IIE Transactions (2003) **35**, 389–403
0740-817X © 2003 "IIE"
0740-817X/03 $12.00 + .00
DOI: 10.1080/07408170390184107

On assemble-to-order systems with flexible customers

S.M.R. IRAVANI[1], K.L. LUANGKESORN[1] and D. SIMCHI-LEVI[2]

[1]*Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA*
E-mail: iravani@iems.northwestern.edu or lluang@northwestern.edu
[2]*Department of Engineering Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
E-mail: dslevi@MIT.edu

Received June 2000 and accepted April 2002

We consider an assemble-to-order system in which each customer order consists of a mix of key and non-key items. Key items are items where the customer is lost if one or more of his key items are not available. The non-key items are items which are not essential, and therefore, customers may ignore those items if they are not available. We also allow substitution in our model, and assume that some proportion of customers accept substitutions for missing items. We then develop a quasi-birth-and-death process which can be used to obtain the performance measures of the assemble-to-order system. We also introduce new measures which evaluate different levels of customer satisfaction. Through a numerical study, we reveal some interesting behaviors of the system and explore the underlying causes.

1. Introduction

Assemble-to-order is a manufacturing strategy where parts and sub-assemblies are produced-to-stock, while the final assembly of products is delayed until customer orders have been received. Customer orders may require the assembly of different sets of items (components). For example, a computer manufacturer such as Dell receives a variety of different orders which consist of different size monitors, different capacity hard drives, different keyboard types and so on. These orders are then assembled from available stocks of parts. In manufacturing systems where: (i) components production times are relatively large compared to assembly times; and (ii) considerable commonality of components exists among orders (products), the assemble-to-order strategy is more likely to emerge (Gerchak and Henig, 1989). This strategy allows manufacturers to achieve a high degree of product variety and quick product delivery while keeping low inventories. See Wemmerlov (1984) for implications of assemble-to-order manufacturing on production planning and its comparison with traditional produce-to-order and produce-to-stock systems.

Because of the highly competitive market, performance measures such as order service level and fill rate have become the most critical performance measures of assemble-to-order systems, and therefore, almost all research efforts in studying assemble-to-order systems have somehow focused on the effects of system parameters on order fulfillment. The major trade-off in inventory models, including assemble-to-order systems, is between the inventory level and the order service level and fill rate. It is intuitively clear that in all inventory systems, increasing inventory levels (base-stock levels) increases order fill rates and service levels. For example, it is commonly accepted (and shown in Song *et al.* (1999) by numerical example) that in an assemble-to-order system item-based and order-based service levels increase as the base-stock levels of items increase. However, in this paper we will show that if customers are very selective, then increasing the base-stock level of items might lead to a decrease in item-based and order-based service levels. In this paper, we also examine the effects of item substitution in assemble-to-order systems. More specifically, we assume that if the manufacturer is not able to fill an order completely, she may be able to persuade the customer to choose a different product mix. While this may seem to improve the overall service, we will show that under certain conditions, it can cause a decline in overall customer satisfaction and/or manufacturer's profit.

2. Selective and flexible customers

In this section we define the concept of customer flexibility and customer selectivity based on the customer's willingness to substitute or ignore some important features of his original order.

Selective customers: A selective customer is a customer who is not going to buy a product unless that product has a set of specific features (key items). The selective customer is sensitive towards his key items, but is flexible regarding his non-key items. For example, consider a customer who is looking for a computer system with a 19″ monitor, a 20 GB hard drive, 128 MB of memory, sound card and speakers. For this customer the 19″ monitor or 20 GB hard drive may be key items that he does not compromise. Therefore, if one of these items is not available, he will not buy the computer. However, if these two (key) items are available, then the customer will buy the system even if the speakers or sound card (non-key items) were not available (since he can use the computer and purchase a sound card or speakers any time in the future).

Flexible customers: A flexible customer is a customer who is willing to compromise his key items. In the above example, if a customer is willing to get a 21″ monitor instead of 19″ monitor when 19″ monitors were not available, then that customer is called a flexible customer.

The above definitions are based on the two concepts of key/non-key items and substitution. Key items of an order are *often* items that provide the required functionality for the product, and therefore, the product will not fulfill its intended purpose without its key items. For example, a computer is not functional without the hard drive, or monitor. However, non-key items are items which are not essential for the purpose of the product. For example a computer can function without accessories (items) such as CD writer or DVD-ROM.

We would like to emphasize that classifying the order of a customer into two classes of key and non-key items must be done not only based on the necessity of the items regarding the functionality of the product, but also based on the customer's preferences. For example, a car with two doors instead of four, or one that was red instead of black will still function; however, having two doors or being red could be a key feature (key item) for relatively young car buyers. Thus, having two doors or being red can be considered as key features (key items) for that group of customers.

The concept of substitution and its effect on the customer satisfaction was another factor that motivated this paper. For example, car dealers who are faced with a backlog of vehicles already on their lots usually try to persuade customers to buy one of their available cars instead of what the customers want and they do not have. Regarding this issue, Hyde (2000) reports that Harold Kutner, the GM's director of worldwide purchasing, said in 2000:

"We know from our surveys that a good percentage of people who buy vehicles from the dealer inventory position walk away and three months after the purchase are dissatisfied because they didn't get something they really wanted."

Thus, it is important to somehow measure the fraction of customers who finally buy a product, but not exactly what they originally had in mind. In this paper we develop a model in order to measure customer satisfaction in different levels such as full satisfaction, key satisfaction and substitution satisfaction. This paper also examines the effects of item substitution in assemble-to-order systems, and shows that while substitution may seem to improve the overall service, it may cause a decline in overall customer satisfaction or system's profit.

3. Literature

Assemble-to-orders are multi-item inventory systems with dependent demands across items. The literature on these systems can be divided into two groups: (i) systems with deterministic supply processes, e.g., deterministic lead times; and (ii) systems with stochastic supply processes. Srinivasan *et al*. (1992), Agrawal and Cohen (1995), Schraner (1996), Hausman *et al*. (1998) and Song (1998, 2000), are examples with assemble-to-order systems with deterministic supply processes in which the inventory of items are continuously reviewed and controlled according to base-stock policies. Zhang (1997), however, assumed that the inventory of items are managed under a periodically reviewed, decentralized order-up-to policies. In his model, the lead times are deterministic and are integer multiples of the base period length.

Assemble-to-order systems with i.i.d. replenishment lead times and Poisson demand were studied by Cheung and Hausman (1995). They derived exact expressions and approximations for service distribution and the expected number of backorders. Zhang (1999) considered a similar problem where each item is supplied by a dedicated production facility with general i.i.d. production times. They obtained the expected total waiting times of each order type to measure the performance of the system. Song *et al*. (1999) studied a similar model but with exponential production (lead) times, and derived the performance measures of the system at item, order and overall system levels under backlog and lost sales assumptions. She also obtained the waiting time distribution of backlogged orders. Glasserman and Wang (1998) analyzed the trade-off between inventory levels and the delivery lead times in a model similar to Zhang (1999). Their study includes models where a given order may include more than one of a given item.

Our model adds two features to the assemble-to-order system studied in Song *et al*. (1999). These new features incorporate *customer preferences* and *flexibility* into the model and create a more realistic assemble-to-order system. We incorporate these two features by using the concepts of key/non-key items and item substitution. The contribution of this paper is not only the analysis of more realistic assemble-to-order systems; but more importantly: (i) introducing new performance measures to evaluate the quality of service in assemble-to-order systems; and (ii) revealing some unexpected behaviors of these systems under certain conditions. We will investigate these unexpected behaviors in order to gain insight on how to effectively manage assemble-to-order systems.

The rest of this paper is organized as follows. In the next sections we first incorporate customer flexibility in assemble-to-order systems, then we formulate the system as a continuous-time Markov chain. We will then show how matrix geometric solution techniques can be used to obtain its steady-state performance. In Section 5.2 we introduce new performance measures for systems with flexible and selective customers, and in Section 6, through our numerical study, we present some insights on how to better manage assemble-to-order systems with flexible and selective customers. Section 7 contains some concluding remarks.

4. Model description

Consider an assemble-to-order system consisting of a set (Ω) of n items (components) ($\Omega = \{1, 2, \ldots, n\}$) from which m orders (products) are assembled, see Fig. 1. Item j is produced in production facility j with rate (capacity) μ_j per unit time according to a produce-to-stock routine with a base-stock level s_j, $j = 1, 2, \ldots, n$. The processing times at facility j is assumed to be exponential random variables.

There are m separate demands. Demand for order type i is stationary and is a Poisson process with an arrival rate λ_i, $i = 1, 2, \ldots, m$. Each demand type i is comprised of a subset of items Ω_i, with certain items, set K_i, being key items and the remaining items, set $\overline{K_i}$, being non-key, where $K_i \cup \overline{K_i} = \Omega_i$ and $K_i \cap \overline{K_i} = \emptyset$. Key items for a selective customer, as we mentioned before, are items which are important, and therefore the selective customer does not compromise or accept any substitution for them. If any item in the set of key items, K_i, is not available, the order (the selective customer) for type i will be lost. However, flexible customers might accept substitution for their key items or ignore them if they are not available. Non-key items, on the other hand, are items for which a customer (flexible or selective) will either accept substitutions or ignore them if they are not available. Therefore, if a non-key item is not available, the customer, flexible or selective, will not be lost. We assume that each demand will request at most one of each item.

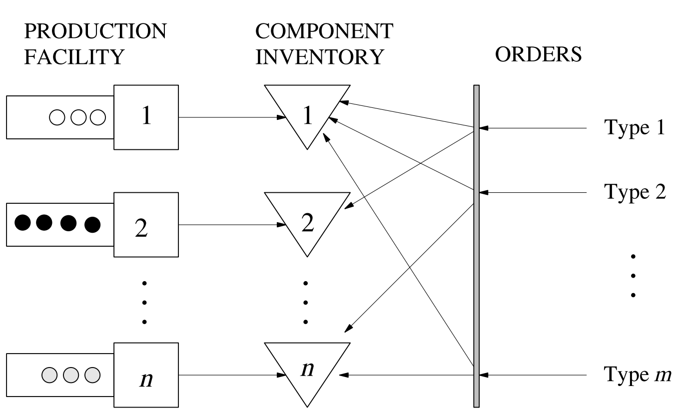

Fig. 1. An assemble-to-order system.

Remark 1. For systems with lost sales, the Total Order Service (TOS) model and the Partial Order Service (POS) model presented in Song *et al*. (1999) are special cases of our model. The TOS model which assumes that an order is either satisfied entirely or rejected entirely represents the special case when $K_i = \Omega_i$, and $\overline{K_i} = \emptyset$ for all $i = 1, 2, \ldots, n$ (i.e., all items are key). On the other hand, the POS model in which an order can be partially satisfied by any combination of the available items represents the special case when $K_i = \emptyset$ and $\overline{K_i} = \Omega_i$ for all $i = 1, 2, \ldots, n$ (i.e., all items in the order are non-key).

Define $I_j(t)$ as the on-hand inventory of item j at time t, and $N_j(t)$ as the number of items in the production facility j at time t, then considering the base-stock level s_j for item j,

$$I_j(t) + N_j(t) = s_j \quad \forall j = 1, 2, \ldots, n,$$

and stochastic process $N = \{N_1(t), N_2(t), \ldots, N_n(t), t \geq 0\}$ will be a continuous-time Markov chain with finite state space $\mathcal{N} = \{\mathbf{n} | \mathbf{n} = (n_1, n_2, \ldots, n_n);\ 0 \leq n_j \leq s_j;\ \forall j\}$. Before we present the transition rate matrix for this Markov chain, we must first introduce substitution probabilities.

4.1. *Substitution probabilities*

We define $p^i_{a,j}(\mathbf{n})$ to be the probability that item a of order i is substituted by item j when: (i) the state of the system is $\mathbf{n}$; (ii) item i is not available; and (iii) item j is available. More specifically, when state of the system is $\mathbf{n}$, the substitution probabilities $p^i_{a,j}(\mathbf{n})$ are defined as follows:

- $p^i_{a,j}(\mathbf{n})$ = The probability that customer type i agrees to receive item j ($j \in \Xi^i_a$) as a substitute for unavailable item a. This means that the customer leaves the system with item j and possibly some other items.
- $p^i_{a,0}(\mathbf{n})$ = The probability that customer type i agrees to ignore item a and does not get any substitute for that item. This means that the customer leaves the system without item a, but with possibly some other items.
- $\overline{p^i_a}(\mathbf{n})$ = The probability that customer type i neither agrees to receive a substitution for item a, nor ignores that item, and therefore the customer is lost. $\overline{p^i_a}(\mathbf{n}) = 0$ for all non-key items ($a \in \overline{K_i}$).

Suppose that when state of the system is $\mathbf{n}$, set $\Xi^i_a(\mathbf{n})$ is the set of all items that are offered to customer of type i as a substitute for item a when item a is not available. Then, we will have,

$$\sum_{j \in \Xi^i_a(\mathbf{n})} p^i_{a,j}(\mathbf{n}) + p^i_{a,0}(\mathbf{n}) + \overline{p^i_a}(\mathbf{n}) = 1, \quad \forall a \in \Omega_i,$$
$$i = 1, 2, \ldots, m, \quad \mathbf{n} \in \mathcal{N}. \tag{1}$$

Note that for all non-key items $a \in \overline{K_i}$, we have $\overline{p_a^i}(\mathbf{n}) = 0$, $\forall \mathbf{n} \in \mathcal{N}$. This is because customers are not lost if their non-key items are not available. Based on the above definitions, it becomes clear that the customer type i is a selective customer if $\overline{p_a^i}(\mathbf{n}) = 1$ for all of his key items ($\forall a \in K_i, \forall \mathbf{n} \in \mathcal{N}$). This means that the customer type i will not accept any substitution for any of his key items. On the other hand, if $\overline{p_a^i}(\mathbf{n}) \leq 1$ for at least one key item $a \in K_i$, then customer type i is a flexible customer who might accept substitution for at least one of his key items, or might ignore that item.

In practice, the set of substitute items $\Xi_a^i(\mathbf{n})$ and therefore the substitution probabilities depend on three major factors:

1. *Manufacturer's substitution strategy*: Although an item can be substituted by several other items, manufacturers often have their own preferences regarding the substitution alternatives. Based on their preferences: (i) they may only offer some of the substitution alternatives to the customer; and (ii) they often offer substitute items in an order that they prefer. For example, suppose unavailable item a can be substituted by available items b or c or d. Based on profits of items b, c and d, the manufacturer may not offer item d as a substitute for a, since it is a key item of a more profitable order. Furthermore, the manufacturer may first suggest item b as a substitute for a, since item b is more profitable than c. If the customer does not accept item b, then the manufacturer suggests item c.
2. *Items availabilities*: Items availabilities effect the substitution probabilities, since: (i) if some substitute items are not available, then they cannot be offered as a substitute; and (ii) items availabilities may influence the manufacturer's decisions on which items to offer as a substitute, and in what order those items must be offered.
3. *Customer preferences*: Customer preferences are the final factor in determining the substitution probabilities, since the customer has the last word on accepting or rejecting the substitutes.

Substitution probabilities can be obtained based on customer surveys which are designed to incorporate the above three factors. As an example, suppose that a customer of type i wants item a, and that item can be substituted by item b, c, or d. Also, assume that if item a is not available, the manufacturer's substitution strategy is to first offer item b as a substitute, then c and then d. Furthermore, suppose that customer survey of 100 potential customers who wanted product i, which includes item a, indicates that: (i) 50 customers accept b as a substitute for a if a is not available; (ii) among the other 50 customers who do not accept b as a substitute for a, 25 customers accept c as a substitute for a; (iii) among the remaining 25 customers who do not accept either b or c as a substitute for a, 10 customers accept d; and (iv) among the remaining 15 customers who do not accept b, c or d, five customers ignore a and buy the other items of their orders and 10 customers reject to buy any items (lost customers). Thus, if item a is not available, but items b, c and d are available, then based on the above manufacturer's substitution strategy and customer preferences, the substitution probabilities can be approximated as $p_{a,b}^i(\mathbf{n}) = 0.50$, $p_{a,c}^i(\mathbf{n}) = 0.25$, $p_{a,d}^i(\mathbf{n}) = 0.10$, $p_{a,0}^i(\mathbf{n}) = 0.05$, and $\overline{p_a^i}(\mathbf{n}) = 0.10$.

Note that the above substitution probabilities are based on the assumptions that: (i) items b, c and d are available; and (ii) the manufacturer offers substitution for item a in the order b, then c and then d. Now suppose that if items a and b are not available, the manufacturer's substitution strategy is to offer c first and then d. Also, assume that the same customer survey indicates that if item a is not available and the customer is first being offered item c as a substitute, then: (i) 45 customers accept c; (ii) among the other 55 customers who do not accept c as a substitute for a, 25 customers accept d as a substitute; and (iii) among 30 customers who do not accept d, 10 customers ignore item a and buy the other items of their order, while 20 customers reject to buy any items (lost customers). Therefore, if items a and b are not available, but items c and d are available, then based on the manufacturer's substitution strategy and customer preferences, the substitution probabilities can be approximated as $p_{a,c}^i(\mathbf{n}) = 0.45$, $p_{a,d}^i(\mathbf{n}) = 0.25$, $p_{a,0}^i(\mathbf{n}) = 0.10$, and $\overline{p_a^i}(\mathbf{n}) = 0.20$.

In systems where unavailable items are substituted, customers who initially want a particular set of items may leave the system with a different set of items. When state of the system is $(\mathbf{n})$, let

- $\mathscr{P}_{(\Omega_i,\Omega_i')}(\mathbf{n})$ = The probability that an arriving customer of type i who initially requested set of items Ω_i accepts substitutions for some of his items and leaves the system with a new set of items $\Omega_i' \in \Psi_i(\mathbf{n})$.

When the state of the system is $\mathbf{n}$, set $\Psi_i(\mathbf{n})$ includes the combination of all different sets of items created by substitution or ignoring of unavailable items requested by the order of type i. For example, consider a flexible customer of type i with $K_i = \Omega_i = \{a, b, c\}$. Suppose at state $\mathbf{n}$ items a and b are not available, but items a', a'' and b' are available, where $\Xi_a^i(\mathbf{n}) = \{a', a''\}$, $\Xi_b^i(\mathbf{n}) = \{b'\}$ and $\Xi_c^i = \emptyset$. Then, at state $\mathbf{n}$, the set of all possible substitutions for order i is $\Psi_i(\mathbf{n}) = \{\{a', b', c\}, \{a'', b', c\}, \{a', c\}, \{a', c\}, \{b', c\}, \{c\}\}$.

The joint probabilities $\mathscr{P}_{(\Omega_i,\Omega_i')}(\mathbf{n})$ can be approximated in two different ways:

1. In systems where there are not many alternatives for substitution, the set of all possible substitutes for each order i (set $\Psi_i(\mathbf{n})$) is a small set. Thus, for those systems it is easier to approximate joint probabilities $\mathscr{P}_{(\Omega_i,\Omega_i')}(\mathbf{n})$ directly from past sales data or customer surveys regarding the small number of substitutions made.

2. In systems where the demand for custom configured products is high and customers have many substitution alternatives, it is not easy to directly estimate the joint probabilities $\mathscr{P}_{(\Omega_i,\Omega_i')}(\mathbf{n})$. This is because the set of joint probabilities $\mathscr{P}_{(\Omega_i,\Omega_i')}$ is very large, due to the large number of different substitution (product configurations) created by different customers. In these systems, it is easier to approximate these joint probabilities based on the substitution probabilities. For example, suppose at state $\mathbf{n}$ the customer type i requests items $\Omega_i = \{a, b, c, d, e, f, g\}$ where his key items are $K_i = \{a, b, c, d\}$, and items a, b, c and f are not available. Furthermore, let at state $\mathbf{n}$ the set of substitute items for a, b, c and f be $\Xi_a^i(\mathbf{n}) = \{a', a''\}$, $\Xi_b^i(\mathbf{n}) = \{b', b''\}$, $\Xi_c^i = \{c'\}$ and $\Xi_f^i(\mathbf{n}) = \{f', f''\}$, respectively. Then the customer accepts configuration $\Omega_i' = \{a', b', d, e, f', g\}$ with probability

$$\mathscr{P}_{(\Omega_i,\Omega_i')}(\mathbf{n}) = p_{a,a'}^i(\mathbf{n}) p_{b,b'}^i(\mathbf{n}) p_{c,0}^i(\mathbf{n}) p_{f,f'}^i(\mathbf{n}). \qquad (2)$$

Equation (2) assumes independence among substitution probabilities. This might not be true in systems where there are not many substitution alternatives, and thus substituting item a with a' or a'' might effect the substitution probabilities $p_{b,b'}^i(\mathbf{n})$ or $p_{c,0}^i(\mathbf{n})$ or $p_{f,f'}^i(\mathbf{n})$. As we mentioned above, for those systems it is better to approximate joint probabilities $\mathscr{P}_{(\Omega_i,\Omega_i')}(\mathbf{n})$ directly from past sales data or customer surveys. However, assuming independence among substitution probability of different items becomes more acceptable in systems that offer a large variety of options for substitution. For example, for their Dimension 8100 desktop, Dell offers nine different options for monitor, four different options for speakers and nine different options for a DVD-ROM or CD-ROM. Depending on their own requirements, customers can make their choice of DVD-ROM or CD-ROM independent of their choice of monitors and their choice of speakers.

To simplify our notation, in the remainder of this paper we suppress $(\mathbf{n})$ in $p_{a,j}^i(\mathbf{n})$, $\mathscr{P}_{(\Omega_i,\Omega_i')}(\mathbf{n})$, $\Xi_a^i(\mathbf{n})$ and $\Psi_i(\mathbf{n})$, and use $p_{a,j}^i$, $\mathscr{P}_{(\Omega_i,\Omega_i')}$, Ξ_a^i and Ψ_i instead.

4.2. *Problem formulation*

In Section 4 we defined the continuous-time Markov chain that represents our assemble-to-order system. In this section, we present the transition rate matrix of the Markov chain. Let, $\mathbf{1}^j$ be an n element row vector with the jth element being 1 and the remaining $n-1$ elements being zeros, $\mathbf{1}^j = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$. Then, the transition rate matrix for the continuous time Markov Chain, $\boldsymbol{A}$, has elements $q_{\mathbf{n},\mathbf{n}'}$, where

$$q_{\mathbf{n},\mathbf{n}'} = \begin{cases} \mu_j; & \text{if } \mathbf{n}' = \mathbf{n} - \mathbf{1}^j, \quad n_j > 0, \quad j = 1, 2, \ldots, n, \\ \lambda_i; & \text{if } \mathbf{n}' = \mathbf{n} + \sum_{k\in\Omega_i} \mathbf{1}^k, \quad n_k < s_k \forall k \in \Omega_i, \quad i = 1, 2, \ldots, m, \\ \sum_{u=1}^m \mathscr{P}_{(\Omega_u,\Omega')}\lambda_u; & \text{if } \mathbf{n}' = \mathbf{n} + \sum_{k\in\Omega'} \mathbf{1}^k, \quad n_k < s_k \forall k \in \Omega', \quad \Omega' \in \bigcup_{i=1}^m \Psi_i, \\ -\sum_{\mathbf{n}''\in\mathscr{N},\mathbf{n}''\neq\mathbf{n}} q_{\mathbf{n},\mathbf{n}''}; & \text{if } \mathbf{n}' = \mathbf{n}, \\ 0; & \text{otherwise.} \end{cases} \qquad (3)$$

In (3), the transition rate μ_j from state $\mathbf{n}$ to $\mathbf{n}' = \mathbf{n} - 1^j$ occurs when the production of item j is completed in production facility j. That reduces the number of items in production facility j by one. Transition rate λ_i from states $\mathbf{n}$ to state $\mathbf{n}' = \mathbf{n} + \sum_{k\in\Omega_i} \mathbf{1}^k$ only occurs if in state $\mathbf{n}$ all the items of order type i are available ($n_k < s_k$ for all $k \in \Omega_i$). In these transitions item substitution does not occur. However, transition rate $\sum_{u=1}^m \mathscr{P}_{(\Omega_u,\Omega')}\lambda_u$ from state $\mathbf{n}$ to $\mathbf{n}' = \mathbf{n} + \sum_{k\in\Omega'} \mathbf{1}^k$ occurs when an order of type u with arrival rate λ_u arrives and, due to unavailability of some items, the customer accepts substitutions (for unavailable items) then leaves the system with the set of items Ω' instead of Ω_u. Note that: (i) $\sum_{k\in\Omega'} \mathbf{1}^k$ refers to the available items k ($n_k < s_k$) that forms the set of Ω' accepted by customer type u; (ii) the summation in $\sum_{u=1}^m \mathscr{P}_{(\Omega_u,\Omega')}\lambda_u$ is due to the fact that customers of different types $u = 1, 2, \ldots, m$ might end up leaving with the same set of items Ω'; and (iii) $\Omega' \in \bigcup_{i=1}^m \Psi_i$ guarantees that all possible substitution scenarios are considered.

The fourth equation in (3) is the diagonal elements of the transition rate matrix which makes the summation of all elements in each row equal zero. Finally, the zeros in matrix $\boldsymbol{A}$ represent state transitions which are not possible.

4.3. *Matrix decomposition approach*

The Markov chain in (3) has $\prod_{j=1}^n (s_j + 1)$ states. This means that the steady-state probability distribution can be obtained by solving a system of $\prod_{j=1}^n (s_j + 1)$ linear equations. However, by rearranging the state space $\mathscr{N}$ it can be shown that the transition rate matrix $\boldsymbol{A}$ then represents a quasi-birth-and-death process. This is because a state transition can only occur if a demand arrives (inventory of items in the order decreases by one) or an item is produced in one of the production facilities (inventory of the item increases by one).

We arrange the states of the system, similar to Song *et al.* (1999), in order $\mathscr{N} = \{N_1, N_2, \ldots, N_n\}$, where N_k is the set of states where there are $n_1 = k$ items in production facility 1. Thus,

$$N_k = \{(k,0,\dots,0,0),\dots,(k,0,\dots,0,s_n),\dots,(k,0,\dots,1,0),\dots,(k,0,\dots,1,s_n),\dots,(k,0,\dots,s_{n-1},0),\dots,(k,0,\dots,s_{n-1},s_n),\dots,(k,s_2,\dots,s_{n-1},0),\dots,(k,s_2,\dots,s_{n-1},s_n)\},$$

and the transition rate matrix $\boldsymbol{A}$ will be

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{A}_0 & 0 & 0 & \dots & 0 & 0 \\ \mu_1\boldsymbol{I} & \boldsymbol{A}-\mu_1\boldsymbol{I} & \boldsymbol{A}_0 & 0 & \dots & 0 & 0 \\ 0 & \mu_1\boldsymbol{I} & \boldsymbol{A}-\mu_1\boldsymbol{I} & \boldsymbol{A}_0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \mu_1\boldsymbol{I} & \mathbf{A}-\mu_1\boldsymbol{I} & \boldsymbol{A}_0 \\ 0 & 0 & 0 & \dots & 0 & \mu_1\boldsymbol{I} & \boldsymbol{A}_1-\mu_1\boldsymbol{I} \end{pmatrix}, \tag{4}$$

where $\boldsymbol{A}$, $\boldsymbol{A}_0$ and $\boldsymbol{A}_1$ are square matrices of order $\prod_{j=2}^{n}(s_j+1)$. We omit the details of these three matrices (to avoid heavy notation); however, these matrices can be easily constructed using (3).

Let $\mathbf{p}_k$ be the steady-state probability vector of states where the inventory on order of item 1 is k. Then the steady-state probability distribution of $\boldsymbol{A}$ is $\mathbf{P} = (\mathbf{p}_0,\mathbf{p}_1,\mathbf{p}_2,\dots,\mathbf{p}_{s_1})$. It can be easily shown that the Markov chain is irreducible, and therefore, the steady-state probability vector $\mathbf{P}$ exists and is unique. The balance equations for the quasi-birth-and-death process are then,

$$\mathbf{p}_0\boldsymbol{A} + \mu_1\mathbf{p}_1 = 0,$$
$$\mathbf{p}_{k-1}\boldsymbol{A}_0 + \mathbf{p}_k(\boldsymbol{A}-\mu_1\boldsymbol{I}) + \mu_1\mathbf{p}_{k+1} = 0 \quad 1 \le k \le s_1,$$
$$\mathbf{p}_{s_1-1}\boldsymbol{A}_0 + \mathbf{p}_{s_1}(\boldsymbol{A}_1-\mu_1\boldsymbol{I}) = 0,$$

with the normalization condition $\sum_{k=0}^{s_1}\mathbf{p}_k\mathbf{e} = 1$, where $\boldsymbol{I}$ is the unit vector and $\mathbf{e}$ is a column vector of ones. The structure of these balance equations are similar to those in Song *et al.* (1999), and thus, the solution can be similarly obtained as:

$$\mathbf{p}_k = \mathbf{p}_{k-1}\boldsymbol{R}_k, \quad k = 1,2,\dots,s_1, \tag{5}$$

where $\boldsymbol{R}_0 = \boldsymbol{I}$,

$$\boldsymbol{R}_{s_1} = -\boldsymbol{A}_0(\boldsymbol{A}_1-\mu_1\boldsymbol{I})^{-1}$$

and $\boldsymbol{R}_k$ for $k = 1,2,\dots,s_1-1$ can be obtained using recursive equation

$$\boldsymbol{R}_k = -\boldsymbol{A}_0[(A-\mu_1\boldsymbol{I}) + \mu_1\boldsymbol{R}_{k+1}]^{-1}$$

Now, using balance Equation (3) recursively, and writing $\mathbf{p}_k$ in terms of $\mathbf{p}_0$, and substituting the results into the normalizing equation we get,

$$\sum_{k=0}^{s_1}\mathbf{p}_k\mathbf{e} = \mathbf{p}_0\sum_{k=0}^{s_1}\prod_{i=0}^{k}\boldsymbol{R}_i\mathbf{e} = 1.$$

The above equation can be used to find probability vector $\mathbf{p}_0$. When $\mathbf{p}_0$ is found, then the steady state probabilities $\mathbf{p}_k$ for $k = 1,2,\dots,s_1$ can be obtained using (5).

We would like to emphasize that while our model assumes Poisson demand arrivals and exponential processing times, the insights gained from the results do not depend on the properties of Poisson arrivals or exponential service times. The insights such as the relative responses of the system to substitutions and/or impact of key-service on profit are mostly the results of the system's dynamics (e.g., order structures, production capacities, etc.) rather than the assumptions regarding Poisson arrival or exponential production times.

5. Performance measures

We evaluate the performance of the system from three different perspectives: (i) items-based performance; (ii) order-based performance; and (iii) system performance. Having probabilities $\mathbf{p}_k$ for all $k = 0,1,2,\dots,s_1$, these performances can be easily measured as we describe in the next sections.

5.1. *Item-based performance*

Item-based performance reflects the availability of an item when requested. In standard inventory models, this is usually measured by F_j, the probability of immediately satisfying a demand for item j. Thus in our model, if I_j is the inventory on hand of item j, then

$$F_j = \Pr\{I_j \ge 0\}, \quad j = 1,2,\dots,n.$$

The above probability is the summation of all steady-state probabilities in which the number of items in the production facility j is less than s_j, and can be easily obtained using steady-state probability vector $\mathbf{p}_k$, $k = 0,1,2,\dots,s_1$.

5.2. *Order-based performance*

The order-based performance represents the customer satisfaction regarding each order type. One of the standard order-based performance measures is service level. Service level of order type i, SL_i, is the probability that a customer is not lost. However, since we included customer preferences and flexibility, we must go beyond standard order-based performance measures and define new measures for served customers. We divide the served customers into three groups, each representing a different level of satisfaction:

- *Fully satisfied customers*: These are the customers who received exactly what they requested. In other words, these are customers who upon arrivals found the inventory of all items in their orders non-empty and are able to receive their complete order. The percentage of customers (orders) of type i who are fully satisfied, FS_i, can be obtained as follow:

$$FS_i = \Pr\{I_j \ge 0; \forall j \in \Omega_i\}, \quad i = 1,2,\dots,m.$$

The right-hand side of the above equation is the summation of the probabilities of all states in which the inventory of each item in order type i is not empty.

- *Key satisfied customers*: These are customers who received all of their key items since they found the inventory of their key items non-empty. Key satisfied customers may get some or all of their non-key items or may accept substitutions for their non-key items. Of course, this group includes fully satisfied customers. Thus, if KS_i represents the percentage of customers of type i who at least get all of their key items, then

$$KS_i = \Pr\{I_j \geq 0; j \in K_i\}, \quad i = 1, 2, \ldots, m.$$

- *Substitution satisfied customers*: These are customers who have accepted substitution for at least one of their key items. These customers upon arrival find the inventory of some of their key items empty, however, they may accept substitutions or ignore them and leave the system with a different set of items. Note that substitution satisfied customers do not include customers who get substitution for some of their non-key items while receiving all of their key items. We did not include these customers in the class of substitution satisfied since these customers have already been counted in the class of key satisfied customers.

  In order to find SS_i, the percentage of customers of type i who accept a substitution for their key items, we will have,

$$SS_i = \sum_{\Omega_i' \subset \Psi_i} \mathbf{P}_{\Omega_i'} \mathscr{P}_{(\Omega_i, \Omega_i')},$$

  where $\mathbf{P}_{\Omega_i'}$ is the probability of system being in a state which allows switches from order type i to a set of items Ω_i' by substitution or ignoring of at least one key item.

We would like to specifically emphasize on the order-based performance measures in three levels; FS_i, KS_i, and SS_i, since we believe that customer satisfaction is the key to success in today's competitive market. In all existing literature on assemble-to-order systems with lost sales, the customers are either considered lost or satisfied. However, in reality, we believe satisfaction has different levels and considering only lost customers as unsatisfied might have long term consequences. We believe recognizing that some customers who are served may be dissatisfied helps managers have a better understanding of the consequences of their decisions, and therefore, helps managers to improve the performance of their system.

5.3. *System performance*

System performance combines the customer satisfactions of different orders into one measure which reflects the overall performance of the system. If FS, KS, and SS are the overall full, key and substitution satisfaction in the system, then

$$SL = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda} SL_i, \quad FS = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda} FS_i, \quad KS = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda} KS_i,$$

$$SS = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda} SS_i,$$

where, $\lambda = \sum_{i=1}^{m} \lambda_i$. As is clear, the overall satisfaction is a weighted average of the order-based satisfactions based on the order arrival rates.

6. Numerical study

In this section, we report the results of the numerical study we conducted. The main purpose is to study the behavior of assemble-to-order systems with selective and flexible customers and gain insight on how to better manage these systems. We present our findings under different scenarios presented in the following sections. First, we show that in some assemble-to-order systems increasing the base-stock levels of some items might, counter to common belief, decrease order-based service levels, and in some cases the total profit. Then, we demonstrate that taking advantage of a customer's flexibility may not only decrease the quality of service, but also may decrease overall profit in the long-run. Finally, we show that decisions regarding the base-stock levels that do not consider different levels of customer satisfaction may become very costly.

6.1. *Base-stock levels versus customer satisfaction*

In Scenario 1 we explore the effects of increasing the base-stock level of an item on all performance measures of the system. Consider an assemble-to-order systems with three ($n = 3$) items which satisfy four different order types ($m = 4$). The production rates for items 1, 2 and 3 are $\mu_1 = 35$, and $\mu_2 = \mu_3 = 80$, respectively. Also, assume that the base-stock levels for those three items are $s_1 = 3$, $s_2 = s_3 = 10$. The data for customers (orders) type 1 through 4 are summarized in Table 1.

In Scenario 1, we explore the effects of increasing the base-stock level s_1 on all performance measures of the system under two different production capacities: (i) $(\mu_1, \mu_2, \mu_3) = (35, 80, 80)$ in Scenario 1a; and (ii) $(\mu_1, \mu_2, \mu_3) = (40, 70, 70)$ in Scenario 1b. In both cases we

Table 1. The assemble-to-order system for Scenario 1

Order type (i)	*Arrival rate* (λ_i)	*Key items* (K_i)	*Non-key items* $(\bar{K}_i)$
1	40	$\{1, 2, 3\}$	–
2	20	$\{2\}$	$\{3\}$
3	20	$\{3\}$	$\{2\}$
4	20	$\{2, 3\}$	–

assume that all customers are selective customers and do not accept any substitutions for their key items ($\overline{p_a^i} = 1$ for all $a \in K_i$, $i = 1, 2, 3, 4$).

As Table 2, Scenario 1a, column $s_1 = 3$ shows, the service level for order of type 1 (SL_1) is significantly lower than the service levels of the other order types. A manager facing this situation can conclude that this is because the service level for item 1, (F_1), is low relative to the other two service levels. This is a reasonable conclusion since item 1 is not only a key item for a type 1 order, but item 1 is only requested by order type 1. Thus, a reasonable decision is to increase s_1 from its current level to, say, $s_1 = 10$. Unfortunately, while this increases the overall performance (FS, KS and SL), it decreases the service level for all order types except order type 1.

The reason that some order-based and item-based performance measures decrease when base-stock level s_1 increases is that by increasing s_1 more customers of type 1 will be served, causing a higher demand seen by other items in the system. This increase in the number of served customers of type 1, who in addition to item 1 require items 2 and 3, creates a higher demand for items 2 and 3. On the other hand, since the base-stock levels of items 2 and 3 are fixed, any increase in demand for items 2 and 3 creates lower service levels for these items (decrease in F_2 and F_3), and consequently, lower service levels for any order which requires these items and does not include item 1.

Now, let's look at Scenario 1a from the perspective of Harold Kutner, the GM's director of worldwide purchasing, who was quoted in Section 2. In his quote, he refers to customers who did not get what they really wanted as *dissatisfied* customers. In Scenario 1a, the percentage of served customers who bought the product, but did not get exactly what they wanted are

$$\frac{SL - FS}{SL} = \frac{0.7051 - 0.7047}{0.7051} = 0.0006,$$

when $s_1 = 3$, and

$$\frac{SL - FS}{SL} = \frac{0.7889 - 0.7188}{0.7889} = 0.0889,$$

when $s_1 = 10$. Therefore, if customers who are not fully satisfied are considered as dissatisfied customers, then,

- When the base-stock level of item 1, s_1, increases from three to 10, the overall service level increases by 0.0838 ($= 0.7889 - 0.7051$). However, the percentage of dissatisfied customers also increases by 0.0883 ($= 0.0889 - 0.0006$). In other words, most of the improvement in service level is due to having more dissatisfied customers.
- One can argue that, although the percentage of dissatisfied customers is increased by 0.0883 in Scenario 1a, the system has a 1.41% more fully satisfied customers ($1.41\% = 0.7188 - 0.7047$). The benefit of a 1.41% increase in the fraction of fully satisfied customers might outweigh the 8.8% increase in the fraction of dissatisfied customers. This might be true; however, we would like to emphasize that an increase

Table 2. Results for Scenarios 1a, 1b, and 1c

Performance measures		*Scenario 1a*		*Scenario 1b*		*Scenario 1c*		
		$S_1 = 3$	$S_1 = 10$	$S_1 = 3$	$S_1 = 10$	$p^1_{1.0} = 0$	$p^1_{1.0} = 0.5$	$p^1_{1.0} = 1$
Items-based	F_1	0.6995	0.9623	0.7833	0.9946	0.6995	0.7151	0.7276
	F_2	0.8994	0.8592	0.8398	0.8062	0.8994	0.8746	0.8534
	F_3	0.8994	0.8592	0.8398	0.8062	0.8994	0.8746	0.8534
Order-based	FS_1	0.5518	0.7003	0.5318	0.6303	0.5518	0.5387	0.5283
	FS_2	0.8067	0.7312	0.6960	0.6342	0.8067	0.7599	0.7204
	FS_3	0.8067	0.7312	0.6960	0.6342	0.8067	0.7599	0.7204
	FS_4	0.8067	0.7312	0.6960	0.6342	0.8067	0.7599	0.7204
	KS_1	0.5518	0.7003	0.5318	0.6303	0.5518	0.5387	0.5283
	KS_2	0.8994	0.8592	0.8398	0.8062	0.8994	0.8746	0.8534
	KS_3	0.8994	0.8592	0.8398	0.8062	0.8994	0.8746	0.8534
	KS_4	0.8067	0.7312	0.6960	0.6342	0.8067	0.7599	0.7204
	SS_1	0.0000	0.0000	0.0000	0.0000	0.0000	0.1257	0.2299
	SL_1	0.5518	0.7003	0.5318	0.6303	0.5518	0.6644	0.7694
	SL_2	0.8994	0.8592	0.8398	0.8062	0.8994	0.8746	0.8534
	SL_3	0.8994	0.8592	0.8398	0.8062	0.8994	0.8746	0.8534
	SL_4	0.8067	0.7312	0.6960	0.6342	0.8067	0.7599	0.7204
System	FS	0.7047	0.7188	0.6303	0.6326	0.7047	0.6714	0.6436
	KS	0.7051	0.7889	0.6789	0.7349	0.7051	0.6900	0.6774
	SS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0503	0.0919
	SL	0.7051	0.7889	0.6789	0.7349	0.7051	0.7403	0.7694

in the number of dissatisfied customers does not always guarantee an improvement in the number of fully satisfied customers. Scenario 1b demonstrates a situation in which an increase in base-stock level for item 1 from three to 10 not only increases the fraction of dissatisfied customers by 6.8%

$$\left(= \frac{0.7349 - 0.6326}{0.7349} - \frac{0.6789 - 0.6303}{0.6789}\right),$$

but also has almost no effect on the overall fraction of fully satisfied customers. This is in addition to the fact that the fraction of fully satisfied customers of types 2, 3 and 4 have already been significantly decreased.

In conclusion, as the above examples show, although increasing the base-stock levels does not have a negative effect on most of the overall system-based performance measures, it can have a negative effect on individual order-based and item-based performance measures. Therefore, since customer satisfaction is the key to success in today's competitive market, managers must be aware of the fact that system-based performance measures are capable of hiding the negative effects of managers decisions on customers' satisfaction. The performance measures such as *FS* and *KS* that measure the satisfaction of the served customers are as important as the standard *SL* that evaluates the percentage of lost customers. These measures have a long-term effect on system's profit, and therefore must be considered in the decisions regarding management of assemble-to-order systems.

6.2. *Customer flexibility versus customer satisfaction*

It is a common practice for manufacturers or retailers to increase their sales by convincing customers to ignore some features (items) of their products (orders) which are not available at that time. This is not an easy task; however, most sales representatives are capable of doing it. For example, it is usually possible for a car dealer to convince some customers to ignores features such as leather seats, or CD players, if these features are not available at the time of the sale. Although this might help increase sales in the short-run, it may effect customer satisfaction, which is often seen as the key for the long-term profitability of the business.

In Scenario 1c, we show the effects of this approach on different levels of service in the system. We assume production capacity $(\mu_1, \mu_2, \mu_3) = (35, 80, 80)$, and base-stock levels $s_1 = 3$, $s_2 = 10$, and $s_3 = 10$ for items 1, 2, and 3, respectively.

Let $p^1_{1,0}$ be the fraction of times that a sales representative is successful in convincing customers of type 1 to ignore item 1. Scenario 1c in Table 2 shows how changes in $p^1_{1,0}$ effect the performance measures of the system at the item and order levels. Having $p^1_{1,0} = 1 (p^1_{1,0} = 0)$ means that the sale representative is always (never) successful in convincing customer type 1 to ignore item 1 and instead of $\Omega_1 = \{1, 2, 3\}$, accept $\Omega_4 = \{2, 3\}$.

As Table 2 depicts, the more successful the sales representative, the higher the decrease in both full and key satisfaction at the order and overall system levels. Furthermore, increasing $p^1_{1,0}$ from zero to one, also has a negative effect on service level, SL_i, for orders $i = 2$, 3, and 4. Thus, as the sales representative is more successful in convincing order type 1 customers to ignore item 1, depending on the product profit margin, long-term revenue across all orders may decrease, not to mention the significant increase in customer dissatisfaction. Therefore, managers of assemble-to-order systems with flexible customers must take into account the negative effects of increasing sales through their sales representatives whose only objectives are to maximize their own sales. One possible remedy to mitigate these negative effects is presented in the next scenario.

Scenario 1d shows how an increase in base-stock level of item 1 can limit the negative effects. As Table 3 depicts, increasing base-stock level of item 1 from three to five improves the system performance measures compared to those in Scenario 1c. Table 3 also implies that the measures of satisfaction such as *FS* and *KS* do not always change in the same direction. As you see in the table, changing $p^1_{1,0}$ and s_1 from zero and three, respectively, to one and five, increases *KS* from 70.51 to 74.27% but decreases *FS* from 70.47 to 68.45%. The reason is that the increase in the base-stock level has more (less) positive effect on *KS* (*FS*) while increase in switching has less (more) negative effect on *KS* (*FS*). This can be observed by comparing Scenarios 1a and 1c.

Managers must determine which of the different performance measures are relevant to the market in question. For example, in Scenario 1d, if the focus is on the immediate sale, the decisions would be to follow this policy (increase base-stock, increase switching) under these conditions. If customers long-term satisfaction and future sales depend on the customer getting the exact order originally sought, this policy may lead to an increase in customer dissatisfaction and it could actually harm future business, despite the short-term benefit of a sale today. Because the two measures of satisfaction do not always

Table 3. Results for Scenario 1d

Performance measures		*Scenario 1d*	
		$p^1_{1,0} = 0$ *and* $s_1 = 3$	$p^1_{1,0} = 1$ *and* $s_1 = 5$
System	*FS*	0.7047	0.6845
	KS	0.7051	0.7427
	SS	0.0000	0.0433
	SL	0.7051	0.7860
	DS	0.0006	0.1291

change in the same direction, it is important that the manager determine which performance measure is relevant to the situation at hand. A wrong choice of a performance measure may lead to a wrong decision by the manager when evaluating ways to improve the system.

6.3. *Base-stock levels versus profit*

As we showed in Section 6.1, although increasing the base-stock level of an item might improve the overall system performance, it can sometimes decrease some item- or order-based performance measures. This may lead to a decrease in profit since in most assemble-to-order systems the profit varies by order types. Scenario 2 presents an example in which an increase in base-stock level of an item actually decreases the total profit.

Consider a simple assemble-to-order system which consists of two order types; 1 and 2, and three items; 1, 2 and 3. Type 1 orders require the set of items $\Omega_1 = K_1 = \{1, 2\}$, and type 2 orders asks for items $\Omega_2 = K_2 = \{2, 3\}$. For simplicity, suppose holding costs for items 1, 2 and 3 are all $h = 1$ per item per unit time, and the profit (revenue minus all costs except holding cost) of satisfying orders 1 and 2 are three and nine per order, respectively. Thus, the total profit per unit time for the system can be obtained as follows:

$$\textit{Total profit per unit time} = 3\lambda_1 SL_1 + 9\lambda_2 SL_2 - h\sum_{i=1}^{3} E[I_i],$$

where $E[I_i]$ is the average inventory of item i. The order arrival rates, production capacities, base-stock levels and the total profit per unit time for the two different scenarios, 2a and 2b are presented in Table 4.

In Scenario 2b, s_1 is increased from six to twelve. As Table 4 shows, although SL_1 and the overall system performance (SL) are improved, the system's profit per unit time declines by 2.6%. The reason behind the decline in profit is that increasing the service level of item 1, leads to serving more type 1 customers. However, serving more type 1 customers causes additional stocks of component 2 to be used, potentially leading to a lower service level for type 2 customers. Since type 2 customers provide higher profits, this results in a lower total profit for the system.

The difference in profit of Scenarios 2a and 2b is not only due to the increase in holding cost in Scenario 2b. Even for $h = 0$, Scenario 2b still has a lower profit than Scenario 2a. Furthermore, this difference becomes larger as the difference in the profits of orders 1 and 2 increase. Under these circumstances, firms often apply yield management techniques to maximize their profit. The model presented in this paper can be used as an effective tool in order to determine the impact of different yield management strategies on service levels of the assemble-to-order systems with flexible customers.

Thus, before taking measures to improve service for some customers in an assemble-to-order systems, it is important to model the system and determine what will be the side effects of changes to the system.

6.4. *Customer flexibility versus profit*

One of the interesting questions regarding assemble-to-order systems with flexible customers is 'When is it beneficial to offer substitution and when should a manager not offer substitution?' One intuitive argument is that one should always offer substitution when a customer is willing to accept it. The logic is that it is better to always serve the current customer rather than attempt to save your resources for a more valuable customer that may or may not come, and that you may be able to serve that customer anyway. However, we will give examples of cases where this is not true.

In Scenario 3, we consider the simple example in Scenario 2 and assume that if item type 1 is not available, then based on a discount plan, all type 1 customers (the flexible customer) accept item 3 as a substitute for an unavailable item 1, and will leave the system with set of components $\Omega_1' = \Omega_2 = \{2, 3\}$. However, if item 1 is not available and item 3 is not offered as a substitute, the type 1 customer is lost. We consider the profit of the substitution satisfied customer to be six due to the discount offered to type 1 customers to induce them to switch. We compare the following two substitution strategies:

1. *Offering substitution:* According to this strategy, the manufacturer always offers item 3 as a substitute for an unavailable item 1. Since all customers accept the offer, then $p^1_{1,3} = 1$.
2. *No substitution:* According to this strategy, the manufacture never offers item 3 as a substitute for item 1 ($p^1_{1,3} = 0$). Thus, when item 1 is not available, any arriving customer of type 1 will be lost.

Table 4. Assemble-to-order system for Scenario 2

Scenario	*Items*			*Orders*		*Base-stock*			*Performance measures*						*Total profit (per unit time)*
									Item-based			*Order-based*		*System*	
	μ_1	μ_2	μ_3	λ_1	λ_2	s_1	s_2	s_3	F_1	F_2	F_3	SL_1	SL_2	SL	
2a	10	20	13	12	12	6	6	6	0.868	0.820	0.949	0.703	0.775	0.739	1000.9
2b	10	20	13	12	12	12	6	6	0.943	0.797	0.954	0.747	0.758	0.753	98.3

Table 5. The assemble-to-order system for Scenario 3

Scenario	μ			λ		*Base-stock*			*Profit*		
	1	*2*	*3*	*1*	*2*	*1*	*2*	*3*	*No substitution*	*Offering substitution*	*Percent difference*
3a	10	20	10	12	12	6	6	6	93.95	92.87	−1.2
3b	10	20	10	12	12	6	7	9	96.05	94.22	−1.9
3c	10	20	13	12	12	6	6	6	100.90	101.15	+0.2
3d	10	20	13	12	12	6	12	6	103.87	102.94	−0.9

Table 5 compares the total profit per unit time under the above two strategies. The total average profit per unit time is computed as follows:

$$\textit{Total profit per unit time} = 3\lambda_1 KS_1 + 9\lambda_2 KS_2 + 6\lambda_1 SS_1 - h\sum_{i=1}^{3} E[I_i].$$

Scenario 3a in Table 5 shows that offering substitution for item 1 leads to 1.2% less profit than the no substitution strategy. The reason is that every time a customer of type 1 receives a substitution, the inventory of item 3 decreases by one. This increases the chance of losing a customer of type 2 due to lack of item 3. Since customer 2 is a more profitable customer, this leads to the overall profit being lower.

In Scenario 3b we investigate whether or not changing the base-stock levels in Scenario 3a provides the system with the opportunity to take advantage of customers' (type 1) flexibility. More specifically, Scenario 3b represents Scenario 3a under the optimal base-stock levels which maximizes profit when substitution is offered. Applying the optimal base-stock levels in Scenario 3a results in 1.4% (= (94.22 − 92.87)/92.87) more profit if substitution is offered. However, even under those optimal base-stock levels, the system is better off not offering substitution. This is because the profit in Scenario 3b under substitution (94.22) is less than the profit when substitution is not offered (96.05).

In Scenario 3a, the production capacity of item 3 ($\mu_3 = 10$) is lower than the demand for item 3 ($\lambda_2 = 12$). Therefore, there is often not enough inventory of item 3 available to satisfy both the demands of type 2 customers and the demands of type 1 customers who request item 3 as a substitute for unavailable item 1. One can argue that when the capacity of item 3 is lower than the demand, no matter what inventory level is used, there is never enough item 3 so that substitution is never a good choice. This is true in Scenario 3a; however, it is not always true. For example, if in Scenario 3a, the substitution profit is 7.5 instead of six, then the profit under a 'no substitution' strategy will become 93.95, while the profit under a 'offering substitution' strategy will be 94.25. As these numbers show, although the capacity of item 3 is lower than the demand, substitution is a better choice than 'no substitution'.

In Scenario 3c we increase the production capacity of item 3 in Scenario 3a from $\mu_3 = 10$ to $\mu_3 = 13 (> \lambda_2)$. As the table shows, the original loss of profit caused by offering substitution changes to a gain of 0.2%. Unlike Scenario 3a, there is now an excess capacity for producing item 3. This means that the system is now capable of satisfying type 2 customers as well as providing enough of item 3 to substitute for unavailable item 1. In other words, the system has now capacity beyond what is necessary to serve its most profitable customers, and thus, this excess capacity can be used to also serve less profitable customers without loss of profit.

Scenarios 3b and 3c show that whether or not substitution can improve the overall profitability is dependent on the environmental parameters such as production capacity and demand arrival rates as well as the control parameters such as base-stock levels and profit. In practice, changing environmental parameters, such as production capacity or customer arrival rates, is often difficult or very costly or even infeasible. For example, increasing production capacity might require additional space or a large investment which might not be justifiable for some companies. On the other hand, control parameters such as base-stock levels, which are effective tools in managing the assemble-to-order systems, are easier to change.

6.4.1. *Random substitution strategies*

In Scenarios 3a to 3c, we have looked at a system which, due a very good discount plan, all customers accept substitution for item 1. We have compared the strategy that offers substitution to all customers with the strategy that never offers substitution. Our model allows us to also analyze strategies that offer substitution to only a portion of customers in a completely *random* fashion. In Scenario 3d we will show that these strategies can lead to greater profit than either offering substitution to everyone or not offering substitution at all. Suppose in Scenario 3d that all customers accept substitution for item 1; however, the system only offers substitution to $p^1_{1,3} \times 100\%$ of type 1 customers in a completely random manner. We call this strategy a *Random Substitution Strategy*. Random substitution strategies control the amount of substitution in the system in a random fashion.

Figure 2 shows the profit per unit time for different values of $0 \le p^1_{1,3} \le 1$ in Scenario 3d. Note that in Fig. 2,

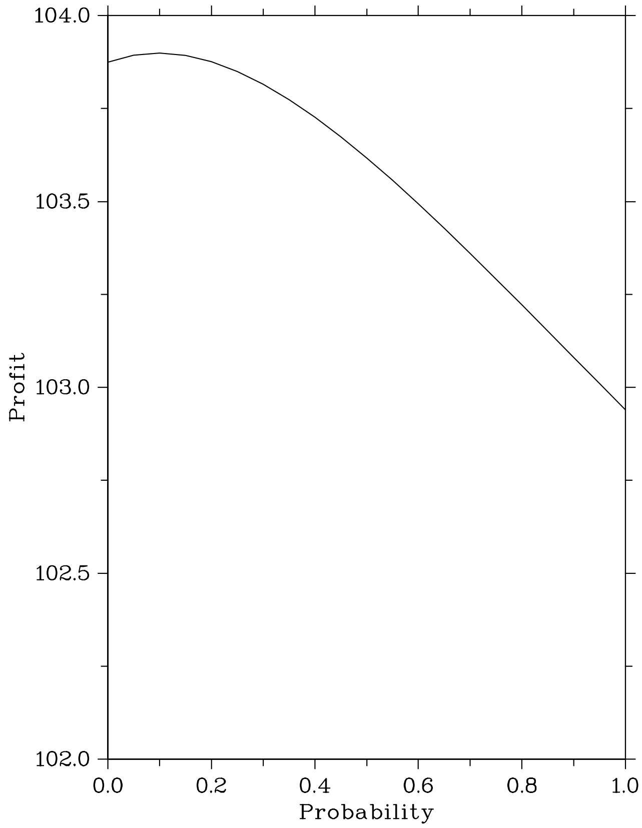

Fig. 2. Profit versus substitution probability $p^1_{1,3}$ for Scenario 3d.

$p^1_{1,3} = 1$ and $p^1_{1,3} = 0$ represent 'Offering substitution to all' and 'No substitution' strategies, respectively, with the profit per unit time of 102.94 and 103.87 (see Table 5). As Fig. 2 shows, the profit under $p^1_{1,3} = 0.1$ (the maximum profit) is higher than profit under $p^1_{1,3} = 1$ and $p^1_{1,3} = 0$. Therefore, if all customers accept substitution for item 1, the system will gain the highest profit (103.93) if it offers substitution to only 10% of its customers in a completely random fashion. For example, a retailer may know that 10% of the customers for order 1 are premium holders of the car company credit card. If these customers are randomly distributed among the arrivals of customers for order 1, then under Scenario 3d, the retailer gets the maximum profit if she offers the (discount) substitution plan only to its card holders.

Random substitution strategies do not use information regarding item inventory levels to decide whether or not to offer substitution to a customer. Thus, in systems where this information is not instantaneously available random substitution strategies can be used to increase the overall profit. Of course, strategies that use this information will perform better than random substitution strategies. However, these strategies have a more complex structure, especially in systems with large number of orders and items.

We conclude this section by emphasizing the necessity of modeling the effects of substitution, system capacity and base-stock levels before making any decision regarding substitution strategies. Since in real world assemble-to-order systems, the number of components and products is large, it is not easy to predict the effects of substitution decisions on the total profit. Thus, it is crucial to use models, such as the one presented in this paper, to examine the effects of item substitution on system performance.

6.5. *Benefit of modeling levels of customer satisfaction*

In this section we present Scenario 4 which emphasizes the benefits of using more realistic models to measure different levels of customer satisfaction in management of assemble-to-order systems. Scenario 4 looks at a case where a manager is faced with the decision regarding the base-stock level of a component. We will show that, if a manager does not model the customer behavior correctly, there is a potential for a loss in profit. More specifically, using a simple example, we show that making decisions for real systems with flexible and selective customers based on existing models which do not incorporate those customers behaviors might be very costly.

Consider a system with four components and four customer types. The order arrival rates are $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (8, 6, 6, 8)$, and the item production capacities are $(\mu_1, \mu_2, \mu_3, \mu_4) = (7, 23, 25, 11)$. Suppose the base-stock level for items 2, 3 and 4 are set to be $s_2 = s_3 = s_4 = 3$, and the manager has to choose the base-stock level for item 1. Table 6 depicts the key and non-key items for each order as well as the profit under both full and key satisfaction. Profit under full satisfaction refers to the case that a customer gets all his key and non-key items, while profit under key satisfaction refers to cases where the customer only gets his key items (since his non-key items are not available). We assume that all customers are

Table 6. The assemble-to-order system for Scenario 4

Customer	*Arrival rate* (λ_i)	*Key items*	*Non-key items*	*Full satisfaction profit*	*Key satisfaction profit*
1	8	$\{1, 2, 3\}$	–	4	–
2	6	$\{4\}$	$\{2\}$	4	3.5
3	6	$\{4\}$	$\{3\}$	4	3.5
4	8	$\{3, 4\}$	$\{2\}$	4.5	4

selective customers, and therefore they are lost if their key items are not available.

We obtain the optimal base-stock level for item 1 using two different models:

- *Model I*: Model I incorporates the behavior of customers regarding their key and non-key items using the model developed in Section 4.3, and therefore gives the correct optimal base-stock level for item 1. The total profit for Model I is calculated based on the profit of key and fully satisfied customers presented in Table 6.
- *Model II*: Model II is based on the Total Order Service (TOS) model presented in Song *et al.* (1999). Model II does not take into account the behavior of customers towards their key and non-key items.

The objective is to compare the optimal base-stock levels obtained by two models and see if modeling customer behavior has any advantages.

Since Model II does not distinguish between key and non-key items, it actually sees the following customer types with corresponding arrival rates:

Customer type $1'$ in Model II is the same as customer type 1 in Model I; however, customer types $2'$ to $5'$ in Model II are different from customer types 2 to 4 in Model I. For example, customer type $3'$ in Model II includes those customers of type 3 in Model I who get all their key and non-key items as well as those customers of type 4 in Model I who only get their key items (when item 2 is not available). Thus, the arrival rates for customers of type $2'$ to 5, $\lambda_{i'}$, will be different from the arrival rates λ_i, and will depend on the availability of non-key items of different orders.

In order to obtain the optimal base-stock level of item 1 using Model II, we first need to obtain the arrival rates $\lambda_{i'}$. These arrival rates depend on the availability of non-key items 2 and 3, F_2 and F_3. On the other hand, F_2 and F_3 depend on the arrival rates $\lambda_{i'}$. One way to overcome this loop is to use the following iterative approach. For the base-stock level $(s_1 = s, s_2 = 3, s_3 = 3, s_4 = 3)$:

Step 0. Use the TOS model and obtain the availability of items 2 and 3 ($F_2^{(0)}$ and $F_3^{(0)}$) for the system presented in Table 6 (assume that all items are key items). Set $k = 0$, and go to Step 1.

Step 1. Use $F_2^{(k)}$ and $F_3^{(k)}$ as estimates for F_2 and F_3, respectively, and obtain the arrival rates $\lambda_{1'}^{(k)}$, $\lambda_{2'}^{(k)}, \ldots, \lambda_{5'}^{(k)}$ according to Table 7. Set $k \leftarrow k+1$, and go to Step 2.

Step 2. Use the arrival rates $\lambda_{1'}^{(k-1)}, \lambda_{2'}^{(k-1)}, \ldots, \lambda_{5'}^{(k-1)}$ in the TOS model, and compute $F_2^{(k)}$ and $F_3^{(k)}$, the new estimates for availabilities of items 2 and 3. If $k = 1$, return to Step 1. Otherwise, go to Step 3.

Step 3. Compute $\Delta = \max_r\{\Delta_r | \Delta_r = |\lambda_r^{(k)} - \lambda_r^{(k-1)}|, r = 1', 2', \ldots, 5'\}$. If $\Delta \le \epsilon = 0.005$, go to Step 4. Otherwise, return to Step 1.

Step 4. For the base-stock levels $(s_1 = s,\ s_2 = 3,\ s_3 = 3,\ s_4 = 3)$, we will have $\lambda_r = \lambda_r^{(k)};\ r = 1', 2', \ldots, 5'$.

Table 7. The customer types seen by Model II

Customer type	*Arrival rate* (λ_i')	*Items*	*Profit*
$1'$	$\lambda_{1'} = \lambda_1$	$\{1,2,3\}$	4
$2'$	$\lambda_{2'} = F_2\lambda_2$	$\{4,2\}$	4
$3'$	$\lambda_{3'} = F_3\lambda_3 + (1-F_2)\lambda_4$	$\{4,3\}$	4
$4'$	$\lambda_{4'} = F_2\lambda_4$	$\{3,4,2\}$	4.5
$5'$	$\lambda_{5'} = (1-F_2)\lambda_2 + (1-F_3)\lambda_3$	$\{4\}$	3.5

To find the optimal base-stock level for item 1 using Model II, we implemented the above iterative approach for different values of s_1. Then, for each value of s_1, we calculated the total profit and picked the one with the maximum profit. This resulted in the optimal base-stock level $s_1^* = 2$ (see Table 8).

On the other hand, using Model I, the total profit under different base-stock levels for item 1 (given $s_i = 3,\ i = 2, 3, 4$) are shown in Table 8. As Table 8 shows, Model I results in an optimal base-stock level of $s_1^* = 3$ while Model II suggests $s_1^* = 2$. If customer behavior follows Model I but the manager does not incorporate this behavior in her model (i.e., uses Model II), she will decide to set s_1 to two which results in a total profit of 72.84 per unit time. However, if the manager uses Model I, she will set s_1 to three which results in a total profit

Table 8. Profit obtained by Model I and Model II

Scenario	μ				λ				*Base-stock*				*Profit (per unit time)*	
	1	*2*	*3*	*4*	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*	*Model II*	*Model I*
1	7	23	25	11	8	6	6	8	1	3	3	3	70.11	70.88
2	7	23	25	11	8	6	6	8	2	3	3	3	**71.29**	72.84
3	7	23	25	11	8	6	6	8	3	3	3	3	71.11	**73.24**
4	7	23	25	11	8	6	6	8	4	3	3	3	70.69	73.12
5	7	23	25	11	8	6	6	8	5	3	3	3	69.69	72.76

73.24 per unit time, which is 0.54% more than 72.84. Thus, in this example, 0.54% can be interpreted as the percentage of the total profit lost per unit time for not directly incorporating customer behavior in the model used to set the base-stock level of item 1.

Note that although 0.54% might seem insignificant, since it is the percent of the profit per unit time, it actually is a large number. For example, Dell's Gross Profit (Anon, 2001) in the fiscal year ending February 2, 2001 was $6443 million. The 0.54% increase in gross profit represented by the above example would translate to a $35 million in annual savings.

7. Conclusion

Assemble-to-order systems have recently become very popular, since they allow manufacturers to achieve a high degree of product variety and quick product delivery while maintaining low inventory. To create a realistic model of assemble-to-order systems our analysis incorporates customer preferences and flexibility. The paper also introduces various measures to evaluate different levels of customer satisfaction, which is important in today's competitive markets. Finally, numerical examples illustrate interesting and unexpected behaviors of these systems and provide useful insights for managers of these systems.

These insights include:

- Although increasing the base-stock levels for some items increases the system overall performances, it does not guarantee an increase in customer satisfaction or profit. In fact, customer dissatisfaction is the overall system performance measure that can become worse as base-stock levels increase.
- Taking advantage of customer flexibility does not necessarily imply an increase in profit, especially in cases where there is not enough production capacity to serve the demand of the flexible customers.
- Various measures of customer satisfaction may respond in different ways to changes in system parameters. Therefore, it is important to identify the specific performance measure(s) which most accurately models customer behavior for the market in question.
- Ignoring customer preferences and flexibility in the decision-making regarding a system's parameters (e.g., setting base-stock levels) can have a significant negative effect on a system's profit.

Acknowledgements

Research supported in part by ONR Contracts N00014-90-J-1649 and N00014-95-1-0232, NSF Contracts DDM-9322828 and DMI-9732795 and a grant from General Motors.

References

Agrawal, N. and Cohen, M. (1995) Optimal material control and performance evaluation in assembly environment with component commonality. Working paper, Decision and Information Sciences Department, Santa Clara University, Santa Clara, CA.

Anon (2001) Dell Computer Corporation: Form 10-K for fiscal year ending February 2, 2001, SEC File number 0-17617.

Cheung, K.L. and Hausman, W. (1995) Multiple failures in a multi-item spare inventory model. *IIE Transactions*, **27**, 171–180.

Gerchak, Y. and Henig, M. (1989) Component commonality in assemble-to-order systems: models and properties. *Naval Research Logistics*, **3**, 61–68.

Glasserman, P. and Wang, Y. (1998) Leadtime-inventory trade offs in assemble-to-order systems. *Operations Research*, **46**, 858–871.

Hausman, W., Lee, H. and Zhang, A. (1998) Joint demand fulfillment probability in a multi-item inventory system with independent order-up-to policies. *European Journal of Operation Research*, **109**, 646–659.

Hyde, J. (2000) GM plans make-to-order vehicles. Associated Press wire story, June 22, 2000.

Schraner, E. (1996) Capacity/inventory trade-off in assemble-to-order systems. Research report, IBM Watson Research Center, Yorktown Heights, NY.

Song, J. (1998) On the order fill rate in a multi-item, base-stock inventory system. *Operations Research*, **46**, 831–845.

Song, J. (2000) A note on assemble-to-order systems with batch ordering. *Management Science*, **46**, 739–743.

Song, J., Xu, S. and Liu, B. (1999) Order-fulfillment performance measure in an assemble-to-order system with stochastic leadtimes. *Operations Research*, **47**, 131–149.

Srinivasan, R., Jayaraman, R., Roundy, R. and Tayur, S. (1992) Procurement of common components in a stochastic environment. Research report, IBM Watson Research Center, Yorktown Heights, NY.

Wemmerlov, U. (1984) Assemble-to-order manufacturing, implication for material planning. *Journal of Operations Management*, **4**, 347–368.

Zhang, A.X. (1997) Demand fulfillment rates in an assemble-to-order system with multiple products and dependent demands. *Production and Operations Management*, **6**, 309–324.

Zhang, R.Q. (1999) Expected time delay in multi-item inventory systems with correlated demands. *Naval Research Logistics*, **46**, 671–688.

Biographies

Seyed Iravani is an Assistant Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. His publications have appeared in *Queueing Systems* and *Operations Research* . He is on the Editorial Board of *IIE Transactions* and he is also an Associate Editor for *Management Science*. His research interests are in the stochastic modeling of manufacturing systems and the analysis and control of queueing systems.

Louis Luangkesorn is a Ph.D. candidate in the Department of Industrial Engineering and Management Sciences at Northwestern University. He holds an M.A. in Science, Technology and Public Policy from The George Washington University and a B.S. in General Engineering

(Management Science) and a B.A. in Political Science from the University of Illinois at Urbana-Champaign. His research interests are in supply chain management and logistics.

David Simchi-Levi is a Professor of Engineering Systems at Massachusetts Institute of Technology. His research currently focuses on developing and implementing robust and efficient techniques for logistics systems. He has published widely in professional journals on both practical and theoretical aspects of logistics and supply chain management. He is the founder and chairman of LogicTools (*http://www.logic-tools.com/*), a software company focusing on developing Decision Support Systems for solving logistics and supply chain management problems. These systems have been used widely to reduce cost and improve service level in large-scale logistics systems. He is the co-author (with Julien Bramel) of *The Logic of Logistics: Theory, Algorithms and Applications for Logistics Management*, published by Springer in 1997. His new book, *Designing and Managing the Supply Chain*: *Concepts*, *Strategies and Case Studies*, written with P. Kaminsky and E. Simchi-Levi was published by Irwin/McGraw-Hill in August 1999. The book received the Book-of-the-Year award and the Outstanding IIE Publication award given in 2000 by the Institute of Industrial Engineers. The book also received the Outstanding First Edition of the Year award given in 2000 by McGraw-Hill. The book was selected by Business 2.0, December 2001 issue, as the best source for slashing time and cost and increasing productivity in the supply chain. The book has been translated to Chinese, Japanese and Korean.

Contributed by the Inventory Department