A Simple and Effective Component Procurement Policy for Stochastic Assembly Systems

JÉRÉMIE GALLIEN   jgallien@mit.edu
Operations Research Center, MIT, 77 Mass Ave, Bldg E40-149, Cambridge, MA 02139, USA

LAWRENCE M. WEIN   lwein@mit.edu
Sloan School of Management, MIT, 50 Memorial Drive, Bldg E53-343, Cambridge, MA 02139, USA

Abstract. We examine the component procurement problem in a single-item, make-to-stock assembly system. The suppliers are uncapacitated and have independent but non-identically distributed stochastic delivery lead times. Assembly is instantaneous, product demand follows a Poisson process and unsatisfied demand is backordered. The objective is to minimize the sum of steady-state holding and backorder costs over a pre-specified class of replenishment policies. To keep the analysis tractable, we impose a synchronization assumption that no mixing occurs between sets of component orders. Combining existing results from queueing theory with original results concerning distributions that are closed under maximization and translation, we derive a simple approximate solution to the problem when lead time variances are identical. In simulations, our derived policy is within 2% of optimal and significantly outperforms policies that ignore either component dependence or lead time stochasticity. It is also quite robust with respect to various model assumptions, except the synchronization one.

Keywords: assembly systems, inventory policies, maximum of random variables, Gumbel distribution, closure under maximization and translation

1. Introduction

In some industries (e.g., consumer electronics), over half the total manufacturing cost of products is attributed to the cost of procuring components. Moreover, because of the increased use of foreign suppliers, most of the manufacturing lead time is typically due to the procurement lead time. In these settings, the component procurement policy can be an important source of competitive advantage. Unfortunately, the assembly process induces dependencies across components that make the component procurement problem very difficult to analyze, particularly in the presence of both demand and procurement variability.

We consider a make-to-stock environment where a manufacturer of a single item procures a variety of components from different suppliers, and instantaneously assembles these components into finished units, which are then placed into a finished goods inventory that services a Poisson demand process. Suppliers are uncapacitated, and each supplier has an associated procurement lead time distribution. We do not attempt to

find the optimal policy for this system (the structure of which is not even known), but rather restrict our attention to the pre-specified class of product base stock policies with component postponement lead times. Under this class of policies, a customer order simultaneously triggers an order for each component after a component-dependent postponement lead time. This particular policy structure allows us to develop an analytically tractable approximation based on a synchronization assumption. We say that the system is *synchronized* if there is no mixing of orders; that is, if components replenishing the same customer order also end up being assembled into the same unit. This assumption is discussed in detail in section 2.3, and in our computational study we investigate how our policy performs in an asynchronized system, where mixing of orders is allowed.

Given the prevalence of stochastic assembly systems in practice, it is not surprising that much has been written on this problem in the operations management literature. In reviewing the literature, we restrict ourselves to systems that are pure make-to-stock or hybrid make-to-stock/assemble-to-order, and focus on one important dimension of the models: whether the suppliers are capacitated (i.e., modeled as single-server queues) or uncapacitated (modeled as infinite-server queues). Although capacitated models are more difficult to analyze and are in some sense more realistic, these two classes of models are complementary in our view: the former case is appropriate when the component orders generated by the assembly system comprise the bulk of the supplier's business (e.g., in a vertically integrated firm, or a devoted supplier to a large manufacturer), and the latter case is appropriate when these orders represent only a small portion of the supplier's workload. In the latter case, the timing of component orders from the assembly facility has a minor impact on the congestion at the supplier's manufacturing facility, and the procurement lead times for the components are reasonably modeled as iid random variables from the assembler's viewpoint. Another case where the suppliers are appropriately modeled by infinite-server queues is when transportation delays account for most of the replenishment lead times; this is common when suppliers are located overseas.

Most of the work in stochastic assembly systems with capacitated suppliers is algorithmic, and is aimed at either the performance analysis of a given policy (e.g., Song et al. [25], Zhang [32], Wilhelm and Som [28], Schraner [22]) or the optimization of a procurement or production policy (Anupindi and Tayur [4], Kushner and Tetzlaff [16]). Glasserman and Wang [10], and to a lesser extent Nemec [18], are able to use asymptotic methods to obtain explicit expressions for performance measures. In addition, Glasserman and Wang [11] use their earlier results to derive simple and effective base stock policies for multi-item systems.

Most of the analysis of assembly systems with uncapacitated suppliers assumes deterministic component lead times (Srinivasan et al. [26], Hausman et al. [14], Tayur [27], Zhang [30,31], Song [23,24], Abhyankar [1] and references therein). This assumption allows some structural results to be derived (Schmidt and Nahmias [21], Rosling [19]) and simplifies the analysis considerably. Nonetheless, with the exception of Zhang [31], these studies provide computational procedures – rather than explicit formulas – for procurement policies. Analyses of the stochastic procurement lead time case include Hopp

and Spearman [15], who consider each assembled unit independently and hence do not track the dynamics of the inventory process, and Cheung and Hausman [6], who derive an exact but computationally intensive expression for the distribution of backorders in a multi-item system with complete cannibalization, which corresponds to an asynchronized system in our terminology.

In summary, the only simple and effective control policy for stochastic assembly systems (to our knowledge) is due to Glasserman and Wang [11], who consider a multi-item system with capacitated suppliers. Our main goal is to develop an analogous result for single-item systems with uncapacitated suppliers; this goal is achieved in equations (19).

In section 2 we describe the model and the class of policies under consideration. Our simple suboptimal solution to the component procurement problem is derived in section 3 using known results from queueing theory and some new results on probability distributions that are closed under maximization and translation. The effectiveness and robustness of our policy are addressed in section 4, where a simulation study is undertaken using industrial data from a Hewlett-Packard facility. Concluding remarks are provided in section 5.

2. The model

2.1. Assumptions

We consider a continuous review inventory system where n components are assembled into a single item. Demand for the end item follows a Poisson process with rate λ. Demand is met whenever possible from on-hand finished goods inventory, while unsatisfied demand is fully backordered.

The replenishment process of each component is uncapacitated, so that each supplier can be viewed as an infinite-server queue, or a "delay box". Each component i has a random replenishment lead time denoted by X_i, $i = 1, \ldots, n$. We assume that $(X_1, \ldots, X_n)$ are mutually independent random variables, but not necessarily identically distributed. Although the distribution of $(X_1, \ldots, X_n)$ is unspecified at this point, assumptions that the lead times are deterministic, follow Gumbel distributions, and follow a generalization of Gumbel distributions are made in sections 3.3, 3.5 and 3.6, respectively.

Because our focus is on the procurement process, we assume that assembly is instantaneous; a detailed specification of the assembly rule is deferred until section 2.3. In this context, complete sets of components are equivalent to finished goods. As a result, this model can also be viewed as an assemble-to-order system where assembly is exclusively triggered by customer demand.

Let Z represent the steady-state net inventory of finished goods, so that $Z^+ = \max(Z, 0)$ is the steady-state inventory of finished goods while $Z^- = \max(-Z, 0)$ is the steady-state order backlog. Let Z_i denote the steady-state inventory of component i that is available for assembly.

Finally, we assume a linear cost structure, where the finished goods inventory hold-ing cost rate is h, the component i inventory holding cost rate is h_i, and the backorder cost rate is b. We assume $h = \sum_i h_i$, so that assembling a complete set of components into a product does not add value. The objective for the optimization problem studied in section 3 is to minimize the long run expected average cost, which is given by

$$C = hE[Z^+] + \sum_{i=1}^{n} h_i E[Z_i] + bE[Z^-].$$
(1)

2.2. Policies

Finding the optimal procurement policy for the model described in section 2.1 is an open problem beyond the scope of this paper. Our approach is to restrict attention to a class of policies with a pre-specified structure, and find the optimal policy parameters within that class. This method is widely adopted in the literature, although usually a component base stock policy is investigated, whereas we study a finished goods base stock policy with component postponement lead times. More specifically, we assume that the finished goods inventory is initially filled to its base stock level S, and each customer order triggers a replenishment order for all components after a component-dependent postponement lead time $\ell_i \geqslant 0$, which is here a discretionary, deliberate delay introduced to reduce holding costs. Figure 1 offers a schematic representation of both the model and the policy parameters ℓ_i.
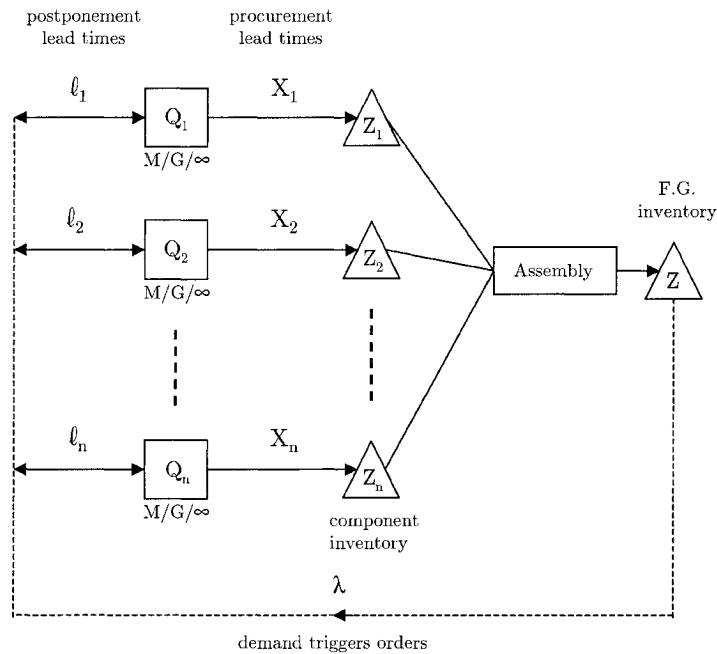

Figure 1. The assembly system.

In the rest of the paper, we refer to the policy just described as $[S, \ell_1, \ldots, \ell_n]$ or $[S, \vec{\ell}\,]$. This class of policies is closely related to the more traditional class of component base stock policies $[s_1, s_2, \ldots, s_n]$ (or $[\vec{s}\,]$), where each component inventory is initially filled to a component-dependent base stock level s_i, and each customer order triggers a replenishment order for all the components. More specifically, if we impose the equivalence relation $s_i = S - \lambda \ell_i$ to hold for each i, then numerical experiments (see [9]) show that the $[S, \vec{\ell}\,]$ policy and its corresponding $[\vec{s}\,]$ policy achieve nearly identical performance. In general, $[S, \vec{\ell}\,]$ is slightly more refined than $[\vec{s}\,]$ because the base stock levels S and $\vec{s}$ are restricted to be integer-valued, whereas the lead times $\vec{\ell}$ are allowed to be continuous. Note that both policies have essentially n parameters, because (see section 3) we can take the optimal value of the smallest ℓ_i to equal zero.

2.3. The synchronization assumption

Under $[S, \vec{\ell}\,]$, every customer order triggers a complete set of component replenishment orders, but the transmissions of these component orders to the suppliers are delayed from that point by the postponement lead times. We assume that the assembly is *synchronized*: each assembly is only performed with components belonging to the same set of replenishment orders. Equivalently, no mixing occurs at the assembly stage between sets of component orders.

If procurement lead times are stochastic, the synchronized system will typically not perform as well as an asynchronized system that employs a first-come first-served (FCFS) assembly rule; i.e., a product is assembled whenever there exists at least one unit of inventory for each component. Note that the synchronized assumption automatically holds under $[S, \vec{\ell}\,]$ for systems with deterministic lead times and for single-item systems with capacitated suppliers, where the suppliers are modeled as single-server queues that employ a FCFS discipline; in both cases, overtaking of component orders is not possible. In fact, the synchronization assumption is closely related to the assumption of uncapacitated vs. capacitated suppliers: while one could argue that suppliers typically satisfy component orders in a first-in-first-out fashion in practice, the reason for this is that suppliers are capacitated, even if these components make up a small part of their business. Although we are unable to provide a compelling justification for the synchronization assumption, we are aware of at least one Japanese company that employs a synchronized system (Sridhar Tayur, personal communication). It may be that, depending upon the details of the information processing system and logistical infrastructure, synchronized systems are sometimes easier to manage than asynchronized systems. As noted earlier, our computational study in section 4 compares the performance of a synchronized system with that of an asynchronized system. Finally, in systems where customized components cannot be ordered until after customer requests are placed, our model applies by setting $S = 0$ and optimizing the postponement lead times; our synchronization assumption is appropriate for such systems.

3. Analysis

3.1. Formulation of the optimization problem

For a synchronized assembly system using $[S, \vec{\ell}]$, the time needed to replenish a complete set of components is $\max_i (X_i + \ell_i)$. As a result, the assembly system can be interpreted as an $M/G/\infty$ queue with arrival rate λ and service times $\max_i (X_i + \ell_i)$. The departures from this queue enter a finished goods net inventory with steady-state level Z and initial value S, which is depleted by the Poisson demand process. Note that the items populating this queueing system (see figure 2) represent complete sets of components, either assembled or unassembled.

This queueing interpretation allows us to express the objective function (1) in terms of the decision variables $(S, \vec{\ell})$. Let Q be the steady-state number of replenishment orders for complete sets of components for which at least one of the n individual component orders has not yet been satisfied. Then Q is exactly the steady-state queue length of the $M/G/\infty$ queue with service times $\max_i (X_i + \ell_i)$ introduced earlier. It is well known that

$$Q \sim \text{Poisson}(\rho), \quad \text{where } \rho = \lambda E\left[\max_i (X_i + \ell_i)\right]. \tag{2}$$

Since the total number of complete sets of components remains constant over time under the $(S, \vec{\ell})$ policy, we have $Z + Q = S$. Therefore, the mean finished goods inventory in steady state is given by

$$E\left[Z^+\right] = e^{-\rho} \sum_{j=0}^{S} (S - j) \frac{\rho^j}{j!}. \tag{3}$$

Taking expected values in the identities $Z = Z^+ - Z^-$ and $Z + Q = S$ yields the mean steady-state backorder level

$$E\left[Z^-\right] = E\left[Z^+\right] - S + \rho. \tag{4}$$

The only remaining term in (1) to study is the mean inventory of unassembled component i, $E[Z_i]$. Because we assume that assembly occurs as soon as possible, another
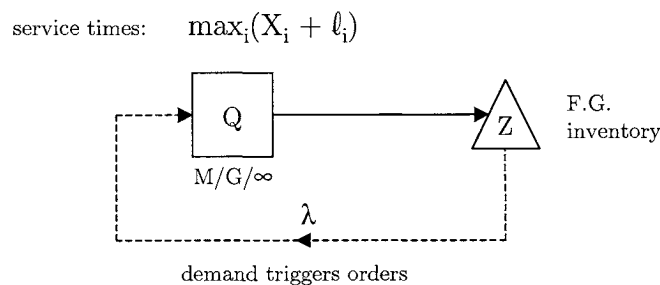

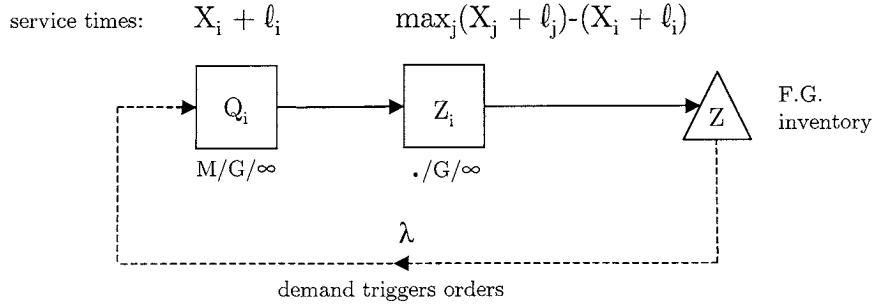Figure 2. Flow of complete sets of components.

service times: $X_i + \ell_i$ $\quad\quad$ $\max_j(X_j + \ell_j)\text{-}(X_i + \ell_i)$

Figure 3. Flow of components.

consequence of synchronization is that the time each component of type i spends in the component inventory before being assembled is $\max_j(X_j + \ell_j) - (X_i + \ell_i)$. Hence, the circulation of the components of type i can be interpreted as the tandem queueing network depicted in figure 3, where each component arrives according to a Poisson process with rate λ, is first serviced by a $M/G/\infty$ queue with service times $X_i + \ell_i$ and steady-state queue length Q_i, and then by a second infinite-server queue with service times $\max_j(X_j + \ell_j) - (X_i + \ell_i)$ and steady-state queue length Z_i. The output of the second queue is placed into the finished goods net inventory, which services the customer demand. Little's formula, which holds even though the service times in this network are not independent across stations, gives when applied to the second queue in figure 3

$$E[Z_i] = \lambda\left(E\left[\max_j(X_j + \ell_j)\right] - E[X_i] - \ell_i\right). \tag{5}$$

Using (1)–(5), rearranging terms and omitting those independent of the decision variables, we can formulate our optimization problem as

$$\min_{S,\ell_1,\dots,\ell_n} C\left(S, \vec{\ell}\right) = (h + b)\left(\rho + e^{-\rho}\sum_{j=0}^{S}(S - j)\frac{\rho^j}{j!}\right) - \lambda\sum_{i=1}^{n} h_i\ell_i - bS$$

subject to: $\rho = \lambda E\left[\max_i(X_i + \ell_i)\right],$ $\tag{6}$

$\quad\quad\quad\quad \ell_i \geqslant 0 \quad \forall i,$

$\quad\quad\quad\quad S$ integer.

3.2. Solving for S in terms of $(l_1, \dots, l_n)$

Solving this constrained mixed nonlinear program analytically is difficult because expressing ρ in terms of $(\ell_1, \dots, \ell_n)$ in closed form is very cumbersome for general lead time distributions. However, we can reduce the complexity of (6) somewhat by observing that the distribution of Q in equation (2) is independent of the base stock level S.

Proposition 1. Let $(\hat{\ell}_1, \ldots, \hat{\ell}_n)$ be a given set of values for the decision variables in (6), and let $S^*_{\hat{\ell}_1, \ldots, \hat{\ell}_n} = \arg\min_S C(S, \hat{\ell}_1, \ldots, \hat{\ell}_n)$ be the optimal value of S for the objective function C given $(\hat{\ell}_1, \ldots, \hat{\ell}_n)$. Then $S^*_{\hat{\ell}_1, \ldots, \hat{\ell}_n}$ is the smallest integer that satisfies

$$P\left(Q \leqslant S^*_{\hat{\ell}_1, \ldots, \hat{\ell}_n}\right) \geqslant \frac{b}{b+h}.$$

Proof. Given the values of $(\ell_1, \ldots, \ell_n)$, the queueing system described in figure 2 is just a single-product version, where the WIP distribution is fully specified by equation (2), of the CONWIP model studied in [20]. Applying proposition 1 of that paper gives the desired result. $\square$

In section 3.3 we use proposition 1 to find the optimal solution to (6) in the deterministic lead time case, by first optimizing over $(\ell_1, \ldots, \ell_n)$ and then optimizing over S for given values of $(\ell_1, \ldots, \ell_n)$. Section 3.4 presents a more general approximate decomposition technique that exploits this idea, which we subsequently apply in sections sections 3.5 and 3.6 to analyze our optimization problem in cases where the lead times are stochastic.

3.3. Deterministic lead times

When the component lead times $(X_1, \ldots, X_n)$ are deterministic, the exact delivery dates of all the components are known as soon as the replenishment orders are sent. In this case, it makes no economic sense to have components from the same set of replenishment orders delivered at different dates: any component delivered in advance would incur unnecessary holding costs because assembly can only occur when the last component is delivered. Therefore, the optimal postponement lead times must satisfy $X_1 + \ell_1^* = \cdots = X_n + \ell_n^*$. The decision variables $(\ell_1, \ldots, \ell_n)$ can thus be reduced to the single variable $\rho = \lambda \max_i(X_i + \ell_i)$, using the transformation

$$\ell_i = \frac{\rho}{\lambda} - X_i \quad \forall i. \tag{7}$$

By (7), the n nonnegativity constraints $\ell_i \geqslant 0 \ \forall i$ now correspond to the single constraint $\rho \geqslant \lambda \max_i(X_i)$. Hence, for the deterministic lead times case, problem (6) can be expressed as

$$\min_{S, \rho} \ C_{\det}(S, \rho) = b(\rho - S) + (h + b)\left(e^{-\rho}\sum_{j=0}^{S}(S - j)\frac{\rho^j}{j!}\right)$$

subject to $\quad \rho \geqslant \lambda \max_i(X_i),$

$\qquad\qquad S$ integer. $\tag{8}$

This optimization problem can be solved analytically, as is shown by the following proposition, which (as with all remaining propositions) is proved in the appendix.

Proposition 2. The optimal value of ρ in problem (8) is $\rho^* = \lambda \max_i(X_i)$.

By (7) and proposition 2, the optimal solution to (8) is

$$
\begin{cases}
\ell_i^* = \max_j(X_j) - X_i \quad \forall i, & \text{(9a)} \\[2mm]
\rho^* = \lambda \max_i(X_i), & \text{(9b)} \\[2mm]
S^* \text{ is the smallest integer that satisfies } e^{-\rho^*} \sum_{j=0}^{S^*} \frac{(\rho^*)^j}{j!} \geqslant \frac{b}{b+h}. & \text{(9c)}
\end{cases}
$$

In words, the solution to the deterministic lead times system sets the postponement lead times so that all components are delivered simultaneously and the component with the longest procurement delay has no postponement lead time. Note that the optimal policy $[S^*, \ell^*]$ depends on the component holding costs $(h_1, \ldots, h_n)$ only through their sum h; the need to consider these costs individually arises in our model only because of procurement stochasticity.

The structure of this policy is strikingly similar to that derived by Zhang [30], who specialized to the single-product case the optimal policy structure found by Rosling [19] for an assembly model with a linear cost structure and deterministic procurement lead times in a discrete-time setting. In particular, Zhang [30] showed that the structure of the optimal policy is entirely determined by the longest procurement delay and its differences with the other procurement delays, as in (9a) and (9b). Moreover, in Zhang's policy the optimal base stock level of the component with the longest procurement delay can be determined by solving a newsvendor problem, of which (9c) is the exact analogue. While it may be interesting to compare (9) with Zhang's policy in more detail, we have chosen not to investigate this issue because the deterministic lead times case only constitutes a building block in our stochastic analysis.

3.4. Approximate component/product decomposition

The analysis in section 3.3 breaks down when the lead times are stochastic. In this subsection we develop an approximate decomposition technique that exploits the relationship between S and $(\ell_1, \ldots, \ell_n)$ revealed in proposition 1.

The partial derivative with respect to ℓ_i of the objective function in (6) is

$$
\frac{\partial C(S, \ell_1, \ldots, \ell_n)}{\partial \ell_i} = (h+b)\left(1 - e^{-\rho} \sum_{j=0}^{S-1} \frac{\rho^j}{j!}\right)\frac{\partial \rho}{\partial l_i} - \lambda h_i. \tag{10}
$$

This expression can be interpreted in terms of the distribution of Q as

$$
\frac{\partial C(S, \ell_1, \ldots, \ell_n)}{\partial \ell_i} = (h+b)\big(1 - P(Q \leqslant S - 1)\big)\frac{\partial \rho}{\partial \ell_i} - \lambda h_i. \tag{11}
$$

Let now $S^*_{\ell_1,\dots,\ell_n} = \arg\min_S C(S, \ell_1, \dots, \ell_n)$. Proposition 1 implies that

$$P\left(Q \leqslant S^*_{\ell_1,\dots,\ell_n} - 1\right) < \frac{b}{h + b} \leqslant P\left(Q \leqslant S^*_{\ell_1,\dots,\ell_n}\right),$$

which suggests the approximation

$$P\left(Q \leqslant S^*_{\ell_1,\dots,\ell_n} - 1\right) \simeq \frac{b}{h + b}. \tag{12}$$

Maintaining the optimality of S given $(\ell_1, \dots, \ell_n)$ and substituting (12) into (11) yields

$$\frac{\partial C(S^*_{\ell_1,\dots,\ell_n}, \ell_1, \dots, \ell_n)}{\partial \ell_i} \simeq h\frac{\partial \rho}{\partial \ell_i} - \lambda h_i. \tag{13}$$

Because the right side of (13) does not depend on S, we decompose the analysis of (6) in the following way.

Approximate component/product decomposition.

1. Solve for $(\ell_1, \dots, \ell_n)$ in

$$\underset{\ell_1,\dots,\ell_n}{\text{Min}} \quad C_{\text{comp}}(\ell_1, \dots, \ell_n) = E\left[\max_i(X_i + \ell_i)\right] - \sum_{i=1}^{n} \bar{h}_i \ell_i \tag{14}$$

$$\text{subject to} \quad \ell_i \geqslant 0 \quad \forall i.$$

2. Let $(\ell_1^*, \dots, \ell_n^*)$ be the solution obtained in step 1, and let $\rho^* = \lambda E[\max_i(X_i + \ell_i^*)]$. Set S^* to be the smallest integer that satisfies

$$e^{-\rho^*} \sum_{j=0}^{S^*} \frac{(\rho^*)^j}{j!} \geqslant \frac{b}{b + h}. \tag{15}$$

   The idea behind the first step above is to minimize a function with partial derivatives given by the right side of (13): the objective function C_{comp} in subproblem (14) has been constructed by integrating (13) using equation (2), dividing by λh and introducing the notation $\bar{h}_i = h_i/h$. The second step of the decomposition is a straightforward application of proposition 1.

   In the rest of the paper, we refer to (14) as the *component subproblem*, since the variables $(\ell_1, \dots, \ell_n)$ in the $[S, \vec{\ell}\,]$ policy are the levers used to differentiate components according to their individual lead times $(X_1 \dots, X_n)$ and relative holding costs $(\bar{h}_1, \dots, \bar{h}_n)$. The values of $(\ell_1, \dots, \ell_n)$ are only specified by (14) up to a common additive constant. To see this, note that $\max_i(X_i + \ell_i + x) = \max_i(X_i + \ell_i) + x \; \forall x \in \mathcal{R}$ and $\sum_i \bar{h}_i = 1$, and therefore $C_{\text{comp}}(\ell_1 + x, \dots, \ell_n + x) = C_{\text{comp}}(\ell_1, \dots, \ell_n) \; \forall x \in \mathcal{R}$. Consequently, we can without loss of generality set $\min_i \ell_i^* = 0$, which is consistent with the deterministic lead time case and allows us to ignore the nonnegativity constraint on $\vec{\ell}$. Lastly, it is not difficult to see that the component subproblem (14) is a convex program,

which allows us to consider only its first-order conditions when seeking its optimal solution in section 3.5.2.

In contrast to the component subproblem, the second step of the decomposition (15) only requires end-item information: the demand rate λ, the finished goods holding and backorder cost rates h and b, and the expected total replenishment lead time $E[\max_i(X_i + \ell_i^*)]$. This observation leads us to refer to (14), (15) as a component/product decomposition.

3.5. Gumbel (CMT1) lead times

3.5.1. Motivation

The main difficulty in solving the component subproblem (14) analytically is to express $E[\max_i(X_i + \ell_i)]$ in terms of the decision variables $(\ell_1, \ldots, \ell_n)$ in closed form. Our approach is to look for families of distributions such that when the component lead times $(X_1, \ldots, X_n)$ follow distributions belonging to these families, calculating $E[\max_i(X_i + \ell_i)]$ becomes an easy problem. More specifically, we would like to identify families $\mathcal{D}$ of distributions that are closed under maximization and translation (CMT): $\mathcal{D}$ is said to be CMT if for any independent distributions[1] $(X_1, \ldots, X_n)$ belonging to $\mathcal{D}$, the distribution of $\max_i(X_i + \ell_i)$ also belongs to $\mathcal{D}$ for any $(\ell_1, \ldots, \ell_n) \in \mathcal{R}^n$. The bottom line here is that working with a CMT family of distributions makes it no harder to calculate $E[\max_i(X_i + \ell_i)]$ than to calculate the expected value of any simple distribution belonging to that family.

In this subsection, we restrict ourselves to the case of continuous uniparametric families of distributions. The next proposition is a characterization result that provides a theoretical background to our interest in the Gumbel distribution; a short discussion of the relevant literature follows.

Proposition 3. The only continuous uniparametric CMT families of distributions with support unbounded from above are the families of Gumbel distributions with the same variance.

Even though a truncated version of it was presented by Gompertz [12] in his study of human mortality as early as 1825, the Gumbel (or double-exponential) distribution is best known as one of the three possible asymptotes in extreme value theory. Interestingly, this distribution is also derived in that setting as the solution to a functional equation, called the Stability Postulate, which is linked to the one we study in the proof of proposition 3. However, we are concerned here with the exact distribution of the maximum of independent but non-identically distributed random variables, whereas classical extreme value theory primarily investigates the limiting distribution of the maximum of i.i.d. random variables. For background and a literature survey on extreme value theory and the Gumbel distribution, see [8,13].

---

[1] For ease of exposition, we use the concepts of distribution and random variable interchangeably, as no ambiguity arises from the present context.

The CMT property of the Gumbel distribution has already been exploited in the literature, and is key to the analytical tractability of the classical multinomial logit (MNL) model for consumer preferences. Moreover, a characterization of the relation between the Gumbel distribution and the MNL has already been obtained by Yellott [29], but his framework (Luce's Choice Axiom and Thurstone's model) is more contextual and less general than the CMT property. For a monograph on the theory behind the MNL, see [3]. Recently, the MNL has also been used in the operations literature by Mahajan and Van Ryzin [17], who investigate the links between consumer choices and retail inventories.

In the rest of this paper we use the notation $CMT1^m(\alpha)$ for the Gumbel distribution with cdf $F_m(x, \alpha) = \exp(-\alpha \, e^{-mx})$, $m > 0$. The mean and variance of $CMT1^m(\alpha)$ are

$$E[X] = \frac{\gamma + \ln \alpha}{m} \quad \text{and} \quad \sigma^2[X] = \frac{\pi^2}{6m^2}, \tag{16}$$

where $\gamma \simeq 0.5772$ is Euler's constant.

3.5.2. Solution of the component/product decomposition

We now assume that there are positive parameters $(\alpha_1, \ldots, \alpha_n)$ such that the component lead times $(X_1, \ldots, X_n)$ satisfy $X_i \sim CMT1^m(\alpha_i) \; \forall i$. From a modeling standpoint, the asymmetric shape of the Gumbel distribution seems well adapted to represent replenishment lead times: it is unimodal, has a very sudden start (typically, replenishment lead times are bounded from below by a physical limit), and a tail decaying more slowly (which accounts for all the problems that can occur during the replenishment process). The fact that the support of $CMT1^m(\alpha_i)$ includes negative numbers is not a major concern here, because $P(X_i \leq 0)$ is typically negligible for parameter values estimated from industrial data. However, a key restriction of $CMT1^m(\alpha_i)$ is that the standard deviations $\sigma[X_1], \ldots, \sigma[X_n]$ must all be equal to the same value $\sigma[X]$, which is dictated by the choice of m via (16). This is the price to pay in order to use the CMT property of the family $CMT1^m$: $\max_i (X_i + \ell_i) \sim CMT1^m(\sum_{i=1}^n \alpha_i \, e^{m\ell_i}) \; \forall(\ell_1, \ldots, \ell_n)$. Taking expected values yields

$$E\left[\max_i (X_i + \ell_i)\right] = \frac{\gamma + \ln\left(\sum_{i=1}^n \alpha_i \, e^{m\ell_i}\right)}{m} \quad \forall(\ell_1, \ldots, \ell_n), \tag{17}$$

which is crucial to the solution of the components subproblem (14). The first-order optimality conditions for the unconstrained version of (14) are

$$\frac{\alpha_i \, e^{m\ell_i}}{\sum_{j=1}^n \alpha_j \, e^{m\ell_j}} = \bar{h}_i \quad \forall i, \tag{18}$$

which yields the solution $\ell_i^* = (1/m)\ln(h_i/\alpha_i) \; \forall i$. By expressing ℓ_i^* in terms of the means $E[X_i]$ and common standard deviation $\sigma[X]$ rather than α_i and m, setting

$\min_i \ell_i^* = 0$ as described at the end of section 3.4, and applying (15), we obtain the policy $[S^*, \vec{\ell}^*]$ solving the component/product decomposition:

$$\begin{cases} \ell_i^* = \max_j \left(E[X_j] - \dfrac{\sqrt{6}}{\pi} \sigma[X] \ln h_j \right) - \left(E[X_i] - \dfrac{\sqrt{6}}{\pi} \sigma[X] \ln h_i \right) \quad \forall i; & \text{(19a)} \\[3ex] \rho^* = \dfrac{\lambda\sqrt{6}\,\sigma[X]}{\pi} \ln\left[\displaystyle\sum_{i=1}^{n} \exp\left(\dfrac{\pi[E[X_i] + \ell_i^*]}{\sqrt{6}\,\sigma[X]} \right) \right]; & \text{(19b)} \\[3ex] S^* \text{ is the smallest integer that satisfies } e^{-\rho^*} \displaystyle\sum_{j=0}^{S^*} \dfrac{(\rho^*)^j}{j!} \geqslant \dfrac{b}{b+h}. & \text{(19c)} \end{cases}$$

Taking the limit $\sigma[X] \to 0$ in (19) yields the solution (9) of the deterministic system, which lends support to the robustness of our approximate component/product decomposition method. In fact, the structure of $\vec{\ell}^*$ in (19a) is similar to that obtained in the deterministic case (9a). However, instead of considering each distribution X_i only through its mean $E[X_i]$, as in (9a), a correction factor $(\sqrt{6}/\pi)\sigma[X]\ln h_i$ is used to take into account both the holding cost rate h_i of each component and the common lead time standard deviation $\sigma[X]$. As expected, components with larger relative holding costs have longer postponement lead times and smaller component inventories. Also, the larger the common lead time standard deviation, the greater the impact of the component holding costs.

As expressed in (19a), the solution $\vec{\ell}^*$ depends on the lead time distributions $(X_1, \ldots, X_n)$ only through their first two moments. In the simulation study in section 4, we implement and test the policy given by (19) even when the replenishment lead times follow different distributions than assumed in this section.

Finally, recall that the $CMT1^m(\alpha)$ distributions assumed in this subsection require that all the lead time variances be identical. In section 3.6 we analyze CMT distributions with more than one parameter in order to enhance the modeling flexibility allowed for the lead time structure. In section 4 we also investigate numerically several straightforward ways of adapting (19) to the heterogeneous lead time variance case.

3.6. CMT2 lead times

By studying CMT families of distributions with two parameters instead of just one, we hope to derive better procurement policies for situations where the lead time variances are very heterogeneous. This analysis also provides the basis for assessing the robustness of policy (19) in the heterogeneous lead time variance case. The particular family under investigation is defined as follows: consider two independent random variables $Y \sim CMT1^m(\alpha)$, $m > 0$, and $W \sim CMT1^k(\beta)$, $k > 0$. Let $CMT2^{m,k}$ be the set of all distributions generated by $\max(Y, W)$ when (α, β) varies in $\mathcal{R}_+ \times \mathcal{R}_+ \setminus (0, 0)$ and m and k are fixed. The distribution in $CMT2^{m,k}$ obtained for a particular choice of (α, β) is denoted $CMT2^{m,k}(\alpha, \beta)$, and its cdf is $F(x, \alpha, \beta) = \exp(-\alpha\, e^{-mx} - \beta\, e^{-kx})$. It is

easy to check that $CMT2^{m,k}$ indeed satisfies the CMT property, which for n mutually independent random variables $(X_1, \ldots, X_n)$ takes the form

$$\left\{ \begin{array}{l} X_i \sim CMT2^{m,k}(\alpha_i, \beta_i) \\ (\ell_1, \ldots, \ell_n) \in \mathcal{R}^n \end{array} \right. \Rightarrow \max_i (X_i + \ell_i) \sim CMT2^{m,k}\left(\sum_{i=1}^{n} \alpha_i \, e^{m\ell_i}, \sum_{i=1}^{n} \beta_i \, e^{k\ell_i} \right). \quad (20)$$

The following proposition shows that the restriction noted earlier on the lead time standard deviations in the $CMT1$ case completely disappears when using $CMT2$ distributions.

Proposition 4. Let $\{(\mu_1, \sigma_1), \ldots, (\mu_n, \sigma_n)\}$ be any finite set of n points with $\sigma_i > 0 \; \forall i$. Then there exist $m > 0, k > 0$ and $\{(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)\}$ defining n random variables $\{X_1, \ldots, X_n\}$ belonging to the same $CMT2^{m,k}$ family such that $E[X_i] = \mu_i, \sigma[X_i] = \sigma_i$ and $X_i \sim CMT2^{m,k}(\alpha_i, \beta_i) \; \forall i$.

Unfortunately, the exact calculation of the first two moments of the $CMT2$ distributions has so far eluded us. However, we propose the following approximations, where $X \sim CMT2^{m,k}(\alpha, \beta)$ and (Φ, ϕ) denote the cdf and pdf of a standard normal random variable:

$$\left\{ \begin{array}{l} E[X] \simeq \dfrac{\sqrt{m^2 + k^2}}{\sqrt{2}\,mk} \ln\left[\exp\left(\dfrac{\sqrt{2}\,k(\gamma + \ln\alpha)}{\sqrt{m^2 + k^2}} \right) + \exp\left(\dfrac{\sqrt{2}\,m(\gamma + \ln\beta)}{\sqrt{m^2 + k^2}} \right) \right]; \quad (21\text{a}) \\[3ex] E[X^2] \simeq \dfrac{\pi^2 + 6(\gamma + \ln\alpha)^2}{6m^2} \Phi\left(\dfrac{\sqrt{6}\ln(\alpha^k/\beta^m)}{\pi\sqrt{m^2 + k^2}} \right) + \dfrac{\pi^2 + 6(\gamma + \ln\beta)^2}{6k^2} \Phi \\[3ex] \qquad \times \left(\dfrac{\sqrt{6}\ln(\beta^m/\alpha^k)}{\pi\sqrt{m^2 + k^2}} \right) + \dfrac{\pi \ln(\alpha^k \beta^m)\sqrt{m^2 + k^2}}{m^2 k^2 \sqrt{6}} \, \phi\left(\dfrac{\sqrt{6}\ln(\alpha^k/\beta^m)}{\pi\sqrt{m^2 + k^2}} \right). \quad (21\text{b}) \end{array} \right.$$

Both equations in (21) are used to estimate parameters in section 4.4. For a justification of these approximations, as well as a discussion of their associated errors, the reader is referred to appendix D.

Assuming now that the component lead times follow $CMT2$ distributions belonging to the same family (i.e., $X_i \sim CMT2^{m,k}(\alpha_i, \beta_i) \; \forall i$), we can use both (20) and (21a) to specify the component/product subproblem

$$\begin{array}{l} \underset{\ell_1, \ldots, \ell_n}{\text{Min}} \; \exp\left(\dfrac{\sqrt{2}\,k(\gamma - m\sum_{j=1}^{n} \bar{h}_j \ell_j + \ln \sum_{j=1}^{n} \alpha_j \, e^{m\ell_j})}{\sqrt{m^2 + k^2}} \right) \\[3ex] \qquad + \exp\left(\dfrac{\sqrt{2}\,m(\gamma - k\sum_{j=1}^{n} \bar{h}_j \ell_j + \ln \sum_{j=1}^{n} \beta_j \, e^{k\ell_j})}{\sqrt{m^2 + k^2}} \right) \qquad\qquad (22) \\[2ex] \text{subject to} \quad \ell_i \geqslant 0 \quad \forall i. \end{array}$$

In contrast to the $CMT1$ case, we have only succeeded in solving (22) numerically.

4. Simulation study

Our simulation study revolves around four main research questions: how well does the theory developed in section 3 work? What is the impact of heterogeneity in the variance of the lead times? How robust are the derived policies with respect to the distributional shapes? What happens when the synchronization assumption is relaxed?

We wrote a program in C++ to simulate the assembly system described in section 2 using standard Monte Carlo techniques, and then designed numerical experiments to answer these four questions. The data is described in section 4.1, the policies under consideration are specified in section 4.2 and the results of the four experiments are contained in sections 4.3–4.6. The 95% confidence intervals in our graphs are not depicted, because our program was designed to reiterate simulation runs (each run was 200,000 days, and the first half of each run was discarded) until the lengths of these intervals were below 1% of the corresponding average simulation value.

4.1. Data

The base case for our numerical experiments uses industrial data pertaining to the Hewlett-Packard Apollo 260 workstation that used to be manufactured in Exeter, NH. The information concerning its 11 main components is displayed in table 1, where the holding cost rates are calculated by multiplying the purchase cost of each component (the costs are disguised for confidentiality purposes) by a 33% per year interest rate. In addition, the end-product demand rate is $\lambda = 1$ unit per day, and the backorder cost rate is $b \simeq 5 \sum_i h_i = \$54.35$ per unit per day. The lead time standard deviations in table 1 are only used in section 4.4, but all the other parameters (costs, demand rates, first moments) are employed throughout our study.

Table 1
Component data.

#	Type	$E[X_i]$ (days)	$\sigma[X_i]$ (days)	Cost ($)	h_i ($/day)
1	CPU	38	9	2070	1.89
2	Monitor	32	7	1436	1.31
3	Hard drive	17	6.5	565	0.51
4	Data drive	17	8.5	750	0.68
5	Floppy drive	31	6	150	0.13
6	CD ROM	31	8	450	0.41
7	Power supply	61	10	478	0.43
8	Graphics AX	59	12	1876	1.71
9	I/O	35	4	150	0.13
10	Memory	57	5	1002	0.91
11	Chassis	49	14	3024	2.76

4.2. Policies

Five policies are investigated in this section. The first three policies are derived from the analysis in section 3. Because there are no simple heuristics for this problem in the literature, we derived the last two policies for comparison purposes by assuming deterministic lead times and component independence, respectively.

Proposed. This is the $[S, \vec{\ell}]$ policy described in (19), which is applied in a straightforward fashion when the variances of the component lead times are all the same. In cases with heterogeneous lead time variances, we set $\sigma[X]$ in (19) equal to $\sigma[X_j]$, where component j achieves the largest value of $E[X_i] - (\sqrt{6}\sigma[X_i]\ln h_i)/\pi$; this is the component with $\ell_i^* = 0$ in the common lead time variance case, and so can be loosely thought of as the bottleneck component. Numerical experiments on a variety of cases (see [9]) show that this policy outperformed by about 3% the policy that set $\sigma[X]$ in (19b) equal to the quantity in the previous sentence, and replaced $\sigma[X]\ln h_i$ in (19a) with $\sigma[X_i]\ln h_i$ $\forall i$.

Numerical-1. When the lead times follow $CMT1$ distributions belonging to the same family, this is the $[S, \vec{\ell}]$ policy obtained by solving numerically the constrained mixed nonlinear program (6), where the expression for $E[\max_i(X_i + \ell_i)]$ is given by (17). This policy is the optimal $[S, \vec{\ell}]$ policy under these distributional assumptions.

Numerical-2. This is the $[S, \vec{\ell}]$ policy obtained when solving numerically the component/product decomposition (that is, solving (22) and applying proposition 1) in situations where the lead times follow $CMT2$ distributions belonging to the same family. Note that the computations involved here are far less intensive than those necessary to derive *numerical-1*, because (22) is a convex nonlinear program with continuous decision variables and unrestrictive nonnegativity constraints.

Deterministic. This is the $[S, \vec{\ell}]$ policy obtained by applying (9) with $E[X_i]$ substituted in for X_i. This is the optimal $[S, \vec{\ell}]$ policy when the lead times are deterministic.

Independent. This is the optimal component base-stock policy $[\vec{s}]$ when assuming that the various component demands are independent (i.e., not linked through the assembly process). Under this simplifying assumption, each component has an $M/GI/\infty$ queue with service times X_i and arrival rate λ that serves its own component inventory R_i, which is initially set to s_i and subsequently depleted by one unit at each arrival to the queue. For each component inventory R_i, the unit holding cost rate is h_i and the unit backorder cost rate is $b_i = bh_i/h$; the contribution of each component to the total cost is $C_i = h_i E[R_i^+] + b_i E[R_i^-]$. By the independence assumption, the optimization over s_i can be carried out separately for each queue. An application of proposition 1 of Rubio and Wein [20] shows that s_i^* is the smallest integer that satisfies

$$e^{-\rho_i} \sum_{j=0}^{s_i^*} \frac{(\rho_i)^j}{j!} \geq \frac{b_i}{b_i + h_i},$$

where $\rho_i = \lambda E[X_i]$.

4.3. Validation of the theory

The main purpose of the first set of experiments was to assess the suboptimality of the *proposed* policy (in this section, optimality is with respect to the class of $[S, \vec{\ell}\,]$ policies). We simulated *proposed* and *numerical-1*, along with *deterministic* and *independent*, in situations with a synchronized assembly rule and $CMT1$ lead times belonging to the same family. More precisely, we considered $CMT1$ lead times with the same expected values as in table 1, but with a common standard deviation across components. The simulated steady-state cost rate for each policy is plotted in figure 4 against the common standard deviation, which ranges from 0 (deterministic lead times) to 12 days; increasing this quantity beyond 12 days generates some negative lead times.

Figure 4 shows that *proposed* is nearly optimal in these cases: for the seven values of the standard deviation, the average suboptimality was 1.1%, and the largest was 1.6% (when the standard deviation was 12 days). In contrast, the performance of *deterministic* and *independent* relative to *numerical-1* and *proposed* rapidly deteriorates as the lead time variability increases. For example, when the standard deviation is 12 days, which represents an average coefficient of variation across component lead times of 0.37 and corresponds roughly to the amount of lead time variability in the Hewlett-Packard data set in table 1, *proposed* outperforms *deterministic* and *independent* by 181% and 215%, respectively. Hence, overlooking the impact of procurement uncertainty and/or dependencies across components is a costly mistake in this setting.
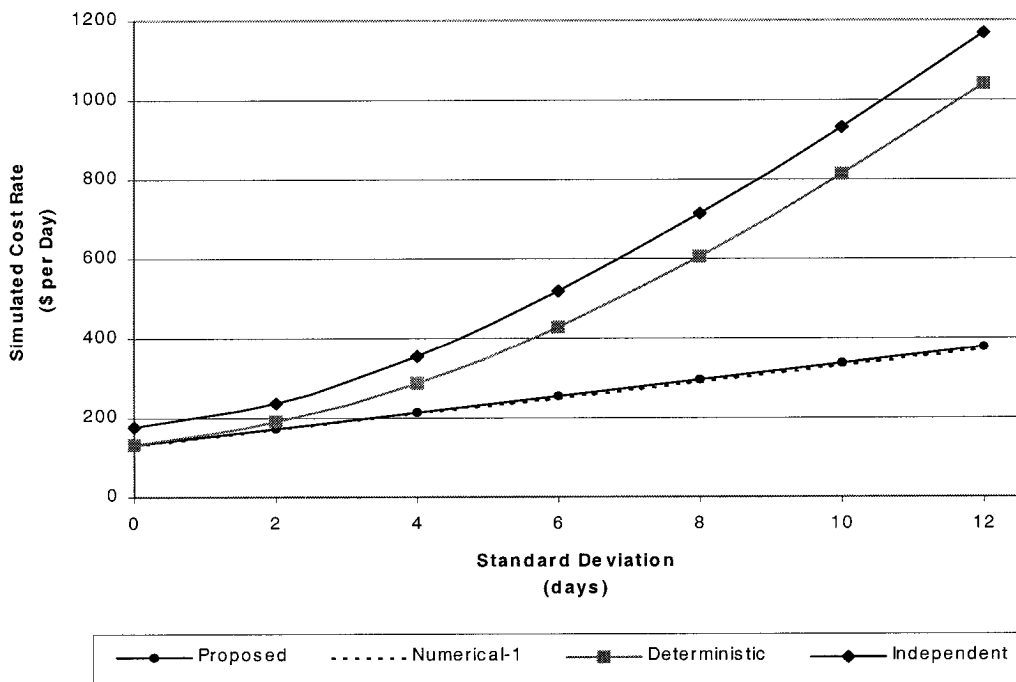

Figure 4. $CMT1$ lead times.

To further test the accuracy of *proposed*, we computed its suboptimality under the worst case value of the lead time standard deviation (12 days) in several other cases: the suboptimality was 0.8% when $b/h = 2$, and was 2.0% and 0.5%, respectively, when λ was reduced (increased, respectively) by 40%.

4.4. Impact of lead time variance heterogeneity

Our next goal was to assess the performance of *proposed* in situations where the lead time standard deviations are heterogeneous. For this purpose, we designed a set of experiments using a synchronized assembly rule and *CMT2* lead times belonging to the same family (the parameters were derived using table 1 and (21)); the simulation results in section 4.3 and the analysis in appendix D suggest that under these conditions, *numerical-2*, which is derived using the component/product decomposition, is close to optimal. Figure 5 plots the steady-state cost rates of *proposed* and *numerical-2* as a function of a procurement variability index. This index is defined as a common fraction across components of their lead time standard deviations in table 1. For example, a procurement variability index of 50% describes a situation where the lead time standard deviation of each component is equal to half of its corresponding value in table 1.

Figure 5 shows that *proposed* performs nearly as well (within 1% for all six values of the procurement variability index) as *numerical-2*. While we have not attempted to construct a perverse test example where *proposed* would fare poorly, the example in table 1 is somewhat devious in that the two components (#7 and #10) with the largest
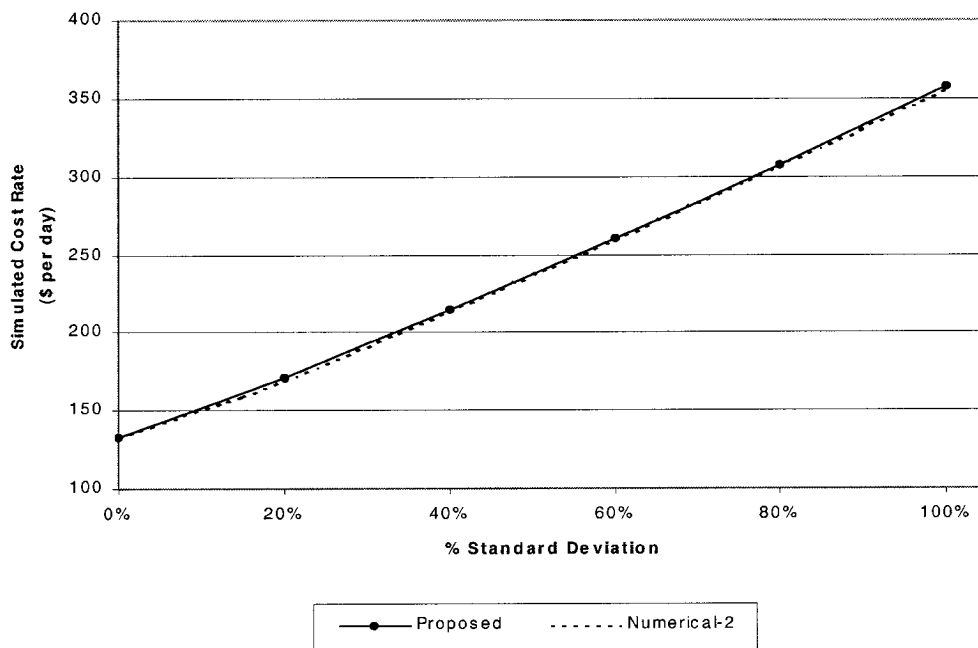

Figure 5. *CMT2* lead times.

bottleneck index $E[X_i] - (\sqrt{6}\,\sigma[X_i]\ln h_i)/\pi$ have much different lead time variances. Hence, figure 5 suggests that *proposed* should be reasonably robust when lead time variances are heterogeneous.

4.5. Robustness with respect to the shape of the lead time distributions

The distributional assumptions required to derive *proposed* involve not only the moments of the lead time distributions, but also their shape. It is therefore appropriate to investigate the robustness of *proposed* with respect to the shape of the lead time distributions. We conducted a third set of simulations on a system almost identical to the one described in section 4.3 (common lead times standard deviation ranging from 0 to 12, synchronized assembly rule), with the only difference that the component lead times followed uniform distributions. The uniform distribution was chosen because it is arguably the distribution that has the least "structure", in that all values of its support are equally likely. The simulated steady-state cost rates of *proposed*, *deterministic* and *independent* are plotted against the lead time standard deviation in figure 6.

The shapes of the curves in figure 6 are very similar to their analogues in figure 4, and *proposed* still rapidly and substantially outperforms *deterministic* and *independent* as the procurement variability increases. However, the superiority of *proposed* is slightly less spectacular here than in the system with *CMT*1 lead times: when the common standard deviation is 12 days, for example, the performances of *deterministic* and *indepen-*
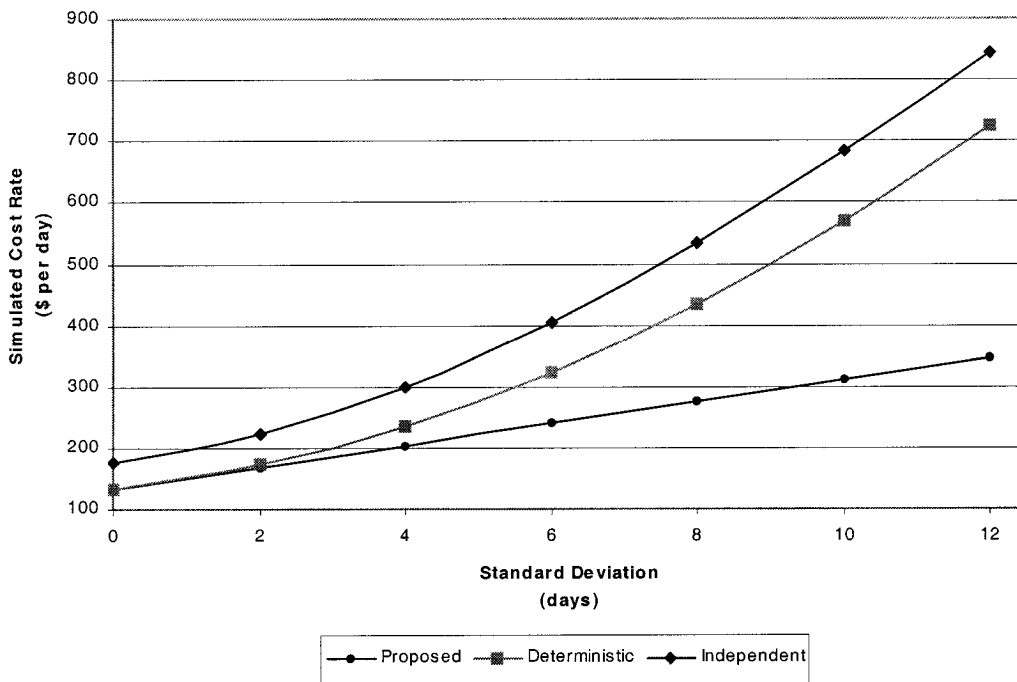

Figure 6. Uniform lead times.

dent are respectively only 108% and 143% worse than that of *proposed* (versus 181% and 215% with *CMT*1 lead times). However, because none of our prior conclusions are qualitatively challenged by these results, and because the uniform distribution is a fairly radical departure from the Gumbel distribution, we conclude that *proposed* is quite robust with respect to the distributional shape of the component lead times.

4.6. Robustness with respect to the synchronization assumption

Our final experiment assesses the impact of relaxing the synchronization assumption. Our approach was to simulate *deterministic*, *independent* and *proposed* under a set of hypotheses identical to that of section 4.3, with the only exception that the assembly operation followed a First-Come First-Served (FCFS) rule: any available component can be used to complete an assembly kit, regardless of what set of replenishment orders the other components in the kit belong to (see section 2.3).

Figure 7 shows that *deterministic* outperforms the other two policies under the FCFS assembly rule. Moreover, *independent* also performs better than *proposed* for values of the standard deviation larger than approximatively six days. Our interpretation of these results, which is based on the cost breakdowns for each policy (see [9]), is the following: in the synchronized assembly case in figure 4, *deterministic* and (to a greater extent) *independent* do not hold enough inventory, and hence incur high back-
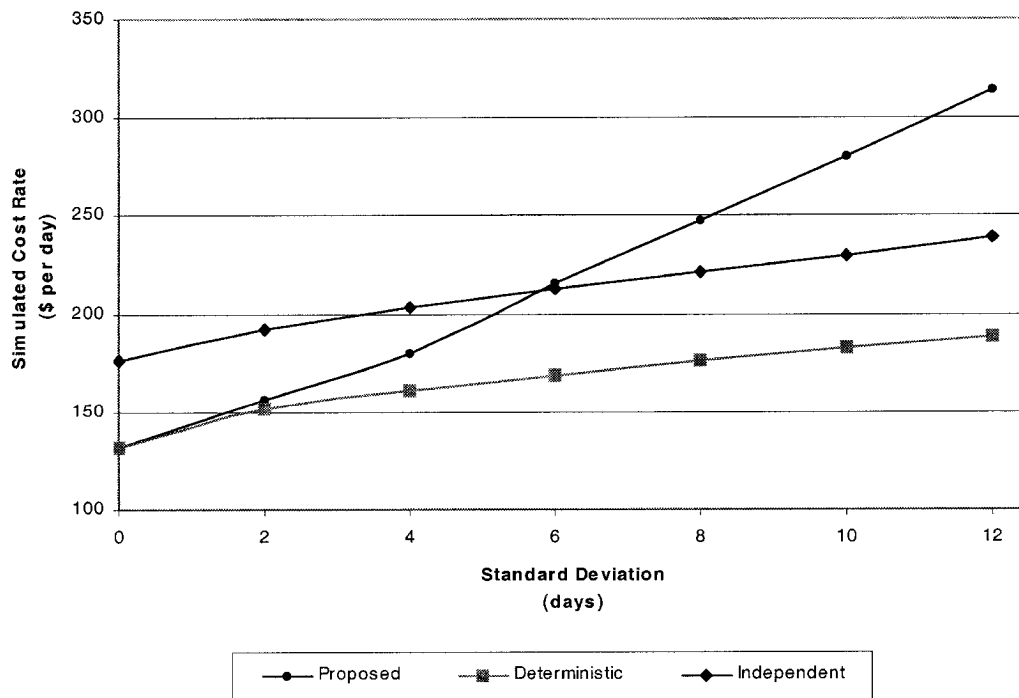

Figure 7. *CMT*1 lead times, FCFS assembly rule.

order costs because they ignore lead time variability and component dependence, respectively. FCFS assembly makes more efficient use of its component inventory than synchronized assembly, and hence requires less inventory. Therefore, *proposed*, which is derived under the assumption of synchronized assembly, overestimates the amount of inventory that is required. *Independent* still underestimates the amount of inventory required and, of the three policies, *deterministic* performs best because its underestimation of inventory due to ignoring variability in the lead times is roughly offset by its overestimation of inventory due to the policy's synchronization assumption. We also compared these three policies under the FCFS assembly rule using the distributional assumptions in section 4.4 (see [9]), and the results agreed qualitatively with figure 7. We conclude that the synchronized vs. FCFS assembly assumption has a critical impact on the relative performance of *proposed*.

5. Conclusion

We have derived in equations (19) a back-of-the-envelope policy for procuring components in a single-item assembly system with uncapacitated suppliers and stochastic procurement lead times. Two elements were key in this analysis. First, we chose to investigate the class of finished goods base stock policies with component postponement lead times, which is amenable to analysis under the synchronized assembly rule, where no mixing occurs between component orders. We could then use results from queueing theory to formulate our problem as a constrained mixed nonlinear program, solve this program exactly in the case when the lead times are deterministic, and develop an approximate decomposition method in the stochastic lead time case. Second, we used functional equations techniques to study a distributional property called closure under maximization and translation (CMT). Restricting our attention to the CMT distributions with one parameter (Gumbel distribution), we derived policy (19) in the case where the component lead times have the same variance. Introducing a class of CMT distributions with two parameters, we also used these techniques to numerically derive policies in cases with lead time variance heterogeneity.

The simplicity of the policy in equations (19) is in stark contrast with the complexity of the results obtained for similar classes of problems in the literature (with the notable exception of Glasserman and Wang's elegant asymptotics), and makes these types of results potentially amenable to implementation. Comparable performance can also be achieved by transforming policy (19) into a conventional component base stock policy with base stock levels $s_i = S - \lambda \ell_i$. Our analysis makes transparent the influence on the proposed procurement policy of model parameters such as the component holding costs and the lead time standard deviations, which is much harder to obtain with numerical results alone. Simulation results using industrial data from a Hewlett-Packard facility demonstrate that our approximate decomposition method works extremely well (less than 2% suboptimality in all cases with a common lead time variance), and allow us to tentatively conclude that our proposed policy in (19) is reasonably robust with respect to both the distributional shapes and the heterogeneity in the lead time variances. They

also show that in cases where the synchronization assumption is valid, our policy should outperform very significantly policies that ignore the stochasticity in the procurement lead times or the dependence among components caused by the assembly process.

However, the simulation results also reveal the potential danger of taking results derived for a synchronized assembly rule and applying them to a system that employs a first-come first-served assembly rule, where mixing of orders is allowed. While this issue does not arise in systems with uncapacitated suppliers with deterministic lead times or in single-item systems with capacitated suppliers modeled as single-server queues, for more complex systems some sort of suboptimal assembly rule, such as the synchronized assembly rule considered here, is usually assumed for the sake of analytical tractability. To the extent that assembly rules employed in practice are more sophisticated than those assumed in mathematical models, great care needs to be taken when attempting to apply the mathematical results to real systems.

There are at least two extensions of this work that would be worthwhile investigating. One generalization is to study multi-item systems; under the synchronized assembly rule, the multi-item system decomposes into independent subsystems (one for each end item), and our analysis carries over directly. However, this assembly rule would not take advantage of the potential benefits of component commonality, and so is not entirely satisfactory in this case. A second direction is to study a system with more serial stages (e.g., assembling integrated circuits into boards, and boards into products) and a capacitated (e.g., single-server queue) assembly process. The analysis in this paper combined with results in [20] make this generalization conceptually – if not computationally – straightforward, if one is willing to restrict to the class of CONWIP policies.

Beyond the context of supply chains, the concept of CMT distributions and the functional equations analysis in sections 3.5.1 and 3.6 may also prove useful for the performance analysis and optimization of PERT networks, reliability systems, telecommunication networks, and other stochastic systems where the maximum of independent but non-identically distributed random variables plays a crucial role.

Acknowledgements

Appendix A. Proof of proposition 2

Differentiating $C_{det}(S, \rho)$ in (8) gives

$$\frac{\partial}{\partial \rho} C_{det}(S, \rho) = b - (h + b) e^{-\rho} \sum_{i=0}^{S-1} \frac{\rho^i}{i!},$$

which can be interpreted in terms of the queue length Q as

$$\frac{\partial}{\partial \rho} C_{\text{det}}(S, \rho) = b - (h + b)P(Q \leqslant S - 1). \tag{A.1}$$

A consequence of proposition 1 is that in the optimal solution (S^*, ρ^*) of (8), we have

$$P(Q \leqslant S^* - 1) < \frac{b}{b + h}. \tag{A.2}$$

Combining (A.1) and (A.2) gives $(\partial / \partial \rho) C_{\text{det}}(S^*, \rho^*) > 0$. Since $C_{\text{det}}(S^*, \cdot)$ is a continuously differentiable function, the only way that this last inequality can occur is if ρ^* does not belong to the interior of the feasible region. Hence, the constraint $\rho \geqslant \lambda \max_i(X_i)$ is binding at ρ^*; i.e., $\rho^* = \lambda \max_i(X_i)$.

Appendix B. Proof of proposition 3

We define a continuous uniparametric family of distributions $(X(\alpha))$ to be any function $F(z, \alpha) : \mathcal{R} \times D \to [0, 1]$, D (interval) $\subset \mathcal{R}$, such that: (i) F is continuous and differentiable with respect to both variables, and its partial derivatives are continuous; (ii) $F(\cdot, \alpha)$ is nondecreasing, $\lim_{z \to +\infty} F(z, \alpha) = 1$ and $\lim_{z \to -\infty} F(z, \alpha) = 0$ $\forall \alpha$; and (iii) F is weakly reducible on the right over uncountable sets, i.e., $F(z, \alpha) = F(z, \beta) \neq 0$ $\forall z \in U$ (uncountable) $\Rightarrow \alpha = \beta$. This definition essentially specify the cdfs in the family through $P(X(\alpha) \leqslant z) = F(z, \alpha)$. It is not the most general definition one could employ, and condition (iii), in particular, is quite restrictive, as it will not allow us to find CMT families containing distributions with support bounded from above. However, our goal here is to derive simple insights rather than find minimal hypotheses. The CMT property is then equivalent to the following system of functional equations:

$$\begin{cases} F\big(z, M(\alpha, \beta)\big) = F(z, \alpha)F(z, \beta), & \text{(B.1a)} \\ F(z - \ell, \alpha) = F\big(z, J(\alpha, \ell)\big) & \forall \ell \in \mathcal{R}, & \text{(B.1b)} \end{cases}$$

where the functions M and J give the values of the resulting distributional parameters. Equation (B.1a) follows from the fact that the product of the cdfs of two independent random variables is the cdf of their maximum. This equation essentially states that for any two distributions in the family (characterized by parameters α and β), their maximum also belongs to it (with a parameter given by $M(\alpha, \beta)$). Because of the associativity property of both the maximum operation and the regular product, it is sufficient to consider only the case of two distributions. Equation (B.1b) states that the distribution obtained when translating by any real number ℓ any distribution in the family (with a parameter α) still belongs to the family, and is characterized by the parameter $J(\alpha, \ell)$.

The resolution of (B.1) is as follows. Equation (B.1a) is known as the Maximum Stability equation, which is a special case of the Generalized Distributivity equation [5, example 6.2.3]. However, we cannot invoke their results because the required hypotheses rule out the case of distributions with support unbounded from

above. Instead, notice that every distribution $F(\cdot, \alpha)$ in the family is non-zero over an uncountable set of values. This and (iii) imply that M is reducible on the right: $M(\alpha, \beta) = M(\alpha, \delta) \Rightarrow F(z, \alpha)F(z, \beta) = F(z, \alpha)F(z, \delta) \; \forall z \Rightarrow \exists U$ uncountable s.t. $F(z, \beta) = F(z, \delta) \; \forall z \in U \Rightarrow \beta = \delta$. Reducibility on the left follows from the commutativity of M. Moreover, because of the associativity of multiplication, (B.1a) and (iii) imply $M(M(\alpha, \beta), \delta) = M(\alpha, M(\beta, \delta))$, so that M also satisfies the so-called associativity equation. The hypotheses of the theorem in [2, section 6.2.2] are thus satisfied, and we can therefore claim the existence of a continuous and strictly monotonic function g such that $M(\alpha, \beta) = g[g^{-1}(\alpha) + g^{-1}(\beta)]$.

Substituting this last expression in (B.1a) where the right side is evaluated at $(g(\alpha), g(\beta))$ gives $F(z, g(\alpha + \beta)) = F(z, g(\alpha))F(z, g(\beta))$; i.e., the function $F(z, g(\cdot))$ satisfies the second Cauchy equation [2, section 2.1.2]. The nontrivial general solution to this equation is $F(z, \alpha) = \exp[c(z)g^{-1}(\alpha)]$. Substituting this into (B.1b) gives $c(z - \ell)g^{-1}(\alpha) = c(z)g^{-1}(J(\alpha, \ell))$. Discarding the trivial solution $g^{-1}(\alpha) = 0$ that implies $F(z, \alpha) = 1 \; \forall z$, we can write

$$c(z - \ell) = c(z)\frac{g^{-1}(J(\alpha, \ell))}{g^{-1}(\alpha)}.$$

Because c is independent of α, the fraction in this last equation must only depend on ℓ, so that $c(z + \ell) = c(z)d(\ell)$ for some function $d(\cdot)$. Thus, if there exists one value of z such that $c(z) = 0$, then $c \equiv 0$. Moreover, evaluation at $z = 0$ gives $c(\ell) = c(0)d(\ell)$, so either $c \equiv 0$ or c satisfies $c(z + \ell) = c(0)^{-1}c(z)c(\ell)$ and $c(z) \neq 0 \; \forall z$. The function $c(\cdot)/c(0)$ is therefore a solution of the second Cauchy equation already encountered above, with nontrivial general solution $c(z)/c(0) = e^{Kz}$. Substitution in F gives $F(z, \alpha) = \exp[c(0)g^{-1}(\alpha)e^{Kz}]$. Provided that $c(0)g^{-1}(\alpha) < 0$ and $K < 0$, this defines a family of Gumbel distributions with variance $\pi^2/6K^2$, which indeed satisfies (B.1) and (i)–(iii). Other cases do not correspond to probability distributions.

Appendix C. Proof of proposition 4

Let $\hat{\sigma} = \max_i \sigma_i$, $\check{\sigma} = \min_i \sigma_i$, $m = \pi/(\hat{\sigma}\sqrt{6})$ and $k = (\pi/\check{\sigma}\sqrt{6})$. Note that $g_{m,k}(\alpha, \beta) = \sigma[CMT2^{m,k}(\alpha, \beta)]$ is a continuous function from $\Omega = \mathcal{R}_+ \times \mathcal{R}_+ \setminus (0, 0)$ into $\mathcal{R}_+$ (this can be seen by considering the integral expression of $g_{m,k}(\cdot, \cdot)$). Moreover, since the standard deviation of a $CMT1^m(\alpha)$ distribution is $\pi/(m\sqrt{6})$, we have $g_{m,k}(x, 0) = \hat{\sigma}$ and $g_{m,k}(0, x) = \check{\sigma} \; \forall x > 0$. By the continuity of $g_{m,k}(\cdot)$ and the connectivity of Ω, there exist $(y_1, z_1), \ldots, (y_n, z_n) \in \Omega$ such that $g_{m,k}(y_i, z_i) = \sigma_i$ $\forall i \in \{1, \ldots, n\}$. If we define $\ell_i = \mu_i - E[CMT2^{m,k}(y_i, z_i)]$, $\alpha_i = y_i e^{m\ell_i}$ and $\beta_i = z_i e^{k\ell_i}$, then the closure under translation property of the $CMT2^{m,k}$ family implies that

$$\begin{cases} E[CMT2^{m,k}(\alpha_i, \beta_i)] = E[CMT2^{m,k}(y_i, z_i)] + \ell_i = \mu_i; \\ \sigma[CMT2^{m,k}(\alpha_i, \beta_i)] = \sigma[CMT2^{m,k}(y_i, z_i)] = \sigma_i. \end{cases}$$

Appendix D. On approximations (21)

D.1. Mean approximation

We adopt a three-step procedure to derive (21a): characterize

$$f_{m,k}(\alpha, \beta) = E\left[CMT2^{m,k}(\alpha, \beta)\right]$$

through functional equations as much as possible, define a class of functions satisfying all these constraints, and perform a simulation-based optimization within that class. These steps are described below in more detail.

Step 1: A system of functional constraints. Using the known expression for the mean of the $CMT1$ distribution and the integral expression of $f_{m,k}(\cdot, \cdot)$, we can write the functional system

$$\begin{cases} f\left(\alpha\, e^{mx}, \beta\, e^{kx}\right) = f(\alpha, \beta) + x \quad \forall x \in \mathcal{R}; \\ f(\alpha, 0) = E\left[CMT1^m(\alpha)\right] = \dfrac{\gamma + \ln \alpha}{m}; \\ f(0, \beta) = E\left[CMT1^k(\beta)\right] = \dfrac{\gamma + \ln \beta}{k}. \end{cases} \tag{D.1}$$

Invoking the homogeneous functional equation theorem [5, section 4.3.1, p. 76], there exists a real function $h_{m,k}(\cdot)$ independent of (α, β) such that

$$f_{m,k}(\alpha, \beta) = \begin{cases} \dfrac{1}{2}\left(\dfrac{\ln \alpha}{m} + \dfrac{\ln \beta}{k}\right) + h_{m,k}\left(\dfrac{\alpha^{1/m}}{\beta^{1/k}}\right) & \text{if } \alpha > 0, \beta > 0; \\ \dfrac{\gamma + \ln \alpha}{m} & \text{if } \alpha > 0, \beta = 0; \\ \dfrac{\gamma + \ln \beta}{k} & \text{if } \alpha = 0, \beta > 0; \\ -\infty & \text{if } \alpha = 0, \beta = 0. \end{cases} \tag{D.2}$$

We now characterize the remaining unknown $h_{m,k}(\cdot)$ in (D.2). Recall that for any two numbers y and w, we have $\max(y, w) = (y + w + |y - w|)/2$. Consider two independent random variables $Y \sim CMT1^m(\alpha)$ and $W \sim CMT1^k(\beta)$. It follows that

$$f_{m,k}(\alpha, \beta) = \frac{1}{2}\left(\frac{\gamma + \ln \alpha}{m} + \frac{\gamma + \ln \beta}{k} + E\left[|Y - W|\right]\right). \tag{D.3}$$

Comparing (D.3) with the solution given by (D.2) yields

$$h_{m,k}\left(\frac{\alpha^{1/m}}{\beta^{1/k}}\right) = \frac{1}{2}\left(E\left[|Y - W|\right] + \gamma\left(\frac{1}{m} + \frac{1}{k}\right)\right). \tag{D.4}$$

The function $h_{m,k}(\cdot)$ in (D.4) is symmetric in Y and W, and therefore

$$h_{m,k}(t) = h_{k,m}\left(\frac{1}{t}\right). \tag{D.5}$$

Let $Y(\theta)$ be the *CMT*1 random variable obtained by shrinking the expected value and the standard deviation of Y by $\theta > 0$, and let $W(\theta)$ be defined analogously. Because this transformation just amounts to changing the unit of measure, $E[\max(Y(\theta), W(\theta))]$ and $E[|Y(\theta) - W(\theta)|]$ are obtained by dividing their original values by the same quantity θ. We have $Y(\theta) \sim CMT1^{\theta m}(\alpha)$, so that

$$h_{\theta m, \theta k}\left(t^{1/\theta}\right) = \frac{1}{\theta} h_{m,k}(t) \quad \forall \theta > 0. \tag{D.6}$$

Because the function $f_{m,k}$ is continuous on $(\mathcal{R}^+)^2 \backslash (0, 0)$ (this can be seen from the integral expression), equation (D.2) implies that

$$\begin{cases} \lim_{\beta \to 0^+} \left[\frac{1}{2} \left(\frac{\ln \alpha}{m} + \frac{\ln \beta}{k} \right) + h_{m,k} \left(\frac{\alpha^{1/m}}{\beta^{1/k}} \right) \right] = \frac{\gamma + \ln \alpha}{m}; \\ \lim_{\alpha \to 0^+} \left[\frac{1}{2} \left(\frac{\ln \alpha}{m} + \frac{\ln \beta}{k} \right) + h_{m,k} \left(\frac{\alpha^{1/m}}{\beta^{1/k}} \right) \right] = \frac{\gamma + \ln \beta}{k}. \end{cases} \tag{D.7}$$

System (D.7) is equivalent to

$$\begin{cases} h_{m,k}(t) = \frac{1}{2} \ln t + \frac{\gamma}{m} + \phi_{m,k}(t) & \text{with } \lim_{t \to +\infty} \phi_{m,k}(t) = 0; \\ h_{m,k}(t) = -\frac{1}{2} \ln t + \frac{\gamma}{k} + \psi_{m,k}(t) & \text{with } \lim_{t \to 0^+} \psi_{m,k}(t) = 0. \end{cases} \tag{D.8}$$

Finally, using the closure under maximization property of the *CMT*1 distribution, we have that $f_{m,m}(\alpha, \beta) = (\gamma + \ln(\alpha + \beta))/m$. Substituting this into the first equation of (D.2) yields the boundary condition

$$h_{m,m}(t) = \frac{1}{m} \left(\ln \left(\frac{1}{t^{m/2}} + t^{m/2} \right) + \gamma \right). \tag{D.9}$$

Step 2: A feasible solution. Let $\Psi(m, k)$ be an arbitrary positive function such that $\Psi(m, m) = m$, $\Psi(m, k) = \Psi(k, m)$ and $\Psi(\theta m, \theta k) = \theta \Psi(m, k)$. Then the system of constraints (D.5), (D.6), (D.8) and (D.9) are satisfied by the class of functions

$$h_{m,k}(t) = \frac{1}{\Psi(m,k)} \ln \left(\left(e^{\gamma/m} \sqrt{t} \right)^{\Psi(m,k)} + \left(\frac{e^{\gamma/k}}{\sqrt{t}} \right)^{\Psi(m,k)} \right). \tag{D.10}$$

More specifically, we have investigated the approximations obtained when using in (D.10) the functions

$$\Psi_{[\omega]}(m, k) = \sqrt[\omega]{\frac{m^\omega + k^\omega}{2}}, \quad \omega \in \mathcal{R},$$

$$\Psi_{\text{MAX}}(m, k) = \max(m, k) \quad \text{and} \quad \Psi_{\text{SQRT}}(m, k) = \sqrt{mk}. \tag{D.11}$$

Step 3: Numerical validation. We assessed the performance of the functions in (D.11) by comparing their values with the results of Monte Carlo simulations for various choices of m, k, α, β. The smallest relative errors were achieved by $\Psi_{[\omega]}(m, k)$. We optimized over ω by minimizing the sum of square distances between the simulated values and the approximations calculated with $\Psi_{[\omega]}(m, k)$ for every point. Consistently across sets, the optimal value of ω was very close to -2, resulting in approximation errors of less than 7% for all the points tested. Substituting $\Psi_{[-2]}(m, k)$ into (D.10), and substituting the resulting function $h_{m,k}(\cdot)$ into the first equation of (D.2) yields after some algebra the first equation in (21).

D.2. Second moment approximation

Equation (21b) is obtained by approximating the two *CMT*1 parent random variables Y and W with independent normally distributed random variables Y_N and W_N having the same first and second moments as Y and W, respectively. We then substitute the expressions giving the moments of Y and W as functions of (m, k, α, β) into the known formula for the second moment of the maximum of Y_N and W_N [7]

$$E\left[\left(\max(Y_N, W_N)\right)^2\right] = \left(E[Y_N]^2 + \sigma^2[Y_N]\right)\Phi(\delta) + \left(E[W_N]^2 + \sigma^2[W_N]\right)\Phi(-\delta)$$

$$+ \left(E[Y_N] + E[W_N]\right)\sqrt{\sigma^2[Y_N] + \sigma^2[W_N]}\,\phi(\delta), \qquad \text{(D.12)}$$

where

$$\delta = \frac{E[Y_N] - E[W_N]}{\sqrt{\sigma^2[Y_N] + \sigma^2[W_N]}}.$$

The largest approximation error for the second moment of $CMT2^{m,k}(\alpha, \beta)$ was 17%.

References

[1] H.S. Abhyankar, A robust, computationally efficient methodology to set service levels for components in assemble-to-order environments, MIT Sloan School of Management, Cambridge, MA (1998).

[2] J. Aczél, *Lectures on Functional Equations and Their Applications* (Academic Press, New York, 1966).

[3] S.P. Anderson, A. de Palma and J.F. Thisse, *Discrete Choice Theory of Product Differentiation* (MIT Press, Cambridge, MA, 1992).

[4] R. Anupindi and S. Tayur, Managing stochastic multiproducts systems: Model, measures, and analysis, Oper. Res. 46 Supp. 3 (1998) S98–S111.

[5] E. Castillo and M.R. Ruiz-Cobo, *Functional Equations and Modelling in Science and Engineering* (Marcel Dekker, New York, 1992).

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

[7] C.E. Clark, The greatest of a finite set of random variables, Oper. Res. 9 (1961) 145–162.

[8] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics* (Krieger, 1987).

[9] J. Gallien, Forthcoming Ph.D. thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA (1999).

[10] P. Glasserman and Y. Wang, Leadtime-inventory trade-offs in assemble-to-order systems, Columbia Business School, New York (1997).

[11] P. Glasserman and Y. Wang, Near-optimal base-stock policies in assemble-to-order systems, MIT Sloan School of Management, Cambridge, MA (1999).

[12] B. Gompertz, On the nature of the function expressive of the law of human mortality, Philos. Trans. Roy. Soc. London Ser. A 115 (1825) 513–580.

[13] E.J. Gumbel, *Statistics of Extremes* (Columbia Univ. Press, New York, 1958).

[14] W.H. Hausman, H.L. Lee and A.X. Zhang, Order fulfillment time reliability in a multi-item inventory system, IEEM Department, Stanford University, Stanford, CA (1995).

[15] W.J. Hopp and M.L. Spearman, Setting safety leadtimes for purchased components in assembly systems, IIE Transactions 25 (1993) 2–11.

[16] H.J. Kushner and U.A.W. Tetzlaff, Control and optimal control of assemble to order manufacturing systems under heavy traffic, Division of Applied Mathematics, Brown University, Providence, RI (1997).

[17] S. Mahajan and G.J. Van Ryzin, Retail inventories and consumer choice, in: *Quantitative Models for Supply Chain Management*, International Series in Operations Research and Management Science (1999) pp. 491–551.

[18] J.E. Nemec, Diffusion and decomposition approximations of stochastic models of multiclass processing networks, Ph.D. thesis, MIT Sloan School of Management, Cambridge, MA (1998).

[19] K. Rosling, Optimal inventory policies for assembly systems under random demands, Oper. Res. 37 (1989) 565–579.

[20] R. Rubio and L.M. Wein, Setting base stock level using product-form queueing networks, Managm. Sci. 42 (1996) 259–268.

[21] C.P. Schmidt and S. Nahmias, Optimal policy for a two-stage assembly system under random demand, Oper. Res. 33 (1985) 1130–1145.

[22] E. Schraner, Capacity/inventory trade-offs in assemble-to-order systems, IBM T.J. Watson Research Center, Yorktown Heights, NY (1997).

[23] J.-S. Song, Evaluation of order-based backorders, IEOR Department, Columbia University, New York (1997).

[24] J.-S. Song, Performance evaluation in an assemble-to-order system with batch ordering, IEOR Department, Columbia University, New York (1998).

[25] J.-S. Song, S.H. Xu and B. Liu, Order-fulfillment performance measures in an assemble-to-order system with stochastic leadtimes, Oper. Res. 97 (1999) 131–149.

[26] R. Srivanasan, R. Jayaraman, J.A. Rappold, R.O. Roundy and S. Tayur, Procurement of common components in a stochastic environment, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY (1998).

[27] S. Tayur, Computing optimal stock levels for common components in an assembly system, GSIA, Carnegie Mellon University, Pittsburg, PA (1995).

[28] W.E. Wilhelm and P. Som, Analysis of stochastic assembly with GI distributed assembly time, Department of Industrial Engineering, Texas A&M University, College Station, TX (1996).

[29] J.I. Yellott, The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement, and the double exponential distribution, J. Math. Psychology 15 (1977) 109–144.

[30] A.X. Zhang, Determining jointly optimal inventory policies in a one-stage assembly system is as simple as solving a Newsboy problem, Department of Information and Operations Management, University of Southern California, Los Angeles, CA (1995).

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

[32] R.Q. Zhang, Time delay in a production–inventory system, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI (1996).