A Separation Principle for Assemble-to-Order Systems with Expediting

Erica L. Plambeck
Amy R. Ward

June 2003
A Separation Principle for Assemble-to-Order Systems with Expediting

Erica L. Plambeck
Graduate School of Business, Stanford University, Stanford, CA 94305
elp@stanford.edu

Amy R. Ward
Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta GA 30332
amy@isye.gatech.edu

Abstract

In an assemble-to-order system, a wide variety of products are rapidly assembled from component inventories, in response to customer orders. Orders must be filled within a product-specific target leadtime. In the event that some of the components required to fill an order are out-of-stock, these components are expedited at a high cost per unit. The objective is to minimize the expected infinite horizon discounted cost of nominal component production and expediting. This discounted formulation captures financial inventory holding costs. The levers for control are (1) sequencing orders for assembly (2) component production (3) component expediting. Under the assumption that expedited components have zero leadtime, the multi-dimensional assemble-to-order control problem separates into single-item inventory control problems. The optimal production and expediting policy for each component is independent of all other components. Hence the literature on single-item inventory management with expediting or lost sales is directly relevant to the control of assemble-to-order systems.

Subject classifications: inventory/production: assemble-to-order with delay constraints and expediting; stochastic optimal control

June 2003
1 Introduction

Motivated by the phenomenal success of Dell computer, companies as diverse as General Electric, American Standard, BMW, IBM, Timbuk2, and National Bicycle are choosing to assemble-to-order. They avoid tying up capital in finished goods inventory that customers may or may not want, and can offer a wide variety of products. However, to fill customer orders within a short leadtime does require inventory: inventory of components that can be rapidly assembled into final products. In the event of a shortage, the manufacturer must dynamically allocate scarce components to customer orders for various products, and/or pay to expedite component production. In practice, most firms expedite production on an ad hoc basis and adopt simple static rules to sequence customer orders for assembly, such as FIFO or proportional allocation (Agrawal and Cohen, 2001; Plambeck, 2001). Even Dell computer has only recently begun to optimize component production and expediting, and the sequencing of orders for assembly dynamically, based on real-time information (Perman, 2001).

Dynamic control of an assemble to order system is challenging because the state space (outstanding orders and their due dates, and the inventory and production status for each component) is very large. For example, the Dell Optiplex facility assembles more than 4,000 product configurations from more than 100 different components (Roberts, 2002).

The structure of an optimal control policy for an assemble-to-order (ATO) system with multiple products is not known. A base stock policy is optimal for managing the inventory of a single item with linear production costs, and a “balanced” base stock policy is optimal for a single-item assembly system with deterministic component leadtimes (Rosling, 1986; Chen and Zheng, 1994). Motivated by these results, ATO researchers have assumed base stock control, then proceeded to characterize system performance and optimize the base stock levels. Song and Zipkin (2001) provide an excellent survey of the literature. Because ATO systems are so complex, most of the results are sophisticated approximations or algorithms.

A common formulation in the ATO literature is: choose component base stock levels to minimize component inventory holding costs subject to a lower bound on the fill rate, the fraction of customer orders assembled within the target leadtime. Under the assumption that component leadtimes are i.i.d. random variables, (Song, 1998) and (Lu. Song and Yao, 2003) show how to compute fill rates, and (Cheng et al., 2002) propose an efficient algorithm for optimizing base stock levels. The assumption of i.i.d leadtimes is reasonable when transportation comprises most of the component leadtime. However, it is often the case that component production is limited by supply contracts or physical capacity constraints. Song, Xu and Liu (1999) and Glasserman and
Wang (1998; 1999) explicitly model the component production facilities as single server queues. Song, Xu and Liu (1999) show how to compute fill rates. Glasserman and Wang (1998) undertake an asymptotic analysis as the fill rate approaches one, and establish a linear relationship between the target maximum delay and the component base stock level. They provide simple expressions for near-optimal base stock levels (Glasserman and Wang, 1999). All of the above “fill rate” models are formulated in continuous time; Hausman, Lee and Zhang (1998) consider periodic review. We will build on the formulation with target leadtimes, assuming that components are expedited to achieve a 100% fill rate.

Two recent papers suggest that base stock control of component inventory is not, in general, optimal; one must account for interactions between components. (Kushner, 1999; Plambeck and Ward, 2003a) The contribution of this paper is to identify a class of assemble-to-order systems for which the control problem separates: under an optimal policy, one manages production and inventory of each component without regard for the status of other components. We require that each customer pays when he places an order, that the leadtime to expedite a component is negligible, and that components must be expedited, if needed, to fill every order within its target leadtime. Then, we can leverage the existing literature on single-item inventory management with expediting or lost sales to characterize the optimal policy for managing each component. In particular, from (Veatch and Wein, 1996) we obtain conditions under which base stock control of component inventory is optimal.

2 Model Formulation

Consider an assemble-to-order system with $J$ components and $K$ products, as shown in Figure 1. Orders for product $k$ arrive according to a stochastic process $D_k(t)$ which denotes the cumulative number of orders for product $k$ up to time $t$, for $k = 1, \ldots, K$ and $t \geq 0$. To assemble a product of type $K$ requires $a_{kj}$ components of type $j$, where $a_{kj}$ is a non-negative integer. Orders for product $k$ must be assembled within the target leadtime $L_k$. In other words, an order arriving at time $t$ must be assembled by time $t + L_k$. Customers pay upon placing an order, as for Dell computers.

We have three levers for control: sequencing orders for assembly, component production planning, and component expediting. First, we can restrict attention to sequencing policies in which orders for each product $k$ are assembled FIFO. It is therefore sufficient to specify $A_k(t)$, the number of product $k$ assembled up to time $t$, for $k = 1, \ldots, K$ and $t \geq 0$. Assembly is instantaneous if the required components are in stock.
The second dynamic control is the nominal production plan \( N_j(t) \) for each component \( j = 1, ..., J \) via the primary (cheap) supply mode. Actual production is stochastic, and may deviate from the plan. The stochastic process \( P_j(t; N_j) \) denotes the cumulative number of components delivered to the assembly facility by time \( t \) via the primary (cheap) supply mode. A per-unit charge of \( c_j(t; N_j) \) is assigned to nominal components of type \( j \) delivered at time \( t \).

As an illustrative example, \( N_j(t) \) could be the number of orders for component \( j \) issued up to time \( t \). Suppose for the moment that transportation leadtimes are i.i.d. random variables \( d_i \), and there is a fixed transportation cost \( K \) and variable cost \( v \) for each order. Let \( T_i \) denote the time at which the \( i^{th} \) order is placed (the \( i^{th} \) jump in the process \( N_j(t) \)), and let \( Q_i = N_j(T_i) - N_j(T_i^-) \) denote the order quantity. Then nominal production is given by

\[
P_j(t; N_j) = \sum_{i=1}^{N_j(t)} \{T_i + d_i \leq t\} Q_i
\]

where \( \{\} \) is the indicator function, and the per-unit cost is

\[
c_j(T_i + d_i, N_j) = v + K/Q_i.
\]

As a second illustrative example, \( N_j(t) \) could be the amount of time that a production facility dedicates to component \( j \) up to time \( t \). (On the flip side, \( t - N_j(t) \) could be interpreted as cumulative idle time). Then, with exponential production times, nominal production is given by

\[
P_j(t; N_j) = Z(N_j(t))
\]

where \( Z \) is a Poisson process. Here, \( N_j(t) \) must be nondecreasing and satisfy \( N_j(t) \leq t \). Finally, note that \( P_j(\cdot; N_j) \) may decrease with time. If \( P_j(\cdot; N_j) \) decreases at time \( t \), \( c_j(t; N_j) < 0 \) should be interpreted as a salvage value.

The third dynamic control is expediting. Additional components may be expedited instantaneously at a high per-unit charge of \( x_j \). \( X_j(t) \) is the cumulative number of type-\( j \) components expedited to the assembly facility by time \( t \), for \( j = 1, ..., J \). The resulting inventory position for
component $j$ at the assembly facility is:

$$I_j(t) = P_j(t; N_j) + X_j(t) - \sum_{k=1}^{K} a_{kj} D_k(t).$$

The physical inventory level is required to be nonnegative:

$$P_j(t; N_j) + X_j(t) - \sum_{k=1}^{K} a_{kj} A_k(t) \geq 0.$$

One cannot assemble components into products before they are delivered to the assembly facility.

A few technicalities are needed. The control processes $(A, N, X)$, the demand process $D$ and the nominal production processes $P$ are defined on the same probability space $(\Omega, \mathcal{F}, \mathcal{P})$. The control processes must be nonanticipating with respect to the filtration $\mathcal{F}$. Furthermore, $(D, A, X)$ are nondecreasing and RCLL (right continuous with left limits). For each $t$, the ongoing demand process $D_k(s)_{s \geq t}$ is independent of $(A(s), N(s), P(s), X(s))_{0 \leq s \leq t}$. Conditional on the planned production $N_j$, the nominal component production and cost processes $(P_j(\cdot; N_j), c_j(\cdot; N_j))$ are independent of $\{(P_l(\cdot; N_l), c_l(\cdot; N_l), N_l)\}_{l \neq j}$.

The objective is to minimize the expected infinite horizon discounted cost of nominal component production and expediting, subject to assembling orders for product $k$ within the target
\[
\min_{A,N,X} E \left[ \sum_{j=1}^{J} \left( \int_0^\infty e^{-\delta t} c_j(t; N_j) dP_j(t; N_j) + \int_0^\infty e^{-\delta t} x_j dX_j(t) \right) \right]
\]
subject to:
\[
A_k(t) \geq D_k(t - L_k) \quad \text{for } t \geq 0 \text{ and } k = 1, \ldots, K
\]
\[
\sum_{k=1}^{K} a_{kj} A_k(t) \leq P_j(t; N_j) + X_j(t) \quad \text{for } t \geq 0 \text{ and } j = 1, \ldots, J
\]
A, N, X are nonanticipating
A, X are nonnegative, nondecreasing and RCLL

where \( D_k(u) = 0 \) for \( u < 0 \) and \( \delta > 0 \) is the discount rate. The integrals should be interpreted as Riemann-Stieltjes integrals in the usual sense. Constraint (3) ensures 100% fill rate within the target leadtime. Constraint (4) dictates that products cannot be assembled without the required components. Supply constraints could be incorporated by setting \( c_j(t,N_j) = \infty \) for any nominal production plan \( N_j \) that is infeasible.

This problem formulation captures financial inventory holding costs incurred when components are produced before they are required for assembly, but not physical inventory holding costs. The formulation implicitly assumes that revenues do not depend upon the assembly sequence \( A \). This is true for Dell, where customers pay upon placing an order. If customers pay upon delivery, the objective function should incorporate a financial order holding cost, which will interfere with the separation obtained in the next section (Plambeck and Ward, 2003a).

3 Separation Principle

We will now show how to separate the assemble-to-order control problem (2)-(6) into \( J \) single-item inventory control problems, one for each component. The first step is to characterize the optimal policy for assembly and expediting.

Proposition 1 Given a nominal production plan \( P \), an optimal policy for assembly sequencing and expediting is:

\[
A^*_k(t) = D_k(t - L_k) \quad \text{for } t \geq 0 \text{ and } k = 1, \ldots, K
\]
\[
X^*_j(t) = \sup_{0 \leq s \leq t} \left[ \sum_{k=1}^{K} a_{kj} D_k(s - L_k) - P_j(s; N_j) \right]^+ \quad \text{for } t \geq 0 \text{ and } j = 1, \ldots, J
\]
Proof: For a given \( t \) and for any \( j = 1, \ldots, J \), the minimum value of \( X_j(t) \) that satisfies (4) is

\[
X_j(t) = \sum_{k=1}^{K} a_{kj} A_k(t) - P_j(t; N_j).
\]

Because the expediting process \( X_j(t) \) must be nonnegative and nondecreasing, the minimal amount of expediting required to satisfy (4) is

\[
X_j(t) = \sup_{0 \leq s \leq t} \left[ \sum_{k=1}^{K} a_{kj} A_k(s) - P_j(s; N_j) \right]^+.
\] (9)

Therefore, \( X_j(t) \) is nondecreasing with \( A_k(s)_{0 \leq s \leq t} \). Assembling an order before its due date tightens constraint (4), increasing expediting if the necessary components are not already in stock. Therefore, an optimal assembly policy is to assemble orders exactly when they are due: \( A_k^*(t) = D_k(t-L_k) \) is the minimal assembly needed to satisfy constraint (3). Substituting \( A_k^*(t) \) into (9) gives (8). Because \( X_j^*(t) \) is the smallest feasible value of \( X_j(t) \) for all \( t \geq 0 \) and \( j = 1, \ldots, J \), it minimizes the infinite-horizon discounted cost of expediting on every sample path \( \omega \in \Omega \).

The optimal expediting process in (8) is unique, but the optimal assembly sequencing process (7) is not. An alternative optimal sequencing policy is to assemble the order with shortest remaining time until its due date if and only if the necessary components are in stock, and the time until the due date is smaller than \( \min_{k} \{L_k\} \). This sequencing policy is the same as global FIFO if and only if all products have exactly the same target leadtime: \( L_k = L \) for \( k = 1, \ldots, K \).

From Proposition 1, the production and expediting costs for component \( j \) are uniquely determined by the nominal production plan for component \( j \), \( P_j \), and do not depend on the production and expediting of other components. Therefore, problem (2)-(6) is separates into \( J \) single-item inventory control problems, one for each component \( j \):

\[
\min_{N_j} \mathbb{E} \left[ \int_0^\infty e^{-\delta t} c_j(t; N_j) dP_j(t; N_j) + \int_0^\infty e^{-\delta t} x_j dX_j^*(t) \right] \tag{10}
\]

subject to:

\[
X_j^*(t) = \sup_{0 \leq s \leq t} \left[ \sum_{k=1}^{K} a_{kj} D_k(s-L_k) - P_j(s; N_j) \right]^+ \text{ for } t \geq 0 \tag{11}
\]

\( N_j \) is nonanticipating \( \tag{12} \)

Surprisingly, incorporating control over expediting in our model of the assemble-to-order system greatly simplifies the optimal control problem.
Separation does not depend on the assumption of instantaneous assembly. We could simply reduce $L_k$ to account for a deterministic assembly time. The strong and essential assumption is the zero leadtime for expediting. A more detailed analysis (Plambeck and Ward, 2003b) establishes a separation principle for assemble-to-order systems in which the expedite leadtime is small compared to the target leadtime and a high volume of orders arrive within the target leadtime. These assumptions are valid for Dell Computer, where target leadtimes are measured in days or weeks, components can typically be expedited within a few hours and order interarrival times are measured in minutes. (Perman, 2001)

4 Discussion

Inventory management with backordering of demand has been studied intensively, but few papers have been written on inventory management with expediting to meet target leadtimes for customers. In particular, none of the papers in the assemble-to-order literature surveyed by (Song and Zipkin, 2001) incorporate expediting. This must be due to the notorious “intractable nature of dual-source models” (Bradley, 2002) because expediting is common in practice. The contribution of this study is to suggest that, surprisingly, expediting simplifies the analysis of assemble-to-order systems.

Having established separability, we can mine the existing literature on single-item inventory management with expediting or, equivalently, lost sales, to obtain optimal policies for assemble-to-order systems with expediting. For example, suppose that the target leadtime $L_k = 0$ and the order processes $D_k$ are independent Poisson processes for $k = 1, \ldots, K$, and that $a_{kj} \in \{0, 1\}$. Then the demand for each component $j$ is also a Poisson process. Furthermore, as in the illustrative example (1) given above, suppose that the cheap mode of supply for component $j$ is a Poisson process $Z_j$ with rate $\mu_j$ and the unit cost is a constant $c_j < x_j$. At each point in time, one must decide whether or not to idle this production process for $j = 1, \ldots, J$. Then from (Veatch and Wein, 1996, pp. 637-638) a base stock policy is optimal. The optimal policy is to idle the cheap production of component $j$ when the inventory position of component $j$ reaches level $b_j$, that is,

$$N_j(t) = \int_0^t 1\{I_j(s) < b_j\} \, ds$$
where the inventory process is given by

\[ I_j(t) = P_j(t; N_j) + \sup_{0 \leq s \leq t} \left[ \sum_{k=1}^{K} a_{kj} D_k(s - L_k) - P_j(t; N_j) \right] - \sum_{k=1}^{K} a_{kj} D_k(t) \geq b \]

\[ P_j(t; N_j) = Z_j(N_j(t)). \]

The base stock level \( b_n \) is given by an expression analogous to (16) in (Veatch and Wein, 1996). Using (Bertsekas, 1995) these results can be extended to a setting in which the service rate \( \mu_j \) can be changed at times when a component is completed or an order arrives, and the unit cost is a convex increasing function of the service rate. The optimal service rate decreases with the inventory position. In the case that \( L_k > 0 \), the optimal production policy may be contingent upon the number of outstanding orders which require component \( k \), and the time remaining until the due date for each of these outstanding orders. Das (1977) analyzed the single-item system with Poisson order arrivals and production, in which customers will wait for \( L > 0 \) time units and then the sale is lost. He assumed base stock control, and characterized the base stock level that minimizes long run average cost. With a discounted formulation, Plambeck and Ward (2003b) prove that base stock control is near-optimal for high volume systems, and characterize the optimal base stock level.

Relevant results on single-item inventory management with uncapacitated component production are in (Daniels, 1962; Nahmias, 1979; Cohen et al., 1988; Lawson and Porteus, 2000).

Separability from expediting may be highly beneficial in decentralized assemble-to-order systems. For example, Dell owns the Optiplex assembly plant, but its suppliers own the component inventory in warehouses next to the plant, and the component production facilities. Dell provides visibility of outstanding orders and the assembly sequence to suppliers, so that they may optimize component production. (Perman, 2001) Separability greatly simplifies information system and coordination requirements: suppliers do not need to know about other suppliers’ component inventory and production.

Acknowledgements: We thank Li Chen, Paul Zipkin, Stephen Smith, Sunil Kumar, Jeannette Song and J. Michael Harrison for helpful suggestions. This research was supported by the National Science Foundation under grant DMI-0239840.
References


