The Effectiveness of Several Performance Bounds for Capacitated Assemble–to–Order Systems

Savas Dayanik
Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ 08544
Tel. (609) 258-5305, e-mail: sdayanik@princeton.edu

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

Susan H. Xu
Smeal College of Business Administration
Penn State University, University Park, PA 16802
Tel. (814) 863-0531, e-mail: shx@psu.edu

November 2001; revised January 2003

Abstract

We consider an assemble–to–order (ATO) system: Components are made to stock by production facilities with finite capacities, and final products are assembled only in response to customers' orders. The key performance measures in this system, such as order fill rates, involve evaluation of multivariate probability distributions, which is computationally demanding if not intractable. The purpose of this paper is to develop computationally efficient performance estimates. We examine several ideas scattered in diverse literatures on approximations for multivariate probability distributions, and determine which approach is most effective in the ATO application. To do so, we first tailor different approximation ideas to the ATO setting to derive performance bounds, and then compare these bounds theoretically and numerically. The bounds also allow us to make connections between capacitated and uncapacitated ATO systems and gain various insights.

1 INTRODUCTION

This paper is concerned with the performance evaluation techniques for the increasingly popular assemble–to–order (ATO) manufacturing systems. There are multiple components and multiple products. Inventories are kept only at the component level; final products are assembled in response to customer orders. Such systems are considered to be ideal for realizing mass customization, because they fully enjoy the benefit of risk pooling at the component level and postpone product differentiation to the latest point in the entire production process. The performance measure of primary interest to management is the order fill rate with a time window, which is the probability of fulfilling a customer order within a prespecified time window, under any given component base-stock levels.

We assume that interarrival times of customer orders are independently and identically distributed. There is a fixed probability that a customer requests any given product. The production of each component is governed by a dedicated production facility with exponential service times, following a base-stock policy. Since each customer order (a final product) typically consists of several components and the product can not be assembled unless all those components are available, the order fill rate involves evaluation of multivariate probability distributions, which is computationally challenging. The problem is further complicated by the fact that different products require different but overlapping sets of components.

When the interarrival and production times are exponentially distributed, which we call the Markovian model, Song, Xu and Liu (1999) present an exact approach for performance evaluation. The method, however, is computationally efficient only for small to medium sized problems. The goal of this paper is to develop easier-to-compute performance estimates to overcome the complexity of the exact approach. For exponentially distributed interarrival times but generally distributed production times, Glasserman and Wang (1998) demonstrate an asymptotic linear relationship between inventory and delivery leadtime at high service level. The current paper has a different focus and uses different approaches. We aim to approximate the service level in all ranges, rather than only at high service level. We refer the reader to the above mentioned papers and Song and Zipkin (2001) for other related literature.

The computational complexity of multivariate probability distributions is well known. Several approaches have been developed in the literature of multivariate statistics. The basic idea of these approaches is to develop bounds that involve solving smaller problems than the original one. If we can develop bounds like these, then for the smaller problems we can apply the results in Song, Xu and Liu. In addition, there has been some recent development in performance bounds in the theory of queues with signals. But can all these approaches apply to our setting? If so, which one is better? These are the questions we aim to answer.

We first tailor different approximation ideas that spread over diverse literatures into the capacitated ATO setting to derive performance bounds, and then compare these bounds theoretically and numerically. As such, the work requires several different kinds of techniques, such as stochastic comparison and queueing network analysis. It also requires extensive computation, including simulation.

The conclusions of our study are clear, which are summarized in the last section. The numerical results also shed light on various managerial insights, such as the quantitative change in system performance as product structure changes. Since the study clarifies the effectiveness of different ideas, we hope the results here will facilitate both real implementations and future research developments.

Throughout the derivation of the bounds, whenever it applies, we also mention the parallel development of similar bounds in the literature on uncapacitated ATO systems (i.e., systems with i.i.d./ leadtimes). We further discuss the connections between the capacitated and uncapacitated ATO systems in Section 6.

The rest of the paper is organized as follows. Section 2 presents a simple ATO system with Markovian structures. We call this the basic model. Section 3 develops several lower bounds for the performance measures of the basic model. These include 1) setwise lower bounds based on the dependence structure of the system; 2) distribution–free Bonferroni–type lower bounds commonly used to bound multivariate distributions (see, e.g., Costigan 1996); 3) setwise–Bonferroni combination lower bounds which combine the setwise and Bonferroni lower bounds to overcome the degeneracy problem (i.e., the null probability as a lower bound) which sometimes occurs in the Bonferroni bounds (also see Costigan 1996); and 4) the signal lower bound based on the study of quasi–reversibility of queueing network with signals (Chao, Miyazawa and Pinedo 1999). We show analytically that the setwise bounds are tighter than the signal bound. We also develop several algorithms to improve the computational efficiency of several bounds.

Section 4 develops several upper bounds for the performance measures of the basic model. These bounds are setwise upper bounds and Frechet–type upper bounds, which use a lower dimensional distribution to bound a higher dimensional distribution (Joe 1997). We show that the setwise bounds are tighter than the Frechet bounds of the same order.

Section 5 reports the numerical results. For small– to medium–size systems with unit-demand and Markovian structures, the bounds are compared against the exact solutions developed in Song, Xu and Liu. Otherwise the bounds are compared against the simulation results. Section 6 discusses extensions to non-Markovian systems and connections with uncapacitated ATO systems. Finally, Section 7 summarizes the key findings.

2 THE BASIC MODEL

Denote by $\{1, ..., J\}$ the component indices. For any subset K of $\Omega = \{1, ..., J\}$, a product is said to be of type–K if it is composed of a unit of each component in K. The overall demand process, $\{D(t), t \geq 0\}$, forms a Poisson process with arrival rate λ. There is a fixed probability, q^K, that a demand is of type–K (i.e., requesting product K). The type–K demand process will be denoted by $\{D^K(t), t \geq 0\}$, which has demand rate $\lambda^K \triangleq q^K\lambda$. (We use subscripts for component types, and superscripts for order types.)

The inventory of component i is replenished by a single-machine production facility i, which operates on a first-come, first-served (FCFS) basis and has i.i.d. exponential processing times with rate μ_i. Let s_i be the base-stock level for component i. Whenever the inventory position of component i is less than s_i, we send a replenishment order to make up the difference. Otherwise, we do not order. The final assembly time is assumed to be negligible relative to the component production times.

A demand for component i that cannot be filled immediately are queued in the backlog queue i with capacity $b_i \leq \infty$ and is filled on the FCFS basis. (Except being a physical characterization of many real systems, b_i can also be used as a measure of customers' patience). If upon the arrival of a customer order, some of the components' queues are full and others are not, we assume that we accept the requests for the components whose queues are not full and reject otherwise. We refer to this scheme as *partial order service* (POS). This is reasonable for distribution systems such as online retailing, in which the a customer order consists of several different items (corresponding to components), and the assembly of a product entails picking out the items in the customer's order and packaging them. An alternative assumption is that we reject the entire order in such a situation, which is called *total order service* (TOS). The TOS scheme is more realistic if there is a physical product to be assembled. However, the POS scheme is more tractable in developing the bounds. For this reason, we focus on the former. Computational results show that it is reliable to use the POS model to approximate the TOS model. See also a discussion in Song, Xu and Liu.

Note that the demand process for component i,

$$D_i(t) = \sum_{K:i\in K} D^K(t),$$

is also a Poisson process, with rate

$$\lambda_i = \sum_{K:i\in K} \lambda^K.$$

Due to the nature of the base-stock policy, the production facility of component i sees the same arrival process as the demand process. Therefore, the supply system of component i constitutes an $M/M/1/N_i$ queueing system with arrival rate λ_i, service rate μ_i and maximum $N_i = s_i + b_i$ jobs.

We now define several important random variables in steady-state:

$$\begin{aligned} IO_i &= \text{inventory on order of component } i, 0 \le IO_i \le N_i, \\ W_i &= \text{delivery or waiting time (queue time plus service time) of component } i, \\ I_i &= \text{on-hand inventory of component } i, \\ B_i &= \text{number of backorders of component } i. \end{aligned}$$

These random variables are related as follows:

$$\begin{aligned} I_i &= (s_i - IO_i)^+, \qquad i \in \Omega, & (1) \\ B_i &= (IO_i - s_i)^+, \qquad i \in \Omega, & (2) \\ W_i &= \sum_{j=1}^{[IO_i+1-s_i]^+} T_{i,j} \mathbf{1}_{\{IO_i < s_i + b_i\}}. \qquad i \in \Omega, & (3) \end{aligned}$$

Here, $(x)^+ = \max\{x, 0\}$ for any real number x, and $\mathbf{1}_E$ is the indicator function of event E. $T_{i,j}$ are i.i.d. exponential random variables with rate μ_i, $i \in \Omega$. The upper limit of the summation in (3) is the position of the last unit in the backlog queue i. Equation (3) can be understood as follows: If $[IO_i + 1 - s_i]^+ = 0$, then the component-i demand is serviced immediately by the existing inventory of component i; otherwise the unit becomes the $(IO_i + 1 - s_i)$th job in the backlog queue i.

Throughout the paper, we use bold-faced letters to denote vectors. For example, $\mathbf{IO} = \{IO_1, IO_2, \ldots, IO_J\}$. Define $\mathbf{W}, \mathbf{I}, \mathbf{B}, \mathbf{s}, \mathbf{b}$ and $\mathbf{N}$ similarly. For a J–dimensional vector $\mathbf{a} = (a_1, a_2, \ldots, a_J)$ and $K \subseteq \Omega$, we let $\mathbf{a}^K = (a_i : \ i \in K)$ and $a^K_{\max} = \max_{i \in K}\{a_i\}$. Also, equality in distribution between two random variables or vectors is denoted by $=_{st}$. Inequalities between vectors are componentwise inequalities.

For any $K \subseteq \Omega$, the key *order-based* performance measures of interest include:

(a) Type–K order fill rate $F^{K,x}$ with window x for any given $x \ge 0$. Let W^K denote the steady-state order delivery time for type–K demand, then

$$F^{K,x} \triangleq P(W^K \le x) = P(\max_{i \in K} W_i \le x). \qquad (4)$$

where W_i is given in (3). When $x = 0$, $F^{K,0}$ is the immediate type–K fill rate, or simply the type–K fill rate, which is the probability that a type–K order is filled immediately. For simplicity, we drop the superscript 0 from its notation. We have

$$F^K \triangleq F^{K,0} = P(\mathbf{I}^K > 0) = P(\mathbf{IO}^K < \mathbf{s}^K), \quad K \subseteq \Omega, \qquad (5)$$

where we apply the relation between I_i and IO_i in (2).

(b) Type–K service level SL^K, which is the probability that a type–K order is accepted as a whole and filled eventually:

$$SL^K \triangleq P(\mathbf{B}^K < \mathbf{b}^K) = P(\mathbf{IO}^K < \mathbf{N}^K), \tag{6}$$

where we apply the relation between B_i and IO_i in (1).

For convenience, we write $F^{\{i\}} \equiv F^i$ and $F^{\{i,j\}} \equiv F^{ij}$, for every $i, j \in \Omega$ (similarly for SL^K). Let F, SL and W be the fill rate, service level, and delivery time of an arbitrary order (regardless of its type), respectively. Then

$$\begin{aligned} F &= \sum_{K \subseteq \Omega} q^K F^K, \\ SL &= \sum_{K \subseteq \Omega} q^K SL^K, \\ P(W \leq x) &= \sum_{K \subseteq \Omega} q^K P(W^K \leq x). \end{aligned}$$

It is clear that the outstanding order vector **IO** determines other performance vectors. For the basic model, assuming b_i can be ∞ for at most one i, Song et al. develop a matrix–geometric solution for the joint distribution of **IO**, which in turn leads to an exact procedure for performance evaluation of other performance measures. Because a performance measure of a type–K order depends only on the joint distribution of $\mathbf{IO}^K = \{IO_i, i \in K\}$, $K \in \Omega$, the computational complexity of the matrix–geometric solution is mainly determined by the maximal cardinality of order types (of course, the cardinality of the state space of $\mathbf{IO}^K$ also plays a role in computational complexity), rather than the total number of components in the system. For example, suppose that in a 20-component system, there are total of 10 different order types with each order type containing at most 4 components. While it would be a daunting task to compute the joint distribution of the 20–dimensional vector **IO** , the matrix–geometric solution will allow us to more efficiently evaluate $\mathbf{IO}^K$ for all 10 order types, with the dimensionality of $\mathbf{IO}^K$ no more than 4. Of course, the procedure becomes less efficient when the maximum number of components in a single product (with upper bound J), the number of different order types (with upper bound $2^J - 1$), or the cardinality of the state space of **IO** is large. Also, the algorithm is only applicable to the systems with Markovian structure and a unit demand for each component.

3 LOWER BOUNDS

This section develops several lower bounds of the outstanding order vector IO^K, $K \subseteq \Omega$, in the basic model.

3.1 Setwise Lower Bounds

The setwise lower bounds are based on the following idea: We construct a new demand process that has the same marginal demand distributions as the original demand process, but is less correlated across components. To this end, we partition the components in a type–K order into several clusters such that the demands within each cluster are identically distributed as its counterpart in the original process, but are independent across different clusters. As a result, a performance vector, say $\overline{\mathbf{IO}}^K$ (with the new demand process), becomes a collection of several independent performance subvectors, with the distribution of each subvector having a lower dimension than the original vector. We shall show that the distribution of $\mathbf{IO}^K$ is bounded below by that of $\overline{\mathbf{IO}}^K$.

Similar bounds have appeared in various applications of multivariate stochastic processes, see, for example, Baccelli and Makowski (1989), Baccelli, Makowski and Schwarz (1989), Mamer and Smith (1993), Nelson and Tawani (1989), Connors and Yao (1996), Song (1998), and Lu, Song and Yao (2003a). The notion of *associated* random variables is essential for the development the bounds in these studies, which, unfortunately, is not applicable to the model setting considered here. The results developed in Xu (1999) and Li and Xu (2000) are applicable, however. To apply those results, we need to introduce the following definitions and properties (see, e.g., Tong (1980) and Shaked and Shanthikumar (1997)).

Definition 3.1. Let $\mathbf{X} = (X_1, \ldots, X_J)$ and $\mathbf{Y} = (Y_1, \ldots, Y_J)$ be two random vectors.

1. $\mathbf{X}$ is said to be larger than $\mathbf{Y}$ in the upper orthant order, denoted as $\mathbf{X} \geq_{uo} \mathbf{Y}$, if for all $\mathbf{x} \in \mathcal{R}^J$,

$$P(\mathbf{X} > \ \mathbf{x}) \geq P(\mathbf{Y} > \ \mathbf{x}).$$

   $\mathbf{X}$ is said to be smaller than $\mathbf{Y}$ in the lower orthant order, denoted as $\mathbf{X} \leq_{lo} \mathbf{Y}$, if for all $\mathbf{x} \in \mathcal{R}^J$,

$$P(\mathbf{X} \leq \ \mathbf{x}) \geq P(\mathbf{Y} \leq \ \mathbf{x}).$$

   If $\mathbf{X} \geq_{uo} \mathbf{Y}$ and $\mathbf{X} \leq_{lo} \mathbf{Y}$, then $\mathbf{X}$ is said to be more positively quadrant dependent (PQD) than $\mathbf{Y}$. Roughly speaking, this means that the tendency that all the elements in $\mathbf{X}$ are small or large is stronger than that of $\mathbf{Y}$.

2. $\mathbf{X}$ is said to be stochastically greater than $\mathbf{Y}$, denoted as $\mathbf{X} \geq_{st} \mathbf{Y}$, if for any increasing function f, $Ef(\mathbf{X}) \geq Ef(\mathbf{Y})$.

It is well-known that (see, for example, Tong 1980):

$$\mathbf{X} \geq_{st} \mathbf{Y} \implies \mathbf{X} \geq_{uo} \mathbf{Y} \text{ and } \mathbf{X} \geq_{lo} \mathbf{Y}. \tag{7}$$

For a given $K \subseteq \Omega$ with size $|K|$, let us consider the subsystem that is only composed of the components in K and call it subsystem $\mathcal{S}^K$. Then $\mathbf{D}^K = \{(D_i(t), i \in K), t \geq 0\}$ is the demand process of subsystem $\mathcal{S}^K$. Let $\{K_1, K_2, \ldots, K_\nu\}$ be a partition of K. We construct a new demand process, $\overline{\mathbf{D}}^K = \{(\overline{D}_i(t), i \in K), t \geq 0\} = \{\overline{\mathbf{D}}^{K_\ell}, \ell = 1, \ldots, \nu\}$, such that the processes $\overline{\mathbf{D}}^{K_1}, \ldots, \overline{\mathbf{D}}^{K_\nu}$ are mutually independent, but $\mathbf{D}^{K_\ell} =_{st} \overline{\mathbf{D}}^{K_\ell}$, $\ell = 1, 2, \ldots, \nu$. That is, the new demand process of each cluster is identically distributed as the original demand process of the corresponding cluster. Denote by $\overline{\mathcal{S}}^K$ a new subsystem which has the demand process $\overline{\mathbf{D}}^K$ but otherwise identical features as in $\mathcal{S}^K$. Let $\mathbf{IO}^K$ be the outstanding order subvector in system $\mathcal{S}^K$, and define other performance subvectors similarly. Use an overline to denote the corresponding performance vector in system $\overline{\mathcal{S}}^K$. According to Xu (1999), $\mathbf{IO}^K$ is more PQD than $\overline{\mathbf{IO}}^K$.

Theorem 3.2. $\mathbf{B}^K$ is more PQD than $\overline{\mathbf{B}}^K$, and $\overline{\mathbf{I}}^K$ is more PQD than $\mathbf{I}^K$, for any K.

Proof. Note that if $\mathbf{X}$ is more PQD than $\mathbf{Y}$, and if f_i, $i = 1, \ldots, J$, are increasing functions, then $(f_1(X_1), \ldots, f_J(X_J))$, is more PQD than $(f_1(Y_1), \ldots, f_J(Y_J))$. Since $B_i = [IO_i - s_i]^+$ is an increasing function of IO_i, $i \in \Omega$, $\mathbf{B}^K$ is more PQD than $\overline{\mathbf{B}}^K$. Now consider $I_i = [s_i - IO_i]^+$. First note that if a random vector $\mathbf{X}$ is more PQD than another random vector $\mathbf{Y}$, then $-\mathbf{Y}$ is more PQD than $-\mathbf{X}$. Therefore $\mathbf{s}^K - \overline{\mathbf{IO}}^K$ is more PQD than $\mathbf{s}^K - \mathbf{IO}^K$. Because $[x]^+$ is an increasing function of x, $\overline{\mathbf{I}}^K$ is more PQD than $\mathbf{I}^K$. ■

The above results allow us to derive lower bounds for the distribution of a performance vector, such as $\mathbf{IO}^K$, $\mathbf{W}^K$, $\mathbf{I}^K$ and $\mathbf{B}^K$. In particular, we have

$$P(\mathbf{IO}^K \leq \mathbf{n}^K) \;\; \geq \;\; P(\overline{\mathbf{IO}}^K \leq \mathbf{n}^K) = \prod_{\ell=1}^{\nu} P(\mathbf{IO}^{K_\ell} \leq \mathbf{n}^{K_\ell}), \tag{8}$$

$$P(\mathbf{W}^K \leq \mathbf{w}^K) \;\; \geq \;\; P(\overline{\mathbf{W}}^K \leq \mathbf{w}^K) = \prod_{\ell=1}^{\nu} P(\mathbf{W}^{K_\ell} \leq \mathbf{w}^{K_\ell}). \tag{9}$$

Instead of developing bounds for every performance measure, here and in the rest of the paper, we shall derive bounds for the order fill rate F^K only. The bounds for other performance measures can be obtained similarly. Note that $F^K = SL^K$, $K \subseteq \Omega$, for the lost-sales system. In addition, with everything else equal, the service level for the partial backlogging system with $N_i = s_i + b_i$ equals the service level for the lost-sales system with base-stock level N_i, $i \in \Omega$.

For any given set K and a scalar m, $m < |K|$, we say $K(m)$ is an order–m partition of K if

$$K(m) = \{K_1, ..., K_\nu : \; |K_\ell| = m, \; \ell \neq j, \; |K_j| = |K| - m(\nu - 1) \text{ for some } j\}.$$

A lower bound resulting from an order–m partition is called an order–m setwise lower bound. Let $\mathcal{P}_m(K)$ be the set of all order–m partitions of K. Then, from (5) and Theorem 3.2, the best order–m lower bounds for F^K is:

$$F^K \quad = \quad P(\mathbf{IO}^K < \mathbf{s}^K) \geq \max_{K(m)\in\mathcal{P}_m(K)} \left\{ \prod_{K_\ell \in K(m)} F^{K_\ell} \right\}. \tag{10}$$

Observe that a setwise lower bound never degenerates, and reduces to the exact solution under independence.

Although any higher–order setwise lower bound is tighter than the first–order setwise lower bound, it is generally not true that the lower bound will become tighter as m increases. For example, the best third–order bound may not be tighter than the best second–order bound. For this reason, we shall pay special attention to the first– and second–order bounds of F^K, which are given by:

$$F^K = P(\mathbf{IO}^K < \mathbf{s}^K) \geq \prod_{j\in K} P(IO_j < s_j) \tag{11}$$

$$F^K \geq \max_{K(2)\in\mathcal{P}_2(K)} \left\{ \prod_{K_\ell \in K(2)} P(\mathbf{IO}^{K_\ell} < \mathbf{s}^{K_\ell}) \right\}. \tag{12}$$

The first–order bounds are the simplest, which require only the marginal distributions of IO_i, $i \in K$. However, the bounds may become loose for large $|K|$. The second–order bound requires evaluating the joint distributions of $\mathbf{IO}^{ij}$, $i, j \in K$, for which the matrix–geometric solution developed by Song, Xu and Liu is especially efficient. Here $\mathcal{P}_2(K)$ is also called the pair partition of K, in which we partition K into pairs. In the case that $|K|$ is odd, there exists one $K_j \in \mathcal{P}_2(K)$ that contains only one component. After F^{ij}, $i, j \in K$, are computed, finding the best order-2 bound, given in (12), becomes a nonbipartite weighted matching problem, which can be solved *exactly* by existing algorithms in the combinatorial optimization literature. See a similar discussion in Song (1998).

The procedure may become impractical if we have to evaluate a large set of F^{ij} for all pairs $(i, j) \subseteq K$. In addition, for the higher–order partitions ($m > 2$), there do not exist efficient algorithms to obtain the optimal partition of $\mathcal{P}_m(K)$, even if we are able to compute F^{K_ℓ} for all $K_\ell \in \mathcal{P}_m(K)$. Therefore, we propose a heuristic partition of K which is based on arrival rates λ^M, $M \subseteq \Omega$, rather than m–variate distributions of outstanding orders. This algorithm is also the basis of Algorithm 3.6, where we derive the combination bound for F^K (see Section 3.3). This algorithm at each step selects order type M, $|M| = m$, which has the highest *aggregated order arrival rate* among the unsolicited components in K. The following is a greedy procedure to find such a partition (Assume $|K| > 3$; recall that such a heuristic is valuable only when $|K|$ is large).

Algorithm 3.3. A Greedy Heuristic for the mth Order Partition of K, $|K| > m$.

1. Set $n = 1$ and $K(n) = K$.

2. Let
$$M^*(n) = \text{argmax}\Bigg\{ \sum_{L'\subseteq M\subseteq\Omega} q^M : L' \subseteq K(n), |L'| = m \Bigg\}, \tag{13}$$
and set $K(n+1) = K(n) - M^*(n)$.

3. If $|K(n+1)| > m$, set $n = n+1$ and go to step 2. Otherwise, set $M^*(n+1) = K(n+1)$ and $P^*_m(K) = (M^*(1), ..., M^*(n+1))$. STOP.

The above algorithm is motivated by the following reasoning: Among all unsolicited components $L' \subseteq K(n)$ in the nth iteration, one expects that the components in set $M^*(n)$, defined as in (13), induce the highest correlation among components of $\mathbf{IO}^{M^*(n)}$, compared with that of other $\mathbf{IO}^{L'}$, $L' \subseteq K(i)$, $|L'| = m$. Also, (13) means that the order type-$M^*(n)$ is quite frequent, so it is wise to trace the information of $\mathbf{IO}^{M^*(n)}$. The heuristic then utilizes this information.

Example 3.4. Consider a six-component system with arrival rates q^K given by

$$q^{1234} = 0.1,\ q^{2356} = 0.3,\ q^{23} = 0.2,\ q^{24} = 0.4.$$

Here we abbreviate the notation of a set $\{i_1, i_2, ..., i_k\}$ by $i_1 i_2 ... i_k$. All other q^K's are zero. We wish to obtain a good second–order partition of set $(1, 2, 3, 4)$. By Algorithm 3.3,

$$\max_{(j,k)} \Bigg\{ \sum_{(j,k)\subseteq M\subseteq(1,2,3,4)} q^M \Bigg\} = q^{1234} + q^{2356} + q^{23} = 0.6.$$

Hence $M^*(1) = (2,3)$ and $M^*(2) = (1,4)$, i.e., the partition is $\{(2,3),(1,4)\}$. Note that even though q^{24} is the largest among the pairs in $(1,2,3,4)$, (2,4) is not kept as a pair. This is because that we expect correlation of $\mathbf{IO}^{23}$ to be stronger than that of $\mathbf{IO}^{24}$ due to the aggregated order types 1234, 2356 and 23. ■

In Section 7, we will discuss the extension of setwise lower bounds to other types of ATO systems. In particular, the setwise lower bound is still valid for the $M^X/M/1$ type of ATO systems. The next example illustrates the first-order bound for such a system.

Example 3.5. Suppose the interarrival time between two orders is exponentially distributed with rate λ and batch size vector $\mathbf{X}$ has the distribution

$$P(\mathbf{X}^K = \mathbf{x}^K, \mathbf{X}^{\Omega-K} = \mathbf{0}) = q^K \prod_{i\in K}(1-p_i)p_i^{x_i-1}, \quad \mathbf{x}^K \geq \mathbf{1},\ K \subseteq \Omega. \tag{14}$$

This assumption implies that, with probability q^K, an order will bring a geometric batch X_i to facility i for each $i \in K$, $K \in \Omega$. Assume $N_i = \infty$, $i \in \Omega$. Then, by (11), the first–order setwise lower bound is

$$F^K \geq \prod_{i \in K} P(IO_i^a + X_i \leq s_i) = \prod_{i \in K} \sum_{x_i=1}^{s_i} P(IO_i^a \leq s_i - x_i)(1-p_i)p_i^{x_i-1}.$$

Let

$$\rho_i = \frac{1}{(1-p_i)\mu_i} \sum_{K: i \in K} \lambda q^K = \frac{\lambda_i}{(1-p_i)\mu_i} < 1,$$

then the probability mass of IO_i^a is given by (see, for example, Gross and Harris 1974)

$$\begin{aligned} P(IO_i^a = 0) &= 1 - \rho_i \\ P(IO_i^a = n) &= (1-\rho_i)[p_i + (1-p_i)\rho_i]^{n-1}[(1-p_i)\rho_i], \quad n > 0, \ i \in \Omega. \qquad \blacksquare \end{aligned}$$

3.2 Bonferroni-Type Bounds

The lower bounds developed in Section 3.1 depend heavily on PQD properties of performance vectors. It is interesting to compare them with the commonly used, distribution–free, *Bonferroni–type* bounds. By distribution–free, we mean that the construction of the bounds applies to any distribution. In particular, let $K = \{i_1, ..., i_{|K|}\}$, then the Bonferroni–type lower bounds of orders 1–3 on F^K are (see Costigan 1996 for references):

$$\begin{array}{lcll} F^K & \geq & \sum_{i \in K} F^i - (|K| - 1) & \text{(first order)} \\ F^K & \geq & \sum_{j=2}^{|K|} F^{i_{j-1}i_j} - \sum_{j=2}^{|K|-1} F^{i_j} & \text{(second order)} \\ F^K & \geq & \sum_{j=3}^{|K|} F^{i_{j-2}i_{j-1}i_j} - \sum_{j=3}^{|K|-1} F^{i_{j-1}i_j} & \text{(third order).} \end{array} \tag{15}$$

A major shortcoming of a Bonferroni–type lower bound is that it may degenerate (to a non-positive number), especially for large $|K|$. In addition, it does not yield the exact solution under independence.

Cheng et al. (2002) use this type of bounds in a system with i.i.d./ leadtimes, and Zhang (1999) apply similar bounds to a system with sequential general production times.

3.3 Setwise–Bonferroni Combination Bounds

For large $|K|$, the degeneracy issue of Bonferroni–type bounds can be resolved by the approach of *setwise–Bonferroni combination* bounds (Costigan 1996) that first partition K into subsets $\{K_1, \ldots, K_\nu\}$, then apply Bonferroni–type bounds to each F^{K_j}, $j = 1, \ldots, \nu$. Next we give an algorithm that computes the second order combination bounds. To simplify notation, we assume $|K| = m\nu$, that is, K can be partitioned into ν subsets with each subset having cardinality m.

Algorithm 3.6. The Second-Order Combination Lower Bound of F^K, $|K| = m\nu,\ m > 2$

1. Apply Algorithm 3.3 to obtain the mth order partition $P_m^*(K) = (M_1^*, \dots, M_\nu^*)$ of K. Let $M_j^* = (i_{1,j}, \dots, i_{m,j}),\ j = 1, 2, \dots, \nu$.

2. Use (15) to obtain the second–order Bonferroni–type lower bound for $F^{M_j^*}$, $j = 1, \dots, \nu$.

3. The second–order combination lower bound for F^K is given by

$$\prod_{j=1}^{\nu} \left[\sum_{\ell=2}^{m} F^{i_{\ell-1,j}\, i_{\ell,j}} - \sum_{\ell=2}^{m-1} F^{i_{\ell,j}} \right].$$

3.4 Signal Lower Bound

Chao et al. (1999) show that a system with several parallel $M/M/1$ queues and simultaneous arrivals can be modeled as a queueing network with concurrent movements by means of *positive signals*. The authors also derive a *stochastic upper bound* for the joint queue lengths that has the product–form distribution. Their approach can be adapted to our system of parallel $M/M/1/N$ queues and simultaneous arrivals to obtain a stochastic upper bound, denoted by $\widehat{\mathbf{IO}}$, for $\mathbf{IO}$. This, in turn, can be used to derive a lower bound for F^K, which is termed the *signal* lower bound.

In Appendix A, we provide a signal–queue formulation of the supply system of our model with independent geometric batch arrivals having parameters p_i, $i = 1, \dots, J$. The unit–batch case is obtained by taking $p_i \equiv 0$, $i = 1, \dots, J$. We can therefore show, using the results of Appendix A, that $\widehat{\mathbf{IO}}$ has the distribution

$$P(\widehat{\mathbf{IO}} = \mathbf{n}) = \prod_{i \in \Omega : \widehat{\rho}_i \neq 1} \frac{(1 - \widehat{\rho}_i)\widehat{\rho}_i^{\,n_i}}{1 - \widehat{\rho}_i^{\,N_i+1}} \prod_{i \in \Omega : \widehat{\rho}_i = 1} \frac{1}{N_i + 1}, \quad \text{where } \widehat{\rho}_i = \frac{\sum_{K : i \in K} \alpha_i^K}{\mu_i}, \tag{16}$$

and nonnegative numbers α_i^K, $i \in \Omega$, $K \subseteq \Omega$, are obtained by solving the *the traffic equations* (A.7), and also that $\mathbf{IO} \leq_{st} \widehat{\mathbf{IO}}$. Since stochastic order is closed under marginalization, we have $\mathbf{IO}^K \leq_{st} \widehat{\mathbf{IO}}^K$. From (7), this further implies $\mathbf{IO}^K \leq_{lo} \widehat{\mathbf{IO}}^K$. Thus we obtain the *signal* lower bound for F^K:

$$F^K = P(\mathbf{IO}^K < \mathbf{s}^K) \geq P(\widehat{\mathbf{IO}}^K < \mathbf{s}^K) = \prod_{i \in \Omega : \widehat{\rho}_i \neq 1} \frac{1 - \widehat{\rho}_i^{\,s_i}}{1 - \widehat{\rho}_i^{\,N_i+1}} \prod_{i \in \Omega : \widehat{\rho}_i = 1} \frac{s_i}{N_i + 1}. \tag{17}$$

Note that if $N_i = \infty$, $\widehat{\mathbf{IO}}_i$ is well defined only when $\widehat{\rho}_i < 1$.

Example 3.7. Consider a two-component system with $\lambda^1 = 3$, $\lambda^2 = 6$, $\lambda^{12} = 3$, $\mu_1 = 12$ and $\mu_2 = 20$. Also let $s_i = 4$ and $N_i = \infty$, $i = 1, 2$. Let us first compute the signal lower bound for type-12 order. By (A.7),

$$\begin{aligned}\alpha_1^1 &= \lambda^1 = 3,\ \alpha_1^{12} = \lambda^{12} = 3 \text{ and } \widehat{\rho}_1 = \frac{\lambda^1 + \lambda^{12}}{\mu_1} = 0.5 \\ \alpha_2^2 &= \lambda^2 = 6,\ \alpha_2^{12} = \frac{\alpha_1^{12}}{\widehat{\rho}_1} = \frac{3}{0.5} = 6 \text{ and } \widehat{\rho}_2 = \frac{\alpha_2^2 + \alpha_2^{12}}{\mu_2} = \frac{6+6}{20} = 0.6\end{aligned}$$

This yields the signal lower bound of F^{12} as $(1-(0.5)^4)(1-(0.6)^4) = 0.816$. Note that the first–order setwise lower bound of F^{12} is $(1-(0.5)^4)(1-(0.45)^4) = 0.899$, which is significantly better than the signal lower bound. It can be computed that the setwise upper bound (see Section 4.2) of F^{12} is $(1-(0.5)^4)(1-(0.3)^4) = 0.930$. The closeness of the setwise lower and upper bounds indicates the high quality of the setwise bounds. Furthermore, if the arrival rate of the type–2 order were 15 instead of 6, then $\widehat{\rho}_2 > 1$ and the signal bound no longer exists. The setwise lower bound in this case is $(1-(0.5)^4)(1-(0.9)^4) = 0.322$. ■

This example illustrates that the signal lower bound may degenerate. It also raises the following question: What is the relationship between a setwise lower bound given by (10) and the signal lower bound given by (17)? We obtain the following result:

Theorem 3.8. Let $\mathbf{IO}^I = (IO_1^I, \ldots, IO_J^I)$ consist of independent components with $IO_i =_{st} IO_i^I$. That is, for $\lambda_i = \sum_{K:i\in K} \lambda^K$ and $\rho_i = \lambda_i/\mu_i$,

$$P(\mathbf{IO}^I \le \mathbf{n}) = \prod_{i\in\Omega,\rho_i\neq 1} \frac{1-\rho_i^{n_i}}{1-\rho_i^{N_i}} \prod_{i\in\Omega,\rho_i=1} \frac{n_i}{N_i+1},$$

Also let $\widehat{\mathbf{IO}}$ have the distribution specified by the RHS of (16). Then

$$\mathbf{IO} \le_{lo} \mathbf{IO}^I \le_{st} \widehat{\mathbf{IO}}. \tag{18}$$

Proof. From (8) we know that $\mathbf{IO}$ is more positively lower orthant dependent than $\mathbf{IO}^I$. Thus it is sufficient to show that $\widehat{\mathbf{IO}}$ stochastically dominates $\mathbf{IO}^I$. We first show, by induction, that $\alpha_i^K \ge \lambda^K$, for any $i \in K = \{i_1^K, \ldots, i_{|K|}^K\}$. From (A.7), $\alpha_{i_1^K}^K = \lambda^K$. Now assume that $\alpha_{i_{\ell-1}^K}^K \ge \lambda^K$. Then

$$\alpha_{i_\ell^K}^K = \begin{cases} \alpha_{i_{\ell-1}^K}^K\left(1+\frac{1}{\widehat{\rho}_{i_{\ell-1}^K}}\right) \ge \alpha_{i_{\ell-1}^K}^K \ge \lambda^K, & \text{if } \ell = 2,\ldots,|K|,\ N_\ell < \infty \\ \frac{\alpha_{i_{\ell-1}^K}^K}{\widehat{\rho}_{i_{\ell-1}^K}} \ge \alpha_{i_{\ell-1}^K}^K \ge \lambda^K, & \text{if } \ell = 2,\ldots,|K|,\ N_\ell = \infty \end{cases}$$

Since the above equation is true for every $K \subseteq \Omega$, $\sum_{K:i\in K} \alpha_i^K \ge \sum_{K:i\in K} \lambda^K = \lambda_i$. Now, a sample path coupling argument can be used to show that, in two parallel $M/M/1/N$ queueing systems, if the arrival rate to each queue in one system is larger than that in another system, with all other parameters identical, then the stationary queue length vector in the former system stochastically dominates its counterpart in the latter system. ■

The theorem implies that the first–order setwise bound, developed via the study of the dependent structure of parallel $M/M/1/N$ queues with simultaneous arrivals, is *tighter* than the signal bound. Since a higher order setwise lower bound is always better than the first–order setwise lower bound, we conclude that a setwise lower bound of any order is stochastically bigger than the signal lower bound. It is also worth mentioning that, when $N_i = \infty$, $i \in \Omega$, the signal bound is exactly the same asymptotic bound obtained by Schwartz and Weiss (1993) via the asymptotic analysis using large deviations.

4 UPPER BOUNDS

This section introduces two classes of upper bounds, Frechet–type upper bounds and setwise upper bounds.

4.1 Frechet Upper Bounds

The Frechet–type upper bounds are distribution–free bounds. Let $G_i,\ i \in K$, be given univariate distribution functions. Let $\mathcal{G}(G_i,\ i \in K)$ be the class (called the Frechet class) of $|K|$–variate distributions with given marginals $G_i,\ i \in K$. The first–order Frechet upper bound of $\mathcal{G}(G_i,\ i \in K)$ is given by

$$G(\mathbf{x}^K) \leq \min_{i \in K} G_i(x_i), \text{ for any } G \in \mathcal{G}(G_i,\ i \in K),\ \mathbf{x}^K = (x_i : i \in K).$$

It is known that $\min_{i \in K} F_i(x_i)$ is a proper $|K|$–variate distribution function (Joe 1997). Notice that the above bound only requires the information of the univariate marginal distribution functions. For this reason, we call it the first–order Frechet upper bound.

Using this result, the first–order Frechet upper bound for the type–K order fill rate satisfies

$$F^K \leq \min_{i \in K} F^i. \tag{19}$$

(Song (1998) obtains the same bound for systems with deterministic leadtimes.) The bound is rather intuitive, it states that an order fill rate cannot be greater than the fill rate of its each individual component. The first–order Frechet upper bound is expected to perform satisfactorily under high correlation and high component fill rates. For example, for a two-component system with 90% fill rate for each component, the maximal percentage error of the first order Frechet upper bound is

$$\frac{\min\{F^1, F^2\} - F^{12}}{F^{12}} \leq \frac{\min\{F^1, F^2\} - F^1F^2}{F^1F^2} = \frac{0.9 - (0.9)^2}{(0.9)^2} = 11\%$$

Similarly, we can assess the maximal percentage error of the first–order setwise lower bounds, which is expected to perform well under lower correlation:

$$\frac{F^{12} - F^1F^2}{F^{12}} \leq 1 - F^1 = 10\%.$$

The above expressions imply that one can expect that both Frechet upper bound and setwise lower bound perform relatively well with high component fill rates, regardless of the degree of correlation. This perhaps explains why the independent approximation used in Glasserman and Wang works well under high fill rate.

Now let $\mathcal{G}(G^M, M \subseteq K,\ |M| = m)$ be the Frechet class with given m–variate marginals G^M. Then

$$G(\mathbf{x}^K) \leq \min_{M\subseteq K} G^M,\ \text{for any } G \in \mathcal{G}(G^M,\ M \subseteq K,\ |M| = m).$$

The above bound attempts to use an m–variate marginal distribution to bound from the above the $|K|$–variate distribution. This result leads to the mth–order Frechet upper bound for the type–K order-fill rate:

$$F^K \leq \min_{M\subseteq K, |M|=m} F^M.$$

For $|K| = 3$, the second–order Frechet upper bound of $F^{i_1i_2i_3}$ can be strengthened as follows:

$$F^{i_1i_2i_3} \leq \min\{F^{i_1i_2}, F^{i_1i_3}, F^{i_2i_3}, a_4\},$$

where $a_4 = 1 - F^{i_1} - F^{i_2} - F^{i_3} + F^{i_1i_2} + F^{i_1i_3} + F^{i_2i_3}$, comes from the identity $\overline{F}^{i_1i_2i_3} = 1 - F^{i_1} - F^{i_2} - F^{i_3} + F^{i_1i_2} + F^{i_1i_3} + F^{i_2i_3} - F$; since $\overline{F}^{i_1i_2i_3} \geq 0$, $F^{i_1i_2i_3} \leq a_4$.

Note that in order to obtain the mth–order Frechet upper bound, one needs to evaluate F^M for all $M \subseteq K$ and $|M| = m$ (assuming the queues are not symmetric). This task becomes formidable when the number of distribution functions to be evaluated becomes large. As a possible approximation of the mth order Frechet upper bound, we propose the following approximate mth order Frechet upper bound, which only needs to evaluate a *single* m–variate joint distribution function. In the algorithm, we attempt to sequentially select m components that have the smallest component-fill rates; in the case that several candidate components have identical component fill rates, we choose the component that has the weakest correlation with the previously selected component. This is because an order fill rate is bounded above by these low component fill rates; furthermore, due to the lower orthant property of **IO**, the fill rate of a less correlated pair of components is smaller than the fill rate of a more correlated pair of components, given that the corresponding marginals of the two pairs are identical (see Xu (1999), for a detailed account on how correlated arrival processes lead improved system performances.)

Algorithm 4.1. The Approximate mth-Order Frechet Upper Bound, $m \geq 2$

1. Compute F^i, $i \in K$, and $\widehat{\lambda}^{ij} = \sum_{i,j \in M} \lambda^M, \quad i, j \in K$. Also set $n = 1$, $M = \emptyset$ and $K(n) = K$.

2. Let $U \triangleq \text{argmin}_{i \in K(1)}\{F^i\}$. If $U = i(1)$ is a singleton, let $M \cup i(1) \to M$, $K(2) = K(1) - i(1)$, $n = 2$ and go to Step 3. If U is not a singleton, let $\widehat{\lambda}^{i(1),i(2)} = \min_{i,j \in U}\{\widehat{\lambda}^{i,j}\}$, where ties are broken arbitrarily. Let $M \cup \{i(1), i(2)\} \to M$. If $m = 2$, go to Step 4; otherwise let $K(3) = K(1) - i(1) - i(2)$, $n = 3$ and go to Step 3.

3. Let $U \triangleq \text{argmin}_{i \in K(n)}\{F^i\}$. If $U = i(n)$ is a singleton, let $M \cup i(n) \to M$. If $n = m$, go to Step 4, otherwise set $K(n+1) = K(n) - i(n)$, $n \to n+1$ and repeat Step 3. If U is not a singleton, let $\widehat{\lambda}^{i(n-1),i(n)} = \min_{j \in U}\{\widehat{\lambda}^{i(n-1)j}\}$. Again, ties are broken arbitrarily. Let $M \cup i(n) \to M$. If $n = m$, go to Step 4; otherwise let $K(n+1) = K(n) - i(n)$, $n \to n+1$ and repeat Step 3.

4. Compute the type-M fill rate F^M, which is an approximate mth order Frechet upper bound.

4.2 Setwise Upper Bounds

The *setwise upper bounds* are aimed at improving the Frechet upper bounds without substantially increasing computational complexity. Here is the idea: Let $F^{M^*(1)}$ be the mth–order Frechet upper bound of F^K, that is, $F^{M^*(1)} = \min_{M \subseteq K}\{F^M : \quad M \subseteq K, \ |M| = m\})$. For any $M \subseteq \Omega$, $M \cap M^*(1) \neq \emptyset$ and $M \cap (K - M^*(1)) \neq \emptyset$, we treat such a type-$M$ order as a type $M \cap M^*(1)$ order, that is, we accept component i, whenever possible, if $i \in M \cap M^*(1)$ and discard the component otherwise. Clearly, the demand process for components in $M^*(1)$ is unchanged, but the new demand process for components in $K - M^*(1)$ becomes *stochastically smaller* (which can be shown easily using a stochastic coupling argument), and the two demand processes are *independent.* As a result, the number of outstanding orders of components in $K - M^*(1)$ is also stochastically smaller. The new demand process for the components in $K - M^*(1)$ is Poisson with the adjusted rates:

$$\widehat{\lambda}^M = \sum_{M \subseteq L' \subseteq \Omega - M^*(1)} \lambda^{L'}, \qquad M \subseteq K - M^*(1). \tag{20}$$

Let $F^{K-M^*(1)}(2)$ be the fill rate of type $K - M^*(1)$ order corresponding to the new arrival process, with all other parameters kept the same as in the original system. Then we obtain:

$$F^K \leq F^{M^*(1)} F^{K-M^*(1)}(2). \tag{21}$$

We can repeat the above procedure, and develop an upper bound for $F^{K-M^*(1)}(2)$, if $|K - M^*(1)| > m$. The algorithm below gives the mth–order setwise upper bound for F^K, $K \subseteq \Omega$ and $|K| \geq 2$.

Algorithm 4.2. The mth–order Setwise Upper Bound of F^K, $|K| \geq 2$

1. Set $n = 1$, $K(n) = K$, $\Omega(n) = \Omega$.

2. For each $M \subseteq K(n)$, let

$$\lambda^M(n) = \sum_{M \subseteq L' \subseteq \Omega(n)} \lambda^{L'} \quad \text{and} \quad \lambda_i(n) = \sum_{i \in M \subseteq K(n)} \lambda^M, \quad i \in K(n).$$

   Let $\rho_i(n) = \lambda_i(n)/\mu_i$. If $m = 1$, for each $i \in K(n)$, compute

$$F^i(n) = \left\{ \begin{array}{ll} \dfrac{1 - (\rho_i(n))^{s_i}}{1 - (\rho_i(n))^{N_i+1}}, & \text{if } \rho_i(n) \neq 1 \\ \dfrac{s_i}{N_i + 1} & \text{if } \rho_i(n) = 1 \end{array} \right\}.$$

   If $m > 1$, apply the matrix–geometric solution developed by Song, Xu and Liu to compute $F^M(n)$ for each $|M| = m$ and $M \subseteq K(n)$, using the new arrival process with arrival rate $\lambda^M(n)$.

3. Let

$$M^*(n) = \text{argmin}\{F^M(n) : \ M \subseteq K(n), \ |M| = m\}.$$

   Ties are broken arbitrarily. Set $K(n+1) = K(n) - M^*(n)$ and $\Omega(n+1) = \Omega(n) - M^*(n)$. If $|K(n+1)| \leq m$, let $M^*(n+1) = K(n+1)$ and $F^{M^*(n+1)}(n+1) = F^{M^*(n+1)}(n)$ and go to Step 4; otherwise go to Step 2.

4. Compute the setwise upper bound as:

$$F^K \leq \prod_{n=1}^{\lceil \frac{K}{m} \rceil} F^{M^*(n)}(n), \tag{22}$$

   where $\lceil a \rceil$ is the smallest integer greater than or equal to a. Stop.

Because $F^{M^*(1)}(1) = \min\{F^M : \ M \subseteq K, |M| = m\}$ is the mth order Frechet upper bound, the next claim follows (22) immediately.

Theorem 4.3. The mth order setwise upper bound is tighter than the Frechet upper bound of the same order.

Since the setwise upper bound yields the exact solution under independence, we expect the improvement of the setwise upper bound over the Frechet upper bound of the same order to be significant when demand correlation is low. Also, if computing the *exact* mth order Frechet upper bound turns out to be impractical, Algorithm 4.1 may be inserted in Step 2 of Algorithm 4.2 to obtain the approximate mth order setwise upper bound.

We now illustrate the above algorithm via an example.

Example 4.4. Consider a four-component system with the following parameters:

$$\lambda^{123} = 1,\ \lambda^{134} = 1, \lambda^{12} = 2, \lambda^{24} = 1, \lambda^{1} = 5, \lambda^{2} = 2, \lambda^{3} = 5, \text{ and } \lambda^{K} = 0 \text{ otherwise;}$$
$$\mu_i = 10,\ s_i = 4 \text{ and } N_i = \infty,\ i = 1, 2, 3, 4.$$

We wish to use Algorithm 4.2 to compute the first–order setwise upper bound of F^{123}.

In Step 1, we set $n = 1$, $K(1) = K = \{1, 2, 3\}$, $\Omega(1) = \Omega = \{1, 2, 3, 4\}$.

In Step 2, we set $\lambda^M(1) = \lambda^M,\ M \subseteq \Omega(1)$. We also obtain

$$\lambda_1(1) = \sum_{1 \in M \subseteq \Omega^{(1)}} \lambda^M = \lambda^{123} + \lambda^{134} + \lambda^{12} + \lambda^{1} = 9 \text{ and } \rho_1(1) = 0.9.$$

Similarly, $\lambda_2(1) = 5,\ \rho_2(1) = 0.5,\ \lambda_3(1) = 7$ and $\rho_3(1) = 0.7$. Hence the component fill rates satisfy:

$$F^1(1) = F^1 = 1 - (0.9)^4,\ F^2(1) = F^2 = 1 - (0.5)^4,\ F^3(1) = F^3 = 1 - (0.7)^4.$$

In Step 3, we obtain $\operatorname{argmin}\{F^i(1), i = 1, 2, 3\} = 1$. Set $K(2) = K(1) - \{1\} = \{2, 3\}$ and $\Omega(2) = \Omega(1) - \{1\} = \{2, 3, 4\}$. Since $|K(2)| = 2 > m = 1$, we return to Step 2.

In Step 2, we compute

$$\begin{aligned} \lambda_2(2) &= \sum_{2 \in M \subseteq \Omega(2)} \lambda^M = \lambda^2 + \lambda^{24} = 3, \quad \rho_2(2) = 0.3, \\ \lambda_3(2) &= \sum_{3 \in M \subset \Omega(2)} \lambda^M = \lambda^3 = 5, \quad \rho_3(2) = 0.5. \end{aligned}$$

This gives $F^2(2) = 1 - (0.3)^4$ and $F^3(2) = 1 - (0.5)^4$. Go to Step 3.

In Step 3, we obtain $\operatorname{argmin}\{F^i(2), i = 2, 3\} = 3$. Set $K(3) = K(2) - \{3\} = \{2\}$ and $\Omega(3) = \Omega(2) - \{3\} = \{2, 4\}$. Since $|K(3)| = 1$, set $M^*(3) = K(3) = 2$ and $F^2(3) = F^2(2) = 1 - (0.3)^4$ and go to Step 4.

In Step 4, we compute the first order setwise upper bound as

$$F^{123} \leq F^1(1)F^3(2)F^2(3) = (1 - (0.9)^4)(1 - (0.5)^4)(1 - (0.3)^4) = 0.3198.$$

and then stop. The corresponding Frechet upper bound is $1 - (0.9)^4 = 0.3439$. The first order setwise *lower bound* is $F^K \geq F^1F^2F^3 = 0.245$. ■

Example 4.5. Suppose the interarrival time between two orders is exponentially distributed with rate λ and batch size vector $\mathbf{X}$ has the distribution

$$P(\mathbf{X}^K = \mathbf{x}^K, \mathbf{X}^{\Omega - K} = \mathbf{0}) = q^K \prod_{i \in K} (1 - p_i) p_i^{x_i - 1}, \quad \mathbf{x}^K \geq \mathbf{1},\ K \subseteq \Omega. \tag{23}$$

This assumption implies that, with probability q^K, an order will bring a geometric batch X_i to facility i for each $i \in K$, $K \in \Omega$. Assume $N_i = \infty$, $i \in \Omega$. Then, the first–order setwise lower bound is

$$F^K \geq \prod_{i \in K} P(IO_i^a + X_i \leq s_i) = \prod_{i \in K} \sum_{x_i=1}^{s_i} P(IO_i^a \leq s_i - x_i)(1-p_i)p_i^{x_i - 1}.$$

Let

$$\rho_i = \frac{1}{(1-p_i)\mu_i} \sum_{K: i \in K} \lambda q^K = \frac{\lambda_i}{(1-p_i)\mu_i} < 1,$$

then the probability mass of IO_i^a is given by (see, for example, Gross and Harris 1974)

$$\begin{aligned} P(IO_i^a = 0) &= 1 - \rho_i \\ P(IO_i^a = n) &= (1-\rho_i)[p_i + (1-p_i)\rho_i]^{n-1}[(1-p_i)\rho_i], \quad n > 0,\ i \in \Omega. \qquad \blacksquare \end{aligned}$$

5 NUMERICAL RESULTS

In this section we conduct numerical comparisons of the bounds developed in the previous sections under various system configurations. Both backlogging and lost–sales cases are considered. Subsections 5.1-5.3 examine how product variety, facility utilization and component base-stock levels affect bounds on order performance measures in systems with three, six and nine components, respectively. Numerical results are tabulated and gathered in Appendix B. In all these examples, we assume the Markovian property.

In Tables 4–6, we report percentage errors for various lower and upper bounds on the immediate system-based fill-rate F for various production systems. Those tables also contain the exact immediate overall fill rate and the overall order serviceability for comparison.

Tables 1–3 describe the parameter choices for the production systems. The parameters of the production systems are designed such that the service levels are the same for both partial backlogging and lost-sales cases. Since the fill rates and service levels are the same in a lost-sales case, the percentage errors of various bounds on the service levels in the partial backlogging cases are the same as those of immediate fill rates in the lost-sales cases. The percentage error for a bound is calculated by [(bound − exact quantity)/exact quantity] × 100%. As such, the percentage error for a lower bound is negative and an upper bound positive.

The legends used in the tables are as follows: “U” = Utilization (“h” = high, “l” = low), “C” = Correlation (“h” = high, “l” = low), “P” = Policy (“b” = backlogging, “l” = lost–sales), “Fill Rate” = Exact overall immediate order fill rate, “Serv’blty” = Exact overall serviceability, “Signal” = Signal Bound, “Bf” = Bonferroni, “CMB” = Combination, “StL” = Setwise Lower, “EStL” = Exact Setwise Lower, “GStL” = Greedy Setwise Lower, “StU” = Setwise Upper, “Frcht” = Frechet.

5.1 Three Component Systems

We start with a small system with three components. We consider two levels of product variety and call them *Order Profile 1* and *Order Profile 2*. e believe that the quality of bounds is mainly affected by traffic intensity and correlation level of product demand. Therefore, for each order profile, we consider both the backlogging and lost–sale cases and also vary machine utilization and demand correlation from high to low. This gives us totally sixteen cases. For both order profiles we assume $\lambda_1 + \lambda_2 + \lambda_3 = 8.1$. Other parameters are specified in Table 1.

Table 4 summarizes the test results of the sixteen cases. From the table, we observe that the first–order setwise lower bound (StL1) outperforms all other first–order bounds. The first–order combination lower bound (CMB1) does better than Bonferroni lower bound (Bf1) on average, especially when Bf1 is likely to degenerate, which often occurs for the system with high utilization, high demand correlation and backlogging. It also appears that the signal lower bound cannot compete with other bounds.

The second-order bounds perform better than the first–order bounds, as expected, and they are quite satisfactory on average. There are no significant differences between the average performance of the second–order bounds, partly because a three–component system is not large enough to observe the differences. Also, it appears that for the small system the first order bound gives the reliable prediction of the exact fill rate.

The results indicate that the quality of the first– and second–order lower bounds improves as the fill rate increases or demand correlation decreases. With the high fill rate, all lower bounds, except for signal bounds, perform extremely well, regardless of the correlation level. However, with the low fill rate, their performance is sensitive to demand correlation.

The setwise upper bounds perform better than the Frechet upper bounds, but only marginally. Note that the second–order upper bound significantly improves its first–order counterpart.

Overall, it appears that for a fixed fill rate, the lower bound performs better than the upper bound of the same order.

5.2 Six Component Systems: The PC Examples

We next consider a six-component PC assembly system under different parameter settings. We consider both the unit demand and geometric batch demand cases. In each case, the total arrival rate of orders equals $\lambda = 2$. We again examine the backlogging and lost–sale cases and vary utilization from high to low. This results in eight different system configurations, with system parameters specified in Table 2. The base-stock levels and production queue capacities are chosen in such a way that the overall serviceability of orders are at least as high as 0.90. Table 5 summarizes the results.

We again observe superior performances of the setwise lower and upper bounds compared against other bounds. The combination lower bound performs better than the Bonferroni lower bound. Notice that the difference between the first order and second order bounds becomes more pronounced for this mid-sized problem. Thus, when the problem size grows, it becomes necessary to compute the second order bound when the fill rate is low and demand correlation is high.

The numerical results also show the greedy second–order setwise lower bound (GStL2), computed by Algorithm 3.3, performs extremely well, as they give identical results compared against the exact second–order setwise lower bound (EStL2) using (12).

The setwise upper bound performs slightly better than the Frechet upper bound. In general, the upper bound is of high quality only when the fill rate is very high.

5.3 Nine Component Systems

We now present the experiments for a nine-component system with different order profiles. Under both scenarios, $\lambda_1 + \lambda_2 + \cdots + \lambda_9 = 37.8$. Other parameter selections are summarized in Table 3. Note that production capacity N_i is chosen so that the overall serviceability of component i is at least 0.85, for $i = 1, \ldots, 9$. Table 6 summarizes our findings.

The signal lower bound is not satisfactory in general. Degeneracy of Bonferroni lower bounds occurs quite frequently for this large system. In such a case, the combination bound does better than the Bonferroni bound. This is mainly because the combination bound uses the Bonferroni bounds on smaller subsets of components, which reduce the possibility of degeneracy.

The second–order lower bounds are more accurate than the first–order lower bounds (with the exception that first order setwise lower bound does better than the second order Bonferroni bound if the latter degenerates). The second–order exact setwise lower bound is the best among all the lower bounds. The greedy setwise lower bound is quite close to the exact setwise lower bound. It is

also worth noting that the second–order combination bound does almost as good as the second order setwise bounds in some cases.

The setwise upper bounds are better than the Frechet bounds. The difference is more pronounced in the first–order case.

6 Extensions and Connections with Uncapacitated ATO Systems

To what extent the bounds developed in this paper for the basic capacitated ATO system, modeled as a set of $M/M/1/N$ queues, can be generalized to other types of ATO systems in which demand

and service processes follow general distributions? We answer this question for both the capacitated system (with the supply system for each component modeled as a single server queue) and the uncapacitated system (with the supply system for each component modeled as an infinite-server queue or a finite-server loss system).

Consider the setwise bounds first. For the capacitated ATO system, the setwise lower bounds for $M/M/1/N$ queues, $N \leq \infty$, are rested on a fundamental property of a multivariate Poisson process: It is known that a multivariate Poisson process generated by exponential interarrival times and parameter set $\{q^K, K \subseteq \Omega\}$ is positively quadrant dependent (PQD), regardless of the distribution of $\{q^K, K \subseteq \Omega\}$. In addition, the QPD property of a multivariate Poisson process is inherited by the inventory on order vector **IO**. It can be shown that the above conclusion still holds even if the service time of each queue follows a general distribution. As a result, the setwise lower bound is still valid for the $M/G/1$ type of ATO systems (in fact, it is also true for the $M^X/G/1$ type of ATO systems, where X is the batch size).

However, if we move away from the multivariate (compound) Poisson demand process, then the dependence structure of the demand process, and henceforth **IO**, is no longer independent of the parameter set $\{q^K, K \subseteq \Omega\}$. Indeed, counterexamples exist that show that simultaneous demand may even result in negatively dependent inventory on order vector **IO**, for certain multivariate renewal processes. In order to derive the setwise lower bound for the $G/G/1$ type of ATO systems, we require an additional condition that the demand type indicator vector $\mathbf{X} = (X_1, ..., X_J)$, whose distribution is given by $\{q^K, K \subseteq \Omega\}$ (i.e. $P(\mathbf{X}^K = 1, \mathbf{X}^{\Omega-K} = 0) = q^K$, $K \subseteq \Omega$), be more PQD than $\mathbf{X}^I$, where $\mathbf{X}^I$ is the independent counterpart of $\mathbf{X}$. Under this condition, Xu (2002, Section 5.3) shows that the setwise lower bound holds for the $G/G/1$ type of ATO systems. The performance bounds for the ATO system with batch demands can be found in Li and Xu (2001) and Xu (2002, Section 5.2).

As mentioned earlier, the setwise lower bounds have been established in the literature for the uncapacitated, $M/G/\infty$ type ATO systems. Again, these results rely on the properties of multivariate Poisson processes. If we move away from the multivariate Poisson demand process, then two complications arise. First, as in the capacitated system, the multivariate non-Poisson renewal process specified by general inter-arrival time distribution G and the parameter set $\{q^K, K \subseteq \Omega\}$ may not be PQD. Consequently, **IO** may not be PQD. Second, the i.i.d./ leadtimes may cause orders to cross over, that is, the orders for each component may be received in a different sequence as they are placed. This makes it difficult to have the sample-path representation of the **IO** process, an essential step to carry out the dependence analysis of a performance vector. However, the following partial results are known: If $\mathbf{X}$ is more PQD than $\mathbf{X}^I$, then the setwise bounds exist for the $G^X/M/N/N$, $N \leq \infty$, and $G/D/\infty$ types of uncapacitated ATO systems (see Xu 2002, Section

6).

Besides the setwise lower bounds, other types of bounds developed in this paper can also be generalized to various types of capacitated and uncapacitated systems. The Bonferroni-type lower bounds and the Frechet upper bounds are distribution free and therefore can be used for systems with uncapacitated leadtimes. The setwise upper bounds in Section 4.2 can also apply to various capacitated and uncapacitated leadtimes by adjusting the arguments, particularly for the systems mentioned above. However those results have not appeared in the literature. The combination lower and upper bounds and various heuristics for improved bounds can also be adapted to systems with capacitated and uncapacitated leadtimes. The signal lower bound relies on the Markovian assumption and can only be derived, similar as in Section 3.4, for the $M^X/M/1/N$ system. As shown in Section 4.2, the setwise upper bound is tighter than the first-order Frechet upper bound.

The evaluations of these bounds, on the other hand, are non-trivial. This is because, unlike the $M/M/1/N$ type of ATO systems which can be evaluated via the matrix-geometric solution, most of non-Markovian queues defy exact solutions, even for single-server queues (e.g., the $G/G/1$ queue). However, our analysis indicates that, assuming good bounds are available for individual queues, the first-order lower bound can produce an adequate surrogate for the fill-rate of a non-Markovian ATO system, especially under high fill-rate.

The applicability of the first-order setwise bounds to both the capacitated and uncapacitated systems allows us to gain insights into how different supply systems affect ATO performance measures. For example, consider two ATO systems having the same arrival process and the same component base-stock levels s_i. Since only first-order bound is considered, we assume there is no limit on the backorder queue size, i.e., $b_i = \infty$. In the first system, the production facility of component i is a FCFS exponential server with mean $1/\mu_i$. Thus, the supply system for component i is an M/M/1 queue. As a result,

$$F^i = P(I_i > 0) = P(IO_i < s_i) = 1 - \rho_i^{s_i}.$$

So, the system fill rate (fill rate of an arbitrary order)

$$F \approx \sum_K q^K \prod_{i \in K} (1 - \rho_i^{s_i}). \tag{24}$$

In the second system, each order of component i experiences a leadtime with mean $1/\mu_i$ and a general distribution. Thus, the supply system for component i is an M/G/∞ queue. Use a tilde to distinguish the performance measures of this system, we have

$$\tilde{F}^i = P(\tilde{I}_i > 0) = P(\tilde{I}O_i < s_i) = e^{-\rho_i} \sum_{k=0}^{s_i - 1} \frac{\rho_i^k}{k!},$$

and

$$\tilde{F} \approx \sum_K q^K \prod_{i \in K} (\sum_{k=0}^{s_i-1} \frac{\rho_i^k}{k!}) e^{-\rho_i}. \tag{25}$$

Note that

$$e^{-\rho_i} \sum_{k=s_i}^{\infty} \frac{\rho_i^k}{k!} = e^{-\rho_i} \rho_i^{s_i} \sum_{k=0}^{\infty} \frac{\rho_i^k}{(k+s_i)!} \leq e^{-\rho_i} \rho_i^{s_i} \sum_{k=0}^{\infty} \frac{\rho_i^k}{k!} = \rho_i^{s_i}.$$

So $\tilde{F^i} \geq F^i$ for all i, and consequently $F \leq \tilde{F}$. That is, other things being equal, the fill-rate of the uncapacitated system is higher than that of the capacitated system.

As illustrated above, using the first-order approximation, the comparison of the two types of ATO systems reduces to the comparison of the marginal distributions. Thus, up to this approximation, the qualitative insights one gains from the single-component systems on the effects of the supply systems would carry through to ATO systems. As shown in Chapter 7 of Zipkin (2000), the difference between the two single-component systems is the same as that between a single-server queue and an infinite-server queue. In the first, the asymptotic behavior of the system is exponential, but in the second, it is normal. Also, the uncapacitated system can handle increases in load much more easily. For an uncapacitated system, the cost is proportional to the square root of the mean demand rate, but in a capacitated system, the cost is convex in the mean demand rate. Because of the multiplications of the marginal distributions, we expect that these effects are not less in the ATO systems.

The applicability of various bounds on the two types of ATO systems also suggests that the approximation-based optimization techniques developed for ATO systems with uncapacitated leadtimes are likely to be applicable in the capacitated systems, or vice versa. Indeed, it can be shown that, if we only consider one final product, then the greedy algorithm developed in Song and Yao (2002) for the problem of minimizing inventory holding cost subject to a fill rate constraint in the uncapacitated system can be applied here. Similarly, the techniques developed in Cheng et al. (2002), Lu and Song (2002) and Lu et al. (2003b) for the uncapacitated systems can be adapted to the corresponding capacitated systems as well.

7 Concluding Remarks

7.1 Summary

We provided an in-depth comparison of the effectiveness of several bounding ideas to estimate the order fill rates in capacitated assemble–to–order systems. In particular, we considered four lower bounds, including the setwise lower bound based on dependence structure of outstanding orders, the distribution–free Bonferroni bounds, the combination bound aimed at using a setwise partition of a high–dimensional joint distribution to deal with the degeneracy of the Bonferroni bound, and the signal bound originated from the study of the quasi–reversibility of queueing networks. We also considered two upper bounds: The Frechet upper bound that uses the distribution of a subvector to bound the original distribution, and the setwise upper bound based on the refinement of the Frechet upper bound. We introduced several algorithms to improve the computational efficiency in calculating the bounds.

We showed analytically that the setwise bound of any order is tighter than the signal bound. Numerical tests also indicated that the performance of the signal bound is very disappointing. The Bonferroni bound provides satisfactory performance only for the system with few components or with high fill rate, otherwise its performance degrades caused by degeneracy. The combination bound outperforms the Bonferroni bound as the setwise partition approach reduces the likelihood of degeneracy of the Bonferroni bound. The setwise bound in most cases performs the best compared against all other bounds. Generally, we found that the performances of all lower bounds, except for the signal bound, are satisfactory for the system with a moderate fill rate (e.g., greater than 80%).

We also showed analytically that the setwise upper bound is tighter than the Frechet upper bound of the same order. However, numerical results showed that the difference is negligible. Overall, the quality of upper bounds is not as good as that of lower bounds and it is a reliable indicator of the performance measure only when the fill rate is very high (e.g., greater than 95%).

Finally, we discussed possible extensions of these bounds to systems with non-Markovian features. We also made connections with the uncapacitated ATO systems, which allowed us to gain qualitative insights with regard to the effect of different supply systems on ATO system performance.

Acknowledgment: This research was supported in part by NSF grants DMI-9700037, DMI-9812994, DMI-9896339 and DMI-0084922. The authors thank the Senior Editor and the referees for their insightful comments.

APPENDIX

A A Signal Queue Approach

Consider the supply system in our basic model with batch arrivals, with the batch size distribution

$$P(\mathbf{X}^K = \mathbf{x}^K, \mathbf{X}^{\Omega-K} = \mathbf{0}) = q^K \prod_{i\in K}(1-p_i)p_i^{x_i-1}, \quad \mathbf{x}^K \geq \mathbf{1},\ K \subseteq \Omega,$$

for some $0 \leq p_i < 1$, $i = 1, \ldots, J$ (allowing $p_i = 0$ will take us back to the unit–batch case). In other words, with probability q^K, an order brings a geometric batch X_i to facility i for each $i \in K$, $K \in \Omega$.

This supply system can be formulated as a queueing network with a single class of customers denoted by c and $2^J - 1$ classes of positive signals denoted by K, $K \subseteq \Omega$. Node i has a single server with service time exponentially distributed with rate μ_i and a finite buffer with capacity N_i, $i \in \Omega$. Suppose $K = (i_1^K, \ldots, i_{|K|}^K)$. Class–$K$ positive signals arrive at node i_1^K from outside according to a Poisson process with rate λ^K. The effect of a class–K positive signal on node i_ℓ^K, $\ell = 1, \ldots, |K|$, is as follows: It adds $\min(X_i, N_i - n_i)$ class–c customers to node i_ℓ^K, if, upon arrival, the signal finds $0 \leq n_i \leq N_i$ customers at node i, and then leaves for node $i_{\ell+1}^K$ (with probability 1) as a class–K positive signal. Here and in the sequel, we denote $i_{|K|+1}^K = 0$, $K \subseteq \Omega$ and node 0 is the outside world. In this way, the arrival of a class K positive signal will simultaneously create an arrival at each node in K, $K \in \Omega$. We can describe the dynamics of node i by the following quantities: For every $n_i,\ n_i' = 0, \ldots, N_i$, $i \in \Omega$, and $u = c$ or K, $K \subseteq \Omega$, define

$$\begin{array}{rcl}
p_{iu}^A(n_i, n_i') & = & \text{The probability that a type–}u\text{ arrival to node }i\text{ changes its state from }n_i\text{ to }n_i', \\
q_{iu}^D(n_i, n_i') & = & \text{The transition rate of node }i\text{ from }n_i\text{ to }n_i'\text{ due to a type–}u\text{ departure,} \\
f_{iu,u'}(n_i, n_i') & = & \text{The probability that a type–}u\text{ arrival to node }i\text{ changes the state of the node} \\
 & & \text{from }n_i\text{ to }n_i'\text{ and immediately trigger a type-}u'\text{ departure.} \\
r_{iu,ju'} & = & \text{The probability that a type–}u\text{ departure from node }i\text{ joins node }j\text{ as a type-}u' \\
 & & \text{arrival.}
\end{array}$$

The above quantities becomes

$$p_{jK}^A(n_j, n_j') = \left\{ \begin{array}{ll} (1-p_j)p_j^{n_j - n_j' - 1}, & \text{if } 0 \leq n_j \leq N_j - 2 \text{ and } n_j + 1 \leq n_j' \leq N_j - 1 \\ p^{N_j - n_j - 1}, & \text{if } 0 \leq n_j \leq N_j - 1 \text{ and } n_j' = N_j \\ 1, & \text{if } n_j = n_j' = N_j \end{array} \right\},$$

$$f_{jK,K}(n_j, n_j') = \left\{ \begin{array}{ll} 1, & p_{jK}^A(n_j, n_j') > 0 \\ 0, & \text{otherwise} \end{array} \right\}, \quad r_{iK,jK} = \left\{ \begin{array}{ll} 1, & i = i_\ell^K,\ j = i_{\ell+1}^K, \ell = 1, \ldots, |K| \\ 0, & \text{otherwise} \end{array} \right\},$$

and $q_{ic}^D(n_i, n_i - 1) = \mu_i,\ 0 < n_i \leq N_i,\ i \in \Omega$. Any parameter that is not explicitly mentioned above is set to zero.

Let $\boldsymbol{\alpha}_i = \{\alpha_i^K, i \in K \subseteq \Omega\}$, and α_i^K be the arrival rate of type K positive signals to node i, $i \in K$. At this stage we treat α_i^K as a dummy variable and will determine it later by the traffic equations (A.5). If we let $\widehat{\rho}_j \triangleq \sum_{K:j\in K} \alpha_j^K/[(1-p_j)\mu_j]$, then the stationary distribution of the number of customers in node j will be given as

$$\pi_j^{\boldsymbol{\alpha}_j}(0) = \begin{cases} \dfrac{1-\widehat{\rho}_j}{1-\widehat{\rho}_j\Big[p_j+(1-p_j)\widehat{\rho}_j\Big]^{N_j}}, & \text{if } p_j+(1-p_j)\widehat{\rho}_j \neq 1, \\ \dfrac{1}{1+(1-p_j)\widehat{\rho}_j N_j}, & \text{if } p_j+(1-p_j)\widehat{\rho}_j = 1, \end{cases} \tag{A.1}$$

and, for every $\ell = 1, 2, \ldots$

$$\pi_j^{\boldsymbol{\alpha}_j}(\ell) = \begin{cases} \pi_j^{\boldsymbol{\alpha}_j}(0)\Big[p_j+(1-p_j)\widehat{\rho}_j\Big]^{\ell-1}(1-p_j)\widehat{\rho}_j, & \text{if } p_j+(1-p_j)\widehat{\rho}_j \neq 1, \\ \pi_j^{\boldsymbol{\alpha}_j}(0)(1-p_j)\widehat{\rho}_j, & \text{if } p_j+(1-p_j)\widehat{\rho}_j = 1. \end{cases} \tag{A.2}$$

To proceed, we need the following background, taken from Chao et al.

Definition A.1. (Quasi–reversibility of node i with signals) Let S_i be the state space of node i and T_i the collection of class types. If there exist two sets of nonnegative numbers $\{\alpha_i^u\}$ and $\{\beta_i^u\}$ such that

$$\sum_{n_i'\in S_i} q_{iu}^A(n_i, n_i') = \alpha_i^u, \quad n_i \in S_i,\ u \in T_i, \tag{A.3}$$

$$\sum_{n_i'\in S_i} \pi_i^{\boldsymbol{\alpha}_i}(n_i')\left[q_{iu}^D(n_i', n_i) + \sum_{u'\in T_i} q_{iu'}^A(n_i', n_i) f_{iu',u}(n_i', n_i)\right] = \beta_i^u \pi_i^{\boldsymbol{\alpha}_i}(n_i), \quad i \in S_i,\ u \in T_i, \tag{A.4}$$

then queue i with signals is said to be quasi–reversible with respect to $\{q_{iu}^A, f_{iu,u'} : u, u' \in T_i\}$ and $q_{iu}^D, u \in T_i\}$.

Theorem A.2. If each node i with signals, $i = 1, \ldots, J$, is quasi–reversible with $\boldsymbol{\alpha}_i = \{\alpha_i^u, u \in T_i\}$ that is the solution of the traffic equations

$$\alpha_i^u = \sum_j \sum_{u'\in T_j} \beta_i^u r_{ju',iu}, \quad i = 1, \ldots, J,\ u \in T_i, \tag{A.5}$$

then the queueing network with signals has the product–form stationary distribution

$$\pi(\mathbf{n}) = \prod_{i=1}^{J} \pi_i^{\boldsymbol{\alpha}_i}(n_i), \quad \mathbf{n} = (n_1, \ldots, n_J), \tag{A.6}$$

where $\pi_i^{\boldsymbol{\alpha}_i}$ is the stationary distribution of $q_i^{\boldsymbol{\alpha}_i}$, $i = 1, \ldots, J$.

It can be checked that the nodes of our network are not quasi–reversible. However, by introducing additional Poisson departures of class-c and class-K entities, we can obtain a quasi–reversible network. More precisely, if we let $q^D_{jc}(0,0) \triangleq \mu_j p_j$ and $q^D_{jc}(N_j,N_j) \triangleq \mu_j\left[p_j+(1-p_j)\widehat{\rho}_j\right]$, $j=1,\ldots,J$; and for all K and $j \in K$,

$$q^D_{jK}(0,0) \triangleq \left[1+\frac{1}{(1-p_j)\widehat{\rho}_j}\right]\alpha^K_j, \quad \text{and} \quad q^D_{jK}(n_j,n_j) \triangleq \left[1+\frac{1}{(1-p_j)\widehat{\rho}_j}-\frac{1}{\widehat{\rho}_j}\right]\alpha^K_j,$$

then LHS of (A.4) is satisfied with

$$\beta^c_j \triangleq \mu_j\left[p_j+(1-p_j)\widehat{\rho}_j\right],\ j=1,\ldots,J, \quad \text{and} \quad \beta^K_j \triangleq \left[1+\frac{1}{(1-p_j)\widehat{\rho}_j}\right]\alpha^K_j,\ K\subseteq\Omega,\ j\in K.$$

If we denote the outstanding order vector of the new system by $\widehat{\mathbf{IO}}$, then Theorem A.2 implies that the stationary distribution of $\widehat{\mathbf{IO}}$ is of the product–form (A.6); and α^K_i, $i \in K \subseteq \Omega$, are determined by the traffic equations (A.5):

$$\alpha^K_j = \begin{cases} \lambda^K, & \text{if } j = i^K_1, \\ \left[1+\dfrac{1}{(1-p_{i^K_{m-1}})\widehat{\rho}_{i^K_{m-1}}}\right]\alpha^K_{i^K_{m-1}}, & \text{if } i^K_m = j,\ m=2,3,\ldots,|K|,\ N_j < +\infty, \\ \dfrac{\alpha^K_{i^K_{m-1}}}{\widehat{\rho}_{i^K_{m-1}}}, & \text{if } i = i^K_m,\ \ m=2,3,\ldots,|K|,\ N_i=\infty, \\ 0, & \text{otherwise.} \end{cases} \tag{A.7}$$

The traffic equations for α^K_i can be solved as follows: It is immediate from (A.7) that $\alpha^K_1 = \lambda^K$ if $1 \in K \subseteq \Omega$ and zero otherwise. Once we know α^K_1 for each $K \subseteq \Omega$, we can also compute $\widehat{\rho}_1 = \alpha_1/\mu_1 = \sum_{K:1\in K}\alpha^K_1/\mu_1$. Then we can compute α^K_2 based on α^K_1, $K\subseteq\Omega$ and $\widehat{\rho}_1$, and, later, $\widehat{\rho}_2 = \sum_{K:2\in K}\alpha^K_2/\mu_2$, and so on.

Since the new network has additional departures of positive signals, which subsequently will introduce more customers into the network, one can show, via sample path construction, that $\mathbf{IO} \leq_{st} \widehat{\mathbf{IO}}$.

B Tables

Order Profile	Utilization	Correlation	Policy
#1 Order Types: $\{1, 2, 3, 123\}$	High $\rho_j = 0.90$ $j = 1, 2, 3$	High $q^i = 0.05$, $i = 1, 2, 3$ $q^{123} = 0.85$	Backlogging $N_j = 7$, $s_j = 4$
			Lost Sales $N_j = 7$, $s_j = 7$
		Low $q^i = 0.25$,$i = 1, 2, 3$ $q^{123} = 0.25$	Backlogging $N_j = 7$, $s_j = 4$
			Lost Sales $N_j = 7$, $s_j = 7$
	Low $\rho_j = 0.50$ $j = 1, 2, 3$	High $q^i = 0.05$, $i = 1, 2, 3$ $q^{123} = 0.85$	Backlogging $N_j = 7$, $s_j = 4$
			Lost Sales $N_j = 7$, $s_j = 7$
		Low $q^i = 0.25$,$i = 1, 2, 3$ $q^{123} = 0.25$	Backlogging $N_j = 7$, $s_j = 4$
			Lost Sales $N_j = 7$, $s_j = 7$
#2 Order Types: $\{1, 2, 3,$ $12, 13, 23,\ 123\}$	High $\rho_j = 0.90$ $j = 1, 2, 3$	High $q^i = 0.05, i = 1, 2, 3$ $q^{12} = q^{13} = q^{23} = 0.225$ $q^{123} = 0.175$	Backlogging $N_j = 7$, $s_j = 4$
			Lost Sales $N_j = 7$, $s_j = 7$
		Low $q^i = 0.25, i = 1, 2, 3$ $q^{12} = q^{13} = q^{23} = 0.083$	Backlogging $N_j = 7$, $s_j = 4$
			Lost Sales $N_j = 7$, $s_j = 7$
	Low $\rho_j = 0.50$ $j = 1, 2, 3$	High $q^i = 0.05, i = 1, 2, 3$ $q^{12} = q^{13} = q^{23} = 0.225$ $q^{123} = 0.175$	Backlogging $N_j = 7$, $s_j = 4$
			Lost Sales $N_j = 7$, $s_j = 7$
		Low $q^i = 0.25, i = 1, 2, 3$ $q^{12} = q^{13} = q^{23} = 0.083$	Backlogging $N_j = 7$, $s_j = 4$
			Lost Sales $N_j = 7$, $s_j = 7$

Table 1: Parameter Settings: Three–component Systems

Batch Size	Utilization	Policy
Unit Batch Sizes Order Types: $\{25, 35, 125, 136, 1345, 1346\}$ with arrival probabilities, q^K, 0.10, 0.40, 0.15, 0.10, 0.20, 0.05, respectively	High $\rho_i = 0.90$, $i = 1, \ldots, 6$	Backlogging $N_i = 7$, $s_i = 4$, $i = 1, \ldots, 6$
		Lost Sales $N_i = 7$, $s_i = 7$, $i = 1, \ldots, 6$
	Low $\rho_i = 0.50$, $i = 1, \ldots, 6$	Backlogging $N_i = 7$, $s_i = 4$, $i = 1, \ldots, 6$
		Lost Sales $N_i = 7$, $s_i = 7$, $i = 1, \ldots, 6$
Geometric Batch Sizes with parameters, p_i, $i = 1, \ldots, 6$, 0.5, 0.75, 0.4, 0.8, 0.2, 0.85, respectively, with the same order profile as above	High $\rho_i = 0.90$, $i = 1, \ldots, 6$	Backlogging $s = (15, 30, 15, 40, 10, 50)$ $N = (30, 60, 30, 80, 20, 110)$
		Lost Sales $s_i = N_i$, $i = 1, \ldots, 6$ $N = (30, 60, 30, 80, 20, 110)$
	Low $\rho_i = 0.50$, $i = 1, \ldots, 6$	Backlogging $s = (15, 30, 15, 40, 10, 50)$ $N = (30, 60, 30, 80, 20, 110)$
		Lost Sales $s_i = N_i$, $i = 1, \ldots, 6$ $N = (30, 60, 30, 80, 20, 110)$

Table 2: Parameter settings: Six–component Systems

Order Profile	Utilization	Policy
#1 Order Types: $\{123, 345, 567, 789$ $12345, 34567, 56789\}$ with arrival probabilities, q^K, 0.08, 0.12, 0.16, 0.04, 0.2, 0.2, 0.2, respectively	High $\rho_j = 0.90$ $j = 1, \ldots, 9$	Backlogging $N_j = 7$, $s_j = 4$
		Lost Sales $N_j = 7$, $s_j = 7$
	Low $\rho_j = 0.50$ $j = 1, \ldots, 9$	Backlogging $N_j = 7$, $s_j = 4$
		Lost Sales $N_j = 7$, $s_j = 7$
	asymmetric $\rho_j = 0.45 + 0.05j$ $j = 1, \ldots, 9$	Backlogging $N_j = 7$, $s_j = 4$
		Lost Sales $N_j = 7$, $s_j = 7$
#2 Order Types: $\{123, 345, 567, 789$ $23, 34, 45, 56, 67, 78\}$ with arrival probabilities, q^K, 0.08, 0.12, 0.16, 0.04, 0.1, 0.1, 0.1, 0.1, respectively	High $\rho_j = 0.90$ $j = 1, \ldots, 9$	Backlogging $N_j = 7$, $s_j = 4$
		Lost Sales $N_j = 7$, $s_j = 7$
	Low $\rho_j = 0.50$ $j = 1, \ldots, 9$	Backlogging $N_j = 7$, $s_j = 4$
		Lost Sales $N_j = 7$, $s_j = 7$
	asymmetric $\rho_j = 0.45 + 0.05j$ $j = 1, \ldots, 9$	Backlogging $N_j = 7$, $s_j = 4$
		Lost Sales $N_j = 7$, $s_j = 7$

Table 3: Parameter Settings: Nine–component Systems

					Lower Bounds								Upper Bounds			
					First Order				Second Order				First Order		Second Order	
Factor			Exact		Order Profile #1											
U	C	P	Fill Rate	Serv'blty	Signal	Bf1	CMB1	StL1	Bf2	CMB2	EStL2	GStL2	StU1	Frcht1	StU2	Frcht2
h	h	b	0.3810	0.8185	-80.34	-76.23	-48.27	-27.12	-16.49	-17.53	-17.53	-17.53	58.34	58.47	20.94	20.99
		l	0.8185	0.8185	-43.26	-5.54	-4.07	-3.40	-1.32	-2.14	-2.14	-2.14	11.91	11.91	5.30	5.30
	l	b	0.5234	0.8831	-58.57	-13.48	-7.49	-2.96	-4.03	-1.92	-1.92	-1.92	13.14	15.37	4.92	5.67
		l	0.8831	0.8831	-32.23	-1.03	-0.63	-0.45	-0.28	-0.29	-0.29	-0.29	3.62	3.73	1.67	1.72
l	h	b	0.8762	0.9902	-92.93	-3.99	-3.32	-3.00	-0.88	-1.86	-1.86	-1.86	7.42	7.42	3.27	3.27
		l	0.9902	0.9902	-65.70	-0.07	-0.07	-0.06	-0.02	-0.04	-0.04	-0.04	0.60	0.60	0.29	0.29
	l	b	0.9179	0.9942	-56.89	-0.66	-0.47	-0.38	-0.17	-0.24	-0.24	-0.24	2.34	2.54	1.09	1.18
		l	0.9942	0.9942	-26.29	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.19	0.19	0.09	0.09
Factor			Exact		Order Profile #2											
U	C	P	Fill Rate	Serv'blty	Signal	Bf1	CMB1	StL1	Bf2	CMB2	EStL2	GStL2	StU1	Frcht1	StU2	Frcht2
h	h	b	0.4177	0.8466	-82.07	-44.77	-14.16	-10.19	-3.55	-2.11	-2.11	-2.11	42.35	44.54	4.66	4.66
		l	0.8466	0.8466	-45.15	-1.98	-1.12	-0.99	-0.22	-0.25	-0.25	-0.25	8.14	8.19	1.22	1.22
	l	b	0.5476	0.8974	-33.35	-7.82	-0.65	-0.65	0.00	0.00	0.00	0.00	4.45	10.27	0.00	0.00
		l	0.8974	0.8974	-11.27	-0.26	-0.07	-0.07	0.00	0.00	0.00	0.00	1.31	2.08	0.00	0.00
l	h	b	0.8926	0.9915	-76.38	-1.31	-0.91	-0.85	-0.15	-0.22	-0.22	-0.22	5.29	5.44	0.81	0.81
		l	0.9915	0.9915	-34.27	-0.06	-0.06	-0.06	-0.07	-0.07	-0.07	-0.07	0.43	0.47	0.14	0.14
	l	b	0.9278	0.9951	-18.11	-0.14	-0.05	-0.05	0.00	0.00	0.00	0.00	0.81	1.44	0.00	0.00
		l	0.9951	0.9951	-3.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.10	0.00	0.00

Table 4: Three–component Systems: Percentage Errors of Bounds for Overall Immediate Fill Rates

				Lower Bounds								Upper Bounds			
				First Order				Second Order				First Order		Second Order	
Factor		Exact		Basic Model with Unit Batch Sizes											
U	P	Fill Rate	Serv'blty	Signal	Bf1	CMB1	StL1	Bf2	CMB2	EStL2	GStL2	StU1	Frcht1	StU2	Frcht2
h	b	0.3329	0.8053	-99.03	-68.27	-55.70	-18.26	-23.57	-13.76	-6.16	-6.16	68.21	80.93	20.47	21.42
	l	0.8053	0.8053	-73.49	-4.33	-3.08	-2.04	-1.29	-0.98	-0.75	-0.83	12.52	13.68	5.06	5.15
l	b	0.8647	0.9898	-98.02	-3.06	-2.46	-1.99	-1.19	-1.04	-0.96	-0.96	7.86	8.84	3.03	3.11
	l	0.9898	0.9898	-78.44	-0.06	-0.05	-0.05	-0.05	-0.05	-0.04	-0.04	0.63	0.63	0.24	0.24
Factor		Exact		Basic Model with Geometric Batch Sizes											
U	P	Fill Rate	Serv'blty	Signal	Bf1	CMB1	StL1	Bf2	CMB2	EStL2	GStL2	StU1	Frcht1	StU2	Frcht2
h	b	0.4196	0.9415	-100.00	-54.57	-33.61	-13.00	n/a	n/a	n/a	n/a	56.18	59.01	n/a	n/a
	l	0.9415	0.9415	-91.74	-0.43	-0.34	-0.28	n/a	n/a	n/a	n/a	3.54	3.54	n/a	n/a
l	b	0.9742	0.9999	-100.00	-0.08	-0.06	-0.05	n/a	n/a	n/a	n/a	1.24	1.32	n/a	n/a
	l	0.9999	0.9999	-89.17	0.00	0.00	0.00	n/a	n/a	n/a	n/a	0.00	0.01	n/a	n/a

Table 5: Six–component Production Systems: Percentage Errors of Bounds for Overall Immediate Fill Rates

				Lower Bounds								Upper Bounds			
				First Order				Second Order				First Order		Second Order	
Factor		Exact		Order Profile #1											
U	P	Fill Rate	Serv'blty	Signal	Bf1	CMB1	StL1	Bf2	CMB2	EStL2	GStL2	StU1	Frcht1	StU2	Frcht2
h	b	0.2402	0.7382	-99.88	-100.00	-72.62	-43.29	-58.07	-30.18	-26.99	-29.47	132.33	151.39	66.41	66.41
	l	0.7382	0.7382	-91.90	-12.33	-7.17	-5.94	-4.15	-2.89	-3.49	-3.91	23.14	24.09	14.50	14.50
l	b	0.8174	0.9848	-99.68	-7.87	-5.56	-4.98	-2.32	-2.02	-2.76	-3.17	14.30	15.15	9.03	9.03
	l	0.9848	0.9848	-87.09	-0.12	-0.11	-0.11	-0.03	-0.04	-0.05	-0.06	1.13	1.15	0.76	0.76
a	b	0.5138	0.8934	-99.80	-51.46	-24.08	-18.53	-17.11	-9.07	-10.30	-13.27	40.87	41.91	19.74	20.98
	l	0.8934	0.8934	-91.67	-2.54	-1.76	-1.62	-0.76	-0.63	-0.79	-1.24	6.44	6.60	3.34	3.57
Factor		Exact		Order Profile #2											
U	P	Exact	Serv'blty	Signal	Bf1	CMB1	StL1	Bf2	CMB2	EStL2	GStL2	StU1	Frcht1	StU2	Frcht2
h	b	0.3602	0.8232	-93.56	-65.43	-51.50	-14.82	-7.33	-7.33	-3.85	-3.91	58.46	67.55	11.47	11.47
	l	0.8232	0.8232	-52.87	-3.01	-2.32	-1.50	-0.19	-0.19	-0.33	-0.36	10.52	11.26	2.83	2.83
l	b	0.8776	0.9917	-78.81	-2.14	-1.82	-1.44	0.00	0.00	-0.22	-0.22	6.48	7.23	1.83	1.83
	l	0.9917	0.9917	-30.08	-0.11	-0.11	-0.10	0.00	0.00	0.00	0.00	0.42	0.44	0.14	0.14
a	b	0.6375	0.9362	-89.56	-17.29	-12.44	-5.76	-1.18	-1.18	-1.11	-1.11	18.04	21.05	4.21	4.61
	l	0.9362	0.9362	-45.98	-0.73	-0.64	-0.51	0.00	0.00	0.00	0.00	2.69	2.92	0.76	0.78

Table 6: Nine–component Systems: Percentage Errors of Bounds for Overall Immediate Fill Rates

References

[1] Baccelli, F. and A. M. Makowski. 1989. Multidimensional Stochastic Ordering and Associated Random Variables. *Oper. Res.*, **37**, 478-487.

[2] Baccelli, F., A. M. Makowski and A. Schwartz. 1989. The Fork-Join Queue and Related Systems with Synchronization Constraints: Stochastic Ordering and Computable Bounds. *Adv. Appl. Prob.*. **21**, 629-660.

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

[4] Connors, and D.D. Yao. 1996. Methods for Job Configuration in Semiconductor Manufacturing. *IEEE Transactions on Semiconductor Manufacturing* **9**, 401-411.

[5] Chao, X., M. Miyazawa and M. Pinedo. 1999. *Queueing Networks: Customers, Signals, and Product Form Solutions.* Wiley & Sons.

[6] Costigan, T. 1996. Combination Setwise–Bonferroni-Type Bounds. *Naval Research Logistics* **43**, 59-77.

[7] Flatto, L. and S. Hahn. 1984. Two Parallel Queues Created by Arrivals with Two Demands I. *SIAM J. Appl. Math.* **44**, 1041-1052.

[8] Glasserman, P. and Y. Wang. 1998. Leadtime-Inventory Trade-Offs in Assemble-to-Order Systems. *Operations Research*, **46**, 858-871.

[9] Hausman, W., H. Lee and A. Zhang. 1998. Joint Demand Fulfillment Probability in a Multi-component Inventory System with Independent Order-up-to Policies. *Euro. J Oper. Res.* **109**, 646-659.

[10] Joe, H. 1997. *Multivariate Models and Dependence Concepts*, Chapman & Hall.

[11] Li, H. and S. H. Xu 2000. On the Dependence Structure and Bounds of Correlated Parallel Queues and Its Applications to Synchronized Stochastic Systems. *J. of Appl. Prob.,* **37**, 1020-1043.

[12] Lu, Y. and J.-S. Song. 2002. Order-Based Cost Optimization in Assemble-to-Order Systems. Working paper, Graduate School of Management, University of California, Irvine.

[13] Lu, Y., J.-S. Song, and D.D. Yao. 2003a. Performance Analysis of Multiproduct Assemble-to-Order Systems with Uncertain Leadtime, *Operations Research*, March-April, forthcoming.

[14] Lu, Y., J.-S. Song, and D.D. Yao. 2003b. Backorder Minimization in Multiproduct Assemble-to-Order Systems. In preparation (near completion).

[15] Mamer, J. and S. Smith. 1993. Job Fill Inventories for Sequences of Jobs. Working paper, UCLA Graduate School of Management.

[16] Nelson, R. and A. N. Tantawi. 1988. Approximate Analysis of Fork/Join Synchronization in Parallel Systems. *IEEE Trans. Comput*, **37**, 739-743.

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

[18] Rolski, T. 1983. Upper Bounds for Single Server Queues with Doubly Stochastic Poisson Arrivals. *Math. of Oper. Res.*, Vol 11, 442-450.

[19] Shaked, M. and J. G. Shanthikumar 1997. Supermodular stochastic order and positive dependence of random vectors, *J. Multivariate Analysis*, **61**, 86-101.

[20] Schwartz, A. and A. Weiss. 1993. Induced Rare Events: Analysis via Large Deviations and Time Reversal. *Adv. Appl. Prob.* **25**, 667-689.

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

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

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

[24] Song, J.-S. and P. Zipkin. 2001. Supply Chain Operations: Assemble-to-Order Systems. Chapter XIII in *Handbooks in Operations Research and Management Science*, **XXX**, *Supply Chain Management*, T. de Kok and S. Graves, eds, North-Holland, forthcoming.

[25] Tchen, A. H. 1980. Inequalities for distributions with given marginals. *Ann. Probab.* **8**, 814-827.

[26] Tong, Y.L. 1980. *Probability Inequalities in Multivariate Distributions*, Academic Press, New York.

[27] Topkis, D.M. 1978. Minimizing a Submodular Function on a Lattice, *Operations Research,* **26**, 305-321.

[28] Wright, P. 1992. Two Parallel Processors with Coupled Inputs. *Adv. Appl. Prob.* **24**, 986-1007.

[29] Xu, S. H. 1999. Structural Analysis of a Queueing System with Multi-classes of Correlated Arrivals and Blocking. *Operations Research*, **47**, 263-276.

[30] Xu, S. H. 2002. Dependence Analysis of Assemble-to-Order Systems. In *Supply chain structures: Coordination, Information and Optimization*, J.S. Song and D.D. Yao (eds), Kulwer Academic Publishers, 359-414.

[31] Zhang, R. 1999. Expected Time Delay in a Multi-Item Production-Inventory System with Correlated Demands. *Navel Research Logistics*, **46**, 671-688.

[32] Zipkin, P. H. 2000. *Foundations of Inventory management.* McGraw-Hill, New York.