An Approach to Develop Base-Stock Inventory Control Policy In Serial Production Systems With Some Stochastic Features

Kadir Ertogral
George R. Wilson
Lehigh University

Report No. 98T-002

Table of Contents

List of Figures

ABSTRACT

In this report, we suggest an integrated approach for the development of a base stock inventory control policy in serial production systems with certain stochastic features. The stochastic features include Poisson demand process, random processing times and yield problem. We assume that there are several products to be produced in the system. Research activities are three fold : (1) Lot sizing and estimation of transit times through production stages (2) Developing the analytical model for base stock inventory control policy (3) Approximate optimization of base stock levels using a heuristic.

Chapter 1
Introduction And Literature Review

1.1 Introduction

The production, use or the distribution of inventories are activities in which almost all organizations are involved. On the average, 30% of current assets and 90 % of the working capital of a typical company in the United States are invested in inventories (Silver and Peterson, 1985). Thus, the inventory management has become one of the key issues to be addressed for a business to survive.

A multi-echelon inventory system is a network of inventory holding facilities organized into different echelons. Goods flow from one facility to another at a lower echelon. A major issue in the design of a multi-echelon inventory system is how to manage the flow of inventories so as to balance different considerations, such as economies of scale, inventory carrying cost, and service level.

A common example of a multi-echelon system is one with a number of retail outlets which directly satisfy customer demand. These retailers act as customers for a higher level wholesale or production operation. There could be more than two layers of echelons in the system. There has been particular interest in this type of multi-echelon distribution system. A production example of a multi-echelon inventory system is the one in which several final products are made in a common production facility. Inter-stage inventories decoupling the production stages are viewed as the levels of a multi-echelon inventory system.

The basic functions of an inventory policy are to determine (1) how often to review the inventory status, (2) how much to order, (3) when to place an order. The review interval, the order quantity, and the reorder level (inventory position at which an order must be placed) are usually designated by R, Q and s respectively. The inventory policies are either continuous review policies, such as (s,Q) policy (where an order of size Q is placed when the inventory position falls to or below s), or periodic review policies like the (R,S) policy (where an order is placed each R units of time to bring the inventory up to a value S). In multi-echelon systems, ordering decisions at a certain level of the hierarchy determine the demand at the next higher level. Because of this demand and supply dependency, any good inventory policy should be a coordinated one.

Some of the reasons that have made multilevel inventory control problems difficult relative to single level ones can be summarized as follows;

1) The demand process : The demand process and the inventory policy at one level of a multilevel system determine the demand process at the next higher level. Thus, even if the lowest level has a well behaved demand process, the supply and demand process will be more complex as we move up the hierarchy.

2) Treatment of shortages : Shortages at one level can cause shortages at lower levels.

3) Simultaneous optimization : Optimality is rarely achieved by determining the optimal policy at each level separately because of the demand and the shortage dependencies mentioned above.

4) Computing requirements : Due to its complexities, an analysis of the multilevel problem requires much more computational effort than a single level problem.

There are several structures of multilevel inventories and many ways to organize and control these structures. In this report, we are concerned with base stock inventory control policy development in serial production systems with some stochastic features such as, random processing time, demand and yield. The base stock control policy prescribes that each production stage start production whenever the associated inventory level drops to or below a predetermined base stock level and stop production when the inventory level reaches the base stock level. Thus, this policy implies continuous review. This type of policy can be classified as a "pull" system because replenishment orders are pulled down by the lower echelons from their replenishment sources.

The base stock control policy requires that only a single parameter, base stock level, be determined and results in a limited amount of communication between echelons. Fredergruen and Zipkin (1986 a,b) show that a base stock policy is optimal for capacitated single-stage systems. Clark and Scarf (1960) establish the optimality of base-stock policies in serial, uncapacitated multistage systems. Finally, Veatch and Wein (1994) show experimentally that base-stock policies are often close to optimal for a class of two-stage capacitated systems. These results, and the ease of implementation, made us decide to develop an approach to find base-stock inventory control policy for serial production systems.

We also needed to make multi-item lot sizing decisions at each echelon level which yield transit times at each production stage. These transit times are necessary inputs in developing a base stock control policy in our approach. Hence, we devote the first part of the solution procedure to the lot sizing decision.

1.2 Literature review

For the lot sizing part of our approach we adopted a modified version of a method presented by Karmarkar et al. (1992). They suggest multi-item batching heuristics which try to minimize the queuing delays in a single server queuing system with Poisson arrivals of multiple items and general service time. They approximate multiple Poisson arrivals with a single arrival stream from a Poisson distribution and, thus, the system is characterized as M/G/1. They claim that taking the minimization of queuing delays of the batches as the objective in batching yields better results in terms of cash flow compared to the conventional trade off between setup and inventory holding cost. Further discussion on this method is given in chapter 3.

In the remainder of this chapter, relevant literature on multilevel inventory control is discussed. First, we introduce some literature on general muti-echelon systems and, then, discuss research on serial systems.

1.2.1 General multi-echelon systems

In the area of multi-echelon inventory control, systems with deterministic demand have been studied for the last three decades. The problem studied often is a two echelon problem ; the, so called, one-warehouse n-retailer problem. Schwarz (1973) proved that if an optimal policy exists, then it has to satisfy the following conditions;

1. Deliveries are made to the warehouse only when the warehouse has zero inventory and at least one retailer has zero inventory.

2. Deliveries are made to any given retailer only when that retailer has zero inventory.

3. All deliveries that are made to any given retailer between successive deliveries to the warehouse are of equal size.

He also suggested some heuristic procedures for determining the lot sizes at each retailer and at the warehouse. His policies were nested stationary policies (a nested policy is one where each facility orders each time its immediate supplier does and perhaps at other times too, while a stationary policy is one where each facility orders at equally spaced points in the time and in equal amounts). Up on testing of these heuristics against analytical lower bounds, he found that they gave near optimal solutions.

Graves and Schwarz (1977) examined optimal and near-optimal continuous review policies for deterministic arborescent inventory systems (where an echelon inventory point can be supplied only by a single echelon inventory point, while it can supply more than one echelon inventories). They extended the conditions for optimality developed by Schwarz (1973).

Roundy (1985) examined one-warehouse n-retailer problems where the demand is constant and there is linear holding cost. In his literature review he mentions that optimal policies seem to be very difficult to compute. He came up with a new class of policies and proved that the cost of a policy in this class is within 2 % of the optimal solution. Park and Kim (1987) developed an inventory model for two echelon distribution systems assuming periodic ordering at both levels and a constant, deterministic demand. Johnson and Silver (1987) examined the redistribution at the lower level echelon (branch warehouses) due to out of balance inventory. The redistribution they considered is a complete redistribution of all lower level facility inventories one period before the order cycle. Muckstadt and Roundy (1987) studied the problem of coordinating the purchase and shipment of items in a one-warehouse, n-retailer system. Their model includes positive echelon holding cost, fixed cost for ordering and shipping each-item, and a fixed joint item order cost. They also assumed that a stationary nested policy is followed. Finally, Williams (1981) discusses seven heuristic algorithms used for deterministic distribution scheduling in arborescent networks and joint deterministic production distribution scheduling in conjoined assembly arborescent networks.

In multilevel inventory distribution systems under probabilistic demand, work again has been mostly on the one-warehouse, n-retailer problem. Federgruen and Zipkin (1984a) considered a central depot supplying several locations. Assuming that the central depot carries no inventory, they developed a dynamic programming model to minimize the expected total cost. Deuermeyer and Schwarz (1981) developed an analytical model for estimating the expected fill rate for a one-warehouse, n-retailer system as a function of the system parameters. This model was based on an exact, single facility (R,Q) model of Hadley and Whitin (1963). Rosenaum (1981) developed a heuristic model which combines the service levels at the two echelons. This heuristic minimizes the company safety stock, while ensuring that a specific percentage of the customer demand will be filled from on hand inventory. Sand (1981) developed two methods for predicting the demand on the secondary level. He compared his results to the data obtained from simulation and concluded that the results are quite accurate.

Svoronos and Zipkin (1991) published an approach to approximate the system performance measures in an arborescent multi-echelon system. They assume stochastic transit times of the parts between echelons and a Poisson demand process arising at the lowest echelon. They considered the base stock control policy. We adopt their approach in this paper to estimate the system performance measure for a given set of base stock levels since the approach has the closest set of assumption to those of our problem. More explanations of this approach is given in chapter 4. Zipkin (1991) extends this approach to allow compound Poisson demand distribution.

1.2.2 Serial Systems

There has been relatively less work done on particular serial systems compared to the work for general multi-echelon systems. Clark and Scarf (1960) published one of the first papers on serial and assembly inventory systems. Their model dealt with periodic review policies under stochastic demand. They presented the echelon inventory and echelon inventory holding cost concepts for the first time. They assumed constant demand originating at the lower level only. They also assumed the cost of purchasing and shipping of an item to be linear without any setup cost. The only exception to this assumption was at the highest echelon where a setup cost is permitted. Clark and Scarf (1962) also give an approximate solution to a two-level problem with setup cost at both levels. Their results only apply to finite horizon problems. Federgruen and Zipkin (1984b) extend the Clark and Scarf (1960) approach to infinite horizon problems to find an optimal solution for a two-level serial system and approximate policies for a two-level arborescent system.

Debodt and Graves (1985) present a continuous review inventory model for a multistage serial inventory system where the demand for the end item is stochastic and stationary. They provide approximate performance measures under a nestedness assumption ; whenever a stage receives a shipment a batch must be send to its downstream stage. Bardinelli (1992) studies installation (R,Q) policies in serial systems with Poisson demand. By assuming that the inventory position at each stage is non-negative, he provides exact, long-run average holding and backorder cost expressions. Glasserman and Tayur (1996) present an approximation for serial production-inventory system with limited capacity. They assumed fixed lead times. They adopt a base-stock inventory control policy and suggest that the distribution of echelon inventory can be approximated as a sum of exponentials which is the basis of their analysis. They optimize the base-stock levels for multistage examples using their approximation model and compare their results to simulation results. Their results turn out to be very close to simulation results in terms of system inventory holding cost. Glasserman (1997) gives more detailed explanation of their approximation model for single stage and serial systems which also accommodates imperfect production.

Chapter 2

Problem Description and Summary of The Solution Approach

2.1 Problem description and objective

The production system for which we will develop a base-stock inventory control policy is assumed to have the following properties ;

1. It is composed of serially connected production stages.
2. There are multiple products to be produced in the system.
3. The setup times associated with each type of product at each stage are significant and constant.
4. The processing time per unit product at each stage is a random variable with a general distribution.
5. There are yield problems in the system. Yield is a unit by unit binomial process.
6. There is an infinite supply of raw material before the first stage for each type of product.
7. Demand arises only at the lowest inventory point after the final stage and it has a Poisson distribution.
8. If a demand can not be satisfied at the moment it arises it is backordered.
9. At each stage only one unit of semifinished product from the previous stage is necessary for the current stage to start processing.
10. There are no breakdowns in the system.

The following figure is an example of a four-stage system showing the labeling we use for production stages and inventory points;

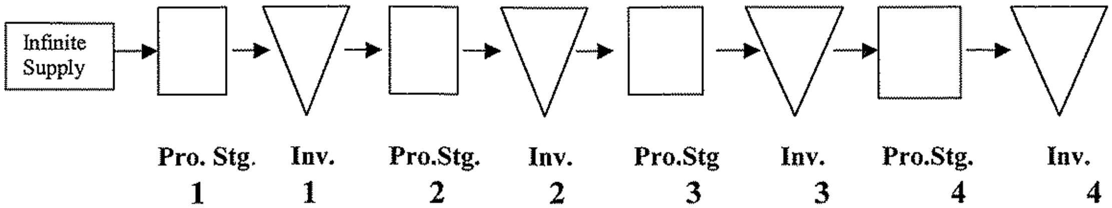


Figure 2-1 Representation of a four-stage system.

The objective of this report is to develop an approach to determine the base-stock levels at each inventory point for each type of product which minimizes the system inventory holding cost and, at the same time, achieves a certain fill rate or, in other words, service level at the end product inventory. Other type of restrictions, such as forcing the expected inventory level at an inventory area to be less that a certain value can easily be incorporated in our approach as we explain in chapter 5.

Our problem description is quite comprehensive in terms of the assumptions. We include many features that a real problem can have in our problem description, such as random processing times, yield problem, random demand, multiple products and limited production capacity.

2.2 Summary of the solution approach

Our approach to the solution of the problem, described above, is complete in that it starts with lot sizing using processing and setup times, and demand data, continues with establishing the relation between base-stock levels and fill rate, and ends with approximately optimizing the base-stock levels at each stage by using a heuristic search method. There has been some work in the literature dealing with one of the first two parts of our approach but not all of the three parts, collectively. Thus, our approach has three main steps as described in the following subsections.

2.2.1 Estimation of the moments of transit times

The transit time for a part is defined as the time period that elapses between the release of the part from the previous inventory and the entrance of the part to the current inventory. That is, it is composed of queuing time and service time of the part at the current production stage. The approach taken here requires that the first two moments of the transit times of each type of part at each stage be known or estimated. It is clear that the transit times will be affected by the lot sizing decisions since the lot sizes are important factors determining both the setup time share of a unit product, which can be found by dividing the setup time by the lot size, and the queuing time of each lot of products.

The lot sizing decision part of our problem can be classified as multi-product, multilevel, stochastic lot sizing. This problem is a very complicated one. We needed to simplify it since the main purpose of this step of the approach merely is to get good estimates of the first two moments of the transit times under steady state conditions provided by using certain lot sizes. One may even skip this step if good estimates of these moments are known from a simulation study or historical data. Therefore, we ignored the multilevel property of the lot sizing problem.

There is an important fact regarding the entire analysis. If the real distribution of the lead times, the transit times plus the delay due to lack of inventory at the previous inventory point, can be approximated satisfactorily by a gamma distribution, our analysis will be exact. The deviation from the true values will, thus, be dependent on the following two items;

1. How good a gamma distribution can represent the real distribution of the lead times.
2. How good our estimates of the first two moments of the lead times are.

After finding the estimates of the first two moments of the lead times for each product at each stage under steady state, we carry out the analysis product by product and consider the interaction among products and capacity limitation as we find the estimates. We address the lot sizing issue and the estimation of the moments of the transit times in chapter 3.

2.2.2 Establishing the analytical relationship

The analytical relationship to be established here is between the base stock levels (the decision variables) at each stage for each product and the fill rate, the system performance measure, for each product. As we stated in the literature review part, we will apply a method suggested by Svoronos and Zipkin (1991) to establish the relationship. We modified this analytical model since our problem setup differs from the one they assumed.

We selected to use fill rate as the system performance measure. Any other system performance measure can easily be replaced or used together with the customer fill rate in our approach. Because the analytical model provides the state of the system at all stages, our search heuristic can easily accommodate any restriction based on these states, such as inventory level or backorder level restriction at any stage. We explain the analytical model in chapter 4.

2.2.3 Approximate optimization of base-stock levels

Here the optimization means the determination of the base-stock levels which minimize the system inventory holding cost. The single restriction of assuring a predetermined fill rate needs to be taken into account in the cost minimization process.

We suggest a heuristic search procedure that accomplishes this approximate optimization part of the approach. The heuristic utilizes the analytical relationship introduced in subsection 2.2.2 to calculate the system inventory holding cost and the customer satisfaction rate for a given set of base-stock levels at each stage. We discuss this search heuristic in chapter 5.

Chapter 3
Lot Sizing and Estimation of The Moments of Transit Times

3.1 Introduction

As we mentioned in chapter 2.2.1 estimates of the first two moments of the transit times are the inputs for the second step of our approach and lot sizes directly affect transit times. We consider one stage at a time as we make lot sizing decision in our multi stage system where the processing times has general distribution. As an approximation we will model each stage as an M/G/1 system and utilize a published method by Karmarkar et al. (1992) and Karmarkar et al. (1985) which relates the lot sizes to waiting time in the queue and in the system in the M/G/1 system.

In general, lot sizing models have traditionally aimed at a tradeoff between inventory holding and setup costs. They concentrate on finished goods and echelon stocks and overlook the congestion and queuing delays. These delays, in turn, increase the lead time. Long lead times mean high levels of work-in-process inventories and poor inventory "turn" which is defined as 1/(lead time) in Karmarkar et al. (1985). They also cause high levels of safety stock since the variance of the lead time increases proportionally to lead time. As a result, the competitiveness of a firm can be harmed.

It is usually assumed that the performance of a manufacturing system with delays and queues is primarily due to dispatching and sequencing at production resources. But in fact the lot sizing policy applied in the system is another major factor affecting the queuing behaviors in the system. Large lot sizes cause queue buildups at production resources since large lots tie up the resources for long time periods. Reducing the lot sizes will be helpful up to a certain point. After this point, reduction of lot sizes will cause longer queue buildups and delays because of the frequent setups leading to high workloads at production resources and, after a threshold, explosively increasing queuing times.

The model presented here ignores the setup costs and concentrates on queuing delays, whereas the traditional EOQ models do not consider the queuing delays and their consequences in the system. Rummel (1989) shows that the cost models based on the minimization of queuing delays and lead times are more representative in terms of cash flow than the traditional ones mentioned above.

The results of the published model by Karmarkar et al. (1992) is quite different than the usual EOQ models. They show that lot sizes are linear in setup times and processing rates and increase explosively with total utilization. The queuing delays are also approximately linear in setup times and convex increasing in total utilization.

3.2 Model formulation

In the published model by Karmarkar et al. (1992), it is assumed that the demand and processing times are constants. We use the expected value of processing times and demand as an approximation since we are just looking for rough estimates of the moments of transit times using relatively "good" lot sizes. Subsequently, we relax this assumption and use random processing times as we calculate the transit times corresponding to the lot sizes calculated. We account for the yield factor as we calculate the demand for each stage. We divide demand arising at the inventory of the last stage by the multiplication of the yield factors of all down stream stages, including the current stage, to find the effective demand for a stage.

What follows is the model and analysis for a single stage which has to be repeated for each stage.

Let i = 1..n index values for individual products and define

D_i : Expected demand for item i (units/time)

P_i : Processing rate for item i (units/time).

Q_i : Batch size for item i.

τ_i : Setup time for item i.

X_i : Processing time for a batch of item i.

The expected processing time for a batch of item i is given by

$$X_i = \tau_i + \left(\frac{Q_i}{P_i}\right)$$

Let $\lambda_i = \left(\frac{D_i}{Q_i}\right)$ denote the number of batches per time unit for item i. If we assume that a batch arriving at a facility is selected randomly, the probability of selecting a batch of item i is ;

$$\pi_i = \frac{\lambda_i}{\sum_j \lambda_j}$$

and the mean service time is ;

$$E[X] = \sum_i \pi_i X_i$$

or

$$E[X] = \frac{\sum_i \{(\frac{D_i}{Q_i})(\tau_i + \frac{Q_i}{P_i})\}}{\sum_i \frac{D_i}{Q_i}}$$

The mean waiting time in queue for this system is given by the Pollaczek-Khinchin formula (Buzacott, J.A. and Shanthikumar, J.G. (1993)) as

$$E[W(Q)] = \frac{\lambda E[X^2]}{2(1-\rho)}$$

where the traffic intensity $\rho = (\lambda E[X])$; $\lambda = \sum_i \lambda_i$ and $E[X^2] = \frac{\sum_i \lambda_i X_i^2}{\lambda}$. The average time spent in the system by a batch is ($X_i + E[W(Q)]$).

Here, the problem is to determine the batch sizes for each item that minimizes the queuing delays of batches. This problem is represented by the following non linear programming model ;

$$W^* = \min_{Q_i > 0} E[W\{Q)]$$

S.T.

$$\sum_i \{ (\frac{D_i}{P_i}) + (\tau_i \frac{D_i}{Q_i}) \} < 1$$

The constraint is equivalent to $\rho < 1$.

3.3 Derivation of the heuristic rule for lot sizing

Following are the dimensionless quantities defined to simplify and clarify the relationships involved in the derivation :

$u_i = \frac{D_i}{P_i}$, the work center utilization due to item i

$u = \sum_i u_i$, the total utilization of the work center

$q_i = \frac{Q_i}{(D_i \tau_i)}$, a dimensionless batch size.

$p_i = \frac{Q_i}{(P_i \tau_i)} = q_i u_i$, batch size expressed as the ratio (run time/setup time)

$s_i = (\frac{1}{q_i})$ and $s = \sum_i s_i$, the proportions of time spent in setups for item i, and in total, respectively.

$t_i = (\frac{\tau_i}{\sum_j \tau_j})$, a dimensionless setup time.

$w = \frac{W}{\sum_j \tau_j}$, a dimensionless waiting time in the queue.

Using this notation we can write the dimensionless mean queue time as (following Karmarkar et al. (1992))

$$w = \frac{\sum_i t_i \{ (\frac{1}{q_i}) + 2u_i + u_i^2 q_i) \}}{2 \{1 - \sum_i (\frac{1}{q_i + u_i}) \}}$$

$$= \sum_i \left(\frac{t_i (s_i + u_i)^2}{2s_i \{ 1 - \sum_j (s_j + u_j) \}} \right)$$

From this expression, it is apparent that it is enough to specify u_i and t_i to describe a problem instance. The domain of w as a function of s_i is the interior set

$$S = \{ s : \sum_i s_i \leq 1 - \sum_i u_i , s_i \geq 0 \text{ for all i} \}$$

denoted by int S. Now we can state the problem as

$$w^* = \min_{s \in \text{int } S} w(s)$$

Karmarkar et al. (1992) derive three closed form heuristic rules to find the lot sizes based on bounds on w* and certain relationships that they establish, instead of using an optimization technique to solve the model above. We elected to use the second heuristic rule among the three since it outperforms the other heuristic rules and gives results close to optimal in the experiments included in the paper, specially at utilization levels 70% or higher. We give a brief discussion of this second heuristic rule.

Some relationships and bounds on w* that are presented and proved in the paper are

$$2w^* = t_i\{(\frac{u_i}{s_i^*})^2 - 1\} \text{ for each i}$$

$$w_L = \frac{2\sum_i t_i u_i}{(1-u)}$$

$$w_U = \frac{2\sum_i t_i u_i}{(1-u)^2}$$

$$\text{and } \; w_L \leq w^* \leq w_U$$

The heuristic rule we seleced is obtained by assuming that utilization levels are high and that $w^* \gg t_i$. Using the relationship in the first equation above, this assumption leads to the following

$$t_i(\frac{u_i}{s_i})^2 = 2w^*$$

or

$$s_i = u_i\sqrt{\frac{t_i}{2w^*}}$$

The heuristic rule is then obtained by substituting w_U as an estimate for w^* ;

$$s_i = u_i\sqrt{\frac{t_i}{2\,w_U}} = \frac{u_i(1\text{-}u)}{2}\sqrt{\frac{t_i}{\sum_j t_j u_j}}$$

After the proper substitution, this expression translates into the following one that yields the lot size, Q_i ;

$$\mathbf{Q}_i = \frac{\mathbf{2P}_i\tau_i}{\mathbf{(1\text{-}u)}}\sqrt{\frac{\sum_j \tau_j \mathbf{u}_j}{\tau_i}}$$

3.4 Estimation of the moments of transit times

We use the heuristic rule described in the previous section to find the lot sizes for each product at each stage. This rule is based on the assumption that the processing rates or the processing times are constant. We substitute the expected values of processing rates as constant processing rates when we calculate the lot sizes. After finding the lot sizes, we no longer assume constant processing times as we continue with the calculation of the transit times, the time period that elapses between the release of a unit product from the previous inventory and the entrance of the unit product to the current inventory. In this section we explain how to get the first two moments of the transit times of a unit of a product given that we use the lot sizes calculated. We define the moments for the processing time of a batch of item i which we use to calculate E[W] and Var[W] later on. The first moment is

$$E[X_i] = Q_i E[r_i] + \tau_i$$

where r_i represents the random processing time of unit of item i. Here, we assume that r_i's are independently distributed, identical random variables. $E[X_i^2]$ is found using the variance of X_i, $Var[X_i]$;

$$E[X_i^2] = Var[X_i] + (E[X_i])^2$$

and the variance of X_i is;

$$Var[X_i] = Q_i Var[r_i]$$

$E[X_i^3]$ is found using the regular expectation formula by assuming that X_i's will have normal distribution ($X_i \sim N(E[X_i], Var[X_i])$). This assumption is justified according to the central limit theorem since X_i is simply the summation of many independent random variables. Hence ;

$$E[X_i^3] = \int_{-\infty}^{\infty} X_i^3 \, f(X_i) \, dX_i$$

where $f(X_i)$ designates the density function of X_i. The equations that follow yield the first two moments and the variance of the waiting time of a batch in the M/G/1 system (Buzacott, J.A. and Shanthikumar, J.G. (1993)). ;

$$E[W] = \frac{\lambda E[X^2]}{2(1-\rho)}$$

$$E[W^2] = \frac{\lambda E[X^3]}{3(1-\rho)} + \frac{\lambda^2 E[X^2]^2}{2(1-\rho)^2} + \frac{\lambda E[X^2]E[X]}{(1-\rho)}$$

$$Var[W] = E[W^2] - (E[W])^2$$

where $E[X]$, $E[X^2]$ and $E[X^3]$ are the first, second and the third moments of the processing times of a common batch, respectively. These moments are found using the equation

$$E[X^k] = \frac{\sum_i \lambda_i E[X_i^k]}{\lambda}.$$

The average transit time of a unit of item i, T_i, e consists of waiting time of a batch, setup time for item i, waiting time for the processing of half of a batch of item i and the processing time of a unit of item i. That is ;

$$T_i = W + \tau_i + \sum_{i=1}^{\frac{Q_i}{2}+1} r_i$$

Using this equation, we find $E[T_i]$ and $E[T_i^2]$ as follows ;

$$E[T_i] = E[W] + \tau_i + \left(\frac{Q_i}{2} + 1\right) E[r_i]$$

$$Var[T_i] = Var[W] + \left(\frac{Q_i}{2} + 1\right) Var[r_i]$$

$$E[T_i^2] = Var[T_i] + (E[T_i])^2$$

We are now ready to continue with the second step of our solution approach since the first two moments of transit time of a unit for each product at each stage and the demand rates are all we need as inputs for the second step.

3.5 A numerical example

We give a numerical example here to illustrate the calculations for a stage. The calculations go are identical across all stages. The difference in calculations for different stages, obviously, will be the processing times and demand rates for each product.

In this example, we have a four-stage system with three different products being produced. Stage 4 is the final stage in the production process. We show the calculations for stage 2. The data for the example follows :

Available weekly capacity : 5 days/week x 8 hours/day x 60 min./hour = 2400 minutes.

Demand : 600 units/week , 190 units/week, 150 units/week for product 1, product 2 and product 3, respectively.

Stage 2	**Exp. proc. time**	**Var. of the unit proc. time**	**Process Rate**	**Setup times**
Prod. 1	.6 min/unit.	.09 min^2.	1.6667 units/min	10 min.
Prod. 2	1.4 min/unit.	.49 min^2.	.7143 units/min	15 min.
Prod. 3	1.6 min/unit.	.64 min^2.	.625 units/min	35 min.

	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**
Yield rate	.9	.8	.85	.7

Table 3-1 Data of the example problem

First we calculate the effective demand rates, D_i's, at stage 2 that account for loss due to the yield problem;

$$D_i = \frac{original\ demand\ of\ product\ i}{Product\ of\ the\ yield\ rates\ of\ all\ down\ stream\ stages\ including\ stage\ 2}$$

$$D_1 = 300 / (.8 \times .85 \times .7) = 630.2 \text{ units/week}$$

$$D_2 = 190 / (.8 \times .85 \times .7) = 399.2 \text{ units/week}$$

$D_3 = 150 / (.8 \text{ x } .85 \text{ x} .7) = 315.1$ units/week

The utilizations due to each product at stage 2 are ;

$$u_i = \frac{D_i}{P_i}$$

$$u_1 = 630.2 / (1.6667 \text{ x } 2400) = .157$$

$$u_2 = 399.2 / (.7143 \text{ x } 2400) = .233$$

$$u_3 = 315.1 / (.625 \text{ x } 2400) = .210$$

Total unitization at stage 2, $u = .60$. Now, we can calculate the lot sizes for each product using the expression given in previous section ;

$$Q_i = \frac{2P_i\tau_i}{(1\text{-}u)}\sqrt{\frac{\sum_j \tau_j u_j}{\tau_i}}$$

$$Q_1 = \frac{2 \text{ x } 1.6667 \text{ x } 10}{(1\text{-}.6)}\sqrt{\frac{(10\text{x } .157 + 15\text{x } .233 + 35\text{x } .210)}{10}} = 92.85 \simeq 93 \text{ units}$$

$$Q_2 = \frac{2 \text{ x } .7143 \text{ x } 15}{(1\text{-}.6)}\sqrt{\frac{(10\text{x } .157 + 15\text{x } .233 + 35\text{x } .210)}{15}} = 48.74 \simeq 49 \text{ units}$$

$$Q_3 = \frac{2 \text{ x } .625 \text{ x } 35}{(1\text{-}.6)}\sqrt{\frac{(10\text{x } .157 + 15\text{x } .233 + 35\text{x } .210)}{35}} = 65.14 \simeq 65 \text{ units}$$

Next we find the first two moments and the variance of the processing time of a batch for each product corresponding to the lot sizes above;

$$E[X_i] = Q_iE[r_i] + \tau_i \qquad Var[X_i] = Q_iVar[r_i]$$

$$E[X_1] = 93 \text{ x } .6 + 10 = 65.8 \text{ min} \qquad Var[X_1] = 93 \text{ x } .09 = 8.37 \text{ min}^2$$

$$E[X_2] = 49 \text{ x } 1.4 + 15 = 83.6 \text{ min} \qquad Var[X_2] = 49 \text{ x } .49 = 24.01 \text{ min}^2$$

$$E[X_3] = 65 \text{ x } 1.6 + 35 = 139 \text{ min} \qquad Var[X_3] = 65 \text{ x } .64 = 41.60 \text{ min}^2$$

$$E[X_i^2] = \text{Var}[X_i] + (E[X_i])^2$$
$$E[X_1^2] = 8.37 + 65.8^2 = 4338.01 \text{ min}^2$$
$$E[X_2^2] = 24.01 + 83.6^2 = 7012.97 \text{ min}^2$$
$$E[X_3^2] = 41.6 + 139^2 = 19362.60 \text{ min}^2$$

$$E[X_i^3] = \int_{-\infty}^{\infty} X_i^3 f(X_i)\, dX_i$$
$$E[X_1^3] = 286542.54 \text{ min}^3$$
$$E[X_2^3] = 590298.76 \text{ min}^3$$
$$E[X_3^3] = 2702966.17 \text{ min}^3$$

$E[X_i^3]$'s were found by applying the regular expectation formula using Maple.

Demand rates in batches for each product are;

$$\lambda_1 = 630.2 / 93 = 6.7763 \text{ batches/week}$$
$$\lambda_2 = 399.2 / 49 = 8.1469 \text{ batches/week}$$
$$\lambda_3 = 315.1 / 65 = 4.8477 \text{ batches/week}$$

Total demand , λ = 19.771 batches/week. The calculations for the moments of the processing time of a common batch are

$$E[X] = \frac{(6.7763 \text{ x } 65.8) + (8.1469 \text{ x } 83.6) + (4.8477\text{x}139)}{19.771} = 91.062 \text{ min}$$

$$E[X^2] = \frac{(6.7763 \text{ x } 4338.01) + (8.1469 \text{ x } 7012.97) + (4.8477\text{x}19362.60)}{19.771}$$
$$= 9124.16 \text{ min}^2$$

$$E[X^3] = \frac{(6.7763 \text{ x } 286542.54) + (8.1469 \text{ x } 590298.76) + (4.8477\text{x}2702966.17)}{19.771}$$
$$= 1004196.67 \text{ min}^3$$

The traffic intensity is $\rho = \frac{\lambda E[X]}{2400} = \frac{19.771\text{x}91.062}{2400} = .75$. Using the moments found above and the traffic intensity, we calculate the first two moments and the variance of the waiting time of a batch as follows;

$$E[W] = \frac{(19.771/2400) \times 9124.16}{2x(1-.75)} = 150.3 \text{ min}$$

$$E[W^2] = \frac{(19.771/2400)x1004196.67}{3x(1-.75)} + \frac{(19.771/2400)^2x9124.16^2}{2x(1-.75)^2} + \frac{(19.771/2400)x9124.16x91.082}{(1-.75)}$$

$$= 83605.44 \text{ min}^2$$

$$Var[W] = 83605.44 - 150.3^2 = 61015.35 \text{ min}^2$$

Finally, we are ready to find the first two moments of the transit times of a unit for each product at stage 2 :

$$E[T_1] = 150.3 + 10 + (\tfrac{93}{2} + 1)x.6 = 188.8 \text{ min}$$

$$Var[T_1] = 83605.44 + (\tfrac{93}{2} + 1)x.09 = 83609.71 \text{ min}^2$$

$$E[T_1^2] = 83609.71 + 188.8^2 = 119255.14 \text{ min}^2$$

$$E[T_2] = 150.3 + 15 + (\tfrac{49}{2} + 1)x1.4 = 201 \text{ min}$$

$$Var[T_2] = 83605.44 + (\tfrac{49}{2} + 1)x.49 = 83617.93 \text{ min}^2$$

$$E[T_2^2] = 83617.93 + 201^2 = 124018.93 \text{ min}^2$$

$$E[T_3] = 150 + 35 + (\tfrac{65}{2} + 1)x1.6 = 238.9 \text{ min}$$

$$Var[T_3] = 83605.44 + (\tfrac{65}{2} + 1)x.64 = 83628.88 \text{ min}^2$$

$$E[T_3^2] = 83628.88 + 238.9^2 = 140701.54 \text{ min}^2$$

It can be seen that the waiting time of a batch dominates the results in this example.

We have also calculated the waiting time of a batch corresponding to two sets of lot sizes. In the first set, lot sizes are fixed at approximately half of the lot sizes we calculated in the example and in the second set, approximately three times those values. The lot sizes and the corresponding waiting times are the following;

I:	$Q_1 = 45$	$Q_2 = 25$	$Q_3 = 32$	$E[W] = 282.12$ min
II:	$Q_1 = 300$	$Q_2 = 150$	$Q_3 = 200$	$E[W] = 237.02$ min

This brief sensitivity analysis shows the fact, explained in Karmarkar et al. (1992), that having smaller lot sizes than the optimal ones affects the waiting time much more drastically than having larger lot sizes. The waiting time calculated in the example was $E[W] = 150.3$ min.

Chapter 4
The Base-Stock Inventory Control Policy

4.1 Introduction

In this part of the report, we present the model that relates the base-stock policy parameters to the system performance measure, namely the fill rate. This model is a modified version of a published model by Svoronos and Zipkin (1991) for arborescent network structure. We modified the model to take the yield issue into account and used the modified model to analyze our serial system.

As we stated earlier we assume that demand occurs at the last stage inventory and it is a Poisson process. Each stage follows a base-stock, or (S-1,S), or one-for-one replenishment policy. This policy requires a single parameter $S \geq 0$, base-stock level for each stage.

If we assume that a system with four stages starts with inventory levels equal to base stock levels at each stage, every single occurrence of demand at the last stage causes an order to be placed with the inventory of the third stage. In turn, this order becomes a demand for the third stage and an order is immediately placed with the second stage inventory, etc. Thus, every demand at the last stage results in an order being placed against each stage inventory in the system, immediately.

A demand at the last stage inventory is satisfied immediately if there is inventory or after a delay if backordered. But for the previous stages, a demand is always satisfied after a time period. This time period involves both a delay time associated with the previous inventory and a transit time due to the production activities. Both the delay time and the transit time are stochastic variables. The transit times in our model are assumed to have a gamma distribution.

Svoronos and Zipkin (1991) assume that the orders are processed sequentially and do not cross in time; that is, a FIFO queuing discipline is followed. This is not an issue in our model since we analyze the system one product at a time after finding estimates for the transit times at steady state and, hence, there is no difference among orders.

Another important assumption of theirs is that the transit times are independent of the demands and orders in the system. This assumption is justified if we suppose that there are many orders in the system at steady state so that our units comprise a quite small portion of the total workload. It is as if the transit times are determined by observing a queuing system at steady state and we do not influence the transit system by the decisions we are making. The estimates of the moments of the transit times that we found in section 3.1. are of this nature.

Some useful features of the model are that the lead times are assumed to be stochastic and, not only the mean of transit times is important in the model, but also the variance of transit times matters.

4.2 Notation

λ : Expected demand rate.

T_i : Transit time before stage i, the random time from the release of a unit by the inventory of stage i-1 until the receipt of the unit at the inventory of stage i.

F_{T_i} : Distribution of T_i.

(T_i's represent the transit times when the system is in equilibrium, so F_{T_i} is the steady state distribution of transit times.)

S_i : Base-stock level for inventory of stage i. (The decision variables of the model)

What follows are the random variables describing the equilibrium behavior of the system. These variables are dependent on transit times and base stock levels.

D_i : Delay after inventory of stage i, the time from the order of a unit by the inventory of stage i+1 until the release of the unit by the inventory of stage i.

L_i : Total lead time of stage i.

$$= D_{i-1} + T_i$$

Let F_{D_i} and F_{L_i} denote the corresponding distributions.

I_i : Inventory at stage i.

B_i : Backorders at the inventory of stage i.

K_i : Number of outstanding orders at stage i.

$$= S_i - I_i + B_i$$

The densities for these variables are denoted by g_{I_i}, g_{B_i} and g_{K_i} respectively. For any x, let $[x]^+ = \max\{x,0\}$

4.3 Analyses for a single stage

We will drop the index i as we analyze a single stage. The steady state behavior of the system is described by the random variables I, B, and K = S - I + B with densities g_I , g_B and g_K . The delay time that an order waits as a backorder is another variable that we are concerned with. Let D denote a customer delay in steady state and F_D its distribution.

The following are the two key results presented and proved in Svoronos and Zipkin (1991) :

a) The variable K has the same distribution as the *lead-time-demand* , the number of demands in a random time with distribution F_L .

b) The variable B has the same distribution as the *order-delay-demand* , the number of demands in a random time with distribution F_D .

These results can be represented exactly in terms of transforms: Let τ_B and τ_K denote the z transforms of the densities g_B and g_K, respectively, and Υ_L and Υ_D Laplace transform of the distributions F_L and F_D, respectively. Then ;

$$\tau_K(z) = \Upsilon_L[\lambda(1\text{-}z)]$$
$$\tau_B(z) = \Upsilon_D[\lambda(1\text{-}z)]$$

Given F_L, we can compute τ_K using the first equation above. Then, inversion of τ_K gives us g_K with which we obtain the densities of I and B since $I = [S\text{-}K]^+$ and $B = [K\text{-}S]^+$. By using the density of B, g_B, we can find F_D using the second equation above.

This procedure involves many transform inversions which could turn out to be difficult. Therefore, in the next section we explain an alternative approximation method that uses only the first two moments of transit times.

4.4 Analyses for the serial system using two-moment approximation

We can approximate the procedure explained in the previous section by assuming a parametric family of densities for g_{K_i}. Each g_{K_i}, then, is characterized by specifying the first two moments at each stage. Here, we select the negative binomial family. This is equivalent to assuming that lead times have gamma distribution from the first of the transform equations above.

For a single stage, the following equations can be derived using both transform equations :

$$E[K] = \lambda \, E[L]$$
$$E[K(K-1)] = \lambda^2 \, E[L^2]$$
$$E[B] = \lambda \, E[D]$$
$$E[B(B-1)] = \lambda^2 \, E[D^2]$$

If we assume that we know $E[L]$ and $E[L^2]$, we can find $E[K]$ and $E[K(K-1)]$ from the first two equations above. This also yields $V[K]$. Since now we know $E[K]$ and $V[K]$ and assume that g_K has a negative binomial distribution, we can calculate the parameters n,p for g_K in the following way :

$$E[K] = \frac{n\,p}{(1-p)}$$
$$V[K] = \frac{n\,p}{(1-p)^2}$$

From these equations we have

$$p = 1 - \frac{E[K]}{V[K]}$$
$$n = \frac{(1-p)}{p} \, E[K]$$

Now, we are ready to find the moments for B using the relation $B = [K\text{-}S]^+$ since we know g_K. Let $\epsilon = \frac{p}{(1-p)}$ and let G_K denote the complementary cumulative distribution of K. Then, using the standard formulas for expected values, and after some modifications, we get the following equations :

$$E[B] = (n\epsilon - S)\ G_K(S-1) + (1+\epsilon)\, S\, g_K(S)$$

$$E[B(B-1)] = [n\,(n+1)\,\epsilon^2 - 2\,n\,\epsilon\, S + S\,(S+1)]\ G_K(S-1) + [(n+1)\,\epsilon - (S+1)]\,(1+\epsilon)\, S\, g_K(S)$$

We use g_K and G_K, that we already know, to evaluate these equations. We substitute E[B] and E[B(B-1)] into the corresponding equations given before to find E[D] and $E[D^2]$ and, then, V[D].

The linkage between stages are provided by the following equations;

$$E[L_i] = E[D_{i-1}] + E[T_i]$$

$$V[L_i] = V[D_{i-1}] + V[T_i]$$

For the first stage ;

$$E[L_1] = E[T_1]$$

$$V[L_1] = V[T_1]$$

since an infinite supply is assumed before the first stage and, hence, there is no delay before the first stage.

Starting with the first stage, we can calculate the $E[D_1]$ and $V[D_1]$ using the method described since we know $E[L_1]$ and $V[L_1]$. Given $E[D_1]$ and $V[D_1]$, we can find $E[L_2]$ and $V[L_2]$. Using $E[L_2]$ and $V[L_2]$ we calculate $E[D_2]$ and $V[D_2]$. We iterate in this fashion until the last stage where we finally get the values of $E[K_m]$ and $E[B_m]$, where m is the number of stages. These values are plugged into the following equation that yields the fill rate (FR), our system performance measure, as a percentage :

$$FR = [1- \frac{E[B_m]}{E[K_m]}] \text{ x}100$$

This equation gives us the fill rate since $E[K_m]$ is the expected number of outstanding orders in steady state at the last stage and $E[B_m]$ is the backordered portion of $E[K_m]$ that is not satisfied directly from inventory.

The main advantage of this two-moment approximation method is that it only requires first two moments of the transit times to be known (or estimated). How good the approximation performs is entirely dependent on how well a gamma distribution represent the lead times in the system.

4.5 Addition of yield factor to the model

We have to modify the two-moment approximation model since our problem description additionally involves yield issue. Yield was assumed to be a unit-by-unit binomial process in the problem description section. That is, a unit can turn out to be good or bad after processing with the probabilities of p and (1-p) respectively and there is no correlation among units in terms of these probabilities. We assume these probabilities are stage dependent.

If we consider a single stage, the number of units to be processed until we obtain a good unit is, excluding the good unit, negative binomial variable by definition. Let y_i denote the probability of obtaining a good unit and N the number of units to be processed to get a good unit. The probability equation of negative binomial variable (X) is

$$P(X = x) = \binom{x+k-1}{k-1} p^k (1-p)^x$$

where x is the number of failures before the kth success in Bernoulli trials with success probability p and failure probability (1-p). Mean and variance of X are

$$E[X] = \frac{k(1-p)}{p}$$
$$V[X] = \frac{k(1-p)}{p^2}$$

In our case, $p = y_i$ and k=1. Using these equations, we find E[N] and V[N] as follows ;

$$E[N] = \frac{(1-y_i)}{y_i} + 1 = \frac{1}{y_i}$$

The first term in the equation above is the expected number of bad units before encountering a good one and plus one is for the good unit and the variance of N is ;

$$V[N] = \frac{(1-y_i)}{y_i^2}$$

The time that it takes for a good unit to pass through a stage is the summation of N transit times. This is a summation of a random number of random variables. Hence, the expected time that it takes for a good unit to pass through a stage (expected adjusted transit time ,E[AT]) is

$$\begin{aligned} E[AT] &= E[N]\ \ E[T] \\ &= \tfrac{1}{y_i} E[T] \end{aligned}$$

and the variance of AT is

$$V[AT] = V[T]\ E[N] + V[N]\ \left(E[T]\right)^2$$

This variance is analogous to the variance of lead time demand in inventory models where both lead time and demand are random variables. Here, lead time corresponds to N and demand to transit time.

By replacing the transit time with the adjusted transit time in the model we take the effect of yield on the transit times into consideration. Another modification we need to make is to adjust the demand on the inventories of each stage. Since there is a yield problem, the demand arising at an inventory is the original demand divided by the multiplication of yield factors of all down stream stages. Following is the expression that gives this demand value for the inventory of stage i ;

$$\lambda_i = \frac{\lambda}{\prod\limits_{k=i+1}^{m} y_k}$$

This demand value is basically the required input to the binomial yield processes, through the down stream stages, whose expected output is equal to the original demand, λ.

Chapter 5

A Search Heuristic To Find "Good" Base-Stock Policy Parameters

5.1 Introduction

Base-stock control policy model we presented in the previous section is a highly nonlinear one due to many random variables, such as inventory levels, backorders, number of outstanding orders, involved in the model and the interaction of these variables. Our attempts at solving this nonlinear model using LINGO, a linear and nonlinear model solver, yielded no results because of the inability of the software to deal with this nonlinear system. It is not difficult to guess that there would not be many solvers that can solve this model, if there is any at all. We think that troublesome complexity associated with the model is due to the constraints that enforce a certain fill rate involve both cumulative complementary distribution functions and density functions of the negative binomial variables. The objective also involves these functions. Therefore, we resort to developing a search heuristic for the solution. Solution means setting the base sock levels, the decisions variables, to certain values so that we minimize the system inventory holding cost as we keep the fill rate above a predetermined level.

The search heuristic involves two main parts. The first part initializes the system by setting the base-stock levels so high that there wont be any delay while the second part tries to reduce these levels.

5.2 Phase one : Initialization

Phase one of the heuristic simply finds a set of base-stock levels as a starting point for phase two. This starting point is the set of base-stock levels that results in zero delay time at every inventory in the system. In another words, in this part of the heuristic the base-stock levels are set so high that all demand at any inventory in the system is satisfied directly from the inventory with a very high probability. This means lead time for a stage will be equal to the transit time of the stage since there would be no delay time. Fill rate will obviously be very close to 100 % after applying the first part of the heuristic.

Initialization of the system is accomplished stage by stage, starting with the first stage, by increasing the base-stock levels gradually until the delay associated with the inventory of the stage becomes approximately zero. The delay associated with the previous inventory is set to zero as we carry out the calculations for the current stage since it will be approximately zero after initialization.

5.3 Phase Two: Search

Phase two of the heuristic attempts to reduce the base-stock policy parameters, systematically, in order to decrease the expected system inventory holding cost while it obeys the restriction that the fill rate must be greater than or equal to a predetermined value.

In this second part of the heuristic, at each iteration, the associated system inventory holding cost per percent (INCPP) of the fill rate is found as follows

$$\text{INCPP} = \frac{\text{CURRENT SYSTEM INVENTORY HOLDING COST}}{\text{CURRENT FILL RATE}}$$

Then, each base-stock level is reduced, in turn, by a certain step size and the *gain* in terms of the expected system inventory holding cost is calculated for each reduction. Since each reduction causes the fill rate to drop, also calculated is the *loss* due to each reduction. This loss is defined as follows

$$\text{LOSS} = \text{INCPP x AMOUNT OF PERCENT DROP IN THE FILL RATE}$$

The reduction that has the highest gain to loss ratio and that does not violate the fill rate restriction is carried out. For a certain reduction step size, the search procedure iterates until there is no reduction that yields a fill rate greater than the required one. At this point, the step size is reduced and the iterations restart with the new step size. The search procedure stops when the step size is one and there is no possible reduction of base-stock policy parameters. The flow charts of the first and second phase of the heuristic are given below.

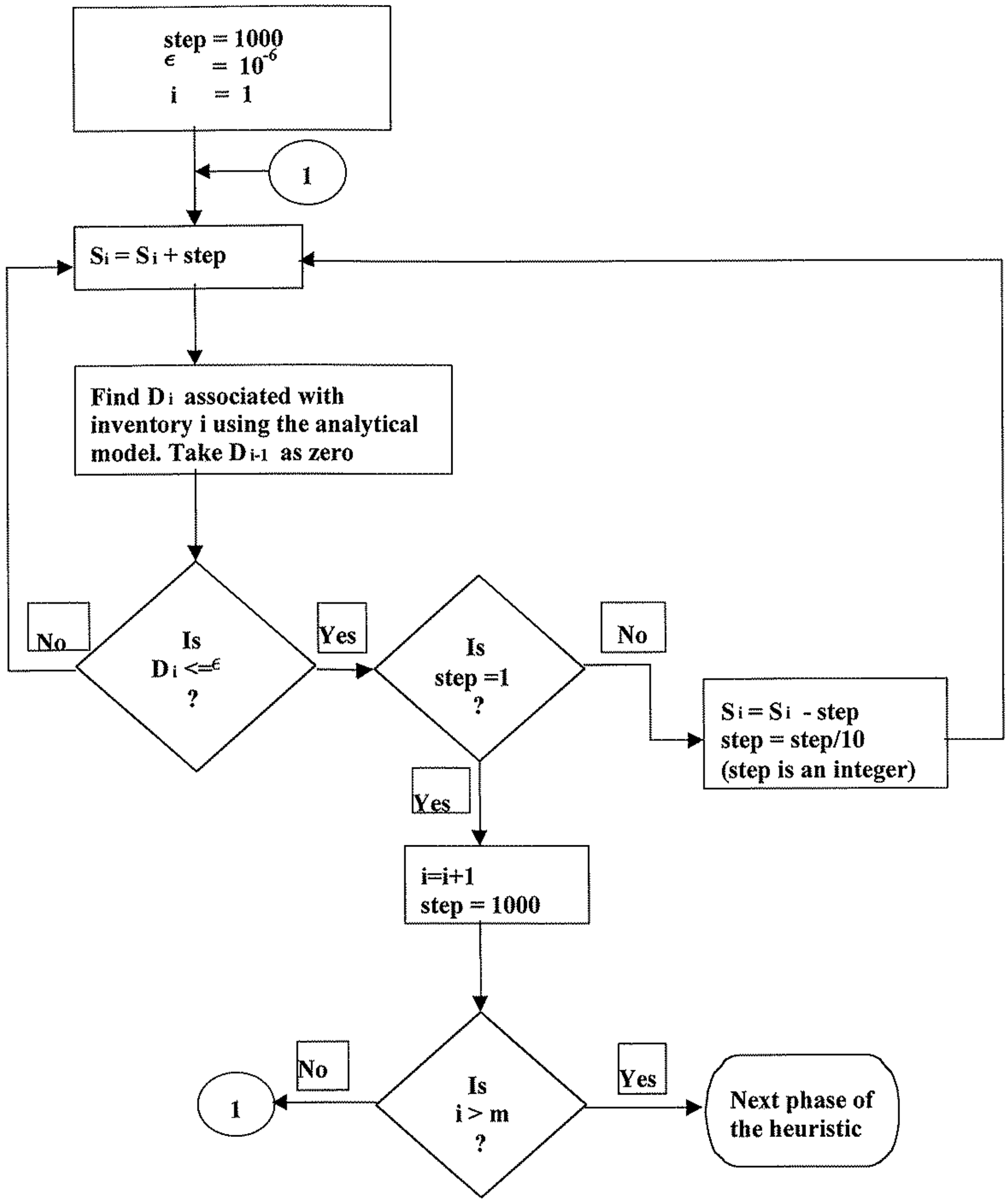
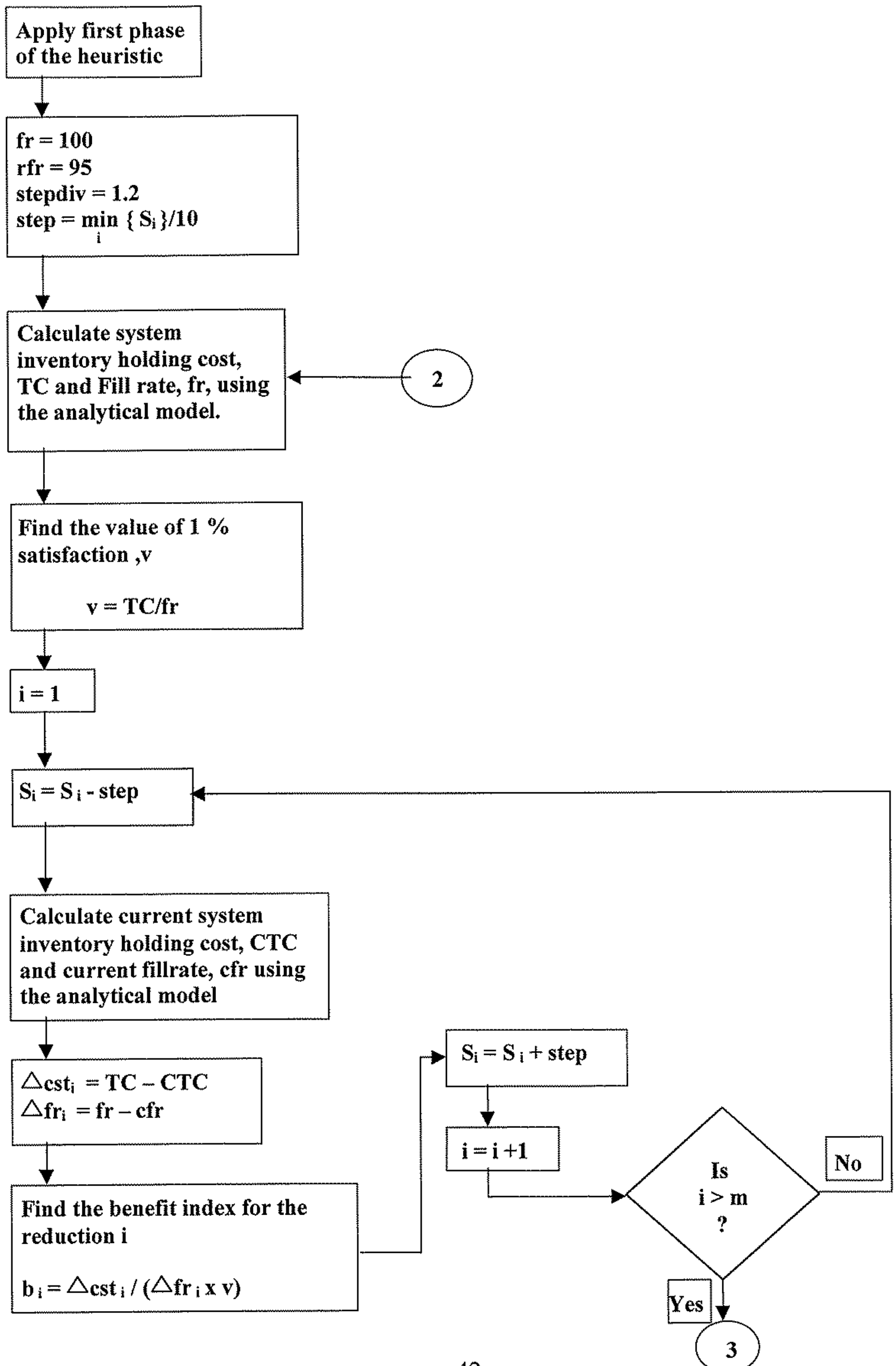


Figure 3-1 Flow chart of the phase one of the heuristic.


Apply first phase of the heuristic
fr = 100
rfr = 95
stepdiv = 1.2
step = min_i { S_i }/10
Calculate system inventory holding cost, TC and Fill rate, fr, using the analytical model.
2
Find the value of 1 % satisfaction ,v
v = TC/fr
i = 1
S_i = S_i - step
Calculate current system inventory holding cost, CTC and current fillrate, cfr using the analytical model
△cst_i = TC – CTC
△fr_i = fr – cfr
Find the benefit index for the reduction i
b_i = △cst_i / (△fr_i x v)
S_i = S_i + step
i = i +1
Is
i > m
?
No
Yes
3

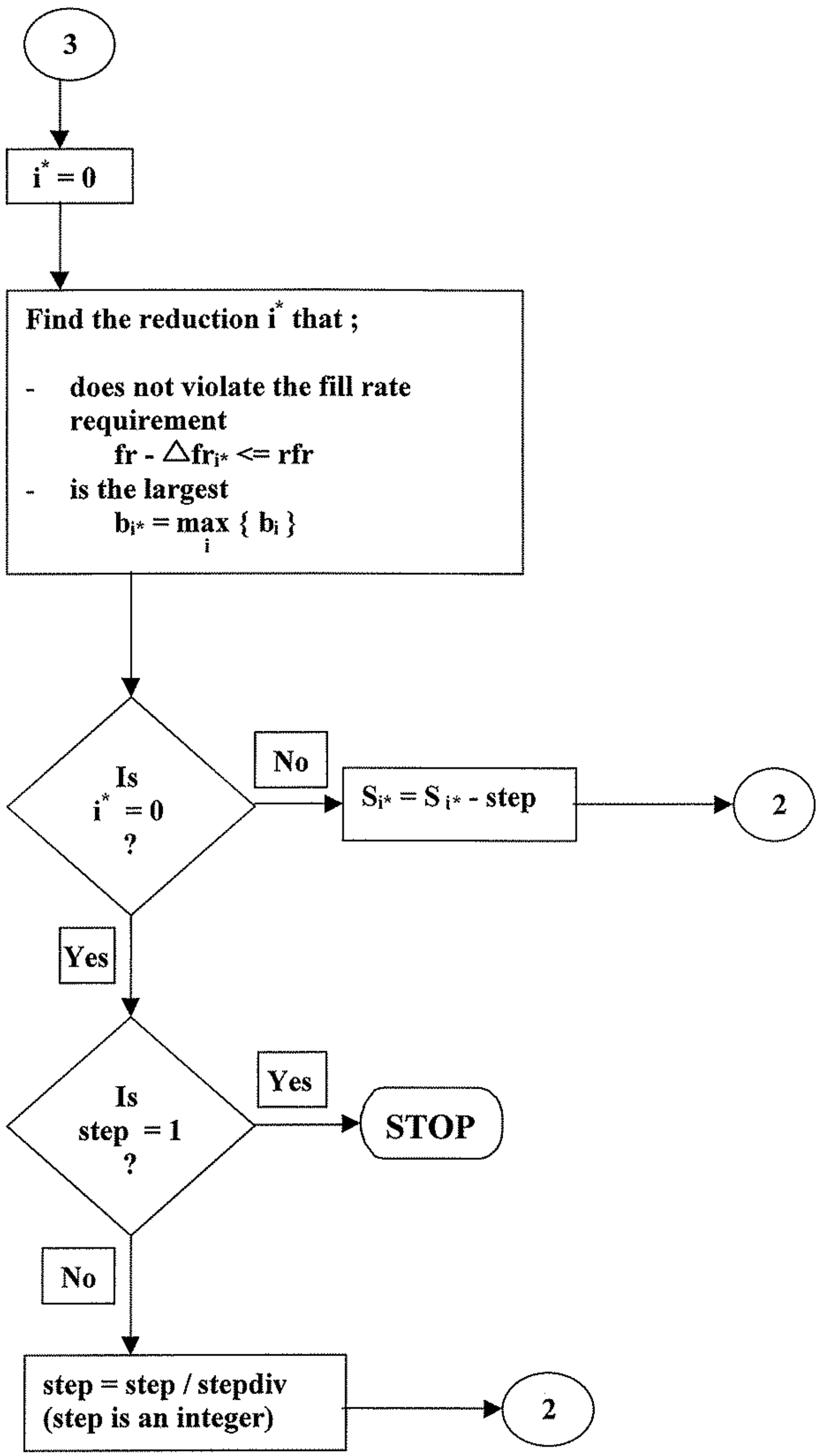


Figure 3-2 Flow chart of the phase two of the heuristic.

Notation used in the flow charts is explained as follows;

i ; Index for the stages of the system.

m ; Number of stages.

S_i ; Base-stock level at stage i.

D_i ; Delay associated with the inventory of stage i.

fr ; Fill rate of the system.

rfr ; Required fill rate.

crf ; Current fill rate of the system after the reduction of a base-stock level.

TC ; Total system inventory holding cost.

CTC ; Current total system inventory holding cost after the reduction of a base-stock level.

$\triangle cst_i$; Reduction in total system inventory holding cost due to the decrease in base-stock level i.

$\triangle fr_i$; Reduction in fill rate of the system due to the decrease in base-stock level i.

stepdiv ; step division factor which is used to divide the step size in order to get the new step size.

Chapter 6

Experiments And Simulation :

6.1 Experimentation With The Analytical Model

We have coded the analytical model and the heuristic search algorithm that uses the analytical model in Fortran 77. We have used Fortran 77 so that we could include the binomial probability subroutine from the IMSL library of Fortran 77. This subroutine returns a probability from the Binomial distribution for any given set of distribution parameters. We also have coded a small program in C to produce random problems for a four-stage system where the closest stage to the infinite supply is labeled as stage 1.

We have produced 20 random problems using our small program. Each problem instance includes demand rate and, for each stage, alpha and beta values that determine the transit time, unit inventory holding cost and yield rate. We fixed the range [3-7] for demand rate, [2-5] for alfa, [2-4] for beta and [.65-.95] for yield rate. Unit inventory holding cost for stage 1 ranges from 5 to 30 and inventory cost for the following stages are found by adding a value in the range [1-25] to the previous stage unit inventory holding cost. Table 4-1 shows these problems.

We have run the program for these random problems. The fill rate is fixed at 95 % for all the problems. Output of the program is the base-stock levels for each stage that are approximately optimized using the search heuristic described earlier, expected inventory and backorder levels at each stage and total system inventory holding cost. We compare these results with the simulation results in the next section.

Problem	Demand	Alpha	Beta	Unit Inv. Cost	Yield Rate (%)
1	3	4 4 3 3	4 3 4 3	29 44 68 77	88 82 76 86
2	6	4 3 2 2	3 4 3 2	18 30 49 69	93 90 82 85
3	7	4 3 2 2	4 2 2 4	19 49 58 79	71 78 71 73
4	6	4 5 4 2	3 4 2 3	6 20 26 53	67 89 89 88
5	5	3 2 3 2	3 4 3 4	12 38 65 83	91 81 82 75
6	7	3 4 2 2	4 3 4 4	26 52 78 84	74 85 73 71
7	4	4 5 3 4	3 2 2 4	28 33 48 61	66 65 91 82
8	5	5 3 3 3	3 2 4 2	18 23 48 71	81 76 85 78
9	5	2 4 4 5	3 2 2 4	13 20 45 51	66 87 84 82
10	6	2 5 3 2	2 4 3 3	24 43 58 72	72 69 75 91
11	3	4 5 2 3	2 3 2 3	12 26 55 61	70 82 79 72
12	4	3 5 3 2	3 3 4 3	30 47 77 87	73 73 81 72
13	3	5 4 2 3	2 3 3 4	16 26 31 42	71 72 83 68
14	7	4 4 3 3	3 2 2 2	28 53 71 99	89 85 73 90
15	7	2 4 5 4	3 4 2 4	19 26 34 46	80 79 91 85
16	3	3 3 2 5	2 3 3 2	25 43 57 82	80 92 85 81
17	7	5 4 3 2	3 4 2 4	18 41 47 61	86 72 90 86
18	7	4 3 5 3	3 3 3 4	16 30 37 44	75 89 79 69
19	4	5 4 3 5	2 2 4 2	23 52 82 99	71 75 94 81
20	7	5 4 5 4	4 2 2 3	13 35 63 80	70 94 70 90

Table 4-1 Data of 20 random problems

6.2 Simulation Comparisons

We have build up a simulation model of the base-stock inventory policy for a four-stage serial system using SIGMA (Schruben, L.W. (1995)), a system simulation software. We have simulated the system using the data set of each problem for 7500 time units and calculated the average inventory levels for each stage. In the following table we give the results from the analytical model and the simulation experiments.

	Exp. last Stage Inventory			System inventory holding cost		
Problem	Anal. Mod.	Simulation	% error	Anal. Mod.	Simulation	% error
1	48.80	46.40	5.18	24887.62	23778.68	4.66
2	64.73	63.12	2.55	71760.52	71742.03	0.03
3	184.70	178.99	3.19	138040.40	137542.80	0.36
4	89.96	87.14	3.24	43322.45	40551.22	6.78
5	127.45	123.64	3.09	85837.09	85563.04	0.32
6	191.72	186.57	2.76	191042.40	190693.50	0.18
7	115.02	108.85	5.67	76854.78	7640.32	0.54
8	59.63	57.85	3.08	26409.42	26273.53	0.52
9	167.49	159.77	4.83	41439.79	41092.49	0.85
10	57.42	53.26	7.81	20355	19937.78	2.09
11	110.03	112.62	2.30	121638.80	121544.30	0.08
12	77.10	74.36	3.70	41833.66	41621.33	0.51
13	82.04	79.22	3.56	137742.40	137360.20	0.28
14	113.11	110.83	2.59	49890.82	49795.87	0.19
15	69.92	65.82	6.22	48049.30	47622.34	0.90
16	186.53	176.73	5.56	63721.33	63306.31	0.66
17	53.83	50.94	5.68	32710.17	32458.19	0.78
18	144.40	139.55	3.48	114448.80	114151.60	0.26
19	252.80	243.47	3.83	102758.90	102237.60	0.51
20	70.49	66.23	6.43	91763.70	91261.66	0.55
		Average error	4.237		Average error	1.052

Table 4-2 Results from analytical and simulation models

We conservatively fixed the simulation warm-up period at 750 time units for each problem based on our observation of the output plots of different problems. We selected the inventory level of the last stage and the system inventory holding cost as the criterion to compare the analytical results with simulation results since these two are the most important and most representative measures in terms of approximation ability of the analytical model.

For all the problems, fill rate of the simulation experiment was greater than required 95%. The error values in the table are given by

$$\% \text{ error} = \frac{100*|\text{Simulation result-Analytical result}|}{\text{Simulation result}}$$

These results suggest that the analytical model yields good approximations, especially for system inventory holding cost. For all the problems, the analytical model slightly overestimated both the expected last stage inventory and system inventory holding cost. This could be because of an overestimation of the first two moments of transit times in the analytical model when we adjust them to account for the yield effect.

The analytical approximation was relatively poor for the problems where the search heuristic yielded high expected backorder levels relative to base-stock levels in some stages. This is understandable because high backorder levels lead to non-gamma lead time distributions and the analytical model is based on the assumption that the lead time distribution can be closely approximated by a gamma distribution. As a remedy we can force the backorder not to be too high for any stage while we are doing the search. This will smooth out the backorders across stages and help the analytical approximation improve. Moreover, since in a production environment, the required fill rate generally is expected to be greater then 95%, the chance of finding solutions with high backorder levels at any stage will be small.

6.3 An Improvement To The Search Heuristic

In the phase two of the heuristic, we select the reduction that has the highest gain to loss ratio, as we defined earlier. To add in some diversity to the search such that the procedure is less likely to settle for an inferior local optimal solution, we suggest random selection among feasible reductions. This selection, however, is not completely random. We give weights to the candidate feasible reductions proportional to their respective gain to loss ratios so that the larger the gain to loss ratio of a reduction is, the better chance it has to be selected. We include this modified version of the code in the appendix as well.

For any problem, we iterate a predefined number of times and at each iteration we find a solution to the problem. At the end of the iterations we get the output as the minimum cost solution among the solutions produced. This means the improvement comes along with a computational cost.

We have modified the code of the search heuristic and run this improved version of the heuristic for seven problems among the problems used before. We set the number of iterations to 20 for all the problems. The following table shows system inventory holding cost from the original and the improved versions of the search heuristic.

Problem	Previous result	Improved result	% improvement
2	71760.52	70680.52	1.5
3	138040.4	134712.5	2.4
5	85837.09	82724.17	3.6
7	76854.78	69166.76	10
8	26409.42	22943.72	13.1
12	41833.66	38439.65	8.1
13	137742.4	137692.4	.04
		Average	5.53

Table 4-3 Improved results

The result table basically shows that the improved version of the heuristic has the potential to give better results. It is not difficult to see that the improvement can increase as we increase the number of iterations which translates into computational cost. For example, in our experiment, the computational cost of the improved version of the heuristic is 20 times that of the original heuristic since we set the number of iterations to 20 in the improved version.

6.4 Conclusion

The problem of the determination of a base stock inventory control policy for a production system with several stochastic features is a very complex one due to the many related random variables and their interaction. In this report, we have tried to simplify some of the decisions involved in an approach to identify base stock inventory control policies for serial production systems.

Our approach provides decisions that one needs to make in a realistic setting when determining an inventory control policy in a production system: lot sizing decision, modeling an inventory policy and fixing the parameters of an inventory policy to yield cost effective results.

We have used published models as a basis for the first two parts of our approach, after some modifications and additions to tailor them to our problem. As part our approach, we have developed a heuristic search method based on marginal benefit analysis to set base stock levels in the new model. The analytical model provided excellent approximation results when compared with those of a simulation model.

REFERENCES :

Bardinelli, D.R., (1992), " A model for continuous-review pull policies in serial inventory systems", *Operations Research* 40, 142-156.

Buzacott, J.A. and Shanthikumar, (1993), J.G., Stochastics Models of Manufacturing Systems, Prentice Hall, Englewood Cliffs, NJ .

Clark, A.J. and Scarf, H., (1960), "Optimal policies for multi-echelon inventory problem", *Management Science* 6, 475-490.

Clark, A.J. and Scarf, H., (1962), "Approximate solutions to a simple multiechelon inventory problem", Studies in Applied Probability and Management Science, 88-110, Arrow, K.J., Karlin, S., and Scarf, H. (eds.), Stanford University Press, Stanford, CA.

Debodt, M.A. and Graves, S.C., (1985), "Continuous review policies for a multi-echelon inventory problem with stochastic demand", *Management Science* 31, 1286-1299.

Deuermeyer, B.L. and Schwartz, L.B., (1981), "A model for the analysis of system service level in warehouse-retailer distribution systems : The identical retailer case", Multi-level production/inventory control systems: Theory and practice, North Holland, Amsterdam.

Federgruen, A. and Zipkin, A., (1984a), "Approximation of dynamic multilocation production and inventory problems", *Management Science* 30, 69-84.

Federgruen, A. and Zipkin, A., (1984b), "Computational issues in an infinite horizon multiechelon inventory model', *Operations Research* 32, 818-836.

Federgruen, A. and Zipkin, A., (1986a), "An inventory model with limited production capacity and uncertain demands, I: The average cost criterion", *Mathematics of Operations Research* 11, 193-207.

Federgruen, A. and Zipkin, A., (1986b), "An inventory model with limited production capacity and uncertain demands, I: The discounted cost criterion", *Mathematics of Operations Research* 11, 208-215.
Glasserman, P., (1997), "Bounds and asymptotic for planning safety stocks", *Operations Research* 45, 244-257.

Glasserman, P. and Tayur, S., (1996), "A simple approximation for a multistage capacitated production-inventory system", *Naval Research Logistics* 43, 41-58.

Graves, S.C. and Schwartz, L.B., (1977), "Single cycle continuous review policies for arborescent production/inventory systems", *Management Science* 23, 529-541.

Hadley, G. and Whitin, T.M., (1963), "Analysis of inventory systems", Prentice Hall Englewood Cliffs, NJ.

Johnson, H. and Silver A.E., (1987), "Analysis of a two echelon inventory control system with complete redistribution", *Management Science* 33, 215-227.

Karmarkar, U.S., Kekre, S. and Kekre, S., (1985), "Lot sizing in multi item, multi machine job shops", *IIE Transactions* 17, 290-298.

Karmarkar, U.S., Kekre, S. and Kekre, S., (1992), "Multi-item batching heuristics for minimization of queueing delays", *European Journal of Operational Research* 58, 99-111.

Kyung, S.P. and Dae, H.K., (1987), "Congruential inventory model for two echelon distribution systems", *Operational Research Society* 38, 643-650.

Muckstadt, J.A. and Roundy, R.O., (1987), "Multi item one-warehouse, multi-retailer distribution system", *Management Science* 32, 1613-1620.

Rummel, J.L., (1989), "Cost models for batching decisions", Unpublished Ph.D. Thesis, William E. Simson Graduate school of Business Administration, University of Rochester, Rochester, N.Y.

Rosenaum, B.A., (1981), "Inventory placement in two echelon inventory system : An application", Multi-level production/inventory control systems: Theory and practice, North Holland, Amsterdam.

Roundy, R.O., (1985), "98% Effective integer ratio lot-sizing for one-warehouse muti-retailer systems", *Management Science* 31, 1416-1429.

Sand, G., (1981), "Predicting demand on the secondary echelon : A case study", Multi-level production/inventory control systems: Theory and practice, North Holland, Amsterdam, 1981.

Schruben, L.W., (1995), Graphical Simulation Modelling and Analysis, The Scientific Press Series, Boyd & Fraser Publishing Company, Danvers , MA.

Schwartz, L.B., (1973), "A simple continuous review deterministic one-warehouse n-retailer inventory problem", *Management Science* 19, 555-565.

Silver, E.A. and Peterson, R., (1985), Decision Systems for Inventory Management and Production Planning, 2nd edition, John Wiley and Sons, N.Y..

Svoronos, A. and Zipkin, P., (1991), "Evaluation of one-for-one replenishment policies for multiechelon inventory systems", *Management Science* 37, 68-83.

Veatch, M.H. and Wein, L.M., (1994) "Optimal control of a two-station tandem production/inventory system", *Operations Research* 42, 337-350.

Williams, J.F., (1981), "Heuristic techniques for simultaneous scheduling of production and distribution in multi-echelon structure : Theory and practice", *Management Science* 27, 336-351.

Zipkin, P., (1991), "Evaluation of base-stock policies in multiechelon inventory systems with compound-poisson demands", *Naval Research Logistics* 38, 397-412.