RC 18971 (82803) 6/18/93
Manufacturing/Mathematics 45 pages

Research Report

Algorithms for the Analysis of Multi-Stage Capacitated Assembly Systems

Pasumarti V. Kamesam

IBM Research Division
T. J. Watson Research Center
Yorktown Heights, NY 10598

Sridhar Tayur

Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburg, PA 15213

IBM Research Division
Almaden • T.J. Watson • Tokyo • Zurich

Algorithms for the Analysis of Multi-Stage, Capacitated Assembly Systems

Pasumarti V. Kamesam
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598

Sridhar Tayur
Graduate School of Industrial Administration
Carnegie Mellon University, Pittsburg, PA 15213.

June 18, 1993

Abstract

A modern manufacturing network, consisting of multiple manufacturing facilities and several external vendors, can be modeled as a multi-stage, capacitated, assembly system. We study one such model assuming centralized control and stochastic end-product demands.

Our analysis and approach is based on a concept of *shortfall*. First, we provide intuition and a clear algorithmic description for computing the optimal base-stock levels for an uncapacitated system. Second, an algorithm for optimal base stock level for a single stage capacitated system is derived. Third, using the proof technique for the single stage and the intuition for the uncapacitated multi-stage systems, we develop a heuristic algorithm for computing base-stock levels for multi-stage capacitated systems. Finally, test cases and computational results are presented for both the uncapacitated and capacitated models and the algorithms are used to analyze the control policies.

Key Words: Capacitated inventory systems, optimal base-stock policies, assembly systems.

1 Introduction

A modern manufacturing network consists of multiple manufacturing facilities and several external vendors. The various components and subassemblies that go into the end product are supplied by the different manufacturing facilities distributed geographically, very often crossing country boundaries. The overall manufacturing cost and flexibility depend on many variables including the uncertainty in demand, the uncertainty in the supply of raw components, capacity limitations, lead times of resupply and transportation costs. In order to improve the overall manufacturing cost and efficiency, it is essential to understand the interrelationships between the cost structure, uncertainty in demand, required service levels and capacity limitations. Models that can describe the above mentioned interrelations are helpful in choosing between alternative control policies. In this paper, we study one such model and provide computational algorithms for it.

Several authors have developed different models that may be employed to analyze the above mentioned problem. We may broadly classify such models into three categories.

1. Finite horizon planning models such as the models in Bitran et al.(1981), Escudero and Kamesam (1992) and others.

2. Infinite horizon inventory control models such as the model and literature survey presented in the rest of this paper.

3. Simulation models that model a given business process and management policies.

Although simulation models play an important role, by themselves they cannot provide the insight necessary to set inventory control policies. It is possible, however, to develop a computational procedure that uses Infinitesimal Perturbation Analysis (IPA) derivative during a simulation to find a good set of parameters for a broad class of policies (see Glasserman and Tayur (1992ab)). However, analytic models are necessary to generate alternative and approximate control policies quickly, which can then be assessed or refined through simulation.

To our knowledge, there is no analytic algorithm (heuristic or exact) that provides solutions to multi-stage capacitated assembly systems with stochastic end-item demand. Motivated by a real application, we study centralized control policies for such a system. Our basic model is a periodic-review, continuous-demand, multi-echelon system with (possibly)

limited production capacity at each stage. Stochastic demands arrive at the lowest stage. Holding and back order costs are linear; there are no fixed ordering costs. The system operates under a base-stock policy. [1] Under a base-stock policy, the operation of each node is determined by a target level of inventory. As demands deplete inventories, each node produces goods to restore inventories to their target levels. When production capacity is limited, it may take several periods of production to offset demand in a single period. Our analysis is based on a concept of *shortfall*. *Shortfall*, at any stage, is defined as the difference between the echelon base stock level and the echelon inventory. A base-stock policy attempts to reduce the shortfall to zero, while never driving it below zero. We note that shortfalls occur for two reasons – non zero lead times and finite capacity.

The contributions of this paper are as follows. First, we define the concept of a shortfall process and intuitively interpret existing results for uncapacitated inventory models and give a clear algorithmic description; this provides a basis for developing algorithms for capacitated systems. Second, we develop a closed form solution for a capacitated single stage inventory system. Third, we provide an algorithm for multi-stage capacitated assembly systems. Finally, we apply these algorithms to several test cases and derive some insight from the computational results.

1.1 Literature Review

Clark and Scarf (1960) developed a periodic review inventory control model for a serial system without setup costs. By using a discounted cost framework, they established that an *order up to* policy at each node is indeed optimal. Federgruen and Zipkin (1984) and Zipkin (1984) extended these results further. Muckstadt et al. (1984) conducted a computational study using the Clark and Scarf model. A continuous review version of the Clark and Scarf model is studied by De Bodt and Graves (1985). Schmidt and Nahmias (1985) present an in depth analysis of an assembly structure with only two inputs, again by using the discounted cost framework. Rosling (1989) showed that under some initial conditions, an assembly system can be reduced to a serial system with modified lead times so that the results of Clark and Scarf (1960) may be applied to this equivalent serial system.

[1] We do not claim that base-stock policies are optimal in the general setting; however, their simplicity makes them attractive. Also, for high-volume products with reasonably long intervals between ordering opportunities, orders are placed every period and (s, S) policies reduce to base-stock policies.

Cohen and Lee (1988) described a model of a production and distribution network in which manufacturing is modelled by a single node. Their work uses the earlier results of Karmarkar(1987) and Zipkin(1986). It also differs from much of the earlier work in an important way in that decentralized control is assumed and the model itself is a framework that combines separate models of production and distribution. Cohen and Moon (1990) present a supply chain planning model that can be used to study production scale and scope economics.

Lee, Billington and Carter (1991) present a model for supply chain management which assumes decentralized control at each node in the manufacturing network, controlled by periodic review order up to inventory policies. Once the service levels are set for each node, the overall relationships between cost and service can be obtained by applying this model. Although capacity considerations are not addressed, they allow for uncertainty in the supplier lead times.

The literature on inventory control systems and production-distribution systems is extensive and hence we limit our review to the work that is closely related to the theme of this paper. Similarly, there is a vast body of literature on single location production-inventory systems (addressing many aspects of interest) that is not reviewed here.

The Clark and Scarf model and many of its extensions, including Rosling (1989) analyze the model within a discounted cost framework. These results are fairly involved and further, the computational procedures are not easy to describe or program. Langenhoff and Zijm(1989) on the other hand analyze the assembly system inventory control problem in an average cost framework. This analysis leads to an exact decomposition of the assembly system into several single location problems. Even this decomposition is not easy to handle, but Van Houtum and Zijm (1990) describe computational approximations that lead to a simplified computational procedure. Our analysis of the multi-echelon production-inventory system is based on the analysis of Rosling (1989) and Langenhoff and Zijm (1989).

Rosling (1989) (extending Clark and Scarf 1960) has shown that under the cost conditions stated above along with the backlogging and some other minor assumptions, an *order up to* (or base stock) policy is optimal for uncapacitated *assembly* systems with stationary demand. It is well known that *order up to* policies are optimal for uncapacitated single stage systems with non-stationary demand (Karlin and Iglehart 1960). The same is true for capacitated single stage systems with stationary demand (Federgruen and Zipkin 1986a);

however, no algorithm was provided by these authors. As mentioned earlier, in section 3 we will describe an algorithm to compute the critical *order up to* number. Results of this type hold for more general cost structures and for some demand distributions (Karlin 1960), and under discounted cost criteria (Rosling 1989, Federgruen and Zipkin 1986b, Karlin and Iglehart 1960, Clark and Scarf 1960).

1.2 Summary of Paper

The organization of this paper is as follows. In section 2, we begin with a clear definition of the problem and assumptions. A brief summary of the relevant results from Rosling (1989) and Langenhoff and Zijm (1989) are presented along with a clear algorithmic description of the computational procedures. In section 3, we extend the results of section 2 to a system with a single *capacitated* node. Section 4 further extends the results of section 3 to capacitated multi-stage systems. The models of sections 2, 3 and 4 are applied to the analysis of several test cases. These computational results are presented in section 5. Section 6 concludes by suggesting some directions for further research.

For easy reference, all the algorithms described in this paper are summarized below.

- `Algorithm 1.` Given a set of base-stock levels, describes how to operate an uncapacitated assembly system.
- `Algorithm 2.` Gives a generic method to compute the optimal order up to levels for uncapacitated serial systems.
- `Algorithm 3.` Computes the optimal base-stock levels for an uncapacitated serial system.
- `Algorithm 4.` Computes *Type-1* and *Type-2* service obtained, given a set of base-stock levels.
- `Algorithm 5.` Approximates any distribution by a weighted erlang, or a hyper-exponential distribution.
- `Algorithm 6.` Computes the optimal base-stock level for a single stage capacitated model.
- `Algorithm 7.` Computes the first two moments of a random variable with distribution $F(x+C)$, where $F(x)$ is known and C is a given constant.

- `Algorithm 8.` Computes the optimal base stock levels for a two stage serial system with limited capacity at both stages.
- `Algorithm 9.` Computes base-stock levels for a capacitated assembly system.

2 Model for the Uncapacitated Assembly System

Consider a periodic review inventory control model for a production system with the following assumptions (Assumption 4 below will be relaxed in a later section). Each node with a predecessor represents a sub-assembly, and nodes with no predecessor represent components.

1. Echelon Inventory is reviewed periodically to place orders if necessary.
2. An assembly system in which each node has a unique successor (which implies a single end product).
3. There is a constant lead time between the nodes, which is an integral multiple of the review period.
4. Each node is assumed to have no capacity restriction.
5. Demands occur only for the end product. These demands are assumed to be stochastic and independent from one period to the next.
6. The demands that cannot be satisfied from the stock on hand are backlogged. There is a linear penalty cost associated with the unfilled demand. In section 3, we will define type-1 service and type-2 service (fill rate) and address the problem of computing them.
7. There are holding costs at each node, and the echelon holding cost at each node is positive. There are no fixed (set-up, review, or order) costs (see below) and the variable purchasing cost is linear [2].

Our goal is to minimize the long-run average cost of operating this system, subject to the penalty cost or a constraint on the fill rate or the type-1 service. In order to review some

[2]The purchasing cost will never enter the analysis because we are dealing with infinite horizon average cost and back orders are allowed.

of the known results from the literature, we will first introduce the following definitions.

Definition 1: The *Inventory position* of any node is defined as (stock on hand) + (stock on order) - (back orders).

Definition 2: The *Echelon Inventory* of a node in a serial or assembly system is defined as all the stock on hand at that node plus the stock in transit to or on hand at any node down stream minus the backlogs at the most down stream location.

Definition 3: The *Echelon Inventory Position* of a node denotes the echelon stock plus all the material on order.

Definition 4: An *order up to policy* with a critical number x is one where an order of size $x - y$ is placed if the (echelon) inventory (stock, or position, depending on the situation) y is such that $(y < x)$ at the time of review. No order is placed if $y \geq x$.

Rosling (1989) has also shown that under some regularity conditions, the optimal inventory control policies of the assembly system are equivalent to those of a pure serial (every node has a unique predecessor) system. So, we adopt the following strategy to develop an overall computational procedure.

- In order to minimize the long run average cost, we can restrict our attention to *order up to* policies at each node in the system.
- If a *type-1 service* or *fill rate* constraint is necessary, we can proceed as follows. Compute the order up to policies for a given penalty cost on back orders. Compute the achieved fill rate. If the fill rate is not satisfactory, adjust the penalty cost and recompute the optimal policies. This process may be repeated for many different values of the penalty cost.
- If the system is capacitated, reduce it to an equivalent uncapacitated system, as described in sections 3 and 4.
- Following Rosling (1989), convert the uncapacitated assembly system to an equivalent uncapacitated serial system.

- Compute the *order up to* policies of the uncapacitated serial system as per Langenhoff and Zijm (1989).

- Use the *order up to* values obtained from the equivalent serial system to operate the assembly system by using `Algorithm 1` described in section 2.1.

In order to state these computational procedures clearly, we will first introduce the following notation.

Notation:

- N: the number of stages (node 1 is the end product).

- $(a)_+ = \max(a, 0)$.

- $P(n)$: the set of nodes that immediately precede node n.

- $S(n)$: the unique successor of node n.

- L_n: the (constant) lead time from n to $S(n)$, as a multiple of the review period.

- $f(x)$: the probability density of the demand in a time period.

- $F(x)$: the distribution function of the demand in a period.

- S_n: the *order up to* level for node n.

- d_t: demand realized in period t.

- $\hat{d}_i$: the cumulative demand during periods $t-1, t-2, \ldots, t-(l_N + l_{N-1} + \ldots + l_{i+1})$.

2.1 Equivalence of assembly system to a serial system

Theorem 1: (Rosling (1989)):
Let $M_n = L_n + M_{S(n)}$, be the cumulative lead time from node n to the end product. Re-label the nodes in the ascending order of M_n; thus, after re-labeling,

$$M_N \geq M_{N-1} \geq \ldots \geq M_1. \tag{1}$$

Let

$$l_k = M_{k+1} - M_k. \tag{2}$$

For the nodes with $l_k = 0$, collapse them into a single node, and add the respective echelon holding costs. (This idea will be crucially used in section 4 to develop a heuristic algorithm for capacitated systems). The assembly system with the above assumptions is equivalent to a serial system where the lead time between nodes $k+1$ and k is l_k.

Figure 1 gives examples of two assembly systems and their equivalent serial systems.

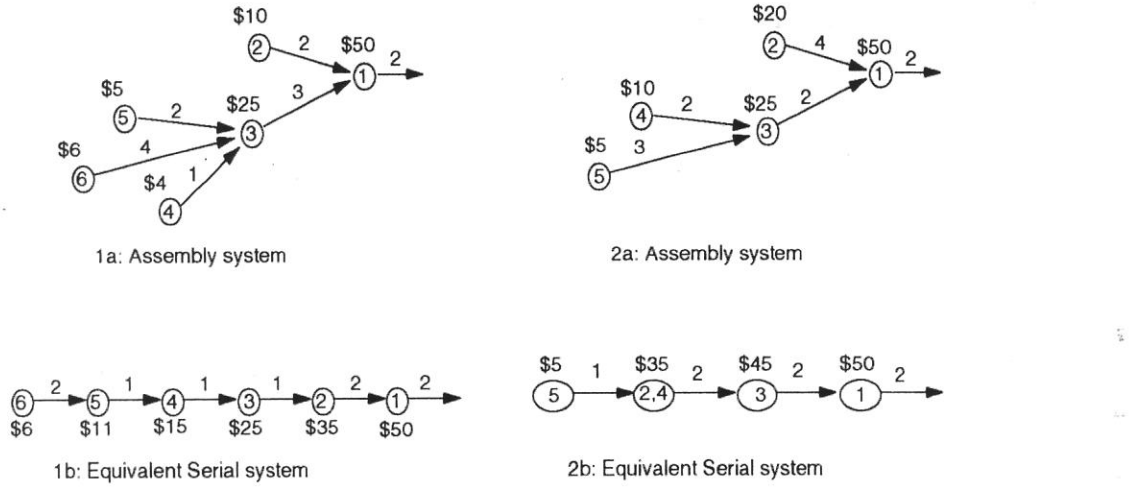


Figure 1: Equivalent Serial and Assembly Systems

The above result is stated without proof, but the proof (see Rosling(1989)) consists of the following steps.

- Showing that a modified *order up to* policy for an assembly system is equivalent to an order up to policy for the equivalent serial system.
- Showing that the cost function for the two systems (the original assembly, and the corresponding serial) are identical.
- The stated equivalence depends on a *long run balance* condition as well.

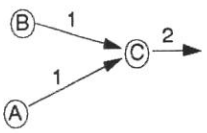


Figure 2:Two Stage Assembly System

The basic intuition in collapsing an assembly system to a serial system is as follows. Consider a two-stage assembly system with three nodes A, B and C, where at time zero there are no units of A, B or C. Let nodes A and B feed into node C, and let $L_A = L_B = 1$ as shown in figure 2. If in any period, there is more of type A than of type B, these will incur holding costs but not help in making more C's; the extra A's could have been ordered later. In fact, the ordering of A must be perfectly synchronized with the ordering of B to minimize holding excess of A. Thus, we can lump A and B as one component, and add up the respective holding costs, and run the system as a two-stage two node serial system.

Next consider the situation where $L_A = 2$ with $L_B = 1$ with order up to levels S_B and S_A. We recognize that in every period, B should input exactly what A had input in the last period and no more. Thus, given the order up to values, in an assembly system, nodes should not blindly raise the echelon inventory to their respective critical numbers; they should also coordinate with other nodes that have the same successor. In a general N node system, only node 1 and the node with the longest lead time should attempt to bring their inventories to the order up to levels blindly; the others need to coordinate as in `Algorithm 1` below.

begin Algorithm 1

- Input to stage N an amount to bring its echelon inventory to S_N. In any period t, this amount will exactly equal the demand in the previous time period, namely , d_{t-1}.

- For stage $N-1$, if $l_{N-1} = 0$, we have $S_N = S_{N-1}$ and we input d_{t-1}. If however, $l_{N-1} > 0$, then we input at stage $N-1$ an amount that brings the echelon inventory at stage $N-1$ to a minimum of S_{N-1} and $S_N - (d_{t-1} + \ldots d_{t-l_{N-1}})$, that is, to the minimum$(S_{N-1}, S_N - \hat{d}_{N-1})$. This is the coordination alluded to earlier.

- Similarly, for any stage $k > 1$, we need to input an amount that brings the echelon inventory to the minimum of $(S_k, S_{k+1} - \hat{d}_{l_k}, \ldots, S_N - \hat{d}_{N-1})$.

- For stage 1, we simply attempt to bring its echelon inventory up to S_1.

end Algorithm 1

The above line of argument is possible because all nodes are uncapacitated, and any amount of material desired, if available, can be input into the system in any period. Thus, at every node the material input is the demand that has already occurred in a previous

period. The pipeline inventories are therefore simply sums of demands that have occurred–all independent and identically distributed–and this fact is crucially used in obtaining the distribution of shortfalls which are then used in our algorithms to compute the order up to values and service levels. In the presence of capacity constraints, however, this simple argument breaks down and needs to be modified. We will construct (in section 3) an equivalent multi-stage uncapacitated system for a capacitated single stage and use the above intuition and procedure.

Next we define the cost functions and state results that make it possible to develop an algorithm for uncapacitated system, based on a level by level decomposition.

2.2 Average cost analysis of a serial system

As stated earlier, let F be the distribution function of end product demand per period and F_l represents the distribution function of the l-period cumulative demand. Since the optimal policies will be computed in terms of *echelon stocks*, we need to account for all the costs in like manner. Hence inventories present at node n or in transit between echelon n and echelon $n-1$ are charged a holding cost at a rate of $\sum_{i=n}^{N} h_i$ per unit, per period. A penalty cost p per unit per period is incurred whenever the demand for end product is not met. Let $L_i(x_i)$ represent the expected echelon holding and penalty costs, where x_i is the stock on-hand at node i. Note that order costs need not be considered because on the average we produce what is demanded (stationary demand and no lost sales) and we are computing the average cost. Then, similar to the familiar news-person formulae, we can write

$$
\begin{aligned}
L_1(x_1) &= \sum_{n=1}^{N} h_n \int_0^{x_1} (x_1 - u) dF(u) + p \int_{x_1}^{\infty} (u - x_1) dF(u) - \sum_{n=2}^{N} h_n x_1 \ \text{ if } x_1 \geq 0 \\
L_1(x_1) &= p \int_0^{\infty} (u - x_1) dF(u) - \sum_{n=2}^{N} h_n x_1 \ \text{ if } x_1 < 0 \\
L_n(x_n) &= h_n x_n \ \ \forall\, n = 2, \ldots, N.
\end{aligned}
$$

The next two results are crucial to the analysis of uncapacitated serial systems.

Theorem 2: (Langenhoff and Zijm(1989)):

Consider a policy which, at the beginning of every period, increases the echelon inventory position of each echelon n to y_n , $(n = 1, 2, \ldots, N)$. Let $D^N(y_1, y_2, \cdots, y_N)$ be the associated average cost (which is defined only on $\{(y_1, y_2, \cdots, y_N) | y_1 \leq y_2 \leq \cdots \leq y_N\}$ because, the *echelon inventory position* at node n cannot be less than the *echelon inventory position* at node $n-1$) [3] Then

$$D^N(y_1, y_2, \cdots, y_N) = C_1(y_1) + \cdots + C_N(y_1, y_2, \cdots, y_N) \tag{3}$$

where

$$\begin{aligned} C_1(y_1) &= \int_0^\infty L_1(y_1 - u_{l_1}) dF_{l_1}(u_{l_1}) \\ C_n(y_1, \cdots, y_n) &= \int_0^\infty L_n(y_n - u_{l_n}) dF_{l_n}(u_{l_n}) \\ &+ \int_{y_n - y_{n-1}}^\infty \Big\{ C_{n-1}(y_1, \cdots, y_{n-2}, y_n - u_{l_n}) - C_{n-1}(y_1, \cdots, y_{n-2}, y_{n-1}) \Big\} dF_{l_n}(u_{l_n}) \\ &\quad \forall n = 2, \ldots, N. \end{aligned}$$

Theorem 3: (Langenhoff and Zijm(1989)):
Define for each $n = 1, 2, \ldots N$,

$$D^n(y_1, y_2, \cdots, y_n) = C_1(y_1) + \cdots + C_n(y_1, y_2, \cdots, y_n). \tag{4}$$

The functions $D^n(y_1, y_2, \cdots, y_n)$ are convex for all $n = 1, 2, \ldots, N$.

As mentioned above, $D^N(y_1, y_2, \cdots, y_N)$, is defined only on $\{(y_1, y_2, \cdots, y_N) | y_1 \leq y_2 \leq \cdots \leq y_N\}$ and hence we can restrict the search to this set. The following is a generic description of an algorithm to compute the optimal *echelon order up to* levels.

- **begin algorithm 2**

  1. set $n \leftarrow 1$. Let $S_1 \leftarrow argmin_{y_1} D^1(y_1)$.
  2. set $n \leftarrow n+1$. if $(n > N)$ stop. Let $(S_1, \cdots, S_{n-1})$ minimize $D^{n-1}(y_1, \cdots, y_{n-1})$. Let $S_n \leftarrow argmin_{y_n} D^n(S_1, \cdots, S_{n-1}, y_n)$.

[3] Note that in our case we do not use the term *inventory position*, but *inventory* because we label stages to include the lead time. To be clear, in our paper l_i is the lead time *of* stage i and not *to* stage i. Our entire analysis is similar to Langenhoff and Zijm(1989) but for this difference in terminology.

3. If $S_n \geq S_{n-1}$, go to 2. Else let $k \geq 1$ be the smallest index such that $S_k > S_n$. Set $S_i \leftarrow S_n, i = k, k+1, \cdots, n-1$. Go to 2.

- **end algorithm 2**

Theorem 4: (Langenhoff and Zijm(1989)):
Algorithm 2 yields the global minimum of $D^N(y_1, y_2, \cdots, y_N)$ and the associated policy is average cost optimal.

Most of the work in `Algorithm 2` is in carrying out the one dimensional minimizations in step 2. Theorem 2 above states that the functions $D^n(y_1, y_2, \cdots, y_n)$ are indeed convex. Hence it would seem that these minimizations should be straight forward. This, however, is not the case. These functions are not easy to compute and further they are defined only recursively (see Theorem 1). Rest of this section is devoted to describing the required computational procedures to carry out these one dimensional minimizations efficiently. The following additional notation is needed.
Given a constant $a \geq 0$ and a random variable X with distribution function F, let $F^a(x)$ be defined as follows:

$$F^a(x) = F(x+a) \ if \ x \geq 0, \tag{5}$$

$$F^a(x) = 0 \ if \ x < 0. \tag{6}$$

Also, let $*$ denote the familiar convolution operator on two distribution functions, say F and G. Then,

$$F * G(x) = \int F(x-y)dG(y). \tag{7}$$

Similarly, we may define

$$F^a * G(x) = \int_0^x F(x+a-y)dG(y). \tag{8}$$

In order to minimize the one dimensional convex functions in `Algorithm 2` above, we can begin by setting their partial derivatives to zero and then attempt to solve the resulting equation. The following theorem gives the expressions for these partial derivatives of the recursively defined cost functions $D^n(y_1, \cdots, y_n)$.

Theorem 5: (Van Houtum and Zijm (1991)):

Let $(S_1, S_2, \cdots, S_{n-1}) :=$ argmin $D^{n-1}(y_1, y_2, \cdots, y_{n-1})$ for each $n \geq 1$. Then

$$\frac{\partial}{\partial y_n} D^n(S_1, \cdots, S_{n-1}, y_n) = J_n^n(y_n) \quad for \quad y_n > S_{n-1}$$

$$\frac{\partial}{\partial y_n} D^n(S_1, \cdots, S_{n-1}, y_n) = J_k^n(y_n) \quad for \quad S_{k-1} < y_n \leq S_k$$
$$\frac{\partial}{\partial y_n} D^n(S_1, \cdots, S_{n-1}, y_n) = J_1^n(y_n) \quad for \quad y_n \leq S_1$$

where,

$$J_1^n(y) = \Big\{-(p + \sum_{j=n+1}^{N} h_j) + (p + \sum_{j=1}^{N} h_j)\Big\}(F_{l_n} * \cdots * F_{l_1+1})(y)$$

and for $2 \leq k \leq n$,

$$J_k^n(y) = \Big\{-(p + \sum_{j=n+1}^{N} h_j) + (p + \sum_{j=1}^{N} h_j)\Big\}$$
$$((\cdots((F_{l_n} * \cdots * F_{l_k})^{y-S_{k-1}} * F_{l_{k-1}})^{S_{k-1}-S_{k-2}} * \cdots F_{l_2})^{S_2-S_1} * F_{l_1+1}(S_1)$$

Theorem 5 is the heart of the optimization routine, and the structure of $J_k^n(y)$ provides us the intuition to deal with more complex cases, like the capacitated system studied in section 3. Recall that at the end of each time period, all stages attempt to restore their inventories to their respective base-stock levels. However, this may not always be possible: in an uncapacitated multi-stage system, this is because of lead times between stages; in capacitated systems, it is because of capacity restriction as well as the lead times between stages. We call the difference between base-stock levels and the echelon inventory in any time period as the *shortfall*. As we note below, expressions like $((\cdots((F_{l_n} * \cdots * F_{l_k})^{y-S_{k-1}} * F_{l_{k-1}})^{S_{k-1}-S_{k-2}} * \cdots F_{l_2})^{S_2-S_1} * F_{l_1+1}(S_1)$ implicitly contain the shortfall distribution.

1. Consider the case $N = 2$. Then, at the end of each time period, the total inventory in the system is S_2 (because of infinite capacity). However, the inventory in stage 1 may be less than S_1, because the demand could not have been fully satisfied from the material available from stage 2. This happens only when the material in the pipeline, say P'_{12}, between stages 2 and 1 exceeds $S_2 - S_1$. Let $P_{12} = (P'_{12} - (S_2 - S_1))^+$ denote this excess. Then, the inventory (or backlog) in stage 1 is $S_1 - P_{12}$; thus the shortfall at stage 1 is P_{12}. The distribution of P_{12} is $(F_{l_2})^{S_2-S_1}$, as what is in the pipeline between stages 2 and 1 is exactly the sum of demands of some consecutive l_2 time periods.

2. Recall that if $S_n < S_k$ for some $k < n$ (but $S_n \geq S_{k-1}$), then we set $S_k = S_{k+1} = \ldots = S_n$ (see `Algorithm 2`); i.e;, we lower the values of $S_k, \ldots, S_{n-1}$ to S_n. This has the interpretation that stages k through n are linked with no *buffer* between the

stages. Thus, in an N stage system, having $S_n = \ldots = S_k$ implies that the costs are such that there is no safety stock (buffer) allocated to each of the stages $k, \ldots, n$, but only as a group.

3. Following the above two observations, the expression $((\cdots((F_{l_n} * \cdots * F_{l_k})^{y-S_{k-1}} * F_{l_{k-1}})^{S_{k-1}-S_{k-2}} * \cdots F_{l_2})^{S_2-S_1} * F_{l_1+1}(S_1)$ denotes the probability that the shortfall in stage 1 is less than or equal to S_1 if stages $1 \leq i \leq k-1$ have order up to levels S_i and stages $k \leq i \leq n$ have an order up to value y.

4. In the equation for $J_1^n(y)$ above, the term $(F_{l_n} * \cdots * F_{l_1+1})(y)$ arises because we have $S_1 = \ldots = S_n$. To be clear, this is the case where all the nodes are going to have just one order up to level, namely S_N. See item 2 above.

5. Thus, in a two stage system,

$$\begin{aligned}(F_{l_2})^{S_2-S_1} * F_{l_1+1}(S_1) &= 1 - \text{Prob(back order)} \\ &= \textit{type 1} \text{ service.}\end{aligned}$$

Similarly, the distribution of backlog is $((F_{l_2})^{S_2-S_1} * F_{l_1+1})^{S_1}(x)$, and so

$$\begin{aligned}\frac{\int_0^\infty (1 - ((F_{l_2})^{S_2-S_1} * F_{l_1+1})^{S_1}(x))dx}{\int_0^\infty (1 - F(x))dx} &= \text{fill rate} \\ &= \text{type 2 service}\end{aligned}$$

Similar expressions for an N stage system are derived in a following subsection.

6. If we could find similar expressions for a capacitated system, then the results of this section will carry over easily; this is done in sections 3 and 4.

7. The cost of operating the system can be computed easily as also the distribution of inventories at various stages. For example, in stage 1 of a two stage system, the inventory in stage 1 is always between zero and S_1, and its distribution is that of $(S_1 - P_{12})^+$. We will go into these details in subsection 2.4.2.

2.3 Algorthmic Description

Now that we have specified how to compute the partial derivatives (theorem 5), we can make `Algorithm 2` more specific. Let the penalty cost on back orders be p. (Recall that if

a type-1 service or a fill rate requirement is specified, we can achieve the specified service level by adjusting the penalty parameter appropriately. In the next section we will describe a computational procedure to compute the fill rate). The following algorithm computes the optimal *echelon order up to* policies $S_1, S_2, \ldots, S_N$.

begin algorithm 3

1. • set $p \leftarrow p_0$.

2. • set $n \leftarrow 1$
   • set $S_1 \tilde{\leftarrow} y_1$, where y_1 solves $F_{l_1+1}(y_1) = \frac{p+\sum_2^N h_j}{p+\sum_1^N h_j}$

3. • set $n \leftarrow n+1$
   • $S_1, S_2, \cdots, S_{n-1}$ are known.
   • set $k \leftarrow n$
   • **do** while $(k \geq 2)$
     – compute $\frac{\partial}{\partial y_n} D^n(S_1, \cdots, S_{k-1}, y_n)|_{S_{k-1}} = J_k^n(S_{k-1})$
     – if $J_k^n(S_{k-1}) > 0$ set $k \leftarrow (k-1)$; else goto 4.
   • **end** while

4. • if $k = 1$ solve for y, $(F_{l_n} * \cdots * F_{l_1+1})(y) = \frac{p+\sum_{n+1}^N h_j}{p+\sum_1^N h_j}$
   • if $k \geq 2$ solve for y

$$((\cdots((F_{l_n} * \cdots * F_{l_k})^{y-S_{k-1}} * F_{l_{k-1}})^{S_{k-1}-S_{k-2}} * \cdots * F_{l_2})^{S_2-S_1}) * F_{l_1+1})(S_1)$$
$$= \frac{p+\sum_{n+1}^N h_j}{p+\sum_1^N h_j}$$

   • set $S_n \leftarrow y$

5. • **do** for j = k,k+1, ...,n-1
     – set $S_j \leftarrow S_n$
   • **end**

6. • if $n < N$ go to step 3 above

end algorithm 3

Note the intuitive pattern of `Algorithm 3`. Given an N stage system, consider the following N sub-systems: first sub-system contains only node 1, the second sub-system contains nodes 1 and 2, ..., and the N^{th} sub-system is the entire system. For sub-system n, we have

$$\text{Probability } \{ \text{ (shortfall in stage 1 + demand during } (l_1 + 1) \text{ time periods)} \leq S_1 \} = \frac{p + \sum_{n+1}^{N} h_j}{p + \sum_{1}^{N} h_j}.$$

Thus, we find (recursively) numbers $S_1, \ldots, S_N$ that satisfy the above relationship. We will see that the same relationship holds in a single stage capacitated system (section 3), and we will exploit it to solve multi-stage capacitated system (section 4).

2.4 Cost, Service Levels and Fill Rate Computation

Although the model stated above uses a penalty cost on back orders, we would like to be able to estimate the service level achieved by the chosen *echelon order up to* levels $(S_1, S_2, \ldots, S_N)$. In the literature, service levels are computed in many different ways. In this paper we show how to estimate service levels; the type-1 service defined as the probability of no back order in a time period at stage 1 and the *fill rate* (type-2 service) defined as

$$\text{fill rate} = \frac{E(backorders)}{E(demand)}. \tag{9}$$

2.4.1 Type-1 Service and Fill Rate(Type-2 Service)

The following algorithm computes the type-1 service obtained for echelon order up to levels $S_1, \ldots, S_N$. Note that G_n below is the distribution of (shortfall + pipeline stock) at node n.

begin Algorithm 4

- $G_N(x) = F_{l_N}(x)$
- $G_{N-1}(x) = (G_N)^{S_N - S_{N-1}} * F_{l_{N-1}}(x)$
- $\vdots$

- $G_n(x) = (G_{n+1})^{S_{n+1}-S_n} * F_{l_n}(x)$
- $\vdots$
- $G_1(x) = (G_2)^{S_2-S_1} * F_{l_1}(x)$
- Type-1 Service $= (G_1 * F)(S_1)$.

end Algorithm 4

We recognize that the backlog distribution is $(G_1 * F)^{S_1}(x)$ and so

$$\text{fill rate} = \frac{\int_0^\infty (1 - (G_1 * F)^{S_1}(x))dx}{\int_0^\infty (1 - F(x))dx}. \tag{10}$$

At this point, it is worth pointing out a few subtleties in the type 1 service and *fill rate* computation discussed above. The main question is, when is a demand considered to be satisfied (or on back order). In the computation above, a demand is considered to be satisfied if the demand is met in the same time period in which the demand occurs; otherwise it is counted as a back order.

One variation that is of interest is as follows. We may consider a demand to be satisfied if it is met within say, k time periods after the demand occurs. Let n' be such that $\sum_1^{n'} l_i \geq k \geq \sum_1^{n'-1} l_i$ and let $k' = \sum_1^{n'} l_i - k$. Then,

$$\text{type 1 service} = (G_{n'} * F_{k'})^{S_{n'}-S_{n'-1}} * F(S_1). \tag{11}$$

A similar analysis leads to an expression for fill rate.

A closer look at algorithms 2 and 3 reveals the following. The optimal *order up to* levels are calculated by using a penalty cost on the back orders, and then the fill rate and the type-1 service is computed as an after the fact. Computational results and further remarks are presented in section 5.

2.4.2 Long Run Expected Average Cost

In order to compute the expected average cost, we need to find the inventories at each stage; which can be computed from the backlog distribution discussed in the previous subsection.

Because the system is uncapacitated, the amount of echelon inventory at stage N is always S_N. Define P_{i+1i} as the amount of material in the pipeline between stages $i+1$ and i

in excess of $S_{i+1}-S_i$ (analogous to P_{12} defined earlier). The echelon inventory (or backlog) in stage $N-1$ is $S_{N-1}-P_{NN-1}$; the distribution of P_{NN-1} is $(F_{l_N})^{S_N-S_{N-1}}(x) = G_N^{S_N-S_{N-1}}(x)$. Thus, the echelon inventory distribution for any stage can be recursively computed as in `Algorithm 4`. In fact, the echelon inventory (or backlog) in stage n has the distribution $S_n - P_{n+1n}$ where P_{n+1n} has a distribution $G_{n+1}^{S_{n+1}-S_n}$.

The average echelon inventory in any stage and the backlog at stage 1 can be computed by algorithms 5 and 6 described in the following subsections. The computation of the total cost then is straight-forward.

2.5 Fitting an Approximate Distribution if the first Two Moments are Known

Algorithm 3 is described in terms of $F(x)$, the distribution function of the one period demand. But in many cases, the only information available to us may be the estimates of mean and variance. In such cases, it is a common practice to use a suitable approximation – usually a weighted Erlang – of the unknown density. `Algorithm 5` in `Appendix 1` gives such an approximation from Tijms(1986).

Note that in `Algorithm 3` we repeatedly encounter random variables of the type $H^a(x)$ where a is a known constant and $H(x)$ itself is a mixture of *erlangs* as output by `Algorithm 5`. We need a procedure to compute the mean and variance of such a random variable. This computation is also required in capacitated systems. Such an algorithm (`Algorithm 7`) is described in `Appendix 1`. The optimal order up to values and the costs can then be computed by using algorithms 5 and 7.

3 Single Node Capacitated Model

The previous section described an algorithm that computes the optimal *order up to* policies for an assembly or a serial system. Each node in the system, however, was assumed to be uncapacitated. Determining what the optimal inventory control policy is and computing such a policy is far more difficult when there are capacity restrictions. In this section we will concentrate on a manufacturing system with a single capacitated node with lead time of zero; thus, in this section, we capture only the effects due to capacity and not non zero

leadtimes [4]. Rest of the assumptions are the same as those stated at the beginning of section 2.

Federgruen and Zipkin (1986ab) investigated a capacitated single stage system under discounted cost and average cost criteria and showed (under a fairly general cost structure) that a base-stock policy is optimal. However, they did not provide any algorithm to compute the critical number or the optimal cost. We provide here an algorithm to compute the optimal policy under the average cost criteria and our approach differs significantly from that in Federgruen and Zipkin(1986ab). Further, we are interested here in bringing the analysis of single-stage capacitated systems closer to uncapacitated systems, so that multi-stage capacitated systems can be analyzed using methods similar to uncapacitated systems.

We begin by recognizing an equivalence between a dam model and the capacitated inventory model operated by a base stock policy. This and the discussions of section 2 motivate us to analyze the inventory problem *not* by the traditional approach that tracks *(inventory on hand - back orders)*, but rather by considering the *shortfall* process. The shortfall in any period here is the *amount on order that has not yet been produced* because of the capacity restriction; this is not to be mistaken for the back orders that a customer faces.

It also turns out that the single-stage capacitated model with an order up to value z and capacity C can be replaced by a specially structured, uncapacitated infinite-stage model with a lead time of one time period between stages k and $k-1$ for $k \geq 2$ operated by a *specially structured* order up to policy of the following type: If stage one has an order up to level of z (based on inventory), then stage 2 has an order up to level (based on echelon inventory) of z+C, and stage 3 has an order up to level of z+2C, and so on (see figure 3). This observation allows us to use the results of previous sections with certain modification.

The equivalence of the inventory problem to the dam model is explained first. This will provide us with the necessary and sufficient conditions for the stability of the system under consideration. An algorithm that leads to an optimal solution and a computational approach are derived in the next two subsections.

[4] The case of non zero leadtimes is straightforward as we allow backlogging.

3.1 The Dam Model

Dam models or storage processes, studied in Prabhu (1965), in discrete time have the following structure. In every period there is some input to the dam (rainfall, for example) that increases the content (in terms of height) of the dam; at the end of the period an amount of water is released out of the dam (for irrigation, say) that is the minimum of the water content and a maximum possible release. Significant analysis has been done on the steady state distribution of the dam content, (random) times between the dam is empty and (random) times to cross different levels. We will shortly show the connection between our problem and a model in dams. We can, therefore, use many of the results from storage processes to analyze the capacitated inventory model.

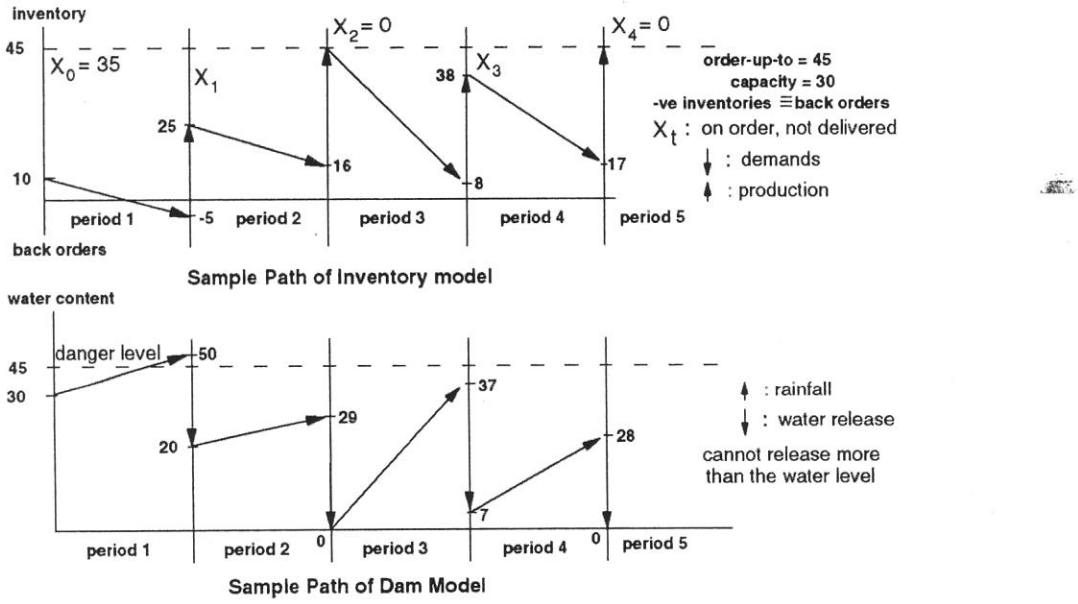


Figure 3 : Equivalence of Inventory and Dam Models

Figure 3 shows the sample path of a typical single stage capacitated inventory system under periodic review that is operated by a base stock policy where excess demand is backlogged. The capacity (C) is 30, the order up to level (z) is 45 and the inventory at time 0 (I_0) is 10. Let $d_1 = 15, d_2 = 9, d_3 = 37, d_4 = 21$ be the demands in the first four periods.

Figure 3 also shows the sample path of a dam that has an infinite height, a release capability of at most C and an initial water level of 35. Let the rainfall in the first four periods be 15, 9, 37, and 21. The dam releases as much water as it can, and if the water level is less than C, the dam goes empty. The equivalence of the two sample paths is straight

Table 1: Comparison of dam and inventory models.

Dam Model	Inventory Model
Maximum release	Capacity
water content	amount not yet produced (shortfall)
empty dam	order up to level achieved
rainfall	demand
danger level crossed	back orders

forward. If Z_n is the content of the dam in period n just after release, then it satisfies

$$Z_n = (Z_{n-1} + d_{n-1} - C)_+ \tag{12}$$

and if X_n is the amount on order in period n that has not yet been produced, it satisfies

$$X_n = (X_{n-1} + d_{n-1} - C)_+. \tag{13}$$

Note that $\{X_n, n = 1, 2, \ldots\}$ is a Markov chain. This motivates us to study the capacitated inventory system in terms of the process X_n, and provide results in terms of the steady state distribution of $X = lim_{n\to\infty} X_n$. Table 1 summarizes the equivalence between the capacitated inventory model and the dam model.

If $K(x)$ is the shortfall distribution in the capacitated inventory model, the main result of this section is that `Algorithm 6` yields the optimal order up to level.

begin Algorithm 6

- **Step 1:** Compute the distribution $K(\cdot)$;
- **Step 2:** Solve for z in $(K * F)(z) = \frac{p}{p+h}$.

end Algorithm 6

Intuitively, we are adding two independent random variables: (1) demand in a period and (2) the amount on order at the beginning of the period that has not yet been produced. Penalty p is incurred if this sum crosses z and is proportional to the excess, and a holding

cost (h) is imposed if the sum is less than z and is proportional to the amount on hand at the end of the period. Note that $K(\cdot)$ does not depend on z.

It remains to compute $K(\cdot)$; for F discrete or Erlang, K is known (Prabhu(1965)). For other distibutions, in particular for a mixture of Erlangs, we give a constructive description of $K(\cdot)$ in terms of $F(\cdot)$ and C. This same construction brings us closer to the analysis of uncapacitated multi-stage systems.

3.2 A Special Infinite Stage Uncapacitated System

We recognize that the effect of a capacity restriction on a single stage is equivalent to a delay in producing the amount depleted by the demands. There is a random time associated with 'recovering' to a state where the on hand inventory at the beginning of a period equals the order up to level. This effect can be mimicked by considering an infinite-stage *uncapacitated* serial system that is operated in the special manner described below. The following additional notation is used in this section.

Notation:

${}^{N}Y_n^j$: random variable corresponding to the amount on order that has not yet arrived at stage j in period n in a system that has a total of N stages.

${}^{N}Y^j$: $lim_{n\to\infty}\, {}^{N}Y_n^j$.

$H_{N-1}(x)$: distribution of ${}^{N}Y^1$.

$D^N(z, z+C, \ldots, z+(N-1)C)$: long run average cost of operating the N-stage system where the echelon inventory order up to levels are $z, z+C, \ldots, z+(N-1)C$.

Note that the *order up to* policies at all stages of the specially constructed infinite stage system are in terms of echelon inventory.

3.2.1 Construction

Consider an infinite-stage uncapacitated serial system with a lead time of one time period between stages k and $k-1$ for $k \geq 2$ operated by a *specially structured* order up to policy of the following type: If stage one has an order up to level of z (based on inventory), then stage 2 has an order up to level (based on echelon inventory) of z+C, and stage 3 has an order up to level of z+2C, and so on. This is shown in Figure 3. The holding cost at stage 1 is h and the penalty cost for back orders at stage 1 is p. There are no holding costs at

other stages. The system operates as follows:

- at the beginning of period n the amount on hand at stage 1 is $(z -^{\infty} Y_n^1)$ while the echelon inventory at stage j for $j \geq 2$ is $z + (j-1)C -^{\infty} Y_n^j$;

- a demand d_n occurs at stage 1 causing the echelon inventory at all stages to drop by d_n;

- supply from stage $j+1$ to stage j $(j \geq 1)$ is made to bring the echelon inventory level of stage j as close to $z + (j-1)C$ as possible.

It is important to notice that, in any time period n, stage $j+1$ can supply j (for $j \geq 1$) a maximum of C in this infinite-stage system, not because of any capacity restriction, but rather because the on hand inventory at stage $j+1$ for $j \geq 1$ is exactly C at all times.

Proposition 1 *In any time period, the shortfall at stage 1 in the infinite-stage uncapacitated system operated by the above policy is equal to the shortfall in a capacitated single-stage system operated by the base stock policy.*

Proof At time zero, start the infinite-stage system with all the stages at their order up to levels. If ${}^{\infty}Y_n^j$ is the amount on order that is yet to arrive at stage j at the beginning of period n, it satisfies the recurrence

$$ {}^{\infty}Y_n^j = ({}^{\infty}Y_{n-1}^j + d_{n-1} -^{\infty} S_{n-1}^{j+1})_+ \tag{14} $$

where ${}^{\infty}S_{n-1}^{j+1}$ is the maximum possible stock that level $j+1$ can supply level j in period $(n-1)$. Since, the system is operated in the special way described above, we have for every fixed t, ${}^{\infty}Y_t^j = {}^{\infty}Y_t^l \ \forall j, l$. Therefore,

$$ {}^{\infty}S_{n-1}^{j+1} = (z + jC -^{\infty} Y_{n-1}^{j+1}) - (z + (j-1)C -^{\infty} Y_{n-1}^j) = C \tag{15} $$

In particular, ${}^{\infty}Y_n^1 = ({}^{\infty}Y_{n-1}^1 + d_{n-1} - C)_+$ as desired.

Q.E.D.

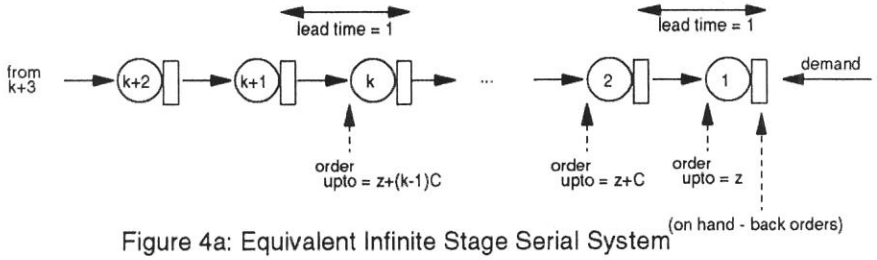


Figure 4a: Equivalent Infinite Stage Serial System

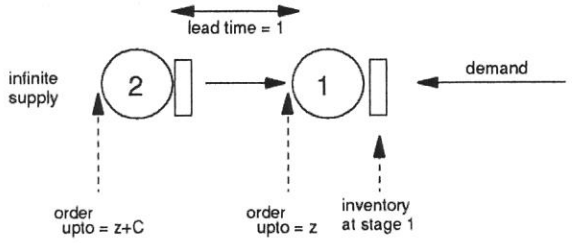


Figure 4b: Truncated System with N = 2

Figure 4: Specially Constructed Uncapacitated serial systems

We now analyze the specially constructed uncapacitated infinite stage system by considering its finite truncations, with say, N stages. Figure 4 shows the case when N=2. Recall that the holding cost in stages $k \geq 2$ is zero. Let $D^N(z, z+C, \ldots, z+(N-1)C)$ represent the long run average cost of operating a N-stage truncated system with the specially structured policy. Also, $N = 1$ is equivalent to an uncapacitated single stage system, and the dummy stages add on additional 'capacity' in a monotone manner.

3.2.2 The ShortFall Distribution

We can compute the steady state distribution of ${}^{N}Y^{1}$ as follows.

$$\begin{aligned}
{}^{N}Y^{N} &= 0 \\
P({}^{N}Y^{N-1} \leq x) &= P(demand \leq C + x) \\
&= F(C + x) \\
&= F^{C}(x) \\
P({}^{N}Y^{N-2} \leq x) &= P(demand \leq (C + x - {}^{N}Y^{N-1})) \\
&= P(demand + {}^{N}Y^{N-1} \leq (C + x)) \\
&= (F^{C} * F)^{C}(x)
\end{aligned}$$

$$\vdots$$

$$\begin{aligned} P({}^{N}Y^{1} \leq x) &= (\ldots(F^{C} * K)^{C} * F)^{C} \ldots F)^{C}(x) \\ &= H_{N-1}(x) \end{aligned}$$

It is clear that for an $N+1$ truncation we simply replace $H_{N-1}(x)$ with $H_N(x)$. In general, $H_N(x) = (H_{N-1} * F)^C(x)$. Thus, ${}^{N}Y^{1}$ are monotone in N as previously stated. This also implies that K can be approximated by $H_N(x)$ for sufficiently large N. The computational work to do so is similar to those needed to compute S_k in `Algorithm 3` (step 4 for $k \geq 2$). As stated earlier, this is accomplished for the case when F is a mixture of Erlangs by `Algorithm 7`.

3.2.3 Cost Function

To obtain the cost function for the truncated system, let $G_{N-1}(u)$ be the distribution of (on hand inventory - back orders)) at stage 1 in a N-truncated serial system. The cost function is simply the expected value of the standard news person problem, the expectation taken over the amount of stock at stage 1. We assume here that $F(\cdot)$ is continuous.

$$\begin{aligned} D^{N}(z, z+C, \ldots, z+(N-1)C) = \quad & \int_0^z (h \int_0^u (u-x) dF(x) \\ & + p \int_u^\infty (x-u) dK(x)) dG_{N-1}(u) \\ & + \int_{-\infty}^0 (p \int_0^\infty (x-u) dF(x)) dG_{N-1}(u) \end{aligned}$$

Making a change of variables, and rewriting the cost function in terms of ${}^{N}Y^{1}$ allows us to use the monotone convergence theorem. Thus,

$$\begin{aligned} D^{N}(z, z+C, \ldots, z+(N-1)C) = \quad & \int_{0+}^z (h \int_0^{z-u} (z-u-x) dF(x) \\ & + p \int_{z-u}^\infty (x-(z-u)) dK(x)) dH_{N-1}(u) \\ & + \int_z^\infty + H_{N-1}(0)(h \int_0^z (z-x) dF(x) \\ & + p \int_z^\infty (x-z) dK(x)) \end{aligned}$$

By definition, ${}^{N}Y^{1} \geq 0$ a.s. for all N. If a limit exists, it must satisfy $H_\infty(x) = (H_\infty * F)^C(x)$ and $H_\infty(\infty) = 1$. If the expected demand in a period is less than C, the limit

does exist because the dam model is stable under this condition. The uniqueness follows from Prabhu (1965). That the cost functions converge to the true cost function now follows from the *monotone convergence theorem* (Durett (1991)). To prevent the possibility of some mass "disappearing" to infinity, and hence have a cost of infinity, we can enforce that the convergence ${}^{N}Y^{1} \uparrow {}^{\infty}Y^{1}$ be *tight*, namely that there exists a function $g(x) \geq 0$, $g(x) \to \infty$ as $|x| \to \infty$ so that $limsup_{n\to\infty} \int g(x)dH_n(x) < \infty$ (or simply assume that the second moment of ${}^{\infty}Y^{1}$ is finite).

Thus, the cost of operating a capacitated system with an order up to policy with critical number z is given by

$$\begin{aligned} D^{\infty}(z, z+C, \ldots,) = \quad & \int_{0+}^{z}(h\int_{0}^{z-u}(z-u-x)dF(x) \\ & +p\int_{z-u}^{\infty}(x-(z-u))dF(x))dH_{\infty}(u) \\ & +\int_{z}^{\infty}(p\int_{0}^{\infty}(x-(z-u))dF(x))dH_{\infty}(u) \\ & +H_{\infty}(0)(h\int_{0}^{z}(z-x). \end{aligned}$$

That $D^{\infty}(z, z+C, \ldots)$ is a convex function of z can be verified easily. This will help find the global optimum. Differentiating once, we have

$$\begin{aligned} \int_{0+}^{z}((h+p)F(z-u)-p)dH_{\infty}(u) \quad & +\int_{z}^{\infty} p\int_{0}^{\infty} dF(x)dH_{\infty} \quad +H_{\infty}(0)(F(z)(p+h)-p) \\ = \quad & (p+h)(H_{\infty} * F)(z) - p; \end{aligned}$$

the second derivative is positive as $(H_{\infty} * K)(z)$ is a distribution function. Thus, to compute the optimal value of z, we need to solve $(p+h)(H_{\infty} * K)(z) - p = 0$. Note that if z_N^* is the optimal solution to the N truncated problem (where an order up to policy is optimal, as can be verified easily), we have $z_1^* \leq \ldots \leq z_N^* \ldots \leq z_{\infty}^* < \infty$, implying that due to capacity restrictions, the base stock has increased. A straight forward argument will also show that if $C_1 \geq C_2$ where $C_i, i = 1, 2$ are capacities of two systems facing the same demand and cost structure, system 1 will have a lower (no greater) critical number. Note that $K(\cdot) = H_{\infty}(\cdot)$.

3.3 Computing Order up to policies for the Single Stage System

In order to compute the optimal policy with algorithm 6, we first need to compute the shortfall distribution $K(.)$. In some cases, it is possible to get a closed form expression for

$K(.)$.

Example 1: *If the demand is exponentially distributed with mean rate* λ *(*$F(u) = 1 - e^{-\lambda u}$, $u \geq 0$*) and the capacity is* C*, then the shortfall distribution is given by*

$$K(x) = 1 - e^{-\theta(x+C)} \ (x \geq 0, \ \lambda C > 1) \tag{16}$$

where θ *is the largest positive root of the equation*

$$\theta = \lambda - \lambda e^{-(\theta+C)}. \tag{17}$$

Thus, the optimal order up to policy has a critical number z^**, and is obtained by solving*

$$(K * F)(z^*) = \frac{p}{p+h} \tag{18}$$

In general, we cannot hope to find $K(.)$ exactly and need to find an approximation. We expect that for some reasonable n, $H_n(\cdot)$ is approximately $K(\cdot)$. Suppose we had a procedure that provided us with the first two moments of $F^C(\cdot)$, where $F(\cdot)$ is a weighted erlang as output by `Algorithm 5`. Then we can approximate $F^C(\cdot)$ by weighted erlangs using `Algorithm 5`, and recursively find $H_n(\cdot)$. In fact, the same procedure can be used to find moments of $F_{l_k}^{S_k - S_{k-1}}$ as required in the algorithms for uncapacitated systems. `Algorithm 7` in `Appendix 1` provides us with the first two moments of F^C if F is weighted erlang.

4 Capacitated Model for a Multi-Stage System

This section will first present the extensions of section 3 to serial systems; the extension to assembly systems follows naturally. Let $C^i, i = 1, \ldots N$, be the capacities for the N stages. These are all *real* stages. Let $z_i, i = 1, \ldots, N$ be the echelon order up to levels. As before, we reserve S_i for the optimal values of the z_i.

4.1 Preliminary Observations

To orient oneself, first consider a system which has $C^N < \infty$ and all other $C^i = \infty$, as shown in figure 9 below. We can now replace the Nth node by an infinite series of dummy nodes, all uncapacitated, as in section 3. The results of sections 2 and 3 now imply that this

multi-stage system can be solved by methods of section 2 with one modification: replace $G_N(\cdot) = F_{l_N}^N(\cdot)$ with $G_N(\cdot) = K_N * F_{l_N}^N(\cdot)$. Note that $K_N(\cdot)$ is the unique distribution function that solves

$$(K_N * F)^{C^N}(x) = K_N(x) \tag{19}$$

and represents the shortfall at stage N due to finite capacity. To be clear, `Algorithm 3` is unchanged but for the last computation (i.e of S_N) where we include K_N in the shortfall distribution.

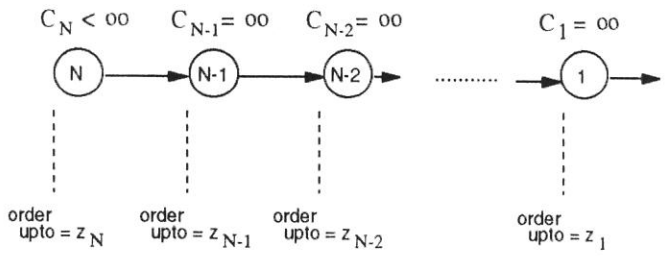
Unfortunately, other capacitated cases are not amenable to such straightforward analysis. However, some conclusions about optimal base-stock levels can still be drawn in special cases of multi-stage capacitated systems. The next result shows that if capacities increase with the stage index, then it is never optimal to hold more safety stock between each pair of stages than the downstream stage can use in a single period:

Proposition 2 Suppose $C^1 \leq C^2 \leq \cdots \leq C^N$. If $S_{i+1} - S_i \geq C^i$, $\forall i = 1, \ldots, N-1$, then reducing each $S_{i+1} - S_i$ to C^i, leaving S_1 fixed, decreases costs.

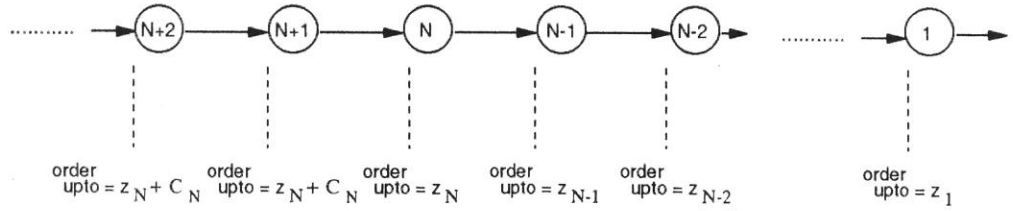
Suppose now that the capacity levels are subject to control, possibly within a range of values. For example, it might be possible to physically re-allocate capacity from one stage to another, or else a stage may modify its policy, choosing a maximum production level less than its capacity. This has the same effect as changing some C^i. The following result gives a necessary condition for a set of optimal capacity levels.

Proposition 3 An optimal $(C^1, \ldots, C^N)$ satisfies $C^{i+1} \leq C^i$, $i = 1, \ldots, N$. More precisely, given any set of $(C^1, \ldots, C^N)$, if $C^{i+1} \geq C^i$ then replacing C^{i+1} with C^i does not increase costs.

The above propositions are stated here without proofs but the proofs can be found in Glasserman and Tayur(1992b).



Serail System where node N is capacitated



Serail System where all nodes are uncapacitated

Figure 9: Equivalent Capacitated and Uncapacitated Systems

4.2 A Two-Stage Serial System with Limited Capacity in Each Stage

In general, the capacities and the order up to values interact in an intricate manner. We consider a two-stage system, where both the stages are capacitated, with a lead time of one between the two stages to illustrate the complexity (figure 10). The extension to an N stage capacitated system follows directly. As in the case of a single stage capacitated system, each of the two nodes here can be replaced by an infinite sequence of dummy nodes, all uncapacitated, whose order up to levels differ from their downstream stage by exactly C^i.

This leads to an *assembly* system, infinitely long, but *uncapacitated.* Thus, we have a sequence of order up to values along one branch of the assembly: $z_1 + C^1, \ldots, z_1 + (k-1)C^1, \ldots$, and a sequence of order up to values: $z_2, \ldots, z_2 + C^2, \ldots, z_2 + (l-1)C^2 \ldots$ along the the other branch. These branches meet at stage 1 which has order up to level z_1, as shown in figure 10b.

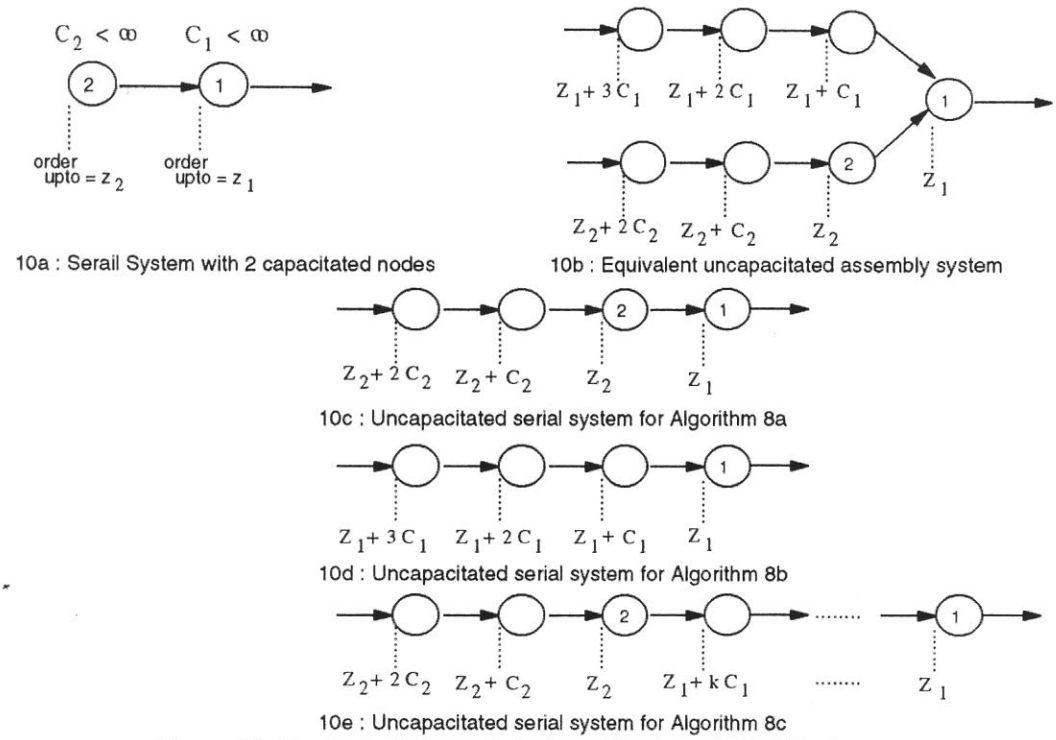


Figure 10: Equivalent Capacitated and Uncapacitated Systems

Now, following the arguments of `Algorithm 1`, the assembly system of figure 10b can be converted to a serial system. To do this, however, we first need the values for z_1 and z_2. We, therefore, start with some initial guess values for z_1 and z_2. These initial values of z_1 and z_2 are then modified iteratively. In the process of iteratively improving the values of z_1 and z_2, we need to contend with several distinct cases. In each case, the uncapacitated assembly (figure 10b) is converted to a different serial system (figure 10c, 10d or 10e). The following cases arise.

$$\mathbf{A}: \; C^1 \geq C^2 \quad \text{and} \quad z_1 + C^1 \geq z_2$$
$$\mathbf{B}: \; C^1 = C^2 \quad \text{and} \quad z_1 + C^1 < z_2$$
$$\mathbf{C}: \; C^1 > C^2 \quad \text{and} \quad z_1 + C^1 < z_2$$
$$\mathbf{D}: \; C^1 < C^2$$

i.e; the first three cases correspond to $C^1 \geq C^2$. Algorithm 8 describes how to find S_1 and S_2 in each case.

begin algorithm 8

1. • start with initial guess values z_1 and z_2.

2. • **A** : $C^1 \geq C^2$ **and** $z_1 + C^1 \geq z_2$

   – In this case, one sequence of order up to values will always dominate the other. We need to consider the sequence that has the lower values. This implies that we need to consider a serial system with order up to values: $z_1, z_2, z_2 + C^2, z_2 + 2C^2 \ldots$, as shown in figure 10c. This case can be solved by the algorithms described in sections 2 and 3.

   begin algorithm 8a

$$\begin{aligned} F_{l_1+1}(S_1) &= \frac{p + h_2}{p + h_1 + h_2} \\ (K_2 * F)^{S_2 - S_1} * F_{l_1+1}(S_1) &= \frac{p}{p + h_1 + h_2} \end{aligned}$$

   end algorithm 8a

   Note that $K_2(\cdot)$ satisfies $(K_2 * F)^{C^2}(x) = K_2(x)$.

   – if $S_2 \leq S_1 + C^1$, we are done; return.

   – if $S_2 > S_1 + C^1$, S_1 and S_2 are the new guess values of z_1 and z_2. go to step 3 (case **B**) or step 4 (case **C**) as appropriate.

3. • **B** : $C^1 = C^2$ **and** $z_1 + C^1 < z_2$

   – This case can be solved by considering a serial system whose order up to values have the the sequence: $z_1, z_1 + C^1, z_1 + 2C^1 \ldots$ (see figure 10d), which is equivalent to a single stage capacitated system. So `Algorithm 5` can be used with $C = C^1$ as the capacity. We have:

   begin algorithm 8b

$$(K_1 * F)^{C^1} * F_{l_1+1}(S_1) = \frac{p}{p + h_1 + h_2}$$

   end algorithm 8b

   $K_1(\cdot) = (K_1 * F)^{C^1}(\cdot)$.

   – set $S_2 = S_1 + C^1$; return.

4. • **C** : $C^1 > C^2$ **and** $z_1 + C^1 < z_2$

   – with the initial values z_1 and z_2 (from `Algorithm 8a` or otherwise), find k such that

$$z_1 + (k-1)C^1 \leq z_2 + (k-2)C^2 \text{ and } z_1 + kC^1 \geq z_2 + (k-1)C^2. \quad (20)$$

- consider a serial system with the sequence $z_1, z_1 + C^1, \ldots z_1 + (k-1)C^1, z_2 + (k-1)C^2, z_2 + (k)C^2, \ldots$ (see figure 8e). This leads to the following algorithm:
  begin algorithm 8c

$$F_{l_1+1}(S_1) = \frac{p + h_2}{p + h_1 + h_2}$$

$$(\ldots((K_2 * F)^{S_2 - S_1 - (k-1)(C^2 - C^1)} * F)^{C^1} \ldots F)^{C^1} * F_{l_1+1}(S_1) = \frac{p}{p + h_1 + h_2}$$

  end algorithm 8c

  $K_2(\cdot)$ satisfies $(K_2 * F)^{C^2}(x) = K_2(x)$. F^{c^1} occurs $(k-1)$ times in the second equation of `Algorithm 8c`.

- if the values S_1 and S_2 satisfy equation (20), we are done; return.
- else find a new value of k that satisfies equation (20) for the new values of S_1 and S_2 and rerun `Algorithm 8c`. Repeat until the assumption of k is consistent with S_1 and S_2; return.

5. • **D** : $C^1 < C^2$
   - The method of analysis here is similar to the case $C^1 \geq C^2$. Note that Proposition 3 should be appropriately used here; in fact, we may reduce C^2 to equal C^1 and use cases **A** or **B** above. We will skip the details.

end algorithm 8

A numerical example illustrating the algorithm is presented in section 5.3 below.

4.3 The General Capacitated Assembly System

A similar approach to the one above leads to the algorithm for an N stage capacitated assembly system.

begin algorithm 9

1. First convert the N stage capacitated assembly problem to an uncapacitated assembly problem. This is done by replacing all capacitated nodes by an infinite sequence of uncapacitated nodes.

2. Guess values for $z_1, \ldots, z_N$.

3. Next, following `Algorithm 1`, create a serial system. Solve the resulting serial problem that is uncapacitated but for the highest node. If the optimal solution obtained satisfies the same *serial structure* as obtained by the guess of $z_1, \ldots, z_N$, go to step 4. If not, use the optimal values obtained as a guess for $z_1, \ldots, z_N$ and go to step 3.

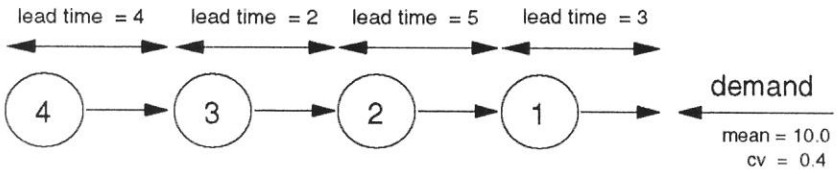
4. Output the solution obtained as the required order up to values.

end algorithm 9

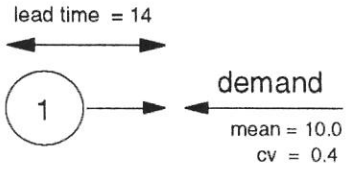
5 Computational Results

We have developed a software implementation of all the algorithms described in the previous sections. These algorithms were then applied to the inventory analysis of several test problems. Computational results with these test problems are summarized in this section.

5.1 Uncapacitated Serial system: Test Case 1



5a : Serail Uncapacitated system



5b : Single Node with Lead Time = Sum of Lead Times in 5a.

Figure 5: Uncapacitated Test Case 1

Test case 1 (see figure 5) is a serial assembly system. Fixing the demand distribution parameters as shown in figure 5 and setting the fill rate requirement to 0.95, we have computed the echelon order up to levels by varying the Echelon costs, as shown in table 3. Although echelon inventory levels did not vary as much as we expected, it is interesting to note that as the *added cost* of down stream stages decreased, more of the inventory buffer

4	3	2	1	4	3	2	1	
Echelon Costs				Echelon *Order up to*				Exp. Cost
30	100	200	330	192.0	140.0	120	62.4	34518.1
65	165	265	330	186.5	140.0	123	68.6	41343.6
80	160	245	330	185.6	142.0	124	66.5	40914.8
150	250	300	330	182.0	144.6	131	75.0	48950.6

Table 2: Sensitivity of Echelon Inventories to Echelon Costs

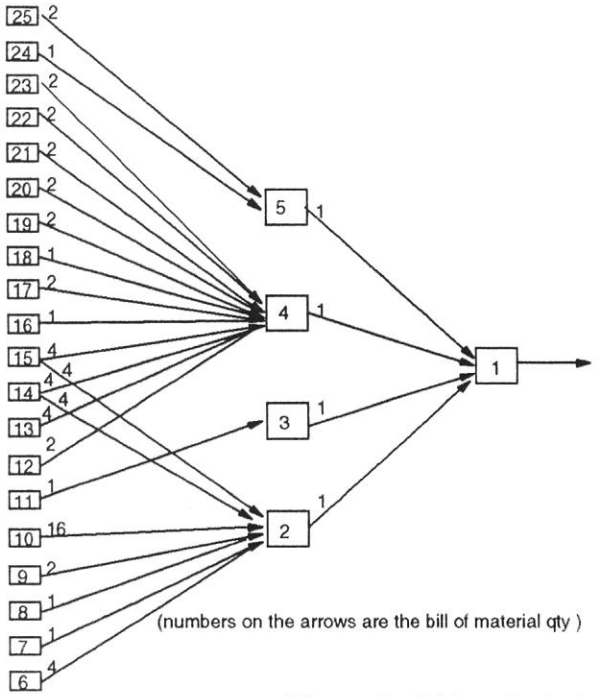
in the system is moved to the down stream stages. Although the echelon inventory levels did not very too sharply, there is a significant change in the expected cost of operating the system as the *added cost* of the down stream stages decreased and the *added cost* of the upstream stages increased.

Next, replace the serial system in figure 5a by a system with a single node, but with the same total lead time as shown in figure 5b. Note that this has the effect of forcing all safety stock to be held in the form of the end item. With an echelon cost of $ 330 and fill rate = 0.95, (same as the echelon cost and fill rate for the end item in the serial system 5*a*. we find that the echelon order up to level = 167.2. We note that this is smaller than the echelon up to level of stage 4 for all the test cases shown in table 3.

The computational results thus far, give us an insight into the sensitivity of the echelon inventory levels to changes in echelon costs. In the next test case, we test the sensitivity of the inventory cost to required service levels and demand variability.

5.2 Uncapacitated Assembly system: Test Case 2

Test Case 2 describes a representative part of a real manufacturing network that produces complex electronic assemblies. The manufacturing network and the problem data are as shown in figure 6.



Part No.	Part Type	Mfg. location	Lead Time (weeks)
1	end item	1	2
2	card asm.	2	4
3	card asm.	2	5
4	card asm.	2	4
5	card asm.	2	4
6	module	3	10
7	card	4	3
8	module	5	9
9	module	5	10
10	module	5	15
11	card	4	3
12	module	5	11
13	module	3	10
14	module	3	10
15	module	3	10
16	module	5	9
17	module	3	11
18	card	4	3
19	module	5	10
20	module	5	10
21	module	5	14
22	module	5	13
23	module	5	13
24	module	3	12
25	module	3	12

Figure 6 : A Manufacturing Network

In this manufacturing system, semiconductor chip modules and electronic cards are assembled together to produce four subassemblies, which together with other miscellaneous parts (not shown here) are assembled to make the final assembly (the end product). The manufacturing in this case is physically distributed across 5 different manufacturing plants. The lead times shown in figure 6 are adjusted to account for the transit times between locations as well. It is to be noted that the lead times for the semiconductor modules are very long.

We made several runs with algorithm 2 to compute the optimal echelon *order up to* levels and the related costs for a specified fill rate. The cost/service trade-off curves at various levels of demand variability were obtained and the results are summarized in figure 7. The following comments are in order.

- The manufacturing network described in figure 6 is not an assembly network (note that part numbers 14 and 15 do not have a unique successor). We replicated the nodes 14 and 15 to overcome this difficulty.

- The unit costs of all the parts have to be appropriately scaled to account for the bill of material quantities shown in figure 6. This problem will be more complicated if the

lead times are not deterministic.

- The costs shown in figure 7 are the inventory asset value and they *do not* include the back order penalty costs.

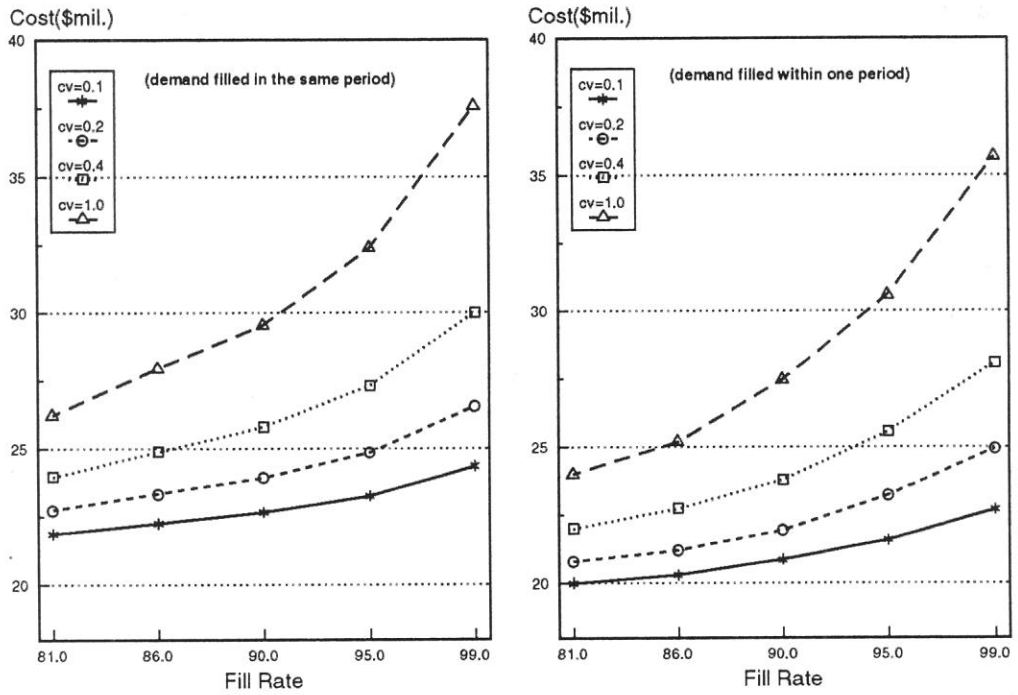


Figure 7 : Cost-Service Tradeoff Curves

We can make a number of remarks by comparing the different curves presented in Figure 7.

If the coefficient of variation is low, then the inventory cost curves are relatively flat as the fill rate increases from 80% to 95% . This is partly explained by the fact that the demand variance and hence the safety stock is low. But the main reason for these relatively flat curves is the fact that the resupply lead times in figure 6 are very high and hence the pipeline stock dominates the total inventory whenever the demand variability is low. This process, however, is reversed if the demand variability is high as evidenced by the case where coefficient of variation is 1.0. In all cases, the slope of the curve increased sharply for fill rates higher than 0.95. This is to be expected.

The cost/service trade off curves were computed by measuring the fill rate for demand filled with in the same period and demand filled within one time period. These fill rates were computed as explained in section 2.4. The trade-off curves for demand filled within

one period are systematically below the corresponding curves when demand is to be filled in the same period. The shape of the curves curves, however, remained the same. This phenomenon deserves an explanation. Recall that algorithm 3 works with a penalty cost and not with a fill rate constraint. The fill rates shown in figure 7 were obtained by the algorithms described in section 2.4. When we change the fill rate definition from immediate fill to fill within one time period, we expected to see some of the safety stock shift from the end item to the subassembly stage. This, however, does not happen with the algorithms described in section 2.

Figure 8 shows a further break down of the total inventories. All the inventories are grouped into inventories of end item (1), sub assemblies (part numbers 2,3,4 and 5), cards (part numbers 7, 11 and 18) and modules (all the rest). It is worth noting that the safety stocks of the long lead time parts (the modules) increased fairly slowly as the demand variability increased. This can be explained by the stationarity of the demand distribution. Also, at high levels of demand variability, most of the safety stock is in the form of end items and sub assemblies.

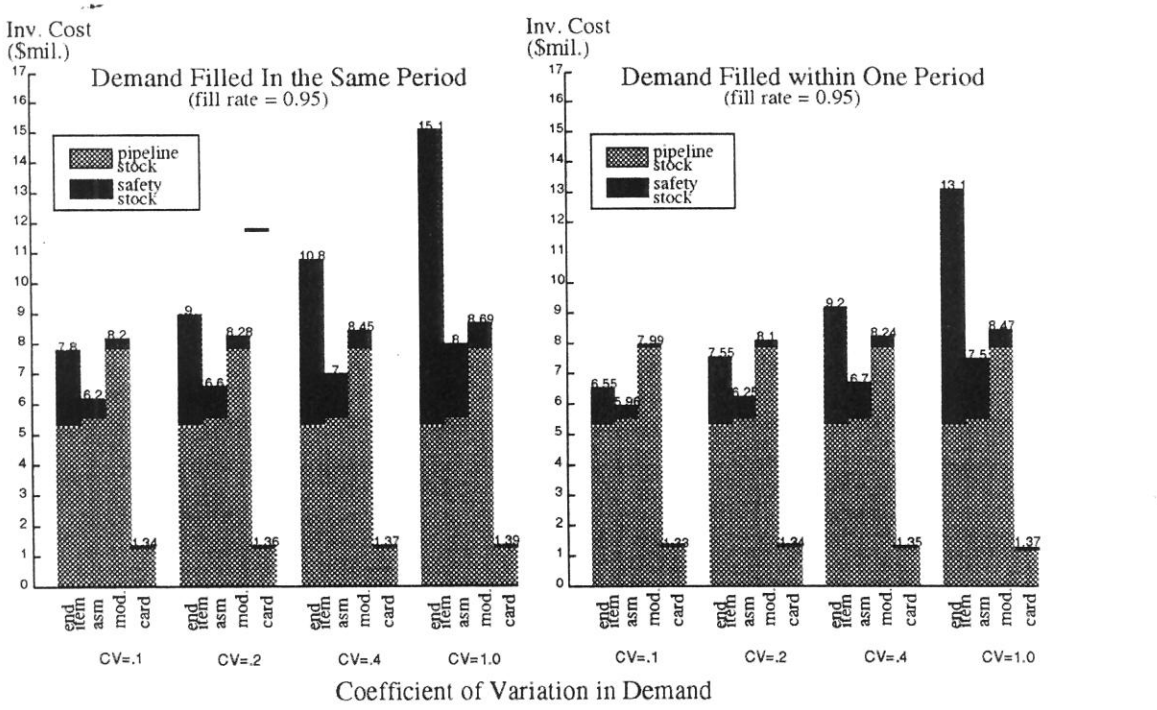


Figure 8 : Inventory Profile Analysis for the Manufacturing Network in Figure 6

Capacity	55	60	65	70	75	80	85
θ	0.0035	0.0063	0.0085	0.010	0.012	0.013	0.014
$\hat{\mu}$	233.72	109.36	68.20	47.84	35.79	27.88	22.33
$\hat{\sigma^2}$	11685.4	5468.1	3410.06	2392.0	1789.4	1394.1	1116.9

Table 3: Roots (θ) of equation (22); Means ($\hat{\mu}$) and Variances $\hat{\sigma^2}$ of shortfall when rate = 0.02.

serv. level	capacity	55	60	65	70	75	80	85
0.8								
	S_1	437	287	235	209	195	195	195
	S_2	492	347	300	279	267	257	249
	Cost	7001	5345	4749	4486	4316	4204	4124
0.9								
	S_1	519	354	295	268	252	242	242
	S_2	574	414	360	338	327	321	312
	Cost	8634	6664	5943	5653	5492	5387	5297

Table 4: Order up to levels and Costs for a two-stage system.

5.3 Systems with Limited Capacity

Computational results for a two-stage system with equal capacities and exponential demand with rate 0.02 (λ) are presented below. The holding costs are $h_1 = 20$ and $h_2 = 10$. Rather than have a penalty cost, we placed a type-1 service requirement of 0.8 and 0.9 respectively. To find the equivalent penalty cost p, we simply use type-1 service $= \frac{p}{p+h_1+h_2}$. θ is the root of the equation in example 1 of section 3. In table 3, $\hat{\mu}$ and $\hat{\sigma^2}$ represent the mean and variance of the shortfall distribution. Table 4 shows the order up to values as we change capacities as well as the type-1 service levels.

The basic observations, as expected, are the following.

1. The optimal costs drop off sharply first as capacities increase, and then level off.

2. The order up to values follow a similar pattern as the optimal costs.

3. Increasing service level increases costs and order up to levels.

The following relationship between capacity and order up to levels can be seen in the computational results obtained. Note that in this example, because $C^1 = C^2$, we have $K_1() = K_2()$. Thus, we would first use `Algorithm 8a` to find S_1 and S_2. If $S_2 < S_1 + C^1$, we are done. If $S_2 > S_1 + C^1$, we would use `Algorithm 8b` to find S_1 and then simply set $S_2 = S_1 + C^1$. Because $C^1 = C^2$, we do not use `Algorithm 8c` here.

1. At low capacities, the difference between the order up to levels $(S_2 - S_1)$ is exactly the capacity. This means that the holding cost in stage 2 is such that if capacity at stage one was larger, we would keep more material between stages 2 and 1. In this range of capacities, therefore, we would use `Algorithm 8b`.

2. As capacity increases, $S_2 - S_1$ drops below C^1. Here, we use `Algorithm 8a`.

3. As the service level is increased, the relationship $S_2 - S_1 = C^1$ continues to hold for higher values of C^1.

It can be easily verified that if h_2 is reduced, keeping other parameters the same, then the relationship $S_2 - S_1 = C^1$ will continue to hold for higher values of C^1.

Other insights obtained in capacitated models with larger number of stages are similar to those described in Glasserman and Tayur (1992a).

6 Conclusions and Research Directions

We address the problem of inventory control in multi-stage production-distribution systems. This paper emphasizes the computational algorithms that are required for the above mentioned analysis. Besides providing the necessary algorithms, they are also applied to the analysis of several example test problems. The computational results of section 5 provide much insight into the behavior of these control policies.

Using a *shortfall* process, we explained the algorithm for uncapacitated systems. Extending this intuition, we provided algorithms for capacitated systems. These algorithms for the capacitated case offer a promising start although much work still remains to be

done in this area. Extensions to problems with nonstationary demand and or decentralized inventory control are of great practical importance.

Acknowledgements : The first author would like to express his thanks to Dr. A. Clark for bringing references 18 and 26 to his attention.

7 References

1. BITRAN, G.R., E.A. HAAS AND A.C. HAX, "Hierarchical Production Planning: A single stage system", **Operations Research** 29, 1981 pp 717-743.

2. CLARK A. AND H. SCARF, 'Optimal policies for a Multi-Echelon Inventory Problem', **Management Science** 6, 1960 pp 474-490.

3. COHEN M.A. AND H.L.LEE, 'Strategic Analysis of Integrated Production-Distribution Systems: Models and Methods' **Operations Research** 36, 1988, pp216-228.

4. COHEN M.A. AND S. MOON, 'Impact of Production Scale Economics, Manufacturing Complexity, and Transportation Costs on Supply Chain Facility Networks' **Journal of Manufacturing and Operations Management** 3, 1990, pp269-292.

5. DEBODT M. AND S.C. GRAVES, 'Continuous review policies for a Multi-Echelon Inventory Problem with Stochastic demand', **Management Science** 31, 1985, 1286-1299.

6. DURETT, R., 'Probability : Theory and examples', Wadsworth and Brooks/Cole series, 1991.

7. ESCUDERO, L.F. AND P.V. KAMESAM, "MRP modelling via Scenarios" RC 17982, IBM T.J.Watson Research Center, Yorktown Heights, NY., 1992. Also in *Optimization in Industry*, by Ciriani,T.A. and R. C. Leachman (ed), Wiley, London, 1993.

8. FEDERGRUEN A, SCHWEITZER, P.J. AND TIJMS, H., 'Denumerable Undiscounted Semi-Markov Decision Processes with Unbounded Rewards' **Mathematics of Operations Research** 8, 1983, 298-314.

9. FEDERGRUEN A AND P. ZIPKIN, 'Approximations of Dynamic Multilocation Production and Inventory Problems', **Management Science** 30, 1984, 69-84.

10. FEDERGRUEN A AND P. ZIPKIN, 'Computational Issues in an Infinite Horizon Multi-Echelon Inventory Model', **Operations Research** 32, 1984, 218-236.

11. FEDERGRUEN A AND P. ZIPKIN, 'An Inventory Model with Limited Production Capacity and Uncertain Demands. I: The Average Cost Criterion', **Mathematics of Operations Research** 11, 1986, 193-207.

12. FEDERGRUEN A AND P. ZIPKIN, 'An Inventory Model with Limited Production Capacity and Uncertain Demands. II: The Discounted Cost Criterion', **Mathematics of Operations Research** 11, 1986, 208-215.

13. GLASSERMAN, P. AND S. TAYUR, 'Sensitivity Analysis for Base Stock Levels in Multi-Echelon Production/Inventory Systems', Technical report, GSIA, Carneigie Mellon University, Pittsburg, PA, June 1992.

14. GLASSERMAN, P. AND S. TAYUR, 'The Stability of a Capacitated, Multi-Echelon Production Inventory System under a Base-Stock Policy', Technical report, GSIA, Carneigie Mellon University, Pittsburg, PA, August 1992.

15. GRAVES, S.C., 'Safety Stocks in Manufacturing Systems', **Journal of Manufacturing and Operations Management**, 1, 1988, pp67-101.

16. KARLIN S., 'Dynamic Inventory Policy with Varying Stochastic Demands', **Management Science** 6, 1960, pp231-258.

17. KARMARKAR U, 'Lot Sizes, Lead times and In-Process Inventories', **Management Science** 33, 1987, pp409-418.

18. LANGENHOFF, L.J.G. AND W.H.M. ZIJM, 'An Analytical Theory of Multi-Echelon Production/Distribution Systems', Working paper, Eindhoven University of Technology, Department of Mathematics and Computer Sciences, 1989. to appear in **Statistica Neerlandica**.

19. LEE H.L., C. BILLINGTON AND B. CARTER, 'Gaining Control of Inventory and Service through Design for localization', Working paper 1991.

20. MORTON, T., 'The Non-stationary Infinite Horizon Inventory Problem', **Management Science** 24, 1978, 1474-1482.

21. MUCKSTADT, J., M. LAMBRECHT AND R. LUYTEN, ' Protective Stocks in Multi-stage Production Systems ' **International Journal of Production Research**, 1984, vol 6, pp. 1001-1025.

22. PRABHU, N.U., 'Queues and Inventories', John Wiley, 1965.

23. ROSLING, K., 'Optimal Inventory Policies for Assembly Systems Under Random Demands', **Operations Research** 37, 1989, 565-579.

24. SCHMIDT, C. AND S. NAHMIAS, 'Optimal Policy for a Two-Stage Assembly System Under Random Demand', **Operations Research** 33, 1985, 1130-1145.

25. TIJMS, H.C., 'Stochastic Modelling and Analysis: A Computational Approach', John Wiley, 1986.

26. VAN HOUTUM, G.J. AND W.H.M. ZIJM, 'Computational Procedures for Stochastic Multi-Echelon Production Systems', Working paper, Eindhoven University of Technology, Department of Mathematics and Computer Sciences, 1991.

27. ZIPKIN, P., 'Models for Design and Control of Stochastic, Multi-Item Batch Production Systems', **Operations Research** 34, 1986, pp91-104.

28. ZIPKIN, P., 'Critical Number Policies for Inventory Control with Periodic Data', **Management Science** 35, 1989, pp71-80.

A Appendix 1

In this Appendix, we describe two approximation algorithms that are needed for Algorithm 3 and others as well. Let μ and σ^2 be the mean and variance of a random variable X with unknown density function $f(x)$. Let

$$cv^2 = \frac{variance}{mean^2}$$

Algorithm 5 gives us a two moment approximation scheme (see Tijms (1986)).

begin algorithm 5

1. if ($cv^2 > 1$) then
   - approximate $f(x)$ by $\left\{q\lambda_1 e^{-\lambda_1 x} + (1-q)\lambda_2 e^{-\lambda_2 x}\right\}$, where
   - $\lambda_1 = \frac{2}{mean}\left\{1 + \sqrt{\frac{(cv^2-0.5)}{(cv^2+1)}}\right\}$
   - $\lambda_2 = \frac{4}{mean} - \lambda_1$
   - $q = \frac{\lambda_1(\lambda_2 * mean - 1)}{\lambda_2 - \lambda_1}$
2. else
   - choose k s.t. $\frac{1}{k} \leq cv^2 \leq \frac{1}{k-1}$
   - approximate $f(x)$ by $\left\{q\frac{\lambda^{k-1}x^{k-2}e^{-\lambda x}}{(k-2)!} + (1-q)\frac{\lambda^k x^{k-1}e^{-\lambda x}}{(k-1)!}\right\}$, where
   - $q = \frac{k*cv^2 - \sqrt{k(1+cv^2) - k^2 cv^2}}{1+cv^2}$
   - $\lambda = \frac{k-q}{mean}$

end algorithm 5

Algorithm 7 computes the first two moments of F^C, where $F(.)$ is a weighted Erlang. To be clear, we will assume that if cv^2 of F is less than 1, it is approximated by a mixture of an Erlang (k) and an Erlang $(k-1)$; if the cv^2 of F is greater than 1, then it is approximated by a hyper-exponential, as in Algorithm 5 above. Let μ_{1,x^c} and μ_{2,x^c} be the first two moments of F^C. The quantities $\lambda, \lambda_1, \lambda_2, q, k$ have the same meaning as in Algorithm 5.

begin algorithm 7

- $cv^2 \leq 1$.

$$\begin{aligned}
\mu_{1,x^C} &= qA(k-1,\lambda,C) + (1-q)A(k,\lambda,C) \\
\mu_{2,x^C} &= qB(k-1,\lambda,C) + (1-q)B(k,\lambda,C)
\end{aligned}$$

where

$$\begin{aligned}
A(k,\lambda,C) &= -C(1-G_{k-1}(C)) + \frac{k}{\lambda}(1-G_k(C)) \\
B(k,\lambda,C) &= C^2(1-G_{k-2}(C)) - \frac{2Ck}{\lambda}(1-G_{k-1}(C)) + \frac{k(k+1)}{\lambda^2}(1-G_k(C))
\end{aligned}$$

where

$$G_k(C) = 1 - \sum_{j=0}^{k-1} e^{-\lambda C}\frac{(\lambda C)^j}{j!}$$

- $cv^2 \geq 1$.

$$\begin{aligned}
\mu_{1,x^C} &= \frac{q}{\lambda_1}e^{-\lambda_1 C} + \frac{1-q}{\lambda_2}e^{-\lambda_2 C} \\
\mu_{2,x^C} &= \frac{2q}{\lambda_1^2}e^{-\lambda_1 C} + \frac{2(1-q)}{\lambda_2}e^{-\lambda_2^2 C}
\end{aligned}$$

end algorithm 7

end Appendix 1.

Copies may be requested from:

IBM Thomas J. Watson Research Center
Distribution Services F-11 Stormytown
Post Office Box 218
Yorktown Heights, New York 10598