A Componentwise Index of Service Measurement in Multi-Component Systems

Sridhar Seshadri,[1] Jayashankar M. Swaminathan[2]

[1] *Stern School of Business, New York University, New York, New York 10012*

[2] *Kenan-Flagler Business School, University of North Carolina,*
Chapel Hill, North Carolina 27599

Received 22 March 2000; revised 14 June 2002; accepted 26 July 2002

DOI 10.1002/nav.10050

Abstract: In this paper we present a componentwise delay measure for estimating and improving the expected delays experienced by customers in a multi-component inventory/assembly system. We show that this measure is easily computed. Further, in an environment where the performance of each of the item delays could be improved with investment, we present a solution that aims to minimize this measure and, in effect, minimizes the average waiting time experienced by customers. © 2002 Wiley Periodicals, Inc. Naval Research Logistics 50: 184–194, 2003.

Keywords: multi-component inventory/assembly system; random lead times; fixed budget; greedy allocation

1. INTRODUCTION

Improving customer waiting times in a multi-item inventory/assembly system is a challenging task since demands for different items may be correlated to one another. This is particularly true in environments such as PC and electronic assembly where randomness in the delivery time may primarily be due to randomness in the kitting process. Researchers have worked on different aspects of this problem including performance evaluation, developing approximations for performance measures and system optimization subject to budgetary constraints (see Hausman [6], Song [10], Glasserman and Wang [5] and Song, Xu, and Liu [11], for example). In such an environment, it becomes difficult for managers to allocate resources to improve the delivery performance with regard to different items because their performance may be interrelated. Zhang [12] recently wrote an article in this context that has important implications for customer order management in many industries. Zhang showed that in a multi-item inventory system, when demands are correlated across items, the expected time delay experienced by customers for order fulfillment will be overestimated if the individual item demands are assumed to be

Correspondence to: S. Seshadri

independent. This result was proved by Zhang under certain assumptions. In this paper we show that the same result can be obtained under minimal assumptions. The main results in this paper are as follows: (1) We devise a componentwise measure for order fulfillment that correctly *estimates* the expected delay experienced by customers; (2) we show that this index has useful properties for improving customer service; and (3) we present an optimal greedy allocation algorithm for allocating resources to minimize the expected waiting times subject to a capital budget by utilizing the componentwise measures. It turns out that, under certain conditions, the optimal allocation is such that the item that can deliver the maximum marginal reduction in the componentwise indices is chosen as the candidate for investment at each step.

The rest of the paper is organized as follows. In Section 2, we present the model and show that minimal assumptions are needed for the result established by Zhang [12]. In Section 3, we present the componentwise index for measuring the waiting times. In Section 4, we consider the problem faced by a firm that has to minimize the expected customer waiting times subject to a budget constraint on the dollars available to reduce the delay. We provide an optimal greedy allocation algorithm for budget allocation. We provide our conclusions in Section 5.

2. MODEL

In this paper we consider a multi-item assembly system where we primarily focus our attention on a make to order environment in which items are ordered based on the actual customer demand. The set of items demanded by customers is denoted as $\Omega = \{1, 2, \ldots, N\}$. Customers demand a kit of items, where a kit is a subset of Ω with cardinality greater than one. We assume that once all the items in the kit are received, the order is ready for delivery to the customer. Without loss of generality, let S_j denote the jth subset of Ω, and let the collection of these subsets be all possible subsets of Ω. Let $|S_j|$ denote the cardinality of subset S_j and S_0 be the empty set. The kit corresponding to S_j will be called the type-j kit. We shall observe the order fulfillment process from some convenient starting point in time denoted as time zero.

Let $w_{i,k}$ be the waiting time for producing or procuring the kth type i item demanded by customers. Define $w_{i,0} = 0$ and $w_{0,n} = 0$ for all $n \geq 0$. The $w_{i,k}$'s will be called item-delays to distinguish from the delay of the kit or kit-delay. We use the term kit-delay synonymously with the term customer-delay. We make no assumptions about $w_{i,k}$'s. We provide a small three item example to illustrate the ideas as we develop the expression for the kit-delay experienced by customers.

Consider a system with three items 1, 2, and 3. In this system, for example, $w_{1,3}$ would be the waiting time associated with the third unit of item 1 and $w_{3,1}$ would be the waiting time associated with the first unit of item 3. There are four types of kits that are demanded by customers in this system, namely, $S_1 = (1, 2)$, $S_2 = (2, 3)$, $S_3 = (1, 3)$, and $S_4 = (1, 2, 3)$. Also assume that during the horizon T we observe demands for 5 kits in the following order, S_2, S_3, S_1, S_4, S_2, and the associated item-delays are given as follows: $w_{1,1} = 5$; $w_{1,2} = 4$; $w_{1,3} = 6$; $w_{2,1} = 3$; $w_{2,2} = 7$; $w_{2,3} = 4$; $w_{2,4} = 4$; $w_{3,1} = 2$; $w_{3,2} = 4$; $w_{3,3} = 5$; $w_{3,4} = 6$. These values are shown in Table 1.

Kits of each type are numbered in increasing sequence as 1, 2, ... as they are demanded by customers. Let $R_j(i, k)$ denote the index of the type j kit that has the kth type i item in it (else it equals to zero). For the system above for example, $R_1(1, 2) = 1$ because the first type S_1 kit contains the second unit of item 1. Similarly, let $Q_j(i, n)$ denote the index of the type i item in the nth type j kit (else equal to zero if the type i item is not part of this kit). So, $Q_2(3, 2) = 4$ because the 4th unit of item 3 is presented in the second unit of kit S_2 [clearly $Q_2(3, 1) =$

Table 1. Wait times and associated variables.

	Item wait times			Kits and R-values			
	1	2	3	S_1	S_2	S_3	S_4
$Order1 - S_2$		**3** $Q_2(2, 1) = 1$	**2** $Q_2(3, 1) = 1$		1 $R_2(2, 1)$ $R_3(3, 1)$		
$Order2 - S_3$	**5** $Q_3(1, 1) = 1$		**4** $Q_3(3, 1) = 2$			1 $R_3(1, 1)$ $R_3(3, 2)$	
$Order3 - S_1$	**4** $Q_1(1, 1) = 2$	**7** $Q_1(2, 1) = 2$		1 $R_1(1, 2)$ $R_1(2, 2)$			
$Order4 - S_4$	**6** $Q_4(1, 1) = 3$	**4** $Q_4(2, 1) = 3$	**5** $Q_4(3, 1) = 3$				1 $R_4(1, 3)$ $R_4(2, 3)$ $R_4(3, 3)$
$Order5 - S_2$		**4** $Q_2(2, 2) = 4$	**6** $Q_2(3, 2) = 4$		2 $R_2(2, 4)$ $R_2(3, 4)$		
Total delay	**15**	**18**	**17**	**7**	**9**	**5**	**6**

1 because the first unit of item 3 appeared in the first unit of kit S_2]. The complete set of nonzero values for Q and R variables for the above example is given below (also see Table 1):

$$R_1(1, 2) = 1; \quad R_1(2, 2) = 1; \quad R_2(2, 1) = 1; \quad R_2(3, 1) = 1;$$
$$R_2(2, 4) = 2; \quad R_2(3, 4) = 2; \quad R_3(1, 1) = 1; \quad R_3(3, 2) = 1;$$
$$R_4(1, 3) = 1; \quad R_4(2, 3) = 1; \quad R_4(3, 3) = 1.$$

$$Q_1(1, 1) = 2; \quad Q_1(2, 1) = 2; \quad Q_2(2, 1) = 1; \quad Q_2(3, 1) = 1;$$
$$Q_2(2, 2) = 4; \quad Q_2(3, 2) = 4; \quad Q_3(1, 1) = 1; \quad Q_3(3, 1) = 2;$$
$$Q_4(1, 1) = 3; \quad Q_4(2, 1) = 3; \quad Q_4(3, 1) = 3.$$

The delay experienced by a customer for the fulfillment of an order is not the sum of the $w_{i,k}$'s for that kit but the maximum of the delays of all items requested in the kit by that customer. To define the kit-delay in terms of the item-delay let $M_i(T)$ units of item i be demanded during the time interval $[0, T]$, $i = 1, 2, \ldots, N$. In the above example, $M_1(T) = 3$, $M_2(T) = 4$; $M_3(T) = 4$. Then the sum total of delay experienced by customers during $[0, T]$, say $D(T)$, is given by

$$D(T) = \sum_{i=1}^{N} \sum_{k=1}^{M_i(T)} \sum_{j>0} \max_{l \in S_j} \{w_{l,Q_j(l,R_j(i,k))}\}/|S_j|. \tag{1}$$

This expression looks exceedingly complex, but the idea behind it is quite simple but radical. In words, all it says is "to each item supplied to a customer assign a number that is equal to the kit-delay divided by the number of items in the kit containing this item." Thus, when all these numbers are added up over all items we automatically obtain the sum of all kit-delays. Notice that a given item goes into only one kit, therefore, only one term is nonzero in the innermost sum on the right-hand side of (1). Let the expression in the "max" operation in (1), i.e., $\max_{l \in S_j}\{w_{l,Q_j(l,R_j(i,k))}\}/|S_j|$, be written as a function: $g(i, k, j)$. In our example, when $i = 1$ and $k = 1$, the kit in question is $j = 3$, and the expression $g(1, 1, 3)$ evaluates to 5/2. Similarly, $g(1, 2, 1) = 7/2$, $g(1, 3, 4) = 6/3$, $g(2, 1, 2) = 3/2$, $g(2, 2, 3) = 7/2$, $g(2, 3, 4) = 6/3$, $g(2, 4, 2) = 6/2$, $g(3, 1, 2) = 3/2$, $g(3, 2, 3) = 5/2$, $g(3, 3, 4) = 6/3$, $g(3, 4, 2) = 6/2$. These add up to 27. Clearly in the example, the total delay experienced by the customers = max(3, 2) + max(5, 4) + max(4, 7) + max(6, 4, 5) + max(4, 6) = 3 + 5 + 7 + 6 + 6 = 27 units. The reader may wonder why do we not write this directly, instead of using the indices? The problem is that a computer or an accountant going over records essentially does what we did to compute the customer delays. Moreover, it is when writing out the customer delay in this fashion that we realize as we explain below that there is a simple but powerful bookkeeping technique that uses the same indices to determine which item was responsible for the kit-delay.

Now we turn to the problem faced when the only data we have is the sum of the item-delays for each item. It is clear that $D(T)$ as defined above is smaller than the sum of the w's, i.e.,

$$D(T) \leq \sum_{i=1}^{N} \sum_{k=1}^{M_i(T)} w_{i,k}. \tag{2}$$

This extends Theorem 1 of Zhang [12], who obtain similar results with Poisson demand processes. In the three-item example the sum of all item-delays is 50, which is the right-hand side of (2) whereas $D(T)$ is only 27. Therefore, it would be erroneous to set the kit-delay equal to the sum of the delays of the items that constitute the kit. In fact, any customer order fulfillment measure that is an increasing function of $D(T)$ will be overestimated if it is computed based on the componentwise *sums* of the $w_{i,k}$. Similarly, if the delay "penalty" assessed to each item in a kit is set proportional to its delay, then the sum total penalty over all kits supplied will only reflect the total of item-delays and not the total of kit-delays. Therefore, the utility of such a metric for improving waiting time performance is questionable.

We focused our attention on total delays, but the result follows for average delay once the number of arrivals or total demand for items is given. For example, if $O_j(T)$ represents the number of orders for type j kit during the horizon then the average delay over orders, $D(T)/\Sigma_j O_j(T)$ is less than or equal to $\Sigma_{i=1}^{N} \Sigma_{k=1}^{M_i(T)} w_{i,k}/\Sigma_j O_j(T)$. This result also extends to performance measures that are constructed using subsequences of delays over a given time interval and does not depend on any assumptions regarding the ordering or order fulfillment process. It turns out that if we attempt to correct the over estimate, we end up with a measure that is otherwise of interest for determining how to reduce the order fulfillment delay. This is explained in the next section.

3. COMPONENTWISE DELAY INDEX

Let us consider a componentwise delay index that is constructed as follows: The kth type i item is assigned a delay penalty $d_{i,k}$ equal to $w_{i,k}$ if that item is the one that delayed the delivery of the kit else it is assigned a delay penalty equal to zero. The reader should refer to Table 2 where we show the assignment of delays for the example in Table 1. The idea is once again

Table 2. Wait times and corresponding delay indices.

	Item wait times and **delay indices**			Kit			
	1	2	3	S_1	S_2	S_3	S_4
$Order1 - S_2$		3	2		1		
		3	**0**				
$Order2 - S_3$	5		4			1	
	5		**0**				
$Order3 - S_1$	4	7		**1**			
	0	**7**					
$Order4 - S_4$	6	4	5				1
	6	**0**	**0**				
$Order5 - S_2$		4	6		2		
		0	**6**				
Total delay index	**11/3**	**10/4**	**6/4**	**7**	**9**	**5**	**6**

simple: The item that delayed the kit should bear *all* the penalty for the delay. This concept has important implications for providing incentives to vendors or for system improvement as will be explained in Section 4. The component delay index ϕ_i could be expressed for each item or more generally a component of a kit as follows:

$$\phi_i(T) = \sum_{k=1}^{M_i(T)} d_{i,k}/M_i(T).$$

Then consider the sum of the componentwise delay penalties,

$$S(T) = \sum_{i=1}^{N} \phi_i(T)M_i(T).$$

It can be verified that the expected value of $S(T)$ is equal to the expected value of $D(T)$. In the three item example, the total customer delay is as expected equal to the sum of the delay penalties (see Table 2).

There is a discrepancy between $D(T)$ and $S(T)$ if and only if two or more items have the identical maximum delay in a kit. Since this event is unlikely to happen when item-delays are continuous random variables the two expectations, $S(T)$ and $D(T)$ can be assumed to be identical in most situations. For the sake of completeness, we suggest that, in case of ties, the penalty should be equally distributed to the item types involved in the tie. Thus, we have presented a componentwise measure that accurately captures the expected delay experienced by customers. The reader is also referred to Seshadri and Shanthikumar [7] and references therein about similar problems of estimating and using the maximum of several random variables when kits have to be produced subject to random yield. In the next section, we discuss how such a measure can be used by a manager to improve the performance of a multi-item assembly system.

4. SERVICE IMPROVEMENT

In this section we study the problem of allocating resources to improve the delivery performance of a multi-product inventory system when there is randomness in item delivery times. The normal but heuristic approach is to compute the average delays for each item and to allocate the available resources to each item in proportion to its average delay. In this section we show how the customer service index developed in the previous section can be used to derive the optimal allocation. To motivate this discussion, based on the delay penalties shown in Table 2, it seems that item 1 should be the candidate for improvement. We shall prove that if only a small change can be made to all the delays observed of any single item—and this can be done without affecting the observed delay of other items—then the item whose delay penalty can be reduced the most is the right candidate for improvement. This idea is formalized below.

Consider a firm that makes items to order. For each item i, let the $w_{i,k}$'s constitute a stationary and ergodic sequence. Let the stationary delay distribution be $F_i(x)$. Assume that item delivery times are independent across different types of items. Let X_i have the distribution $F_i(x)$. Assume that the expectation and the second moment of X_i for $i = 1$ to N are finite. Consider the sequence of the delay penalties $d_{i,k}$ $k = 1, 2, \ldots$ for item i. We shall show that the average of the delay penalties of item i converge to a constant, i.e., $\text{E}(d_i)$. In particular, this will imply that the delay index $\phi_i(T)$ for item i converges to $\text{E}(d_i)$.

Let Ω comprise the space of doubly infinite sequences of real numbers. The delay sequences are defined on the probability space given by the triple, (Ω, F, P). The σ-field is generated by cylinders, or sets of the form $(\omega : (w_{i,n}(\omega) \cdots w_{i,n+k-1}(\omega)) \in \mathbf{G})$ with $\mathbf{G}$ a k-dimensional Borel set (Billingsley [2, p. 18], Darrett [1]). To model the assignment of delays to customer orders, we pair each delay with a random variable that takes the value of the kit type. These random variables form the sequence $\{K_n(i)\}$. The values these random variables can take are $\{1, 2, \ldots, 2^{\Omega}\}$. To keep the analysis simple, we will give sketch of the proof when there are only two types of items that go into type 1 kits. Assume that items 1 and 2 are included in type 1 kits.

We assume that the sequences $\{(K_n(1), w_{1,n})\}$ and $\{(K_n(2), w_{2,n})\}$ are ergodic. Moreover, given that $K_n(i) = 1$, $w_{i,n}$ is independent of the sequence $\{(K_n(3-i), w_{3-i,n})\}$, $i = 1, 2$. If this assumption is not made, then the item delays of different items can be correlated even if the delay sequences are independent—for example, the firm can try to match the item delays across items that go into a kit. We show that the delays of type 1 items that go into a type 1 kit forms an ergodic sequence. So do the delays for type 2 items that go into type 1 kits. The pairs of these delays, consisting of the delay for item 1 and the other for item 2, also are ergodic. Thus, the maximum of these delays for each pair forms an ergodic sequence. The final result follows by extending this idea for different types of kits. We refer the reader to theorems that support each of these claims in standard textbooks.

Define the function $g(x, y)$ that takes the value y if $x = 1$ and the value zero otherwise. The first claim is that the sequence $\{g(K_n(j), w_{j,n})\}$ is ergodic for each $j = 1, 2$. This claim follows from Breiman [3, Proposition 6.31, p. 119]. Consider only the thinned sequence that has nonzero elements of $\{g(K_n(j), w_{j,n})\}$. The elements of this sequence are $\{w_{j,m(j,n)}\}$, where $\{\ldots m(j, -1), m(j, 0), \ldots\}$ are the nonzero elements, namely, the type j item delays for those items that go into a type 1 kit. The second claim is that this sequence is also ergodic. This claim follows from Sigman [9, Theorem 5.3, p. 111]. A direct proof also can be given by noting that the sequence given by the function $\phi(w_{j,n}, w_{j,n+1}, \ldots) = g(K_n(j))I\{\text{first } g(K_m(j)w_{j,m}) > 0 \text{ is } \leq a_1, \text{ second } g(K_m(j)w_{j,m}) > 0 \text{ is } \leq a_2, m \geq n\}$ is ergodic from Breiman [3, Proposition 6.31, p. 119]. This implies that its average over n converges almost everywhere to $\Pr\{w_{j,m(j,1)} \leq a_1, w_{j,m(j,2)} \leq a_2\}$ times the fraction of type 1 items that go into type 1 kits. Similar conclusion holds if k *nonzero* delays (instead of the first two) are used to define the function $\phi(w_{j,n}, w_{j,n+1}, \ldots)$. The claim then follows from exercise 15 in Breiman [3, p. 120] that $X_1, X_2, \ldots$ is ergodic if and only if for k-dimensional Borel set $\mathbf{G}$ and every k

$$\frac{1}{n}\sum_n I_G(X_n, X_{n+1}, \ldots, X_{n+k-1}) \rightarrow_{\text{a.s}} P((X_1, X_2, \ldots, X_k) \in \mathbf{G}).$$

The third claim is that the sequence $\{(w_{1,m(1,n)}, w_{2,m(2,n)})\}$ is ergodic. It is easy to show that the sequence is stationary due to assumption that the two random variables $w_{1,m(1,n)}$ and $w_{2,m(2,n)}$ are independent of one another for each n. Thus, by use of the ergodic theorem from Breiman [3, Theorem 6.21, p. 113],

$$\frac{1}{n}\sum_n I((w_{1,m(1,n)}, \ldots, w_{1,m(1,n-k+1)}, w_{2,m(2,n)}, \ldots\; w_{2,m(2,n+k-1)}) \in \mathbf{G}) \rightarrow$$

$$P((w_{1,m(1,n)}, \ldots, w_{1,m(1,n-k+1)}, w_{2,m(2,n)}, \ldots\; w_{2,m(2,n+k-1)}) \in \mathbf{G}|\Gamma\},$$

where Γ is the sigma field of invariant sets and $\mathbf{G}$ is a k-dimensional Borel set. However,

$$\begin{aligned} &P((w_{1,m(1,n)}, \ldots, w_{1,m(1,n-k+1)}, w_{2,m(2,n)}, \ldots\; w_{2,m(2,n+k-1)}) \in \mathbf{G}|\Gamma) \\ &\quad = P((w_{1,m(1,n)}, \ldots, w_{1,m(1,n-k+1)}) \in \mathbf{G}|\Gamma, (w_{2,m(2,n)}, \ldots\; w_{2,m(2,n+k-1)}) \in \mathbf{G}) \\ &\qquad P((w_{2,m(2,n)}, \ldots\; w_{2,m(2,n+k-1)}) \in \mathbf{G}|\Gamma) \\ &\quad = P((w_{1,m(1,n)}, \ldots, w_{1,m(1,n-k+1)}) \in \mathbf{G}|\Gamma)P((w_{2,m(2,n)}, \ldots\; w_{2,m(2,n+k-1)}) \in \mathbf{G}|\Gamma) \\ &\quad = P((w_{1,m(1,n)}, \ldots, w_{1,m(1,n-k+1)}) \in \mathbf{G})P((w_{2,m(2,n)}, \ldots\; w_{2,m(2,n+k-1)}) \in \mathbf{G}) \\ &\quad = P((w_{1,m(1,n)}, \ldots, w_{1,m(1,n-k+1)}, w_{2,m(2,n)}, \ldots\; w_{2,m(2,n+k-1)}) \in \mathbf{G}), \end{aligned}$$

where the first equality is obtained by repeated conditioning, the second is due to the independence of the two thinned delay sequences, the third is via the ergodicity of each sequence, and the fourth uses independence. Application of exercise 15 from Breiman [3, p. 120] proves the third claim. The fourth and final claim that $\{I(w_{1,m(1,n)} > w_{2,m(2,n)})\max(w_{1,m(1,n)}, w_{2,m(2,n)})\}$ is ergodic follows from claim 3 and Proposition 6.31 from Breiman [3, p. 119].

Based on these four claims, it follows that the average of the delay penalties for each item i (which is the sum of the average of the delay penalties for each kit that requires item i) converges to a constant that we represent as $\mathrm{E}(d_i)$. Moreover, the constant $\mathrm{E}(d_i)$ is given by the intuitive weighted average of the expected delay penalty contributions due each type of kit where the weights are the relative frequency with which item i is required by that type of kit.

Due to this result, not only is it meaningful to refer to the expected value of the delay indices, but also when the number of kits is large, we can focus our attention on reducing the kit-delay based on the delay indices. Assume that the firm has a fixed budget, denoted as B, for improving the performance of the system. The budget could represent money that can be invested to improve supplier coordination, or to improve internal information technology infrastructure or simply to find a more reliable and agile supplier. Typically suppliers of each of the items may be different firms or facilities (certainly true for PC assembly where motherboard, hard drive, memory, and peripherals are sourced from different suppliers). As a result, we assume that the allocation is separable, i.e., b_i dollars are spent on improving the delivery performance of item i. Further, we assume that the objective of the firm is to minimize the average delay experienced by customers. We assume that the investment of b_i in item i leads to an almost sure reduction of the random variables $w_{i,k}$'s by a scale factor of $\delta_i(b_i)$, i.e., for each k, the new random variable is given by $w_{i,k} * \delta_i(b_i)$. Further, let $\delta_i(0) = 1$ and δ_i be a strictly decreasing convex function of b_i. The decreasing assumption follows because more investment is expected to lead to smaller delays. The function $\delta_i(\,)$ is assumed to be convex because it is reasonable to expect that it becomes more and more difficult to improve performance as one gets closer and closer to a perfect supply situation.

We assume that the fraction of orders for kits of type j converges as T increases to a limit. Let the fraction of orders that include item i be $\lambda_i = \lim_{T\to\infty} \frac{M_i(T)}{\sum_j O_j(T)}$. Given the above assumptions, the optimization problem can be formulated as follows:

$$\min_{b_i} \frac{D(T)}{\sum_j O_j(T)} \qquad \text{s.t.} \qquad \sum_{i=1}^{N} b_i \le B,$$

which is equivalent to

$$\min_{b_i} \frac{\sum_{i=1}^{N} \sum_{k=1}^{M_i(T)} d_{i,k}}{\sum_j O_j(T)} \qquad \text{s.t.} \qquad \sum_{i=1}^{N} b_i \le B.$$

For large T, the objective function is equivalent to

$$\min_{b_i} \sum_{i=1}^{N} \lambda_i E(d_i) \qquad \text{s.t.} \qquad \sum_{i=1}^{N} b_i \le B,$$

where we have used the fact that when T is large $\lim_{T\to\infty} \frac{\sum_{k=1}^{M_i(T)} d_{i,k}}{\sum_j O_j(T)} = \lim_{T\to\infty} \frac{M_i(T)}{\sum_j O_j(T)} \cdot \frac{\sum_{k=1}^{M_i(T)} d_{i,k}}{M_i(T)} = \lim_{T\to\infty} \frac{M_i(T)}{\sum_j O_j(T)}\, \phi_\mathrm{i}(\mathrm{T}) = \lambda_\mathrm{i}\mathrm{E}(\mathrm{d_i})$. Note that the estimate for $E(d_i)$ at time T is the componentwise delay index $E(d_i) \approx \phi_i(T) = \frac{\sum_{k}^{M_i(T)} d_{i,k}}{M_i(T)}$. Then the optimization problem is conveniently approximated as

$$\min_{b_i} \sum_{i=1}^{N} \lambda_i \phi_i(T) \qquad \text{s.t.} \qquad \sum_{i=1}^{N} b_i \le B. \tag{3}$$

Let $I(i, k)$ denote the type of kit which contains the kth type i item (zero otherwise).

LEMMA 1: $\mathbf{E} \sum_{i=1}^{N} \sum_{k=1}^{M_i(T)} d_{i,k}$ is jointly convex and strictly decreasing in b_i.

PROOF: Note that the delay experienced by customers can be written as follows: $\sum_{i=1}^{N} \sum_{k=1}^{M_i(T)} d_{i,k} = \sum_{i=1}^{N} \sum_{k=1}^{M_i(T)} \{\max_{l\in S_{I(i,k)}}\{\delta_l(b_l) w_{l,Q_{I(i,k)}(l,R_{I(i,k)}(i,k))}\}\}/|S_{I(i,k)}|$. Since $\max_l\{\delta_l(b_l) w_{l,Q_{I(i,k)}}(l, R_{I(i,k)}(i, k))\}$ is jointly convex in b_l and because convexity is preserved under expectation, $\mathbf{E} \sum_{i=1}^{N} \sum_{k=1}^{M_i(T)} d_{i,k}$ is convex in b_i. Further, since $\delta_l(b_l)$ is a strictly decreasing function in b_l, $\mathbf{E} \sum_{i=1}^{N} \sum_{k=1}^{M_i(T)} d_{i,k}$ is decreasing in b_i. □

THEOREM 2: The optimal allocation of budget to the different items is a convex optimization problem.

PROOF: From Lemma 1, the objective function given in (3) is convex and strictly decreasing in b_i. □

This result can be generalized as follows. Let $\{X(\theta),\ \theta \in \Theta\}$ be a family of random variables. $\{X(\theta),\ \theta \in \Theta\}$ is said to be stochastically decreasing and convex almost every where if there exist $\{\hat{X}(\theta),\ \theta \in \Theta\}$ such that $X(\theta) =_{st} \hat{X}(\theta)$ for each $\theta \in \Theta$ and $\hat{X}(\theta)$ is decreasing and convex in θ [8]. Thus, the theorem will continue to hold if the stationary distribution of the delay of item i is given parametrically by a family of random variables $\{W_i(\theta)\}$, where θ corresponds to the amount invested to reduce the delay of item i, and this family is stochastically decreasing and convex almost every where, see for example Theorem 6.D.7 in [8].

4.1. Special Cases

In this section, we consider the cases when the item delays are either deterministic or have the exponential distribution.

Deterministic Case: Let us assume that there is only one type of kit, and the delays are given by $w_i(b_i)$ where the delay is a convex decreasing function of the budget allocated. Then the problem is

$$\min_{b_i} \lambda_i \phi_i(T) \qquad \text{s.t.} \qquad \sum_{i=1}^{N} b_i \leq B.$$

Due to the fact that delay is deterministic, the item that has the maximum delay, say item i (where i = argmax w_i), has a customer service index value equal to w_i and the rest of the indices are zero. Based on our analysis expressed in Theorem 2, an ϵ investment should be made in that item so that its delay becomes at most equal to the delay of the item that has the next smallest delay. This procedure is repeated until the budget is exhausted. Each of the above delay functions is convex and decreasing in the allocations made. This property is preserved under the max operator. Therefore, the objective function is convex in the allocations. The minimization problem is separable; therefore, a sequential greedy algorithm will find the optimal solution.

Exponential Case: Consider a two-item case in which customers always order the two items together as a kit. Let the item delays be independent and exponentially distributed. Let the stationary item delays be denoted as w_1 and w_2 and the mean delay for the two types of items be $\frac{1}{\lambda_1}$ and $\frac{1}{\lambda_2}$. Further, assume that budget allocation to item i leads to a decrease in the parameter of the distribution, namely, λ_i, but does not otherwise change the distribution of the delay. The expected delay experienced by a customer is given by

$$\mathbf{E} \max\{w_1, w_2\} = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2}.$$

Table 3. Derivatives with respect to the componentwise indices.

$\dfrac{\partial \phi_1}{\partial \lambda_1} \approx -\dfrac{1}{\lambda_1^2} + \dfrac{1}{(\lambda_1 + \lambda_2)^2} - \dfrac{2\lambda_2}{(\lambda_1 + \lambda_2)^3}$	$\dfrac{\partial \phi_2}{\partial \lambda_1} \approx \dfrac{2\lambda_2}{(\lambda_1 + \lambda_2)^3}$
$\dfrac{\partial \phi_1}{\partial \lambda_2} \approx \dfrac{1}{(\lambda_1 + \lambda_2)^2} + \dfrac{\lambda_1 - \lambda_2}{(\lambda_1 + \lambda_2)^3}$	$\dfrac{\partial \phi_2}{\partial \lambda_2} \approx \dfrac{-2\lambda_1}{(\lambda_1 + \lambda_2)^3} - \dfrac{\lambda_1^2 + 2\lambda_1\lambda_2}{(\lambda_2\lambda_1 + \lambda_2^2)^2}$

The expected value of the delay penalties for the two items are given by

$$\mathbf{E}(d_1) = \mathbf{E}\{w_1; w_1 > w_2\} \qquad \text{and} \qquad \mathbf{E}(d_2) = \mathbf{E}\{w_2; w_2 > w_1\}.$$

Since $\mathbf{E}(d_i) \approx \phi_i(T)$ for large T, upon simplification, we get

$$\phi_1(T) \approx \frac{\lambda_2}{\lambda_1 + \lambda_2}\left(\frac{1}{\lambda_1 + \lambda_2} + \frac{1}{\lambda_1}\right) \qquad \text{and} \qquad \phi_2(T) \approx \frac{\lambda_1}{\lambda_1 + \lambda_2}\left(\frac{1}{\lambda_1 + \lambda_2} + \frac{1}{\lambda_2}\right).$$

Observe that the sum of $\phi_i(T)$'s is equal to the average delay in the system. Next, the derivative of the average delay with respect to λ_1 is given by

$$\frac{\partial \mathbf{E}D(T)}{\partial \lambda_1} = -\frac{1}{\lambda_1^2} + \frac{1}{(\lambda_1 + \lambda_2)^2}.$$

Note that

$$\frac{\partial \phi_1(T)}{\partial \lambda_1} \approx -\frac{1}{\lambda_1^2} + \frac{1}{(\lambda_1 + \lambda_2)^2} - \frac{2\lambda_2}{(\lambda_1 + \lambda_2)^3} \qquad \text{and} \qquad \frac{\partial \phi_2(T)}{\partial \lambda_1} \approx \frac{2\lambda_2}{(\lambda_1 + \lambda_2)^3}.$$

Observe that the *sum* of these partial derivatives is equal to the partial derivative of $\mathbf{E}\max(w_1, w_2)$ with respect to λ_1. But the derivative of the index of the first item is not equal to the derivative of the expected delay. The reason this happens in this example is that a reduction in the delay of the first item not only decreases the index of the first item but also increases the index for the second item. We can cater to such a situation by maintaining a record of the decrease as well as the increase in the customer service indices for all items due to the reduction in the delay of each item. For example, in the above exponential case the derivatives are as shown in Table 3. This table can be used as the format for bookkeeping purposes. For example, each time a kit is delivered, we determine which item has the maximum delay (say item 1). Then, ask whether a reduction of, say 1%, in the delay of that item will lead to a similar reduction in the delay experienced by the customer. The actual percentage of the delay, say $x\%$, that is reduced is recorded as a negative number under the column for item 1, whereas the remaining percentage, $(1 - x)\%$, as a positive number under the column for item 2. The sum across a row of the table gives the marginal benefit of reducing the delay of that item. This can be used to determine the optimal allocations.

5. CONCLUSIONS

In this paper, we consider the issue of improving customer waiting times in a multi-item inventory/assembly system where the delivery performance needs to be improved through appropriate financial investment. In the process, we derive a new componentwise delay index whose sum provides waiting time experienced by customers. We present an optimal greedy allocation algorithm for allocating resources to minimize the expected customer waiting times subject to a capital budget. We also present conditions under which the optimal allocation is such that a component yielding the maximum marginal reduction in the componentwise indices should be chosen as the candidate for investment at each step.

ACKNOWLEDGMENTS

The authors wish to thank the associate editor and two referees for their valuable comments on an earlier version of the paper which helped us improve the presentation of the paper to a great extent. The research of the second author was funded in part by NSF CAREER Award # 9984252.

REFERENCES

[1] R. Durrett, Probability: Theory and examples, 2nd ed., Wadsworth, Belmont, CA, 1996.
[2] P. Billingsley, Ergodic theory, Wiley, New York, 1965.
[3] L. Brieman, Probability (republication), SIAM, Philadelphia, 1992.
[4] B. Fox, Discrete optimization via marginal analysis, Manage Sci 13(3) (1966), 210–216.
[5] P. Glasserman and Y. Wang, Leadtime-inventory trade-offs in assemble-to-order systems, Oper Res 46(6) (Nov–Dec 1998), 858–871.
[6] W.H. Hausman, H.L. Lee, and A. Zhang, Joint demand fulfillment probability in a multi-item inventory system with independent order-up-to policies, Eur J Oper Res 109(3) (Sep 1998), 646–659.
[7] S. Seshadri and J.G. Shanthikumar, Allocation of chips to wafers in a production problem of semiconductor kits, Oper Res 45(2) (Mar–Apr 1997), 315–321.
[8] M. Shaked and J.G. Shanthikumar, Stochastic orders and their applications, Academic Press, San Diego, CA, 1994.
[9] K. Sigman, Stationary marked point processes: An intuitive approach, CRC Press, Boca Raton, FL, 1995.
[10] J.S. Song, On the order fill rate in a multi-item, base-stock inventory system, Oper Res 46(6) (Nov/Dec 1998), 831–845.
[11] J.S. Song, S.H. Xu, and B. Liu, Order-fulfillment performance measures in an assemble-to-order system stochastic leadtimes, Oper Res 47(1) (Jan–Feb 1999), 131–149.
[12] R.Q. Zhang, Expected time delay in a multi-item production-inventory system with correlated demands, Nav Res Logistics 46(6) (1999), 671–688.