{L}_i]$. Let $\tilde{f}$ and $\tilde{B}$ denote the fill rate and the number of backorders in the new system. Then,

$$\begin{aligned} f &\leq \tilde{f}, \\ B &\geq_{\text{st}} \tilde{B}. \end{aligned}$$

Thus, in contrast to the standard single M/G/∞ queueing system, leadtime variability degrades performance here.

Since evaluating $\mathsf{E}[B]$ is hard, the paper develops simple upper and lower bounds on $\mathsf{E}[B]$ and uses them as surrogate objectives in the following optimization problem:

$$\begin{aligned} &\min \mathsf{E}[B(s_1, ..., s_m)] \\ &s.t. \quad c_1 s_1 + \cdot + c_m s_m \leq C. \end{aligned} \tag{11}$$

Greedy algorithms are developed, and numerical results indicate that these solution techniques are fairly effective.

The paper considers another optimization problem that minimizes the average component inventory costs subject to a required fill rate. Approximating the constraint by using (7) yields a separable convex programming problem, which can be solved via a greedy algorithm. Numerical results show that this lower bound approach usually results in an order fill rate (in the original system) that is substantially higher than the required service level. However, the greedy algorithm has considerable advantage in computation time. Therefore, it can be used to quickly generate an initial solution, followed by a neighborhood search to find the best solution.

The extension of this analysis to multiple products turns out to be far from routine. Lu et al.

(2003a) derive the joint generating function of X:

$$\begin{aligned}\psi(z_1,...,z_m) &:= \mathsf{E}[\prod_{j=1}^{m} z_j^{X_j}],\\ &= \exp\left[\sum_{K\in\mathcal{K}} \lambda^K \int_0^\infty (\psi^{Z^K}(G_1(u)+z_1G_1^c(u),...,G_m(u)+z_mG_m^c(u))-1)du\right].\end{aligned}$$

In the special case of unit demands (all $Z_i \equiv 1$), the generating function takes the following form:

$$\psi(z_1,...,z_m) = \exp[\sum_{K\in\mathcal{K}} \lambda^K \int_0^\infty (\prod_{j\in K}[G_j(u)+z_jG_j^c(u)]-1)du],$$

which corresponds to a multivariate Poisson distribution. Thus, we obtain a generalization of (10):

$$X_i = \sum_{K\in\mathcal{K}_i} \sum_{\mathcal{S}\ni i,\mathcal{S}\subseteq K} N(\lambda^K\theta_{\mathcal{S}}^K)$$

where all the Poisson variables are independent, and for any subset $\mathcal{S}$ of K,

$$\theta_{\mathcal{S}}^K = \int_0^\infty \prod_{j\in\mathcal{S}} G_j^c(u) \prod_{j\in K\setminus\mathcal{S}} G_j(u)du.$$

The effort required to evaluate the performance measures is linear in the number of products. Unfortunately, it is again exponential in the number of components. On the other hand, the generating function can be used to obtain simple expressions for the means, variances and covariances. Let $\mathcal{K}_{ij}$ denote the family of subsets that contain both i and j. Then,

$$\begin{aligned}\mu_j &:= \mathsf{E}[X_j] = \mathsf{E}[L_j]\sum_{K\in\mathcal{K}_j}\lambda^K\mathsf{E}(Z_j^K),\\ \sigma_j^2 &:= \mathsf{Var}[X_j] = \mathsf{E}(X_j) + \sum_{K\in\mathcal{K}_j}\lambda^K[\mathsf{E}((Z_j^K)^2)-\mathsf{E}(Z_j^K)]\int_0^\infty [G_j^c(u)]^2du,\\ \sigma_{ij} &:= \mathsf{Cov}[X_i,X_j] = \sum_{K\in\mathcal{K}_{ij}}\lambda^K\mathsf{E}(Z_i^KZ_j^K)\int_0^\infty G_i^c(u)G_j^c(u)du \quad i\neq j,\end{aligned}$$

This suggests using the multivariate normal distribution with these moments to approximate X. This is still a difficult calculation, and the paper develops several further approximations.

One approximation is based on an upper bound on the covariance:

$$\sigma_{ij} \leq \eta_i\eta_j,$$

where

$$\eta_i := \left[\sum_{K\in\mathcal{K}_i}\lambda^K\mathsf{E}((Z_i^K)^2)\right]^{1/2}\left[\int_0^\infty (G_i^c(u))^2du\right]^{1/2}.$$

It turns out that the normal calculations become easy with each σ_{ij} replaced by $\eta_i \eta_j$. This is called the *factorized* normal approximation.

A second approximation applies (6) to a *pairwise* partition of $\mathcal{I}$. The calculation then reduces to the evaluation of bivariate normal distributions.

A numerical study compares the factorized normal approximation, the pairwise approximation and the marginal lower bound (7). The pairwise approximation is the most accurate, but the other two methods require less computational effort, and their accuracy is reasonably good.

Lu et al. (2003b) focus on the following optimization problem for the same multiproduct model:

$$\begin{aligned} &\min_{\mathbf{s}} \sum_K w^K \mathsf{E}[B^K(\mathbf{s})] \\ &s.t. \quad c_1 s_1 + \cdot + c_m s_m \leq C. \end{aligned} \tag{12}$$

where $\mathbf{s} = (s_1, ..., s_m)$, and $w^K \geq 0$ is a weighting factor for the average type-K backorders. To find the optimal solution, the authors study two surrogate problems, based on upper- and lower-bound approaches to approximate the objective function. Both surrogate problems are of the same structure. Let n be the number of demand types (or products). Then the optimal base-stock levels can be determined by solving $n!$ minimizations problems of the following form:

$$\begin{aligned} &\min_{\mathbf{y}} \sum_{\ell=1}^{n} v_\ell y_\ell \\ &s.t. \quad \sum_{\ell=1}^{n} \tau_\ell (y_1 + \cdots + y_\ell) \leq C, \quad \mathbf{y} \geq \mathbf{0}. \end{aligned} \tag{13}$$

Each of these problems can be solved by greedy methods. Heuristic algorithms are also developed to speed up the computation. Numerical results indicate that these solution techniques are quite effective.

Lu and Song (2002) formulate an unconstrained cost-minimization problem for the multi-product, unit-demand system. Let b^K be the backorder cost rate for each backlogged customer order of type-K, and let J_i be the steady-state number of units of item i that have been put aside and committed to demands which are backlogged due to the unavailability of other items. The expected total average cost under any base-stock policy $\mathbf{s} = (s_1, ..., s_m)$ is:

$$\begin{aligned} C(\mathbf{s}) &= \sum_i h_i \mathsf{E}[I_i(s_i) + J_i(\mathbf{s})] + \sum_K b^K \mathsf{E}[B^K(\mathbf{s})] \\ &= \sum h_i s_i + \sum_K \tilde{b}^K \mathsf{E}[B^K(\mathbf{s})] - \sum_i h_i \mathsf{E}[X_i], \end{aligned}$$

where

$$\tilde{b}^K \quad = \quad b^K + \sum_{i \in K} h_i.$$

The paper compares the above formulation with item-based formulations – that is, treat the system as a set of independent single-item systems. Let b_i be the unit backorder cost for item i backorders. The item-based optimization problem is

$$\min_{\mathbf{s}} \sum_{i=1}^{m} (h_i \mathsf{E}[I_i(s_i)] + b_i \mathsf{E}[B_i(s_i)]) \quad = \quad \sum_{i=1}^{m} (h_i s_i + (h_i + b_i)\mathsf{E}[B_i(s_i)]) - \sum_i h_i \mathsf{E}[X_i].$$

This problem is separable across i. The problem for each i is a newsvendor-type problem and its solution can be obtained easily.

Lu and Song show that, if we set

$$b_i = \sum_{K \in \mathcal{K}_i} \frac{\lambda^K}{\lambda_i} (b^K + \sum_{j \in K, j \neq i} h_j),$$

the result of this item-based calculation is an upper bound on $\mathbf{s}^*$. A different choice of b_i yields a lower bound for the upper bound. Moreover, using the upper bound as a starting point, $\mathbf{s}^*$ can be obtained in a greedy fashion by employing recently developed optimization techniques for discretely convex functions.

4.5 Capacitated Stochastic Leadtimes

Song *et al.* (1999) consider a multi-product, unit-demand model. The supply system of each component i is modeled as single exponential processor with rate μ_i and a finite backlog buffer of capacity $b_i \geq 0$. The finite buffer works as follows: A demand for component i that cannot be filled immediately goes to the backlog queue i, provided the queue is not full. The demand will be shipped out (or put aside) as soon as a unit of item i becomes available. When a demand arrives and finds any of its items' backlog queues full, it signals the customer that a long wait is likely, and the customer decides to leave. Thus, the buffer sizes can be viewed as measures of customer impatience. (When $b_i = \infty$ for all i, unfilled demands are backlogged. When $b_i = 0$ for all i, unfilled demands are lost.)

Two blocking mechanisms are considered when an incoming demand finds the backlog queue for at least one of its components full:

- *Total order service (TOS):* If a type K order sees at least one of its component's backlog queue is full, then the order is lost entirely. In other words, a type K order must be accepted

as a whole. This model is valid for the assemble-to-order environment and also for some make-to-stock systems.

- *Partial order service (POS):* When a type K order arrives and the backlog queue i is full for $i \in K' \subset K$, then the order for items in K' is lost, whereas the order for items in $K - K'$ is satisfied, either immediately or in the future. This model fits many distribution systems, where customers often accept partial shipments of finished goods.

Thus, in the POS model, customer impatience is associated with individual items, while in the TOS model, it is associated with the whole order.

The paper shows that the outstanding-order vector $X(t)$ is an irreducible continuous-time Markov chain with finite state space. Its unique stationary distribution can be obtained through the *matrix-geometric* solution of a *quasi birth-and-death* (QBD) process (Neuts (1981), Chapter 3). The paper derives the exact expressions for $f^{K,w}$ and $E[B^K]$ in terms of this solution.

Numerical experiments indicate that the results from the POS model provide reliable estimates of their counterparts in the TOS model. So, it is sufficient to focus on one order service scheme. Also, with moderate traffic the finite-buffer model provides an accurate approximation for the infinite-buffer model.

Iravani et al. (2000) employ a similar modeling framework and technique to study a system with flexible customers. Each K is partitioned into two subsets K^1 and K^2. K^1 contains the "key" components of a type-K demand. If any components in K^1 are not available, a type-K demand is lost. On the other hand, if some "non-key" components – those in K^2 – are not available, then a type-K customer may accept substitutions or even ignore them. Specifically, for $i \in K$, there is a probability p_{ij}^K that customers of type-K will accept component j if component i is not available. If $p_{ij}^K = 0$ for all $j \in \mathcal{I} \setminus \{i\}$, then customer type-$K$ accepts no substitutes for component i.

Glasserman and Wang (1998) model the supply system as a set of M/G/1 queues (G/G/1 queues for the single-product case). Assuming the fill rate $f^{K,w}$ remains high, the paper investigates the tradeoff between the delivery time window w and the total base-stock units $s = s_1 + s_2 + \cdots + s_m$. It is intuitively clear that, fixing the fill rate, s increases as w decreases. The key result of the paper is that this relationship is asymptotically linear, provided the ratios $k_i = s_i/s$ are kept constant as s increases. The two parameters of this linear relationship can be determined exactly (through analysis of the arrival and service times' cumulant generating functions) or approximately (from their moments).

Let U_i denote the random processing time of a unit of component i, and V_i be the interarrival time of demand for component i. It is shown that for large s or w, the item fill rate can be approximated by

$$1 - f_i^w \quad \approx \quad C_i e^{-\gamma_i w - \beta_i s_i}.$$

for some constants C_i, γ_i and β_i. For Poisson demand process (which is necessary for the multi-product case),

$$C_i = \lambda_i^{-1}(\lambda_i + \gamma_i)(1 - \lambda_i \mathsf{E}[Z_i]\mathsf{E}[U_i])(\psi'_{Z_i}(\beta_i)\psi_{U_i}(\gamma_i)(\lambda_i + \gamma_i) - 1)^{-1}.$$

Recall that Z_i is the demand batch size for component i. The constants γ_i and β_i can be obtained by solving some equations involving the cumulant generating functions of the input random variables. (The cumulant generating function of a random variable Y is defined by $\psi_Y(\theta) = \log \mathsf{E}[e^{\theta Y}]$.) They can also be approximated by the first two moments of these random variables as follows:

$$\gamma_i \approx -\frac{2\mathsf{E}[U_i]\mathsf{E}[Z_i] - \mathsf{E}[V_i]}{\mathsf{E}[Z_i]\mathsf{Var}[U_i] + \mathsf{Var}[Z_i](\mathsf{E}[U_i])^2}$$

and

$$\beta_i \approx \mathsf{E}[U_i]\gamma_i + \frac{1}{2}\mathsf{Var}[U_i]\gamma_i^2.$$

Let

$$\gamma^K = \min_{i \in K}\{\gamma_i\} \quad \text{and} \quad \mathcal{W}^K = \{i \in K : \ \gamma_i = \gamma^K\}.$$

Define $\alpha_i = k_i \beta_i$ and let

$$\alpha^K = \min_{i \in K}\{\alpha_i\} \quad \text{and} \quad \mathcal{S}^K = \{i \in K : \ \alpha_i = \alpha^K\}.$$

Then, when the time window w is long, the order fill rate can be approximated by

$$[1 - f^{K,w}] \approx \sum_{i \in \mathcal{W}^K} C_i e^{-\gamma^K w - \alpha_i s}.$$

When the total base-stock units s is high,

$$[1 - f^{K,w}] \approx \sum_{i \in \mathcal{S}^K} C_i e^{-\gamma_i w - \alpha^K s}.$$

Assuming that all products' fill rates are the same and high, and the items' base-stock levels change in constant proportions, this approximation suggests that, when the time window w is changed, the base-stock levels should be varied according to the component-level tradeoff rule

$$\Delta s_i = -\frac{\gamma_i}{\beta_i}\Delta w$$

to maintain the same order fill rate. Numerical experiments show that under certain conditions this linear rule provides satisfactory results.

This result, however, depends strongly on the assumption of finite capacity. (It is easy to show that, in a simple single-item system with constant leadtimes, the relationship is nonlinear.)

Wang (1999) applies this result to an optimization problem to minimize average inventory cost subject to a fill-rate constraint. The paper focuses on a single-product system and solves a surrogate problem with closed-form solution. In particular, it shows that there exists an index k, such that the following solution is effective:

$$\tilde{s}_i \quad = \quad \frac{1}{\beta_i} \log \frac{\theta_k C_i e^{-\gamma_i w}}{h_i/\beta_i}, \qquad \text{for } i \leq k \qquad \text{and} \qquad \tilde{s}_i = 0 \qquad \text{for} \qquad i > k,$$

where

$$\theta_j = \frac{\sum_{i=1}^{j} (h_i/\beta_i)}{\delta - \sum_{i=j+1}^{m} C_i e^{-\gamma_1 x}}$$

and h_i is the unit holding cost of component i.

Dayanik et al. (2001) examine several ideas scattered in diverse literature on approximations for multivariate probability distributions, and determine which approach is most effective in estimating performance in capacitated ATO systems. Tailoring different approximation ideas to the ATO setting, they derive several performance bounds, such as setwise bounds based on the dependence structure of the system, distribution-free Bonferroni-type bounds commonly used to bound multivariate distributions, Frechet-type bounds, and bounds that are combinations of these previous ones. The paper compares these bounds both analytically and numerically. The general conclusion is that the setwise bounds are most effective.

Xu (2001) summarizes several performance bounds for ATO systems based on stochastic comparison techniques.

To summarize, research on continuous-time models has made major strides in the last few years in developing robust analytical tools for design and control of ATO systems. Both exact and asymptotic results as well as bounds and approximations have been developed. The methods, however, heavily depend on the detailed model assumptions. So, in applying these methods, one should be careful about which model framework fits best in the particular application.

5 Research on System Design

Another line of research aims to understand the broad issues involved in product and process design. An overview is given by Nevins and Whitney (1989) and a review of the research literature by Krishnan and Ulrich (2001). This work, of course, considers a range of production modes, not just ATO. Here, we focus on research that explicitly treats the product- and component-variety issues posed by ATO systems.

This research tends to suppress most of the detail of the operational models discussed above. Instead, it aims to approximate the operational cost of a system by means of simple functions.

Fisher et al. (1999) develop a model for system design in one specific industry, automobile brakes, and test it empirically. This model represents the total operational costs (and design costs as well) by affine functions of demand. Brakes differ from each other on one critical dimension, their rotor diameters. Depending mainly on its weight, a car requires rotors of at least a certain diameter. The model aims to determine the optimal number of brakes for a given family of cars. Under certain simplifying assumptions, the optimal number of brakes is proportional to a simple index, the square root of (total demand times the range of car weights). The paper then tests the model using data from six companies, half American and half Japanese. The index accurately predicts the actual variety in brakes.

Ramdas and Sawhney (2001) develop a model to redesign a product line. First, they develop a method to estimate the impact on revenues due to product-line extensions. Second, they outline a method to estimate the operational-cost impact. This method includes terms reflecting the scale economies due to component commonality in the new products. Third and finally, they combine these methods into an integer-programming model to select the optimal line extensions. The paper reports a case study based on data from a wristwatch manufacturer.

Krishnan et al. (1999) develop another product-line design model. Although it focuses on product-development costs, it includes functions that represent part-commonality effects in operational costs. The end result, again, is an optimization model. Krishnan and Gupta (2001) use a similar model to evaluate "product platforms", sets of components and subassemblies shared across whole families of products. They identify conditions under which such platforms may, and may not, be beneficial.

The issue of component commonality is related to the broader issue of modular design. Baldwin and Clark (2000) provide an overview of this concept. Thonemann and Brandeau (2000) develop

a detailed model to optimize the level of part commonality.

These few works are, in our view, best seen as initial forays into largely uncharted territory. The design process involves many factors in addition to operational costs. Although the phrase "design for manufacturing" represents a recognition of the importance of such costs, we do not yet understand how best to organize design resources to take them into account, along with other critical factors. Future research, we hope, will shed more light on this matter.

6 Summary and Future Directions

As we have seen above, recent research has made considerable progress in developing analytical methods for ATO systems. We now have tractable methods to estimate and improve performance, at least for some systems. Those methods have led to some interesting and useful managerial insights. Much work remains, nevertheless. Here we point out a few areas where further research is needed.

6.1 Optimal Policies

As indicated above, little is known about the forms of optimal policies for multi-period models. The research to date mostly assumes particular policy types. It would be valuable to learn more about truly optimal policies. Even partial characterizations would be interesting. Also, better heuristic policy forms would be useful.

6.2 Tractable Methods for Large-Scale Systems

Many real ATO systems contain hundreds of components and thousands of products. A division of Hewlett-Packard, for example, uses over 100 PC components grouped into eight component families to make their computers. Such a system poses a considerable computational burden on existing models and solution methods. Even data estimation is no trivial task.

A number of approaches might improve matters. One approach is to seek model formulations with special structures that allow efficient evaluation of the performance measures. Another approach is to develop decomposition and approximation schemes allowing algorithm "scalability" to large data sets. Sections 3 and 4 reviewed several ideas in the recent developments. Still, better methods of this sort would be most welcome.

6.3 Demand Distributions

Nearly all the models in this chapter assume stationary data. However, short product life cycles imply time-varying or state-dependent demand. It is desirable that practical models in the future allow for such complex demand models.

6.4 Shifts in Supply Chain Structures and Costs

The pressure to streamline supply chain flows and to increase supply chain efficiency and reduce cost has led many manufacturers to outsource some (even all) steps of assembly operations (mostly product configuration and customization), usually to their distributors, who might in turn delegate part of the final assembly to the retailers. Hence the ATO problem might be encountered by multiple players in the same supply chain. (Chapters VII and VIII of this volume discuss the issues involved in multi-player supply chains.)

In addition, new issues and management practices continue to emerge. Manufacturing capacity in assembly, once owned and concentrated at the manufacturer, is now shifted downstream in the supply chain and becomes distributed among the players and more flexible and less expensive to expand at the same time. Therefore, the scheduling of short-term flexible capacities to meet temporary product demand fluctuations or product mix changes, and the coordination of capacities at different stages of a system, will continue to be important research problems.

The manner in which supply-chain players share the financial risks and costs might also be changing. A particular example is the "price protection" contract between the manufacturer and its distributor. Designed to protect the distributor from rapid price declines and shift the price decline risk to the manufacturer, the price protection policy changes the traditional definition of inventory holding cost so that much of the inventory hold cost might be shifted to either the supplier or the customer. Furthermore, the cost relationship between the supplier and the buyer may be made more complex by vendor-managed inventory (VMI) programs. Since the inventory holding cost is a critical parameter in ATO models, the change in the cost structure might affect the solution and the recommendation significantly. (Chapters VII and VIII of this volume discuss these and other issues involved in multi-player supply chains.)

6.5 Product Design Implications

Model-based research on product design, as suggested in Section 5, is at an early stage. Unfortunately, we do not yet understand detailed operational models well enough to derive from them

simple, empirically testable and usable cost models. For the time being, therefore, empirical models must rely on ad hoc cost functions with little basis in theory. We see many opportunities for future research to help bridge this gap.

Acknowledgement. We would like to thank Alex Zhang for helpful discussions on this topic.

References

[1] Agrawal, m. and Cohen, M., Optimal Material Control and Performance Evaluation in an Assembly Environment with Component Commonality. *Naval Research Logistics* **48** (2001), 409-429.

[2] Bagchi, U. and Gutierrez, G., Effect of Increasing Component Commonality on Service Level and Holding Cost, *Naval Research Logistics*, **39** (1992), 815-832.

[3] Baker, K., Magazine, M. and Nuttle, H., The Effect of Commonality of Safety Stock in a Simple Inventory Model, *Mgmt. Sci.*, **32** (1986), 982-988.

[4] Baldwin, C. and Clark, K., *The Power of Modularity*, MIT Press, Cambridge, MA, 2000.

[5] Brumelle, S. and Granot D., The Repair Kit Problem Revisited, *Operations Research* **41** (1993), 994-1006.

[6] Chen, F. and Zheng,Y.-S. Lower Bounds for Multi-echelon Stochastic Inventory Systems, *Management Science*, **40** (1994), 1426-1443.

[7] Cheng, F., Ettl, M., Lin, G.Y., and Yao, D.D., Inventory-Service Optimization in Configure-to-Order Systems, *Manufacturing & Service Operations Management* **4** (2002), 114-132.

[8] Cheung, K.L. and Hausman, W., Multiple Failures in a Multi-Item Spare Inventory Model, *IIE Transactions*, **27** (1995), 171-180.

[9] Dayanik, S., Song, J.-S. and Xu, S.H. The Effectiveness of Several Performance Bounds for Capacitated Assemble-to-Order Systems, working paper (2001), Graduate School of Management, University of California, Irvine, CA 92697.

[10] de Kok, A and Visschers, J., Analysis of Assembly Systems with Service Level Constraints. *International Journal of Production Economics* **59** (1999), 313-326.

[11] Eynan, A., The Impact of Demand's Correlation on the Effectiveness of Component Commonality. *Int. J. Prod. Res.* **34** (1996), 1581-1602.

[12] Eynan, A. and M. Rosenblatt, Component Commonality Effects on Inventory Costs. *IIE Transactions* **28** (1996), 93-104.

[13] Fisher, M., Ramdas, K. and Ulrich, K., Component sharing in the management of product variety: A study of automotive braking systems. *Management Science* **45** (1999), 297-315.

[14] Gallien, J. and Wein, L., A Simple and Effective Component Procurement Policy for Stochastic Assembly Systems, *Queueing Systems* **38** (2001), 221-248.

[15] Gerchak, Y. and Henig, M., An Inventory Model with Component Commonality, *Operations Research Letters* **36** (1986), 61-68.

[16] Gerchak, Y. and Henig, M., Component Commonality in Assemble-to-Order Systems: Models and Properties, *Naval Research Logistics* **36** (1989), 61-68.

[17] Gerchak, Y., Magazine, M. and Gamble, A., Component Commonality with Service Level Requirements, *Mgmt. Sci.* **34** (1988), 753-760.

[18] Glasserman, P. and Wang, Y., Leadtime-Inventory tradeoffs in Assemble-to-Order Systems, *Operations Research* **46** (1998), 858-871.

[19] Graves, S., A Multi-Item Inventory Model with a Job Completion Criterion. *Mgmt. Sci.* **28** (1982), 1334-1336.

[20] Ha, A., Inventory Rationing in a Make-to-Stock Production System with Several Demand Classes and Lost Sales. *Mgmt. Sci.* **43** (1997), 1093-1103.

[21] Hausman, W.H., Lee, H.L., and Zhang, A.X., Joint Demand Fulfillment Probability in a Multi-Item Inventroy System with Independent Order-up-to Policies, *European J. of Operational Research* **109** (1998), 646-659.

[22] Hillier, M. Component Commonality in Multi-Period Assemble-to-Order Systems, *IIE Transactions*, **32** (2000), 755-766.

[23] Hillier, M. Component Commonality in a Multi-Period Inventory Model with Service Level Constraints. *International J. of Production Research* **37** (1999), 2665-2683.

[24] Iravani, S., Luangkesorn, K. and Simchi-Levi, D. On Assemble-to-Order Systems with Flexible Customers. Working paper, Northwestern University, 2000. Forthcoming, *IIE Transactions*.

[25] Kerwin, K., At Ford, E-commerce Is Job 1, *Business Week*, February 28, 2000, 74-78.

[26] Krishnan V. and Gupta, S. Appropriateness and impact of platform-based product development. *Management Science*, **47** (2001), 52-68.

[27] Krishnan V., Singh, R. and Tirupati, D. A model-based approach for planning and developing a set of technology-based products. *Manufacturing & Service Operations Management* **1** (1999), 132-156.

[28] Krishnan, V. and Ulrich, K. Product Development Decisions: A Review of the Literature. *Management Science* **47** (2001), 1-21.

[29] Lu, Y. and Song, J.-S. Order-Based Cost Optimization in Assemble-to-Order Systems. Working paper (2002), IBM Watson Research Center, Yorktown Hights, NY 10598.

[30] Lu, Y., Song, J.-S. and Yao, D.D., Order Fill Rate, Leadtime Variability, and Advance Demand Information in an Assemble-to-Order System. *Operations Research* **51** (2003), March-April.

[31] Lu, Y., Song, J.-S. and Yao, D.D., Backorder Minimization in Multiproduct Assemble-to-Order Systems. Working paper (2003), IBM Watson Research Center, Yorktown Hights, NY 10598.

[32] Mamer, J. and Smith, S., Optimizing Field Repair Kits Based on Job Completion Rate. *Mgmt. Sci.* **28** (1982), 1328-1333.

[33] Mamer, J. and Smith, S., Job Completion Based Inventory Systems: Optimal Policies for Repair Kits and Spare Machines. *Mgmt. Sci.* **31** (1985), 703-718.

[34] Mamer, J. and Smith, S., Inventories for Sequences of Multi-Item Demands. Chapter 12 in J.S. Song and D.D. Yao, eds. *Supply Chain Structures: Coordination, Information and Optimization.* Kluwer Academic Publishers, Norwell, MA, 2001, 415-437.

[35] Neuts, M. *Matrix Geometric Solutions in Stochastic Models.* Johns Hopkins University Press, Baltimore, MD, 1981.

[36] Nevins, J. and Whitney, D. *Concurrent Design of Products and Processes*, McGraw-Hill, NY, 1989.

[37] Ramdas, K. and Sawhney, M. A cross-functional approach to evaluating multiple line extensions for assembled products. *Management Science*, **47** (2001), 22-36.

[38] Rosling, K., Optimal Inventory Policies for Assembly Systems under Random Demands, *Operations Research*, **37** (1989), 565-579.

[39] Schmidt, C. and Nahmias, S., Optimal Policy for a Two-Stage Assembly System Under Random Demand. *Operations Research* **33** (1985), 1130-1145.

[40] Sherbrooke, C., *Optimal Inventory Modeling of Systems*, Wiley, New York, 1992.

[41] Smith, S., Chambers, J. and Shlifer, E., Optimal Inventories Based on Job Completion Rate for Repairs Requiring Multiple Items, *Management Science*, **26** (1980), 849-852.

[42] Song, J.-S., On the Order Fill Rate in a Multi-Item, Base-Stock Inventory System, *Operations Research*, **46** (1998), 831-845.

[43] Song, J.-S., A Note on Assemble-to-Order Systems with Batch Ordering, *Management Science*, **46** (2000), 739-743.

[44] Song, J.-S., Order-Based Backorders and Their Implications in Multi-Item Inventory Systems, *Management Science* **48** (2002), 499-516.

[45] Song, J.-S., Xu, S.H. and Liu, B. Order Fulfillment Performance Measures in an Assembly-to-Order System with Stochastic Leadtimes, *Operations Research* **47** (1999), 131-149.

[46] Song, J.-S., Yano, C. and Lerssrisuriya, P., Contract Assembly: Dealing with Combined Supply Leadtime and Demand Quantity Uncertainty, *Manufacturing & Service Operations Management* **2** (2000), 287-296.

[47] Song, J.-S. and Yao, D.D., eds. *Supply Chain Structures: Coordination, Information and Optimization.* Kluwer Academic Publishers, Norwell, MA, 2001.

[48] Song, J.-S. and Yao, D.D., Performance Analysis and Optimization in Assemble-to-Order Systems with Random Leadtimes. *Operations Research* **50** (2002), 889-903.

[49] Swaminathan, J. and Tayur, S., Managing Broader Product Lines through Delayed Differentiation Using Vanilla Boxes, *Management Science* **44** (1998), S161-S172.

[50] Swaminathan, J., S. Tayur, Stochastic Programming Models for Managing Product Variety, Chapter 19 in S. Tayur, R. Ganeshan and M. Magazine, eds., *Quantitative Models for Supply Chain Management*, Kluwer, Boston, 1999.

[51] Thonemann, U. and Brandeau, M., Optimal Commonality in Component Design. *Operations Research*, **48** (2000), 1-19.

[52] Topkis, D. *Optimal Ordering and Rationing Policies in a Nonstationary Dynamic Inventory Model with n Demand Classes, Management Science* **15** (1968), 160-176.

[53] Van Mieghem, J. and N. Rudi, Newsvendor Networks: Inventory Management and Capcity Investment with Discretionary Activities, working paper, Northwestern University, 2001. Forthcoming, *Manufacturing & Service Operations Management*.

[54] Veinott, A. The Optimal Policy for a Multi-Product, Dynamic, Nonstationary Inventory Problem. *Management Science* **12** (1965), 206-222.

[55] de Vericourt, F., Karaesmen, F. and Dallery, Y., Optimal Stock Rationing for a Capacitated Make-to-Stock Production System, working paper, Laboratoire Productique et Logistique, Ecole Centrale de Paris, 1999. Forthcoming, *Management Science*.

[56] Wang, Y. Near-Optimal Base-Stock Policies in Assemble-to-Order Systems under Service Levels Requirements, working paper, MIT Sloan School, 1999.

[57] Wemmerlov, U., Assembly-to-order manufacturing: Implications for Materials Planning. *Journal of Operations Management*, **4** (1984), 347-368.

[58] Xu, S.H. Dependence Analysis of Assemble-to-Order Systems. Chapter 11 in J. Song and D. Yao, eds. *Supply Chain Structures: Coordination, Information and Optimization.* Kluwer Academic Publishers, Norwell, MA, 2001, 359-324.

[59] Zhang, A.X. Optimal Order-Up-To Policies in an Assemble-To-Order System with a Single Product, Working paper, School of Business Administration, University of Southern California, 1995.

[60] Zhang, A.X., Demand Fulfillment Rates in an Assemble-to-Order System with Multiple Products and Dependent Demands, *Production and Operations Management* **6** (1997), 309-324.

Figure 1. Assemble-to-Order System

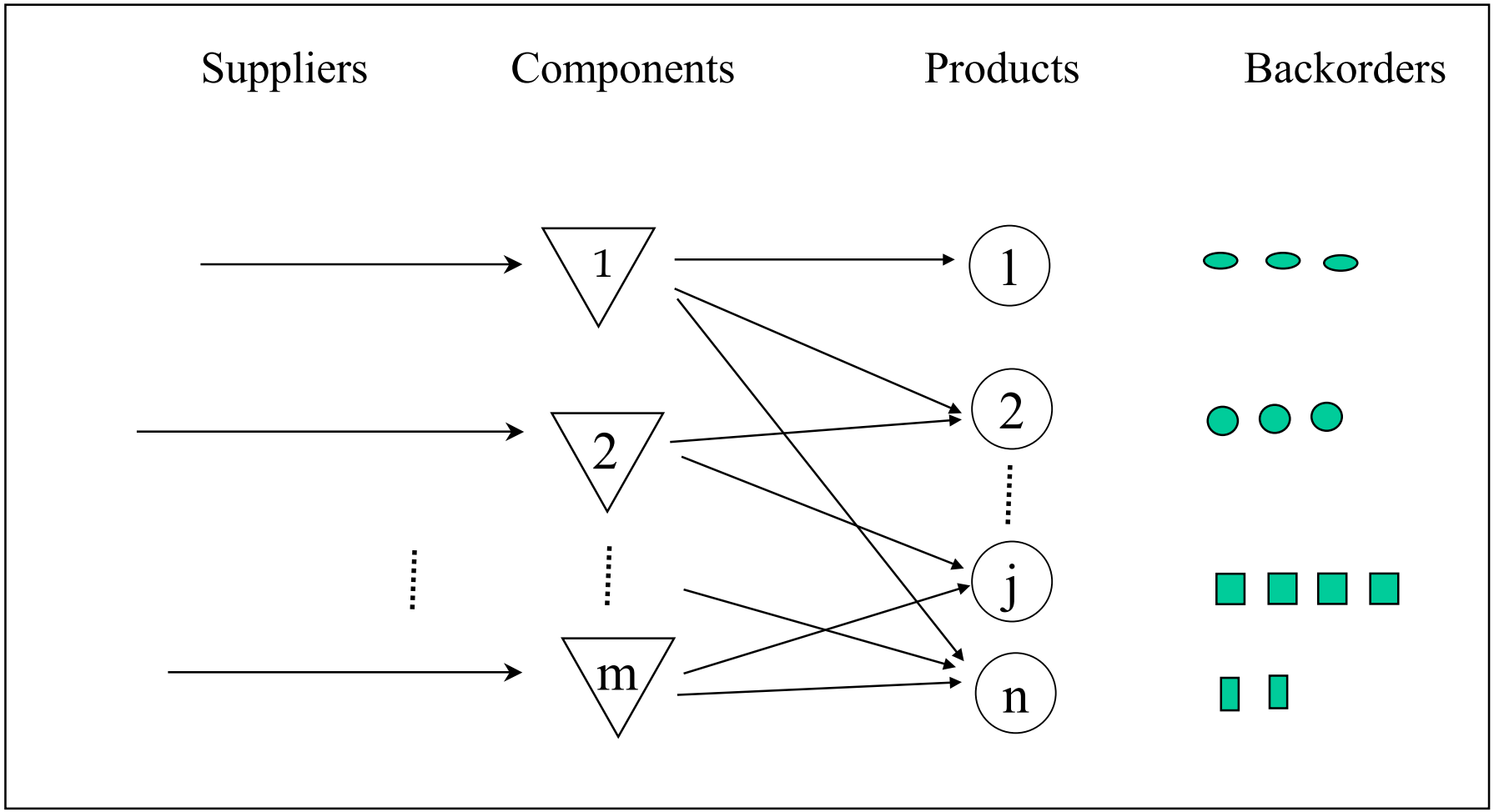

Figure 2. Special Cases