MANUFACTURING DECISIONS UNDER UNCERTAINTY : MODELS AND METHODOLOGY

by

Sriram Dasu

B.Tech., Indian Institute of Technology (1980)

P.G.D.M.,Indian Institute of Management (1982)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS OF THE
DEGREE OF

DOCTOR OF PHILOSOPHY
IN MANAGEMENT
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September, 1988

Signature of Author ....
Sloan School of Management
September, 1988

Certified by ....
Gabriel R. Bitran
Thesis Supervisor

Accepted by ....
Chair, Doctoral Program Committee

MANUFACTURING DECISIONS UNDER UNCERTAINTY: MODELS AND METHODOLOGY

by

Sriram Dasu

Submitted to the Sloan School of Management
in September, 1988 in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy in Management

ABSTRACT

In this thesis we study two manufacturing problems. Both the problems were motivated by our interaction with a manufacturer of semi-conductor chips. In the first problem we are concerned with production planning in an environment of joint replenishment, variable yields, and substitutable demands. We formulate the planning problem as a profit maximizing convex stochastic program. Based on the structure of the problem we propose heuristics for managing inventories in this environment. We also propose and test approximation procedures for solving the stochastic program.

In the second part of this thesis we develop a decomposition procedure for analyzing networks of queues in which the interarrival times of all external arrivals to the network and, service distributions are of the phase type. For this purpose, we study the superposition of phase renewal processes, queues with non-renewal arrivals and the departure process from queues. We also propose and test non-renewal approximations for the superposition of renewal processes.

Thesis Supervisor: Dr. Gabriel R. Bitran

Title: Professor of Management Science

Acknowledgements

I am extremely grateful to Professor Gabriel Bitran, for his invaluable guidance, support and encouragement. I am indebted to him for the interest he took in my progress at M.I.T. I have benefitted immensely from his knowledge of operations management, and his ability to identify and structure problems that are of significance to managers. I am touched by the genuine concern he has for the welfare of his students.

I would like to thank Professor Charles Fine, Professor Lode Li and Professor Thomas Magnanti for serving on my thesis committee. I am grateful to Professor Thomas Magnanti and Professor Robert Freund for developing my interest in operations research. I am also grateful to Professor S.C. Graves for his advice and support.

I thank my good friends Anna, Devanath, G.K., Prakash, Raman and Thin-Yin for all their help. I am grateful to Rama for her love, care, tolerance and understanding. I could not have written this thesis without her support.

I am grateful to M/A- Com for allowing me to study their operations, and for their support. I am especially grateful to Dr Demir Zoroglu.

To My Parents

&

Rama

TABLE OF CONTENTS

I. INTRODUCTION

In this thesis we study two problems. Both problems were motivated by our study of a semi-conductor manufacturing operation. In this industry, products and technology change rapidly. The production process is complex and often not very well understood. As a result, there is high variability in production yields.

Semi-conductor chips are produced in wafers. Each wafer may contain several thousand chips. In the facility that motivated this research, it was not uncommon for chips from the same wafer to exhibit different electrical properties. The electrical properties of a chip determine its end use. Often, a chip that is not useful for its intended use may be suitable for some other application. In other words demands are substitutable. Further, the mix of chips obtained from a wafer varied from one production lot to the next. As a result we have a production system in which the yields are variable, demands are substitutable and chips are jointly replenished.

In the first part of the thesis we develop a mathematical model to aid managers in deciding (a) how many wafers to produce each period, and (b) how to allocate chip inventories to end-users. We formulate the problem as a convex stochastic program. Unfortunately, the size of this program grows very rapidly, therefore, we propose approximation procedures for solving the stochastic program, and heuristics for managing the chip inventories.

The wafer fabrication facility is shared by many product families. This facility contains many machines which are very expensive and several product families compete for these resources. As a result we observe queues of production lots in front of these machines. One of the factors managers consider while determining the appropriate capacity levels, is the trade-off between work-in-process costs, lead times and cost of additional capacity. To support this decision process, we need models that provide some insight into the long term behavior of the shop. In this context, queueing network models of the wafer fabrication facility have been useful (Bitran and Tirupati 86). In the second part of this thesis we develop methodology for analyzing networks of queues.

In general, queueing networks are difficult to analyze. However, over the past decade researchers have successfully developed accurate, yet simple approximations procedures for evaluating the performance of these systems - the parametric decomposition method (Kuehn 79, Shantikumar and Buzacott 81, Whitt 83, Bitran and Tirupati 86). These procedures have primarily focused on the average work-in-process levels in the queueing net work.

At the same time, another group of researchers (Neuts 81, Lucantoni and Ramaswami 85) have furthered the theory for analyzing queues with a class of distributions called phase type distributions. This class of distributions is quite versatile; exponential, hyper-exponential, Erlangs and mixture of Erlangs are special cases of phase distributions.

In the second chapter of our thesis we bring these two bodies of literature together. We show how the theory of phases can be incorporated into the decomposition framework. As a result of this marriage we are able to compute performance measures such as the variance of the number in queue, distribution of the number in queue and waiting time moments. In addition, the phase methodology enables us to get a better understanding of the basic processes arising in queueing networks, such as superposition of renewal processes and departure streams.

II. ORDERING POLICIES IN AN ENVIRONMENT OF STOCHASTIC YIELDS AND SUBSTITUTABLE DEMANDS

1. INTRODUCTION

In almost all manufacturing environments the quality of the output is determined by whether or not the product meets the customer's requirements. In some cases a defective good can be repaired and in other cases it may have to be discarded. There are many situations where a product which does not meet its specification may be appropriate for some other use. For instance in the electronic industry the quality of capacitors and resistors is determined by their tolerance. Another example is in the metal cutting industry. The usage that an electronic connector finds, is in part determined by the precision of its threads and the accuracy of the metal plating. This phenomena is quite common in the semi-conductor industry. Frequently from one wafer (or lot) several grades of circuits are obtained.

In this paper we are concerned with situations where there is a hierarchy in the grades of outputs. We assume that a higher grade product can be substituted for a product lower in the hierarchy. Under these circumstances the manufacturer may occasionally down-grade a product rather than back-order the demand for a lower grade item . This type of action may be motivated by a variety of reasons, for example - to prevent customer dissatisfaction ; to reduce set-up costs; or to reduce inventory costs.

We are interested in identifying appropriate ordering policies in an environment of stochastic yields and substitutable demands. Our interest in this problem resulted from a specific application in the semi-conductor industry. For the sake of concreteness we will describe this application in greater detail. Although the discussion is motivated by a specific application, a wide variety of inventory - production problems have a similar structure. This claim is particularly applicable to the semi-conductor industry (Leachman).

In the semi-conductor industry, products and manufacturing technologies change rapidly. The manufacturing process is complex and often not very well understood. Consequently yields vary significantly. Decision aids that facilitate inventory management in this uncertain environment can play an important role in improving profitability.

Semi-conductor chips are produced in wafers. Each wafer contains several chips. The number of chips produced per wafer depends on the complexity of the circuit on the chip, and can vary from 10 to 100,000. The production of wafers is essentially a flow process involving several steps. To start, disks (wafers) of either Silicon or Gallium Arsenide are cut from ingots. Next, several layers of semi-conducting material are deposited on top of the wafer. At the end of the fabrication process wafers are cut into individual chips - also called dice - and delivered to the assembly stage. The chip is then assembled into a package. For a more detailed description of the production process the reader is referred to Kothari (84) and Bitran and Tirupati (88).

The company we are studying manufactures devices for both military and commercial applications. The particular facility that motivated this study produces diodes (an electronic valve). It is perhaps the largest facility of its kind in the world. Each wafer yields approximately 5000 diodes. Typically, diodes from the same wafer exhibit different electrical properties. The application (end-use) that a diode finds depends on its electrical properties. The number of diodes of each type obtained from a wafer is a random variable. The distribution of these random variables depends on the process employed to produce the wafer. A diode that does not meet its intended specification can often be useful for some other application.

In this facility diodes are produced for approximately 100 different applications. These applications can be grouped into 12 families. Within each family there is a hierarchy. If the electrical performance of a diode meets the requirements of an application then it satisfies the requirements of all applications lower in the hierarchy. Thus the demands are substitutable in this hierarchical sense.

In a general context considered in this paper we have (i) processes, (ii) items and (iii) applications or customers. A process corresponds to the production process for a wafer, and an item corresponds to a diode. Associated with each process is a set of items that can be produced using that process. The actual yields are random variables. If we produce η units, then we get ηp_i units of item i. Where p_i - the fraction of type i items - is a random variable. The distribution of these random variables is determined by the production process employed.

We assume that customers can be grouped into families. Within each family there is a hierarchy such that an item that is suitable for a member in the family is also suitable for all members lower in the hierarchy. The relationship between production process, items and customers is illustrated by Fig. 1.

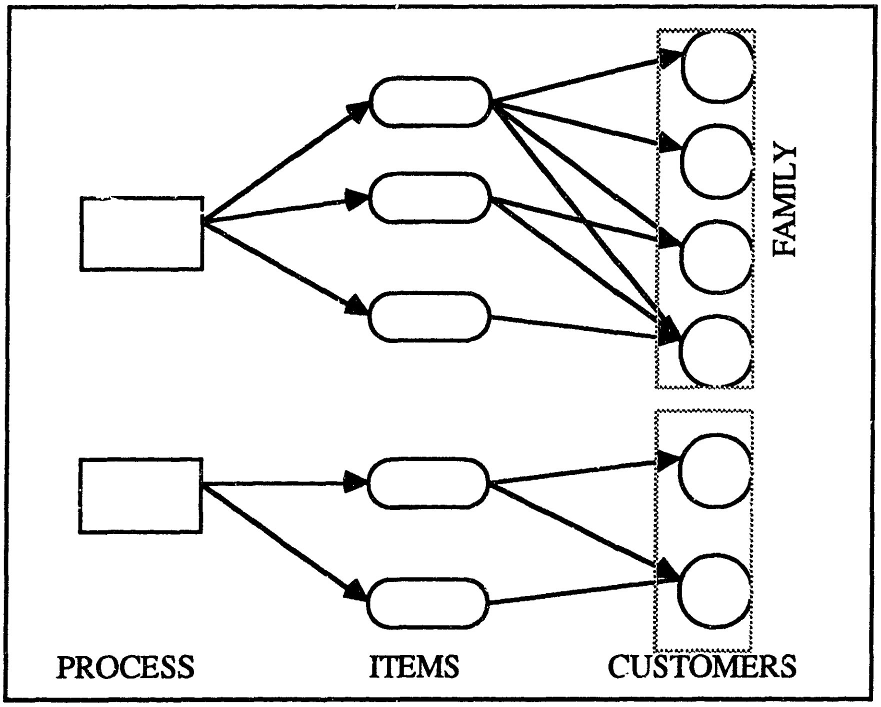


Figure 1. Relationship between Processes, Items and Customers

Our objective in this paper is to develop a model to assist managers in deciding how many units to produce in each period using each process . Once the yields are known the model must assist managers in deciding how to allocate the items to customers. One of our tasks is to develop an understanding of how to manage the item inventories.

We have formulated the problem as a profit maximizing convex stochastic program, and have developed approximation procedures for solving it. Based on the structure of a two period problem we propose heuristics for allocating items to customers in a multiperiod setting.

The structure of this paper is as follows. In section 2 we review literature dealing with production planning when yields are uncertain. In section 3 we describe in greater detail the main assumptions on which our model is based. We present the formulation in section 4. In sections 5 and 6 we look at the structure of the problem. We first show that it is a convex program. We then derive the structure of the optimal inventory allocation policy for a two period problem. Based on this result we propose heuristics for problems with longer horizons. In section 7 we explore computational procedure for solving the model. We begin by studying solution approaches for single family problems and finally indicate how to solve multifamily multiperiod problems with capacity constraints.

2. LITERATURE REVIEW

There is a vast body of literature dealing with inventory and production problems. An implicit assumption in most of the inventory models is that the yield is 100% or is deterministic and known. There are very few papers dealing with variability in yields. Vienott(60) provides an excellent review of early literature on basic lot-sizing models. Silver(76) derives the economic order quantity when the quantity received from the supplier does not match the quantity ordered. He permits the probability density of the quantity received to be a function of the quantity ordered. He then analyses two different cases
(i) where the standard deviation of the quantity received is independent of the order size
(ii) where the standard deviation of the quantity received is proportional to the order size. Under the assumption of constant deterministic demand the economic order quantity in both cases is shown to be dependent upon only the mean and standard deviation of the quantity received.

Kalro and Gohil(82) extend Silver's model to the case where the demand during the stock-out period is either partially or completely backordered. In both cases the results obtained are shown to be extensions of the well known results of the lot sizing problem with backordering. Shih(80) analyzed a single period inventory model with random demand, variable yield and no ordering cost.

Mazzola, McCoy, and Wagner(87) consider a multi-period problem. They derive an EOQ model when the production yield follows a binomial distribution and backlogging of demand is permitted. They test several heuristic applications of this lot size problem to discrete time problems. The most promising heuristics are based on adjusting deterministic lot-sizing policies as calculated either optimally using the Wagner-Whitin algorithm or approximately using the Silver-Meal heuristic. The quantity q_t produced in period t is such that Prob(# of good pieces $\geq Q_t \mid q_t$ produced) $\geq \alpha$. In this expression Q_t is the deterministic lot size and α is a parameter associated with the service level. They have tested these heuristics for different demand patterns and find them on an average to be within 2.8% of optimality.

Lee and Yano have modelled a multi-stage single period single product problem. They assume that the proportion of defective pieces produced at each stage is a random variable. The decision variables are the quantities to be produced at each stage. These decisions are to be made after you know the number of good pieces produced by the previous stage. At each stage you incur production and holding costs. Unsatisfied demand results in penalties in the form of back-order costs. They show that the cost incurred at each stage is a convex function of the quantity produced at that stage. Consequently the optimal policy at each stage is specified by a critical number. If the number of good pieces from the previous stage exceeds this critical number then the quantity produced at that stage is equal to the critical number . Else the quantity produced equals the yield of the previous stage.

The by-product problem has been studied by Pierskalla and Deuermeyer(78). They consider the control of a production system that consists of two processes that produce two products. Process A produces both products in some fixed (deterministic) proportion while process B produces only one product. They assume that the demands are random. They formulate the problem as a convex program and derive some properties of the optimal policy. They show that the decision space (inventory levels) can be divided into 4 regions depending on whether or not a production process is used.

There is a large body of literature on computational procedures for solving stochastic programs. One of the earliest works in this field is due to Dantzig(55). He posed a multi-time period stochastic program. Since then several strategies have been proposed for solving these problems. In general it is very difficult to obtain exact solutions. The emphasis has been on developing good approximations. In principle if the random variables are discrete, then the problem of finding the first time period decision variables to optimize the expected value of the objective function reduces to one of solving a large linear-program. If the random variables have a continuous distribution then a popular strategy is to approximate the continuous problem by a sequence of discrete problems.

Olsen(76) proved that the sequence of solutions of the discrete approximations would converge to the solution of the continuous problem under relatively mild conditions. Birge and Wets(86) provide a very good review of approximation schemes for stochastic optimization problems.

To the best of our knowledge no one has modelled a production process where the demands are substitutable and the yields are stochastic. In a sense our model generalizes the notion of 'defect' by permitting a number of grades of quality. We do restrict the analysis to a particular substitution structure. However, this structure is likely to be the most natural and common substitution structure.

3. MODELLING ASSUMPTIONS

In order to develop a mathematical model we had to make a few key assumptions regarding demand patterns and the nature of variability of yields. In this section we will state and provide some justification for our assumptions.

The customers for the diodes are manufacturers of electronic goods who specify their requirements over a horizon of 4 to 5 months. The delivery schedule specifies the quantity to be dispatched each week. Accordingly we assume that the demands are dynamic and deterministic. The uncertainties in the demand are small and we ignore them.

There are at least two models for yield variability. We call them additive and multiplicative models. In the additive model if we produce h units then the number of type i items we get is given by $hp_i + e_i$. Where p_i is a constant that depends on the process and e_i is a random variable. In the multiplicative model the yield is simply given by hp_i. Here p_i is a random variable. We, of course, require that the sum of p_is be less than or equal to one. We adapt the multiplicative model and assume a finite number of outcomes. This choice is consistent with the motivating example.

The objective of our model is to maximize expected profits. Customers higher in the hierarchy pay a higher price. The costs include production, holding and backorder costs. There are no set-up costs.

4. PROBLEM FORMULATION

For ease of presentation we first formulate and analyze a single period problem with zero lead time. We also ignore capacity constraints. We then extend our analysis to multi-family multi-period problems and indicate how to incorporate capacity constraints.

Since we start with no capacity constraint the problem separates by families. Therefore in this section we look at a single family problem. We further simplify the problem by assuming that there are only two customers in the family. We number the members of the family in descending order - the highest member of the family is numbered #1. We also assign numbers to items. The number of an item designates the highest member in the family that it is suitable for. Thus item #1 can be used for customer 1 or 2, whereas item #2 can be used only for customer 2.

We assume that production is initiated at the beginning of the day and is completed by evening and, then the yields are known. In the evening, items are allocated to customers. The overall objective of our problem is to maximize profits. We assume that customer 1 pays a higher price than customer 2. All surplus inventories can be sold at a salvage price. We incur backorder, inventory holding and production costs. There are no set-up costs. In this framework, in the morning we must decide the number of units to produce, and in the evening optimally allocate the items. We are now ready to formulate the single period single family problem.

DESCRIPTION OF VARIABLES:

A_{ij}	: The number of 'i' items allocated to customer 'j'. Since item 2 does not meet customer 1's specification we do not have A_{21}.
$B_{t,j}$	: Backorders of customer 'j' carried into period t
C_i	: Salvage price for item 'i'
$D_{t,j}$	: Demand from customer j in period t
$\mathbf{E}_p$	: Expectation with respect to the yields.
$I_{t,i}$	: Inventory of item 'i' carried into period t
K	: Production cost per unit
N	: Number of units produced
p_m (q_m)	: Fraction of item 1s (2s) - m th outcome of the random variables
S_j	: Price paid by customer j
β_j	: Backorder cost for customer 'j'
Bold	: Bold letters denote vectors (except for the expectation operator)

The periods are indexed backwards. The index indicates the number of periods remaining in the horizon. In this notation period 1 is the last period in the horizon, and period 2 is the penultimate period.

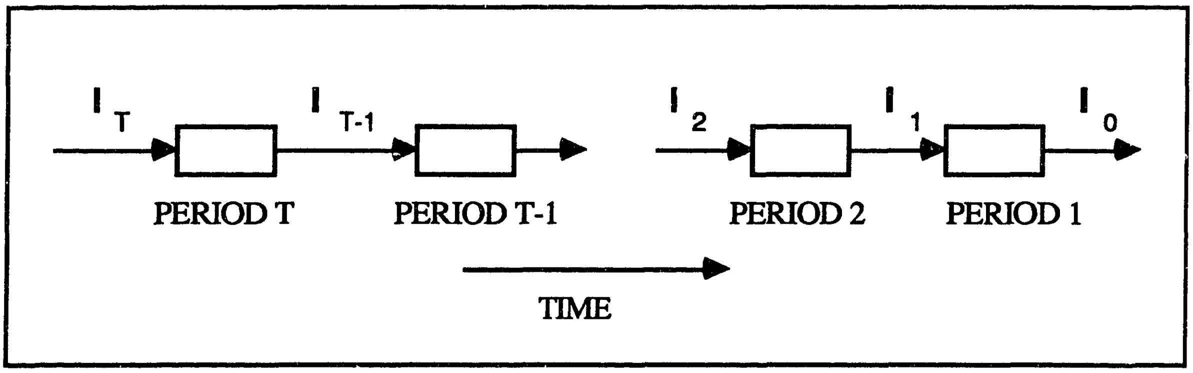


Figure 2. Indexing of Time

MORNING PROBLEM

$$F_1(I_1, B_1) = \text{Max } E_p\, G_1(p_m, N, I_1, B_1) - KN$$

S.T.

(1) $N \geq 0$

EVENING PROBLEM

$$G_1(p_m, N, I_1, B_1) = \text{Max } S_1A_{11} + S_2(A_{12}+A_{22}) + \sum_{i=1}^{2} C_i I_{0,i} - \sum_{j=1}^{2} \beta_j B_{0,j}$$

S.T.

(2) $A_{11} + A_{12} + I_{0,1} = Np_m + I_{1,1}$

ITEM INVENTORY

BALANCE

(3) $A_{22} + I_{0,2} = Nq_m + I_{1,2}$

(4) $A_{11} + B_{0,1} = D_{1,1} + B_{1,1}$

DEMAND CONSTRAINTS

(5) $A_{12} + A_{22} + B_{0,2} = D_{1,2} + B_{1,2}$

(6) $A_{11}, A_{12}, A_{22}, B_{0,1}, B_{0,2}, I_{0,1}, I_{0,2} \geq 0$ NON-NEGATIVITY

CONSTRAINTS

Constraints (2) and (3) are inventory balance constraints. The right hand side is the inventory of each item available at the end of the day. Recall that in the evening we know the quantity produced and the yields. Constraints (4) and (5) account for the customers' demands. The right side is the net demand and the left hand side shows how it is satisfied (or backordered).

5. CONVEXITY

We first show that the one period problem is a convex stochastic program.

Proposition 1: For any outcome $\mathbf{p_m}$ G($\mathbf{p_m}$,.) is a concave function of N, , $\mathbf{I_1}$, $\mathbf{B_1}$ and $\mathbf{D_1}$

Proof : G($\mathbf{p_m}$,.) is the right hand side parametric of a maximization problem with a concave objective function and linear constraints. Therefore G($\mathbf{p_m}$,.) is concave.

Proposition 2: The morning problem is a convex problem.

Proof: By proposition 1 G($\mathbf{p_m}$,.) is concave, implying that $\mathbf{E_p}$ G($\mathbf{p_m}$, .) is concave. We are therefore, maximizing a concave function over a convex set. •

Proposition 3: $F_1(\mathbf{I_1}, \mathbf{B_1})$ is a concave function.

Proof: Please see Lemma 1 in appendix

These propositions continue to be valid even if demand is not deterministic. The demand substitution structure can also be fairly general and we can impose capacity constraints. The structure, however, would be destroyed if set-up costs are significant and have to be incorporated. The propositions can also be extended to the multi-period problem.

In a multi-period problem the objective of the evening problem is changed to

$$S_1 A^{t+1}{}_{11} + S_1(A^{t+1}{}_{12} + A^{t+1}{}_{22}) - \sum_{i=1}^{2} h_i I_{t,i} - \sum_{j=1}^{2} \beta_j B_{t,j} + F_t(I_t, \mathbf{B}_t)$$

In this expression h_i is the holding cost for item i

To prove that $F_T(\mathbf{I}_T, \mathbf{B}_T)$ and $G_T(\mathbf{p}_m,N,\mathbf{I}_T, \mathbf{B}_T)$ are concave functions we use an inductive argument. Assume that $F_t(\mathbf{I}_t, \mathbf{B}_t)$ is concave, then by proposition 1 is $G_{t+1}(\mathbf{p}_m,N,\mathbf{I}_{t+1}, \mathbf{B}_{t+1})$. If $G_{t+1}(\mathbf{p}_m,N,\mathbf{I}_{t+1}, \mathbf{B}_{t+1})$ is concave, then by Lemma 1 $F_{t+1}(\mathbf{I}_{t+1}, \mathbf{B}_{t+1})$ is also concave. Since $F_1(\mathbf{I}_1, \mathbf{B}_1)$ is concave Propositions 1, 2,and 3 are valid for a multi-period problem.

6. ITEM ALLOCATION PROCESS

In a single period problem the allocation process is straight forward. For ease of exposition, assume that the starting inventories and back-orders are zero. Let $S_1+ \beta_1 \geq S_2 + \beta_2$, and $C_1 \geq C_2$. The first inequality ensures that the cost of foregoing a sale to customer 1 is greater than that for customer 2. The second inequality requires the salvage price of item 1 to be higher than that for item 2. We will also require $S_j + \beta_j \geq C_j$. Consequently it is suboptimal to backorder demand for customer j and hold inventories of item j. Under these conditions we will first allocate item j to customer j. If these yields are adequate to meet the demands then we do not have to allocate any item #1s to customer 2, and all excess inventories will be salvaged. The only situation in which we will consider allocating item #1to customer 2 is when item #1 yield exceeds customer 1's demand and item #2 yield is less than customer 2's demand. In this situation it will be optimal to down-grade item #1 if the salvage price C_1 is less than $S_2 + \beta_2$. Else, it is optimal to hold inventories of item #1 and backorder customer 2's demand. Henceforth we will assume that it is optimal to down-grade. The allocation process is illustrated by fig 3.

Figure 3 is useful for two reasons. We use it to develop a line search procedure for identifying the optimal production quantity. It is also be used to describe the down grading process in a two period problem. The x-axis in fig.3 is item 1 yield and the y-axis is item 2 yield. The space is divided into five regions. In each of the regions the objective value of the evening problem is linear. Of course, over the entire space the evening objective value is (piecewise linear) concave (Proposition 1).

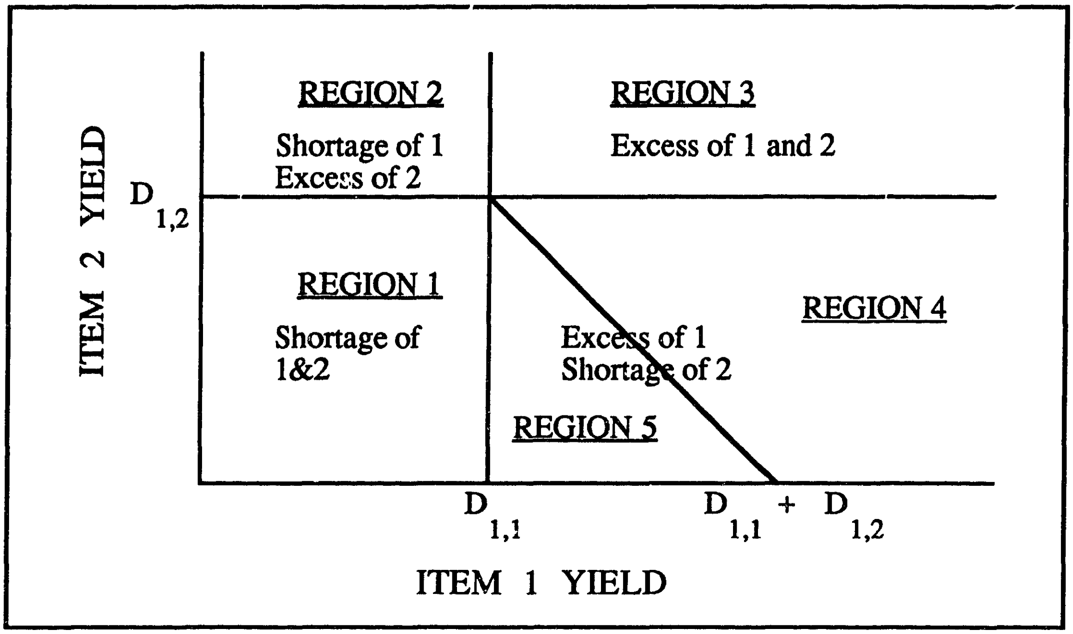


Figure 3. Item Allocation Process

In region 1 the demand for both the items exceeds the yields and therefore we do not downgrade item #1. In region 2 item #2 yield exceeds the demand from customer 2 and there is a shortage of item #1. In region 3 the yields for both the items exceeds the demands from their respective customers. In regions 4 and 5 we enjoy an excess of item #1 but the yield of item #2 is inadequate to satisfy customer 2's demand. In region 4 and 5, under the assumptions stated in the previous paragraph we will downgrade item #1. In region 4 the excess of item #1 yield is adequate to cover the shortfall in item #2 yield, as a result customer 2's demand will be satisfied. On the other hand in region 5, excess yield of item #1 is inadequate to meet the residual demand from customer 2. Hence, in region 5 some of customer 2's demand goes unsatisfied.

Next, consider the allocation process in a two period problem. Assume that the yields of the penultimate period are known and we have to make the allocation decisions. In our model, costs and sale prices do not change from period to period. Hence we will not back-order demand from customer j and hold inventories of item j. Again the only case that is interesting is when we have an excess of item 1 and a shortage of item 2. In a single period problem if it is optimal to down-grade then we down-grade until either customer 2's demand is fully satisfied, or we run out of item 1. In general, this is not the case in a two period problem. We may down-grade only some of the excess and start the final period with item 1 inventory and back-orders of customer 2.

Once again, assume that we start period 2 with zero inventories and back-orders. Let the the yields for item 1 be x and for item 2, y. Further let $x > D_{2,1}$ and $y < D_{2,2}$. Define $\varepsilon = x - D_{2,1}$, and $U = D_{2,2} - y$. Thus after we allocate item 1 to customer 1 and item 2 to customer 2 we will be left with ε units of item 1 and U units of unsatisfied customer 2 demand. At this point every unit of item 1 allocated to customer 2 will have the following consequences:

(1) Increase period 2 profits by $S_2 + ß_2 + h_1$

(2) Inventory of item 1 will decrease by 1 unit

(3) Backorders of customer 2 will decrease by 1 unit.

(4) As a result of (2) and (3), period 1 profit will change from $F_1(\varepsilon,0, 0,U)$ to $F_1(\varepsilon-1,0, 0, U-1)$.

Since $F_1(.,.)$ is a concave function it is optimal to down-grade so long as the marginal increase in the current periods' profits exceeds the marginal decrease in the final periods' profits. If we are to say anything more about the down-grading process it is necessary to understand the structure of $F_1(., .)$. Fortunately, it has an intuitively appealing structure. On the basis of this structure, we propose heuristics for managing the inventories in a multi-period problem.

In the next sub-section we study the structure of $F_1(., .)$. This analysis is carried out primarily to motivate heuristics for allocating items in multi-period problems.

6.1 Structure of $F_1(I,B)$

Throughout this sub-section we assume:

(1) $I_{1,j} * B_{1,j} = 0$ for $j = 1,2$. This is consistent with the observation that in the penultimate period it is not optimal to back-order demand for customer j and carry item j inventory.

(2) $I_{1,j} \leq D_{1,j}$ for $j = 1 ,2$. The other cases are neither interesting nor useful to investigate.

We will also make a minor modification in the notation by including the final periods demand into the arguments of the function $F_1(., .)$. The definition of the function otherwise remains the same.

We first show how to transform a problem with initial inventories and backorders to one without any inventories or backorders (Proposition 4). This permits us to restrict our attention to problems without any starting inventories or backorders. Next, we examine $F_1(D_{1,1},D_{1,2},\mathbf{0},\mathbf{0})$ over the domain $D_{1,1} \geq 0$ and $D_{1,2} \geq 0$. In Proposition 5 we show that doubling demand causes $F_1(.,\mathbf{0},\mathbf{0})$ to double. This shows that in the $(D_{1,1}, D_{1,2})$ space, along any ray emanating from the origin, $F_1(D_{1,1},D_{1,2},\mathbf{0},\mathbf{0})$ is linear. The structure of $F_1(D_{1,1},D_{1,2},\mathbf{0},\mathbf{0})$ is derived in Proposition 6.

<u>Proposition 4</u>:

If $I_{1,j} \leq D_{1,j}$ for j = 1 ,2 then:

$$F_1(D_{1,1}, D_{1,2}, I_{1,1}, I_{1,2}, B_{1,1}, B_{1,2}) \quad =$$

$$S_{1,}I_{1,1} + S_2I_{1,2} + F_1(D_{1,1} - I_{1,1} + B_{1,1},\ D_{1,2} - I_{1,2} + B_{1,2},\ \mathbf{0},\ \mathbf{0})$$

Proof: Let :

$D^*_{1,1} = D_{1,1} - I_{1,1} + B_{1,1}$

$D^*_{1,2} = D_{1,2} - I_{1,2} + B_{1,2}$

$A^*_{11} = A_{11} - I_{1,1}$

$A^*_{22} = A_{22} - I_{1,2}$

Since $I_{1,j} \leq D_{1,j}$ for j = 1 ,2 ; $Np_m \geq 0$ and $Nq_m \geq 0$; in the optimal solution to the evening problem $A_{11} \geq I_{1,1}$ and $A_{22} \geq I_{1,2}$. As a result $A^*_{11} \geq 0$ and $A^*_{22} \geq 0$. Therefore, we can reformulate the evening problem as follows:

$$G_1(\mathbf{p_m}, N, D_{1,1}, D_{1,2}, \mathbf{I_1}, \mathbf{B_1}) =$$

$$\text{Max } S_1(A^*_{11}+I_{1,1}) + S_2(A_{12}+A^*_{22}+I_{1,2}) + \sum_{i=1}^{2} C_i I_{0,i} - \sum_{j=1}^{2} \beta_j B_{0,j}$$

S.T.

$$(2')\quad A^*_{11} + A_{12} + I_{0,1} = Np_m$$

ITEM INVENTORY BALANCE

$$(3')\quad A^*_{22} + I_{0,2} = Nq_m$$

$$(4')\quad A^*_{11} + B_{0,1} = D^*_{1,1}$$

DEMAND CONSTRAINTS

$$(5')\quad A_{12} + A^*_{22} + B_{0,2} = D^*_{1,2}$$

(6') A^*_{11} , A_{12} , A^*_{22} , $B_{0,1}$, $B_{0,2}$, $I_{0,1}$, $I_{0,2}$ ≥ 0 NON-NEGATIVITY

Thus:
$G_1(\mathbf{p_m}, N, D_{1,1}, D_{1,2}, \mathbf{I}_1, \mathbf{B}_1) = S_1 I_{1,1} + S_2 I_{1,2} + G_1(\mathbf{p_m}, N, D^*_{1,1}, D^*_{1,2}, \mathbf{0}, \mathbf{0})$
Consequently:
$F_1(D_{1,1}, D_{1,2}, I_{1,1}, I_{1,2}, B_{1,1}, B_{1,2}) =$
$S_{1,}I_{1,1} + S_2 I_{1,2} + F_1(D_{1,1} - I_{1,1} + B_{1,1}, D_{1,2} - I_{1,2} + B_{1,2}, \mathbf{0}, \mathbf{0})$ •

In the next proposition we will show that in the single period problem with zero inventories and backorders, doubling the demands will cause the optimal solution to double.

Proposition 5
Let $\mu \geq 0$, then $F_1(\mu D_{1,1}, \mu D_{1,2}, \mathbf{0}, \mathbf{0}) = \mu F_1(D_{1,1}, D_{1,2}, \mathbf{0}, \mathbf{0})$

Proof:
Let $\mu A^*_{ij} = A_{ij}$, $\mu I^*_{0j} = I_{ij}$, $\mu B^*_{0j} = B_{0j}$, $\mu N^* = N$.
By following the steps employed in Proposition 4 we can show that :

$$G_1(\mathbf{p_m}, \mu N, \mu D_{1,1}, \mu D_{1,2}, \mathbf{0}, \mathbf{0}) = \mu G_1(\mathbf{p_m}, N, D_{1,1}, D_{1,2}, \mathbf{0}, \mathbf{0}).$$

Therefore:

$$F_1(\mu D_{1,1}, \mu D_{1,2}, \mathbf{0}, \mathbf{0}) = \mu F_1(D_{1,1}, D_{1,2}, \mathbf{0}, \mathbf{0})$$ •

We know that $F_1(D_{1,1}, D_{1,2}, \mathbf{0}, \mathbf{0})$ is concave. Since the number of yield outcomes is finite, the morning problem can be formulated as a finite linear program. Thus $F_1(D_{1,1}, D_{1,2}, \mathbf{0}, \mathbf{0})$ is piece-wise linear concave. The number of linear regions is finite. An immediate consequence of Proposition 5 is that in the $(D_{1,1}, D_{1,2})$ space, along any ray emanating from the origin, $F_1(D_{1,1}, D_{1,2}, \mathbf{0}, \mathbf{0})$ is linear. Therefore, the boundary of each linear region of $F_1(D_{1,1}, D_{1,2}, \mathbf{0}, \mathbf{0})$ must be a ray emanating from the origin. Thus we have shown:

Proposition 6:
There exist a sequence of numbers $0 \leq \pi_1 \leq \pi_2 \leq \pi_3 \leq .... \pi_n \leq \infty$ such that for $\pi_j \leq D_{1,1} / D_{1,2} \leq \pi_{j+1}$, $F_1(D_{1,1}, D_{1,2}, \mathbf{0}, \mathbf{0})$ is linear.

Figure 4 illustrates the structure of $F_1(D_{1,1},D_{1,2},\mathbf{0},\mathbf{0})$. The domain is the non-negative orthant of $(D_{1,1}, D_{1,2})$. Inside each region $F_1(D_{1,1},D_{1,2},\mathbf{0},\mathbf{0})$ is linear. Over the entire domain it is piece-wise linear concave.

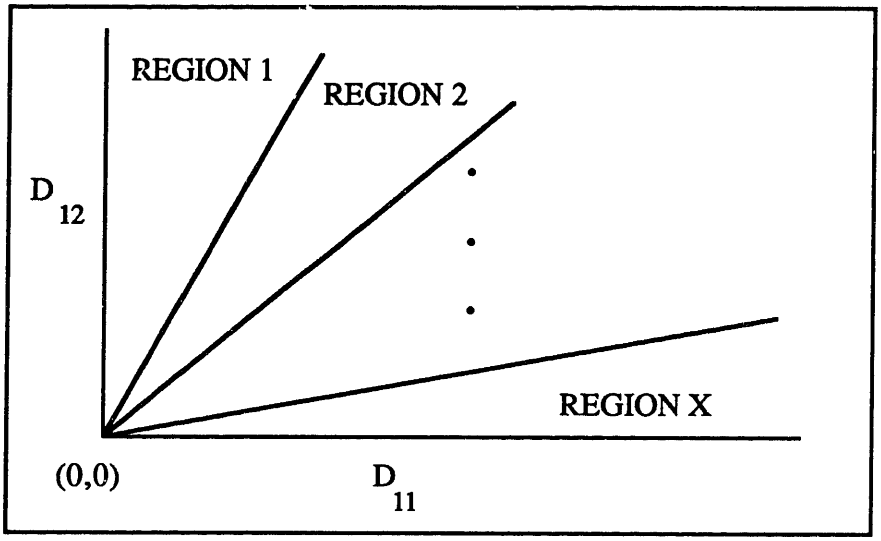


Figure 4. Structure of $F_1(D_{1,1},D_{1,2},\mathbf{0},\mathbf{0})$

6.2 Item Allocation in the Penultimate Period

Recall that after we allocate item 1 to customer 1 and item 2 to customer 2 we will be left with ε units of item 1 and U units of unsatisfied customer 2 demand. We have seen that allocating a unit of item 1 to customer 2 increase the penultimate period profits by $S_2 + ß_2 + h_1$ and changes the final period profits from $F_1(D_{1,1}, D_{1,2}, \varepsilon, 0, 0, U)$ to $F_1(D_{1,1}, D_{1,2}, \varepsilon\text{-}1, 0, 0, U\text{-}1)$. Denote the net change by Δ. Then

$$\Delta = S_2 + ß_2 + h_1 + F_1(D_{1,1}, D_{1,2}, \varepsilon, 0, 0, U) - F_1(D_{1,1}, D_{1,2}, \varepsilon\text{-}1, 0, 0, U\text{-}1)$$

On the basis of Proposition 4

$$\Delta = S_2+ß_2+ h_1 -S_1 + F_1(D_{1,1}\text{-}\ \varepsilon, D_{1,2}\text{-}\ U, \mathbf{0}, \mathbf{0}) - F_1(D_{1,1}\text{-}\ \varepsilon+1, D_{1,2} + U\text{-}1, \mathbf{0}, \mathbf{0})$$

Let $D^*_{1,1}$ and $D^*_{1,2}$ be the net final period demand; i.e.$D^*_{1,1} = D_{1,1}\text{-}\ \varepsilon + \Omega$ and; $D^*_{1,2} = D_{1,2}\text{+}\ U\text{-}\ \Omega$. Ω is the number of item 1s down-graded. By down-grading we are increasing the ratio $D^*_{1,1} / D^*_{1,2}$. Clearly it is optimal to down-grade if $\Delta > 0$. The magnitude of Δ will change whenever the down-grading process moves us from one linear region of the function $F_1(D^*_{1,1}, D^*_{1,2}, \mathbf{0}, \mathbf{0})$ to the next. By proposition 6, this happens every time $D^*_{1,1} / D^*_{1,2}$ exceeds π_j, for some

j. If Δ becomes negative then we stop down-grading. This leads us to the following result:

Proposition 7:

There exists a non-negative number π^* (perhaps $= \infty$) such that it is optimal to continue to down-grade if and only if the following conditions are satisfied:
(1) $D^*_{1,1} / D^*_{1,2} < \pi^*$;
(2) $\varepsilon - \Omega \geq 0$; and
(3) $U - \Omega \geq 0$.

In our problem the mix of the output is random. A priori, we do not know the proportion of each item that we will get in the final product. We only know the likelihood of observing any specific proportion. This probability is independent of the quantity produced. Therefore, it is not surprising that in a single period problem doubling of the demand results in the optimal solution being doubled. Given the nature of yield uncertainty, Proposition 7 is also intuitively appealing. Down-grading alters the proportions of the net demand. It increases the proportion of item 1s that are required in the final period. Thus it is reasonable to stop down-grading if the proportion of item 1s that will be needed increases beyond some critical value π^*. Although the nature of the optimal down-grading policy is intuitive, π^* is difficult to determine. It depends on the cost parameters and the yield distribution. Nevertheless, Proposition 7 is insightful and is the basis for the solution procedures we propose.

We have determined the nature of the optimal down-grading policy for a two-product two-period problem. The results extend to a multi-product problem. Unfortunately this policy is not necessarily optimal if the number of periods exceeds 2. The result extends to a multi-period problem only if the yields are deterministic. With stochastic yields the down-grading process is complex and can not be easily characterized.

In view of the difficulty in describing the optimal down-grading policy for multiple periods we propose heuristics for finite horizon problems. The rules are motivated by Proposition 7.

HEURISTICS:

Heuristic 1:

In period T allocate item k to customer k+i,

if $\sum_{j=0}^{i-1} D^*_{T-1,k+j} / D^*_{T-1,k+i} \leq \pi^*_{k,k+i}$. Where $D^*_{T-1,j}$ is the effective or net demand in period T-1; i.e. $D^*_{T-1,j}$ takes into account the backorders / inventories.

π^* is computed such that $\Pr(\sum_{j=0}^{i-1} p_{k+j} / p_{k+i} \leq \pi^*_{k,k+i}) \leq \alpha$.

Heuristic 2:

Same as Heuristic 1. However, let $\pi^* = p'/ q'$. Where p' and q' are the average yields. In addition hold safety stocks.

[Comments: Several variants of these heuristics can be designed by changing the horizon over which we compute the effective demand]

7.COMPUTATIONAL PROCEDURES

In this section we explore procedures for computing the quantities to be produced. We begin with uncapacitated single family problems, and then discuss problems with capacity constraints. One option for determining production quantities is to solve the problem as a linear program. A definite draw back of this approach is the size of the linear program. Consider a family with five customers and five yield outcomes - a three period problem will have 1250 constraints and 3250 variables. The number of constraints and variables grow exponentially as the number of outcomes increases. The problem size grows equally rapidly if we increase the number of periods.

In view of the computational burden, approximation procedures are necessary for solving this problem. There are several approaches that can be adopted. One strategy is to solve a single period problem. In this method the influence of the future periods is captured by the salvage price of inventories and the cost of backorders. Although the single period problem can be solved fairly easily, the quality of the solutions provided by this approach were found to be very poor.

An approach that is commonly used for solving large stochastic programs involves aggregating the outcomes. When we aggregate, a set of outcomes are replaced by their expected values. As the level of aggregation increases the quality of the solution deteriorates, but the computational burden decreases. Hence, a trade-off needs to be made between the computational difficulty and the accuracy of the solution.

We assume that yields are stochastic in the first period, and are deterministic (equal to their means) in subsequent periods. Thus we are aggregating the outcomes in periods other than the current period. Each period we solve this approximate problem to decide the quantity to be produced. Once the yields are observed, allocations are made on the basis of an allocation heuristic. The problem is solved again the next period with new initial inventories and backorder positions. Thus we solve the problem on rolling horizon basis.

We have tested the performance of the aggregation heuristic on a set of problems. To determine the error due to aggregation, we had to compute the objective value without any aggregation. The size of our test problems was greatly restricted by the rapidity with which the size of the actual problem grows. The test problems consisted of two-product families. In each case there were two possible yield outcomes and the time horizon was 6 periods. To compute the expected profits from the aggregation procedure, we considered every possible set of yield outcomes over the horizon . Table 1 contains the description of the test problems.

The error in the objective value ranged from 3% to 14%. The error increased with increase in the variance of the yields. The variability in yields is higher in problems 5 - 9 as compared to that in problems 1 - 4. The errors in problems 1 - 4 are significantly lower, than the errors in problems 5 - 9 (Table 2). These numbers suggest that the error in the aggregation procedure is significantly influenced by the variability in the yield. The influence of yield variability on the quality of heuristic solutions is further illustrated by test problems 14, 15 and 16. The average yields are the same in these problems. However the variability is greatest in problem 15 and least in 17. The change in the error of the approximation is consistent with the change in variability.

TABLE 1

TEST PROBLEMS

For all the problems listed below the cost parameters were the same. The costs per unit are as follows:

Production Cost: $8

	ITEM 1	ITEM 2
Selling Price:	$20	$15
Back Order Cost:	$2	$1.50
Holding Cost:	$1	$1

All problems were tested over a horizon of 6 periods.

Problem #1:

Stationary Demand, No Capacity Constraint

	ITEM 1	ITEM 2	PROBABILITY
YIELD:			
Outcome 1	0.20	0.28	0.25
Outcome 2	0.18	0.56	0.75
DEMAND:	185	490	

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	7925	7768	2

Problem #2:

Oscillating Demand, No Capacity Constraint

YIELD : Same as Problem # 1

DEMAND:

Period	(1)	(2)	(3)	(4)	(5)	(6)
Item 1	170	200	170	200	185	185
Item 2	490	490	490	490	490	490

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	8052	7746	4

TABLE 1 (Cont'd)

Problem #3:

Stationary Demand, Capacity Constraint

YIELD	Same as Problem #1
DEMAND	Same as Problem # 1
CAPACITY	1200

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	7885	7649	3

Problem #4:

Oscillating Demand, Capacity Constraint

YIELD	Same as Problem # 1
DEMAND	Same as Problem # 2
CAPACITY	1200

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	7839	7647	3

Problem #5:

Stationary Demand, No Capacity Constraint

	ITEM 1	ITEM 2	PROBABILITY
YIELD:			
Outcome 1	0.20	0.28	0.5
Outcome 2	0.18	0.56	0.5
DEMAND:	300	689	

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	8605	7548	11

TABLE 1 (Cont'd)

Problem #6:

Oscillating Demand, No Capacity Constraint

YIELD : Same as Problem # 5

DEMAND:

Period	(1)	(2)	(3)	(4)	(5)	(6)
Item 1	280	320	280	320	300	300
Item 2	689	689	689	689	689	689

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	8750	7550	14

Problem #7:

Stationary Demand, Capacity Constraint

YIELD	Same as Problem # 5
DEMAND	Same as Problem # 5
CAPACITY	1200

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	8244	7194	15

Problem #8:

Oscillating Demand, Capacity Constraint

YIELD	Same as Problem # 5
DEMAND	Same as Problem # 6
CAPACITY	1200

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	7958	7246	9

TABLE 1 (Cont'd)

Problem #9:

Stationary Demand, No Capacity Constraint

	ITEM 1	ITEM 2	PROBABILITY
YIELD:			
Outcome 1	0.35	0.60	0.5
Outcome 2	0.05	0.60	0.5
DEMAND:	200	800	

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	15960	14864	7

Problem #10:

Oscillating Demand, No Capacity Constraint

YIELD : Same as Problem # 9

DEMAND: Period	(1)	(2)	(3)	(4)	(5)	(6)
Item 1	200	200	200	200	200	200
Item 2	750	850	750	850	800	800

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	15960	14728	8

Problem #11:

Stationary Demand, Capacity Constraint

YIELD	Same as Problem # 9
DEMAND	Same as Problem # 9
CAPACITY	1200

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	15866	14435	9

TABLE 1 (Con't'd)

Problem #12:

Oscillating Demand, Capacity Constraint

YIELD	Same as Problem # 9
DEMAND	Same as Problem # 10
CAPACITY	1200

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	15832	14218	11

Problem #13:

Stationary Demand, No Capacity Constraint

	ITEM 1	ITEM 2	PROBABILITY
YIELD:			
Outcome 1	0.25	0.60	0.5
Outcome 2	0.15	0.60	0.5
DEMAND:	200	600	

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	15526	15519	0

Problem #14:

Stationary Demand, No Capacity Constraint

	ITEM 1	ITEM 2	PROBABILITY
YIELD:			
Outcome 1	0.1	0.55	0.5
Outcome 2	0.7	0.25	0.5
DEMAND:	400	400	

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	5125	4440	13

TABLE 1 (Cont'd)

Problem #15:

Stationary Demand, No Capacity Constraint

	ITEM 1	ITEM 2	PROBABILITY
YIELD:			
Outcome 1	0.20	0.50	0.5
Outcome 2	0.60	0.30	0.5
DEMAND:	400	400	

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	11664	10905	7

Problem #16:

Stationary Demand, No Capacity Constraint

	ITEM 1	ITEM 2	PROBABILITY
YIELD:			
Outcome 1	0.30	0.45	0.5
Outcome 2	0.50	0.35	0.5
DEMAND:	200	800	

	EXACT	APPROX	% ERROR
OBJECTIVE VALUE:	12297	11649	5

TABLE 2

EFFECT OF YIELD VARIABILITY ON ERROR

	LOW VARIANCE (PROBLEM #1 - 4)	HIGH VARIANCE (PROBLEM #5 - 9)
Stationary Demand	2%	11%
Oscillating Demand	4%	14%
Stat. Demand with Cap. Constr.	3%	15%
Oscill. Demand with Cap. Constr	3%	9%

7.1 Multi-family Capacitated Problems

Until now we restricted our analysis to an single family problems. We now study a multi-family problems. Henceforth we assume that we have a capacity constraint that restricts the total quantity produced. In the absence of a capacity constraint, the problem separates by families, resulting in several single family problems. We develop a greedy procedure for solving the single period problem. Since the capacity constraint ties all the families together, the size of capacitated problems is significantly larger than that of an uncapacitated problem. For instance a single period problem with four families, each with five customers and five yield outcomes, will have 18,750 constraints and 46,875 variables. The number of constraints and variables grow exponentially as the number of families increases. The problem size grows equally rapidly if we increase the number of periods.

The problem size can be reduced considerably if we relax the capacity constraint. Therefore, an obvious strategy is to relax the capacity constraint and use Lagrangean relaxation based techniques. We adopt this approach, but do not employ linear programming techniques to solve the resulting subproblem. Instead, we develop a greedy procedure. At every stage of our algorithm we compute the marginal value of allocating an additional unit of production to a family. Since the expected value of the evening problem is a concave function of the number of units produced, a greedy procedure for allocating the next unit of production to different families is optimal. Thus we have a simple procedure for solving the single period capacitated problem.

In the next section we describe in greater detail the algorithm for solving the single period multi-family capacitated problem.

7.2 Greedy Procedure for a Single Period Problem

The multifamily capacitated problem is given by:

$$F_1(\boldsymbol{I}_1, \boldsymbol{B}_1) = \text{Max} \sum_{f=1}^{F} \mathbf{E}_p \, G_{f,1}(\mathbf{p}^f{}_m, N_f, \mathbf{I}_{1,f}, \mathbf{B}_{1,f}) - \sum_{f=1}^{F} K_f N_f$$

S.T.

$$\sum_{f=1}^{F} N_f \leq k \qquad \text{[CAPACITY CONSTRAINT]}$$

$$N_f \geq 0$$

The index f denotes the family. The function $\mathbf{E}_p\, G_{f,1}(.)$ represents the expected value of the evening problem for family f. The structure of the evening problems in the multifamily single period case is identical to that of the evening problem in the single family case. Since each $\mathbf{E}_p\, G_{f,1}(.)$ is a concave function whose value depends only N_f and is independent of N_i for $i \neq f$, the capacitated problem is a concave knapsack problem. Thus a greedy procedure which allocates the next unit of production to the family with the highest marginal profit is optimal. We outline below an approximation procedure that allocates δ units of production at each step of the allocation process.

```
STEP 1: Initialization:
            Compute  $v_f = [\partial\, \mathbf{E}_p\, G_{f,1}(.,.) / \partial N_f - K_f]$ at $Nf = qf$
            Let      $q_f = 0$ for all f
        Where :
          $v_f$  : marginal value of allocating a unit of production to family j
          $q_f$  : capacity allocated to family f

STEP 2: Identify the family with highest marginal value
        $f^* = \text{Arg Max}\, [\, v_f\, ]$ ;  If $v_{f*} \leq 0$  STOP  (optimality condition)
STEP 3:  Allocate production to family f*
        $q_{f*} = q_{f*} + \delta'$ ;  Where  $\delta' = \min \{\, \delta\, ;$ Unallocated capacity$\}$
STEP 4:  Check if capacity has been fully allocated.  If so STOP
STEP 5 : Update marginal value for family f*
   Compute:  $v_{f*} = [\partial\, \mathbf{E}_p\, G_{f*,1}(.,.) / \partial N_{f*} - K_{f*}]$ at $N_{f*} = q_{f*}$
        Return to Step 2.
```

For $\delta > 0$ this procedure is an approximation procedure. The final solution may not be optimal. We describe below a procedure for computing v_f.

Consider family f, let π^f_m be the probability that we observe yield outcome m. Then :

$$v_f = \quad [\partial\, E_p\, G_{f,1}(.,p^f_m,.) / \partial N_f \;-\; K_f\,]_{\text{ at } N_f = q_f}$$

$$= \sum_{m=1}^{M_f} \pi^f_m\, [\partial\, G_{f,1}(.,\, p^f_m,.,.) / \partial N_f \;-\; K_f\,]_{\text{ at } N_f = q_f}$$

$$[\partial\, G_{f,1}(.,\, p^f_m,.,.) / \partial N_f]_{\text{ at } N_f = q_f} = (\,\gamma^f_{1,m}\;\; p^f_{1,m}) + (\;\gamma^f_{2,m}\, p^f_{2,m})$$

Where $\gamma^f_{1,m}$ and $\gamma^f_{2,m}$ are the shadow prices of constraints 2 and 3 with $N_f = q_f$. These shadow prices can be determined by solving the corresponding linear program (evening problem). They can also be computed directly.

$\gamma^f_{i,m}$ is the marginal value of an additional unit of item i given that we have produced N_f units and outcome m is observed. Since we know the yield we can determine, under the optimal allocation process, how the next unit of item i will be utilized, and thereby determine its value. For instance if $N_f p^f_{1,m}$ is less than customer 1's demand , then the value of the next unit of item 1 equals the selling price + back-order penalty for customer 1. $p^f_{1,m}$ is the yield of item 1, family f under outcome m.

An algorithm for determining the shadow prices is given in Figure 5. The shadow prices are determined by allocating the items to the customers. We begin from customer 1 and go down the hierarchy. For customer k we first allocate item k. If item k yield is adequate then the shadow price for item k equals its salvage value. If customer k's demand can not be satisfied by item k, then we set item k's shadow price equal to customer k's selling price + back-order cost. Then we go up the hierarchy, starting from item k-1, to see if we can allocate item k -i $(1 \leq i \leq k\text{-}1)$ to customer k. Before we down-grade item k-i we need to check if :

(a) there are any surpluses of item k-i; and

(b) the selling price + backorder cost for customer k exceeds the salvage price of item k-i.
We allocate item k-i to customer k only if the answer to both the questions is yes. If we downgrade, the shadow price of item k-i is updated. The new shadow price can taken on two values:
(1) If the surplus of item k-i equals or exceeds the residual requirements of customer k the shadow prices of all items k-i through k are set equal to the salvage price of item k.
(2) If the surpluses of item k-i are inadequate, then the shadow price of item k-i is set equal to the selling price + backorder cost for customer k.

We terminate the downgrading process for customer k if (1) the salvage price of item k-i exceeds the selling price + backorder cost for customer k ; or (2) there are no more surpluses.

The algorithm requires $O(M_f Z_f^2)$ steps to compute v_f. Where M_f is the number of outcomes for, and Z_f is the number of items in family f.

8. Summary and Conclusions

In this paper we have modelled a production planning problem arising in the semi-conductor industry. The most interesting part of this work was identifying and structuring the problem. In this environment, yields vary significantly and demands are substitutable. In addition, from each production lot we get several different items. We have formulated the problem as a convex stochastic program. On the basis of the structure of a two period problem, we have identified a class of heuristics for allocating items to customers. The heuristics allocate items to customers in a manner that keeps the net demand for the items in balance. The objective of the allocation heuristics is not to allow the net demand for any item to exceed some (predetermined) fraction of the total demand for all the items. These allocation policies are consistent with the nature of the yield uncertainty - multiplicative yields.

We also propose approximation procedures for solving single family finite horizon problems, with and without capacity constraints. We propose to approximate the problem by assuming that yields, in periods other than the first period in the horizon, are deterministic. We have tested this procedure on a sample of problems.

The size of the test problems was constrained by the size of the exact problems. The errors in the objective value ranged between 3% and 14%. The level of uncertainty in the yield had a significant effect on the performance of the heuristic procedure. We have identified a greedy procedure for solving the single period multi-family capacitated problem.

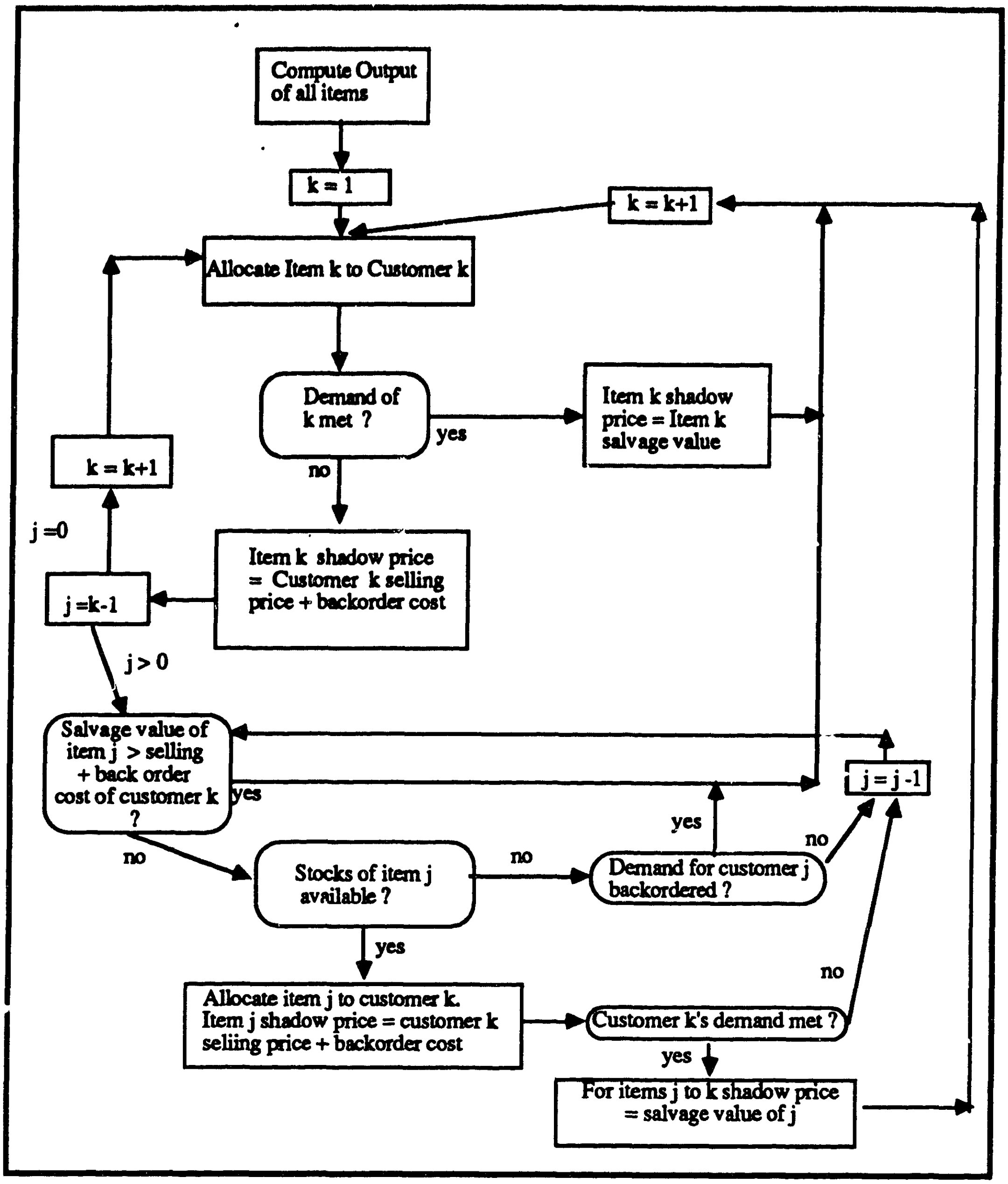

Compute Output of all items
k = 1
k = k+1
Allocate Item k to Customer k
Demand of k met ?
yes
no
Item k shadow price = Item k salvage value
k = k+1
j =0
Item k shadow price = Customer k selling price + backorder cost
j =k-1
j > 0
Salvage value of item j > selling + back order cost of customer k ?
yes
no
j = j -1
Stocks of item j available ?
no
yes
Demand for customer j backordered ?
yes
no
Allocate item j to customer k. Item j shadow price = customer k seliing price + backorder cost
Customer k's demand met ?
no
yes
For items j to k shadow price = salvage value of j

FIGURE 5

ALGORITHM FOR COMPUTING SHADOW PRICES

APPENDIX 1

Proposition 3 claims $F_1(I_1, B_1)$ to be a concave function.

We prove the result for a more general setting.

Lemma 1. Let $h(\mu) = \text{Max } \{L(\mu,N) - cN \mid N \geq 0\}$ If $L(.,.)$ is a concave function then $h(.)$ is also a concave function.

Proof: Let $\mu^* = \beta\mu' + (1-\beta)\mu''$. Where $0 < \beta < 1$. Let N^*, N', N'' be optimal for the problem with μ^*, μ', and μ'' respectively.

$h(\mu^*) = L(\mu^*,N^*) - cN^* \geq L(\mu^*, \beta N' + (1-\beta)N'') - c(\beta N' + (1-\beta)N'')$.

By concavity of $L(.)$ we have

$$L(\mu^*, \beta N' + (1-\beta)N'') - c(\beta N' + (1-\beta)N'')$$
$$\geq \beta\, \{L(\mu',N') - cN'\} + (1-\beta)\, \{L(\mu'',N'') - cN''\} = \beta h(\mu') + (1-\beta)\, h(\mu'').$$

Hence $h(\beta\mu' + (1-\beta)\mu'') = h(\mu^*) \geq \beta h(\mu') + (1-\beta)h(\mu'')$ •

REFERENCES

Birge, J. R., and Wets R. J-B "Designing approximation schemes for stochastic optimization problems, in particular for stochastic programs with recourse, " Mathematical Programming Study 27, 1986, 54 -102

Bitran, G. R., and Tirupati, D., " Planning and scheduling for epitaxial wafer production facilities," Oper. Res., 36, 1988, 34 - 49

Dantzig, G. B., "Linear programming under uncertainty," Management Sci., 1, 1955, 197 - 206

Kalro, A. H., and Gohil, M.M., " A lot size model with backlogging when the amount received is uncertain," Intnl. J. of Prod. Res., 20, 1982, 775 -786

Kothari, V., "Silicon wafer manufacture," Unpublished thesis, Sloan School of Management, May 1984

Leachman, R. C., "Preliminary design and development of a corporate-level production planning system for the semiconductor industry," ORC 86 - 11, University of California, Berkeley, California, 1986

Mazzola, J. B., McCoy, F. W., and Wagner, H. M., "Algorithms and heuristics for variable yield lot sizing," Nav. Res. Log. Quartl., 34, 1987, 67 - 86

Olsen, P., "Multistage stochastic program with recourse: the equivalent deterministic problem," SIAM. J. on Control and Optimization, 14, 1976, 495 - 517

Pierskalla. W. P., Deuermeyer, B. L., "A by-product production system with an alternative, " Mgmt. Sci., 24, 1978, 1373 - 1383

Shih, W., "Optimal inventory policies when stockout results from defective products," Intnl. J. Prod. Res., 18, 1980, 677 - 686

Silver, E. A., " Establishing the reorder quantity when the amount received is uncertain," INFOR, 14, 1976, 32 - 39

Veinott, A. F., Jr., "Status of mathematical inventory theory," Management Sci., 11, 1960, 745 - 777

III. QUEUEING NETWORKS AND THE METHOD OF PHASES

1. INTRODUCTION

Queueing networks have been used to model the performance of a variety of systems such as production job shops, flexible manufacturing systems, and communication networks. In manufacturing environments queueing networks are particularly useful for studying batch shops and closed job shops. The arrival of jobs to the shop is modelled as a random process, and the probability density of the service time is derived from the distribution of the size of jobs that are received by the shop. The estimates of the lead-times and work-in-process obtained from the queueing analysis provide a bench-mark for assessing the performance of the shop. Since queueing analysis estimates the long-term performance of the shop, queueing models are most suitable for supporting strategic and tactical decisions rather than operational decisions. In recent years several firms have started using the methodology of queueing networks to model strategic and tactical decisions. For instance, the methodology permits practitioners to assess the impact of new technologies, and the effect of new products on work-in-process and lead times. Also, it is possible to derive trade-off curves between work-in-process, lead times and capacity [Bitran and Tirupati 86].

Unfortunately queueing networks are difficult to analyze. Exact results exist for a limited class of Jackson type networks [Jackson(63), Kelly(75)]. The reader is referred to survey papers by Lemoine (77) and Disney and Konig (85) for more references and details of exact results in queueing networks. For more general networks several approximation procedures have been proposed in the literature. The approximation schemes can be grouped into 4 categories:

(1) Diffusion approximations;

(2) Mean value analysis;

(3) Operational analysis; and

(4) Decomposition methods.

The approximation procedures that we are studying in this thesis are decomposition methods. The other methods listed above are not directly related to our work and we will not be describing these.

One of the reasons for the difficulty in analyzing queueing networks is the interaction between different nodes in the network. Consider a node somewhere in the middle of a network (node x in Fig 1.1), jobs can arrive to this node from several other nodes in the network, and once they are processed they may be sent off to different nodes. The central ideas of the decomposition procedure is to analyze each node in the network by itself. For this purpose we need to do the following:

Step 1. SUPERPOSITION: Combine the different arrival streams to the node.
Step 2. QUEUEING ANALYSIS: Analyze the queueing effects at the node with the arrival stream derived in Step 1.
Step 3. DEPARTURE PROCESS: Derive the departure process from the node. The departure from this node become arrivals to some other node.

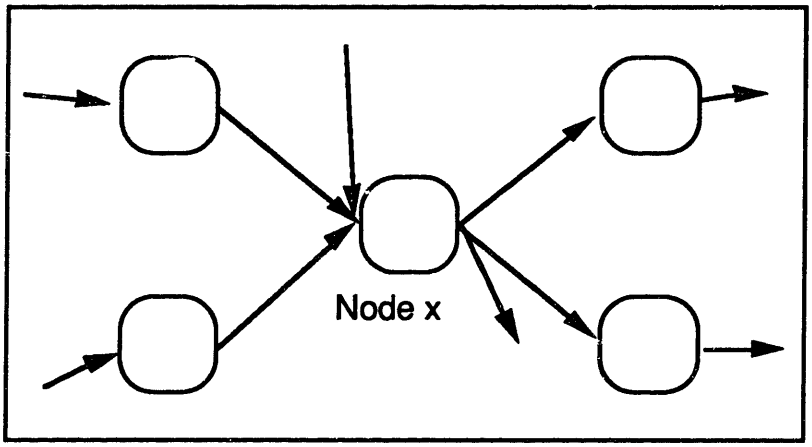


Figure 1.1 Illustration of a Network

This approach of "breaking up" the network into individual nodes and deriving the interactions is the essence of the decomposition approach. The approach is motivated in part by the properties of Jackson type networks. If the interarrival times of all the external arrivals to the network, and all the service times are exponentially distributed, then Jackson(63) showed that each node in the network behaves exactly like an M/M/n queue. Unfortunately, if either the service distributions or the interarrival times are not exponentially distributed, but have general distributions, then not only is exact analysis very difficult, even the 3 steps outlined above are also difficult. Consequently approximation procedures have been proposed to carry out steps 1 through 3.

An approximate procedure that has provided accurate estimates of the mean of the waiting time in queueing networks is the parametric decomposition approach (Whitt 83, Bitran and Tirupati 86). Reiser and Kobayashi(74) and Kuehn(78, 79) were among the first proponents of the parametric decomposition approach. Subsequently, this method has been used by several researchers including Sevick et al(77), Chandy and Sauer(78), Shantikumar and Buzacott(81), Whitt(83), and Bitran and Tirupati(86).

In the next subsection we describe the essential features of the parametric decomposition method. We then outline our approach for analyzing queueing networks and show how we build upon the basic ideas of the parametric decomposition methods.

1.1 Parametric Decomposition Methods

This approach relies on two notions:
(a) The nodes can be treated as being stochastically independent, and
(b) Two parameters - mean and variance - of the arrival and service process are adequate to estimate the performance of the system.

The first assumption is the main decomposition assumption. The second assumption is central to the parametric approach. Since only the first and second moments of the arrival and service process are used for estimating the queueing performance, the parametric approach only keeps track of these parameters while deriving the interaction between stations. This statement is further clarified in the ensuing discussion.

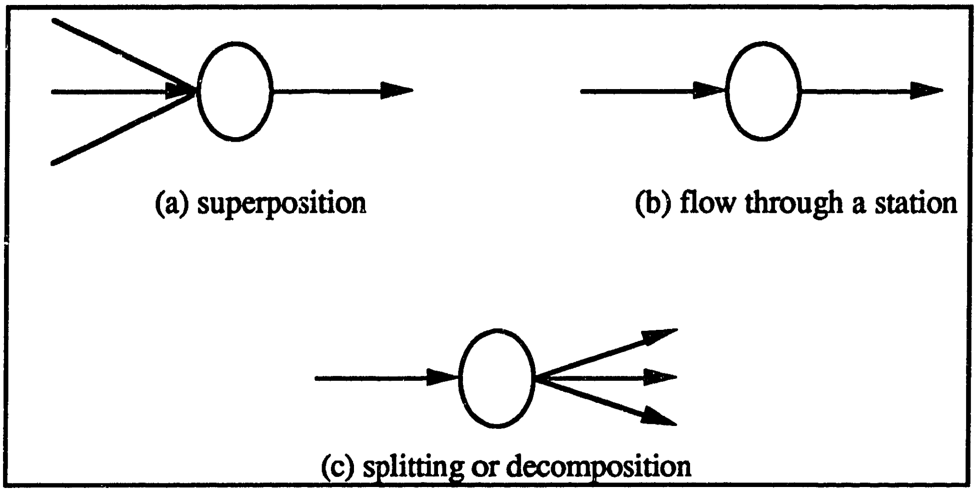
As stated earlier the decomposition methods consist of three stages:
(1) Analysis of interaction between stations.
(2) Analysis of each station.
(3) Composition of the results to obtain the network performance.

The first stage is critical to the decomposition approach and we now describe how it is implemented in the parametric approach in greater detail. The models we describe are due to Shantikumar and Buzacott(81) and Whitt(83). These models are good representatives of the parametric decomposition methods and illustrate the basic ideas of the approach.

1.1.1 Analysis of Interaction between stations in the parametric methods

This step forms the core of the decomposition approach. It in turn involves the three basic steps:
(a) Superposition or merging (arrival).
(b) Flow through a queue or a station.

(c) Splitting or decomposition (departure).



(a) superposition

(b) flow through a station

(c) splitting or decomposition

Figure 1.2 Basic Steps in the Decomposition Method

(a) *Superposition*:: In this step different streams arriving at a station are merged into a single stream. Each stream is characterized by two parameters, the mean arrival rate and the squared coefficient of variation (scv) of the arrival process. The parameters of the superposed process are approximated as a linear combination of the parameters of the individual streams.

(b) *Flow through a queue:* The parameters of the departure process from the station are estimated. Clearly if the utilization of the station is less than one, the mean departure rate must equal the mean arrival rate. The scv of the departure stream is estimated to be a convex combination of the scv of the service process and the arrival process. Thus the scv of the departure stream is a linear function of the scv of the arrival stream and the service process.

(c) *Splitting:* The output from the station is split into a number of substreams. The splitting process captures alternate routing of different product types through the facility. Once again a linear system of equations describes the relationship (an approximation) between parameters of the departure stream prior to splitting and after splitting.

Since at each of the three basic steps the scv of the output process is a linear function of the scv of the input process, combining the approximations for the three basic steps described above, leads to a system of linear equations. Solving these equations yields the first and second moment of the arrival process at each station in the network.

These two parameters are then used to estimate performance measures such as the first two moments of the waiting and queue length at each station. The parametric decomposition approach is computationally attractive because it only entails solving a system of linear equations.

For more details of this approach please refer to the works by Shantikumar and Buzacott(81), Whitt(83a) and Bitran and Tirupati(86).

1.1.2 Assumptions Underlying the Parametric Decomposition Approach

Consider the main assumptions underlying the parametric decomposition method.
(1) Departure Process: The departure process from each queue is approximated by a renewal process, which is distributed as the stationary interval of the actual departure process (please note that the departure process from a queue is in general not a renewal process. In other words, adjacent interdeparture times are not independent of each other). In the parametric method the squared coefficient of the departure process is assumed to be equal to the scv of the stationary interval. However, the scv of the stationary interval is not computed, instead it is approximated. The approximation for the scv of the stationary interval used in the parametric approach is exact under the following conditions:
i) utilization of the server is either 0 or 1,
ii) the queue is an M/M/n queue.

(2) Superpositioning : Each individual stream being combined is assumed to be a renewal process, and the combined process is approximated by a renewal process. The combined process is in general not a renewal process. The superposition of renewal processes is a renewal process if and only if each arrival stream is a Poisson process. Thus in the parametric decomposition approach the point process resulting from the superpositioning of renewal processes is approximated by a renewal process.

If we are going to approximate a point process by a renewal process then we must identify some basis for comparing the approximation with the actual process. The superpositioning of renewal process results in a point process in which each interarrival interval is identically distributed but the intervals are not independent of each other. Thus the first moment of the approximating process is straightforward to compute. In fact the mean arrival rate of the approximating renewal process should be the same as the mean

arrival rate of the superposed process. It is the second moment of the approximating renewal process that is difficult to compute. Whitt (82) and Albin (84) consider two basic procedures for approximating superposition of renewal process. These approaches have been called micro and macro approaches. In the micro approach, we take a microscopic view and try to match the behavior of the superposed process over a short interval of time. In the micro approach the scv of the approximating renewal process is set equal to the scv of the stationary interval of the superposed process. The macro approach on the other hand tries to match the behavior of the superposed process over a very long time horizon. Hence, the macro approach is also called the asymptotic method. When these methods were tested in queues, it was found that an approach that combines both the basic methods produced the best results. Albin refers to this combined approach as the hybrid approach. In the hybrid approach the scv of the approximating renewal process is set equal to a convex combination of the asymptotic scv and the stationary interval scv. The asymptotic scv is fairly easy to compute (Whitt 82). However, if the interarrival times of the renewal process have general distributions, the scv of the stationary interval is difficult to obtain. Hence, in the parametric decomposition approach the stationary interval scv is approximately computed. The approximation for the scv of the stationary interval are exact if the interarrival times are distributed as one of the following:
(i) Erlang or exponential,
(ii) shifted exponential, or
(iii) hyperexponential.

(3) Queueing Analysis: Once the mean and the scv of the arrival stream are computed, Kraemer -Lagenbach - Belz (76) approximate formulae, are used to estimate the waiting time.

1.1.3 Phase Distributions and Queueing Networks

Although the parametric decomposition method makes strong assumptions, it provides very good estimates of performance measures such as the the mean waiting time and mean queue length. This suggests that analyzing the queueing phenomena at each node independently is likely to provide good estimates of not only the average waiting time but also other performance measures such as the variance of the waiting time and the waiting time distribution; provided the 3 basic steps of the decomposition procedure can be carried out accurately. Motivated by this observation we have identified an approach for analyzing networks of queues. Our procedure is a decomposition procedure in that each node is analyzed independently. However, we require the distribution of the service times and the interarrival times to belong to a class of distributions called Phase type distributions (Neuts 81). A phase type distribution is the distribution of the time to leave a finite transient Markov chain. In the next section we provide a formal definition of the phase type distribution.

The requirement, that all distributions in the queueing network are of the phase type is not a severe restriction, as the phase type distributions are dense in the set of all probability distributions on $[\, 0 \,, \infty)$. Theoretically, any distribution can be approximated arbitrarily closely by a phase type distribution. The first step in our scheme is to identify phase type distributions that approximate the interarrival time distributions and the service time distributions. Once this is done, in this thesis we show how to :
(1) Superpose different arrival streams; and
(2) Analyze queues with non-renewal arrivals.

In our procedure, the first and second steps of the decomposition framework can in principle be carried out exactly. However, if the number of processes being superposed is large then the exact representation of the superposed process becomes very large and poses computational difficulties. We therefore, propose procedures for approximating the superposed process. In the parametric decomposition approach the superposed process is approximated by a renewal process. In our framework we can use non-renewal approximations for the superposed process. Our approximations are likely to more closely resemble the superposed process, because the latter is not a renewal process. We can use non-renewal approximations, because we can analyze queues with non-renewal arrivals.

The third step of the decomposition procedure concerns the departure stream. Recall, that in the parametric approach, the departure stream is approximated by renewal process which is distributed as the stationary interval of the departure process. As a by product of analyzing phase type queues we identify the distribution of the stationary interval. In fact, the stationary interval is also a phase type distribution. Whereas in the parametric decomposition approach, the second moment of the stationary interval of the departure process is computed approximately, we derive the exact distribution of the stationary interval. Later in this thesis we show how to develop non-renewal approximations for the departure process. We also show that the stationary interval distribution is a good approximation for the departure process. Table 1.1 below summarizes the differences between the parametric decomposition approach and the method we are proposing.

In this thesis we also show that, if the arrival process to a queue and the service process have phase type representations, then under the stationary interval assumption made in the decomposition approach, the departure stream also has a phase type representation. In addition, the superposition of phase type processes is also a phase type process. (A phase type process need not be a renewal process.) Thus if all the external arrivals to the network and the service distributions have phase type representation, then under the renewal approximation for the departure process, all flows in the network have phase type representations. Given these properties of phase distributions and that Erlang and hyper-exponential distributions are special cases of phase distributions, we believe it is appropriate to study queueing networks with phase distributions. Several other factors reinforce our interest in this class of distributions. Before we enumerate these factors we digress briefly to define more precisely the class of distributions of interest.

TABLE 1.1
COMPARISON OF PARAMETRIC METHOD AND PHASE METHOD

CHARACTERISTIC	PARAMETRIC METHODS	PHASE
1) External arrivals and service distributions.	General distributions. Need only first and second moments of each process.	Phase type distributions.
2) Departures from queues.	Approximated by a renewal process. Approximately compute variance of stationary interval.	Can be approximated by a non-renewal process. Derive exact distribution of stationary interval, and the correlation between adjacent interdeparture intervals.
3) Superposing of arrival streams.	Superposed process approximated by a renewal process.	Exact representation of the superposed process. If size of exact representation is too large, then use non-renewal approximations.
4) Queueing analysis.	Approximate.	Exact.
5) Performance measures.	Primarily mean waiting time.	Derive mean waiting time and all higher moments. Distribution of number in queue.

2. PHASE DISTRIBUTIONS

We adopt the notations and definitions of Neuts(81). Much of the analysis in this thesis is based on his work. For the sake of completeness we reproduce some of the essential properties of phase distributions.

Definition: A probability distribution F(.) on $[0, \infty)$ is a distribution of the phase type if and only if it is the distribution of the time until absorption in a finite Markov process of the type defined below.

Consider a Markov process on the states $\{1,2,\ldots, m+1\}$ with infinitesimal generator:

$$Q = \begin{vmatrix} T & T^0 \\ 0 & 0 \end{vmatrix}$$

Where the m x m matrix T satisfies $T_{ii} < 0$, for $1 \le i \le m$, and $T_{ij} \ge 0$ for $i \ne j$. $Te + T^o = 0$, and the initial probability vector of Q is given by $(\mathbf{a}, a_{m+1})$. e is a column vector of 1s. States 1 through m are all transient, so that absorption into state m+1, from any initial state is certain. A useful equivalent condition is given by the following property.

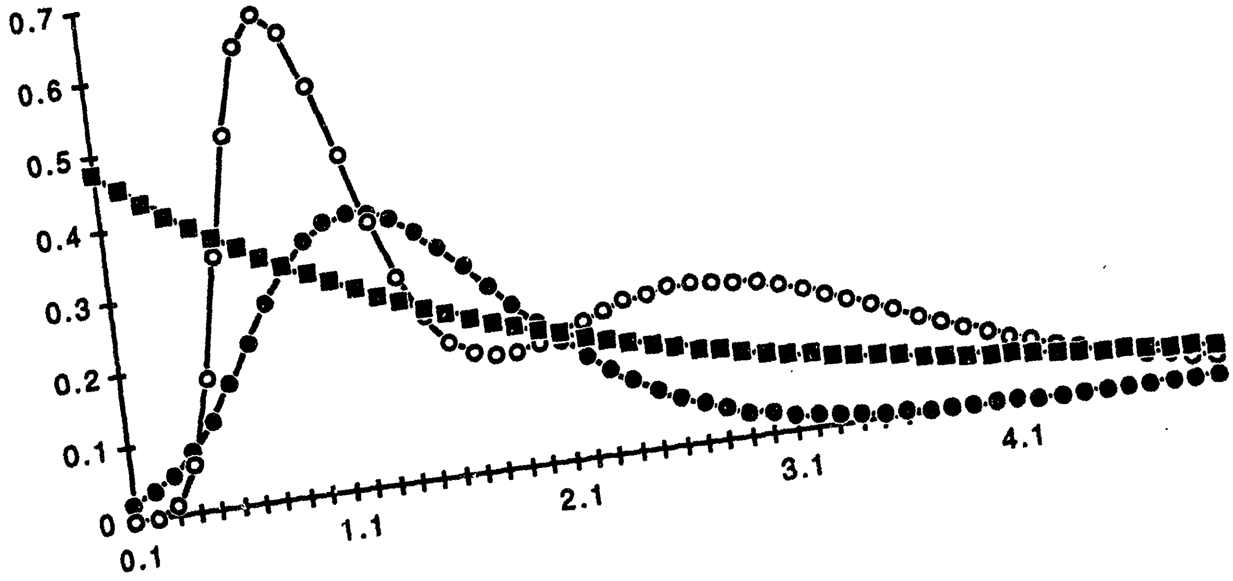
Property 1. The states 1 through m are transient if and only if the matrix T is nonsingular.
Proof: Lemma 2.2.1 page 45 Neuts(81).

Examples of Phase type distributions: Erlang distributions, mixtures of Erlang distributions, exponential distributions, hyperexponential distributions are phase type distributions. Figure 2.1 graphs some phase type distributions. Figure 2.2 depicts the Markov chain of a phase type distribution with the following representation.

a = (0, 1, 0) (initial probabilities)

$$T = \begin{vmatrix} -3 & 3 & 0 \\ 0 & -2 & 2 \\ 1 & 0 & -3 \end{vmatrix} \qquad T^0 = \begin{vmatrix} 0 \\ 0 \\ 2 \end{vmatrix}$$

Figure 2.1 Illustration of Phase distributions



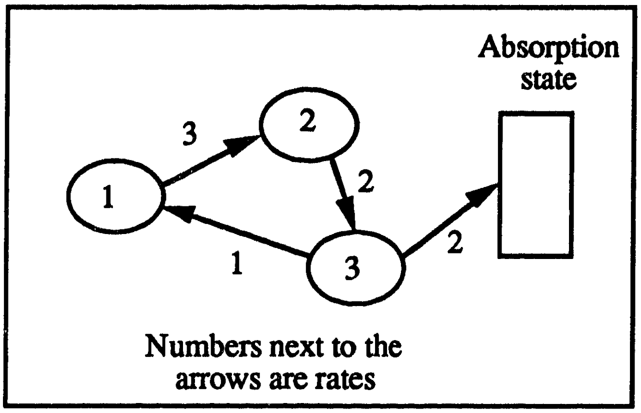


Figure 2.2 Markov Chain of a Phase Distribution

Property 2. The probability distribution F(.) of the time until absorption in the state m+1, corresponding to the initial probability vector $(\mathbf{a}, a_{m+1})$ is given by

$$F(x) = 1 - \mathbf{a}\exp(Tx)\mathbf{e}.$$

Proof : Lemma 2.2.2 page 45 Neuts(81).

Property 3. The moments of F(.) are all finite and are given by

$$m_i = (-1)^i i!(\mathbf{a}T^{-i}\mathbf{e}) \quad \text{for } i \geq 0$$

The elements of the matrix T^{-1} have a useful probabilistic interpretation. Element $-Tij$ is the expected amount of time spent in state j before absorption, given that we started in state i. We make use of this fact several times in this paper.

Property 4. The Laplace-Stieltjes transform f(s) of F(.) is given by

$$f(s) = a_{m+1} + \mathbf{a}(sI - T)^{-1}T^o \quad \text{for Re } s \geq 0$$

2.1 Phase Renewal Process

If the interarrival times are independent and identically distributed as a phase type distribution then we describe the arrival process as a phase renewal process.

3. MOTIVATION FOR PHASE TYPE DISTRIBUTIONS

As stated earlier, in this thesis we restrict our attention to queues where the flows and the service processes are of the phase type (PH). In this section we provide some more reasons for our interest in these distributions.

Phase type distributions are dense in the set of all probabilities on $[0, \infty)$ [Neuts(81)]. Several qualitative features such as bimodality, and skewness can be readily captured by phase type distributions. This does not mean all distributions can easily be represented by phases. Neuts reports difficulties in approximating distributions with steeply increasing or decreasing regions, or with regions of constancy. Nevertheless, at the very least the phase type distributions augment the family of distributions that can be conveniently analyzed. Also, at present in the decomposition approach only two moments are being used to characterize the various processes. Much of the work in approximating point processes has also focussed on two parameter approximations [Chandy and Sauer(78), Kuehn(79)]. In that context phases could be viewed as a procedure for enlarging the number of moments being represented. Higher moments can be fairly easily computed for phase distributions.

Further, it is possible to represent point processes other than renewal processes using phase type representation. Hence we can provide more general approximations for the arrival process (not just renewal approximations).

PH/PH/n queues are a natural generalization of M/M/n queues, and can be analyzed as quasi-birth and death processes. The transition matrix of the Markov chain, describing the queueing system has a special diagonal structure. Based on this structure several computational procedures have been developed (Lucantoni and Ramaswami, Neuts 87, Seelan, and Takahashi and Takami 76). This structure is retained even if the arrival process is not a renewal process. Consequently it is possible to extend the existing techniques to solve phase type queueing systems where the arrival process is non-renewal. If instead, we had general distributions, such analysis would be extremely difficult. This is very valuable since the arrival process at a station in a queueing network is unlikely to be a renewal process.

As stated earlier, the assumptions made in the decomposition approach are very strong. Yet, the estimates of performance measures obtained by this procedure are accurate

[Whitt(83b), Bitran and Tirupati(86)]. In particular, the decomposition approach provides good estimates of the first moment of the number of jobs in the queue. Theoretical justification of the observed accuracy require further study of two basic processes:
(1) Departure processes from queues and;
(2) Superposition of point processes and approximating point processes by renewal processes.

In this respect too, phase distributions are useful. As compared to general distributions, it is relatively easy to study the superposition process. Several parameters such as moments of the time until the nth arrival in the superposed process can be computed. This not only allows us to assess the accuracy of the approximations, but also provides valuable insights into the behavior of the superposed process. The accuracy of the approximations can be assessed either by studying the associated queueing systems or by comparing the counting process generated by the two point process being studied. Such comparisons would be extremely difficult for distributions that are not phase type.

For PH/PH/1 queues we can compute several performance measures such as waiting time distributions, probability of a departing customer leaving the system empty and idle period distributions. With two moment approximations these measures can not be computed. Please note that the knowledge of the idle period distribution and the probability of a departing customer leaving the system empty are adequate to characterize the stationary interval of interdeparture times.

There is definitely a price to be paid for getting a richer description of the processes in the queueing network. The computational procedures we are proposing require more effort than the parametric decomposition approach. It is unlikely to be suitable for evaluating communication networks with thousands of nodes. However, the computational burden may not be prohibitive if we are analyzing manufacturing systems with 50 machine centers.

4. CONTENTS OF THIS CHAPTER

In the next section we describe our algorithm for analyzing queueing networks with phase type distributions. In this section we do not derive any results. The objective of this section is to illustrate to the reader the steps of our procedure.

In section 6 we study a queueing system where the arrival process is the superposition of two or more phase renewal processes. For this queueing system we examine alternate solution strategies. We derive the moments of the waiting time and the number in system, as observed by each customer class. We derive a system of linear differential equations which describe the waiting time distribution. Since the waiting time distributions are difficult to compute we propose approximations. For this purpose we derive the tails of the waiting time distribution and the number in queue.

In section 7 we look at approximations for superpositions of renewal processes. The size of the exact representation grows exponentially in the number of processes superposed. However, as the number of superposed processes increases, the arrival process begins to look like a Poisson process. Therefore, it should be possible to develop approximations which have small representations, yet emulate the behavior of the superposed process. The basic idea is to approximate the superposition process by a phase process in which there are multiple 'absorption' states. Each time we reach an 'absorption' state it is equivalent to an arrival at the queue. As soon as we reach an 'absorption' state we return to the transient states . Associated with each absorption state is a return probability vector. By allowing the return probabilities to be different for different absorption states, we can model non-renewal arrival processcs. We refer to these as general phase processes (GPH). We derive a general procedure for computing the variance of the stationary interval . We give an iterative procedure for computing the moments of the distribution of the time until the nth arrival. At this point we do not have a formal procedure for approximating superposed processes. But we show through a few numerical examples the feasibility of the proposed approach.

Finally in section 8, we derive the stationary interval of the departure process and the covariance between adjacent departure intervals. For a few queues we have computed the correlation between adjacent departure intervals. These correlations are very small suggesting that the departure process may be well approximated by a renewal process.

5. QUEUEING NETWORKS WITH PHASE TYPE DISTRIBUTIONS

Our objective is to develop computationally feasible and accurate procedures for analyzing phase type queueing networks. In this section we illustrate the steps in our approach to solving queueing networks. Since we require all interarrival times and service times to have phase type distributions the first step is to obtain a phase representation of (a) the external arrivals to the network, and (b) the service process at each node. This is a very important component of the model. However, in this thesis we will not concern ourselves with this issue. We merely note that one may approximate an empirical distribution by a phase distribution by matching the first few moments. Presently, in the parametric approach, only the first and second moment of the service distribution and of the external arrivals are taken into account. Thus the scheme we are proposing at the least can be viewed as a parametric approach that takes into account moments higher than the second moment. Once the description of the relevant distributions is obtained, we need to analyze the queue at each machine center or node. We propose the following scheme for analyzing the queues at each machine center:
(1) The individual arrival streams assumed to be phase renewal processes are superposed.

> (a) If the exact representation of the superposed arrival process is very large then we find a general phase process with smaller representation which has the same stationary interval and asymptotic second moment as the superposed process. A general phase process need not be a renewal process. In section 7 we discuss in greater detail the approximations for superposed processes.

(2) The queue with the superposed arrival stream is analyzed. We compute the moments of the waiting time, waiting time distribution and the idle period distribution. In the parametric decomposition approach only the first moment of the waiting time distribution is computed.

(3) Finally, the phase distribution describing the stationary interval of the departure stream is computed.

To illustrate the algorithm we now work through an example. For this purpose consider a machine center somewhere in the middle of a network (Fig 5.1). Let us assume that jobs arrive to this machine center from two sources. We call the two streams x and y.

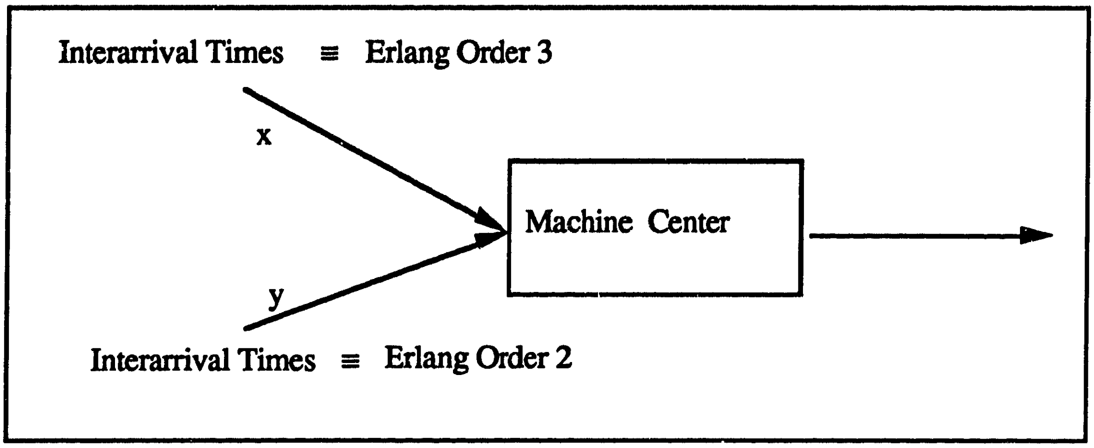

Interarrival Times ≡ Erlang Order 3
x
Machine Center
y
Interarrival Times ≡ Erlang Order 2

Figure 5.1 Arrivals to the Machine Center

Let the interarrival times of x type jobs be distributed as Erlang order 3 with a mean interarrival time of 3, and the interarrival time of y type jobs be distributed as Erlang order 2 with a mean interarrival time of 2. Further, assume that the service time for both these jobs is distributed as Erlang order 2 with a mean of 1. For these parameters, the utilization of the machine center is 0.833. Let [a, T_x, $T^o{}_x$], [b, T_y, $T^o{}_y$], and [c, S , S^o] be the phase representations of x type, y type arrivals and the service process respectively. Then these matrices will be as follows:

$$T_x = \begin{vmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{vmatrix} \qquad T_x^0 = \begin{vmatrix} 0 \\ 0 \\ 1 \end{vmatrix}$$

$$T_y = \begin{vmatrix} -1 & 1 \\ 0 & -1 \end{vmatrix} \qquad T_y^0 = \begin{vmatrix} 0 \\ 1 \end{vmatrix}$$

$$S = \begin{vmatrix} -2 & 2 \\ 0 & -2 \end{vmatrix} \qquad S^0 = \begin{vmatrix} 0 \\ 2 \end{vmatrix}$$

The corresponding initial probability vectors are : a: (1,0,0); b: (1,0) ; c: (1,0).

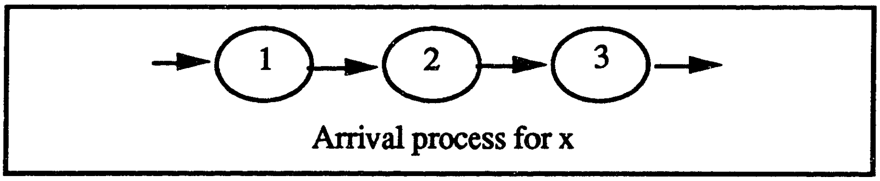


Figure 5.2 Arrival Phase for x type jobs

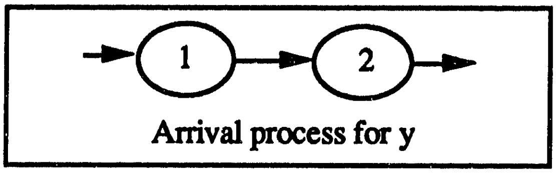


Figure 5.3 Arrival Phase for y type jobs

STEP 1: Superpose the arrival process

Figure 5.2 depicts the Markov chain of the superposed arrival process. The nodes in the superposed process are partitioned into blocks. Movement from one block to the next block corresponds to an arrival. Each node in the Markov chain is doubly indiced as (i,j). i is the state of arrival process y, and j is the state of arrival process x. The nodes within a block are numbered lexicographically: (1,1) = 1, (1,2) = 2, (1,3) = 3, (2,1) = 4 etc. The infinitesimal generators for the superposed process are given by the following matrices

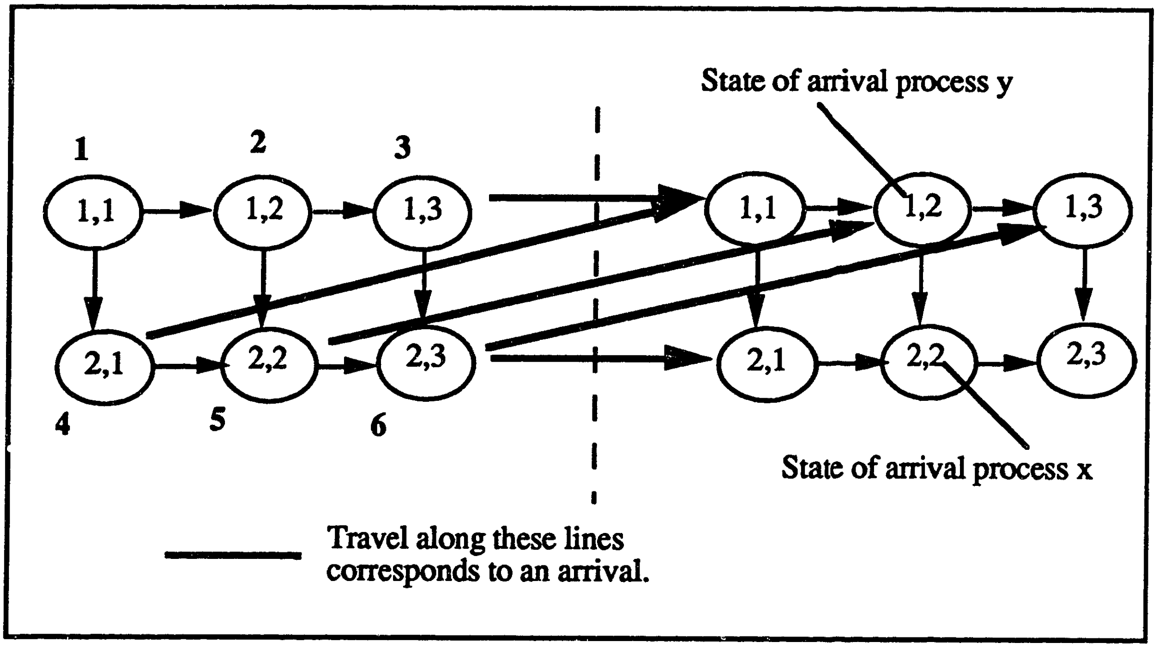

State of arrival process y
1
2
3
1,1
1,2
1,3
1,1
1,2
1,3
2,1
2,2
2,3
2,1
2,2
2,3
4
5
6
State of arrival process x
Travel along these lines corresponds to an arrival.

Figure 5.4 Superposed Arrival Process

$$T = \begin{vmatrix} -2 & 1 & 0 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 & 1 & 0 \\ 0 & 0 & -2 & 0 & 0 & 1 \\ 0 & 0 & 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & -2 \end{vmatrix} \qquad T^0 = \begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{vmatrix}$$

The matrix T gives the rates within the same block, and T^o gives the rates into the next block.. Matrix T is the Kronecker sum of matrices T_x and T_y . We denote this relationship as $T = T_y \oplus T_x$. Matrix $T^o = T^o{}_y\, b \oplus T^o{}_x a$.

STEP 1a : Approximations for the superposed process If the size of the superposed process is very large then we have to find a suitable approximation for the superposed. For this purpose we need to know the variance of the stationary interval and the asymptotic variance of the superposed process. The derivation of the distribution of the superposed process is given in section 7. In that section we also show how to compute the variance of the time until the nth arrival of the superposed process. Here we only show how to compute the stationary interval distribution and the asymptotic variance.

Let us assume that the arrival process has been going on for a long time. Further assume that at time zero we have an arrival. The distribution of the time until the next arrival is the stationary interval distribution of the superposed process. In section 7 we show that the stationary interval is a phase distribution with the following representation $[\beta , B', B^0]$. $B' = T^o$ and $B^0 = T^o e$.

$$B' = \begin{vmatrix} -2 & 1 & 0 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 & 1 & 0 \\ 0 & 0 & -2 & 0 & 0 & 1 \\ 0 & 0 & 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & -2 \end{vmatrix} \qquad B^0 = \begin{vmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 2 \end{vmatrix}$$

β, the initial probability vector equals $P_x\, a \otimes w + P_y\, v \otimes b$. In this expression P_x (P_y) is the probability that the arrival at time zero was an x (y) type arrival. Since the probability of an arrival being an x type arrival is proportional to the arrival rate of x type jobs, in our example P_x equals 0.4 and $P_y = 0.6$. Consider the phase process describing the arrival process for x type jobs (Fig. 5.2). Each time an x type job arrives the phase process returns to state 1. Thus the arrival process circulates through states 1-2 -3. v is the vector of steady state probabilities of the Markov chain that depicts this circulation. Vector w is similarly defined for y type jobs. In our example v = (1/3, 1/3, 1/3) and w = (1/2, 1/2). We use the symbol $\otimes$ to denote Kronecker multiplication. Thus $\beta = P_x\, a \otimes w + P_y\, v \otimes b = (0.4, 0.2, 0.2, 0.2, 0, 0)$

It is well known that the asymptotic limit of the scv of the superposed process equals $P_x c_x + P_y c_y$. In this expression c_x (c_y) is the scv of the interarrival distribution of job x (y). For more details regarding the asymptotic limit please see Whitt (82). For our example the asymptotic scv = 0.4 x 0.333 + 0.6 x 0.5 = 0.4333

As stated earlier, in section 7 we also show how to compute the variance of the time until the nth arrival at the queue (either x or y type) starting from time zero. This enables us to compute the correlation between adjacent interarrival intervals. The basic idea behind the approximation scheme is to design a phase process which has a representation that is smaller than that of the actual superposed process but has the same stationary interval and asymptotic scv, and if possible the same correlation between adjacent intervals.

Step 2: Queueing Analysis: The queue process at the machine center is depicted by a Markov chain that has the following generator:

$$Q = \begin{vmatrix} A0 & C0 & 0 & 0 & 0 \\ A1 & B & C & 0 & 0 \\ 0 & A & B & C & 0 \\ 0 & 0 & A & B & C \end{vmatrix}$$

Where A0, A, A1, B, C, C0, O are matrices.

$$\begin{aligned} A0 &= T_y \oplus T_x = T \\ A1 &= I \otimes S^o \\ A &= O_{Ty} \oplus O_{Tx} \oplus S^o c \\ B &= T_y \oplus T_x \oplus S \\ C &= T_y{}^o b \oplus T_x{}^o a \oplus O_s \\ C0 &= (T_y{}^o b \oplus T_x{}^o a)c \end{aligned}$$

The Markov chain depicting the queueing process is given in Figure 5.5. This chain is divided into blocks corresponding to the number in the system. Within each block, the nodes have 3 indices (i,j,k) corresponding to the state of the arrival process y, arrival process x, and the service process. Once again, the nodes are numbered lexicographically: (1,1,1) =1, (1,1,2) = 2, (1,2,1) = 2, etc. Matrix B gives the rates within each block, matrix C gives the rates into the block on the right (corresponds to an arrival), and matrix A gives the rates into the block on the left (corresponds to a service completion).

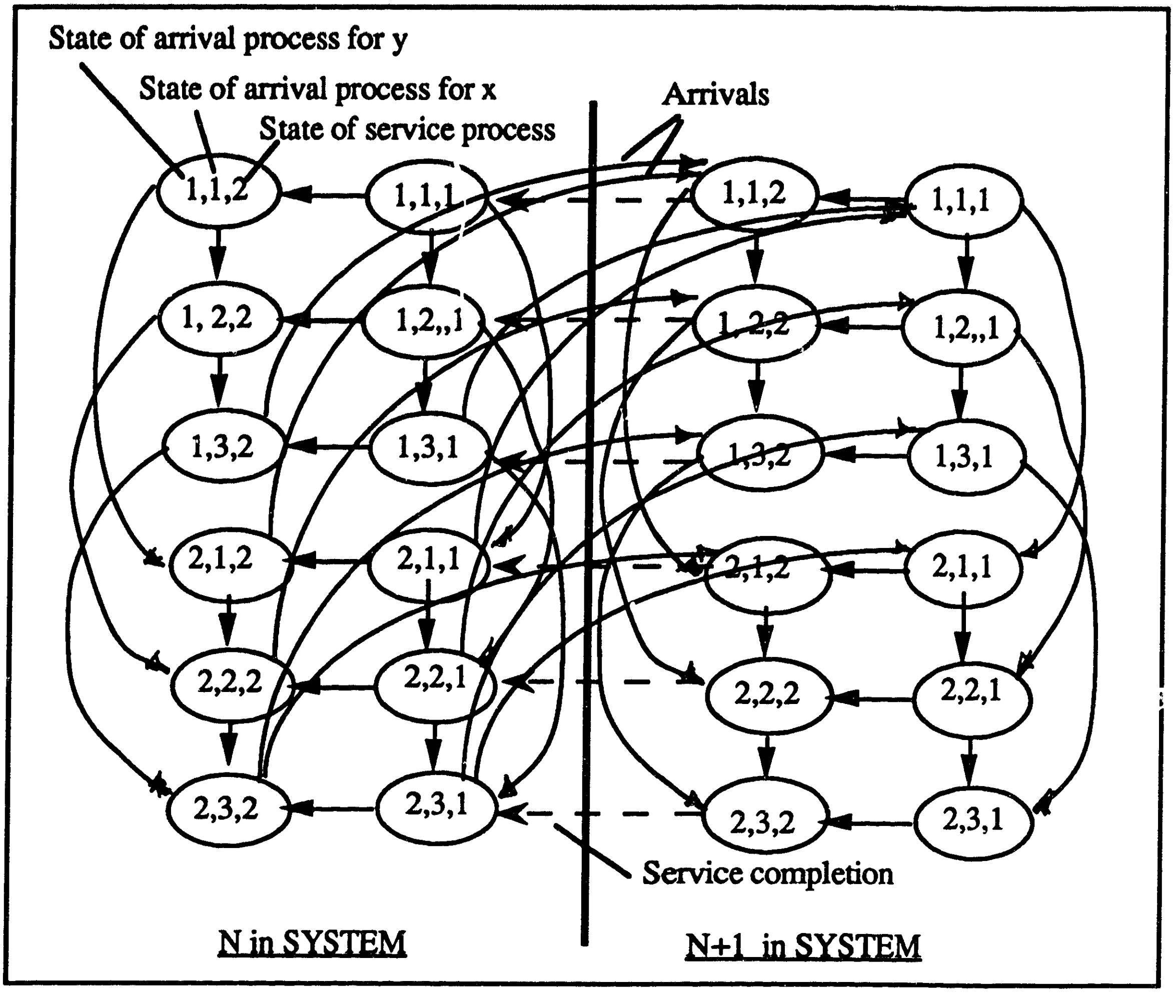


Figure 5.5 Markov chain of queueing system

The steady state probabilities of the number in system and the performance measures are obtained by following these 3 steps:

(1) Find the matrix R which solves the equation $R^2A + RB + C = 0$. The matrix R is computed through a simple recursive process. Let D be a diagonal matrix formed with the diagonal elements of B. Since the diagonal elements of B are strictly negative, D has an inverse. We can rewrite the polynomial equation as

$$R = -R^2AD^{-1} - RBD^{-1} - CD^{-1} .$$

Let R_n be the value of R at the nth iteration, of the iterative algorithm :

$$R_n = -R_{n-1}{}^2AD^{-1} - R_{n-1}BD^{-1} - CD^{-1}.$$

The algorithm is initialized by setting R_0 equal to CD^{-1}. The iterative procedure is terminated when max $\left|R_n^{i,j} - R_{n-1}^{i,j}\right| < \varepsilon$. Where $R_n^{i,j}$ is the i,j element of matrix R_n and e is a small constant.

(2) We let U_n be the vector of steady state probabilities for the block which corresponds to the system having n jobs (Figure 5.5). Then U_0 and U_1 are obtained by solving the system of linear equations

$$U_0\, A0 + U_1\, A1 = 0;$$
$$U_0\, C0 + U_1\, (B + R\, A) = 0\, ;$$
$$U_0\, e \;= 1 - \rho\, .$$
$$\text{For } n > 2,\;\; U_n \;= U_1 R^{n-1}.$$

ρ is the utilization of the server.

(3) Performance Measures: The average number in system is given by :

$$U_1(I - R)^{-2}e,$$

and the second moment of the number in system is

$$U_1(\, I + 2R(I - R)^{-1})\; (I-R)^{-2}e.$$

We can also compute the distribution and the moments of the number of jobs found in the system by each job class. For instance the the probability that job type x finds n jobs in the system upon arrival is given by

$$V_x(n) \;=\; L_x\;\; U_n\, [\, O_y \oplus T_x{}^o a \oplus O_s]\; [\, e_T \otimes I_s]\, .$$

Where L_x is the mean interarrival time for job type x . The distribution of the number of jobs an arriving job finds is given by:

$$V(n) \;= (L_x + L_y)^{-1}\; (L_x L_y)\, U_n\, [T_y{}^o b \oplus T_x{}^o a \oplus O_s]\; [\, e_T \otimes I_s]\, .$$

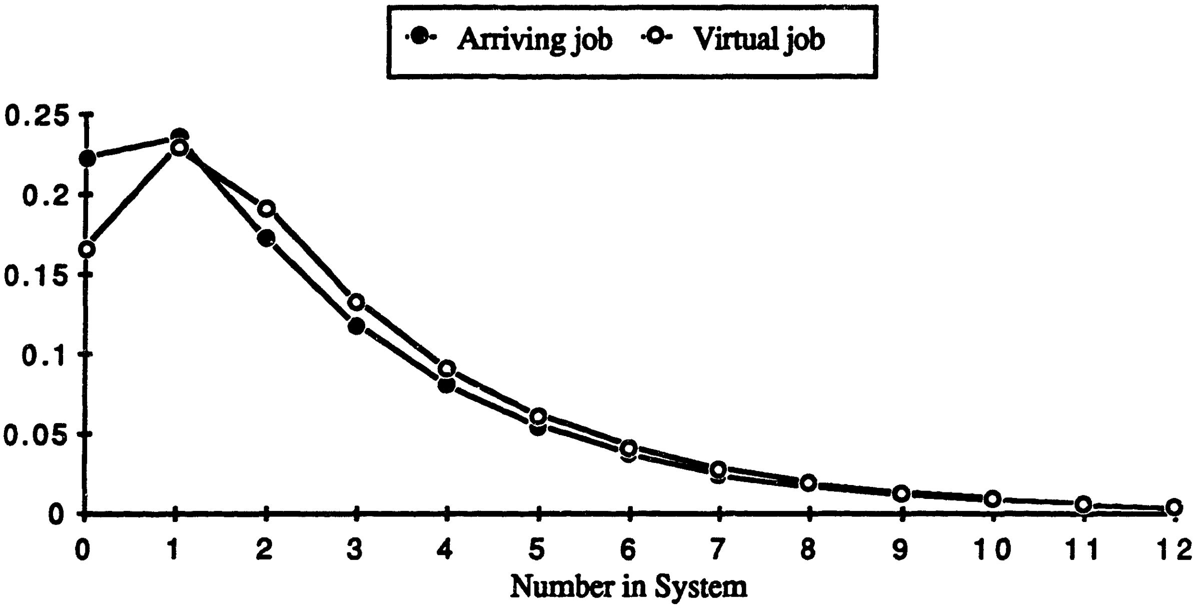
The distributions of the number in queue as seen by a virtual job and by an arriving job in our example are graphed in Figure 5.6.

Step 3. Departure Process: We approximate the departure process by its stationary interval. To derive the stationary interval of the departure process we need to know the probability of a departing job leaving the system empty and the state of the arrival process given that the system just become empty. In other words we need to know the probability of the system becoming empty and the idle period distribution. Since the probability of a departing job leaving the system empty is the same as that of an arriving job finding the system empty, V(0) is the probability of departing job leaving the system empty.

$$V(0) = (L_x + L_y)^{-1}\, (L_x L_y)\, U_0\, (T_y{}^o \oplus T_x{}^o\,)\, e.$$

For our example the probability of a departing job leaving the system empty is 0.224

Figure 5.6 Number in System



Let å denote the state of the arrival process at the instant the system becomes empty. Then å equals $u_1[I_T \otimes S^o] / \{u_1[I_T \otimes S^o]\, e\}$. Thus the idle period is a phase distribution with the following representation : [å, $T_y \oplus T_x$], $(T_y{}^o \oplus T_x{}^o)\, e$]. Since the stationary interval of the departure process is with probability V(0) the concatenation of the idle period and the service process, and with probability 1 -V(0) the same as the service process, the stationary interval is a phase with the following representation: [ø, G, G^o]. Where

$$ø = ç(å\,, 0) + (1-ç)\,(0,\beta\,)$$

$$G = \begin{vmatrix} (T_1 \oplus T_2) & (T_1{}^o \oplus T_2{}^o) e\beta \\ O & S \end{vmatrix}$$

$$G^o = \begin{vmatrix} O \\ S^o \end{vmatrix}$$

Returning to our example, the state of the arrival process at the instant the system becomes empty was computed to be (0.1135, 0.1544, 0.1807, 0.1699, 0.1895, 0.1920). Hence the idle period is distributed as the following phase:
Initial probability å = (0.1135, 0.1544, 0.1807, 0.1699, 0.1895, 0.1920).

$$B^* = \begin{vmatrix} -2 & 1 & 0 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 & 1 & 0 \\ 0 & 0 & -2 & 0 & 0 & 1 \\ 0 & 0 & 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & -2 \end{vmatrix} \qquad B^{*0} = \begin{vmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 2 \end{vmatrix}$$

Since the probability of the system becoming empty at a departure instant is 0.224, the stationary interval of the departure process has the following representation:
ø = (.0254, .0346, .0405, .0381, .0424, .043, .776, 0)

$$G = \begin{vmatrix} -2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & -2 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -2 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{vmatrix} \qquad G^0 = \begin{vmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{vmatrix}$$

In section 8 we also derive the correlation between adjacent departure intervals. We have computed this correlation for a some queues and found the correlation to be extremely small (≈ 0.03). Suggesting that the stationary interval is a good descriptor for the departure process.

This completes the analysis of the queueing process at a node in the network. The departure process from this node becomes the arrival process to some other node, and the same analysis is carried out at the other nodes. Figure 5.7 summarizes the steps of the decomposition procedure.

In the next section we derive the formulae used to compute the performance measures of the queue.

6. COMPUTATIONAL PROCEDURES FOR PH,PH/PH/1 QUEUES

Consider a machine center to which jobs arrive from two other machine centers. We will assume that the departure process from each machine center is a phase renewal process. Let the inter-departure times from machine center i be a phase distribution characterized by [α_i, T_i, T^o_i]. Let the service process at machine center 3 be the phase distribution [β, S, S^o]· Let Ti be an m_i x m_i matrix and S an n x n matrix. Further assume that the representations are irreducible.

Consider the Markov chain describing the queueing process. This chain has the states (n,i,j,k). n signifies the number in the system, i the state of the arrival process 1, j state of the arrival process 2, and k is the state of the service process. The states are listed in that lexicographic order.

The generator for this Markov process has the following structure

$$Q = \begin{vmatrix} A0 & C0 & 0 & 0 & 0 \\ A1 & B & C & 0 & 0 \\ 0 & A & B & C & 0 \\ 0 & 0 & A & B & C \end{vmatrix}$$

Where A0, A, A1, B, C, C0, O are matrices.

$$A0 = T_1 \oplus T_2$$
$$A1 = I \otimes S^o$$
$$A = O_{T1} \oplus O_{T2} \oplus S^o\beta$$
$$B = T_1 \oplus T_2 \oplus S$$
$$C = T_1{}^o\alpha_1 \oplus T_2{}^o\alpha_2 \oplus O_s$$
$$C0 = (T_1{}^o\alpha_1 \oplus T_2{}^o\alpha_2)\beta$$

$A \oplus B$ is the Kronecker sum of the matrices A and B

$A \otimes B$ is the Kronecker product of the matrices A and B

O_s (O_{T1}) is a zero matrix of dimension S (T_1)

PROPOSITION 1:

Let $A^* = A + B + C$, and π solve $\pi A^* = 0$, $\pi e = 1, \pi \geq 0$. Then the above Markov chain is positive recurrent if $\pi Ce < \pi Ae$.

Proof : Theorem 1.7.1 page 32 Neuts(81)

COROLLARY 1

Let the mean service time be μ and the mean inter-arrival times be L_1 and L_2 for the two arrival streams. If $\mu^{-1} > L_1^{-1} + L_2^{-1}$ then the Markov chain is positive recurrent.

Proof : Let $A^* = A + B + C$, and

$A = O_{T1} \oplus O_{T2} \oplus S^o\beta$

$B = T_1 \oplus T_2 \oplus S$

$C = T_1^o\alpha_1 \oplus T_2^o\alpha_2 \oplus O_s$

Therefore

$A^* = (T_1 + T_1^o\alpha_1) \oplus (T_2 + T_2^o\alpha_2) \oplus (S + S^o\beta).$ [eq 1]

Let f solve $f(S + S^ob) = 0$, $fe = 1$, $f \geq 0$, then f is the vector of steady state probabilities of the phase renewal process corresponding to the service processes. Consider the matrix $-S^{-1}$, element (i,j) of this matrix is the expected amount of time spent in state j before reaching the absorption state, given that we begin in state i. Hence $f = (\beta S^{-1}) / (\beta S^{-1}e)$. Similarly let g_i solve $g_i (T_i + T_i^o\alpha_i)$, $g_i e = 0$, $gi \geq 0$.

Implying $g_i = (\alpha_i T_i^{-1}) / (\alpha_i T_i^{-1} e)$.

By [eq 1] and property of Kronecker sum of matrices (Bellman 60) we have

$(g_1 \otimes g_2 \otimes f) A^* = 0$, $(g_1 \otimes g_2 \otimes f)e = 1$, $g_1 \otimes g_2 \otimes f \geq 0$.

Therefore $\pi = g_1 \otimes g_2 \otimes f$

$\pi Ae = (g_1 \otimes g_2 \otimes f)(O_{T1} \oplus O_{T2} \oplus S^o\beta)e = fS^o\beta e = fS^o$

Since $Se + S^o = 0$. $S^{-1} S^o = -e$. Implying that $fS^o = -(\beta S^{-1}e)^{-1}$

Therefore $\pi Ae = -(\beta S^{-1}e)^{-1} = \mu^{-1}$.

Similarly $\pi Ce = L_1^{-1} + L_2^{-1}$. ♦

Hence forth we will assume that the Markov chain is positive recurrent and aperiodic.

The steady state probabilities for this Markov chain are given by a nonnegative vector U that solves $UQ = 0$, $Ue = 1$, $U \geq 0$. For convenience partition the vector U in to blocks corresponding to the number in system; i.e $U = (u_0, u_1, u_2, .. u_j ..)$

6.1. Solution Strategies

There are at least two strategies for finding the steady state probabilities for our queueing system:

(1) Iterative Procedures;

(2) Matrix Geometric Procedures.

We use the latter technique for finding the steady state probabilities and so provide only a brief review of the iterative procedures. For more details of these techniques the reader is referred to paper by Seelan (86), and Takahashi and Takami (76).

(1) ITERATIVE PROCEDURES:

In this approach the system of equations $UQ = 0$, $Ue = 1$, $U \geq 0$ are solved using iterative procedures such as successive over relaxation or aggregation-disaggregation (Tijms 86, Takahashi and Takami 76, Seelan 86). Seelan has developed and implemented an iterative procedure for Ph/Ph/c queues which incorporates the basic principles of successive overrelaxation and aggregation-disaggregation. In aggregation - disaggregation algorithms all the micro-states corresponding to a specific number of customers in the system are lumped together and viewed as a macro state. The system of equations describing the steady state probabilities of the Markov chain with only macro-states has a structure like that of an M/M/c queue. At each step of the aggregation-disaggregation algorithms, a solution is found for the aggregated states, and then for each block of disaggregated micro-states an attempt is made to derive a solution that is consistent with aggregate solution. Seelan adapts the basic ideas of disaggregation-aggregation algorithms by making the relaxation factor in the point successive over relaxation method depend on the aggregate solution. With his algorithm, Seelan analyzed 30,000 Ph/Ph/c queues, with finite and infinite waiting rooms. The largest queue he analyzed involved finding a solution for a system of 150,000 linear equations. In every case his algorithm converged.

(2) MATRIX GEOMETRIC TECHNIQUES :

Let R be the minimal non negative solution to the following Matrix polynomial equation:

$$R^2A + RB + C = 0 \qquad \text{[eq 2]}$$

R is a square matrix of size $m_1 \times m_2 \times n$. The element (i,j,k; p,q,r) is the expected amount of time spent in state (p,q,r) with n +1 jobs in the system before the first return to a state with n jobs in the system. The chain begins in state (i,j,k) with n jobs in the system (Neuts 81). Since the matrices S, T_1 and T_2 are irreducible we will assume that R is strictly positive. Since the queue is positive recurrent and aperiodic the spectral radius of R lies strictly between 0 and 1 (Neuts 81).This approach makes use of the fact that the microstates, when partitioned into blocks on the basis of the number in queue are related to each other through the equation

$$u_n = u_1 R^{n-1} \quad \text{(Chung 67, Neuts 81)} \qquad [\text{eq } 3]$$

Hence, instead of solving the large system of equations, $UQ = 0$, $Ue = 1$, $U \geq 0$, iterative techniques are used to solve for the Matrix R (Asmussen 87, Lucantoni and Ramaswami 85, Latouche 87).

Recall that R is the minimal nonnegative solution to the matrix polynomial equation $R^2A + RB + C = 0$. The diagonal elements of matrix B are strictly negative. Let D be a diagonal matrix formed with the diagonal elements of B. Then D has an inverse. We can rewrite the polynomial equation as $R = -R^2AD^{-1} - RBD^{-1} - CD^{-1}$. Since the spectral radius of R is less than 1 a simple iterative technique described below converges to the correct value of R (Asmussen):

Let R_n be the value of R at the nth iteration, then the iterative algorithm is:

$R_n = -R_{n-1}{}^2AD^{-1} - R_{n-1}BD^{-1} - CD^{-1}$. The algorithm is initialized by setting R_0 equal to CD^{-1}.

This method has several advantages as compared to the iterative techniques. First, the storage space and the computational burden are significantly lower in this approach. Second, several queueing performance measures such as the moments of the number in queue can be expressed in terms of the matrix R.

In the remainder of this section we assume that the steady state probabilities have been computed using one of the techniques described above. We then show how to compute the performance measures of the queue based on this information.

6.2. Distribution Of The Number In Queue

The distribution of the number as seen by a virtual customer is given by the vector U. We can also compute the distribution of the number of jobs as seen by each job class. Let Vj(n) be a vector such that its kth element is the probability that an arriving job of class j finds n jobs in the system and the service process is in phase k.

PROPOSITION 2 :

$$Vj(0) = Lj\ u_0 [\partial j(1)T_1{}^o\alpha_1 \oplus \partial j(2)T_2{}^o\alpha_2]\, e \qquad [\text{eq } 5]$$

$$Vj(n) = Lj\ u_n [\partial j(1)T_1{}^o\alpha_1 \oplus \partial j(2)T_2{}^o a_2 \oplus O_S] [\, e_T \otimes I_S] \qquad [\text{eq } 6]$$

e_T = column vector of 1s of dimension m1 x m2

I_s = Identity matrix of size n

$\partial j(h)$ = 1 if j = h ; other wise 0

Lj = Mean inter-arrival time for arrival stream j.

If we allow V(n) to be the vector of probabilities that an arriving customer (arbitrary class) finds n in the system . Then :

$$V(n) = (L_1 + L_2)^{-1} (L_1 L_2)\, u_n [T_i{}^o\alpha_1 \oplus T_2{}^o\alpha_2 \oplus O_s] [e_T \otimes I_s] \qquad \text{[eq 7]}$$

Proof: Consider [eq 5]. $[\partial j(1)T_1{}^o\alpha_1 \oplus \partial j(2)T_2{}^o\alpha_2]\, e$ is the arrival rate of job class j given that the system is empty and u_0 is the probability that the system is empty. Lj , the mean inter-arrival time is the normalizing constant. [eq 6] and [eq 7] are similarly derived. Note that Vj(0) is a scalar.

6.3. Average Number In System

If we explicitly solve for the steady state probabilities than it is straight forward to compute the moments of the number in queue as seen by a virtual customer, or an arriving customer. The moments can also be expressed in terms of the matrix R. Recall that the spectral radius of R lies strictly between 0 and 1. Let N denote the expected number in system and N^2 the second moment of the expected number in system. Then :

$$N = \sum_{n=1}^{\infty} n u_n e = \sum_{n=1}^{\infty} n u_1 R^{n-1} e = u_1 (I - R)^{-2} e \qquad \text{[eq 8]}$$

$$N^2 = \sum_{n=1}^{\infty} n^2 u_1 R^{n-1} e = u1(I + 2R(I - R)^{-1}) (I-R)^{-2} e \qquad \text{[eq 9]}$$

6.4. Average Number Of Jobs Found By Each Class

Let Nj be the average number of jobs found by a job belonging to class j. Then

$$Nj = \sum_{n=1}^{\infty} n Vj(n) e = Lj\, u_1 [I - R]^{-2} [\partial j(1)T_1{}^o\alpha_1 \oplus \partial j(2)T_2{}^o\alpha_2 \oplus O_s]\, e \qquad \text{[eq 10]}$$

6.5. Waiting Time Distributions

From the steady state probabilities one can readily compute the moments of the waiting time as seen by (1) virtual customer (2) an arriving job and (3) by job class.

However it is difficult to compute the waiting time distribution. The waiting time distribution can be analyzed as a Markov chain with a single absorption state 0 (number in system) and infinitely many transient states (n,k). Where n is the number in system and k is the state of the service phase. Since under the first come first served priority rule the waiting time is not influenced by subsequent arrivals we need not worry about the state of the arrival process.

The generator for the Markov chain corresponding to the waiting time distribution has the following structure :

$$Q^* = \begin{vmatrix} 0 & 0 & 0 & 0 & . \\ A'e & B' & 0 & 0 & . \\ 0 & 0 & A' & B' & . \\ 0 & 0 & 0 & A' & . \end{vmatrix}$$

Where $A' = S^o\beta$

$B' = S$

The starting probabilities for this Markov chain are given by Vj(n) and V(n) for the waiting time distribution for job class j and for an arbitrary job , respectively. In the case of a virtual job the starting probability is given by $u_n[\, e_T \otimes I_s \,]$. This representation of is useful primarily for deriving the tails. It is otherwise computationally intractable.

Perhaps the easiest procedure for computing the waiting time distribution is through a system of linear differential equations.

PROPOSITION 3

Let $Z_{ij}(x)$ be a vector of size n and $W_{ij}(x)$ a scalar such that

$$W'_{ij}(x) = - W(x)\,[\, T1 \oplus T2]_{ij} - Z_{ij}(x)\, S^o$$

$$Z'_{ij}(x) = Z_{ij}(x)S + W(x)\,[T_1{}^o a_1 \oplus T_2{}^o a_2]_{ij}\beta. \qquad \text{[eq 11]}$$

$$Z_{ij}(0) = 0 \;\; ; \text{ and } \;\; W(0) = u_0$$

Where : W(x) is a vector whose elements are $W_{ij}(x)$. W'(x) is the derivative of W(x) with respect to x. Z'(x) is the derivative of Z with respect to x. $[\, T1 \oplus T2]_{ij}$ is the (i,j)th element of the matrix $[\, T1 \oplus T2]$.

Then the probability of the system having x units of work is given by:

$$W^*(x) = W(x)e \qquad \text{[eq 12]}$$

The probability that an arriving customer waits x units is given by

$$E(x) = (L_1 + L_2)^{-1} (L_1L_2) W(x) [T_1{}^o\alpha_1 \oplus T_2{}^o\alpha_2] e \qquad \text{[eq 13]}$$

Proof:

We will derive this result via the Chapman Kolmogorov equations. The term $W_{ij}(x)$ has the probabilistic interpretation that a virtual job finds x units of work in the system and the arrival process is in state (i,j). Consequently:

$$W'_{ij}(x) = \sum_{p\ q} (W_{pq}(x)\text{Prob(of going from state p,q to i,j)}$$

$$+ \int_0^x \sum_{p\ q} W_{pq}(u) \text{ Prob[arrival occurs when in state p,q and arrival leads into state i, j]}$$

$$\text{Prob[additional work brought in by new arrival} = x - u\text{] du}$$

This can be rewritten as :

$$W'_{ij}(x) = - W(x) [T1 \oplus T2]_{ij}$$

$$- \int_0^x W(x) [T_1{}^o a_1 \oplus T_2{}^o a_2] \beta EXP(S(x-u))So\ du \qquad \text{[eq 14]}$$

$$\text{Let } Z_{ij}(x) = \int_0^x W(x) [T_1{}^o a_1 \oplus T_2{}^o a_2] \beta EXP(S(x-u))So\ du$$

$$\text{Then } Z'_{ij}(x) = Z_{ij}(x)S + W(x)[T_1{}^o\alpha_1 \oplus T_2{}^o\alpha_2]_{ij}\beta. \qquad \text{[eq 15]}$$

Eq 14 and Eq 15 prove the proposition.

Although all the moments of the waiting time can be computed it is not easy to compute the waiting time distribution . Hence it necessary to develop approximations for the waiting time distribution. We are in a particularly good position to develop approximations, because of our ability to compute the moments of the waiting time and the probability of an arriving customer finding the system empty. The quality of the approximations can be improved if we also know the tail of the waiting time distribution.

The knowledge of the tail is also useful if we wish to assess the probability of a job spending a very long time in the queue. Typically service level constraints are stated in terms of the tails of the waiting time distribution.

The system of equations $UQ = 0$, $Ue = 1$, $U \geq 0$ is of infinite dimension for a queue with infinite waiting room. In order to solve these equations using iterative techniques such as those used by Seelan, we need to truncate the system. The knowledge of the tails of the number in queue, is useful for determining the error resulting from truncation. For all these reasons in the next sub-section we will derive the tails of the number in queue and the waiting time distribution.

6.6 Tails Of The Number In System

In this sub-section we will derive the tail distributions by employing techniques developed by Takahashi (81) .

We already know that $u_n = u_1 R^{n-1}$, and R is a positive matrix. Thus by Perron - Frobenius theory (Gantmacher 59)

$$\text{Lim}_{n \to \infty} \; u_n \to K^* ß^{n-1} + O(ß^n) \qquad \textbf{[eq 16]}$$

Where ß is the Perron -Frobenius eigenvalue of R and K* is a function of u_1 and the right and left Frobenius eigenvectors of R.

For the purpose of deriving K* and ß we will state with out proof two technical results.

PROPOSITION 4 :

Let $\quad M(t) = d^{-1} (t^2 A + \iota (B + dI) + C).$ **[eq 17]**

Here t and d are positive scalars. d is chosen to be sufficiently large so that M(t) is non-negative. Since all the representations are irreducible we will assume that there exists an integer n such that $[M(t)]^n$ is positive. Let $ø(t)$ be the Perron Frobenius eigenvalue of M(t). Then the unique solution to $ø(t) = t$ in the open interval (0,1) is the Perron Frobenius eigenvalue of R.

Proof: Neuts(1980) lemma 1.3.4 page 17.

Recall that the Laplace -Stieltjes transform of a phase distribution with representation [a, T, T^o] is given by $F^*(s) = a(sI - T)^{-1}T^o$. We will be assuming that s is real. Clearly $F^*(s)$ is a monotone function of s. For $s> 0$ $F^*(s)$ is less than 1. $F^*(s)$ takes on finite values for $s > y$, where y is some negative number. For s in $(y,0)$ $F^*(s) > 1$.

Given any $t > 0$ let $\Omega(t)$ be such that $F^*(\Omega(t)) = 1/t$. Since $F^*(.)$ is monotone there is one to one correspondence between t and $\Omega(t)$. If $t > 1$ $\Omega(t) > 0$ and if $t< 1$ then $\Omega(t) <0$.

Further, let $g(t) = a\,(\Omega(t)I - T)^{-1}$

PROPOSITION 5:

For $t > 0$

(1) $g(t)$ is positive.

(2) For the matrix $(t\,T^o a + T)$ $g(t)$ is an eigenvector and the corresponding eigenvalue is $\Omega(t)$ i.e., $g(t)(t\,T^o a + T) = \Omega(t)g(t)$.

Proof: The positivity of $g(t)$ follows from the fact that $(sI - T)^{-1}$ is positive for $s > y$. Claim (2) is derived by direct substitution. (For more details please refer to Proposition 2.1 Takahashi 81).

Let $\Omega_i(t)$, and $g_i(t)$ correspond to arrival process i ; and $\Omega_s(t)$ and $g_s(t)$ correspond to the service process.

PROPOSITION 6:

The Perron-Frobenius eigenvalue of R denoted by ß is such that $0 < ß< 1$ and

$$\Omega_1(1/ß) + \Omega_2(1/ß) + \Omega s(ß) = 0 \qquad \text{[eq 18]}$$

Proof: Define the matrix $M^*(t) = (tA + B + (1/t)C)$. Then $M^*(t)$ can be rewritten as :

$$M^*(t)=(T_1 + (1/t)T_1{}^o a_1) \oplus (T_2 + (1/t)T_2{}^o a_2) \oplus (S +tS^o b). \qquad \text{[eq 19]}$$

(Note that this rearrangement of terms is the same as that used in eq 1). Recall from Proposition 1 that if (∂_i, f_i) are the eigenvalue and eigenvector for matrix X_i, then for $X_1 \oplus X_2$, $\partial_1 + \partial_2$ is an eigenvalue and $f_1 \otimes f_2$ is the corresponding eigenvector (Bellman 61).

The above fact, together with Proposition 5 leads us to the result that an eigenvalue of $M^*(t)$ is $\Omega_1(1/t) + \Omega_2(1/t) + \Omega s(t)$, and the corresponding eigenvector is $g_1(1/t) \otimes g_2(1/t) \otimes g_s(t)$. We denote this eigenvector by $g(t)$ and the eigenvalue by $\Omega(t)$. Note that for $t > 0$, $g(t)$ is positive.

Any eigenvector of $M^*(t)$ is also an eigenvector of $M(t)$. Therefore $g(t)$ is a positive eigenvector of $M(t)$. The corresponding eigenvalue is $d^{-1}g(t)t + t$. By the uniqueness of positive eigenvectors of the matrix $M^*(t)$ (Gantmacher 59) $d^{-1}g(t)t + t$ must be the Perron-Frobenius eigenvalue of $M^*(t)$. In other words $\phi(t) = d^{-1}g(t)t + t$. But by Proposition 4 the Perron-Frobenius eigenvalue of R is equal to t that solves $\phi(t) = t$. We therefore require $g(\beta) = 0$. This proves the proposition.

PROPOSITION 7:

$\text{Lim}_{n \rightarrow \infty} \; u_n \rightarrow [u_1,z] \, \beta^{n-1} f^* + O(\beta^n)$

where z is the right Perron-Frobenius eigenvector of R, β is the Perron-Frobenius eigenvalue of R, and f^* is the left eigenvector of R. f^* is normalized such that $f^*e = 1$. $[u_1,z]$ is the inner product of u_1 and z

Proof: Follows directly from the Perron-Frobenius theorem.

PROPOSITION 8:

$f^* = -\beta(1-\beta)^{-3}\Omega_1(1/\beta)\Omega_2(1/\beta)\Omega_s(\beta)g_1(1/\beta) \otimes g_2(1/\beta) \otimes g_s(\beta)$ [eq 20]

Proof: To establish this claim we need to show that f^* is a positive eigenvector of R and that $f^*e = 1$.

f^* is positive because all terms in eq 20 are positive, except $\Omega_s(\beta)$.

Since R solves $R^2A + RB + C = 0$, a positive eigenvector of R corresponding β is an eigenvector for $M(\beta)$. By the uniqueness of the positive eigenvector for $M(\beta)$ (up to a scalar multiple) we require $g(\beta)$ to be the eigenvector for R. Hence f^* is an eigenvector for R.

At this point all that remains to be shown is that $f^*e = 1$. For this we merely note that $g_i(t)e = (t - 1)/(t\,\Omega_1(t))$. The rest follows from the property of Kronecker products .

Remark: Proposition 6 can be derived without using Proposition 4. Proposition 6 can be derived from Perron-Frobenius theory using arguments based on the uniqueness of positive eigenvectors and the fact that R solves the matrix polynomial equation.

In this sub-section we have derived the tail of the number in queue as seen by a virtual customer. The key to this is deriving the Perron-Frobenius eigenvalue of R. To compute ß we need to invert the Laplace transform of the phase distributions. Since the phase distributions are a subset of the distributions with rational Laplace -Stieltjes transforms, to compute the tail probabilities we have to solve a system of polynomial equations.

At this point we are in a position to derive the tail distributions as seen be an arriving job, and also by each job class.

6.7. Tails Of The Waiting Time Distribution

The knowledge of the tail of the waiting time distribution is useful in answering questions regarding the likelihood of a job spending a long time in the system. It is also helpful in developing approximations for the waiting time distribution. This is particularly valuable because it is difficult to compute exactly the waiting time distribution.

Recall that the waiting time distribution can be analyzed as a Markov chain with a single absorption state 0 (number in system) and infinitely many transient states (n,k). Where n is the number in system and k is the state of the service phase. The generator for this Markov chain has the structure:

$$Q^* = \begin{vmatrix} 0 & 0 & 0 & 0 \\ A'e & B' & 0 & 0 \\ 0 & 0 & A' & B' \\ 0 & 0 & 0 & A' \end{vmatrix}$$

Where $A' = S^o\beta$

$B' = S$

The starting probabilities for this Markov chain are given by Vj(n) and V(n) for the waiting time distribution for job class j and for an arbitrary job , respectively. In the case of a virtual job the starting probability is given by $u_n[\, e_T \otimes I_s \,]$. We will derive the tail of the waiting time distribution for an arbitrary customer.

By [eq 7] we have:

$$V(n) = (L_1 + L_2)^{-1} (L_1 L_2)\, u_n\, [T_1{}^o a_1 \oplus T_2{}^o a_2 \oplus O_s]\, [\, e_T \otimes I_s]$$

Combining this with Proposition 6 gives us :

$$\mathrm{Lim}_{\, n \to \infty}\ V(n) \to$$

$$(L_1 + L_2)^{-1} (L_1 L_2)\ [u_1, z]\, ß^{\,n-1} f^*\, [T_1{}^o a_1 \oplus T_2{}^o a_2 \oplus O_s]\, [\, e_T \otimes I_s] + O(ß^n) \qquad \text{[eq 21]}$$

If we now substitute for f^* (as given by eq 20) in to eq 21 and noting that $g_i(t)\ T_i{}^o = 1/t$ we get :

$$\mathrm{Lim}_{\, n \to \infty}\ V(n) \to\ k^* ß^{n-1}\, g_s(ß)\ + O(ß^n)$$

By a straight forward application of theorem 6.1 of Takahashi(81) we get the following proposition

PROPOSITION 9

$$\mathrm{Lim}_{\, t \to \infty}\ E(t) \to$$

$$-ß\Omega_s(ß)\, (1 - ß)^{-2} (L_1 + L_2)^{-1} (L_1 L_2)\ [u_1, z]\ \mathrm{EXP}[\Omega_s(ß)t] + O(e^{-\Omega s(ß)t}) \qquad \text{[eq 22]}$$

Remark: The methodology used to derive Proposition 9 can be readily extended to derive the tails of the waiting time distribution for each job class. (Only the constants will change. The exponential rate will be the same)

6.8 Departure Process

As stated earlier we approximate the departure process by its stationary interval (a renewal approximation). For this purpose we need to know the probability of a departing customer leaving the system empty and the idle period distribution.

IDLE PERIOD DISTRIBUTION

PROPOSITION 10

The idle period is distributed as the following phase distribution:

$$[\mathring{a}\, , T_1 \oplus T_2\, , (T_1{}^o \oplus T_2{}^o\,)e]$$

where å , the initial probability vector is given by :

$$\text{å} \;=\; u_1[I_T \otimes S^o] \,/\{u_1[I_T \otimes S^o]\, e\} \qquad \text{[eq 23]}$$

Proof: To establish this proposition we merely need to validate å. Recall that the state space of the Markov chain describing the queueing system is lexicographically arranged. Accordingly elements in å are doubly indiced as (i,j). Where i is the state of arrival process 1 and j is the state of the second arrival process. The numerator of eq 23 is the rate in to each micro state corresponding to an empty system. The denominator is a normalizing constant.

PROBABILITY OF A DEPARTING CUSTOMER LEAVING THE SYSTEM EMPTY

PROPOSITION 11:

Let ç be the probability that a departing job leaves the system empty.

$$\text{ç} = (L_1 + L_2)^{-1} (L_1 L_2)\, u_0\, (T_1{}^o \oplus T_2{}^o)\, e \qquad \text{[eq 24]}$$

Proof:

$(L_1 + L_2)^{-1} (L_1 L_2)\, u_0\, (T_1{}^o \oplus T_2{}^o)\, e$ is the probability that an arriving customer finds the system empty. The result follows from the fact that the probability of a departing customer leaving the system empty is the same as that of an arriving customer finding the system empty.

STATIONARY INTERVAL OF THE DEPARTURE PROCESS

PROPOSITION 12:

The stationary interval of the departure process is given by the distribution [ø, G, G^o]. Where

$$\text{ø} = \text{ç}(\text{å}\,,0) + (1-\text{ç})\,(0,\beta\,)$$

$$G \;=\; \begin{vmatrix} (T_1 \oplus T_2) & (T_1{}^o \oplus T_2{}^o)e\beta \\ O & S \end{vmatrix}$$

$$G^o = \begin{vmatrix} O \\ S^o \end{vmatrix}$$

Proof : This phase is the concatenation of the idle period distribution and the service distribution. The rest follows directly from Propositions 10 and 11.

7. SUPERPOSITIONING PHASE RENEWAL PROCESSES

In the previous section we studied a queueing system with an arrival process that was the superpositioning of several renewal processes. The analysis was based on exact representation of the superposed process. Unfortunately, the size of the exact representation grows very rapidly. For instance, if we superpose two phase renewal processes of sizes m and n, the exact representation has mn nodes. Thus the size grows at an exponential rate. Although the size of the representation grows very rapidly, we know that as more and more processes are superposed, the superposed process begins to look like a Poisson process - this process is generated by a phase process with one node! This fact is encouraging if we want to approximate the large phase process by a smaller one. The utility of the result of course depends on how rapidly the superposed process converges to a Poisson process. In this section we provide computational results which provide some insights into this question. In order to develop approximations for a point process such as the superposition of renewal processes, we must first identify criteria for measuring how closely the approximation emulates the actual process. We use the behavior of the second moment as the basis for evaluating the quality of the approximation.

In the parametric decomposition method the superposition of renewal processes is approximated by a renewal process. The first moment of the approximating renewal process is set such that the arrival rate of the superposed process and that of the approximating process are the same. For the second moment, Whitt(82) tried two options - the stationary interval variance and the asymptotic variance. Whitt's experimentation with these two approximation procedures showed that neither procedure dominated the other in predicting performance measures of queues. The actual values (estimated via simulations) fell between the measures predicted by these approximations. This led him to consider approximations obtained by combining these two procedures. The hybrid approximations have been investigated by Albin(80,82). She found significant improvement in predictions with these composite procedures. It was observed that asymptotic methods works better when utilizations are high. In fact the asymptotic method is exact in the limit as the utilization approaches one. On the other hand the stationary variance is exact in the limit as the number of superposed processes approaches infinity. In the hybrid approach the variance of the approximating renewal process lies between the asymptotic value and the stationary interval value. The hybrid variance is a convex combination of the other two variances. The weighting factor takes into account the utilization of the server and the number of processes being superposed.

In this section we compute the stationary interval variance, the asymptotic variance, and the variance of the time until the nth arrival for the superposed process. This study was undertaken to get a feel for how rapidly the superposed process approaches the Poisson process, and to see if it is possible to find smaller phases that behave like the larger process. Our approximation differs from the conventional procedures in that we do not require the approximation to be a renewal process. Our objective is to design a process with a small representation that has the same stationary interval and asymptotic variance as the process it is approximating. Our focus has been on the second moment of the superposed process and we have not determined the effect of higher moments on the performance of the queue. The main reason for this attention on the second moment is the accuracy of the parametric decomposition approach.

We begin this section by showing how to represent superposition of phase renewal processes. Based on this discussion we identify how to represent more general point processes using phases. We call these processes general phase processes. We then derive the stationary interval distribution of superposed process and the moments of the time until the nth arrival. This is followed by a discussion of our computational study. Finally we test our approximation approach on two superposed processes. Preliminary testing of the approximations are very encouraging.

7.1 Representation of Superposition of Phase Renewal Processes

Let us first consider a renewal process with interarrival times distributed as the mixture of an exponential distribution and an Erlang order two distribution. Figure 7.1 depicts the Markov chain for this phase distribution. Node I is the absorption state.

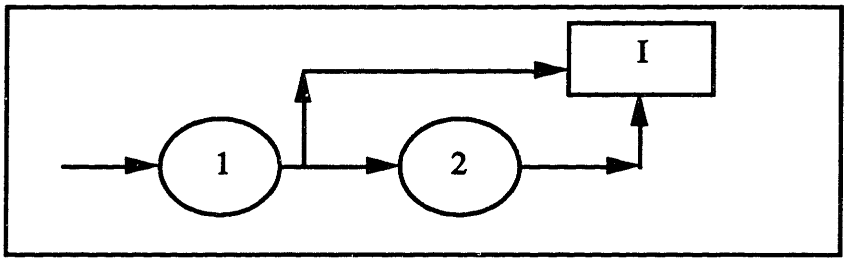


Figure 7.1 Markov chain for a phase distribution

The phase representation for this distribution is : [a, T, T^o] ; a = (1,0)

$$T = \begin{vmatrix} -1 & 0.5 \\ 0 & -1 \end{vmatrix} \quad T^0 = \begin{vmatrix} 0.5 \\ 1 \end{vmatrix}$$

The arrival process generated by this renewal process is a pure birth process. The corresponding Markov chain is shown in figure 7.2. It consists of blocks corresponding to the number of arrivals that have occurred. Each time we cross the vertical line and enter a new block there is an arrival. We can exit a block from either node 1 or node 2. Regardless of which node we exit from, we always start from node 1.

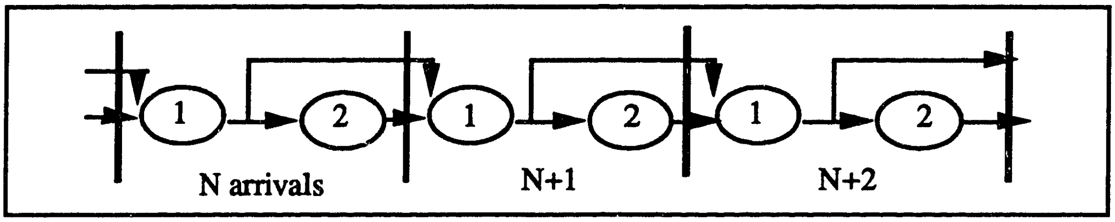


Figure 7.2 Phase renewal process.

Now consider the superposition of two phase renewal processes. Let us assume that for both the processes the interrenewal times are distributed as Erlang order 2 , perhaps with different rates. The states in superposed process are indiced as (i,j), where i is the state of the first process and j the state of the second process. Further these states are lexicographically numbered; i.e (1,1) =1, (1,2) = 2, (2,1) = 3 .... The birth process generated by this superposed process is shown in figure 7.3.

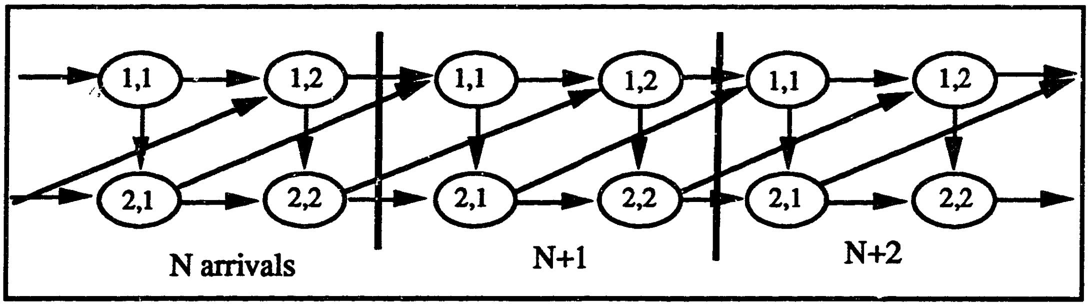


Figure 7.3 Birth process generated by the superposition of two phase renewal processes

The essential difference between figure 7.2 and figure 7.3 is that in figure 7.3 where we start in the next block depends on the exit node in the previous block. For instance if we exit from node (2,1) we reach node (1,1), but if we exit from node (2,2) we go either to (1,2) or (2,1). This illustrates the well known fact that the superposition of renewal processes is in general not a renewal process.

To represent the superposed phase renewal process we need two matrices. The first matrix corresponds to the rates within a block, and the second matrix gives the rates into the next block. Let [a, T, T^o] and [b , S, S^o] be the representations of the two renewal processes being superposed. Then the rates within the each block are given by $B^* = T \oplus S$, and the rates into the next block are given by $B^{*o} = T^o a \oplus S^o b$.

Let us briefly return to the phase renewal process. In figure 7.1 if we allow node I to be an instantaneous state, then the birth process can be described as follows: each time we reach node I an arrival occurs, and we return immediately to node 1. Since the difference between Figure 7.2 and 7.3 is that the starting node in fig 7.3 is not independent of the exit node, by analogy, we need many instantaneous states to describe the birth process generated by the superposition of phase renewal processes. For our example we need 2 instantaneous states. Figure 7.4 shows the representation of the superposition process.

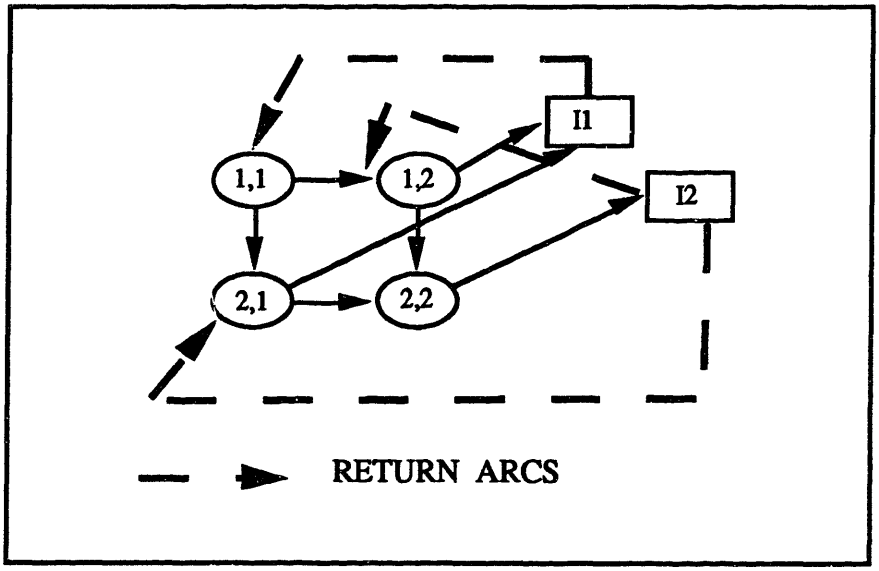


Figure 7.4 Representation of superposition two Erlang order 2s.

We call a phase process which consists of a finite transient Markov chain connected to multiple instantaneous states (as in fig 7.4) a general phase process (GPH). Associated with each instantaneous state is a vector of initial probabilities. The birth process generated by this general phase process is as follows : each time we enter an instantaneous state there is an arrival (or birth), and we immediately return to node j of the transient part with a probability given by the initial probability vector associated with the instantaneous state. This is a natural generalization of phase renewal process. We can represent many processes including alternating renewal processes, and the superposition of phase renewal processes as general phase processes. We use general phase processes to build non-renewal approximations for superposed renewal processes.

As stated earlier, in this thesis our approximations are based on the behavior of the second moment of the superposed process. For this purpose we now derive the stationary interval of the superposed process.

7.2 Stationary Interval of the Superposed Process

Let us assume (A1) that the arrival process has been going on for a sufficiently long time, and (A2) that at time zero there is an arrival. Further, assume (A3) that we neither know when the previous arrival occurred nor the type of arrival that has occurred at time zero (by the type of arrival we mean the renewal process that is associated with the arrival). Then, the distribution of the time until the next arrival is the stationary interval distribution of the superposed process.

Clearly, at time zero we enter a new block (figure 7.3) and the stationary interval is the time to leave this block. To derive the distribution of the stationary interval we need to know the initial states; i.e. we need the state of the arrival process at time = 0^+. This can be deduced through two different approaches. In the first, approach we focus on the discrete time Markov chain that keeps track of the state of the superposed process at arrival instants. This embedded Markov chain is irreducible and aperiodic, if each of the superposed renewal processes is irreducible. Hence forth we assume that the embedded Markov chain is indeed irreducible and aperiodic. In order to derive the stationary interval distribution we need the steady state probabilities of this embedded Markov chain. For this purpose we need to derive the transition matrix of the embedded Markov chain. In the second approach, we derive the state of the arrival process at time = 0^+ by conditioning on the type of arrival that occurs at time zero. The second approach results in a simpler procedure for calculating the initial probabilities. We, nevertheless, describe both the approaches as we use these results in the subsequent discussion.

PROPOSITION 13. Let P be the transition Matrix for the embedded discrete time Markov chain that keeps track of the state of the superposed process at arrival instants. Then $P = -(B^*)^{-1}(B^{*o})$.

Proof: Consider the discrete time Markov chain that keeps track of the state of the superposed process at each transition epoch; i.e. instants at which we move from one node to the other in figure 7.3. The transition matrix for this Markov chain can be divided into two parts. The first deals with the transitions into states within the same block. These transition probabilities are given by $(-D^{-1}B^* + I)$. Where D is a diagonal matrix formed by the diagonal elements of B*. The diagonal elements of B* are strictly negative. The second part deals with the transition probabilities into the states in the next block. These transition probabilities are given by $-D^{-1}B^{*o}$.

Let us assume temporarily that each state in the next block is an absorption state, and that we are in state i (in the current block) with probability ßi. Let **ß** be the vector with elements ßi. Then the probability of eventually reaching state j in the next block is given by

$$\sum_{n=0}^{\infty} -\beta(-D^{-1}B^* + I)^n D^{-1}B^{*o}.$$

Since the block we enter at t = 0 is transient.

$$\sum_{n=0}^{\infty} (-D^{-1}B^* + I)^n = (B^*)^{-1} D \text{ and}$$

$$\sum_{n=0}^{\infty} -\beta(-D^{-1}B^* + I)^n D^{-1}B^{*o}. = -\beta (B^*)^{-1}(B^{*o}).$$

Hence, given that our current state vector is ß, our state vector when we reach the next block is $-\beta(B^*)^{-1}(B^{*o})$. This shows that P, the transition matrix of the Markov chain that keeps track of the state of the arrival process at arrival instants is $-(B^*)^{-1}(B^{*o})$. •

PROPOSITION 14: The stationary distribution of the superposed renewal process is a phase distribution with the following representation : (ç , B*, $B^{*o}e$). The initial probability vector ç is the unique solution to $uP = u$, $ue = 1$, $u \geq 0$.

Proof : We have shown in Proposition 13 that P is the transition matrix for the embedded Markov chain that keeps track of the state of the arrival process at arrival instants. ç is the vector of steady state probabilities for this chain. B^* is the generator for the transitions within each block, and $B^{*o}e$ is the column of rates into the next block. This completes the proof of Proposition 14.

Now, consider the second method for computing the steady state probabilities. For this purpose let us number the renewal processes that are being superposed. We first condition on the type of arrival that occurs at time zero. Without loss of generality let us assume that it is a type 1 arrival; i.e., the arrival at time 0 was due to renewal process 1. This would mean that at time 0^+ renewal process 1 would be in its initial state 'a'. If, in addition we know the state of the second renewal process, then we can determine the state of the superposed process at 0^+. To deduce the state of the second renewal process we define (a) a new embedded discrete time Markov chain and (b) a continuous time Markov process.

(a) The Markov chain : This chain is a discrete time Markov chain that keeps track of the state of renewal process 2 at renewal epoches for process 1. We would like to note that for both renewal processes, the mean interrenewal time is strictly positive.

(b) The Markov process: This Markov process is generated by renewal process 2. In this process, we move through the 'transient' states of the phase distribution (b, S, S^o) and upon reaching the instantaneous state we immediately return to the 'transient' part. The node that we reach upon exiting from the instantaneous state is determined by the initial probability vector b. Thus, in this Markov process we keep circulating between the transient states and the instantaneous state. The infinitesimal generator for this Markov process is given by $\Omega = S + S^o b$.

Let Π be the transition matrix which keeps track of the state of arrival process 2 at renewal instants for process 1. Then the state of arrival process 2 at time 0^+ (under the assumption that we have a type 1 arrival at time 0) is given by a vector V which is the unique solution to $f\Pi = f$, $fe = 1$, $f \geq 0$. In other words V is the steady state probability vector for the Markov chain (a). We now show that V is also the unique solution to $f\Omega = 0$, $f \geq 0$.

PROPOSITION 15: Let V be the unique solution to $f\Pi = f$, $fe = 1$, $f \geq 0$. Then V is also the unique solution to $f\Omega = 0$, $f \geq 0$.

Proof: It is well known (Çinlar 75) that $f\Omega = 0,\ f \geq 0$ if and only if

$$f\left[\sum_{n=0}^{\infty} t^n\Omega^n/n!\right] = f,\ fe = 1, f \geq 0 \text{ for every t.}$$

Let $\phi(t)$ be the density function of the interrenewal times for process 1, then

$$\Pi = \int_0^{\infty}\left[\sum_{n=0}^{\infty} t^n\Omega^n/n!\right]\phi(t)\,dt.$$

Therefore it follows that if V solves $f\Pi = f,\ fe = 1, f \geq 0$ then V is also the solution to $f\Omega = 0,\ f \geq 0$. •

Thus we have shown that at time 0^+ the Markov process generated by renewal process 2 (please see definition (b) above) is also in its steady state.

COROLLARY 2: Given that the arrival at time zero is a type 1 arrival, the state of the arrival process at time 0^+ is given by $a \otimes V$.

The probability that the arrival at time zero is type 1 is proportional to the arrival rate of process 1. If we let πi be the probability that the arrival at time 0 is an i type arrival, then

$$\pi 1 = L2/(L1 + L2), \text{ and } \pi 2 = L1/(L1 + L2).$$

L1 and L2 are the mean interarrival times for processes 1 and 2, respectively. By unconditioning on the type of arrival at time zero we can deduce the following proposition:

PROPOSITION 16: The state of the arrival process at time 0^+ is given by a vector

$$ç = \pi 1\, a \otimes V + \pi 2\, W \otimes b.$$

In this expression W is the steady state probability for the Markov process generated by renewal process 1 (analogous to definition (b) above).

In this subsection we have derived the stationary interval distribution. Each interarrival time in the superposed process is distributed as the stationary interval; i.e, they are identically distributed. This results can be deduced from Proposition 14. If we do not know when the first arrival after 0 occurs, then the state of the arrival process when that arrival occurs is also ç (by Proposition 14). However, the interarrival times are not independent of each other. In other words if we know the length of the first interarrival time after 0, then the distribution of the second interarrival time is no longer the same as the stationary interval distribution. This is how the superposed process differs from a renewal process. A natural measure of the 'non-renewalness' of the superposed process is the correlation between adjacent interarrival intervals. In particular, we study the correlation between the first and second interarrival times after time zero. To compute the correlation we need the second moment of the time until the second arrival after time zero. In the next subsection we show how to compute the second moment of the time until the nth arrival after time zero.

7.3 Second Moment of the n^{th} Arrival Time

In this subsection we continue to make assumptions A1 through A3 stated in subsection 7.2. We now derive the second moment of the n^{th} arrival time. Let τ_n denote the time of the nth arrival. We begin with the second arrival time; i.e.τ_2 . Consider Fig 7.5; the second arrival occurs when we enter the third block .

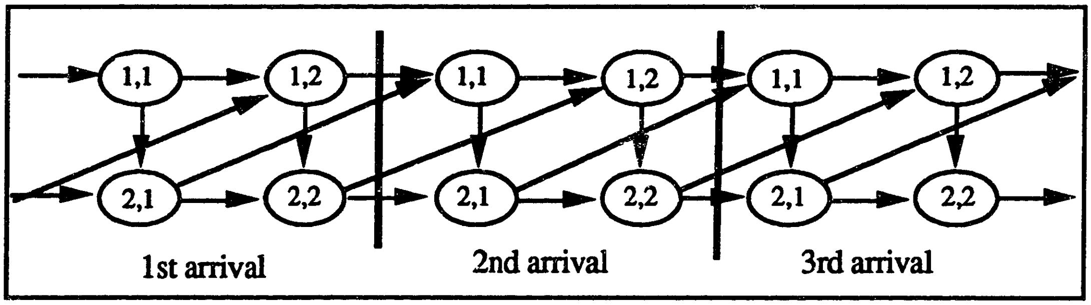


Figure 7.5 Birth process generated by the superposition of two phase renewal processes

Therefore, the time for the second arrival is distributed as a phase distribution with the following representation ((ç,0), B2, $B2^0$).

$$B2 = \begin{vmatrix} B^* & B^{*0} \\ 0 & B^* \end{vmatrix} \qquad B2^0 = \begin{vmatrix} 0 \\ B^{*0}e \end{vmatrix}$$

Since the second moment of a phase distribution with representation (a, T, T^0) is $2a\,T^{-2}e$, the second moment of the random variable τ_2 is

$$2ç\,(B^*)^{-2}e\ -\ 2ç(B^*)^{-2}B^{*0}(B^*)^{-1}e - 2ç(B^*)^{-1}B^{*0}(B^*)^{-2}e\ .$$

In this expression, the first term is the second moment of the stationary interval. By Proposition 13 and 14 $ç(B^*)^{-1}B^{*0} = -ç$. Thus the third term is also equal to the second moment of the stationary interval. We denote the second moment of the stationary interval by μ''. The above expression can be rewritten as : $2\mu'' - 2ç(B^*)^{-2}B^{*0}(B^*)^{-1}e$.

Let
$$k = B^{*0}(B^*)^{-1}$$
$$\emptyset(n) = 2\,ç\ (-1)^n(B^*)^{-2}(k)^n$$
$$S(\tau_n) = \text{Second moment of } \tau_n$$

Then using arguments similar to those employed above we can show:

$$S(\tau_n) = n\,\mu'' + \sum_{j=1}^{n}(n-j)\emptyset(j)\ .$$

7.4 Non-renewal Approximations for the Superposed Process

Consider the superposition of three identical renewal processes. The interarrival times for these process are distributed as Erlang order 3. Figure 7.6 depicts the Markov chain for the superposed process. Since all three processes are identical, the state space of the superposed process is defined slightly differently. Each node has three indices - (x_1, x_2, x_3) - x_i is the number of processes in state i. For instance, (3,0, 0) would mean that all three Erlang processes are in their first stage.

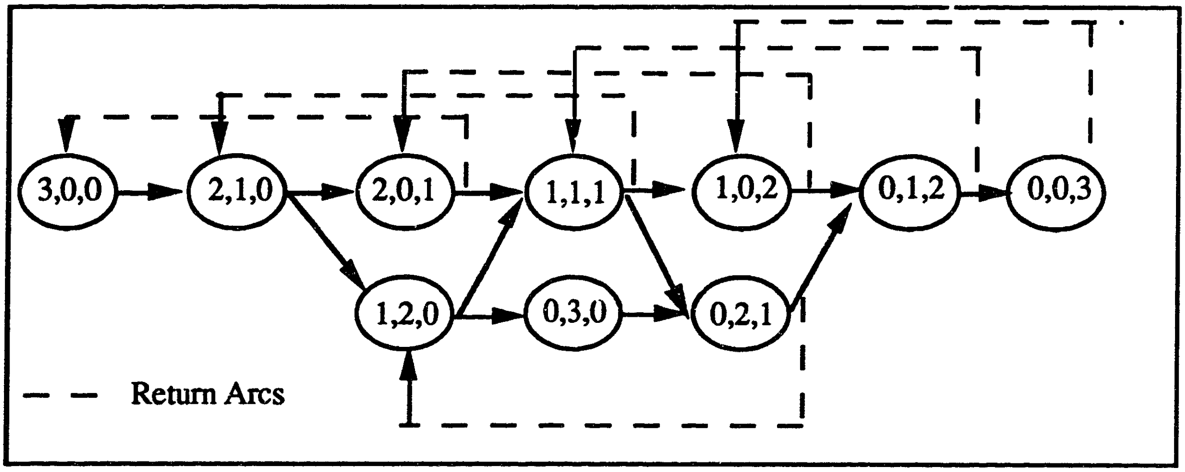
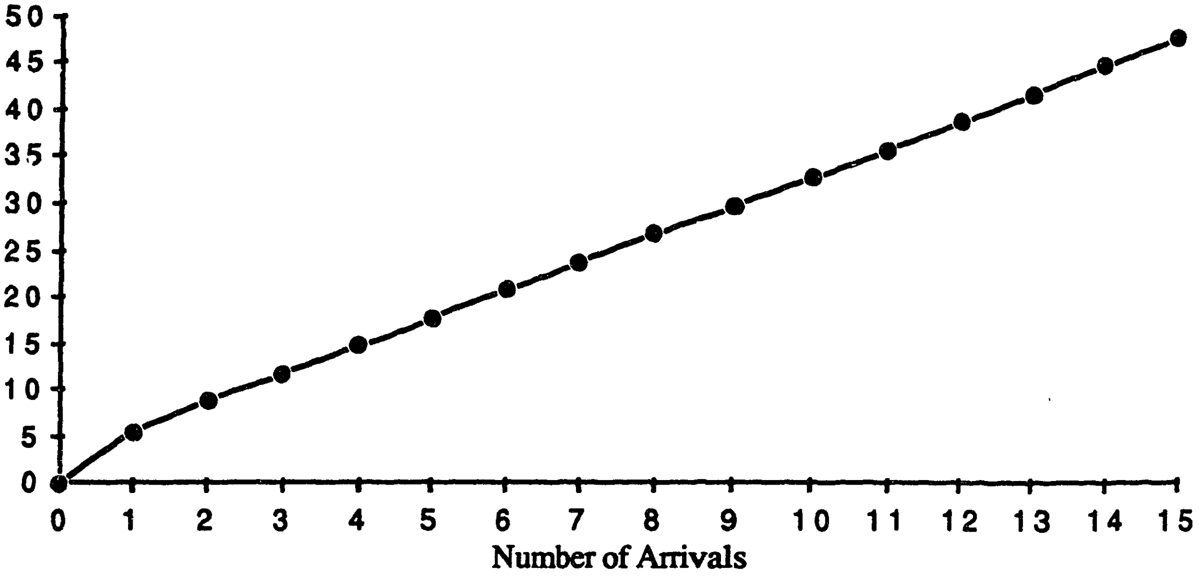


Figure 7.6. Illustration of the Superposition of 3 Erlang Order 3s

For the superposed process we have computed the variance of the time until the nth arrival, starting from time zero. The variances have been plotted in figure 7.7. Note that

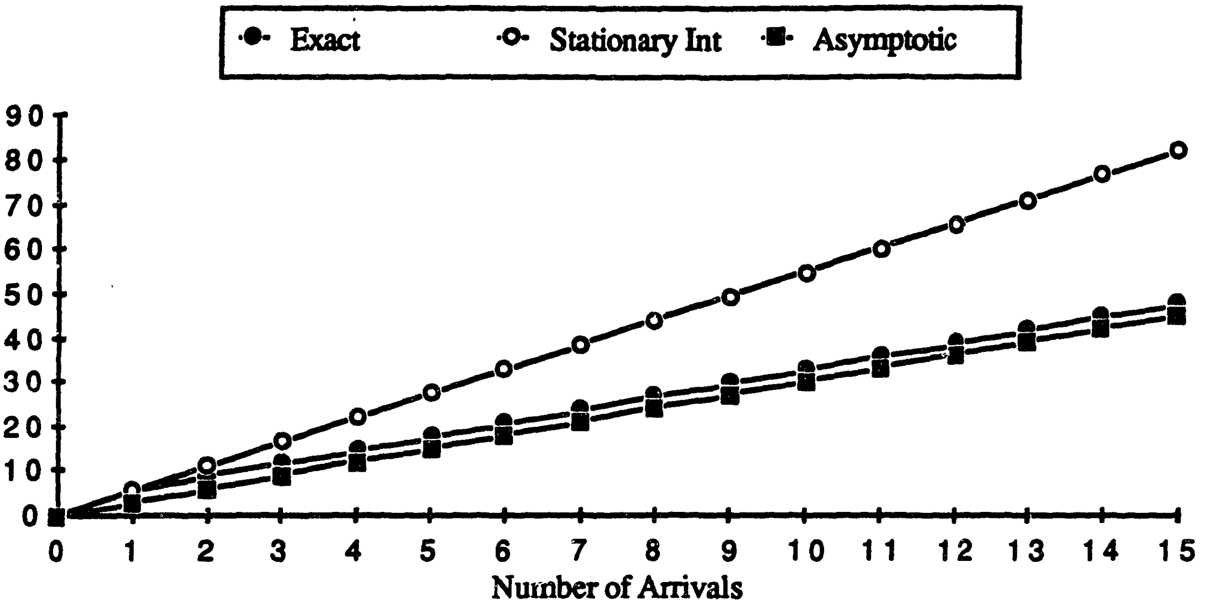
Figure 7.7. Variance of the time of the nth arrival



this figure is non-linear. A similar graph for a renewal process would be linear. The starting slope of the curve (fig 7.7) is the stationary interval variance, and the limiting slope is the asymptotic variance. If we approximate the superposed process by a renewal process with variance equal to the variance of the stationary interval, we would get a straight line going through the origin with slope equal to the initial slope in fig 7.7. On the other hand if the variance of the approximating renewal process equals the asymptotic variance, the corresponding line would have a slope equal to the asymptotic variance of the superposed process. In figure 7.8 we have plotted these approximations. In this figure we see that the stationary interval overestimates the variance and the asymptotic limit underestimates the variance. Therefore, it is not surprising that hybrid approaches (Albin 83), in which the variance of the approximating renewal process is set equal to the convex combination of the asymptotic variance and the stationary interval, work well. In the queueing network analyzer (QNA, Whitt), a different approximation is used for different performance measures. In other words, to compute the average number in queue one approximation for the variance is used, and to estimate the second moment of the number in queue, yet another approximation is used. Since, we are not constrained to use renewal approximations, we would like to identify a general phase of smaller size that tracks the variance of the larger process (shown in fig 7.7).

The key questions are: (a) Is there a general phase process with a smaller representation that can track the variance ?and (b) How do we identify / develop this phase ? At this point we can not provide a definitive answer to either question. However, in the ensuing discussion we provide reasons that lead us to believe that we can find smaller phases. Returning to our example (fig 7.6), observe that the structure of the superposed process is fairly repetitive. Based on this structure we designed a smaller phase (fig 7.9) that captures the essential features. The parameters of this design are p - the probability of going from node 2 to node 3, and q - the probability of exiting from node 3. For different values of p and q we computed the stationary interval and the asymptotic variance. It is interesting to note that neither the mean nor the asymptotic variance changed with p and q. The squared coefficient of variation was always equal to 0.333. However, p and q altered the stationary interval variance and the correlation between adjacent interarrival intervals. Note that in fig 7.6 and 7.9 each exit arc returns to a node that is three 'levels' behind. The asymptotic variance of these processes depends entirely on the number of levels that you are thrown back upon exiting.

Figure 7.8 Renewal Approximations



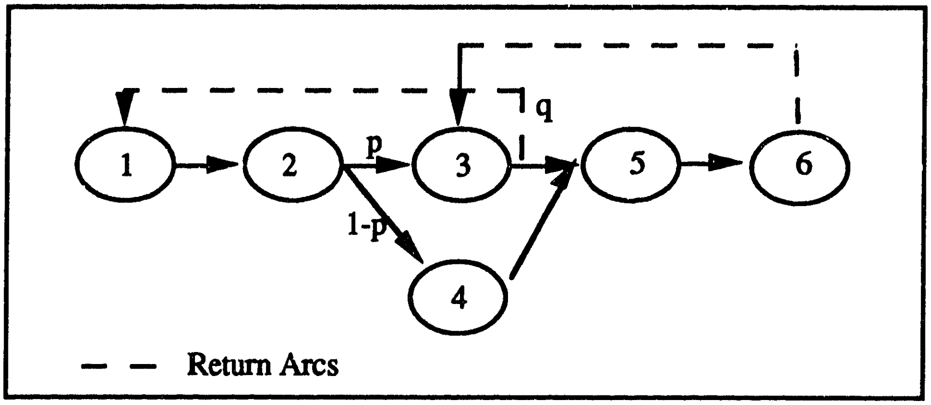


Figure 7.9. Approximation for the superposed process

With p and q equal to 0.6, the stationary interval and the asymptotic variance of the approximating process (Fig 7.9) and the actual process (Fig 7.6) were found to be the same. In fact, the smaller phase closely tracks the variance of the larger process (fig 7.10). We tested the approximation in a queue. Not only did the approximation provide accurate estimates of the first and second moments of the number of jobs in the system (Table 7.1), even the distribution of the number of jobs in the system was effectively the same (Fig 7.11 a, b, c). This accuracy was not effected by the utilization of the server. We have seen in section 6 that the tail of the number in system is geometric. It was interesting to note that the geometric rates for the approximation and the actual queue were the same.

	UTILIZATION					
	0.5		0.75		0.9	
	Mean	Var	Mean	Var	Mean	Var
Exact	0.811	1.0882	2.114	5.3887	6.254	41.8324
Approximation	0.804	1.0601	2.178	5.4115	6.370	41.6396
Error (%)	0.75	2.58	3.02	0.429	1.85	0.4

TABLE 7.1. Evaluation of the Accuracy of the Approximation - Number in System

We used the same design (fig 7.9) with p = 0.3, and q = 0.75 to approximate the superposition of 4 Erlang order 3s. Once again the quality of the approximation, both in terms of tracking the variance (Fig 7.12) and the performance in the queue (Fig 7.13 a, b) was very good.

Figure 7.10 Tracking of Variance for the Superposition of 3 Erlang Order 3s

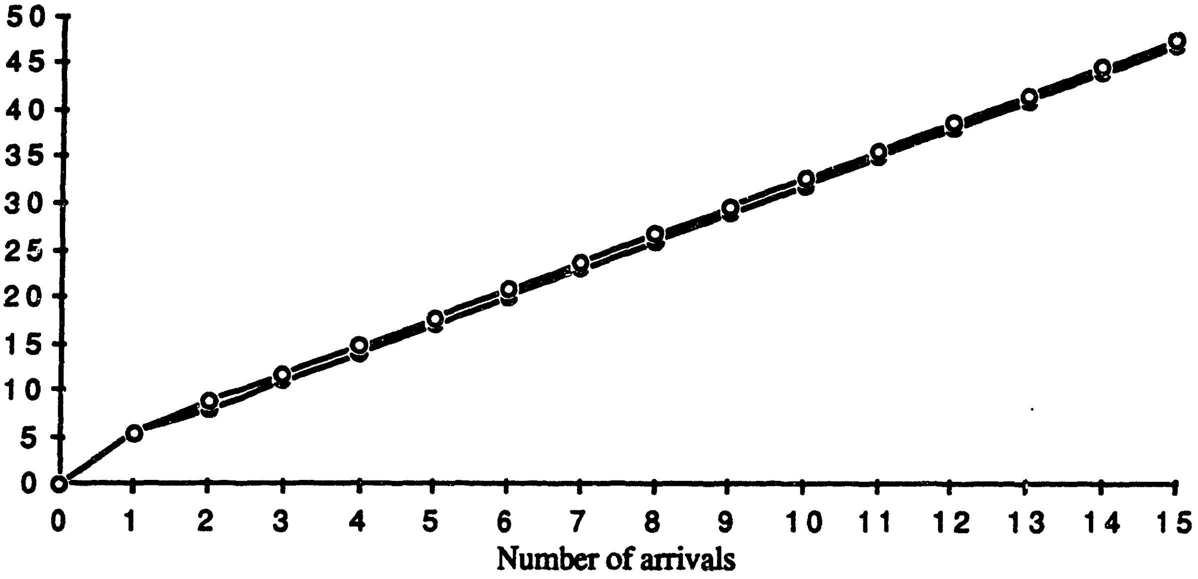


Figure 7.11 a Utilization of Server = 0.5

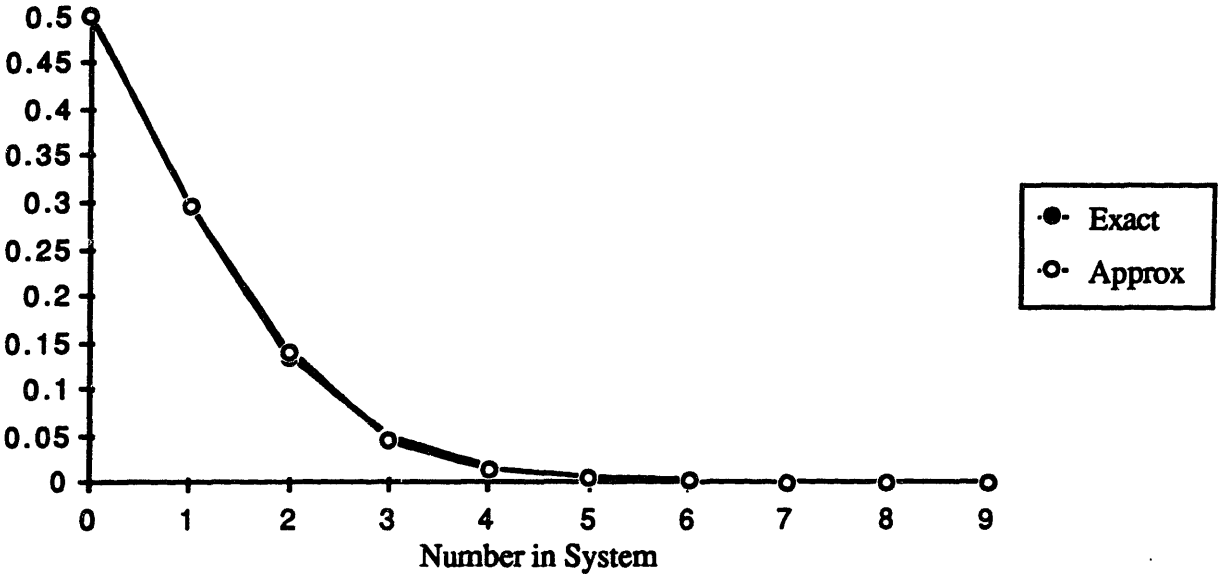


Figure 7.11 b Utilization of Server = 0.75

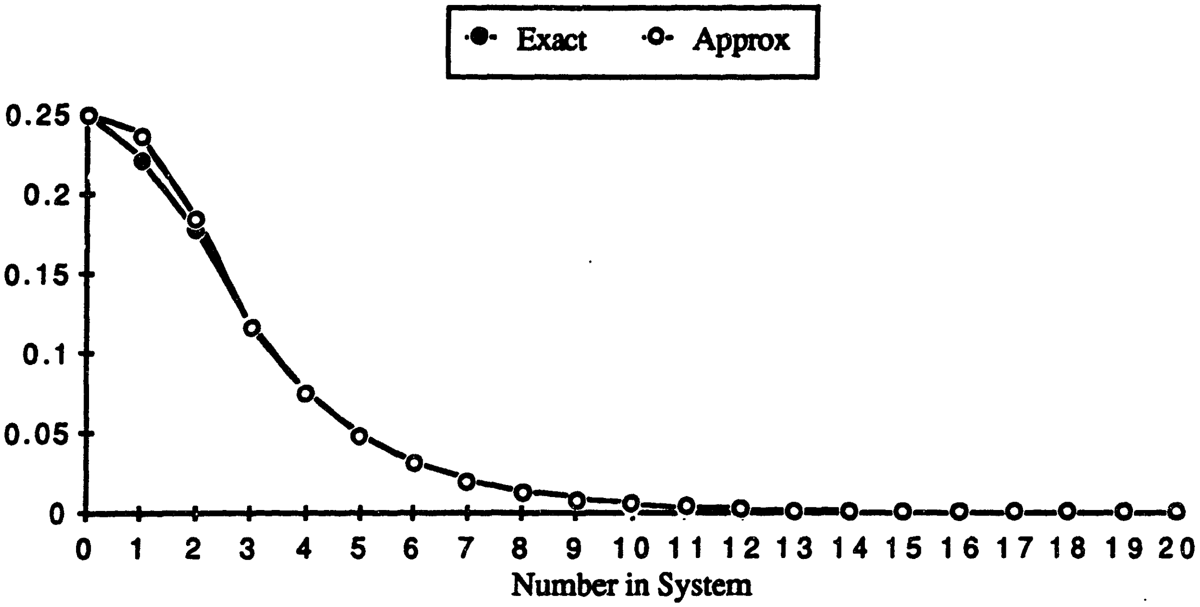


Figure 7.11 c, Utilization of server = 0.9

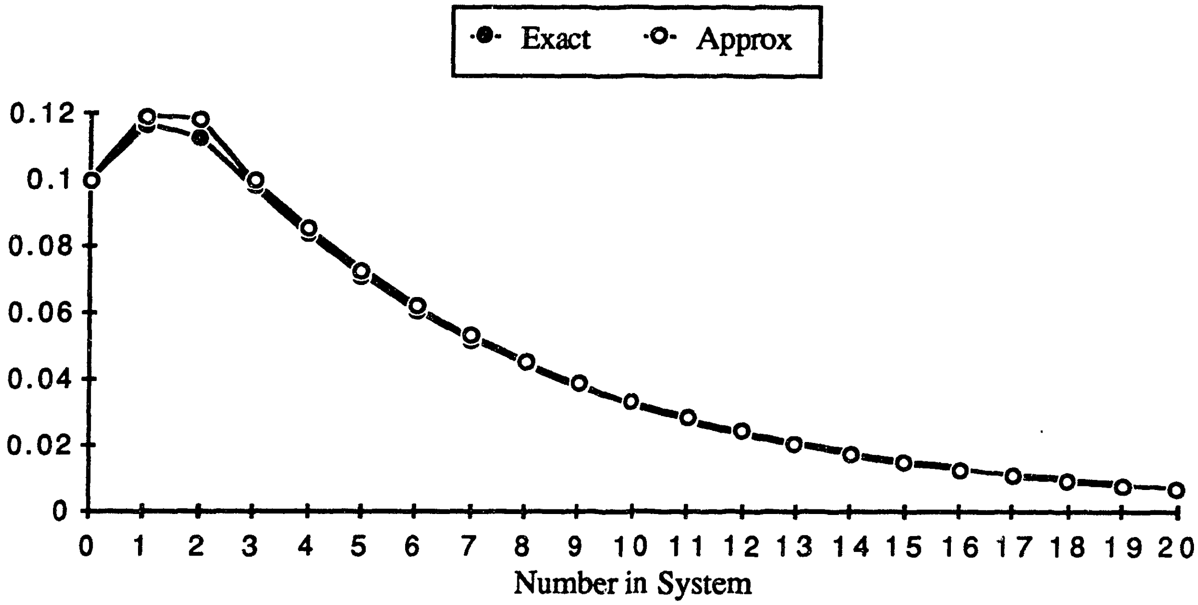


Figure 7.12 Tracking of Variance - Superposition of 4 Erlang order 3s

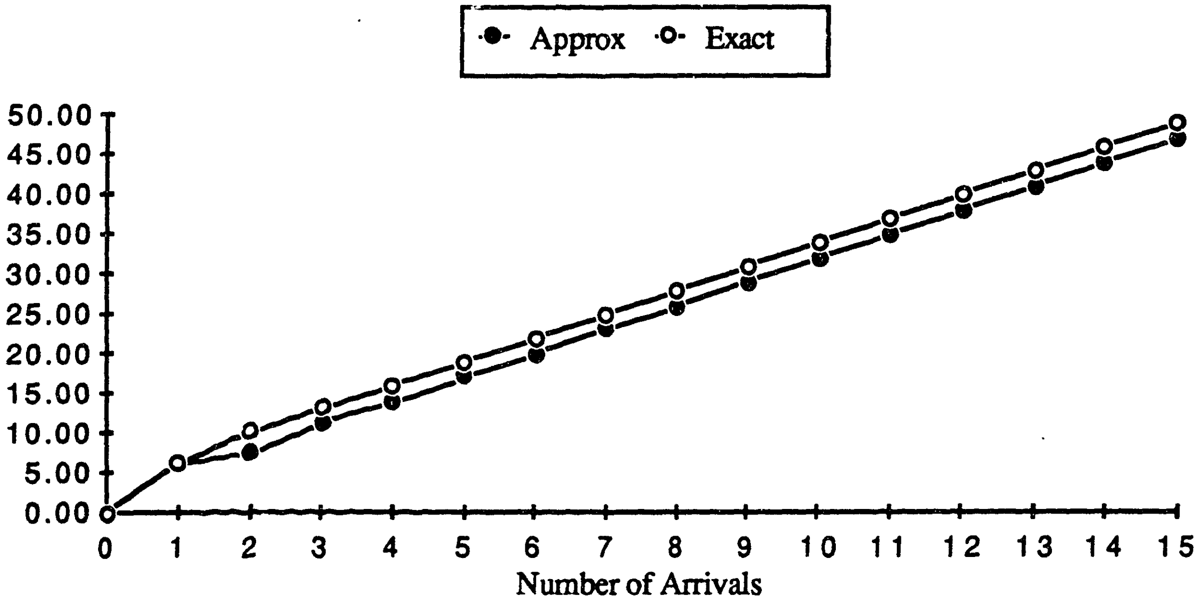


Figure 13. a Utilization of server = 0.75

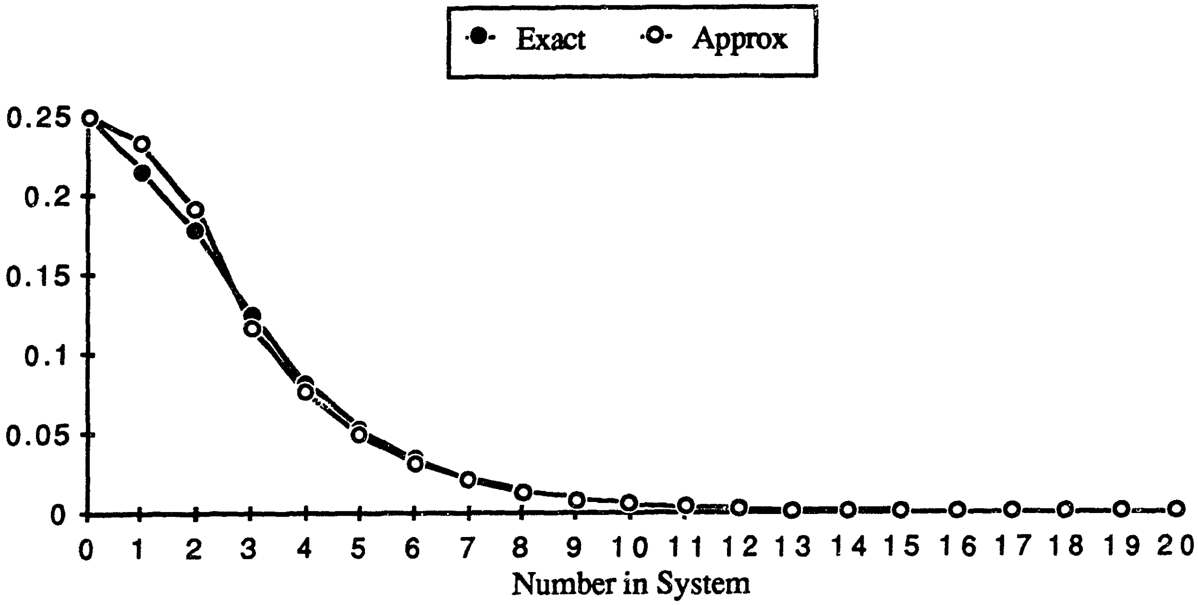
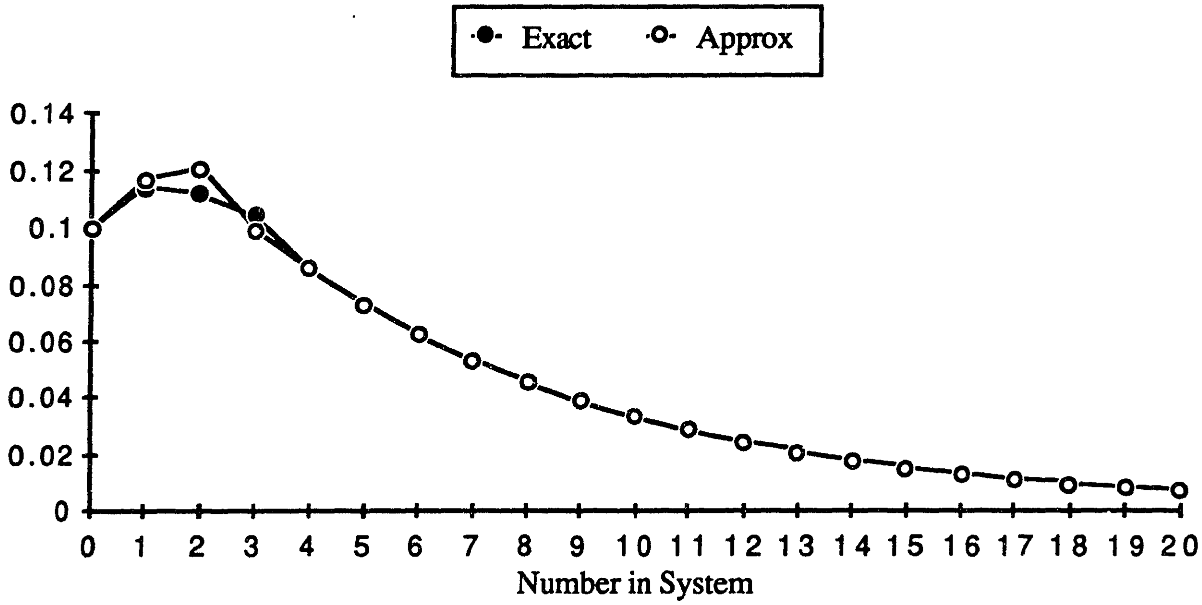


Figure 7.13.b Utilization of server = 0.9



Although the accuracy of the approximations in the two cases that we tested are very encouraging, we do not have a formal procedure for automating the process of identifying good approximations. One obvious strategy is to develop a catalog of designs. Associated with each design would be a set of parameters (analogous to p and q) and a domain of asymptotic and stationary squared coefficient of variations that can be attained by altering the parameters.

Let us return to the question of whether we can find a smaller phase with the same stationary interval and asymptotic variance as the larger process. We know that the asymptotic limit of the squared coefficient of variation of the superposed process is a convex combination of the squared coefficient of variation of the renewal processes being superposed. Hence, to represent the asymptotic limit we do not need a phase larger than the largest renewal process being superposed. We still have to worry about the size of the stationary interval. For this purpose, we computed the squared coefficient of variation of the stationary interval, and the correlation between adjacent intervals for the superposition of some Erlang processes. Table 7.2 gives the scv of the stationary interval for the superposed process and Table 7.3 contains the correlation between adjacent arrival intervals. It is encouraging to note that the stationary interval can also be approximated by Erlang processes of small order (the scv of Erlang order k is 1/k). An interesting question that arises is : can we approximate the superposed process by a renewal process if the correlation between adjacent intervals falls below a threshold. We intend to address this question in our future work.

Erlang Order	Number of Superposed Processes				
	2	3	4	5	6
Order 2	0.625	0.6945	0.7402	0.7731	0.7983
Order 3	0.5093	0.6071	0.6711	0.7167	0.7699
Order 4	0.4551	0.5691	0.6485	0.6949	
Order 5	0.4224	0.5470	0.6285		

TABLE 7.2 SCV of the Stationary Interval of Superposed Erlangs

Erlang Order	Number of Superposed Processes				
	2	3	4	5	6
Order 2	- 0.10	- 0.104	- 0.0979	- 0.0902	- 0.083
Order 3	- 0.2034	- 0.1885	- 0.1647	- 0.1442	- 0.1167
Order 4	- 0.2924	- 0.2427	- 0.1933	- 0.1761	
Order 5	- 0.3662	- 0.2819	- 0.2348		

TABLE 7.3 Correlation between adjacent intervals of Superposed Erlangs

8. DEPARTURES FROM PHASE QUEUES

A by-product of analyzing phase queues using the techniques discussed in section 6, is the idle period distribution and the probability of a randomly chosen customer leaving behind an empty system. These parameters enabled us to determine stationary distributions of the inter-departure time (Proposition 12, Section 6). We propose to approximate the departure by a renewal process distributed as the stationary interval of the departure process. Other approximations can be developed if we can compute the distributions or the moments of the sum of n inter-departure intervals. These distributions are closely related to the remaining busy period . Unfortunately busy period analysis is difficult. However, we can compute the correlation between the first and second interdeparture times.

8.1. Related literature

One of the earliest papers on output processes is that of Burke(56) who showed that the output of a stable M/M/n system is Poisson. A rather complete discussion of Poisson processes in queueing networks is found in Melemad(79). Daley(75) has reviewed results concerning departure process of general queueing systems G/G/n. This review has been updated by Disney and Konig(85).

The description of the output process is usually in terms of the interdeparture intervals, tacitly assumed to be stationary. Most of the results in this area characterize the distribution of inter-departure intervals. Very few results are available which study the covariance of inter-departure intervals, and the variance of the counting process (counts of the number of departure over some interval).

Finch(59) derived a simple expression for the joint distribution of two consecutive inter-departure times for M/G/1 queues. Daley(68) analyzed the covariance between pairs of inter-departure intervals for M/G/1 and G/M/1 queues. He also showed that in a stable M/G/1 system the covariance of adjacent intervals is at most $0.5\, e^{-1}$. This bound is achieved in an M/D/1 system with utilization approaching 1. However, this covariance can be arbitrarily close to -1. Disney and de Morias(1976) explore the auto-covariance of the output from an M/E_k/1/L queue.

In this section we derive the second moment of the time of the second departure for phase type queues. This enables us to compute the correlation between adjacent departure intervals.

8.2. Stationary Interval Distribution of the Departure Process.

Let us assume (B1) that the queue has been operating for a sufficiently long time , and (B2) that at time zero a job departs from the system. Further, assume that we do not know the history of the departure process prior to time 0. Under these assumptions the distribution of the time till the first departure after time zero is the stationary interval distribution of the departure processes. In section 6 (Proposition 12) we have derived the distribution of the stationary departure interval. The interdeparture intervals, in general are neither identically distributed nor are they independent of each other. However, Whitt (83) has found the stationary interval to be a good descriptor of the departure process. We have computed the correlation between adjacent departure intervals for a few queues. The correlations were very small, suggesting that the stationary interval is a good approximation for the departure process.

8.3 Second Moment of the Second Departure.

In this section we assume that the steady state probabilities of the number in queue as seen by an arriving job, and a virtual job have already been computed.

Let τ_2 : Time of second departure after time 0 (random variable)

z : Number of jobs in system at 0+ (random variable)

$S(\tau_2)$: Second moment of τ_2

π_n : Probability that an arriving job finds n jobs in the system

μ_s' : Average service time

μ_s'' : Second moment of the service time

(T, T^0) : Representation of the arrival process (need not be a renewal process)

(β, S, S^0) : Representation of the service process

c0 : State of the arrival process at time 0+, given that there are zero jobs in the system.

c1 : State of the arrival process at time 0+, given that there is one job in the system .

We consider three cases for deriving the second moment of the time to the second departure after time 0.

Case 1: In the first case, there are 2 or more jobs in the system at time 0+. In which case the random variable τ_2 is the sum of two service times. Therefore,

$$S(\tau_2 \mid z \geq 2) = 2\mu_s'' + 2(\mu_s')^2 .$$

Case 2: In this case at time 0+ there is only 1 job in the system. Recall from section 6 that the states of the Markov process depicting the queueing system are indiced by the states of the arrival process and the state of the service process. They are numbered lexicographically. Under this notation, the state of the queueing process at time 0+ is given by

$$c1 = U2[I \otimes S^0 \beta] / U2[I \otimes S^0 \beta]e .$$

U2 is the steady state probability vector for the block corresponding to 2 jobs in the system in the Markov chain depicting the queueing system (Section 6.1).

For ease of exposition we adopt the following convention. Assume that at time zero job #0 departs and we number the jobs in the sequence in which they are processed. Thus, in case 2 processing of job #1 begins at time 0+, and job #2 has not yet arrived at the machine center. Now there are two possibilities:

(i) Job #2 arrives before job #1 is completed ;

(ii) Job #1 is completed before job #2 arrives.

The Markov chain describing the time until the departure of job #2 can be divided into four blocks (Figure 8.1).

(a) Initially we have the superposition of the service process of job #1 and the arrival process of job #2;

(b) If the service of job #1 is completed prior to the arrival of job #2, then we have to account for the remaining portion of the arrival process for job #2;

(c) The third block corresponds to the case when the arrival of job #2 occurs prior to the service completion of job #1. This block consists of the remainder of the service of job #1; and

(d) Finally, we have the service of job #2.

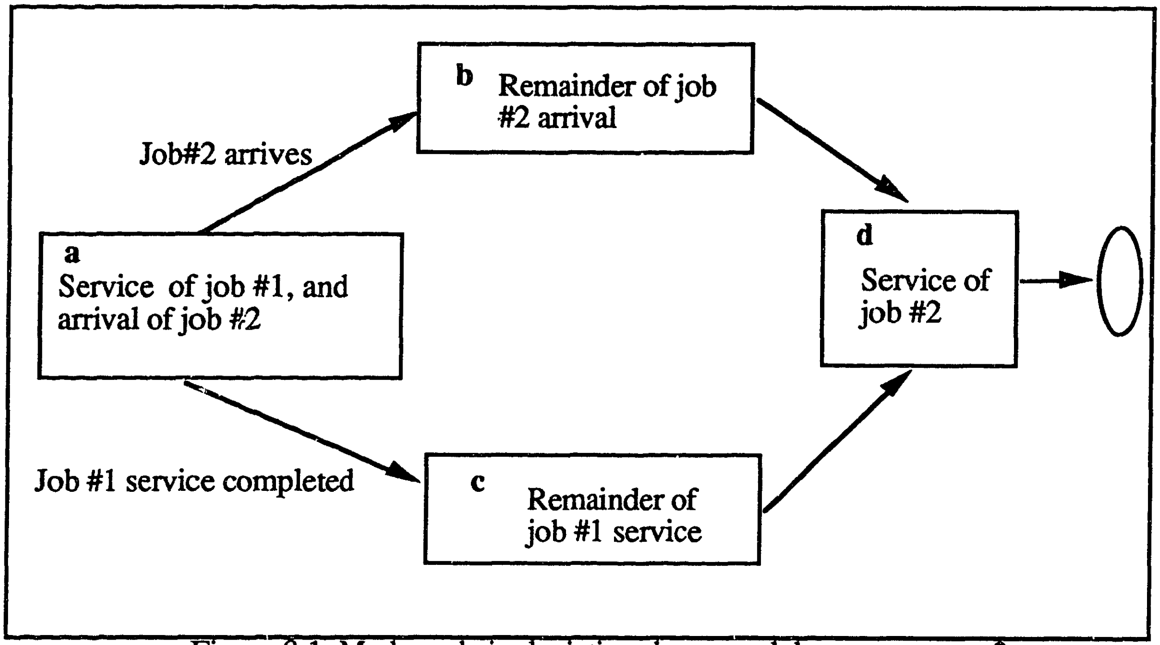

b Remainder of job #2 arrival
Job#2 arrives
a Service of job #1, and arrival of job #2
d Service of job #2
Job #1 service completed
c Remainder of job #1 service

Figure 8.1 Markov chain depicting the second departure - case 2

The distribution of the time of departure of job #2 is distributed as a phase with the representation (g', G', G'^{0}). g' = (c1,0,0,0)

$$G' = \begin{vmatrix} B & B_S^0 & B_T^0 & 0 \\ 0 & T & T^0\beta & 0 \\ 0 & 0 & S & S^0\beta \\ 0 & 0 & 0 & S \end{vmatrix} \qquad G'^0 = \begin{vmatrix} 0 \\ 0 \\ 0 \\ S^0\beta \end{vmatrix}$$

In these matrices, $B = T \oplus S$, $B_S{}^0 = I \otimes S^0$, $B_T{}^0 = T^0 e \otimes I$. Recall that the second moment of a phase distribution (g', G', G'^{0}) is $2\, g\, (G')^{-2} e$. Therefore,

$$S(\tau_2 \mid z = 1) = \mu_s'' + \mu_b'' - 2\,c1\,B^{-1}\,B_T{}^0\,S^{-2}\,e - 2\,c1\,B^{-1}\,B_S{}^0\,T^{-2}\,e$$

$$+ 2\,\mu_s'\,\mu_b' + 2\,\mu_s'\,B^{-1}\,B_T{}^0\,S^{-1}\,e + 2\,\mu_s'\,B^{-1}\,B_S{}^0\,T^{-1}\,e$$

$$- 2\,c1\,B^{-2}\,B_T{}^0\,S^{-1}\,e - 2\,c1\,B^{-2}\,B_S{}^0\,T^{-1}\,e.$$

In this expression $\mu_b' = -\,c1\,B^{-1}\,e$; $\mu_b'' = 2\,c1\,B^{-2}\,e$.

Case 3: In this case job #0 leaves behind an empty system, and so we have to wait for job #1 to arrive. The Markov chain depicting the departure time of job #2 is obtained by concatenating the remainder of the arrival time of job #1 to the Markov chain of the previous case. Figure 8.2 shows the Markov chain for the departure time of job #2.

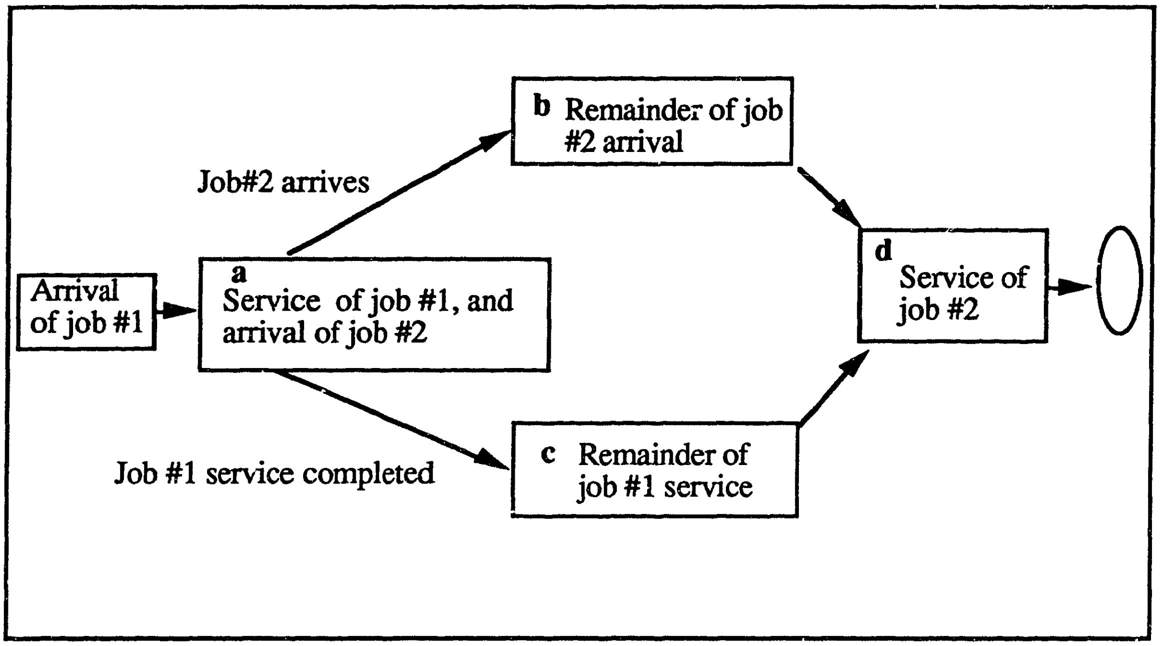

b Remainder of job #2 arrival
Job#2 arrives
d Service of job #2
Arrival of job #1
a Service of job #1, and arrival of job #2
c Remainder of job #1 service
Job #1 service completed

Figure 8.2 Markov chain depicting the second departure - case 3

The departure time for job #2 is a phase distribution with representation (g'', G'', G''^0). $g'' = (c0, 0)$

$$G'' = \begin{vmatrix} T & G^{00} \\ 0 & G' \end{vmatrix} \qquad G''^0 = \begin{vmatrix} 0 \\ G'^0 \end{vmatrix}$$

In this expression $G^{00} = (T^0 \otimes \beta, 0)$. c0 is the state of the arrival process at time 0^+.

$$c0 = U1(I \otimes S^0) / U1(I \otimes S^0)e$$

The second moment is obtained by computing $2g''(G'')^{-2}e$. For simplifying the presentation we define the following intermediary variables:

$$\begin{array}{llcl}
\text{Let} & Y_{b2} & = & -c0\,T^{-1}e \\
& Y_{i2} & = & -Y_{b2}\,B^{-1}B_S^0 \\
& Y_{s2} & = & -Y_{b2}\,B^{-1}B_T^0 \\
& \mu'_{b2} & = & -Y_{b2}\,B^{-1}e \\
& \mu''_{b2} & = & 2\,Y_{b2}\,B^{-2}e \\
& \mu'_{i2} & = & -Y_{i2}\,T^{-1}e \\
& \mu''_{i2} & = & 2Y_{i2}\,T^{-2}e
\end{array}$$

$$\begin{aligned} \mu'_{s2} &= -Y_{s2}\, S^{-1}e \\ \mu''_{s2} &= 2Y_{s2}\, B^{-2}e \\ \mu'_{i} &= -c0\, T^{-1}e \\ \mu''_{i} &= 2\, c0\, T^{-2}e \end{aligned}$$

Then

$$\begin{aligned} S(\tau_2 \mid z=0) = {} & \mu''_i + \mu''_{b2} + \mu''_{s2} + \mu''_{i2} + \mu''_s \\ & + 2\,\mu'_s\mu'_i + 2\,\mu'_s\mu'_{b2} + 2\,\mu'_s\mu'_{s2} + 2\,\mu'_s\mu'_{i2} \\ & - 2\,c0\,T^{-2}(T^0 \otimes ß)B^{-1}\, e + 2\,c0\,T^{-2}(T^0 \otimes ß)B^{-1}B_T{}^0\, S^{-1}\, e \\ & + 2\,c0\,T^{-2}(T^0 \otimes ß)B^{-1}B_S{}^0\, T^{-1}\, e - 2\,Y_{b2}\, B^{-2}B_T{}^0\, S^{-1}\, e - 2\,Y_{b2}\, B^{-2}B_S{}^0\, T^{-1}\, e \end{aligned}$$

In section 6 we derived V(n), the probability that an arriving job finds n other jobs in the system. Since the probability of a departing customers leaving n jobs in the system is the same as the probability of an arriving job finding n jobs in the system, the second moment of the random variable τ_2 is

$$S(\tau_2) = V(0)S(\tau_2 \mid z=0) + V(1)\, S(\tau_2 \mid z=1) + (1 - V(1) - V(2))\, S(\tau_2 \mid z \geq 2)$$

This completes the derivation of the second moment of the time for job #2 to leave the system. Since the mean interdeparture time equals the mean interarrival time, the variance of τ_2 and of the stationary interval can readily be computed. This also allows us to calculate the correlation between the first and second departure intervals after time 0. For a few $E_k/E_k/1$ queues we have computed this correlation. The results are given in Table 8.1 below.

In every case the correlations were extremely small. This suggests that the departure process closely resembles a renewal process. Therefore, the stationary interval seems to be a good approximation for the departure process.

For all the queues tested the correlation was maximum when the utilization is 0.5. This is not entirely surprising. When utilizations are high the server is almost always busy and as a result most interdeparture intervals are distributed as the service distribution. Since service times are independent identically distributed, adjacent interdeparture intervals also appear to be independent. On the other hand when the utilization is small, the interdeparture intervals begin to look like the interarrival intervals. In the queues studied in Table 8.1 the interarrival times are also independent, and identically distributed.

QUEUEING SYSTEM	UTILIZATIONS			
	0.25	0.5	0.75	0.9
$E_3/E_2/1$	0.0527	0.0808	0.0544	0.0239
$E_3/E_3/1$	0.0363	0.0575	0.0395	0.0178
$E_3/E_4/1$	0.0268	0.0398	0.0237	0.0090
$E_2/E_3/1$	0.0125	0.0219	0.0022	0.0015

TABLE 8.1 CORRELATION BETWEEN ADJACENT DEPARTURE INTERVALS

9. SUMMARY AND CONCLUSIONS

In this paper we proposed a new decomposition method for solving networks of queues. We require the distribution of service times and interarrival times to be distributed as phase distributions. This is not a severe restriction as the phase distributions are a dense subset of the family of distributions for nonnegative random variables.

We showed how to carry out the three basic steps of the decomposition method - superposing, queueing analysis, and departure process. We derived the stationary interval of the superposed arrival stream, and the moments of the time until the nth arrival at the queue. We showed how to compute performance measures, including distributions of the number in queue, when the arrival process is not a renewal process. We also derived the stationary interval of the departure process, and the correlation between adjacent departure intervals.

We propose a scheme for approximating large phase processes by smaller process. Preliminary testing of our approach was very encouraging. Unfortunately, we do not have a formal algorithm for approximating the superposed process. We intend to address this aspect of the model in our future work.

REFERENCES

Albin, S.L., "Approximating queues with superposition arrival processes," PhD dissertation, Dept., of IE and OR, Columbia Univ., 1981

Albin, S. L., "Poisson approximations for superposition arrival processes in queues," Mgt. Sc., 28, 2, 1982, 126-137

Albin, S. L., "Approximating a point process by a renewal process, II: Superposition arrival processes to queues," Oper. Res., 32, 1984, 1133 - 1162

Asmussen, S., Applied Probability and Queues. John Wiley, New York, 1987

Bellman, R., Introduction to Matrix Analysis, McGraw-Hill, New York , 1960

Bitran, G. R. and Tirupati, D., "Multiproduct queueing networks with deterministic routings. Decomposition approach and the notion of interference," WP # 1764 - 86. Sloan School of Management, M. I. T., 1986 (To appear in Management Science)

Burke, P. J., "The output of queueing systems," Oper. Res., 4, 1956, 699 - 704

Chandy, K. M. and C. H. Sauer, "Approximate methods for analyzing queueing network models of computer systems," ACM Computing Surveys, vol 10, no 3, 1978, 281-317

Chung, K.L., Markov Chains with Stationary Transition Probabilities, Academic Press, New York, 1967

Çinlar, E., Introduction to Stochastic Processes, Prentice-Hall, Englewood Cliffs, 1975

Daley, D. J., "The correlation structure of the output process of some single server queues," Ann. Math. Stats., 39, 1968, 1007 - 1019

Daley, D. J., "Queueing output processes," Adv. Appl. Prob., vol 8, 1975, 395 - 415

Disney, R. L., and D. Konig "Queueing networks: A survey of their random processes," SIAM Review, 27, 3, 1985, 335-403

Disney, R. L., and DeMorais, P. R. "Covariance properties for the departure process of M/ E_k/ 1/ N queues," AIIE Trans., 8, 1976, 169 - 175

Finch, P. D., "The output process of the queueing system M/ G/ 1," Royal Stat. Soc. Ser. B 21, 1959, 375-380

Gantmacher, F. R., The Theory of Matrices, Vol II, Chelsea, New York, 1959

Jackson, J. R., "Job shop like queueing systems," Mgt. Sc., 10, 1963, 131-142

Kelly, F. P., "Network of queues with customers of different types," Journal of App. Prob., 12, 1975, 542-554

Kraemer, W. and Langenbach-Belz, M. "Approximate formulae for the delay in the queueing system GI/ G/ 1," Congressbook, 8th ITC, Melbourne, 1976, 235.1 - 235.8.

Kuehn, P.J., "Analysis of complex queueing networks by decomposition," 8th ITC, Melbourne, 1976, 236.1 - 236.8

Kuehn, P. J., "Approximate analysis of general queueing networks by decomposition," IEEE Trans. Comm., COM-27, 1, 1979, 113-126

Lemoine, A. J., "Network of queues: A survey of equilibrium analysis," Mgt. Sc., 24, 1977, 464-481

Lucantoni, D. M., and Ramaswami, V. "Efficient algorithms for solving the non-linear matrix equations arising in phase type queues," Stochastic Models, 1, 1985, 29 - 51

Melamed, B., "Characterizations of Poisson traffic streams in Jackson queueing networks," Adv. App. Prob., 11, 1979, 422-438

Neuts, M. F., "Matrix-Geometric Solutions in Stochastic Models, Johns Hopkins Univ. Press, Baltimore, 1981

Neuts, M. F., "A new informative embedded Markov renewal process for the PH/ G/ 1 queue," Adv. Appl. Prob., 18, 1986, 535 - 557

Newell, G. F., Applications of Queueing Theory, Chapman and Hall, London, 1971, chapter 6

Seelan, L. P., "An algorithm for Ph/Ph/c queues," Euro. J. Oper. Res., 23, 1986, 118 - 127

Shantikumar, J. G., and Buzacott, J. A. "Open queueing network models of dynamic job shops," Int. Jour. of Prod. Res., 19, 1981, 255-266

Takahashi, Y., "Asymptotic exponentiality of the tail of the waiting time distribution in a Ph/Ph/c queue," Adv. Appl. Prob., 13, 1981, 619 - 630

Takahashi, Y., and Takami, Y. " A numerical method for the steady state probabilities of GI/G/c queueing system in a general class," J. Oper. Res. Soc. Japan, 19, 1976, 147 - 157

Whitt, W., "Approximating a point process by a renewal process : Two basic methods," Oper. Res., 30, 1982, 125-147

Whitt, W., "The queueing network analyzer," Bell Systems Technical Journal, 62, 1983 2779 - 2815

Whitt, W., "Performance of the queuing network analyzer," Bell Systems Technical Journal, 62, 1983a, 2817 - 2843

Whitt, W., "Approximations for departure processes and queues in series," NRLQ, 31, 1984, 499-521.