ORDER-FULFILLMENT PERFORMANCE MEASURES IN AN ASSEMBLE-TO-ORDER SYSTEM WITH STOCHASTIC LEADTIMES

JING-SHENG SONG
University of California, Irvine, California

SUSAN H. XU
Penn State University, University Park, Pennsylvania

BIN LIU
Chinese Academy of Sciences, Beijing, China
(Received February 1997; revised April 1997; accepted May 1997)

We study a multicomponent, multiproduct production and inventory system in which individual components are made to stock but final products are assembled to customer orders. Each component is produced by an independent production facility with finite capacity, and the component inventory is controlled by an independent base-stock policy. For any given base-stock policy, we derive the key performance measures, including the probability of fulfilling a customer order within any specified time window. Computational procedures and numerical examples are also presented. A similar approach applies to the generic multi-item make-to-stock inventory systems in which a typical customer order consists of a kit of items.

This paper concerns the evaluation and analysis of order fulfillment performance measures for a multi-item, assemble-to-order production/inventory system with stochastic leadtimes.

Order fulfillment performance has become increasingly important as companies that must adapt quickly to market and technology changes move toward "assemble-to-order," as opposed to the traditional "make-to-stock," inventory planning systems. In an assemble-to-order system, products are designed around interchangeable modules, and the company makes and stocks only the modules and other major components. When a customer order arrives requesting a specific kit of modules and components, the company quickly assembles these items and delivers the end product to the customer. Since each customer order requires the simultaneous availability of several items, the questions inventory managers ask most frequently are the following: For any given safety-stock level of each item, what is the probability a demand can be satisfied immediately (order fill rate)? What is the probability a customer order can be met within a time window (customer waiting time distribution; also termed order response time reliability in industry)?

The same issue exists for many multi-item distribution systems as well, including the emerging mail-order businesses (see, for example, Cohen and Lee 1990, Lee and Billington 1992, Song 1998). In particular, a typical customer order to a manufacturer consists of a combination of several finished goods. Reliable and speedy delivery of orders is one of the most crucial factors for customer satisfaction. So, the order-based performance measures, such as the order fill rate and the customer waiting time distribution, are the important ones.

Standard inventory models assume that demands are independent across items, which is a valid assumption for some situations in the make-to-stock environment but not for assemble-to-order systems, in which one must jointly manage inventories and production capacities across various items. Evidently, new models and methods are in demand to address important issues in assemble-to-order systems.

The purpose of this paper is to conduct an exact analysis on a wide range of performance measures in the assemble-to-order systems with sequential, capacitated stochastic production processes. In particular, we model the demand process as a multivariate Poisson process. That is, the overall demand arrives according to a Poisson process, but there is a fixed probability that a demand requests a particular kit of different items. Each item's inventory is controlled by a separate base-stock policy. Demands are filled on an FCFS basis. Demands for an item that cannot be filled immediately are queued in a backlog queue with a certain capacity, whose value may range from zero to infinity. Infinite capacity corresponds to the complete backlog case, whereas zero capacity corresponds to the lost-sale case. The supply system of each item is an independent, single-machine production facility with i.i.d. exponentially distributed processing times.

For any given base-stock policy and backlog queue limits, we present a procedure to evaluate the item-based, order-based, and system-based performance measures, such as fill rate, service level (the probability that an order

Subject classifications: Inventory/production: multi-item, operating characteristics, stochastic. Probability: applications, Markov processes, stochastic model applications.

Area of review: Manufacturing Operations.
will be backlogged and eventually served), waiting time distribution, etc. In terms of computational complexity, this procedure is very efficient to obtain the exact performance of small to medium-sized systems with finite backlog queues. Under certain conditions (e.g., moderate machine utilization), this procedure is also capable of efficiently conducting an approximate analysis on the system with infinite backlog queues. Additionally, the advantage of conducting exact analysis is to enhance our understanding of the inner working of such systems, thereby providing useful insights for effective system design and control. It also provides a benchmark result and insights for developing stochastic bounds, heuristic solutions for systems that are more complex in nature and thus less tractable analytically.

In our analysis, it is essential to obtain the joint equilibrium distribution of the occupancy in the supply system, which subsequently determines other order-based performance measures. The major difficulty in evaluating the joint distribution of the occupancy is a result of correlation of the production facilities, caused by simultaneous arrivals. See, for example, Flatto and Hahn (1984), Shwartz and Weiss (1993), and Wright (1992) for transformation and asymptotic results for the 2-queue system with infinite buffers. The key idea in our solution procedure is that, because of the special structure of the demand process, the occupancy in the supply system can be modeled as a quasi birth-death process. This allows us to develop a matrix-geometric solution for its joint distribution. Another contribution of this paper that distinguishes it from others in the literature is that we are able to derive the exact waiting time distribution for accepted orders. Such information is increasingly important in the assemble-to-order environment. To be competitive, companies often need to guarantee delivery of the product to customers in a specified time window. The waiting time distribution provides the managers the precise likelihood such guarantees can be met (This is termed order response time reliability in Hausman et al. 1998). Since sometimes the performance measure of primary interest is the expected waiting time of an accepted order (as opposed to the distributional probability), we also develop a simpler recursive evaluation procedure for such purpose.

There have been several recent research efforts in studying multi-item inventory systems with correlated demands across items. See, for example, Agrawal and Cohen (1996), Hausman et al., Schraner (1996), Song (1998), and Srinivasan et al. (1992). These studies assume a base-stock policy and a deterministic supply system. For stochastic supply systems, Anupindi and Tayur (1998) develop a simulation model to study both item-based and order-based performance measures in a multiproduct cyclic production system. Cheung and Hausman (1995) consider i.i.d. replenishment leadtimes and a multivariate Poisson demand model. Complete cannibalization is assumed, however, in order to derive the average customer backorders. In addition, two studies have been conducted independently and concurrently with our research, but with different focuses. Zhang (1995) assumes a multivariate Poisson demand process, and each item is supplied by a dedicated facility with general i.i.d. processing times. She uses the expected system-based waiting time (i.e., the expected delay of an arbitrary order) as the sole performance measure. (Our model includes the complete backlog system as a special case, and our primary interest is to obtain the distributions of order-fulfillment performance measures. As a byproduct, we also obtain the distribution and the expected value of the delay of an arbitrary order.) Glasserman and Wang (1998) study a model with the same kind of supply system as in Zhang but a more general demand structure. They study the leadtime-inventory trade-off and show that the relation is linear, in a limiting sense, at high level of service.

In Section 1, we describe the model in detail and introduce the basic notation. In Section 2, we focus our analysis on the assemble-to-order system with total order service (TOS), where an order must be either accepted or rejected as a whole. It is worth mentioning that, in addition to its applicability to some real systems, the fixed backlog buffer sizes can be viewed as a measure of customer impatience. Here, if even one item’s backlog queue is full, it signals the customer that there exists a prospect of a long wait, the customer then decides to leave, without waiting for any items. (Li (1992) makes a similar observation in a single-item system.)

In Section 3, we concentrate on the assemble-to-order environment with partial order service (POS), where a customer request may be only partially accepted (guaranteed to be filled eventually), with the rest rejected because the corresponding backlog queues are full. This model is mostly valid in multi-item, make-to-stock systems in which a typical customer order consists of a kit of items in stock. One can also use this model to measure customer impatience, but here his/her impatience is associated with individual items as opposed to the entire order.

Evidently, the TOS and POS models are identical in the complete backlog case. This raises the possibility of using a POS model to approximate its TOS counterpart. An attractive feature of the POS model, as compared with the TOS model, is that its item-based performance measures can be readily obtained—only through the marginal distribution of the occupancy in each individual production facility. Thus, the item-based performance in the POS model can be conveniently employed to approximate and bound the order-based performances in both POS and TOS models. An interesting question is, of course, when and to what extent the former can provide useful information for the latter (we address this issue further in Section 4.2).

To gain more insight into the model, Section 4 is devoted to the simple two-item system. In particular, we provide details of the general results presented in Sections 2 and 3. Numerical experiments are conducted to illustrate the results and to derive insights. For example, we find that, in general, the result from the POS model provides a
reliable estimate for its counterpart in the TOS model. We also find that with moderate traffic the finite-buffer model provides an accurate approximation for the infinite-buffer model in a broad range of environmental settings. Finally, in Section 5, we make some concluding remarks and discuss future research directions.

1. THE MODEL DESCRIPTION

Now we describe the specific model assumptions and introduce the basic notation. We consider an inventory system of J different items. Let \( \Omega = \{1, 2, \ldots, J\} \) be the set of all item indexes. For any subset \( K \) of \( \Omega \), denote by \( |K| \) the number of elements in \( K \). We consider an infinite planning horizon and assume that the assembly times are negligible (compared to the production times).

The Demand Process

The overall demand process is stationary in time and forms a Poisson process. Each customer requests at most one unit of each item but may require several items simultaneously. In particular, for any subset of items \( K \subseteq \Omega \), we say a demand is of type \( K \) if it requires one and only one unit of each item in \( K \) and 0 units in \( \Omega - K \). We assume that there is a fixed probability that a demand is of type \( K \). Each demand’s type is independent of the other demands’ types and of all other events. When \( K \) contains a single item, say \( i \), we abbreviate type \( K \) by type \( i \). Similarly, we say an order is of type \( ij \) if \( K = \{i, j\} \). Obviously, the demand process for each item is also a Poisson process.

Throughout the paper, we use subscripts to indicate item type and superscripts to denote order type. For any \( \lambda \) in \( \Omega \) and \( K \subseteq \Omega \), let

\[
\lambda^K = \text{overall demand rate},
\]

\[
q^K = \text{probability a demand is of type } K \quad (\sum_{K \subseteq \Omega} q^K = 1),
\]

\[
\lambda^K = \text{demand rate of demand type } K = q^K \lambda,
\]

\[
q_i = \text{probability a demand requires item } i = \sum_{K \subseteq \Omega, i \in K} q^K,
\]

\[
\lambda_i = \text{aggregate demand rate of item } i = \sum_{K \subseteq \Omega, i \in K} \lambda^K = q_i \lambda.
\]

Shipment and Backlogging

Demands are filled on a First-Come-First-Serve (FCFS) basis. When an order arrives and we have some, but not all, of its items in stock we will either ship the in-stock items if partial shipment is allowed or put aside these items as “committed” inventory if partial shipment is not allowed. However, a customer request is considered backlogged unless it can be satisfied completely.

A demand for item \( i \) that cannot be filled immediately is queued in the backlog queue \( i \), which has a capacity \( b_i \geq 0 \), and will be shipped out (or put aside) as soon as a unit of item \( i \) becomes available to fill it. (When \( b_i = \infty \) for all \( i \), all unfilled demands are backlogged. When \( b_i = 0 \) for all \( i \), unfilled demands are lost.) We consider two kinds of lost sales when an incoming order that requests more than one item finds the backlog queue for at least one of its items is full. Suppose \( |K| > 1 \).

a. Total order service (TOS): If a type \( K \) order sees at least one of its item’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 at the manufacturing level and also for some multi-item make-to-stock systems.

b. 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 most assemble-to-order environments at the distribution level, 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 two models are identical in the complete backlogging case.

Replenishment Policy

We assume that there is no economy of scale in replenishment. Each item is controlled by an independent base-stock policy. Let

\[
s_i = \text{the base-stock level for item } i.
\]

That is, at each demand epoch, if the inventory position (i.e., the inventory on hand plus inventory on order minus backorders) of item \( i \) is less than \( s_i \), then order up to \( s_i \). Since there is no economy of scale in replenishment, a base-stock policy would be optimal for each item if we were to manage the system on an item basis and complete backlogging was assumed. Because of its simplicity, we employ this type of policy as a reasonable heuristic.

Without loss of generality, we assume that at time 0 item \( i \) is stocked at level \( s_i \) for all \( i \). Then, each demand for item \( i \) triggers an order for that item until a demand finds the backlog queue is full. Hence, there can be, at most, \( s_i + b_i \) outstanding orders of item \( i \).

The Supply System

Replenishment orders for item \( i \) are sent to a single-machine production facility, say facility \( i \), in which they are processed on a FCFS basis. The processing times at facility \( i \) are i.i.d. exponentially distributed random variables with rate \( \mu_i \), \( i = 1, 2, \ldots, J \). Thus our supply system can be viewed as \( J \) parallel stochastic production facilities, where facility \( i \) accepts Poisson inputs with rate \( \lambda_i \) and contains at most \( s_i + b_i \) outstanding orders at any time.

Performance Measures

We are interested in three levels of performance measures: item-based, order-based, and system-based performance measures.

A. Performance Measures of Item \( i \), \( i \in \Omega \). We first define the item-based performance measures that are common in

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
the standard single-item inventory models. Denote the follow-
lowing quantities in steady-state:
\[ I_i = \text{inventory on hand of item } i, \quad 0 < I_i \leq s_i, \]
\[ B_i = \text{backorders of item } i, \quad 0 \leq B_i \leq b_i, \]
\[ IO_i = \text{inventory on order of item } i, \quad 0 \leq IO_i \leq s_i + b_i. \]

We shall derive the expectations of the above measures. The fill rate of item \( i \) is denoted by
\[ F_i = \text{probability of immediately satisfying a demand for item } i = P(I_i > 0). \]

In contrast to the fill rate, which measures the immediate availability, the service level of an item measures the serviceability of the item:
\[ SL_i = \frac{\text{fraction of item } i \text{ demands satisfied}}{P(B_i < b_i)}. \]

Evidently, \( F_i \leq SL_i \). When \( b_i = 0 \) (the lost sale case), the two measures are identical. When \( b_i = \infty \) (the complete backlogging case), \( SL_i = 1 \) as no demand of item \( i \) is ever lost in this case.

Now let
\[ W_i = \text{waiting time of an accepted request for item } i. \]

We are interested in deriving the distribution of \( W_i \), which allows one to find the probability of filling an item in a specified time window.

**B. Performance Measures of a Type-K Order, \( K \subseteq \Omega \).** The fill rate, service level and waiting time of a type-K order are defined, respectively, by
\[ F^K = \text{joint probability that all items in a type-K order are filled immediately} \]
\[ = P(I_i > 0 : i \in K). \]
\[ SL^K = \text{joint probability that all items in a type-K order are accepted} \]
\[ = P(B_i < b_i : i \in K). \]
\[ W^K = \text{waiting time to fill all items in a type-K order} \]
\[ = \max_{i \in K} W_i. \]

Because of simultaneous demands, the components in the random vector \( \{I_i, i \in K\} \cup \{B_i, i \in K\} \cup \{W_i, i \in K\} \) are dependent. Consequently, one cannot obtain the performances of a type-K order from its item-based counterparts. Indeed, this presents the major challenge of our analysis.

**C. Performance Measures of the System.** Let \( F, SL \) and \( W \) be the fill rate, service level, and waiting time of an arbitrary filled demand (regardless of the type). Then
\[ F = \sum_{K \subseteq \Omega} q^K F^K, \tag{1} \]
\[ SL = \sum_{K \subseteq \Omega} q^K SL^K, \tag{2} \]
\[ P(W \leq x) = \sum_{K \subseteq \Omega} q^K P(W^K \leq x). \tag{3} \]

Because system-based performance measures are readily computable once performance measures of all type-K orders are obtained, in the rest of the paper we are mainly concerned with the item-based and order-based performance measures.

**2. THE TOTAL ORDER SERVICE MODEL**

In this section, we focus on assembly systems. In particular, we assume that an order is either serviced entirely or rejected entirely (total order service). A type-K demand is lost if, upon its arrival, at least one backlog queue \( i \) is full, \( i \in K \). It is worth noting that this feature causes the item-based performance measures to depend on the order-based performance measures.

Following the standard argument in inventory models, it is easy to see that the steady-state net inventory of item \( i \) is given by \( s_i - IO_i \). Thus, the type-K performance measures are determined by the joint distribution of \( (IO_1, IO_2, \ldots, IO_K) \). In this section, we first develop a procedure to compute the joint distribution of \( (IO_1, IO_2, \ldots, IO_K) \). Then, using this distribution, we derive the order-based performance measures for each fixed \( K \subseteq \Omega \), which in turn is employed to obtain the item-based performance measures.

**2.1. The Joint Stationary Distribution of Outstanding Orders**

Observe that \( IO = \{(IO_1(t), IO_2(t), \ldots, IO_K(t)), t \geq 0\} \), where \( IO_i(t) \) is the inventory on order of item \( i \) at time \( t \), is a continuous-time Markov chain with finite state space \( \{n = (n_1, n_2, \ldots, n_K) | 0 \leq n_i \leq N_i, 1 \leq i \leq J \} \) where \( N_i = s_i + b_i \). A state transition can only occur if a demand arrives or an item of an order is released. Specifically, with transition rate \( \lambda^K, \) state \( n \) enters the state \( n' = (n'_1, \ldots, n'_J) \), where
\[ n'_j = \begin{cases} n_j + 1, & \text{if } j \in K \text{ and } n_j < N_j \text{ for all } i \in K, \\ n_j, & \text{otherwise}. \end{cases} \tag{4} \]
and with transition rate \( \mu^K, \) state \( n \) enters the state \( n'' = (n''_1, \ldots, n''_J) \), where
\[ n''_j = \begin{cases} n_j - 1, & \text{if } j = i \text{ and } n_j > 0, \\ n_j, & \text{otherwise}. \end{cases} \tag{5} \]

It is easy to see that this Markov chain is irreducible, so its stationary distribution exists uniquely. Using (4) and (5), one can establish the balance equation for each state (there are \( N_j + 1 \) of them). The stationary joint distribution is then the solution of these balance equations plus a normalization equation, and conventional computational methods for solving linear equations can be employed. However, we find that the special structure of the process allows us to obtain a unified, matrix-geometric solution (defined in Neuts 1981) of the stationary distribution of \( IO \), as explained below.

Let us order the state space of \( IO \) in the lexicographic order \( (I_1, I_2, \ldots, I_K) \), where \( I_k \) is the collection of the states when the inventory on order of item \( k \) is \( k \). That is, \( I_k \) is the \( N_j + 1 \) dimensional vector.
Under the ordering \((I_0, I_1, \ldots, I_{N_1})\), \(I_0\) can be viewed as a two-dimensional Markov chain, with the sizes of the first and second dimensions being \(N_1 + 1\) and \(\Pi'_{j=2} (N_j + 1)\), respectively. Let \(p_i\) be the stationary probabilities of those states in \(I_0\). Then one can conveniently represent the stationary distribution of \(I_0\) by

\[
p = (p_0, p_1, \ldots, p_{N_1}).
\]

Since \(IO_i\) in a single transition can increase (decrease) by at most 1, \(I_0\) can be regarded as a quasi birth-and-death (QBD) process (Neuts 1981, Chapter 3). It is straightforward to show that the infinitesimal generator of the chain \(I_0\) is a block-tridiagonal matrix

\[
\tilde{Q} = \begin{pmatrix}
A & A_0 & 0 & 0 & \cdots & 0 & 0 \\
\mu_1 & A_0 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & \mu_1 & A_0 & \cdots & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & \cdots & \mu_1 & A_0 & \cdots \\
0 & 0 & \cdots & \cdots & 0 & \mu_1 & A_1 - \mu_1 I
\end{pmatrix}
\]

where \(A, A_0\) and \(A_1\) are square matrices of order \(\Pi'_{j=2} (N_j + 1)\). (The matrices \(A\) and \(A_0\) can be constructed according to (4) and (5).) Generally speaking, they are quite sparse, especially if there are only a few identifiable demand types, i.e., \(q^h > 0\) for only a few \(Ks\). To avoid heavy notation, we omit the details of these two matrices. We will provide details for the 2-item system in Section 4.)

Therefore, we can express the steady-state balance equations in the matrix form:

\[
p_0 A + \mu_1 p_1 = 0,
\]

\[
p_k A_0 + p_k (A - \mu_1 I) + \mu_1 p_{k-1} = 0, \quad 1 \leq k < N_1,
\]

\[
p_{N_1-1} A_0 + p_{N_1} (A_1 - \mu_1 I) = 0.
\]

The above balance equations then lead to:

\[
p_{N_1} = -p_{N_1-1} A_0 (A_1 - \mu_1 I)^{-1} = p_{N_1-1} R_{N_1},
\]

\[
p_k = p_{k-1} R_k, \quad k = 1, 2, \ldots, N_1 - 1,
\]

where \(R_k\) can be found recursively using

\[
R_{N_1} = -A_0 (A_1 - \mu_1 I)^{-1},
\]

\[
R_k = -A_0 [(A - \mu_1 I) + \mu_1 R_{k+1}]^{-1}, \quad k = N_1 - 1, \ldots, 1;
\]

\[
R_0 = I.
\]

Now, \(p_0\) satisfies

\[
p_0 (A + \mu_1 R_1) = 0,
\]

and the normalization equation

\[
\sum_{i=0}^{N_1} p_i e = p_0 \sum_{i=0}^{N_1} \tilde{R}_i e = 1,
\]

where \(e\) is column vector of ones, and

\[
\tilde{R}_k = \prod_{i=0}^{k} R_i, \quad k = 0, 1, \ldots, N_1.
\]

**Remark 2.1.** In terms of computational complexity, the direct approach solves \(\Pi'_{j=1} (N_j + 1)\) linear equations, which by using Gaussian elimination requires

\[
(\Pi'_{j=2} (N_j + 1))^{3/2} + (\Pi'_{j=2} (N_j + 1))^{2}
\]

operations (multiplications; we ignore additions). See Issacson and Keller (1966). The matrix-geometric solution approach, on the other hand, involves \(N_1 + 1\) inversions and \(2N_1\) multiplications of matrices of order \(\Pi'_{j=2} (N_j + 1)\), which results in

\[
(3N_1 + 1)\left(\prod'_{j=2} (N_j + 1)\right)^{3/2} + (\prod'_{j=2} (N_j + 1))^{2}
\]

operations. Thus, the matrix-geometric approach is faster than the direct approach by a factor of approximately \((N_1 + 1)^{3/2}/9\). In fact, since any \(IO_i\) can serve as the first dimension of the QBD process, we can renumber the item indices if necessary so that \(N_1 = \max_i N_i\) and, consequently, accelerate the computation by a factor of approximately \((\max_i N_i + 1)^{3/2}/9\).

In addition, as a QBD process is solvable as long as the second dimension of the process is finite, our approach can be applied to obtain the exact analysis for the system with one of the item completely backlogged. This would correspond to the practical situation in which a company wishes to capture 100% demands of one of its key products. This approach is also capable of providing an approximate solution to the system with more than one item completely backlogged (see Section 4.2 for a discussion on buffer size truncation and for numerical results). In most practical situations (e.g., with moderate traffic intensity), one can replace an infinite backlog queue buffer with a finite one of rather moderate size without sacrificing computational accuracy.

For large-scale systems, we propose the following:

(i) Although the total number of potential demand types can be large, by the Pareto phenomenon, often a large portion of the total dollar volume of sales is accounted for by a small number of demand types. When this is true, one can concentrate only on these few demand types. Consequently, the system of balance equations will be relatively easy to solve.

(ii) Sometimes one can partition \(\Omega = \{1, 2, \ldots, J\}\) into several disjoint sets such that their associated \(IOs\) in different sets are either independent or weakly dependent. This will be the case if the total arrival rate of multi-item orders belonging to different sets is small. Then one can use those disjoint sets to partition the chain \(I_0\) into several smaller-dimensional chains and derive their solutions separately.
2.2. Order Fill Rates and Service Levels

The order fill rates and the service levels can be calculated directly from the stationary probability vector \( \mathbf{p} \). In particular, the type-\( K \) order fill rate is
\[
F^K = P(I_j > 0; j \in K) = P(IO_j < s_j; j \in K),
\]
and the service level of type-\( K \) orders is
\[
SL^K = P(B_j < b_j; j \in K) = P(IO_j < s_j + b_j; j \in K).
\]

2.3. Type-\( K \) Waiting Time Distribution

Recall that \( W^K \) represents the waiting time of a type-\( K \) order that is accepted as a whole. We assume that the joint distribution of \( \mathbf{IO}^K := \{IO_j; j \in K\} \) is known via the joint distribution of \( \mathbf{IO} = \{IO_j, j \in \Omega\} \). The basic idea of finding the distribution of \( W^K \) is to condition on the state that a type-\( K \) demand observes upon its arrival. \( \mathbf{IO}^K = \mathbf{n}^K := (n_j, j \in K) \).

First note that a type-\( K \) order will be filled only if none of the backlog queues \( i, i \in K \) is full (i.e., \( IO_j < s_j + b_j, \quad i \in K \)) when the order arrives, so we shall focus only on this set of states, namely,
\[
C^K = \{n^K: n_i < s_i + b_i, \quad i \in K\}.
\]
Define
\[
\tilde{p}(n^K) = P(\mathbf{IO}^K = n^K|n^K \in C^K).
\]
That is, \( \tilde{p}(n^K) \) is the probability that an accepted type-\( K \) order observes the system in state \( n^K \). For any subset of \( K \), say \( L \subset K \), we define
\[
C^K(L) := \{n^K \in C^K: n_i < s_i, \quad i \in L, \quad j \in K \setminus L\}.
\]
In other words, \( C^K(L) \subset C^K \) is the collection of states such that item \( j, j \in K \setminus L \), of a type-\( K \) order will be filled immediately, whereas item \( i, i \in L \), of the order will join the backlog queue \( i \). In particular, let \( \emptyset \) be the empty set. Then \( C^K(\emptyset) \) is the set of states in which all items of a type-\( K \) order will be filled immediately. Conditioning on \( \mathbf{IO}^K = n^K \), we obtain
\[
P(W^K \leq x|\mathbf{IO}^K = n^K) = \sum_{n^K \in C^K(\emptyset)} P(W^K \leq x|n^K)\tilde{p}(n^K)
\]
\[
\quad + \sum_{L \subset K} \sum_{n^K \in C^K(L)} P(W^K \leq x|\mathbf{IO}^K = n^K)\tilde{p}(n^K).
\]
Since \( W^K = 0 \) when \( n^K \in C^K(\emptyset) \), one has
\[
P(W^K \leq x|\mathbf{IO}^K = n^K) = \begin{cases} 1, & n^K \in C^K(\emptyset) \end{cases}.
\]
On the other hand, if \( n^K \in C^K(L) \), the fill time of the type-\( K \) order will be the time to fill all items in set \( L \):
\[
\{W^K|\mathbf{IO}^K = n^K\} = \max_{i \in L}\{W_i|\mathbf{IO}^K = n^K\}, \quad n^K \in C^K(L).
\]
Observe that, with \( n^K \in C^K(L) \) and \( i \in L \), if a new type \( i \) demand observes state \( IO_i = n, s_i \leq n < s_i + b_i \), then there will be \( n - s_i \) orders in the backlog queue \( i \). Thus the new demand will become the \((n - s_i + 1)\)st backlogged order in backlog queue \( i \) whose waiting time is just the sum of \( n - s_i + 1 \) exponential random variables with rate \( \mu_i \), which has an Erlang-(\( n - s_i + 1, \mu_i \)) distribution. Moreover, the waiting times of the items in \( L \) are conditionally independent. Letting \( G_{\mu}(x) \) be the cumulative distribution function of an Erlang-(\( n, \mu \)) random variable, we get
\[
P(W^K \leq x|\mathbf{IO}^K = n^K) = P(\max_{i \in L} W_i \leq x|\mathbf{IO}^K = n^K) = \prod_{i \in L} G_{\mu_i}^{-1}(x), \quad n^K \in C^K(L),
\]
where
\[
G_{\mu}(x) = 1 - \sum_{k=0}^{n-1} \frac{(\mu x)^k e^{-\mu x}}{k!}.
\]
Substituting (10) and (12) into (9) yields
\[
P(W^K \leq x) = \sum_{n^K \in C^K(\emptyset)} \tilde{p}(n^K)
\]
\[
+ \sum_{L \subset K} \sum_{n^K \in C^K(L)} \prod_{i \in L} G_{\mu_i}^{-1}(x)\tilde{p}(n^K).
\]

Sometimes the primary interest is to find the mean waiting time of a type-\( K \) demand, \( E[W^K] \), not its distribution. Here is a simpler procedure for this purpose:
\[
E[W^K] = \sum_{L \subset K} \sum_{n^K \in C^K(L)} E[W^K|\mathbf{IO}^K = n^K]\tilde{p}(n^K)
\]
\[
= \sum_{L \subset K} \sum_{n^K \in C^K(L)} E[V^L(n^K)]\tilde{p}(n^K),
\]
where \( V^L(n) = \max_{i \in L} W_i(n_i - s_i + 1) \), and \( V_i(n), i \in K \), are independent Erlang-(\( n, \mu_i \)) random variables with \( V_i(0) = 0 \). Notice that
\[
V^L(n^K - e_i) = V^{L - \{i\}}(n^K - e_i), \quad n^K \in C^K(L), \quad n^K - e_i \in C^K(L - \{i\}),
\]
where \( e_i \) is the \( i \)th unit vector. Using (16) and the boundary condition
\[
E[V^L(n^K_i)] = \frac{n_i - s_i + 1}{\mu_i}, \quad i \in K,
\]
one can compute the mean of \( V^L(n^K) \) from the recursive equation
\[
E[V^L(n^K)] = \sum_{i \in L} \frac{\mu_i}{\sum_{j \in L} \mu_j} E[V^L(n^K - e_i)].
\]
Let \( B^K \) denote the type-\( K \) backorders, that is, the number of type-\( K \) demands that have been accepted but have not been filled. Then, by Little's law,
\[
\]

Remark 2.2. In fact, \( W^K, K \in \Omega \), follows a phase-type distribution (see Neuts 1981). In other words, \( W^K \) can be viewed as the time until absorption in an absorbing, continuous-time Markov chain with the initial probability vector \( (\alpha, \alpha_{m+1}) \) and the infinitesimal generator

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
where we now specify its parameters $\alpha$ and $T$. Let us arrange all the states in $C^k - C^k(\theta)$ in a lexicographic order as an $m$-dimensional vector $\alpha$, where $m = \Pi_{r \in K} (s_r + b_r) - \Pi_{r \in K} s_r$, the total number of states in $C^k - C^k(\theta)$. Let $\alpha$ be the $i$th element of $\alpha$, whose value is $\tilde{p}(n^\alpha)$, for some $n^\alpha \in C^k - C^k(\theta)$. Let $\alpha_{m-1} = \sum_{n^\alpha \in C^k(\theta)} \tilde{p}(n^\alpha)$. Then $(\alpha, \alpha_{m-1})$ will be the initial probability. Here, the state $m + 1$ is the absorbing state at which all items of a type-$K$ order are filled.

Let $T_{a^*}^{m*}$ be an element of the matrix $T$. Then, if $n^k \in C^k(L), L \subset K$, the Markov chain leaves $n^k$ with rate $\Sigma_{i \in L} \mu_i$. The chain transits to state $n^k - e_i$ with probability $\mu_i / \Sigma_{i \in L} \mu_i, i \in L$, corresponding to machine $i$ completes one unit of item $i$. Therefore, for any $L \subset K$,
\[
T_{n^*}^{n^* - e_i} = \mu_i, \quad n^k \in C^k(L), i \in L.
\]

With $\alpha$ and $T$ specified as above, the distribution of $W^k$ and its expectation $E[W^k]$ can be expressed as
\[
P(W^k \leq x) = 1 - \alpha e^{T \cdot x}, \quad x \geq 0,
\]

where
\[
e^{T \cdot x} = \sum_{j=0}^{\infty} (T \cdot x)^j / j!.
\]

The mean waiting time is
\[
E[W_i] = -\alpha T^{-1} e.
\]

Although relating $W^k$ to a phase-type distribution may help the reader to gain insight in the nature of its distribution, it seems computationally more convenient to use (14) instead of (20) to evaluate the distribution of $W^k$.

### 2.4. Item-Based Performance Measures

For the later purpose of comparison between the total and partial order service models, we now derive the item-based performance measures. Let $p_i(n)$ be the marginal distribution of $IO_i$, $i = 1, 2, \ldots, J$, which can be readily derived via the joint distribution of $IO$. Since
\[
I_i = (s_i - IO_i)^+,
\]
\[
B_i = (IO_i - s_i)^+,
\]

one can also find the distribution and expectation of $I_i$ and $B_i$, but we omit the details here.

To avoid confusion, we remark that the probabilities
\[
P(IO_i < s_i) = \sum_{n=0}^{s_i-1} p_i(n)
\]
and
\[
P(IO_i < s_i + b_i) = \sum_{n=0}^{s_i+b_i-1} p_i(n)
\]
are not the fill rate and service level of item $i$ demands, which consist of demands for item $i$ from all type-$K$ orders such that $i \in K$. In other words, $F \neq F_i$ and $SL \neq SL_i$. In fact, it can be seen that
\[
F_i = \sum_{k \in K} \frac{\lambda^k}{\lambda_i} P(IO_i < s_i, IO_j < s_j + b_j, j \in K, j \neq i),
\]
\[
SL_i = \sum_{k \in K} \frac{\lambda^k}{\lambda_i} SL_i^k.
\]

Next we derive the expressions of the distribution and expectation of $W_i, i = 1, 2, \ldots, J$. To do so, we need to know the distribution of $W_i^k$, the waiting time of an item $i$ demand in an accepted type-$K$ order, $i \in K$. Let
\[
\tilde{p}_i^k(n_i) = P(IO_i = n_i, IO_j < N_i, j \neq i),
\]
\[
P(IO_i < s_i, IO_j < s_j + b_j, j \neq i, j \in K) = \frac{\lambda^k}{\lambda_i} \tilde{p}_i^k(n_i),
\]

where $N_i = s_i + b_i$ and $N_i = \{N_i, i \in K\}$. In words, $\tilde{p}_i^k(n_i)$ is the probability that the item $i$ demand in a type-$K$ order observes $IO_i = n_i$, provided the type-$K$ order is accepted. If the item $i$ demand finds that $IO_i < s_i$, then its waiting time is zero. If the item $i$ demand finds $IO_i = n_i, s_i \leq n_i < N_i$, its waiting time is an Erlang-$(\mu_i, n_i - s_i + 1)$ random variable. We thus obtain
\[
P(W_i^k \leq x) = \sum_{n_i=0}^{N_i-1} \tilde{p}_i^k(n_i) + \sum_{n_i=s_i}^{N_i-1} G_{n_i-s_i+1}(x) \tilde{p}_i^k(n_i), \quad i \in K.
\]

The expected waiting time of the item $i$ demand in an accepted type-$K$ order satisfies
\[
E[W_i^k] = \sum_{n_i=s_i}^{N_i-1} \left( \frac{n_i - s_i + 1}{\mu_i} \right) \tilde{p}_i^k(n_i).
\]

Now,
\[
P(W_i \leq x) = \sum_{k \in K} \frac{\lambda^k}{\lambda_i} P(W_i^k \leq x) = \sum_{k \in K} \frac{\lambda^k}{\lambda_i} \sum_{n_i=0}^{N_i-1} \tilde{p}_i^k(n_i) + \sum_{n_i=s_i}^{N_i-1} G_{n_i-s_i+1}(x) \tilde{p}_i^k(n_i),
\]
\[
E[W_i] = \sum_{k \in K} \left( \frac{\lambda^k}{\lambda_i} \right) E[W_i^k].
\]

### 3. THE PARTIAL ORDER SERVICE MODEL

In this section we consider the assemble-to-order environment at the distribution level, where a customer order typically requests several items, but the partial fulfillment of an order is allowed. That is, an order finds only part of its items’ backlog queues not full, it will stay in the system until these items have been supplied, and the rest
of the order will be left unfilled. We derive both the item-based and the order-based performance measures.

### 3.1. Item-Based Performance Measures

The partial order service feature implies that a type \(i\) demand will be served as long as, upon its arrival, the outstanding orders of item \(i\) is less than \(s_i + b_i\), independent of the status of other facilities. Thus the marginal distribution of \(IO_i\) can be derived without the knowledge of the joint distribution of \(IO\). But \(IO_i\) is just the steady-state occupancy in an \(M/M/1/s_i + b_i\) system with arrival rate \(\lambda_i\) and service rate \(\mu_i\). According to the standard queueing theory, \(IO_i\) follows a **truncated geometric** distribution. That is,

\[
P(IO_i = n) = \frac{(1 - \rho_i)n^{\rho_i}}{1 - \rho_i^{s_i + b_i + 1}}, \quad n = 0, 1, 2, \ldots, s_i + b_i,
\]

where \(\rho_i = \lambda_i/\mu_i\). From (22), it is straightforward to obtain for item \(i\) the average number of backorders and the average inventory on hand:

\[
E[B_i] = E[(IO_i - s_i)^+] = \frac{\rho_i^{s_i + b_i} - (1 - \rho_i^{s_i + b_i + 1})}{1 - \rho_i^{s_i + b_i + 1}},
\]

\[
E[I_i] = E[(s_i - IO_i)^+] = \frac{s_i(1 - \rho_i) - \rho_i^{s_i + b_i + 1}}{1 - \rho_i^{s_i + b_i + 1}}.
\]

The fill rate and the service level of item \(i\) can be expressed as

\[
F_i = P(I_i > 0) = P(IO_i < s_i) = \frac{1 - \rho_i^{s_i + b_i}}{1 - \rho_i^{s_i + b_i + 1}}.
\]

\[
SL_i = P(B_i < b_i) = P(IO_i < s_i + b_i) = \frac{1 - \rho_i^{s_i + b_i}}{1 - \rho_i^{s_i + b_i + 1}}.
\]

As one would expect, \(SL_i \leq F_i\) that is, the item \(i\) service level is no smaller than the item \(i\) fill rate.

Using Little’s law, we obtain the expected waiting time of an accepted request for item \(i\):

\[
E[W_i] = \frac{E[B_i]}{\lambda_i SL_i} = \frac{\rho_i^{s_i + b_i} - (1 - \rho_i^{s_i + b_i + 1})}{\lambda_i(1 - \rho_i)(1 - \rho_i^{s_i + b_i + 1})}.
\]

Clearly, with \(SL_i\) fixed (i.e., \(s_i + b_i = \text{constant}\)), the performance measures such as \(F_i\), \(E[B_i]\), and \(E[W_i]\), will improve as the base-stock level \(s_i\) increases. However, the improved customer service comes at the expense of the higher average inventory level \(E[I_i]\).

Next, let us look at the distribution of \(W_i\), i.e., the time to fill an accepted demand for item \(i\). In fact, this is a special case of Section 2.3 by setting \(K = \{i\}\). That is, the distribution of \(W_i\) can be obtained by conditioning on the number of outstanding orders in facility \(i\) upon the arrival of an accepted demand for item \(i\). In particular, as in (8), define

\[
\tilde{p}_i(n) = P(IO_i = n | IO_i < s_i + b_i) = \frac{(1 - \rho_i)\rho_i^{n}}{1 - \rho_i^{s_i + b_i}}, \quad i = 0, 1, \ldots, s_i + b_i - 1,
\]

which is the distribution of the occupancy in an \(M/M/1/s_i + b_i - 1\) system. If an accepted demand for item \(i\) finds positive inventory on hand, then the waiting time of the demand is zero. Thus

\[
P(W_i = 0) = P(IO_i < s_i | IO_i < s_i + b_i) = \sum_{n=0}^{s_i-1} \tilde{p}_i(n) = \frac{1 - \rho_i^{s_i}}{1 - \rho_i^{s_i + b_i}}.
\]

For \(x > 0\), we condition on the number of outstanding orders of type \(i\) and get

\[
P(W_i \leq x) = P(W_i = 0) + \sum_{n=1}^{s_i + b_i - 1} P(0 < W_i \leq x | IO_i = n) \tilde{p}_i(n).
\]

As we argued before, given \(IO_i = n\), the waiting time of the new type \(i\) demand has an Erlang-\((n - s_i + 1, \mu_i)\) distribution. From (28)–(30), we get

\[
P(W_i \leq x) = \frac{1 - \rho_i^{s_i + b_i}}{1 - \rho_i^{s_i + b_i + 1}} + \sum_{n=s_i}^{s_i + b_i - 1} G_{n,s_i-1}(x) \frac{(1 - \rho_i)\rho_i^{n}}{1 - \rho_i^{s_i + b_i + 1}}, \quad x > 0,
\]

where \(G_{n}(x)\) is given by (13).

**Remark 3.1.** Recall, \(b_i = \infty, i \in \Omega\), corresponds to the complete backlogging model, while \(b_i = 0, i \in \Omega\), corresponds to the lost sales model. Since these are special cases of our model and are of significant interest in practice, they are worth some discussions here.

**Complete Backlogging.** Under the partial service assumption, the item-based performance measures are well known in the standard inventory literature (see Zipkin 1997, Section 7.3.2). In particular, \(IO_i\) now follows a geometric distribution:

\[
P(IO_i = n) = (1 - \rho_i)\rho_i^{n}, \quad n = 0, 1, \ldots,
\]

which is just the limit of (22) as \(b_i \to \infty\). Correspondingly, the performance measures for item \(i\) are the limits (as \(b_i \to \infty\)) of their counterparts in the finite backlog case. It is easy to show that the expressions in (23) and (24) are, respectively, increasing and decreasing in \(b_i\). Therefore, the complete backlogging scheme results in a greater average item-backorders and henceforth a greater average waiting time, but a smaller average inventory on hand. In addition, since (25) is a decreasing function of \(b_i\), the fill rate of item \(i\) in the complete backlog case is smaller than that in the finite backlog case. But \(SL_i\) is increasing in \(b_i\).
Lost Sales. Again, the item-based performance measures are well known in the inventory literature (Zipkin 1999, Section 7.3.4). In particular, the service level and the fill rate in this case are identical. In addition, the item fill rate and average inventory on hand are greater than their counterparts in the (finite and infinite) backlog case, as F_i and E[I_i] are decreasing functions of b_i.

3.2. Order-Based Performance Measures

Obviously, for orders that contain only one item, the order-based performances are just the corresponding item-based performances. Thus, we need to focus only on the order types that contain more than one item. Again, these performance measures depend on the joint distribution of item 1 only, a type-2 customer requires only one unit of item 2, and a type-12 customer asks for one unit of each item. The simplicity of this system allows us to better illustrate the results in Sections 2 and 3, and the numerical examples help us gain more insights from the model.

4.1. The Performance Measures

Clearly, for each model it suffices to get the joint distribution p of (IO_1, IO_2). This, in turn, implies that we only need to specify the square matrices A and A_0 (both with dimension N_2 + 1), in the infinitesimal generator of IO.

The Total Order Service Model. This is the system studied in Section 2. In this case,

$A = \begin{pmatrix}
-\lambda & \lambda^2 & 0 & 0 & \cdots & 0 & 0 \\
\mu_2 & -(\lambda + \mu_2) & \lambda^2 & 0 & \cdots & 0 & 0 \\
0 & \mu_2 & -(\lambda + \mu_2) & \lambda^2 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \cdots & \mu_2 & -(\lambda + \mu_2) & \lambda^2 \\
0 & 0 & \cdots & \cdots & 0 & \mu_2 & -(\lambda + \mu_2)
\end{pmatrix}$

(IO_1, IO_2, \ldots, IO_T). In fact, the basic structure of the infinitesimal generator of the chain IO here is exactly the same as that in Section 2. The differences occur only in certain elements in the matrices A, A_0 and A_1. In particular, the transition (5) remains the same, but the transition (4) now becomes

$n_j' = \begin{cases} 
n_j + 1, & \text{if } j \in K \text{ and } n_j < N_j, \\
n_j, & \text{otherwise}. 
\end{cases}$

As a result, A_1 = A + A_0, and all the procedures and the expressions of the order-K performance measures remain valid here.

With slight modification, the approach of subsection 2.3 can also be used to derive the waiting time distribution of a partially accepted order or a subset of an accepted order. This will be of interest if a company needs to consider its delivery schedule for some key products in an order. For example, suppose for L \subseteq K, one wishes to know the distribution of W^{K(L)}_t, the waiting time to deliver all items in L, in a type K order, given that all items in L are accepted. Then, following the similar arguments that lead to (14), it can be shown that

$p(W^{K(L)} \leq x) = \sum_{n^L \in \text{C}(\emptyset)} \tilde{p}(n^L) + \sum_{L \subseteq K} \sum_{n^L \in \text{C}(L)} \prod_{i \in L} G^i_{L \setminus \{i\}, x} \tilde{p}(n^L), \quad L \subseteq K.$

4.2. Numerical Examples

In this section we present the results of the numerical experiments and discuss our key observations. Our goals are three-fold:

1. To investigate the effects of various system parameters on order-based performances. These parameters include the policy parameters (s = (s_1, s_2) and N = (N_1, N_2)) and the environmental parameters (\rho = (\rho_1, \rho_2), q = (q_1, q_2, q_1^2)). This kind of information will help us gain insight into how the system works. For example, how does the backlog queue capacity, traffic intensity, and demand
Figure 1a. Fill rates of TOS and POS models (symmetric cases).
Figure 1b. Fill rates of TOS and POS models (asymmetric cases).

correlation affect the system performance? Do the order-based performances respond to the changes in the parameters similarly to the item-based performances?

2. To compare the outputs of the TOS model and the POS model under the same set of parameters. The purpose here is to learn under what conditions the two models behave similarly, and whether one model can be used to approximate or bound the other.

3. To understand the effect of the occupancy capacity N on service levels. This will shed light on the effectiveness of using a finite-buffer model to approximate its infinite-buffer counterpart. The findings, obviously, will have important computational implications.

For all the experiments, we fixed $\lambda_1 + \lambda_2 = (1 + q^{15})\lambda = 9$. There are another four parameter vectors that determine the system performance (this explains the system complexity):

- Demand correlation vector $q$: We chose three configurations for this vector: $(0.4, 0.4, 0.2), (0.33, 0.17, 0.5)$, and $(0.1, 0.1, 0.8)$, corresponding to the systems with
Figure 2a. Service levels of TOS and POS models (symmetric cases).
Figure 2b. Service levels of TOS and POS models (asymmetric cases).

- Traffic intensity vector $\rho$: We chose the production rates such that $\rho$ takes values $(0.5, 0.5)$, $(0.9, 0.9)$, and $(0.9, 0.5)$, corresponding to the systems with symmetric and moderate workload, asymmetric workload, and symmetric and heavy workload, respectively.

- Base-stock vector $s$ and production buffer capacity vector $N$: For the symmetric cases, we selected the following configurations of $(s, N)$: $(4, 4, 4, 4)$ (the lost sale case), $(4, 4, 7, 4)$, and $(4, 4, 8, 8)$. For the asymmetric cases, we let $(s, N)$ be $(3, 5, 3, 5)$ (the lost sale case), $(3, 5, 6, 10)$, and $(3, 5, 6, 10)$.

In the graphs, $q^1$ and $q^{12}$ represents $q^1$ and $q^{12}$, respectively, while $F_1$ means $F_1$, and all other notations are similarly defined. Figures 1 through 4 depict the results of fill rates, service levels, expected waiting times and waiting time distributions, respectively, of TOS and POS models under various parameter configurations. Figure 5 presents the system-based performance comparisons of TOS and POS models.
Figure 3a. Expected waiting times of TOS and POS models (symmetric cases).

For each performance measure (say, fill rates), we divide the graphs into two subgroups: symmetric cases (e.g., Figure 1a) and asymmetric cases (e.g., Figure 1b). In each group, the horizontal pairs compare the performances of the different models (i.e., TOS vs. POS) under the same parameter configuration, while the vertical subgroups compare the performances of the same model under the different parameter configurations.

The following summarizes the key observations and their implications.

The Effects of the System Parameters. In general, we found that, in both models, the order-based performance measures respond to the parameter changes in a similar fashion as their item-based counterparts.

1.1. Figure 1 groups (Figure 2 groups) show that item-based, order-based, and system-based fill rates (service levels) are all decreasing (increasing) in N, for fixed s, but increasing (decreasing) in s, for fixed N (not reported here). These observations are not surprising, because larger base-stock levels improve order-fulfillment performance.

Figure 3b. Expected waiting times of TOS and POS models (asymmetric cases).
for accepted orders, while larger buffer sizes improve the serviceability and therefore increase the number of accepted orders.

1.2. We found that the traffic intensity \( \rho \) has a pronounced adverse effect on each performance (see Figures 1a-1 to 1a-8, 2a-1 to 2a-8, 3a-1 to 3a-4). Also, as the traffic intensity increases, the item-based performance provides poorer inference for the order-based performance. For example, under moderate traffic, the item-based bound min \( (F_1, F_3) \) provides a reasonable estimation for \( F^1 \) (see Figures 1a-5 to 1a-8, 1b-1 to 1b-6); but it is a very poor indication of \( F^{12} \) under heavy traffic, especially when \( N \) is large (see Figures 1a-1 to 1a-4). For the asymmetric cases, one should also observe the "bottleneck" effect of the heavy traffic facility on the performance of type-12 orders (see Figures 1b-1 to 4b). This indicates the importance of balanced workloads on the performance of assemble-to-order systems. Further, the traffic intensity also affects the shape of the waiting time distributions (see Figure 4).

1.3. As for the impact of demand correlation, we observed the following. First, while \( q^{12} \) has no effect on the item-based performances for the POS models, it does affect the item-based performances of the TOS models. Figures 1-3 show that as demands become more correlated, the item-based performances of TOS deteriorate, though mildly. Second, in both models one sees that as \( q^{12} \) increases, \( F^K, SL^K, \) and \( P(W^K < X) \), and \( E(W^K) \) decreases (see Figures 1-4, especially Figure 3). In fact, Xu (1999) shows analytically that this property holds for multi-item POS systems. This scenario may be explained by the fact that, as \( q^{12} \) increases, the correlation level of inventory (backlog queue occupancy) of different items increases. Third, in turn, increases the chance that both items are available (neither backlog queue is full) when a type-12 order arrives. Our graphs also show that while the adverse effect of large \( q^{12} \) on \( F^1, SL^1, \) and \( E(W^1), i = 1, 2, \) is rather mild, it has a severe impact on \( F^{12} \) in general and on \( SL^{12} \) and \( E(W^{12}) \) in heavy traffic. Third, the system-based fill rate \( F, \) service level \( SL, \) and waiting time \( E[W] \) deteriorates as \( q^{12} \) increases, most noticeably for large \( N. \)

In summary, it appears that demand correlation improves the order-based performance, whereas it worsens the system-based and item-based (for TOS models) performances. Also, the demand correlation has its greatest impact in heavy traffic.

Model Comparison. Comparing the results of the POS and the TOS models (see horizontal pairs in Figures 1-5), we made the following observations and interpretations:

2.1. Since an order is accepted on the item basis in POS, its item-based fill rates and service levels are higher than their counterparts in TOS. As a result, the POS system has a higher congestion level than that of the TOS system, with the same parameter setting. The high congestion level in POS, however, has an adverse effect on its order fill rates and service level, as it decreases the likelihood that an order is filled immediately or eventually as a whole. Thus, POS has lower order-based and system-based fill rates and service levels, as compared to TOS (e.g., see Figures 5a-1 to 5a-2). Numerical results show that the difference between the order fill rates (service levels) of the two systems increases if \( q^{12} \) or \( \rho \) increases. However, the performance of the two systems are very close over a large set of parameter settings we tested.

2.2. Because POS has a higher congestion level than TOS, the item-based, order-based, and system-based waiting times in POS are stochastically larger than their counterparts in TOS. This is evident from Figures 5a-4 and 5b-4.

2.3. Clearly, for fixed \( q^{12} \) and traffic intensity vector \( \rho, F^K \) and \( SL^K \) in POS are not affected by \( q^1/q^2 \). However, \( F^K \) and \( SL^K \) in TOS vary as \( q^1/q^2 \) varies. Numerical results show that for balanced facility utilizations \( (\rho_1/\rho_2 = 1), \) balanced arrival rates \( (q^1 = q^2) \) yield the highest order fill rate and service level. For unbalanced facility utilizations \( (\rho_1/\rho_2 \neq 1) \), the higher fill rate and service level are reached when \( q^1 = q^2 \) takes a value close to \( \rho_1/\rho_2 \).

2.4. In general, both models demonstrate the same qualitative behavior, and the results from the POS model provides rather reliable estimates for the performance measures in the TOS model for a broad range of parameter settings (see Figure 5).

The Impact of Finite Buffer Sizes. Figures 2a-5 to 2a-8 show that under a moderate traffic intensity, the system with small backlog queue capacities (in our cases, \( b_1 = b_2 = 8 - 4 = 4 \)) can achieve almost 100% service levels, thus its performance (including fill rates, waiting time distributions, etc.) provides an accurate approximation for its counterpart in the infinite queue system. Numerical experiments show that such results hold for a wide range of parameter configurations, as long as the traffic intensity is moderate.

To illustrate the computational efficiency of our procedure for large-scale systems—i.e., systems with high traffic intensities and large backlog queue capacities—we provide two examples.

Example 1. Consider a symmetric POS system with \( \rho_1 = \rho_2 = 0.9 \) (heavy traffic), \( q^1 = q^2 = 0.1, q^{12} = 0.8 \) (high demand correlation), \( s_1 = s_2 = 8, \) and complete backlogging. Suppose that the desired approximate accuracy is at least 99%. To do so, we set

\[
SL = q^1 SL^1 + q^2 SL^2 + q^{12} SL^{12} \geq 0.99,
\]

so that less than 1% of orders are lost. Using the fact that \( (IO_1, IO_2) \) is positively quadrant dependent (Xu 1999), the following inequality holds:

\[
SL^{12} = P(IO_1 < s_1 + b_1, IO_2 < s_2 + b_2) \geq P(IO_1 < s_1 + b_1) P(IO_2 < s_2 + b_2) = (SL^1)(SL^2).
\]

(35)

To ensure (34), it is sufficient to find \( s_1 + b_1 = s_2 + b_2 \) such that

\[
(SL^1)(SL^2) = \left( 1 - \rho_1^{s_1+b_1} + 1 \right) \left( 1 - \rho_2^{s_2+b_2} + 1 \right) \geq 0.99.
\]
With \( p_1 = p_2 = 0.9 \) and \( s_1 = s_2 = 8 \), the above inequality holds when \( b_1 = b_2 = 21 \). Therefore, even under heavy traffic, moderate backlog queue capacities (\( b_1 = b_2 = 21 \)) will be capable of providing an accurate approximation (with the service level more than 99%) to the system with complete backlogging. For this specific problem, the matrix-geometric approach is roughly \((N_1 + 1)^2/9 = 30^2/9 = 100\) times faster than the direct approach (see Remark 2.1). Finally, we note that the above result is also valid for the TOS system, because our numerical examples indicate that the service level of a TOS system is bounded below by its counterpart of a POS system (see Figures 5a-2 and 5b-2).

**Example 2.** Consider an asymmetric POS system with parameter settings similar as in Example 1, except \( p_1 = 0.9 \), \( p_2 = 0.5 \), and \( s_1 = 4, s_2 = 8 \). Again, the desired approximate accuracy is more than 99%. Following the argument of Example 1, if we let \( b_1 = 21 \) (or \( N_1 = s_1 + b_1 = 29 \)), \( b_2 = 3 \) (or \( N_2 = s_2 + b_2 = 7 \)), then \( SL_1 = 0.996, SL_2 = 0.995, SL_{12} \approx (SL_1)(SL_2) = 0.991 \), and hence \( SL \approx 0.99 \). Again, the matrix-geometric solution is roughly 100 times faster than the direct approach.

The above examples illustrate that, in most cases, computational complexity is dominated by the traffic intensity, rather than the buffer size constraint. Therefore, it is safe to say that the finite buffer assumption in our solution procedure does not limit its ability to solve the complete backlogging model that has a small \( J \).

**5. CONCLUDING REMARKS**

We presented a model of assemble-to-order production/inventory systems that includes stochastic processing times. The motivation of building such model was due to the prevalence of stochastic supply times in the computer and semiconductor industries, where assemble-to-order manufacturing has become a common practice. The model was further tailored into two models, one with total order service and the other with partial order service. Exact procedures were developed to evaluate the order-fulfillment performance measures that are of increasing importance to managers. The procedures were illustrated by a two-item system. Numerical examples were presented and their implications were discussed.

We believe that our study is the first exact analysis of this kind. As such, for simplicity, we have made some restrictive assumptions, such as Poisson demand and exponential processing times. While these simplified assumptions are necessary to begin with, it is our hope that our study will inspire some other research efforts for alternative models and solution procedures. (The methodology employed in this paper is applicable in principle to models with phase-type processing time distributions, but this can only add to the computational and notational burden of the model without providing additional insights.)

As a result of multidimensional Markov chains, it is not surprising to note that the solution procedure provided in this paper requires considerable computational effort for
performance evaluation of large-scale systems (i.e., with large $J$ and heavy traffic intensities). Needless to say, efficient approximations and heuristic procedures are in demand. (An important contribution of the current model is to provide a benchmark for testing these approximations.) This, in fact, is one of our ongoing research projects: Since item-based performance measures are much easier to obtain, we intend to develop certain bounds for the order-based performance measures, and these bounds involve only the item-based information. To develop such bounds, structural studies will be carried out.

Another question that interests us is: What is the impact of the standard independent-demand assumption when the demands across items are actually correlated? This line of study will reveal to what extent we can trust such crude assumptions and what is the value of identifying demand types.
Figure 5a-1: System Fill Rates

Figure 5a-2: System Service Levels

Figure 5a-3: Expected System Waiting Times

Figure 5a-4: System Waiting Time Distributions

Legends: s1=s2=4, rho1=rho2=0.9, q1=q2=0.1, q12=0.8

Figure 5a. System-based performance comparisons of POS and TOS models (symmetric cases).

Figure 5b-1 System Fill Rates

Figure 5b-2 System Service Levels

Figure 5b-3: Expected System Waiting Times

System Waiting Time Distributions

Figure 5b. System-based performance comparisons of POS and TOS models (asymmetric cases).
ACKNOWLEDGMENT

Research of the first author was supported in part by NSF grant DMI-9500083.

REFERENCES


