Expected Time Delay in Multi-Item Inventory Systems with Correlated Demands

Rachel Q. Zhang
Department of Industrial and Operations Engineering, University of Michigan,
Ann Arbor, Michigan 48109

Received January 1997; revised November 1998; accepted 2 March 1999

Abstract: This paper considers multi-item inventory systems where a customer order may require several different items (i.e., demands are correlated across items) and customer satisfaction is measured by the time delays seen by the customers. Most inventory models on time delay in the literature assume each demand only requires one item (i.e., demands are not correlated across items or are independent). In this paper, we derive an exact expression for the expected *total* time delay. We show that when items are actually correlated, assuming items are independent leads to an overestimate of the total time delay. However, (1) it is extremely difficult in practice to obtain the demand information for all demand types (especially in a system with tens of thousands of part numbers), and (2) the problem becomes too complicated to be of practical interest when the correlation is considered. We then explore the possibility of including the demand information partially and develop bounds for the time delays. © 1999 John Wiley & Sons, Inc. Naval Research Logistics 46: 671–688, 1999

1. INTRODUCTION

In this paper, we consider a multi-item, continuous review, base stock inventory system. In this system, a customer order may require several different items (i.e., demands are correlated across items). The performance measure we are concerned with is the long-run expected total time delay seen by the customers.

This research was motivated by our experience working with service related industries. In these industries, companies often offer service contracts on their equipment and stock spare parts to support their maintenance function. The spare parts may be manufactured in-house as well as ordered from outside suppliers. When a machine is down, a repair person from the maintenance company is sent in as soon as possible with the parts needed to fix the machine. It is typical that customers who outsource their service have little knowledge about the equipment they are using (e.g., Xerox machines in our offices). Many times, they may not care which part or parts have failed. All they care about is how long the machine remains down. Therefore, down times seen by the customers are used by the customers to measure the service they receive. Furthermore, customers normally do not expect the maintenance companies to specify the service levels for all failure types on their contracts. They either expect the maintenance companies to guarantee the same down time for each repair incident, which is

Contract grant sponsor: National Science Foundation; contract grant number: CAREER DMI-9501740

CCC 0894-069X/99/060671-18

easier to deal with, or the total down time seen by the customers per unit time (e.g., month, year, etc.). On the other hand, the ability of a technician to complete a repair is a function of the availability of the part (or *parts*) needed for that repair job and a repair incident may require several part numbers, i.e., the demand for spare parts is correlated across items.

There has been some research on problems with correlated demands. Smith, Chambers, and Shlifer [10] first introduced the "job-fill" rate criterion, i.e., the fraction of jobs completed without stockout, which is a more appropriate measure than the part-fill rate in many applications. For example, a service representative must have spares for *all* failed components in order to complete a repair, since the equipment remains down if an extra trip is made to obtain parts. Assuming that the service personnel are always able to replenish inventory between jobs, Smith, Chambers, and Shlifer [10], Mamer and Smith [7, 8], Graves [3], March and Scudder [9], and Mamer and Shogan [6] studied the problem of determining the appropriate collection of parts to be carried in a repair kit. Song [11] investigated a continuous review multi-item inventory system where demands are correlated across items and replenishment lead times are constant for all items. She derived the expressions for the job fill rate and studied the relationship between the part fill rate and the job fill rate. Cheung and Hausman [2] considered a base stock spare parts inventory system with multiple failures. They derived exact expressions and approximations for the distribution function and the expected number of backorders. Recently, there have been studies addressing customer waiting time in systems where demands are correlated. In particular, Hausman, Lee, and Zhang [4] considered a periodic review multi-item base-stock system where demands are correlated and the replenishment lead times are constant. They obtained bounds on the customer waiting time distribution and studied the problem of maximizing the probability of meeting customer orders within a prespecified time window. Song, Xu, and Liu [12] studied a multi-item production and inventory problem where demands are Poisson and each item is made by an independent $M/M/1/m$ queue. They derived the order fill rate and the probability of customer waiting times, and developed a procedure to compute these performance measures exactly. Glasserman and Wang [5] studied the tradeoff between inventory levels and the delivery lead times, in a limiting sense, at high service level for various models including the one that allows orders for multiple items.

In this paper, we consider a continuous review multi-item inventory system where a demand may require more than one item. The only major assumptions that we make here are: (1) demand for all demand types is Poisson, and (2) for each item, there is a transit time between the release and receipt of an order due to transportation or processing activities (these transit times may be stochastic as well as deterministic) as defined in Svoronos and Zipkin [13]. We derive expressions for the expected total time delay for two cases: (a) where demands are correlated across items (called exact time delay) and (b) where demands are treated independently although they are actually correlated (called individual time delay). We show that when items are actually correlated, assuming items are independent leads to an overestimate of the total time delay. This suggests that with the same budget, a company may actually offer much higher service than what is stated on the service contract. If they can evaluate customer waiting times accurately, they may be able to reduce inventory or promise better service to the customers.

However, (1) it is extremely difficult, in practice, to obtain the demand information for all demand types, and (2) the evaluation of the exact time delay becomes too complex to be of practical interest when the correlation is considered. Therefore, the goals of this research are to investigate (1) how important the information about correlation among the parts is in decision-making, (2) under what circumstances the relative error caused by ignoring the correlation among the parts is negligible, and (3) if we can consider the information partially to reduce the error. To achieve these goals, we derive the error term and prove that the error would disappear if and only if all demands require only one item at a time. We demonstrate the magnitudes of the relative error by numerical examples and show that the relative error can be large. We then explore the possibility of partially considering the demand information by ignoring the fact that more than l part numbers may be used together. For example, if a customer may request (a) parts 1 and 2, (b) parts 3 and 4, and (c) parts 1, 2, 3, and 4, we only count how many kits containing parts 1 and 2 and how many kits containing parts 3 and 4 are consumed (i.e., ignore the fact that part numbers 1, 2, 3, and 4 may be ordered by a customer altogether). In reality, parts frequently used together may be stocked in the same kit, which makes it possible to obtain the demand information for a kit. This approach has the potential to reduce the complexity of the computation and may lead to heuristic algorithms to approximate the time delay in real systems. However, we show analytically that we may overestimate as well as underestimate the time delay if we consider the correlation partially, depending on how much information about the demand correlation is considered.

In the case where there is a dedicated single machine facility for the production of each item and the processing times are exponential, we develop a procedure to compute the expected total time delay. The goals here are to examine how important the demand information is in decision making and if it is worthwhile to ignore some of the demand information in exchange for data availability and simplicity. We conclude through numerical examples:

1. One may greatly overestimate the time delay if the correlations among the items are ignored. However, the errors get smaller as the time delay decreases.
2. The time delay based on partial demand information offers good approximation, especially when the time delay is not too high. The relative errors decrease as the degree of the correlation decreases or the replenishment lead times decrease.

This paper is organized as follows. In Section 2, we describe the inventory system in detail. We then derive the exact expression for the expected total time delay and compare this with that when items are treated independently. In Section 3, we consider the time delay with partial demand information. In Section 4, we examine a special case where each item is produced by an $M/M/1$ production system. We develop a procedure to compute the time delay and its approximations. We discuss computational issues in nonexponential cases in Section 5. The paper concludes in Section 6.

2. PROBLEM FORMULATION

2.1. Assumptions and Notation

We consider an inventory system of N different items. Let $\Omega = \{1, 2, \ldots, N\}$ be the set of all item indexes and Ω_n be the set of all subsets of Ω containing item n. For any subset $\mathcal{W}$ of Ω, $|\mathcal{W}|$ represents the number of elements in $\mathcal{W}$. We will use boldfaced letters to represent vectors. In particular, $\mathbf{0}$ is a row vector of zeroes, whose dimension will be clear from the context. Throughout the paper, $\mathbf{X}_{\mathcal{W}} = (X_{j_1}, X_{j_2}, \ldots, X_{j_{|\mathcal{W}|}})$ if $\mathcal{W} = \{j_1, j_2, \ldots, j_{|\mathcal{W}|}\}$ and $\mathbf{X} = \mathbf{X}_{\Omega}$.

For modeling purposes, we will make the following assumptions:

- *Same priority for all customers.* We assume that the delay permitted by each customer is proportional to demand. This is the case if service contracts are stated in terms of the number of machines maintained for the customer (e.g., a customer with 10 machines is permitted twice as much delay as a customer with 5 machines).

With this we can consider the total time delay per year based on the aggregated demand.

- *Poisson demand.* This assumption, commonly used to represent demand processes both in research and in practice, represents module failures over time in maintenance related research. Furthermore, we assume each demand may require several items, but at most one unit of each item. We call a demand pattern a type $\mathcal{W}$ if it only asks for all the items in set $\mathcal{W} \subseteq \Omega$. A demand requiring only one item, n, is called a type n demand. If we assume that each type of demand is stationary over time and follows an independent Poisson process, then the total demand for each item forms a Poisson process as well.
- *Independent transit times.* We assume that there is a transit time with positive mean due to transportation or processing activities for each item. This includes constant lead times, iid lead times and other stochastic transit times as defined in Svoronos and Zipkin [13].
- *Continuous review base stock policy for each item.* We assume that whenever the inventory position drops below a fixed reorder point S, an order of size one is placed. If the items are supplied by some production facilities, one job is released to each of the facilities where the items demanded are produced. The facility keeps working until all the jobs in the queue are finished. Then the facility is shut down until a new job is released to the facility. Here we need to point out, this may not be the optimal policy, since one may need to consider the status (e.g., inventory, number of backorders, processing times, etc.) at all the facilities that make the items in a customer order before releasing jobs to the facilities. However, for simplicity and tractability, we use base stock policy. Furthermore, we focus on cases where base-stock levels for all items are nonnegative for all items.
- *Backlogging unfilled demand.* Since we are dealing with service related industries which have service contracts with their customers, unfilled demand is backlogged. A demand is considered filled only when all the items it requires are filled from stock. Demands are filled on a first-come-first-serve basis. If a demand is backlogged, but can be partially filled, the available items will be reserved for the demand. This may not be optimal, since it may make sense to satisfy some demand requiring the items that are available first. However, for simplicity and tractability, we assume first-come-first-serve.

Notation is as follows:

N = number of items,
$\Omega = \{1, 2, \ldots, N\}$ = the set of item indexes,
Ω_n = the set of all subsets of Ω containing item n,
$\lambda_{\mathcal{W}}$ = demand rate of type $\mathcal{W} \subseteq \Omega$, $(\lambda_{\{n\}} \equiv \lambda_n)$ per year,
$\tilde{\lambda}_n = \sum_{\mathcal{W} \in \Omega_n} \lambda_{\mathcal{W}}$ = total demand rate for item n per year,
I_n = net inventory level (on-hand–backorder) of item n upon an arrival,
w_n = waiting time for item n upon arrival,
S_n = order up to level for item n,
$t(\mathbf{S})$ = expected total time delay per year for a given $\mathbf{S} = (S_1, S_2, \ldots, S_N)$,
$t_{ind}(\mathbf{S})$ = expected total time delay per year for a given $\mathbf{S}$ when items are treated independently.

2.2. Exact Time Delay and Individual Time Delay

In this section we derive the exact expression for the expected *total* time delay seen by the customers per year in steady state. For any given **S**, the expected total time delay seen by the customers per year can be expressed as

$$t(\mathbf{S}) = \sum_{n=1}^{N} \lambda_n E(w_n) + \sum_{\substack{\mathcal{W} \subseteq \Omega \\ |\mathcal{W}|>1}} \lambda_{\mathcal{W}} E(\max_{j \in \mathcal{W}} \{w_j\}). \tag{1}$$

That is, if a demand requires only one item, then the expected waiting time is simply $E(w_n)$. However, if a demand asks for more than one item, the waiting time is not the sum of the individual waiting times, but the maximum of the individual waiting times. Hence the expected waiting time for a demand of type $\mathcal{W}$, $|\mathcal{W}| > 1$, is $E(\max_{j \in \mathcal{W}} \{w_j\})$.

If we treat each item independently, i.e., we treat each item demand as a separate demand incident, then the expected total time delay seen by the customers per year becomes

$$t_{ind}(\mathbf{S}) = \sum_{n=1}^{N} \tilde{\lambda}_n E(w_n). \tag{2}$$

We call this the individual expected total time delay (or individual time delay). Here we need to point out that the expected time delay $E(w_n)$ is the same in the exact time delay case and the individual time delay case, because the demand process and lead time are the same for each item in both cases no matter how the items are correlated.

2.3. Comparison between *t*(S) and *t*_{*ind*}(S)

To better understand how $t(\mathbf{S})$ and $t_{ind}(\mathbf{S})$ are related, we examine the two-item case first. Rewriting the exact time delay $t(\mathbf{S})$ as a function of $t_{ind}(\mathbf{S})$ in two-item case yields

$$\begin{aligned} t(\mathbf{S}) &= \lambda_1 E(w_1) + \lambda_2 E(w_2) + \lambda_{\{1,2\}} E(\max\{w_1, w_2\}) \\ &= (\tilde{\lambda}_1 - \lambda_{\{1,2\}}) E(w_1) + (\tilde{\lambda}_2 - \lambda_{\{1,2\}}) E(w_2) + \lambda_{\{1,2\}} E(\max\{w_1, w_2\}) \\ &= t_{ind}(\mathbf{S}) - \lambda_{\{1,2\}} \sum_{j=1}^{2} E(w_j) + \lambda_{\{1,2\}} E(\max\{w_1, w_2\}). \end{aligned}$$

Applying the identity $\max\{w_1, w_2\} = w_1 + w_2 - \min\{w_1, w_2\}$ yields

$$t(\mathbf{S}) = t_{ind}(\mathbf{S}) - \lambda_{\{1,2\}} E(\min\{w_1, w_2\}) \leq t_{ind}(\mathbf{S}). \tag{3}$$

Because the lead times have positive means, $t(\mathbf{S}) = t_{ind}(\mathbf{S})$ if and only if $\lambda_{\{1,2\}} = 0$. Equation (3) indicates that by using $t_{ind}(\mathbf{S})$ one actually counts the smaller waiting time between the two individual waiting times twice when a demand requests two items and both of them are out of stock, in which case the longer waiting time between the two individual waiting times determines the actual waiting time of the demand. Using the same argument, we can extend the results to the N-item case.

THEOREM 1: For given $\mathbf{S}$, $t(\mathbf{S}) \leq t_{ind}(\mathbf{S})$. The equality holds if and only if $\lambda_{\mathcal{W}} = 0$ for all $\mathcal{W} \subseteq \Omega$, $|\mathcal{W}| > 1$. Furthermore,

$$t(\mathbf{S}) = t_{ind}(\mathbf{S}) - \sum_{\substack{\mathcal{W}\subseteq\Omega \\ |\mathcal{W}|>1}} \lambda_{\mathcal{W}}\Delta_{\mathcal{W}} \tag{4}$$

where

$$\Delta_{\mathcal{W}} = \sum_{\substack{i,j\in\mathcal{W} \\ i\neq j}} E(\min\{w_i, w_j\}) - \cdots + (-1)^{|\mathcal{W}|}E(\min_{j\in\mathcal{W}}\{w_j\}) > 0.$$

Theorem 1 shows that we will overestimate the time delay if we treat the items independently while they are actually correlated.

3. TIME DELAY WITH PARTIAL DEMAND INFORMATION

In the previous section, we show that one overestimates the time delay if the correlations of the demands are ignored. However, it is extremely difficult to obtain the demand information for all demand types and the evaluation of the exact time delay can be extremely difficult, especially when N is large. We will see later, even in the case where each item is made by a dedicated $M/M/1$ production system, it is prohibitive to compute the exact time delay for $N \geq 4$. In this section, we explore the possibility to reduce the error by including the demand information partially. First we rewrite $t(\mathbf{S})$ as in Lemma 1 and show that the individual terms in $t(\mathbf{S})$ is in decreasing order in Lemma 2. Then we derive bounds and approximations for $t(\mathbf{S})$.

LEMMA 1: $t(\mathbf{S})$ can be written as

$$\begin{aligned} t(\mathbf{S}) = t_{ind}(\mathbf{S}) - \sum_{\substack{i,j\in\Omega \\ i\neq j}} \tilde{\lambda}_{\{i,j\}}E(\min\{w_i, w_j\}) + \sum_{\substack{i,j,k\in\Omega \\ i\neq j\neq k}} \tilde{\lambda}_{\{i,j,k\}}E(\min\{w_i, w_j, w_k\}) \\ + \cdots + (-1)^{|\Omega|-1}\lambda_{\Omega}E(\min_{j\in\Omega}\{w_j\}) \end{aligned} \tag{5}$$

where $\tilde{\lambda}_{\mathcal{B}} = \Sigma_{\mathcal{B}\subseteq\mathcal{W}\subseteq\Omega}\, \lambda_{\mathcal{W}}$ is the average demand requesting all part numbers in $\mathcal{B}$.

LEMMA 2: For $k = 1, 2, \ldots, |\Omega| - 1$,

$$\sum_{\substack{\mathcal{B}\subseteq\Omega \\ |\mathcal{B}|=k}} \tilde{\lambda}_{\mathcal{B}}E(\min_{j\in\mathcal{B}}\{w_j\}) \geq \sum_{\substack{\mathcal{B}\subseteq\Omega \\ |\mathcal{B}|=k+1}} \tilde{\lambda}_{\mathcal{B}}E(\min_{j\in\mathcal{B}}\{w_j\}).$$

The detailed proof of Lemma 1 is given in the Appendix, and the proof of Lemma 2 is omitted. Lemma 2 indicates that the terms in (5) are decreasing in absolute value.

We now derive bounds for $t(\mathbf{S})$. Let $t^l(\mathbf{S})$ be the first l sums in (5), $l = 1, \ldots, |\Omega|$. Then $t^l(\mathbf{S}) \geq 0$ by Lemma 2. $t^l(\mathbf{S})$ is the exact time delay when $l = |\Omega|$ and the individual time delay when $l = 1$. We show in Theorem 2 that $t^l(\mathbf{S})$ is a lower (upper) bound when l is an even (odd) number for $l < \Omega$ and $t^l(\mathbf{S})$ becomes closer to $t(\mathbf{S})$ as l increases.

THEOREM 2: If $|\Omega|$ is an odd number, then

$$t^2(\mathbf{S}) \le t^4(\mathbf{S}) \le \cdots \le t^{|\Omega|-1}(\mathbf{S}) \le t(\mathbf{S}) = t^{|\Omega|}(\mathbf{S}) \le t^{|\Omega|-2}(\mathbf{S}) \le \cdots \le t^1(\mathbf{S}) = t_{ind}(\mathbf{S}), \quad (6)$$

and, if $|\Omega|$ is an even number, then

$$t^2(\mathbf{S}) \le t^4(\mathbf{S}) \le \cdots \le t^{|\Omega|}(\mathbf{S}) = t(\mathbf{S}) \le t^{|\Omega|-1}(\mathbf{S}) \le t^{|\Omega|-3}(\mathbf{S}) \le \cdots \le t^1(\mathbf{S}) = t_{ind}(\mathbf{S}). \quad (7)$$

Furthermore, $|t(\mathbf{S}) - t^l(\mathbf{S})|$ is decreasing in l.

The proof is in the Appendix.

Theorem 2 tells us that $t^l(\mathbf{S})$ provides better approximation for $t(\mathbf{S})$ as l increases. If we examine $t^l(\mathbf{S})$ closely, we find it is exactly the expected total time delay if we ignore the fact that sometimes more than l part numbers are used together, but take into account the demands requesting l or fewer part numbers. In other words, $t^l(\mathbf{S})$ is the total time delay when the information about the demand correlation is partially considered. The larger the l is, the more information we include and, hence, the more accurate the estimate is. Using $t^l(\mathbf{S})$ as an approximation for $t^l(\mathbf{S})$ should be accurate unless (1) $\lambda_{\mathcal{W}}$ is large, $|\mathcal{W}| > l$ and (2) there is high probability that l or more items stock out when a demand requests l or more items.

However, the larger the l is, the more difficult it is to compute $t^l(\mathbf{S})$. Even though it is still an open question in terms of how $t^l(\mathbf{S})$ can be computed in general cases, future work can be done to develop heuristic algorithms to compute $t^l(\mathbf{S})$ under certain conditions. In this way, we can greatly reduce the complexity of the computation.

Now do we still overestimate the total waiting time by using $t^l(\mathbf{S})$? We may underestimate *and* overestimate the time delay, depending on l since $t^l(\mathbf{S})$ may be an upper bound or a lower bound. This tells us we need to be cautious when considering partial information about the correlation.

By Theorem 2, $|t(\mathbf{S}) - t^l(\mathbf{S})|$ is decreasing in l. Hence, for any given even number l, $4 \le l \le |\Omega| - 1$,

$$t(\mathbf{S}) - t^{l-2}(\mathbf{S}) \ge t^{l-1}(\mathbf{S}) - t(\mathbf{S}) \ge t(\mathbf{S}) - t^l(\mathbf{S}). \quad (8)$$

Combining (8) with (6) and (7), we have

$$\max\{[t^{l-2}(\mathbf{S}) + t^{l-1}(\mathbf{S})]/2,\ t^l(\mathbf{S})\} \le t(\mathbf{S}) \le [t^{l-1}(\mathbf{S}) + t^l(\mathbf{S})]/2 \le t^{l-1}(\mathbf{S}). \quad (9)$$

Following the same argument, for any given odd number l, $3 \le l \le |\Omega| - 1$,

$$t^{l-2}(\mathbf{S}) - t(\mathbf{S}) \ge t(\mathbf{S}) - t^{l-1}(\mathbf{S}) \ge t^l(\mathbf{S}) - t(\mathbf{S})$$

and

$$t^{l-1}(\mathbf{S}) \le [t^{l-1}(\mathbf{S}) + t^l(\mathbf{S})]/2 \le t(\mathbf{S}) \le \min\{[t^{l-2}(\mathbf{S}) + t^{l-1}(\mathbf{S})]/2,\ t^l(\mathbf{S})\}. \quad (10)$$

Inequalities (9) and (10) provide us with better bounds and approximations than (6) and (7). We can improve the bounds as we include more information about the demand correlation as follows:

1. Compute $t_{ind}(\mathbf{S})$ and $0 \le t(\mathbf{S}) \le t_{ind}(\mathbf{S})$;
2. compute $t^2(\mathbf{S})$ and $t^2(\mathbf{S}) \le t(\mathbf{S}) \le [t_{ind}(\mathbf{S}) + t^2(\mathbf{S})]/2$;
3. compute $t^3(\mathbf{S})$ and $t^2(\mathbf{S}) \le [t^2(\mathbf{S}) + t^3(\mathbf{S})]/2 \le t(\mathbf{S}) \le \min\{[t_{ind}(\mathbf{S}) + t^2(\mathbf{S})]/2, t^3(\mathbf{S})\}$;
4. compute $t^4(\mathbf{S})$ and $\max\{[t^2(\mathbf{S}) + t^3(\mathbf{S})]/2, t^4(\mathbf{S})\} \le t(\mathbf{S}) \le [t^3(\mathbf{S}) + t^4(\mathbf{S})]/2 \le t^3(\mathbf{S})$;
5. compute $t^5(\mathbf{S})$ and $t^4(\mathbf{S}) \le [t^4(\mathbf{S}) + t^5(\mathbf{S})]/2 \le t(\mathbf{S}) \le \min\{[t^3(\mathbf{S}) + t^4(\mathbf{S})]/2, t^5(\mathbf{S})\}$;

   $\vdots$

We stop when the upper and lower bounds are close enough.

4. A SPECIAL CASE: EXPONENTIAL PROCESSING TIME

In this section, we explore a special case where each item is made by a dedicated single machine facility and the processing times are exponentially distributed at all facilities. By assuming exponential processing times, we are able to develop a procedure to compute the exact time delay and compare it with the independent time delay and the bounds through numerical examples.

4.1. Exact Time Delay and Individual Time Delay

Recall that the expected total time delay is given by

$$t(\mathbf{S}) = \sum_{n=1}^{N} \lambda_n E(w_n) + \sum_{\substack{\mathcal{W} \subseteq \Omega \\ |\mathcal{W}| > 1}} \lambda_{\mathcal{W}} E(\max_{j \in \mathcal{W}}\{w_j\}). \tag{11}$$

Let μ_n be the mean processing rate for item $n \in \Omega$ and assume $\mu_n > \tilde{\lambda}_n$ so that the system is stable. Let $P_r(I_n)$ represent the probability that the net inventory for item n is I_n upon an arrival and $P_r(\mathbf{I}_{\mathcal{A}})$ the probability that the net inventory for items in $\mathcal{A}$ is $\mathbf{I}_{\mathcal{A}}$ upon an arrival. Rewriting $E(w_n)$ and $E(\max_{j \in \mathcal{W}}\{w_j\})$, we have

$$E(w_n) = \sum_{I_n \le 0} P_r(I_n) E(w_n^{I_n} | I_n), \tag{12}$$

and

$$E(\max_{j \in \mathcal{W}}\{w_j\}) = \sum_{\substack{\mathcal{A} \subseteq \mathcal{W} \\ |\mathcal{A}| \ne 0}} \left[\sum_{\substack{\mathbf{I}_{\mathcal{A}} \le 0 \\ \mathbf{I}_{\mathcal{W}-\mathcal{A}} > 0}} P_r(\mathbf{I}_{\mathcal{A}}, I_{\mathcal{W}-\mathcal{A}}) E(\max_{j \in \mathcal{A}}\{w_j^{I_j}\} | \mathbf{I}_{\mathcal{A}}) \right], \tag{13}$$

where $w_n^{I_n}$ is the waiting time of an arriving customer requesting item n finding I_n inventory in the system. The following lemma gives us the expression for $E(w_n)$.

LEMMA 3: For any given S_n, $P_r(I_n) = \rho_n^{S_n - I_n}(1 - \rho_n)$ and $E(w_n) = \rho_n^{S_n}/\mu_n(1 - \rho_n)$, where $\rho_n = \tilde{\lambda}_n/\mu_n$.

The proof is omitted. To compute $E(\max_{j \in \mathcal{W}}\{w_j\})$, we need to know $E(\max_{j \in \mathcal{A}}\{w_j^{I_j}\}|\mathbf{I}_{\mathcal{A}})$ and $P_r(\mathbf{I}_{\mathcal{A}}, \mathbf{I}_{\mathcal{W}-\mathcal{A}})$. Note that the waiting time of a demand for item j when there are $-I_j$ backorders in the system at the time it arrives is Erlang- $(-I_j + 1)$, i.e.,

$$P_r(w_j^{I_j} > w) = e^{-\mu_j w} \sum_{i=0}^{-I_j} \frac{(\mu_j w)^i}{i!}. \tag{14}$$

Therefore, for any given $\mathbf{I}_{\mathcal{A}}$, the waiting times $w_j^{I_j}$, $j \in \mathcal{A}$, are independent, and, hence,

$$\begin{aligned}
E(\max_{j \in \mathcal{A}}\{w_j^{I_j}\}|\mathbf{I}_{\mathcal{A}}) &= \sum_{j \in \mathcal{A}} E(w_j^{I_j}|I_j) - \sum_{\substack{i \neq j \\ i,j \in \mathcal{A}}} E(\min\{w_i^{I_i}, w_j^{I_j}\}|I_i, I_j) \\
&\quad + \cdots + (-1)^{|\mathcal{A}|-1} E(\min\{w_{j_1}^{I_{j_1}}, \ldots, w_{j_{|\mathcal{A}|}}^{I_{j_{|\mathcal{A}|}}}\}|\mathbf{I}_{\mathcal{A}}) \\
&= \sum_{j \in \mathcal{A}} \int_0^\infty P_r(w_j^{I_j} > w)\, dw - \sum_{\substack{i \neq j \\ i,j \in \mathcal{A}}} \int_0^\infty P_r(w_i^{I_i} > w) P_r(w_j^{I_j} > w)\, dw \\
&\quad + \cdots + (-1)^{|\mathcal{A}|-1} \int_0^\infty P_r(w_{j_1}^{I_{j_1}} > w) \cdots P_r(w_{j_{|\mathcal{A}|}}^{I_{j_{|\mathcal{A}|}}} > w)\, dw \\
&= \sum_{j \in \mathcal{A}} \frac{-I_j + 1}{\mu_j} - \Delta_{\mathbf{I}_{\mathcal{A}}}
\end{aligned} \tag{15}$$

where

$$\begin{aligned}
\Delta_{\mathbf{I}_{\mathcal{A}}} = & \sum_{\substack{i \neq j \\ i, j \in \mathcal{A}}} \sum_{k=0}^{-I_i} \sum_{l=0}^{-I_j} \frac{(k + l)!}{k! l!} \frac{\mu_i^k \mu_j^l}{(\mu_i + \mu_j)^{k+l+1}} \\
& - \cdots + (-1)^{|\mathcal{A}|} \sum_{k_1=0}^{-I_{j_1}} \cdots \sum_{k_{|\mathcal{A}|}=0}^{-I_{j_{|\mathcal{A}|}}} \frac{(k_1 + \cdots + k_{|\mathcal{A}|})! \mu_{j_1}^{k_1} \cdots \mu_{j_{|\mathcal{A}|}}^{k_{|\mathcal{A}|}}}{k_1! \cdots k_{|\mathcal{A}|}!(\mu_{j_1} + \cdots + \mu_{j_{|\mathcal{A}|}})^{k_1 + \cdots + k_{|\mathcal{A}|}+1}}.
\end{aligned} \tag{16}$$

Now it remains to find $P_r(\mathbf{I})$. Since the stochastic process $\mathbf{I}$ is a Markov process, $P_r(\mathbf{I})$ can be computed by solving the following equations for all $\mathcal{B} \subseteq \Omega$:

$$\left(\sum_{\mathscr{W}\subseteq\Omega}\lambda_{\mathscr{W}}+\sum_{n\in\mathscr{B}}\mu_n\right)P_r(I_1, I_2, \ldots, I_N)$$

$$= \sum_{n=1}^{N}\mu_n P_r(I_1, \ldots, I_{n-1}, I_n - 1, I_{n+1}, \ldots, I_N) + \sum_{\mathscr{W}\subseteq\Omega}\lambda_{\mathscr{W}}P_r(I'_1, I'_2, \ldots, I'_N), \quad (17)$$

where

$$I_n\begin{cases} <S_n & \text{if } n \in \mathscr{B}, \\ =S_n & \text{otherwise,}\end{cases} \quad (18)$$

$$I'_n = \begin{cases} I_n + 1 & \text{if } n \in \mathscr{W}, \\ I_n & \text{otherwise.}\end{cases} \quad (19)$$

Of course, we have to truncate the state space at some point in order to compute $P_r(\mathbf{I})$. Using (13) and Lemma 1 along with (15) and (17), we can compute $t(\mathbf{S})$.

With $E(w_n)$ computed, we can easily obtain the individual expect total time delay,

$$t_{ind}(\mathbf{S}) = \sum_{n=1}^{N}\tilde{\lambda}_n E(w_n) = \sum_{n=1}^{N}\frac{\rho_n^{S_n+1}}{1-\rho_n}. \quad (20)$$

4.2. Comparisons between $t(\mathbf{S})$ and $t_{ind}(\mathbf{S})$

To see the difference between $t(\mathbf{S})$ and $t_{ind}(\mathbf{S})$, we examine some three-item examples. Here $\Omega = \{1, 2, 3\}$ and all the possible demand types are $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$, $\{1, 2, 3\}$. In other words, we have seven possible types of demand requiring one item, two different items or three different items, and

$$t(\mathbf{S}) = t_{ind}(\mathbf{S}) - \sum_{i,j\in\{1,2,3\}}(\lambda_{\{i,j\}} + \lambda_{\{1,2,3\}})\sum_{I_i,I_j\le 0}P_r(I_i, I_j)\sum_{k=0}^{-I_i}\sum_{l=0}^{-I_j}\frac{(k+l)!\,\mu_i^k\mu_j^l}{k!\,l!\,(\mu_i+\mu_j)^{k+l+1}}$$

$$+ \lambda_{\{1,2,3\}}\sum_{I_1,I_2,I_3\le 0}P_r(I_1, I_2, I_3)\sum_{k=0}^{-I_1}\sum_{l=0}^{-I_2}\sum_{p=0}^{-I_3}\frac{(k+l+p)!\,\mu_1^k\mu_2^l\mu_3^p}{k!\,l!\,p!\,(\mu_1+\mu_2+\mu_3)^{k+l+p+1}}. \quad (21)$$

Let $\tilde{\lambda}_n = 30$, $\mu_n = 60$, and $S_n = 1$, for $n = 1, 2, 3$. First we treat the items independently and compute $t_{ind}(1, 1, 1)$. In this example, $t_{ind}(1, 1, 1) = 1.5$. For different combinations of λ_1, λ_2, λ_3, $\lambda_{\{1,2\}}$, $\lambda_{\{1,3\}}$, $\lambda_{\{2,3\}}$, and $\lambda_{\{1,2,3\}}$, we compute $t(1, 1, 1)$ and compare $t(1, 1, 1)$ with $t_{ind}(1, 1, 1)$. We report $t(1, 1, 1)$ and the percentage errors caused by using $t_{ind}(1, 1, 1)$ as the time delay while items are correlated in Table 1. Results show that, assuming items are independent while they are actually correlated, one may greatly overestimate the time delay. Furthermore, as the degree of correlation increases, i.e., as $\lambda_{\{1,2,3\}}$ increases, so does the error.

In the same case where $\lambda_{\{1,2,3\}} = 30$ (i.e., all demands require all three items), we also compute the exact time delay and the individual time delay for a range of $\mathbf{S}$ and report the results along with the percentage errors in Table 2. The results clearly indicate that, when items are

Table 1. Comparison of $t_{ind}(\mathbf{S})$ and $t(\mathbf{S})$: Part 1.

λ_1	λ_2	λ_3	$\lambda_{\{1,2\}}$	$\lambda_{\{1,3\}}$	$\lambda_{\{2,3\}}$	$\lambda_{\{1,2,3\}}$	$t(1, 1, 1)$ [% error using $t_{ind}(1, 1, 1)$]
30	30	30	0	0	0	0	1.5 (0)
20	20	20	0	0	0	10	1.3858 (8.24)
10	10	10	0	0	0	20	1.244 (20.58)
0	0	0	0	0	0	30	1.1184 (34.12)

highly correlated, independent time delay significantly overestimates the actual time delay (as high as 60% in our examples). The results also reveal that the relative errors become smaller as the service level increases. This is not surprising because as the service level increases, the possibility of having two or more items out of stock becomes smaller and therefore $t_{ind}(\mathbf{S})$ may be a good upper bound for the exact time delay.

However, difficulty in evaluating the expected total time delay occurs as N increases and it becomes prohibitive to solve a system of equations to obtain $P_r(\mathbf{I})$. From our experience, sometimes it takes almost a day to compute the probabilities on a Sun Sparc 20 workstation when $\mathbf{S}$ is large (e.g., $S_n = 6$) even in three-item cases. In practice, many systems involve thousands of items. Therefore, it is important to see if $t^l(\mathbf{S})$ provides good approximations for the exact time delay.

4.3. Comparisons between t^l(S) and t(S): $N = 3$

In three-item systems, we only need to consider $t^2(\mathbf{S})$, where

$$t^2(\mathbf{S}) = t_{ind}(\mathbf{S}) - \sum_{i,\, j \in \{1,2,3\}} (\lambda_{\{i,\, j\}} + \lambda_{\{1,2,3\}}) \sum_{I_i, I_j \leq 0} P_r(I_i,\, I_j) \sum_{k=0}^{-I_i} \sum_{l=0}^{-I_j} \frac{(k + l)!\, \mu_i^k \mu_j^l}{k!\, l!\, (\mu_i + \mu_j)^{k+l+1}}.$$

Instead of computing $P_r(I_1, I_2, I_3)$, $t^2(\mathbf{S})$ requires only $P_r(I_i, I_j)$, for all $i, j \in \{1, 2, 3\}$. For any $i, j \in \{1, 2, 3\}$, let

$$k = \{1, 2, 3\} - \{i, j\},$$

$$\lambda_i' = \lambda_i + \lambda_{\{i,k\}},$$

Table 2. Comparison of $t_{ind}(\mathbf{S})$ and $t(\mathbf{S})$: Part 2.

S_1	S_2	S_3	$t(\mathbf{S})$	$t_{ind}(\mathbf{S})$ (% error)	$t^2(\mathbf{S})$ (% error)
0	0	0	1.8464	3.0000 (62.48)	1.5372 (16.75)
1	0	0	1.6797	2.5000 (48.84)	1.4954 (10.97)
1	1	0	1.4606	2.0000 (36.93)	1.3421 (8.11)
1	1	1	1.1184	1.5000 (34.12)	1.0355 (7.41)
2	1	1	0.9760	1.2500 (28.07)	0.9297 (4.74)
2	2	1	0.8111	1.0000 (23.29)	0.7800 (3.83)
2	2	2	0.6096	0.7500 (23.03)	0.5863 (3.82)
3	3	2	0.4287	0.5000 (16.63)	0.4193 (2.19)
4	4	2	0.3401	0.3750 (10.26)	0.3368 (0.97)
6	4	2	0.3079	0.3281 (6.56)	0.3070 (0.29)

Table 3. Data sets: $\tilde{\lambda}_n = 30$, $n = 1, 2, 3$.

	λ_1	λ_2	λ_3	$\lambda_{\{1,2\}}$	$\lambda_{\{1,3\}}$	$\lambda_{\{2,3\}}$	$\lambda_{\{1,2,3\}}$
Case 1	10	10	10	10	10	10	0
Case 2	20	20	20	0	0	0	10
Case 3	0	0	0	10	10	10	10
Case 4	0	0	0	5	5	5	20
Case 5	10	10	10	0	0	0	20
Case 6	0	0	0	0	0	0	30

$$\lambda_j' = \lambda_j + \lambda_{\{j,k\}},$$

$$\lambda_{\{i,j\}}' = \lambda_{\{i,j\}} + \lambda_{\{1,2,3\}}.$$

Solving a series of equations for all i, j using λ_i', λ_j' and $\lambda_{\{i,j\}}'$, we can obtain $P_r(I_i, I_j)$. Even though we need to solve three subproblems with respect to {1, 2}, {1, 3}, and {2, 3}, the burden of the computation is *much* lower compared with the computation of the probability involving three items.

For the example in Table 2, we also compute $t^2(\mathbf{S})$ and report the percentage errors of $t^2(\mathbf{S})$. Although $t_{ind}(\mathbf{S})$ does not perform well as we expect, $t^2(\mathbf{S})$ is very accurate, especially as service level becomes high.

In the rest of the section, we concentrate on $t^2(\mathbf{S})$ since we have already made the comparisons between $t(\mathbf{S})$ and $t_{ind}(\mathbf{S})$ in the previous section. We construct some new examples. We fix $\tilde{\lambda}_n = 30$, for $n = 1, 2, 3$, and construct six examples, Cases 1–6, with different combinations of λ_1, λ_2, λ_3, $\lambda_{\{1,2\}}$, $\lambda_{\{1,3\}}$, $\lambda_{\{2,3\}}$, and $\lambda_{\{1,2,3\}}$ (but keep the items identical), as shown in Table 3. In Cases 7–10, we let $\tilde{\lambda}_1 = 40$, $\tilde{\lambda}_2 = 30$, $\tilde{\lambda}_3 = 20$, $\mu_1 = 60$, $\mu_2 = 50$, $\mu_3 = 40$. We construct four examples with different combinations of demand types, as shown in Table 4.

In Cases 1–6, we change the processing times. For $\mu_n = 55, 70, 90$, $n = 1, 2, 3$, we compute $t^2(2, 2, 2)$ and compare $t^2(2, 2, 2)$ with the exact time delay $t(2, 2, 2)$. The results are shown in Table 5. Note that, $t^2(\mathbf{S})$ is the exact time delay in Case 1, where $\lambda_{\{1,2,3\}} = 0$. In Cases 7–10, we change $\mathbf{S}$. For $\mathbf{S} = (3, 2, 1), (1, 2, 3), (2, 2, 2)$, we compute $t^2(\mathbf{S})$ and the exact time delay and compare $t^2(\mathbf{S})$ with the exact time delay. The results are shown in Table 6.

In all the examples, we see the following,

- $t^2(\mathbf{S})$ performs very well, especially when the service level is not too low.
- $t^2(\mathbf{S})$ is improved (1) as $\lambda_{\{1,2,3\}}$ decreases, i.e., the demands that require three items decrease, and (2) as μ_n increases, i.e., the utilization of the facilities decreases.

4.4. Performance of the Bounds: $N > 3$

When $N > 3$, it is extremely challenging to compute the exact time delay $t(\mathbf{S})$. Therefore, we will examine how the bounds are improved as we include more demand information and

Table 4. Data sets: $\tilde{\lambda}_1 = 40$, $\tilde{\lambda}_2 = 30$, $\tilde{\lambda}_1 = 20$, $\mu_1 = 60$, $\mu_2 = 50$, $\mu_3 = 40$.

	λ_1	λ_2	λ_3	$\lambda_{\{1,2\}}$	$\lambda_{\{1,3\}}$	$\lambda_{\{2,3\}}$	$\lambda_{\{1,2,3\}}$
Case 7	20	15	0	5	10	5	5
Case 8	20	10	0	5	5	5	10
Case 9	14	5	0	8	3	2	15
Case 10	15	5	0	5	0	0	20

Table 5. Comparison of $t^2(\mathbf{S})$ and $t(\mathbf{S})$: $\mu_n = 55, 70, 90$ and $\mathbf{S} = \{2, 2, 2\}$.

	$\mu_n = 55$		$\mu_n = 70$		$\mu_n = 90$	
	$t^2(\mathbf{S})$(% error)	$t(\mathbf{S})$	$t^2(\mathbf{S})$ (% error)	$t(\mathbf{S})$	$t^2(\mathbf{S})$ (% error)	$t(\mathbf{S})$
Case 1	1.0124 (0)	1.0124	0.3963 (0)	0.3963	0.1612 (0)	0.1612
Case 2	1.0124 (0.58)	1.0183	0.3963 (0.35)	0.3977	0.1612 (0.26)	0.1616
Case 3	0.9251 (0.93)	0.9338	0.3665 (0.62)	0.3688	0.1497 (0.50)	0.1504
Case 4	0.8686 (2.47)	0.8906	0.3457 (1.85)	0.3522	0.1410 (1.62)	0.1433
Case 5	0.9251 (1.99)	0.9439	0.3665 (1.45)	0.3719	0.1497 (1.25)	0.1516
Case 6	0.8017 (4.89)	0.8429	0.3201 (3.93)	0.3332	0.1300 (3.70)	0.1350

compare the bounds with simulated expected total waiting times. We consider some six item problems where all parts are demanded at the same time, i.e., $\tilde{\lambda}_n = 30$ for all n and $\lambda_{\{1,2,3,4,5,6\}} = 30$. This is the case with the highest level of correlation. For $\mu_n = 50, 70, 90$, $n = 1, \ldots, 6$, we compute the bounds $[0, t_{ind}(\mathbf{S})]$, $[t^2(\mathbf{S}), (t_{ind}(\mathbf{S}) + t^2(\mathbf{S}))/2]$ (bounds with $t^2(\mathbf{S})$), and $[(t^2(\mathbf{S}) + t^3(\mathbf{S}))/2, \min\{(t_{ind}(\mathbf{S}) + t^2(\mathbf{S}))/2, t^3(\mathbf{S})\}]$ (bounds with $t^3(\mathbf{S})$). The length of the intervals and the percentage improvement as more demand information is included are computed and reported in Table 7 along with the bounds. In our simulation, each case was run for 10 replications of 600 time units plus 60 time units of warmup. The simulated expected total waiting time and the 95% confidence interval width are reported in the last column in Table 7. As we can see, the bounds are greatly improved with added demand information. The results agree with the conclusions drawn from the three item examples. Notice that these are the examples with the highest level of correlation. One would expect that the bounds perform much better in cases with moderate or low levels of correlation.

5. NONEXPONENTIAL PROCESSING TIME

The bounds in inequalities (6) and (7) or inequalities (9) and (10) provide us with considerable insight and computational advantage for the exponential case. Even though the derivation of the bounds themselves is not limited to the exponential case, computing the bounds in any nonexponential case is extremely challenging due to the dependency of inventory levels between the supply systems. So far no exact procedure is available that can be used to compute the state probability $P_r(\mathbf{I})$ as in the exponential case. However, the bounds obtained from the exponential case may serve as good approximations for some non-exponential cases, in particular, for some $M/G/1$ type of supply systems.

For any $M/G/1$ queue with an arrival rate λ, the average flow time $E(T)$ can be computed using

Table 6. Comparison of $t^2(\mathbf{S})$ and $t(\mathbf{S})$ under different $\mathbf{S}$.

	$S_1 = 3, S_2 = 2, S_3 = 1$		$S_1 = 1, S_2 = 2, S_3 = 3$		$S_1 = 2, S_2 = 2, S_3 = 2$	
	$t^2(\mathbf{S})$ (% error)	$t(\mathbf{S})$	$t^2(\mathbf{S})$ (% error)	$t(\mathbf{S})$	$t^2(\mathbf{S})$ (% error)	$t(\mathbf{S})$
Case 7	1.5302 (0.27)	1.5343	1.9246 (0.11)	1.9268	1.5951 (0.22)	1.5986
Case 8	1.4947 (0.62)	1.5041	1.8957 (0.27)	1.9009	1.5650 (0.51)	1.5730
Case 9	1.4531 (1.09)	1.4691	1.8522 (0.44)	1.8603	1.5212 (0.89)	1.5348
Case 10	1.4225 (1.59)	1.4455	1.8346 (0.63)	1.8462	1.4959 (1.33)	1.5161

Table 7. Comparison of bounds under different **S** and μ_n.

S	μ_n	$t_{ind}(\mathbf{S})$	$t^2(\mathbf{S})$	$t^3(\mathbf{S})$	bounds with $t^2(\mathbf{S})$	bounds with $t^3(\mathbf{S})$	$t(\mathbf{S})$
(1, 2, 3, 1, 2, 3)	50	3.5280	0.7939	2.5475	[0.7939, 2.1610] 1.3671 (61%)	[1.6707, 2.1610] 0.4903 (64%)	1.821 (0.0905)
	70	1.0364	0.4531	0.7465	[0.4531, 0.7448] 0.2917 (72%)	[0.5998, 0.7448] 0.1450 (50%)	0.6398 (0.0280)
	90	0.4815	0.2526	0.3534	[0.2526, 0.3670] 0.1144 (76%)	[0.3030, 0.3534] 0.0504 (44%)	0.3218 (0.0109)
(2, 2, 2, 2, 2, 2)	50	3.2400	0.6202	2.4212	[0.6202, 1.9301] 1.3099 (60%)	[1.5207, 1.9301] 0.4094 (69%)	1.6405 (0.0916)
	70	0.8269	0.3038	0.6132	[0.3038, 0.5654] 0.2616 (68%)	[0.4585, 0.5654] 0.1069 (59%)	0.4881 (0.0287)
	90	0.3333	0.1383	0.2482	[0.1383, 0.2358] 0.0975 (70%)	[0.1933, 0.2358] 0.0425 (56%)	0.2106 (0.0103)
(3, 3, 3, 2, 2, 2)	50	2.5920	0.7661	1.9006	[0.7661, 1.6791] 0.9130 (35%)	[1.3334, 1.6791] 0.3457 (42%)	1.4042 (0.0839)
	70	0.5904	0.2926	0.4413	[0.2926, 0.4415] 0.1489 (75%)	[0.3670, 0.4413] 0.0743 (50%)	0.3803 (0.0254)
	90	0.2222	0.1245	0.1691	[0.1245, 0.1734] 0.0489 (78%)	[0.1469, 0.1692] 0.0223 (46%)	0.1553 (0.0089)

$$E(T) = \frac{\lambda E(B^2)}{2(1-\rho)} + E(B),$$

where ρ is the system utilization and B the processing time (Buzacott and Shanthikumar [1]). For the same arrival rate λ, there exits an $M/M/1$ queue with parameter μ such that the average flow time in this $M/M/1$ queue, $1/(\mu - \lambda)$, matches the actual average flow time $E(T)$. We compute the bounds for this exponential system and use these bounds to approximate the actual expected total waiting times.

As an example, we consider another distribution commonly used for approximating processing times, the gamma distribution, which is characterized by a shape parameter α and a scale parameter β. For given α and β, the mean and variance are given by $\alpha\beta$ and $\alpha\beta^2$, the coefficient of variation (CV) is $1/\sqrt{\alpha}$, and the average flow time of such an $M/G/1$ queue can be computed as

$$E(T) = \frac{\lambda E(B^2)}{2(1-\rho)} + E(B) = \alpha\beta\left[\frac{\beta(1+\alpha)\lambda}{2(1+\rho)}\right].$$

For numerical comparisons, we consider the examples with the same demand process as in Section 4.4, where $N = 6$, $\tilde{\lambda}_n = 30$ for all n, and $\lambda_{\{1,2,3,4,5,6\}} = 30$. We select α so that $CV = 0.75$, $\sqrt{0.75}$, $\sqrt{1.25}$, 1.25 (note that $CV = 1$ in the exponential case). For different β values, the average flow time $E(T)$ in each queue and the matching parameter μ for the $M/M/1$ queue are given in Table 8.

For $\mu_n = 50, 70, 90$, $n = 1, \ldots, 6$, and $\mathbf{S} = (1, 2, 3, 1, 2, 3), (2, 2, 2, 2, 2, 2), (3, 3, 3, 2, 2, 2)$, $t_{ind}(\mathbf{S})$, $t^2(\mathbf{S})$, and $t^3(\mathbf{S})$ computed using the procedure in Section 4 can be found in Table 7. Again, this is the case with the highest level of correlation. We then simulate these

Table 8. Data for gamma distribution.

α	CV	β	$E(T)$	μ
1.77778	0.75	0.011912	0.05	50
		0.008520	0.025	70
		0.006588	0.016668	90
1.33333	$\sqrt{0.75}$	0.015480	0.05	50
		0.011065	0.025	70
		0.008579	0.016668	90
0.80000	$\sqrt{1.25}$	0.024300	0.05	50
		0.017350	0.025	70
		0.013520	0.016668	90
0.64000	1.25	0.029385	0.05	50
		0.020965	0.025	70
		0.016394	0.016668	90

systems with gamma processing times. Each case was run for 10 independent replications of 600 time units plus 60 time units of warmup. The simulated expected total waiting times and the 95% confidence interval width are presented in Table 9. Due to errors in approximation (i.e., treating processing times as exponential), some $t(\mathbf{S})$ values are above (below) the upper (lower) bounds when $CV > 1$ ($CV < 1$), as one would expect. In general, the upper bounds (lower bounds) provide more accurate information than the lower bounds (upper bounds) when $CV > 1$ ($CV < 1$). So the bounds $t^l(\mathbf{S})$ are probably more accurate than the bounds given by (9) and (10) in nonexponential settings. However, the simulated expected total waiting times fall into the intervals $[t^2(\mathbf{S}), t^3(\mathbf{S})]$ in most of the cases. Even in the cases where the simulated expected total waiting times are below the lower bounds $t^2(\mathbf{S})$ [above the upper bounds $t^3(\mathbf{S})$], the lower bounds $t^2(\mathbf{S})$ [upper bounds $t^3(\mathbf{S})$] provide good approximations as indicated by the percentage errors in Table 9.

In general, the upper bounds (when $CV > 1$) or lower bounds (when $CV < 1$) provide reasonably good approximations in our numerical experiments. We conjecture that the bounds would provide good approximations if the CV of the processing times is not too far away from 1.

Table 9. Simulated expected total waiting times $t(\mathbf{S})$.

CV	μ_n	$\mathbf{S} = (1, 2, 3, 1, 2, 3)$	$\mathbf{S} = (2, 2, 2, 2, 2, 2)$	$\mathbf{S} = (3, 3, 3, 2, 2, 2)$
0.75	50	1.5635	1.3500	1.1519
	70	0.5132	0.3533	0.2738 (−6.87%)
	90	0.2543	0.1423	0.1037 (−20.06%)
$\sqrt{0.75}$	50	1.6671	1.4619	1.2496
	70	0.5654	0.4071	0.3152
	90	0.2822	0.1697	0.1237 (−0.65%)
$\sqrt{1.25}$	50	1.9549	1.7820	1.5374
	70	0.6990	0.5541	0.4349
	90	0.3601(1.86%)	0.2472	0.1834 (7.80%)
1.25	50	2.1254	1.9773	1.7042
	70	0.7751(3.67%)	0.6344(3.34%)	0.5064 (12.86%)
	90	0.3913(9.69%)	0.2818(11.92%)	0.2098 (19.40%)

6. CONCLUSION

This study focused on multi-item continuous review inventory systems where a customer order may require several items and customer satisfaction is measured by the expected total time delay seen by the customers per year. Assuming each type of demand follows an independent Poisson process, replenishment lead times for each item are independent, and base stock policies are used, we formulated the expected total time delay (exact time delay). We then compared the exact time delay with the expected total time delay when the items are treated independently although they are actually correlated (individual time delay) and showed that the individual time delay is an upper bound of the exact time delay. In general, the error can be very big. However, it is extremely difficult to obtain the demand information for all demand types in practice, and the computation involved to evaluate the exact time delay is very complex and prohibitively time consuming when the number of items in the system becomes large. We explore the possibility of including the demand information partially; i.e., we approximate the exact time delay by the bounds.

In the case where each item is manufactured by a dedicated single machine facility and processing times are exponential at all facilities, we develop a procedure to compute the exact time delay and the time delay with partial demand information (bounds). These bounds also provide good approximations for some non-exponential cases. The main conclusions that we drew from this research are:

- One may greatly overestimate the time delay if the correlations among the items are ignored. However, the errors get smaller as the time delay decreases.
- The time delay based on partial demand information offers good approximation, especially when the time delay is not too high. The relative errors decrease as the degree of the correlation decreases or the replenishment lead times decrease.

APPENDIX

PROOF OF LEMMA 1: Let

$$\Delta^l_{\mathcal{W}} = \sum_{\substack{\mathcal{B}\subseteq\mathcal{W}\\|\mathcal{B}|=l+1}} E(\min_{j\in\mathcal{B}}\{w_j\}) - \sum_{\substack{\mathcal{B}\subseteq\mathcal{W}\\|\mathcal{B}|=l+2}} E(\min_{j\in\mathcal{B}}\{w_j\}) + \cdots + (-1)^{|\mathcal{W}|-l-1} E(\min_{j\in\mathcal{W}}\{w_j\}),$$

$l = 2, 3, \ldots, |\mathcal{W}| - 1$, so that

$$\Delta_{\mathcal{W}} = \sum_{\substack{i,j\in\mathcal{W}\\i\neq j}} E(\min\{w_i, w_j\}) - \cdots + (-1)^{l-1}\Delta^l_{\mathcal{W}}.$$

We know that $\Delta_{\mathcal{W}} > 0$ as long as the lead times have positive means. Actually $\Delta^l_{\mathcal{W}} > 0$ when the lead times have positive means. This follows from the fact that $\Delta^l_{\mathcal{W}}$ is the integral of a positive probability function.

Following (4) and substituting $\Delta^l_{\mathcal{W}}$ with

$$\Delta^l_{\mathcal{W}} = \sum_{\substack{\mathcal{B}\subseteq\mathcal{W}\\|\mathcal{B}|=l+1}} E(\min_{j\in\mathcal{B}}\{w_j\}) - \Delta^{l+1}_{\mathcal{W}}$$

for $l = 2, \ldots, |\mathcal{W}| - 2$, we can rewrite $t(\mathbf{S})$ as

$$t(\mathbf{S}) = t_{ind}(\mathbf{S}) - \sum_{\substack{\mathcal{W}\subseteq\Omega\\|\mathcal{W}|>1}} \lambda_{\mathcal{W}}\Delta_{\mathcal{W}}$$

$$= t_{ind}(\mathbf{S}) - \sum_{\substack{\mathscr{W}\subseteq\Omega\\|\mathscr{W}|=2}} \lambda_{\mathscr{W}}\Delta_{\mathscr{W}} - \sum_{k=3}^{N}\sum_{\substack{\mathscr{W}\subseteq\Omega\\|\mathscr{W}|=k}} \lambda_{\mathscr{W}}\Delta_{\mathscr{W}}$$

$$= t_{ind}(\mathbf{S}) - \sum_{\substack{\mathscr{W}\subseteq\Omega\\|\mathscr{W}|=2}} \lambda_{\mathscr{W}}\left[\sum_{i,\,j\in\mathscr{W}} E(\min\{w_i,\ w_j\})\right]$$

$$- \sum_{k=3}^{N}\sum_{\substack{\mathscr{W}\subseteq\Omega\\|\mathscr{W}|=k}} \lambda_{\mathscr{W}}\left[\sum_{i,\,j\in\mathscr{W}} E(\min\{w_i,\ w_j\}) - \Delta^2_{\mathscr{W}}\right]$$

$$= t_{ind}(\mathbf{S}) - \sum_{\substack{i,\,j\in\Omega\\i\neq j}}\left(\sum_{\substack{\mathscr{W}\subseteq\Omega\\i,\,j\in\mathscr{W}}} \lambda_{\mathscr{W}}\right) E(\min\{w_i,\ w_j\}) + \sum_{k=3}^{N}\sum_{\substack{\mathscr{W}\subseteq\Omega\\|\mathscr{W}|=k}} \lambda_{\mathscr{W}}\Delta^2_{\mathscr{W}}$$

$$\vdots$$

$$= t_{ind}(\mathbf{S}) - \sum_{\substack{i,\,j\in\Omega\\i\neq j}}\left(\sum_{\substack{\mathscr{W}\subseteq\Omega\\i,\,j\in\mathscr{W}}} \lambda_{\mathscr{W}}\right) E(\min\{w_i,\ w_j\})$$

$$+ \sum_{\substack{i,\,j,k\in\Omega\\i\neq j\neq k}}\left(\sum_{\substack{\mathscr{W}\subseteq\Omega\\i,\,j,k\in\mathscr{W}}} \lambda_{\mathscr{W}}\right) E(\min\{w_i,\ w_j,\ w_k\}) + \cdots + (-1)^{|\Omega|-1}\lambda_\Omega E(\min_{j\in\Omega}\{w_j\})$$

$$= t_{ind}(\mathbf{S}) - \sum_{\substack{i,\,j\in\Omega\\i\neq j}} \tilde{\lambda}_{\{i,\,j\}} E(\min\{w_i,\ w_j\}) + \sum_{\substack{i,\,j,k\in\Omega\\i\neq j\neq k}} \tilde{\lambda}_{\{i,\,j,k\}} E(\min\{w_i,\ w_j,\ w_k\})$$

$$+ \cdots + (-1)^{|\Omega|-1}\lambda_\Omega E(\min_{j\in\Omega}\{w_j\}).$$

PROOF OF THEOREM 2: To simplify the proof, we let

$$\Sigma_l = \sum_{\substack{\mathscr{B}\subseteq\Omega\\|\mathscr{B}|=l}} \tilde{\lambda}_{\mathscr{B}} E(\min_{j\in\mathscr{B}}\{w_j\})$$

be the lth sum in (5).

Inequalities (6) and (7) follow directly from Lemma 2 and

$$t^l(\mathbf{S}) - t^{l+2}(\mathbf{S}) = \begin{cases} \Sigma_{l+1} - \Sigma_{l+2} \geq 0 & \text{if } l \text{ is an odd number,} \\ \Sigma_{l+2} - \Sigma_{l+1} \leq 0 & \text{if } l \text{ is an even number.} \end{cases}$$

By Lemma 2,

$$|t(\mathbf{S}) - t^l(\mathbf{S})| - |t(\mathbf{S}) - t^{l+1}(\mathbf{S})| = \Sigma_{l+1} - \Sigma_{l+2} + \Sigma_{l+3} - \cdots + (-1)^{|\Omega|-l+1}\Sigma_{|\Omega|}$$
$$- [\Sigma_{l+2} - \Sigma_{l+3} + \cdots + (-1)^{|\Omega|-l}\Sigma_{|\Omega|}] \geq \Sigma_{l+1} - 2\Sigma_{l+2}.$$

For any given $\mathscr{B}$ with cardinality $l + 1$, there exist $l + 1$ subsets of $\mathscr{B}$ with cardinality l and

$$\sum_{\substack{\mathscr{C}\subset\mathscr{B}\subseteq\Omega\\|\mathscr{C}|=l}} \tilde{\lambda}_{\mathscr{C}} E(\min_{j\in\mathscr{C}}\{w_j\}) \geq 2\tilde{\lambda}_{\mathscr{B}} E(\min_{j\in\mathscr{B}}\{w_j\})$$

because $l \geq 1$ and $\tilde{\lambda}_{\mathscr{C}} \geq \tilde{\lambda}_{\mathscr{B}}$. Hence, $\Sigma_{l+1} - 2\Sigma_{l+2} \geq 0$ and $|t(\mathbf{S}) - t^l(\mathbf{S})|$ is decreasing in l by Lemma 2.

REFERENCES

[1] J.A. Buzacott and J.G. Shanthikumar, Stochastic models of manufacturing systems, Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] K.L. Cheung and W.H. Hausman, Multiple failures in a multi-item spare inventory model, IIE Trans 27 (1995), 171–180.

[3] S. Graves, A multiple-item inventory model with a job completion criterion, Manage Sci 28 (1982), 1334–1337.

[4] W.H. Hausman, H.L. Lee, and A.X. Zhang, Joint demand fulfillment probability in a multi-item inventory system with independent order-up-to policies, Eur J Oper Res 109 (1998), 646–659.

[5] P. Glasserman and Y. Wang, Leadtime-inventory trade-offs in assemble-to-order systems, Oper Res 46 (1998), 858–871.

[6] J.W. Mamer and A. Shogan, A constrained capital budgeting problem with applications to repair kit selection, Manage Sci 33 (1987), 800–806.

[7] J.W. Mamer and S.A. Smith, Optimizing field repair kits based on job completion rate, Manage Sci 28 (1982), 1328–1333.

[8] J.W. Mamer and S.A. Smith, Job completion based inventory systems: Optimal policies for repair kits and spare machines, Manage Sci 31 (1985), 703–718.

[9] S.T. March and G.D. Scudder, On optimizing field repair kits based on job completion rate, Manage Sci 30 (1984), 1025–1028.

[10] S.A. Smith, J.C. Chambers, and E. Shlifer, Optimal inventories based on job completion rate for repairs requiring multiple items, Manage Sci 26 (1980), 849–852.

[11] J. Song, On the order fill rate in a multi-item inventory system, Oper Res 46 (1998), 831–845.

[12] J. Song, S. Xu, and B. Liu, Order-fulfillment performance measures in an assemble-to-order system with stochastic leadtimes, Oper Res 47 (1999), 131–149.

[13] A. Svoronos and P. Zipkin, Evaluation of one-for-one replenishment policies for multiechelon inventory systems, Manage Sci 37 (1991), 68–83.