Capacity Allocation and Safety Stocks in Manufacturing Systems

by

Christopher Athaide

Bachelor of Technology in Mechanical Engineering
Indian Institute of Technology, Bombay
May 1985

Master of Science in Operations Research and Statistics
Rensselaer Polytechnic Institute
May 1987

Submitted to the Sloan School of Management and the Operations Research Center
in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy in Operations Research

at the

Massachusetts Institute of Technology

February 1992

Signature of Author_____
Operations Research Center

Certified by_____
Stephen C. Graves
Leaders for Manufacturing Professor of Management Science
Deputy Dean, MIT Sloan School of Management
Thesis Supervisor

Accepted by_____
Richard C. Larson
Professor of Electrical Engineering and Computer Science
Codirector, Operations Research Center

Contents

List of Figures

List of Tables

Capacity Allocation and Safety Stocks in Manufacturing Systems

by

Christopher Athaide

Submitted to the Operations Research Center
on
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

In this thesis, we look at three different problems related to capacity allocation, work-in-process (WIP) inventories, safety stocks and lead times in manufacturing systems.

In Chapter Two, we examine the following question: Consider a manufacturing system with n stages. If there were a fixed number K of capacity units available, how should this capacity be allocated among the different stages so that the total inventory at all the stages is minimized? The exogeneous demand process is characterized by an n-dimensional diffusion process. In this model, the production at each stage in the system is governed by a linear rule. We show that the WIP inventory vector for the n stages can be represented by an n-dimensional stochastic differential equation. We formulate the optimization problem and show that the optimal capacity allocation at each stage is similar to the square root and other formulae obtained for queueing networks. We are also able to obtain expressions for the corresponding lead time at each stage. We then provide tradeoff curves for the WIP inventory, safety stock, base stock and lead time as a function of the total capacity K.

In Chapter Three, we consider a discrete time model of a single production stage with a fixed processing capacity which processes two nonsubstitutable items. After an item is produced at the machine, it enters the final item inventory which is used to satisfy demand for that item. The end item inventory for each item should be set to meet a certain service level. The question that we ask is the following: Should these items be produced on a single machine, or on two separate machines? If we were to consider this as a queueing problem with two servers, and the objective was to minimize the average waiting time, the answer would be to merge the two streams. But if one were to use a different criterion, say, to minimize the total WIP and safety stock, it is not clear which alternative is preferable. If the demand is relatively stable for one item and more variable for the other, it might be preferable to produce the two items on separate machines, because by preserving the stability in production for the

first item, the overall reduction in WIP plus safety stock for the two items combined may be greater. We are able to specify the conditions on the mean and variance of the demand for each item under which either a single machine or two machines are appropriate. We extend this result to the case where there are setup times at each machine. We then look at the case where more than two items are processed at the stage and study how the different items are grouped in an optimal configuration.

In Chapters Two and Three, the control rule that we use to represent the production function was linear. In Chapter Four, we characterize the production function using a capacitated rule and propose a model to study work-in-process (WIP) inventory levels and safety stocks for a single stage. The cumulative demand at the stage is modeled as a diffusion process with drift μ and dispersion σ. The inventory at the stage is of two types: work-in-process inventory and end-item inventory. The end-item inventory, also called safety stock, is used to meet the demand at the stage. The stage processes the WIP at a rate dictated by a certain production rule. This parameterized production rule is fairly general and it subsumes a broad class of rules. When the WIP increases, the production rate increases until it reaches a maximum rate C. The questions that we consider are the following: What should be the level of safety stocks so that a certain level of service is provided? What is the distribution of the WIP inventory in steady state? We represent the WIP inventory process in terms of a stochastic differential equation. We show that under certain conditions on the parameters and the initial state of the process, the WIP inventory process has an ergodic distribution. We then compare the results that we get with the results that have been obtained previously using other models.

Thesis Supervisor: Stephen C. Graves

Title: Leaders for Manufacturing Professor of Management Science
Deputy Dean, Sloan School of Management

Acknowledgements

This work was funded by the National Science Foundation's Strategic Manufacturing Initiative Project "Decision Making in Manufacturing Systems" (NSF Grant DMM-8914181).

I would like to express my deep sense of gratitude to my advisor, Prof. Steve Graves. I thank him for the sage counsel that he provided me throughout this thesis; for his encouragement and his patience with me, especially during the initial period when my output was virtually nought; for allowing me the freedom to pursue a wide range of research topics and for the *Celtics* tickets that he gave me from time to time.

I wish to thank the other two members of my thesis committee, Prof. Gabriel Bitran and Prof. Larry Wein. I thank Gabriel for having taken the time off his impossibly busy schedule to be on my committee and making several insightful and useful comments on earlier drafts on the manuscript. I thank Larry for his comprehensive and extremely valuable feedback on this document, and also for stimulating my interest in this field when he offered his course on Queueing Networks a couple of years ago.

I would like to mention a special word of thanks to Prof. Dick Larson for providing me an education in the *practical* domain of operations research while working on the *Queue Inference Engine* and for giving me the opportunity to work for two summers on different projects. I also thank Prof. Tom Magnanti, Prof. Amedeo Odoni and Prof. Jim Orlin for their genuine concern about my welfare throughout my stay at MIT.

I thank Marcia Chapman and Paulette Mosley, the administrators of the OR center during different periods of my stay, for their work behind the scenes, ensuring every semester that the paperwork went through and for being the buffer between me and the bureaucratic process at MIT. I also thank Michele Brodeur and Laura Terrell for their assistance on more than one occasion.

What really makes the OR Center such a great place are the students. The students at the OR Center are, without exaggeration, the finest that I have met in my life. I thank all the ORC students who have made my stay here extremely pleasant and a fine learning experience.

My roommates, Mehul, Tina, Mikhail and Nicole were extremely tolerant towards a roommate who could, at best, be described as being weird, and I thank them for their friendship and support. My companionship with Mikhail kept me abreast of the latest technological developments (you know, Nintendo, Teenage Mutant Ninja Turtles, Mighty Mouse and the like).

And what can I say to my family -- my mom, my dad, my sister Marilu and my brother Claude, for the countless sacrifices that they have made that enabled me to reach this point? Thank you for everything.

To Mom, Dad, Marilu and Claude

Chapter 1

Introduction

In many manufacturing and service organizations, inventories comprise a significant fraction of a firm's assets. Firms carry inventories for several reasons. One of them is to safeguard against uncertainties that arise during the course of the normal operations of the firm. These uncertainties could occur in different ways. The demand for the product(s) may be uncertain and deciding to produce each unit after an order for the product occurs may be impractical. The lead times for production may be long and customers may decide that they are unwilling to wait. In situations like these, it may be better to meet the demand from stock and produce/order additional quantities to replenish the stock. In these *make-to-stock* systems, the items are substitutable and the items from stock can be used to meet the demand for all the customers. The other sources of uncertainty in the system may include machine breakdowns, variability in the delivery of raw materials and components, etc.

Another reason why firms carry inventory is because of limitations on the capacity or the maximum rate at which the units can be produced. If the demand for the product is seasonal (cyclic), the firm may not be able to meet the demand if it decides to produce units only in periods when the demand occurs. Rather, it may be preferable to produce units at a steady rate even in periods when the demand is low and carry the item as inventory. This also serves the purpose of reducing the fluctuations in the workforce levels; more will be said about this in subsequent chapters when we study the tradeoffs between varying the inventory levels and fluctuations in the levels of the workforce.

Firms also carry inventory to achieve *economies of scale.* It would be very inefficient for a retailer to order items in very small batches from a central warehouse because of a certain fixed overhead associated with transporting the item. There is a considerable body of literature of the *optimal lot-sizing* or the *economic lot-scheduling* problem. (See Elmaghraby [Elm78] for details).

Every dollar tied up in inventory represents a loss, to the firm, of funds that could be invested elsewhere; the forfeited rate of return is equal to the firm's opportunity cost of capital. Managers make significant efforts to minimize the amount of inventory carried on the balance sheet. The methods that they employ range from using sophisticated information systems in order to keep close track of the inventory, to obtaining better forecasts of the demand and reducing the uncertainties in the operating environment. Accountants use *inventory turnover,* which is the ratio of annual cost of goods sold to the average daily inventory, as a measure of how well firms manage their inventory.

Inventories can be classified into different categories depending on the role that they perform or their *raison-d'être.* For example, consider the example of a factory that manufactures lawn mowers. A lawn mower that is manufactured in the winter months is stored in a warehouse in "anticipation" of high sales during the summer months; hence its designation is *anticipation stock.* For a nice discussion on the different classification of inventories, see Graves [Gra88b].

Many production processes can be viewed as networks of stages. The notion of a stage, as we describe it, is fairly general. In a factory, one could model a single machine or a cluster of machines as a stage in the production process. The production output is regarded as end item inventory for the stage, although this output may not be the final product emerging from the factory. At the other extreme, if items produced at factory A are used as components in the assembly of a product in factory B, one could model the whole of factory A as a single stage in the manufacture of the product made in factory B.

A production stage, as we have described above, is modeled as a black box. At any given time, the amount of work-in-process inventory at a stage is known. The raw material or components enter the stage, and some time later (this time could be fixed or variable), the processing of the item is complete. The modeling is done at the macroscopic level. To be more specific, we do not concern ourselves with details of tool changeovers, machine breakdowns, job scheduling, worker unavailability and other elements that would characterize the typical daily operation of a manufacturing facility. The production is measured in terms of the rate at which the work-in-process inventory leaves the stage and is available as end-item inventory.

In studying the behavior of a production stage, we need to find a mechanism to characterize the production rate as a function of the work-in-process inventory. To put it another way, suppose there is no work-in-process inventory at the stage — then the production rate must be zero, since there is nothing to process. If there is an upper bound on the rate at which the work-in-process can be processed, then when the levels of work-in-process inventory are high, the system would be operating at full capacity. But how does one characterize the behavior of the system at intermediate levels of inventory? One approach has been to assume that the system is processing inventory at its maximum rate, whenever there is work-in-process. This is the assumption underlying most queueing models of manufacturing systems.

In this thesis, we use a different approach in modeling the production process. We assume that the production rate is not constant, but proportional to the amount of work-in-process inventory in the system. This approach has been taken before (see Holt et al. [Holt55], Graves [Gra88b] and Karmarkar [Karm89]) and it has been argued that many manufacturing and service facilities behave in this manner. Consider the example of a barber shop. The amount of time that a barber spends with a customer when there is no one else in the shop is typically much more that the amount of time that would be spent when there are a few customers waiting in line. The barber works at a faster rate when there are more customers in the shop because the other customers should not be kept waiting too long. Even though the barber could work

at this rate throughout, he normally works at a much slower rate when there is a single customer.

Another explanation of why systems operate in this manner may be attributed to *Parkinson's law* applied to manufacturing and service facilities. When work-in-process inventory levels are low, there may not be as great a sense of urgency to process material and this may be reflected in a higher fraction of employees going on vacation, a higher proportion of worker hours spent on preventive maintenance etc. But when the levels of inventory rise, associated with this is usually a higher level of backlogs; and there is a desire to expedite orders and bring down inventories to more reasonable levels[1].

In the discussion so far, we have looked at some of the reasons why firms carry inventories and how the production stage is modeled as a black box. We also briefly talked about why the "proportional production rule" could be used to characterize the behavior of many manufacturing and service operations. In the discussion that follows, we shall talk about some of the central ideas in this thesis, namely, work-in-process inventories and safety stocks in manufacturing systems and some of the factors that affect them.

Let us, for the moment, restrict our attention to a single stage in a manufacturing system. Graves [Gra88b] has defined safety stock as the "excess inventories held beyond the minimum inventory level that would be possible in a deterministic and uncapacitated world." If one looks at the definition, there are two reasons why one needs to carry safety stock. The first one is because of the uncertainty or variability within the operating environment. The second reason for safety stock is because of limitations, or constraints on the production capacity. The factors that contribute to the stochastic variability of a manufacturing operation are manifold. As we had briefly stated earlier, the forecasts for the demand are likely to change over time and the production scheduling may have been done previously using different forecasts. The yield of the process may be variable or the production lead times may be random.

[1]The author had a helpful conversation with Prof. Jeff Rummel of Duke University on this subject

There might be uncertainties in the supply of raw materials or components from vendors.

Even if the operating environment were entirely deterministic, there might exist restrictions on production capacity that necessitate safety stock. In a manufacturing operation where several different products compete for limited production resources, safety stock may protect against known surges in demand when the system cannot react quickly enough. Graves calls this the *inflexibility of the manufacturing system.* He proposes a measure of the flexibility of a production stage that we use in later chapters of this thesis. This measure is different from another class of measures of flexibility called *entropic* measures (see Kumar [Kum87]). Other sources of inflexibility may include power outages or other work stoppages, union rules (which may dictate periodic work breaks) and so on.

Safety stocks help "preserve lot size integrity and keep the number of setups from growing, particularly at bottleneck workcenters" (Lambrecht et al. [Lam84]). They also play a smoothing or decoupling role among the different stages in a manufacturing operation (Graves [Gra88b]). This latter function that safety stocks perform, enables us to analyze a multistage operation by looking at each stage in isolation, and letting the production for the downstream stages serve as the demand for the upstream stages. By looking at these stages sequentially, the safety stock requirements at each stage can be determined.

In most of the literature on inventory theory, the paradigm is to set the inventory at that level where the cost of holding inventory plus the production cost is at a minimum. The inventory carrying costs, as well as production costs, are typically convex. While the cost of carrying excess inventory could be determined (it would simply be the dollar value of inventory times the opportunity cost of capital), it is much more difficult to quantify the cost of stockouts. If the firm is unable to meet the demand of certain customers, what is the cost that one associates with the resulting loss of goodwill, if any? It might happen that the customers regard a vendor, who faces frequent stockouts, as being too unreliable and decide to discontinue purchases

from that particular vendor, in which case the cost is the expected present value of the lost profits over a certain time horizon. If the component is used in the assembly of another product, a stockout for that component means delays in the assembly operation for the product, higher work-in-process inventories, idle workers and higher costs due to overtime. These delays could cause ripple effects throughout the system and the effects of these, in terms of costs, could be hard to measure.

In view of this difficulty of quantifying the cost of stockouts, an alternative approach is to impose service level constraints on the system. That is, a certain fraction of the time, say 95%, the demand for the item can be met from the safety stock. This fraction corresponds to the level of service that is being provided. Clearly, as this fraction increases, so will the level of safety stock. The tradeoff now is between the level of safety stocks and the level of service that they provide. This concept of service level has been around for some time now and it serves as "a surrogate for stockout costs in determining inventory policies."(Von Lanzenauer and Noori [Von86]).

In the preceding paragraphs, we had defined safety stocks in a manufacturing system and looked at the role that they play. We had also seen how different factors affect the level of safety stocks. In this thesis, we study in greater detail, some of these factors and develop models to analyze their effects.

The first problem that we consider is a problem related to capacity allocation. Consider a single stage in a manufacturing system. A single stage in the system consists of (i) a production stage (modeled as a black box) and (ii) a buffer containing the end-item inventory (see Figure 1.1). Items produced at the stage are stored in the buffer as end-item inventory and the demand is met from the end-item inventory. We will explain the model in greater detail in subsequent chapters. In Chapter 2, we consider the problem of capacity allocation in a multistage manufacturing system. The level of safety stock at a stage depends on the amount of production capacity at the stage. The greater the production capacity, the lower is the level of safety stock required since the stage is able to react to changes in demand much more quickly. Suppose there are two production stages in tandem as shown in Figure 1.2 and we

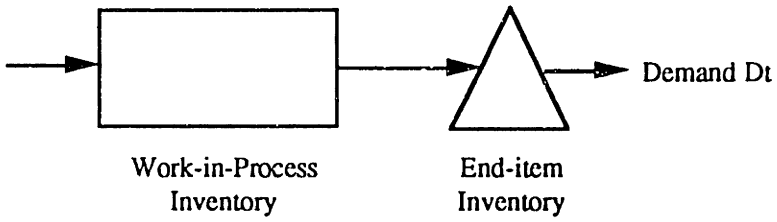

Figure 1.1: A single stage in a manufacturing system

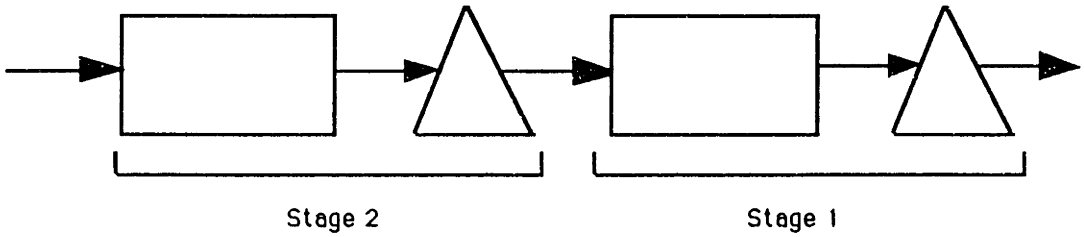

Figure 1.2: Two stages in tandem

have a fixed amount of capacity to assign between the two stages. If the objective is to minimize the total safety stock at both stages, the optimal capacity assignment will ensure that the line is balanced, i.e., neither stage becomes a bottleneck. The question is: What capacity allocation will minimize the total of the work-in-process inventory and safety stock in this configuration. One could extend this problem to a system that comprised a network of stages. We will look at this problem in greater detail in Chapter 2. We will also compare some of the analogous problems that arise in the context of queueing networks.

In Chapter 3, we consider the following problem: suppose that there are two items produced at this single stage. There could be one or multiple machines at this stage and we assume that the units of production capacity are infinitely divisible. That is, given a fixed capacity allocation, the amount of capacity alloted for the production of each item could take on values in a continuum, provided the sum of the capacities equals the total available capacity. The question that the decision maker faces is the

following: Should the two items be processed on two separate, dedicated machines? Or should they be processed on a single flexible machine, and the input streams for the two items be merged? The advantage of having flexible machines is that the total idle time on the two machines is minimized. In the queueing parlance, the sojourn time in the system is minimized.

However, there are benefits from having separate, dedicated lines for each item. First, the cost of installing two dedicated lines may be lower than the cost of installing one flexible line with the same production capacity. In some cases, there might be certain benefits associated with having two separate lines. If the demand for item 1 is fairly stable, i.e., the coefficient of variation is very low, and the demand for item 2 is highly variable, it may be better not to merge the two streams. By merging the two streams, the production output stream for item 1 may become much more variable because of the demand for item 2 is highly variable. Since the production output stream for item 1 is more variable when the two streams are merged, the safety stock level for item 1 may be higher. However, since the lines are dedicated, the idle times on the machines are bound to be higher because of *incompatibility* between the two machines, i.e., if a machine in line 1 is idle, item 2 cannot be processed on this machine. This model is meant to serve as building block for a model that addresses the broader question of dedicated versus flexible factories. We will discuss some of the broader issues that arise in this context in Chapter 2 or 3.

In all of the analysis in Chapters 2 and 3, we assume that the system operates according to a proportional production rule. Earlier, we had argued why a rule of this nature characterizes the actual behavior of many manufacturing systems. The rule that we use is linear, i.e., at a given time, the production rate is set equal to a fixed fraction of the work-in-process inventory at the stage at that time. In other words, if, in every time period, we decide to produce 20% of the work-in-process, then the production, as a function of the work-in-process inventory, is shown in Figure 1.3 . The production function is linear with slope 0.2. The advantages of assuming such a rule are obvious; the resulting system is linear and the analysis is simplified considerably.

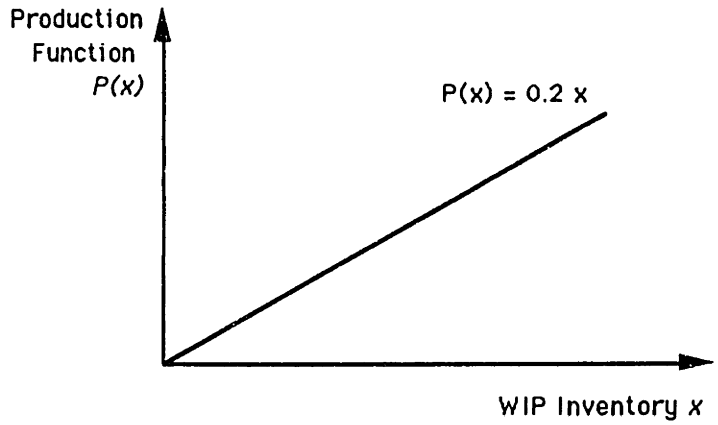

Figure 1.3: The Linear Production rule

(In general, linear systems are far more tractable than nonlinear systems). Besides, a rule of this nature can be shown to be optimal when the production and inventory holding costs are quadratic (See Holt et al. [Holt55] and Graves [Gra88b]). The only drawback of this rule is that the production capacity is assumed to be unbounded. For example, if there are 100 units of work-in-process inventory and one decides to process 20% of the work-in-process , then in that period, 20 units would be assumed to be produced even though the capacity of the stage is, say, 10 units. This can be remedied by an appropriate choice of the slope of the production function. We will talk about this rule in greater detail in subsequent chapters. In Chapter 4, we propose a proportional production rule that explicitly takes into account the constraint on production capacity. Consequently, the resulting rule is nonlinear. Using this rule, the model can be analyzed for a single stage and expressions can be obtained for the mean and the variance of the work-in-process inventory. The linear rule can be seen to be a special case of the nonlinear rule that we propose.

In Chapter 5, we summarize the main results of this thesis and point out some directions for future work.

The main body of this thesis comprises Chapters 2, 3 and 4. It is not necessary to read the chapters in a sequential fashion; they have been written so that they can

be read fairly independent of each other. While the inevitable consequence of this is a certain degree of overlap, we have tried to keep this to a minimum.

To conclude this chapter, this thesis looks at work-in-process inventories and safety stocks in manufacturing systems and the different factors that affect them. We have developed different models to study the effect of these factors. We have formulated models in both discrete time and continuous time. We will justify the assumptions as well as the methodologies when we describe the models in the later chapters. In previous work (see Graves [Gra88b]), for a single stage with the demand process and maximum processing capacity given, the lead time and consequently, the WIP inventory and safety stock at the stage can be determined. In this thesis, we examine how these measures could be affected by changing the available capacity at the stage. We show that the *optimal* capacity allocation for the multistage problem is a closed form expression that is similar to expressions that have been obtained earlier for queueing networks (see Kleinrock [Klei64] and Wein [Wein89]).

In this thesis, we also show that under certain conditions, when many products are being manufactured at a facility, and when the demand for a certain product is stable, it may be advantageous to manufacture this product separately. We argue that the queueing model is inappropriate in this scenario and demonstrate how this could be modeled using the linear production rule that we had described earlier.

We also illustrate how different models of production systems that have been proposed in literature can be viewed as asymptotic or special cases of a production system that we propose. Using this production rule to model system behavior, we obtain a closed form expression for the WIP inventory. It can be shown that the expressions obtained previously using other models correspond to different components of the expression in our model. Besides, the rule is parameterized which allows us to model the system behavior more closely. This is a summary of the major thesis findings and these points will become clearer as we delve into the issues in greater detail in subsequent chapters.

Chapter 2

Capacity Allocation, Work-in-Process Inventories and Lead Times

2.1 Introduction

In many manufacturing organizations, there has been a renewed commitment to excellence in manufacturing over the last decade. In order to evaluate their manufacturing operations, maximize operating efficiency and measure costs, firms have different control mechanisms in place. These mechanisms are used to provide managers with timely feedback, assist in the day-to-day decision making as well as help in devising the overall long-term strategy of the firm.

Included in these mechanisms are several performance measures that enable firms to evaluate performance and one of them is the average inventory level. Inventories comprise a significant amount of a firm's working capital. The benefits of having lower levels of inventory are manifold: lower cost of carrying inventory, smaller factory areas, reduced material handling and transport, smaller percentages of defects etc. Organizations strive to keep their inventories under control and they employ a vast array of methods to achieve this.

Some authors have argued that there are serious flaws in the manner in which costs are allocated. These allocations cause systems to behave in a manner that is far from beneficial. For example, consider an investment in a new machine. There is a desire to keep the utilization of the machine as high as possible in order to justify the investment. But the investment in the machine is a sunk cost and it is independent

of the production volume at the machine. Second, using utilization as a measure of performance implies that the firm can sell everything that it produces. While this may be true, few firms find themselves in this enviable position.

Some authors have argued that the traditional cost accounting systems that are currently in place in most companies are archaic. Kaplan [Kapl83] has argued that besides their inability to provide key nonfinancial information, they also distort product costs. These control systems were developed several decades ago when the nature of the markets, as well as the production techniques, were vastly different. These systems were designed to produce data for external reporting, but they fail to convey the reality of the operational environment. The use of these measures in the decision making process causes managers to choose strategies that are not in the best interest of the organization. Moreover, managers being evaluated on the basis of these measures would adopt a short term focus rather than develop long term strategies. Goldratt [Gold86] has proposed three operational measures of performance in his manufacturing accounting system: throughput, inventories and operational expense.

One measure that has been proposed in the literature, yet seldom used in practice, is lead time. The necessity of a firm to quote delivery dates results in the notion of planned lead times. Of equal importance is the ability to deliver the product within the given time window (see Harrison et al. [Har90]). This measure is different from the traditional measures that are currently in place. In a certain sense, this measure is antithetical to the measure of performance described earlier, namely, machine utilization. Since lead times are proportional to $(1-\rho)^{-1}$ where ρ in the machine utilization, usually associated with high utilizations are high lead times. Karmarkar [Karm90] provides a comprehensive survey on the literature on lead time management. Although this measure by no means encompasses every attribute of an organization, there are many other desirable characteristics associated with systems with low lead times.

In a competitive market, firms need the ability to respond to market changes and customer needs quickly. This is not the same thing as flexibility although many flexi-

ble manufacturing systems have low lead times. Lower lead times also result in lower levels of work-in-process (WIP) inventory and safety stocks. In the model that we present in subsequent sections, we explicitly show how the work-in-process inventory and safety stocks vary with lead times. If the lead time were to be incorporated among the other measures of performance, one may be able to justify an investment in a machine that leads to a substantial reduction in lead times — something that the traditional measures may not provide.

In the case where the quality of the product (in terms of defective or nondefective) can be determined only after the completion of production, lower lead times also provide faster feedback to trace the cause of defects. In addition, since the levels of WIP inventory and safety stock are reduced, the number of defectives in the pipeline are also reduced.

In the preceding paragraphs, we have tried to provide a cogent argument for including lead times among the measures of performance. In the analysis that follows, we use a model where the planned lead time is a decision variable and the lead times are considered when determining the levels of WIP inventory and safety stock. We examine the question of how to allocate capacity among the different stations and how it affects the lead time. We also provide tradeoff curves that describe how the overall processing capacity affects the lead time as well as the WIP inventory and safety stock in the system.

Kleinrock [Klei64] considered the problem for an n-stage Jackson network where $\bar{\mu} = (\bar{\mu}_1, \ldots, \bar{\mu}_n)$ is the vector of mean arrival rates ($\bar{\mu}_i$ being the effective arrival rate to stage i). If $C = (C_1, \ldots, C_n)$ is the corresponding vector of processing rates and the sum of the processing rates must not exceed K, then the allocation of the processing rates that minimizes the expected equilibrium number of customers in the system is given by

$$C_i^* = \bar{\mu}_i + \frac{\sqrt{\bar{\mu}_i}}{\sum_j \sqrt{\bar{\mu}_j}}(K - \sum_j \bar{\mu}_j)$$

This is the well known square root formula. It says that the optimal processing rate

for station i is equal to the effective arrival rate to station i plus an amount that is proportional to the square root of the effective arrival rate to the station.

Wein [Wein89] considered a generalized Jackson network which could be analyzed using a Brownian approximation. This is a two-moment approximation and he showed that the optimal capacity allocation in this formulation is similar to the square root formula, but the square root factor is replaced by another factor that captures the variability. (We will discuss this in greater detail in Section 2.4.)

Bitran and Tirupati [Bitr88] use the open network of queues model to derive tradeoff curves for the lead times and the WIP inventory as a function of the total capacity in the system. In this chapter, we show how one could derive the tradeoff curves for a system (with a linear production rule) that we shall describe in subsequent sections. The outline of this paper is as follows: In section 2.2, we describe the linear model in discrete and continuous time. In section 2.3, we show how this can be extended to the multistage framework. In section 2.4, we obtain the steady state equations and formulate the optimization problem. In section 2.5, we provide a few tradeoff curves as well as a discussion of the overall model.

2.2 The Basic model

2.2.1 The discrete time model

The model that we describe below is due to Graves [Gra88b]. Consider a single stage in a manufacturing system where a single item is produced (see Figure 2.1). The model can be extended to the situation where multiple items are produced at the stage. The single item case that we consider here can be thought of either as a single item being produced, or multiple items that are considered in aggregrate. The inventory at the stage is of two types: work-in-process inventory (WIP) and end-item inventory. The end-item inventory, also known as safety stock, is used to satisfy the demand at the stage and any excess demand is assumed to be backordered. The original model is a discrete time model and was later extended to the continuous

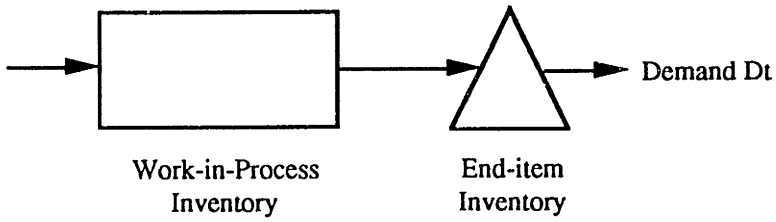

Figure 2.1: A single stage in a manufacturing system.

time framework [Gra88a]. We shall describe the model in discrete time and then show how the extension to continuous time can be performed. We choose to present the model in this manner because we believe that it is much more understandable when presented in this way.

The demand for the item in time period t is D_t and is normally distributed with mean μ and variance σ^2. Moreover, the demands in the different time periods are independent and identically distributed . The process is assumed to have an infinite history and to have reached steady state. The demand in each period occurs at the beginning of the period. Let I_t be the end-item inventory at the start of period t after the demand for the period has been satisfied. Whenever a certain number of units of the item is taken from the end-item inventory in order to meet the demand, an identical amount of raw material is added to the work-in-process inventory. Thus, in period t, D_t units are taken from the end-item inventory to meet the demand and so the amount of raw material R_t released into the system in period t is set equal to D_t. Let X_t be the work-in-process inventory at the start of period t after the amount of raw material R_t is released into the system. Let P_t be the amount of WIP inventory processed in period t and this becomes available at the end of period t. In this model, the control rule that we use for production is linear. The production function is given by

$$P_t = \alpha X_t \qquad 0 \leq \alpha \leq 1 \tag{2.1}$$

The amount produced in period t is a linear function of the amount of WIP at the start of period t. In every period, a fraction α of the WIP at the start of the period is processed. Therefore, the flow balance equations for the WIP inventory and the safety stock, respectively, are

$$X_t = X_{t-1} - P_{t-1} + R_t \tag{2.2}$$

and

$$I_t = I_{t-1} + P_{t-1} - D_t \tag{2.3}$$

Using the fact that $R_t = D_t$ and by adding equations (2.2) and (2.3), we get

$$\begin{aligned} X_t + I_t &= X_{t-1} + I_{t-1} \\ &= B \text{ (constant)} \end{aligned} \tag{2.4}$$

The total of the WIP inventory and safety stock in the system is called the base stock and is denoted by B. Since $R_t = D_t$, from equation (2.2), we get

$$\begin{aligned} X_t &= X_{t-1} - P_{t-1} + D_t \\ &= (1-\alpha)X_{t-1} + D_t \qquad\qquad \text{since } P_{t-1} = \alpha X_{t-1} \\ &= \sum_{s=0}^{\infty}(1-\alpha)^s D_{t-s} \end{aligned} \tag{2.5}$$

since the process has an infinite history. Taking the expected value of both sides in equation (2.5) and using the fact that $E[D_t] = \mu$ for all t, we get

$$\begin{aligned} E[X_t] &= \sum_{s=0}^{\infty}(1-\alpha)^s \mu \\ &= \frac{\mu}{\alpha} \end{aligned} \tag{2.6}$$

In a similar manner, taking the variance of both sides in equation (2.5) and using the fact that $Var[D_t] = \sigma^2$ for all t and the demands in different time periods are independent, we get

$$\begin{aligned} Var[X_t] &= \sum_{s=0}^{\infty}(1-\alpha)^{2s}\sigma^2 \\ &= \frac{\sigma^2}{2\alpha - \alpha^2} \end{aligned} \tag{2.7}$$

From equation (2.5), we see that X_t is the sum of independent, normally distributed random variables and hence, X_t has a normal distribution. Therefore, the mean $E[X_t]$ and the variance $Var[X_t]$ completely characterize the distribution of X_t. As the value of α decreases, both the mean and variance of X_t increase. In other words, as the fraction α of the WIP inventory processed becomes smaller, the inventory level follows the demand less closely and so it fluctuates more widely.

Since $X_t + I_t = \text{constant}$, I_t also has a normal distribution and $Var[I_t] = Var[X_t]$. Now it remains to choose the value of $E[I_t]$, which is equivalent to setting B. When choosing the expected value $E[I_t]$, we would like to ensure that 95% of the time, the demand can be met from the safety stock. For instance, if we set the service level to 0.95, then

$$P[I_t > 0] = 0.95$$

The service level of 95% that we choose here is arbitrary. A firm may decide that a service level of, say, 99.5% is more appropriate and accordingly, would set the levels of safety stock. For the remainder of this discussion, we shall use a service level of 95%, qualifying it with the comment that this choice is a reasonable one in this setting. The value zero corresponds to the 5th percentile of the distribution of I_t and so we set

$$\begin{aligned} E[I_t] &= k_{0.95}\sqrt{\text{Var}(I_t)} \\ &= 1.645\frac{\sigma}{\sqrt{2\alpha - \alpha^2}} \end{aligned} \tag{2.8}$$

Therefore, the level of base stock can be determined from equation (2.4).

$$\begin{aligned} B &= E[X_t] + E[I_t] \\ &= \frac{\mu}{\alpha} + k_{0.95}\frac{\sigma}{\sqrt{2\alpha - \alpha^2}} \end{aligned} \tag{2.9}$$

There is one thing that still remains: How does one choose the value of α? Since the production function in this model is linear, it is unbounded. The idea in this model is to choose the value of α so that the production rate dictated by the model

is consistent with the maximum production capacity. Since $P_t = \alpha X_t$, P_t is normally distributed with mean μ and variance $\frac{\alpha^2\sigma^2}{2\alpha-\alpha^2}$. As α decreases, $Var[P_t]$ decreases. Since the level of base stock B decreases as α increases, to minimize inventory, we would like the value of α to be as large as possible. In this case, we choose the largest value of α such that 95% of the time, the production rate dictated by the model is less than the maximum processing capacity. That is, choose the largest value of α so that

$$E[P_t] + k_{0.95}\sqrt{Var(P_t)} \leq C \tag{2.10}$$

where C is the maximum processing capacity at the stage. Again, the value of 95% that we have chosen here is arbitrary. This is different from the service level for setting safety stocks that we had talked about earlier. We choose a level of 95% here to ensure that the lead times provided by the model are realistic and descriptive of the actual system behavior. Moreover in this analysis, the levels that we choose, both for setting safety stocks (equation (2.8)) and setting lead times (equation (2.10)), are identical, but this need not always be the case. Since in the stage, a fraction α of the WIP inventory is processed, on average, it will take $\frac{1}{\alpha}$ time periods for work to be processed. Therefore, $\frac{1}{\alpha}$ can be interpreted as the lead time for the stage.

2.2.2 The extension to continuous time

We now extend the results of this model to the case of a continuous time framework. In this case, the scenario is almost identical to the one described previously. The cumulative demand upto time t is D_t and this is assumed to be normally distributed with mean μt and variance $\sigma^2 t$, i.e.,

$$D_t = \mu t + \sigma W_t \qquad \text{for all } t > 0 \tag{2.11}$$

where W_t is a standard one-dimensional Brownian motion. In other words, D_t is a diffusion process with drift μ and dispersion σ. Unlike in the discrete time model where μ is the mean demand in a period, μ in the continuous time model is a drift rate

so that the mean demand in the interval $(t, t+dt)$ is μdt. X_t is the WIP inventory at time t and I_t is the end-item inventory at time t. P_t is the instantaneous production rate at time t and it is set equal to αX_t. As in the discrete model, the cumulative amount of raw material released upto time t is equal to D_t, i.e., $R_t = D_t$. Therefore, we can write the flow balance equations for the WIP in the time interval $(t, t+dt)$.

$$\begin{array}{ccccc} \text{Change in WIP} & = & \text{Amount released into} & - & \text{Amount processed} \\ \text{in } (t, t+dt) & & \text{system in } (t, t+dt) & & \text{in } (t, t+dt) \end{array}$$

Therefore,

$$\begin{aligned} dX_t &= dR_t - P_t dt \\ &= dD_t - P_t dt && \text{since } dR_t = dD_t \\ &= \mu dt + \sigma dW_t - \alpha X_t dt && \text{since } P_t = \alpha X_t \\ &= (\mu - \alpha X_t)\, dt + \sigma dW_t, && \mu > 0,\ \sigma > 0 \end{aligned} \tag{2.12}$$

It can be shown that the ergodic distribution of X_t is normal with mean $\frac{\mu}{\alpha}$ and variance $\frac{\sigma^2}{2\alpha}$. In other words, for the continuous time model, the mean level of WIP inventory is

$$E[X_t] = \frac{\mu}{\alpha} \tag{2.13}$$

and the variance of the WIP inventory level is given by

$$Var[X_t] = \frac{\sigma^2}{2\alpha} \tag{2.14}$$

Therefore, the level of base stock is

$$B = \frac{\mu}{\alpha} + k_{0.95}\frac{\sigma}{\sqrt{2\alpha}} \tag{2.15}$$

If we compare the results of the discrete time model and the continuous time model, we see that the mean value for $E[X_t]$ for both models is the same while the variance $Var[X_t]$ differs by a factor of $(1-\frac{\alpha}{2})$. We claim that the results for the base level stock that both models provide are extremely close and the reader is referred to Chapter

4 for a more detailed explanation. In the discrete model, α was the fraction of the WIP inventory that was processed in a given period. Since we could not process more WIP inventory than there was present in the system, α had to be less than one in the discrete model. Moreover, since the lead time is equal to $\frac{1}{\alpha}$, if α was greater than one, the lead time is less than one. If, on average, the WIP inventory became part of the end-item inventory in less than one time period, it simply meant that the choice of the length of a time period in the model was inappropriate. In the continuous time model, P_t is the production rate. Therefore, there is no upper bound on the value of α since P_t is a rate and not the actual amount produced. Besides, we are using the continuous time framework so there is really no problem if the value of the lead time $\frac{1}{\alpha}$ falls below one.

In this chapter, we shall use the continuous time framework in extending the results to the multistage case. The linear production rule facilitates the analysis in the multistage case. The system is much more difficult to analyze when the production function is nonlinear.

2.3 The Multistage System

In this section, we shall develop the model for the multistage system assuming that each stage behaves like the single stage that we had described in the previous section. We shall use the continuous time setting to describe this model.

2.3.1 The model of the multistage system

Consider a network with n stages labelled $1, \ldots, n$. The single stage that we described earlier is used as a building block in this multistage network and the model uses the continuous time framework. The cumulative exogeneous demand upto time t is $\bar{D}_t$ and it is an $n \times 1$ vector where D_{it} is the exogeneous demand for the item produced at the ith stage, $i = 1, \ldots, n$. It can be represented by an n-dimensional diffusion process with an $n \times 1$ drift vector $\mu_{n\times 1}$ and an $n \times n$ dispersion matrix $\Gamma_{n\times n}$. Therefore, the

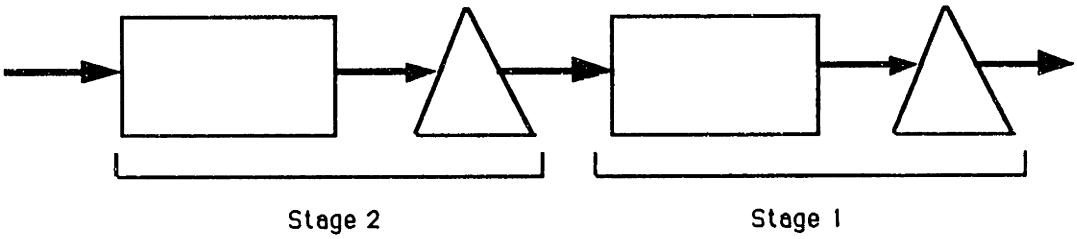

Figure 2.2: A two stage network.

cumulative demand process can be written in differential form as

$$d\bar{D}_t = \mu_{n\times 1}dt + \Gamma_{n\times n}d\bar{W}_t \tag{2.16}$$

where $\bar{W}_t$ is a vector of n independent, one-dimensional Brownian motions.

Let $\bar{X}_t = (X_{1t},\ldots,X_{nt})$ be the vector of work-in-process inventories at time t where X_{it} = WIP inventory at stage i at time t, $i = 1,\ldots,n$. Similarly, let $\bar{I}_t = (I_{1t},\ldots,I_{nt})$ be the vector of end-item inventories at time t. Let $\bar{R}_t = (R_{1t},\ldots,R_{nt})$ be an n-dimensional vector where R_{it} = Cumulative amount released into stage i upto time t, $i = 1,\ldots,n$. Let $\bar{P}_t = (P_{1t},\ldots,P_{nt})$ be the vector of instantaneous production rates. If P_{it} is the instantaneous production rate at stage i at time t, then the amount processed at stage i in the interval $(t, t+dt)$ is $P_{it}dt$.

The relationship between the different stages is given by a matrix $A = [a_{ij}]_{n\times n}$ where a_{ij} = number of units of item i required to process one unit of item j. Consider a two stage network as shown in Figure 2.2. Suppose two units of item 2 are required to assemble one unit of item 1. (When we say item j, we actually mean the item processed at stage j, but since each stage processes only one item, there is no ambiguity.) The *cumulative induced demand* upto time t is given by

$$\begin{aligned}\text{Cumulative induced demand upto time } t &= \bar{D}_t + A\bar{D}_t + A(A\bar{D}_t) + \ldots \\ &= \left[I + A + A^2 + \ldots\right]\bar{D}_t \\ &= (I-A)^{-1}\bar{D}_t \end{aligned} \tag{2.17}$$

In this example, $A = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}$. Suppose there is an exogeneous demand that occurs at stage 1 and no exogeneous demand at stage 2. Then, the end-item inventory at stage 2 is raw material for stage 1. In addition, suppose the exogeneous demand for item 1 upto time t is 10 units. Then the amount taken from the end-item inventory of stage 2 upto time t must be 20 units since two units of item 2 are required to process each unit of item 1. In this example, $\bar{D}_t = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$. Therefore, from equation (2.17), the cumulative induced demand upto time t is given by

$$\begin{aligned} \text{Cumulative induced demand upto time } t &= (I - A)^{-1}\bar{D}_t \\ &= \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 10 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 10 \\ 20 \end{pmatrix} \end{aligned}$$

The cumulative amount released into stages 1 and 2 upto time t are 10 and 20 units respectively. If we write this in vector form, we get

$$\bar{R}_t = (I - A)^{-1}\bar{D}_t$$

i.e., the vector of cumulative releases upto time t must be equal to the cumulative induced demand upto time t. Although the demand at stage 2 upto time t is 20 units, the amount released into stage 1 upto time t is 10 units. We measure the cumulative amount released in terms of the number of units of the item in the stage into which it is released, and not the number of units of the item in the predecessor stage. In this analysis, we allow the end-item inventories at a stage to go negative. In other words, we assume that even if the inventory at the upstream stage falls below zero, units of the end-item inventory can still be released into the downstream stage. We will set base stocks, though, so that the probability of this event is small.

If $\bar{D}_t$ is a diffusion process with drift vector $\mu_{n\times 1}$ and dispersion matrix $\Gamma_{n\times n}$, then the cumulative induced demand process $(I - A)^{-1}\bar{D}_t$ is a diffusion process with drift $\bar{\mu} = (\bar{\mu}_1, \ldots, \bar{\mu}_n)^T = (I - A)^{-1}\mu_{n\times 1}$ and dispersion matrix $\bar{\Gamma} = [\bar{\gamma}_{ij}]_{n\times n} = (I - A)^{-1}\Gamma_{n\times n}$. It should be noted that $\bar{\Gamma}$ is the dispersion matrix for the cumulative

induced demand process. Therefore, the corresponding covariance matrix is $\bar{\Gamma}\bar{\Gamma}^T$ which is symmetric. In differential form, this can be written as

$$\begin{aligned} d\left[(I-A)^{-1}\bar{D}_t\right] &= \left[(I-A)^{-1}\mu_{n\times 1}\right]dt + \left[(I-A)^{-1}\Gamma_{n\times n}\right]d\bar{W}_t \\ &= \bar{\mu}dt + \bar{\Gamma}d\bar{W}_t \end{aligned} \tag{2.18}$$

Let $\frac{1}{\alpha_i}$ be the lead time at stage i, $i = 1, \ldots, n$. Since each stage obeys the linear production rule, we have

$$P_{it} = \alpha_i X_{it} \qquad i = 1, \ldots, n. \tag{2.19}$$

If we let $[\alpha]_{dg} = diag(\alpha_1, \ldots, \alpha_n)$, we can write equation (2.19) in vector form as

$$\bar{P}_t = [\alpha]_{dg}\bar{X}_t \tag{2.20}$$

In the next subsection, we obtain the flow balance equations for the work-in-process and the end-item inventories.

2.3.2 Flow Balance Equations

The flow balance equations for the work-in-process inventories can be written by considering the change in the WIP in the time interval $(t, t+dt)$. For the work-in-process inventory, we have

Change in WIP system in $(t, t+dt)$	=	Amount released into system in $(t, t+dt)$	−	Amount processed in $(t, t+dt)$

Therefore,

$$d\bar{X}_t = d\bar{R}_t - \bar{P}_t dt \tag{2.21}$$

Similarly, the flow balance equations for the end-item inventories can be written as follows:

Change in end-item inventory in $(t, t+dt)$	=	Amount processed in time $(t, t+dt)$	−	Amount released into successor stages in $(t, t+dt)$

Therefore, the stochastic differential equation for the end-item inventory is

$$d\bar{I}_t = \bar{P}_t dt - (I - A)^{-1} d\bar{D}_t \tag{2.22}$$

We choose $d\bar{R}_t = (I - A)^{-1} d\bar{D}_t$. In other words, the amount that is released into the work-in-process inventory at each stage is equal to the induced demand in the time interval $(t, t + dt)$. Adding equations (2.21) and (2.22), we get $d\bar{X}_t + d\bar{I}_t = 0$ or $\bar{X}_t + \bar{I}_t =$ constant for all t. The equations for the work-in-process inventory can be rewritten as

$$\begin{aligned} d\bar{X}_t &= (I - A)^{-1} d\bar{D}_t - \bar{P}_t dt \\ &= \bar{\mu} dt + \bar{\Gamma} d\bar{W}_t - \bar{P}_t dt && \text{from equation (2.18)} \\ &= \left(\bar{\mu} - [\alpha]_{dg} \bar{X}_t\right) dt + \bar{\Gamma} d\bar{W}_t && \text{from equation (2.20)} \end{aligned} \tag{2.23}$$

We have shown that the WIP inventory can be represented by a system of n linear stochastic differential equations, namely (2.23). Since the equations are linear, they are relatively easy to solve. In the next section, we obtain the mean and the variance of the vector $\bar{X}_t$ in steady state.

2.4 The Optimization Problem

In this section, we obtain the mean and the variance of the vector $\bar{X}_t$ in steady state. We do not present the proofs leading to these results here, but have deferred them to the appendix. The interested reader can always refer to them there. The result for the mean of the WIP inventory vector is stated in the following theorem.

Theorem 1 *The mean vector of the WIP inventory level is given by*

$$\begin{aligned} E\left[\bar{X}_t\right] &= [\alpha]_{dg}^{-1} \bar{\mu} \\ &= \left(\frac{\bar{\mu_1}}{\alpha_1}, \ldots, \frac{\bar{\mu_n}}{\alpha_n}\right)^T \end{aligned} \tag{2.24}$$

Proof: See the appendix. □

This result says that the mean WIP inventory level at stage i is equal to $\frac{\bar{\mu}_i}{\alpha_i}$. This is similar to the result for the mean WIP level for a single stage. The only difference is that the mean demand is replaced by the mean induced demand for the stage.

$$E[X_{it}] = \frac{\bar{\mu}_i}{\alpha_i} \qquad i = 1, \ldots, n \tag{2.25}$$

We obtain an expression for the variance of the WIP inventory level at each stage similar to the expression for a single stage that we had obtained earlier. The result is stated in the following theorem.

Theorem 2 *The variance of the WIP inventory level at each stage is given by*

$$Var[X_{it}] = \frac{\bar{\sigma}_i^2}{2\alpha_i} \qquad i = 1, \ldots, n \tag{2.26}$$

where $\bar{\sigma}_i^2 = \sum_{j=1}^{n} \bar{\gamma}_{ij}^2$ *and* $\bar{\Gamma} = [\bar{\gamma}_{ij}]_{n \times n}$ *is the dispersion matrix of the induced demand process.*

Proof: See the appendix. □

The result is similar to the expression for the single stage case. The only difference is that we replace σ^2 in equation (2.14) by $\bar{\sigma}_i^2 = \sum_j \bar{\gamma}_{ij}^2$, $i = 1, \ldots, n$ which are the diagonal terms on the induced demand covariance matrix $(I - A)^{-1}\Gamma[(I - A)^{-1}\Gamma]^T$.

We have obtained expressions for the mean and variance of the inventory level at each stage. As in the case of a single stage, the idea is to choose the value of α_i so that the lead time at each stage is consistent with the production capacity at the stage. Let d_i be the average cost per unit of inventory at stage i, $i = 1, \ldots, n$. We are now in a position to state the basic optimization problem. Suppose there were a total of K units of capacity available. How does one assign this capacity among the n stages so that the total dollar value of the WIP inventory plus safety stock in the system is minimized? And given the different capacities that are assigned to

different stages, what are the lead times for this *optimal* capacity allocation? To state it slightly differently, how does one allocate these K units of capacity among these n stages and choose lead times consistent with the capacity allocation at each stage so that the total dollar value of base stock in the system is minimum? Therefore, the optimization problem (P) is

$$\text{Minimize} \quad \sum_{i=1}^{n} d_i B_i \qquad \text{Problem (P)}$$

subject to

$$E[P_{it}] + k_{0.95}\sqrt{Var(P_{it})} = C_i \qquad i = 1, \ldots, n. \qquad (P1)$$

$$\sum_{i=1}^{n} C_i = K \qquad (P2)$$

We would like to make a few remarks about the optimization problem (P). The decision variables in this problem are C_i and α_i, $i = 1, \ldots, n$. The problem is a NLP and it is very difficult to obtain a closed form solution for the optimal capacity allocation C_i^* and the corresponding lead times $\frac{1}{\alpha_i^*}$, $i = 1, \ldots, n$. Since $E[P_{it}] = \bar{\mu}_i$ and $Var[P_{it}] = \frac{\alpha_i \bar{\sigma}_i^2}{2}$, constraint P1 can be rewritten as

$$\bar{\mu}_i + 1.645\sqrt{\frac{\alpha_i \bar{\sigma}_i^2}{2}} = C_i, \qquad i = 1, \ldots, n$$

Note that we have substituted the value of $k_{0.95} = 1.645$ in the above equation. In equation (2.10), we had said that the value of α_i is chosen so that

$$E[P_{it}] + k_{0.95}\sqrt{Var[P_{it}]} \leq C$$

In the optimization problem, it is not hard to see that this constraint would be tight for any optimal solution. Because the LHS is equal to $\bar{\mu}_i + k_{0.95}\frac{\bar{\sigma}_i}{\sqrt{2}}\sqrt{\alpha_i}$, making the constraint tight by increasing the LHS increases the value of α_i, which decreases the WIP and safety stock at stage i. Hence we write this constraint as a tight constraint in the optimization problem.

In this model, d_i is the cost per unit of inventory at stage i. Given the relationship between the different stages in the network, it may be that the items in one stage are assembled using components from different stages. The cost of a single unit of the assembled item would be greater than the cost of the component parts so the objective function may be regarded as the dollar value of the inventory in the system. Since we are in the continuous time domain, $\bar{P}_t$ is a vector of production rates and so strictly speaking, K is the total capacity rate and C_i is the maximum processing rate at stage i. Henceforth, in this chapter, when we refer to the total capacity K and the capacity at stage i, we are actually referring to a rate or processing capacity per unit time.

Suppose we look at a slightly different problem. Instead of considering the base stock at each stage, suppose we only consider the WIP inventory. So instead of the objective function being $\sum_{i=1}^{n} d_i B_i = \sum_{i=1}^{n} d_i \left(\frac{\bar{\mu}_i}{\alpha_i} + 1.645\sqrt{\frac{\alpha_i \sigma_i^2}{2}} \right)$, we consider the function $\sum_{i=1}^{n} d_i \frac{\bar{\mu}_i}{\alpha_i}$. How can we justify this approach where we merely look at the WIP inventory instead of looking at both the WIP inventory and safety stock in the objective function? For many manufacturing systems, $\frac{\bar{\sigma}_i}{\bar{\mu}_i} \ll 1$, $i = 1, \ldots, n$, that is, the coefficient of variation of the induced demand at each stage is lower than one. Also, the WIP term has a factor of $\frac{1}{\alpha_i}$ and the safety stock has a factor of $\frac{1}{\sqrt{\alpha_i}}$ which is smaller for values of $\alpha_i \in (0, 1]$. If $\alpha_i > 1$, the levels of WIP inventory and safety stocks may be comparable, but the absolute levels for both are lower for higher values of α_i. What we are trying to say is that the WIP component of the base stock is usually larger than the safety stock component. Even if they were the same, both the WIP inventory and the safety stock exhibit the same behavior as a function of α (although they may vary at different rates). The advantage of considering only the WIP for the modified problem is that we are able to get closed form expressions for the optimal capacity allocation and the lead times at each stage for any network configuration. The modified optimization problem (P') is

$$\text{Minimize} \qquad \sum_{i=1}^{n} d_i \frac{\bar{\mu}_i}{\alpha_i} \qquad \text{Problem } (P')$$

subject to

$$\bar{\mu}_i + k_{0.95}\sqrt{\frac{\alpha_i \bar{\sigma}_i^2}{2}} = C_i \qquad i = 1, \ldots, n. \qquad (P'1)$$

$$\sum_{i=i}^{n} C_i = K \qquad (P'2)$$

The result in the next theorem provides the optimal capacity allocation and the corresponding lead times for the problem (P').

Theorem 3 *The optimal capacity allocation for the problem (P') is given by*

$$C_i^* = \bar{\mu}_i + \frac{(d_i\bar{\mu}_i)^{\frac{1}{3}}\bar{\sigma}_i^{\frac{2}{3}}}{\sum_{j=1}^{n}(d_j\bar{\mu}_j)^{\frac{1}{3}}\bar{\sigma}_j^{\frac{2}{3}}}\left(K - \sum_{j=1}^{n}\bar{\mu}_j\right) \qquad i = 1, \ldots, n. \qquad (2.27)$$

and the optimal lead time for this configuration is given by

$$\frac{1}{\alpha_i^*} = \frac{k_{0.95}^2}{2}\left(\frac{\bar{\sigma}_i}{d_i\bar{\mu}_i}\right)^{\frac{2}{3}}\left(\frac{\sum_{j=1}^{n}(d_j\bar{\mu}_j)^{\frac{1}{3}}\bar{\sigma}_j^{\frac{2}{3}}}{K - \sum_{j=1}^{n}\bar{\mu}_j}\right)^2 \qquad i = 1, \ldots, n \qquad (2.28)$$

Proof: See the appendix. □

Looking at equation (2.27), we see that in the optimal capacity assignment, the capacity at each stage is at least equal to the mean production rate at the stage. In order for the entire system to be stable, the total capacity K must be greater that the total mean induced demand at all stages in the system $\sum_i \bar{\mu}_i$. The excess capacity $K - \sum_{i=1}^{n}\bar{\mu}_i$ is distributed among the n stages and the proportionality factor for stage i is

$$\frac{d_i^{\frac{1}{3}}\bar{\mu}_i^{\frac{1}{3}}\bar{\sigma}_i^{\frac{2}{3}}}{\sum_{j=1}^{n} d_j^{\frac{1}{3}}\bar{\mu}_j^{\frac{1}{3}}\bar{\sigma}_j^{\frac{2}{3}}} \qquad (2.29)$$

We can compare this result with the results of Kleinrock [Klei64] and Wein [Wein89]. In the Kleinrock case, the system is a Jackson network with n nodes and the effective

arrival rate vector is $(\bar{\mu}_1, \ldots, \bar{\mu}_n)^T$. The optimal capacity assignment that minimizes the expected equilibrium number of customers in the system is

$$C_i = \bar{\mu}_i + \frac{\sqrt{d_i \bar{\mu}_i}}{\sum_{j=1}^{n} \sqrt{d_j \bar{\mu}_j}} (K - \sum_{j=1}^{n} \bar{\mu}_j)$$

Wein looked at the capacity allocation for a generalized Jackson network which can be analyzed using a Brownian approximation. When the probability density function for the Brownian network has a product form, he showed that the optimal capacity assignment that minimizes the expected number of customers is given by

$$\bar{\mu}_i + \frac{\bar{\sigma}_i \sqrt{d_i}}{\sum_{j=1}^{n} \bar{\sigma}_j \sqrt{d_j}} \left(K - \sum_{j=1}^{n} \bar{\mu}_j \right)$$

The $\bar{\sigma}_i$'s in the Brownian network formulation are not the same as in our network, but they capture different sources of variability. (For further details, refer to Wein [Wein89].) This basically says that in the optimal configuration, the more variable stations get compensated by receiving more capacity.

Let us take a closer look at the proportionality factor in this model as stated in equation (2.29).

$$\frac{d_i^{\frac{1}{3}} \bar{\mu}_i^{\frac{1}{3}} \bar{\sigma}_i^{\frac{2}{3}}}{\sum_{j=1}^{n} d_j^{\frac{1}{3}} \bar{\mu}_j^{\frac{1}{3}} \bar{\sigma}_j^{\frac{2}{3}}}$$

This can be rewritten in terms of the mean and the coefficient of variation $\frac{\bar{\sigma}_i}{\bar{\mu}_i}$ of the induced demand at each stage as

$$\frac{d_i^{\frac{1}{3}} \bar{\mu}_i (CV_i)^{\frac{2}{3}}}{\sum_{j=1}^{n} d_j^{\frac{1}{3}} \bar{\mu}_j (CV_j)^{\frac{2}{3}}}$$

where $CV_i = \frac{\bar{\sigma}_i}{\bar{\mu}_i}$ which is the coefficient of variation of the induced demand for stage i, $i = 1, \ldots, n$. This says that in the optimal assignment, the portion of the excess capacity assigned to stage i is proportional to the mean production rate $\bar{\mu}_i$, the two-thirds power of the coefficient of variation CV_i and the one-third power of the cost of inventory at the stage d_i. For stages where the mean production rate is higher, a

greater portion of the excess capacity is assigned to the stage. If the coefficient of variation of the induced demand at a particular stage is higher, the portion of the excess capacity assigned to that stage is higher. In this model, the optimal capacity measures the variability in terms of the coefficient of variation. For stages where the cost of inventory is higher, the fraction of the excess capacity assigned is greater to allow for lower levels of base stock at those stages.

The lead time for each stage is known once the amount of capacity allocated is known. From equation (2.28), we see that the optimal lead time for stage i is given by

$$\frac{1}{\alpha_i^*} = \frac{k_{0.95}^2}{2}\left(\frac{\sum_{j=1}^{n}(d_j\bar{\mu}_j)^{\frac{1}{3}}\bar{\sigma}_j^{\frac{2}{3}}}{K-\sum_{j=1}^{n}\bar{\mu}_j}\right)^2\left(\frac{\bar{\sigma}_i}{d_i\bar{\mu}_i}\right)^{\frac{2}{3}}$$

For a given network, the expression $\frac{k_{0.95}^2}{2}\left(\frac{\sum_{j=1}^{n}(d_j\bar{\mu}_j)^{\frac{1}{3}}\bar{\sigma}_j^{\frac{2}{3}}}{K-\sum_{j=1}^{n}\bar{\mu}_j}\right)^2$ is constant and so we have the lead time for stage i

$$\frac{1}{\alpha_i^*} \propto \left(\frac{\bar{\sigma}_i}{d_i\bar{\mu}_i}\right)^{\frac{2}{3}}$$

This can be rewritten as

$$\frac{1}{\alpha_i^*} \propto \left(\frac{CV_i}{d_i}\right)^{\frac{2}{3}}$$

In other words, as the coefficient of variation of the demand for stage i increases, the lead time for stage i would increase. As d_i, the cost of inventory at stage i increases, the lead time at the stage $\frac{1}{\alpha_i^*}$ decreases. In the next section, we look at an example as well as tradeoff curves of how the base stock in the system and the lead times vary as the total capacity in the system is varied.

In comparing these results with Kleinrock's expression and Wein's expression for the optimal capacity allocation for queueing networks, a few remarks would be appropriate at this point. First, the system that we have analyzed here is a *make-to-stock* system as opposed to the *make-to-order* systems that we have seen for queueing networks. The make-to-order queueing networks that we described earlier have a product

form, but this is not true of the make-to-stock queueing networks. Second, the model that we present in this paper is tactical in nature, whereas the models for queueing networks are more detailed descriptions of the system behavior at the operational level.

There is another point that needs to be mentioned here. In all of the analysis so far, we have assumed that there is complete buffering between the stages. But is this a valid assumption to make? Or is there another method of releasing inventory from stage to stage that would reduce the total base stock in the system beyond the levels for this method. Two points need to be mentioned here. First, this method of analysing a multistage system is consistent with the role that safety stocks play. In earlier work (see Graves [Gra88b]), it has been argued that safety stocks play a decoupling role between stages. This decoupling function enables us to look at each stage in isolation and use the analysis for a single stage that was discussed earlier.Second, if there were a better method of releasing safety stock from one stage into the WIP inventory of a successive stage, this might lead to nonstationarity that would make the analysis extremely difficult.

2.5 Tradeoff Curves: An Example

In this section, we look at an example of capacity assignment for a multistage network. We also provide tradeoff curves that show how the base stock and lead times vary as a function of the total capacity K.

The network that we consider in this example is shown in Figure 2.3. There are eight stages in this network and the relationship between the different stages is

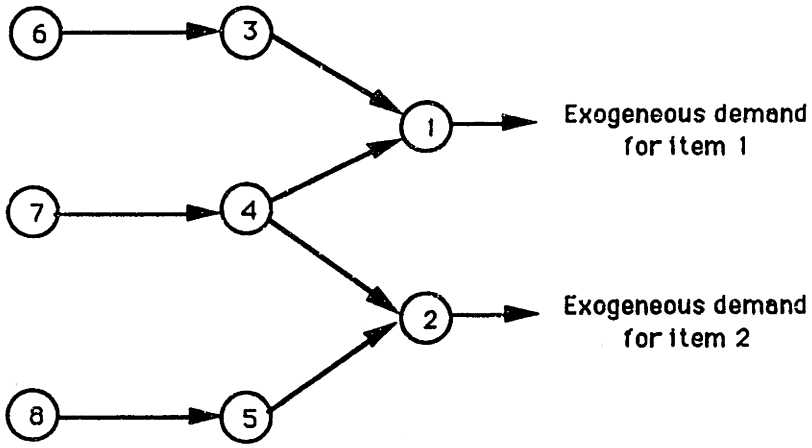

Figure 2.3: A multistage network.

indicated by the matrix A.

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

The exogeneous demand occurs only at stages 1 and 2. The mean and standard deviation of the demand rate for item 1 are 200 and 30 respectively, and the corresponding numbers for item 2 are 200 and 40. For now, let us also assume that the demands are independent. Therefore, the vector $\mu_{n\times 1}$ in this example is $(200\ 200\ 0\ 0\ 0\ 0\ 0\ 0)^T$ and the vector $\Gamma_{n\times n}$ is given by $diag(30\ 40\ 0\ 0\ 0\ 0\ 0\ 0)$. The drift vector $\bar{\mu} = [I - A]^{-1}\mu$ is equal to (200 200 200 400 200 200 400 200). In a similar manner, the dispersion matrix $\bar{\Gamma} = [I - A]^{-1}\Gamma_{n\times n}$ can be computed. The vector of $\bar{\sigma}_i$'s (see Theorem 2) is (30 40 30 50 40 30 50 40). Furthermore, in this example, $d_i = 1$ for $i = 1, \ldots, 8$.

For the problem to be feasible, there must be at least $\sum_{i=1}^{n} \bar{\mu}_i$ units of capacity

available, so $K \geq 2000$. Once the optimal capacity allocation and lead times are determined from Theorem 3, the WIP inventory and safety stocks can be determined. In Figure 2.4, we provide a tradeoff curve which shows how the WIP inventory varies as the total capacity in the system changes. As we expect, when the total capacity K increases, the total WIP inventory in the system decreases. Figures 2.5 and 2.6 show similar tradeoff curves for the total safety stock and the total base stock. If we compare the WIP inventory and safety stock tradeoff curves in Figures 2.4 and 2.5 respectively, we find that the safety stock is about one-tenth of the WIP on average. In this example, the WIP is the dominant component of the base stock and this was the rationale behind us choosing only the WIP component in the objective function of the modified problem (P'). We hasten to add that in this example, the coefficients of variation of the demands for the two items are low, but it can be argued that this is true of many manufacturing systems. Since the safety stock is a small component of the base stock in this example, the tradeoff curves for the WIP inventory and the base stock, in Figures 2.4 and 2.6 respectively, are quite similar.

In Figure 2.7, we look at the tradeoff curve of the total lead time versus the total capacity in the system K. In this example, the total lead time is defined as

$$\max\left(\frac{1}{\alpha_1}+\frac{1}{\alpha_3}+\frac{1}{\alpha_6}, \frac{1}{\alpha_1}+\frac{1}{\alpha_4}+\frac{1}{\alpha_7}, \frac{1}{\alpha_2}+\frac{1}{\alpha_4}+\frac{1}{\alpha_7}, \frac{1}{\alpha_2}+\frac{1}{\alpha_5}+\frac{1}{\alpha_8}\right)$$

In essence, this takes each of the four paths in the network (starting from raw material and terminating at the end-item inventory for items 1 or 2) 1–3–6, 1–4–7, 2–4–7 and 2–5–8 , and sees which one among these has the longest lead time. Another way of looking at this would be the following: Suppose we tag all items at stage 6, 7 and 8 when they enter the work-in-process inventory. On average, how much time would elapse before these items appeared as components in items in the end-item inventory of stages 1 or 2 and the longest of these times is chosen as the total lead time.

For this curve, we see that as the total available capacity increases, the total lead time decreases. When K is low, this lead time could be as high as 40 time units,

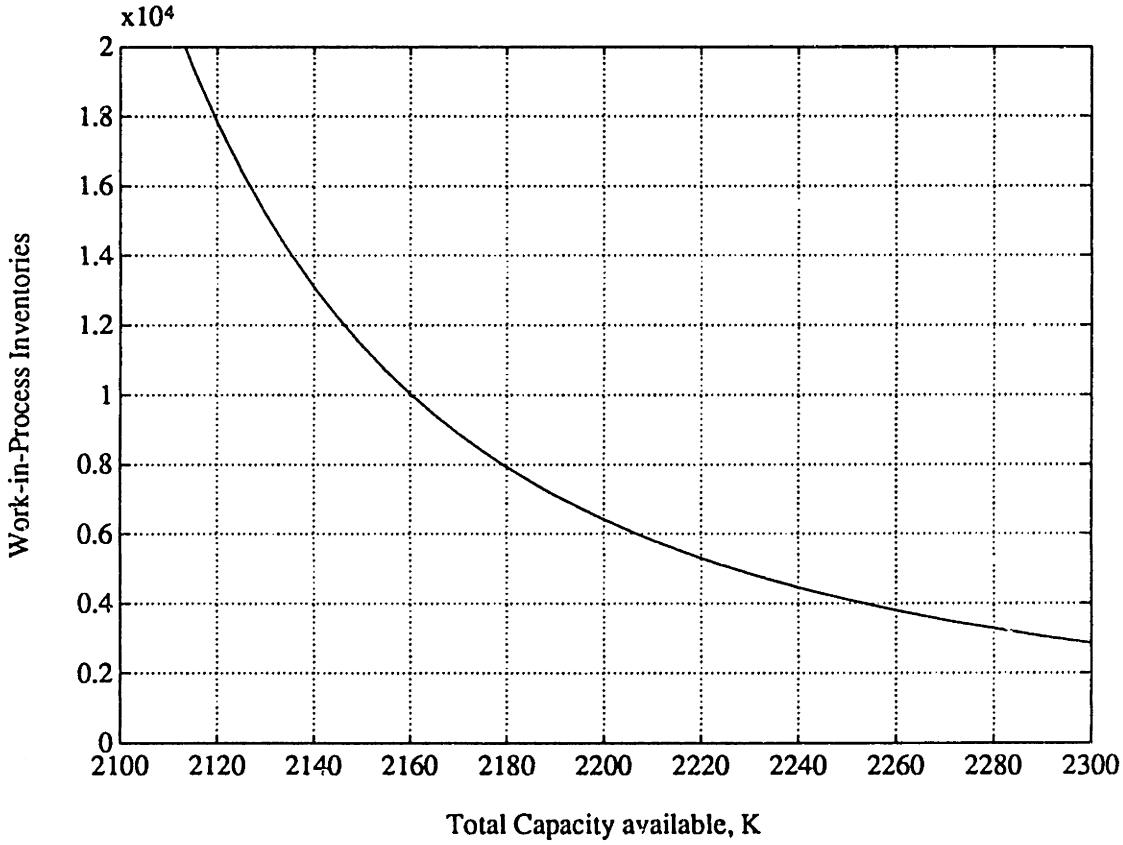

Figure 2.4: Tradeoff curve of total WIP inventory versus total capacity K

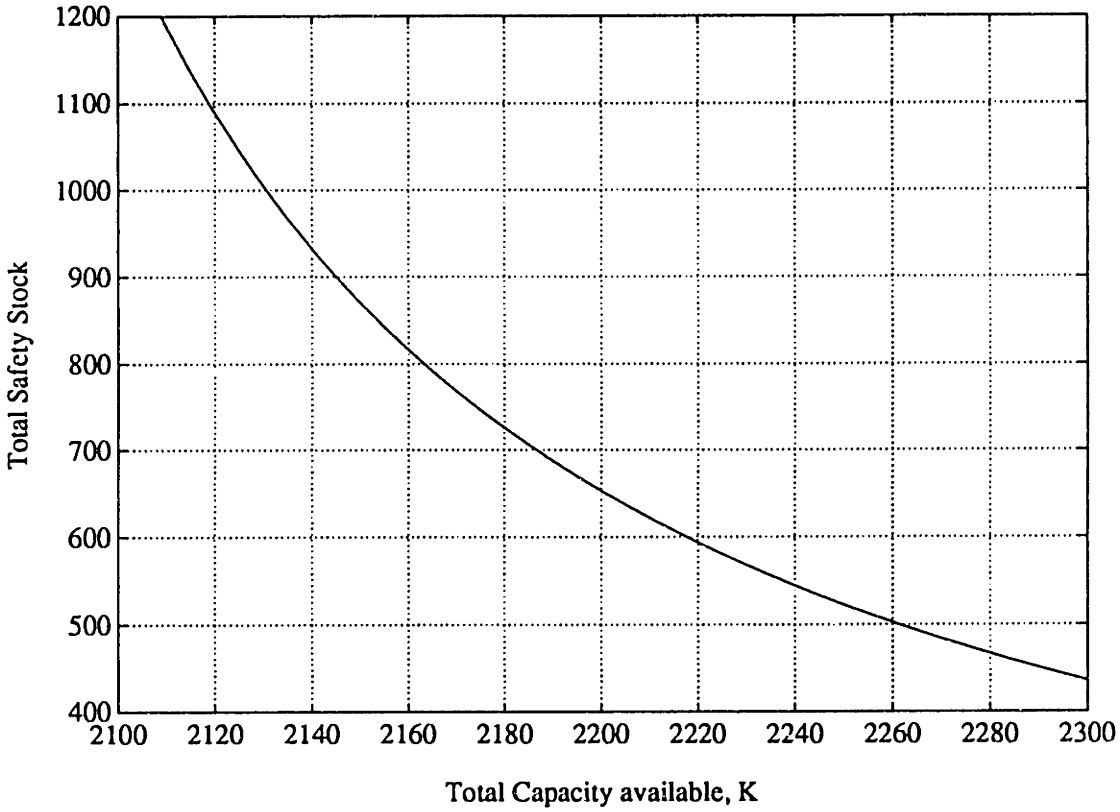

Figure 2.5: Tradeoff curve of total safety stock versus total capacity K

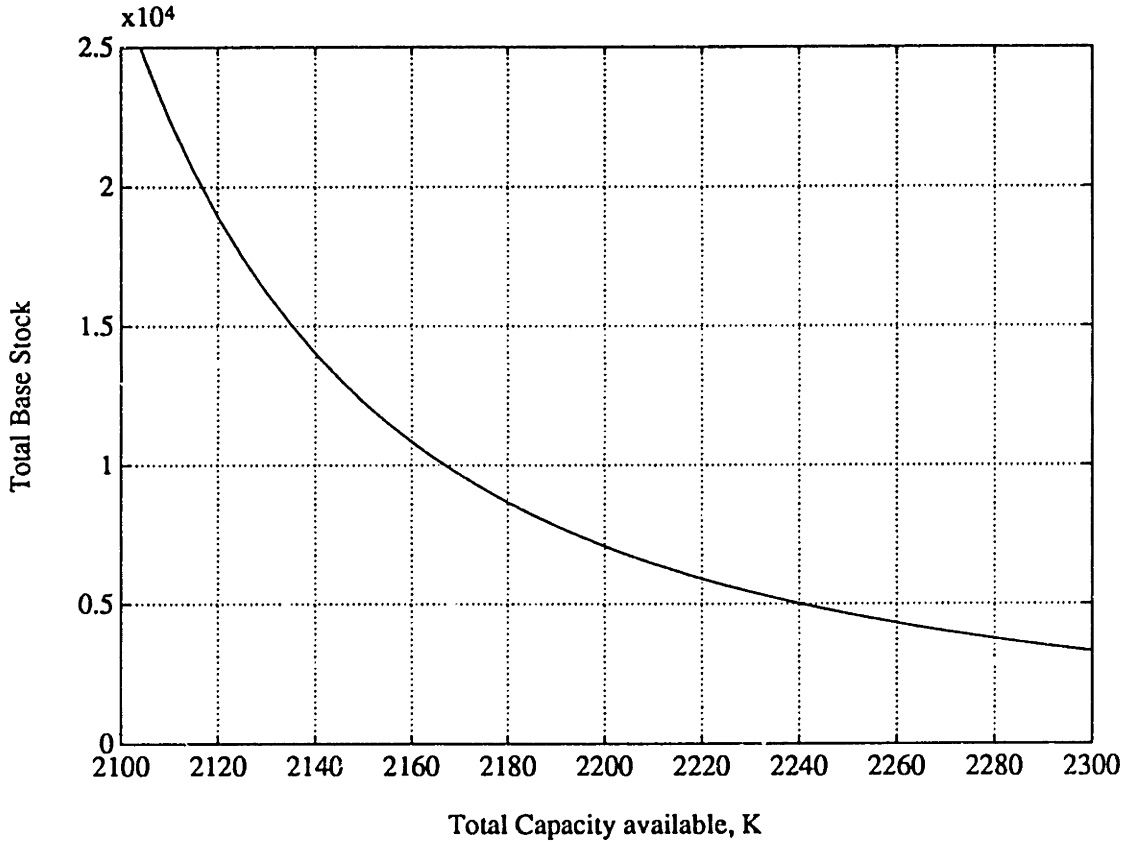

Figure 2.6: Tradeoff curve of total base stock versus total capacity K

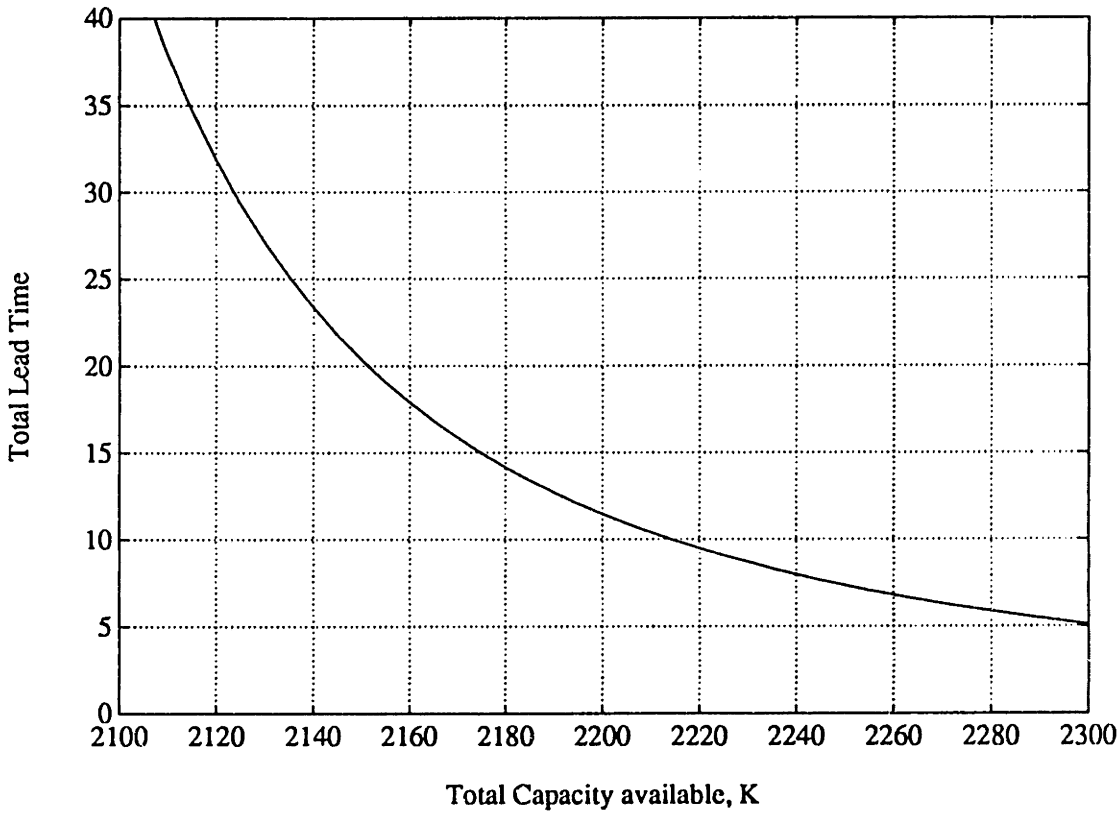

Figure 2.7: Tradeoff curve of total lead time versus total capacity K

which says that as the utilization at the different stages gets very high, so do the lead times.

In the previous section, we looked at a modified version of the optimization problem and found the optimal capacity allocation and lead times. The reason we did this was that we were able to obtain closed form expressions for the optimal decision variables and we argued that there was a good intuitive basis for looking at the modified problem. But we have not yet really answered the question: How do the optimal values for the modified problem compare with those for the original problem? In other words, is this optimum anywhere near the optimum for the problem that we started out looking at?

In Figure 2.8, we have compared the total base stock in the system for the original problem and the modified problem. The optimal solution for the original problem was obtained by writing out the Kuhn-Tucker conditions for this problem, and using a numerical approach to find the optimal decision variables that satisfy these equations. As we can see from the figure, the two curves are almost identical. In the calculations that were performed, the difference between the levels of total base stock for the two methods did not exceed one unit of base stock.

In Figures 2.4–2.7, we had assumed that the demand processes for items 1 and 2 are independent. Suppose the demand processes for items 1 and 2 are correlated. What is the effect of correlation between the demands in the network? When the correlation between the demands is negative, the variance of the demand for items 4 and 7 is lower and hence the base stock for each of these stages would be lower. On the other hand, when the correlation is positive, the variance for these items increases so the base stock increases at each of these stages. In Figures 2.9 and 2.10, we provide tradeoff curves for the total base stock and the total lead time as a function of the coefficient of correlation between the two demands. We see that as the correlation increases, both the total base stock and the total lead time increase. In the case of both these curves, the value of the total capacity is fixed at $K = 2250$.

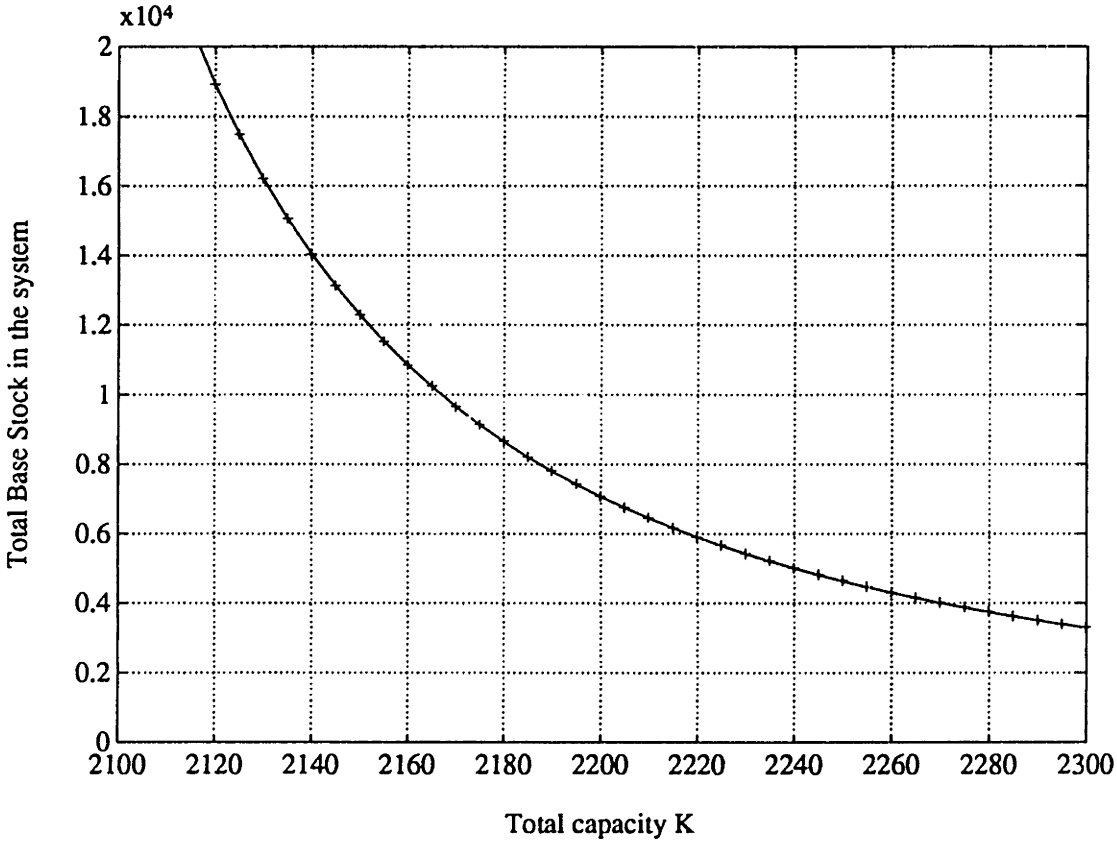

Figure 2.8: Total base stock for the heuristic and optimal capacity allocation.

---------- Heuristic capacity allocation.
++++++++++ Optimal capacity allocation.

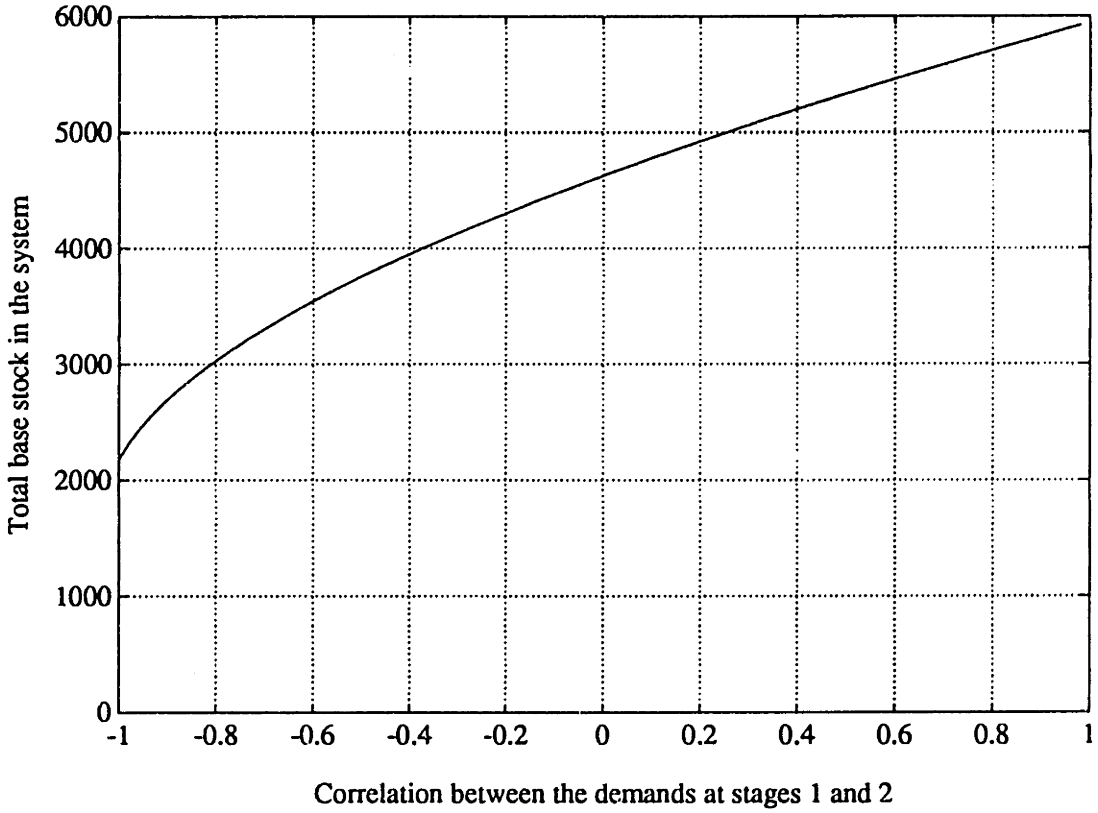

Figure 2.9: Total base stock versus coefficient of correlation between the demands

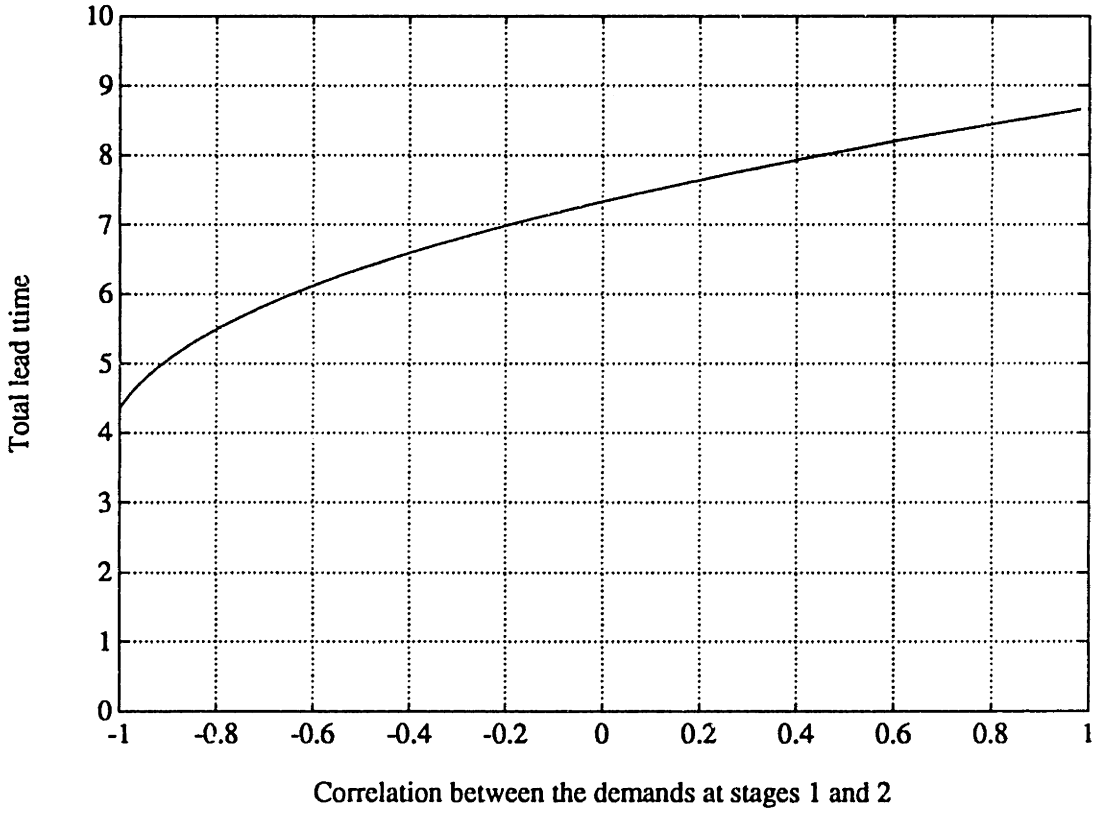

Figure 2.10: Total lead time versus coefficient of correlation between the demands

This completes our discussion of the example. In the next section, we summarize the work in this chapter. The proofs leading to the results are presented in the appendix.

2.6 Concluding Remarks

In this chapter, we have examined the problem of capacity allocation for a multistage network in a manufacturing system. We developed a multistage model and formulated the optimization problem. We then obtained expressions for the optimal capacity allocation and the corresponding lead time at each stage. Later, we showed how these could be used to provide tradeoff curves for the work-in-process inventories, safety stocks, base stocks and lead times.

Appendix

In this section, we derive the steady state results for the linear system of stochastic differential equations in equation (2.23). (See Karatzas and Shreve [Kara88].) Rewriting this system of equations, we have

$$d\bar{X}_t = \left(-[\alpha]_{dg}\bar{X}_t + \bar{\mu}\right)dt + \bar{\Gamma}d\bar{W}_t \qquad X_0 = 0 \tag{2.30}$$

In addition, we have imposed the condition that $X_0 = 0$. Let us define an $n \times n$ matrix Z_t by the following expression

$$Z_t = exp\left[-[\alpha]_{dg}t\right]$$

$$= \begin{pmatrix} e^{-\alpha_1 t} & & 0 \\ & \ddots & \\ 0 & & e^{-\alpha_n t} \end{pmatrix} \tag{2.31}$$

Lemma 4 *The solution to equation (2.30) is given by*

$$\bar{X}_t = Z_t\left[\int_0^t Z_s^{-1}\bar{\mu}ds + \int_0^t Z_s^{-1}\bar{\Gamma}d\bar{W}_s\right] \qquad 0 \leq t < \infty \tag{2.32}$$

Proof: In equation (2.32), Z_t is an $n \times n$ matrix, $\bar{\mu}$ and $\bar{W}_t$ are $n \times 1$ vectors. The first term on the RHS of equation (2.32) is

$$[Z_t]\int_0^t [Z_s^{-1}]_{n\times n}[\bar{\mu}]_{n\times 1}ds$$

and the second term is

$$[Z_t]\int_0^t [Z_s^{-1}]_{n\times n}[\bar{\Gamma}]_{n\times n}[d\bar{W}_t]_{n\times 1}$$

Differentiating both sides of equation (2.32) with respect to t and using the fact that $\dot{Z}_t = -[\alpha]_{dg}Z_t$, we obtain equation (2.30). □

We obtained an expression for $\bar{X}_t$ for all $0 \leq t < \infty$. Now we can obtain the mean and the variance of $\bar{X}_t$ in steady state.

Theorem 5 *The mean of the process in equation (2.30) in steady state is given by*

$$\begin{aligned} E\left[\bar{X}_t\right] &= \bar{\mu} \\ &= (I - A)^{-1}\mu \end{aligned} \tag{2.33}$$

Proof: The solution to equation (2.30) is the expression on the RHS of equation (2.32). Taking the expected value of both sides in equation (2.32), we get

$$E\left[\bar{X}_t\right] = Z_t\left[\int_0^t Z_s^{-1}\bar{\mu}ds\right] \tag{2.34}$$

since $E\left[\int_0^t Z_s^{-1}\bar{\Gamma}d\bar{W}_s\right] = 0$ for all t. Differentiating both sides of equation (2.34) with respect to t, we get

$$\begin{aligned} \frac{d}{dt}\left(E\left[\bar{X}_t\right]\right) &= \dot{Z}_t\left[\int_0^t Z_s^{-1}\bar{\mu}ds\right] + Z_t\left[Z_t^{-1}\bar{\mu}\right] \\ &= -[\alpha]_{dg}E\left[\bar{X}_t\right] + \bar{\mu} \qquad \text{since } \dot{Z}_t = -[\alpha]_{dg}Z_t \end{aligned}$$

In steady state, $\frac{d}{dt}\left(E\left[\bar{X}_t\right]\right) = 0$. Therefore, the result follows. □

Theorem 6 *The variance of the process in equation (2.30) in steady state satisfies the equation*

$$[\alpha]_{dg}\,Var\left(\bar{X}_t\right) + Var\left(\bar{X}_t\right)[\alpha]_{dg} = \bar{\Gamma}\bar{\Gamma}^T \tag{2.35}$$

Proof: The proof of this theorem is similar in spirit to the previous one. Taking the variance of both sides in equation (2.32), we get

$$Var(\bar{X}_t) = Z_t\left[\int_0^t Z_s^{-1}\bar{\Gamma}\bar{\Gamma}^T Z_s^{-1}ds\right]Z_t$$

Differentiating both sides of the above equation with respect to t, we get

$$\begin{aligned} \frac{d}{dt}\left(Var\left[\bar{X}_t\right]\right) &= \dot{Z}_t\left[\int_0^t Z_s^{-1}\bar{\Gamma}\bar{\Gamma}^T Z_s^{-1}ds\right]Z_t + Z_t\left[\int_0^t Z_s^{-1}\bar{\Gamma}\bar{\Gamma}^T Z_s^{-1}ds\right]\dot{Z}_t \\ &\qquad Z_t\left[Z_t^{-1}\bar{\Gamma}\bar{\Gamma}^T Z_t^{-1}\right]Z_t \\ &= -[\alpha]_{dg}\,Var\left[\bar{X}_t\right] - Var\left[\bar{X}\right][\alpha]_{dg} + \bar{\Gamma}\bar{\Gamma}^T \end{aligned}$$

In steady state, $\frac{d}{dt}\left(Var\left[\bar{X}_t\right]\right) = 0$. Therefore, the result follows. □

Theorem 7 *The optimal capacity allocation for the problem (P') is given by*

$$C_i^* = \bar{\mu}_i + \frac{(d_i\bar{\mu}_i)^{\frac{1}{3}}\bar{\sigma}_i^{\frac{2}{3}}}{\sum_{j=1}^{n}(d_j\bar{\mu}_j)^{\frac{1}{3}}\bar{\sigma}_j^{\frac{2}{3}}}\left(K - \sum_{j=1}^{n}\bar{\mu}_j\right) \qquad i = 1,\ldots,n.$$

and the optimal lead time for this configuration is given by

$$\frac{1}{\alpha_i^*} = \frac{k_{0.95}^2}{2}\left(\frac{\bar{\sigma}_i}{d_i\bar{\mu}_i}\right)^{\frac{2}{3}}\left(\frac{\sum_{j=1}^{n}(d_j\bar{\mu}_j)^{\frac{1}{3}}\bar{\sigma}_j^{\frac{2}{3}}}{K - \sum_{j=1}^{n}\bar{\mu}_j}\right)^2 \qquad i = 1,\ldots,n$$

Proof: Let τ_i, $i = 1,\ldots,n$ be the Lagrange multipliers corresponding to the constraints $(P'1)$ and let β be the Lagrange multiplier corresponding to the constraint $(P'2)$. The KKT conditions are

$$-\frac{d_i\bar{\mu}_i}{\alpha_i^2} + \frac{1.645}{2\sqrt{2}}\frac{\bar{\sigma}_i}{\sqrt{\alpha_i}}\tau_i = 0 \qquad i = 1,\ldots,n \tag{2.36}$$

$$-\tau_i + \beta = 0 \tag{2.37}$$

The constraints $\sum_i C_i = K$ and $C_i = \bar{\mu}_i + k_{0.95}\frac{\bar{\sigma}_i}{\sqrt{2}}\sqrt{\alpha_i}$ must be satisfied. The value of β that satisfies these constraints is given by

$$\beta = \left[\frac{k_{0.95}^{\frac{2}{3}}\sum_{j=1}^{n}(d_j\bar{\mu}_j)^{\frac{1}{3}}\,\bar{\sigma}_i^{\frac{2}{3}}}{(K - \sum_{j=1}^{n}\bar{\mu}_j)}\right]^3$$

Substituting this value of β in equation (2.36) (since $\beta = \tau_i$ for all i), we get the optimal values of α_i^* and C_i^*. □

Chapter 3

Capacity Allocation and Manufacturing Flexibility

3.1 Introduction

A Flexible Manufacturing System (FMS) can be broadly defined as a network or group of automated workstations that are connected by a material handling system which is used to transport parts between different workstations. These automated workstations are capable of producing jobs with diverse characteristics and in general, an FMS is appropriate for production systems where there is a high overall mix of products that are produced in small to medium size batches.

There are several reasons in favor of using an FMS. They are able to achieve efficiency and high utilization levels that one normally associates with a flow shop or an assembly line. At the same time, they provide the flexibility of a job shop. Since they are capable of producing a wide variety of job types, an FMS can be extremely useful when the product life is short or changes in the design are frequent. In this case, if an investment were to be made in a machine that is designed specifically for the manufacture of a product with a short life, it may be hard to recoup the initial investment over the life of the product. Besides, it is not clear what the residual value of the machine would be at the end of the product life given the obsolescence of the product.

While there are several advantages of using an FMS, there are some disadvantages that have somewhat restricted their use. One of them is the high initial investment necessary to set up the system. The costs of operating the system may be higher

as well. These may take the form of higher retooling costs, operating training costs[1] and maintenance costs due to the increased system complexity. There is an added dimension of complexity because of information management. Since an FMS is designed to handle diverse part types, information tracking and control policies for the system are much more difficult. The nonuniformity of job types and variability in the machine operation times makes job scheduling a much harder task.

There is a sizeable body of literature that deals with modeling the behavior of FMS. One approach is to model the FMS as a closed network of queues. In this approach, there are a fixed number of workstations and each workstation has one or more identical servers. Each station has a queue with a FCFS discipline and the service times are exponentially distributed. There is a fixed quantity of jobs in the system and the routing matrix is independent of the system state. It is possible to compute the queue length density function and other useful performance measures such as the utilization, mean queue lengths and throughput. This approach has been taken by several authors (see Gordon and Newell [Gord67], and Vinod and Solberg [Vin85]).

An alternative approach is to model the system as an open queueing network (see Buzacott and Shanthikumar [Buza80]). Stecke [Stec83] formulates a number of production planning problems for an FMS as nonlinear 0-1 mixed integer programs. Stecke and Solberg [Stec81] examine some loading and control policies using a simulation study. Schweitzer and Seidmann [Schw89] consider a problem where the processing rate at each station is a decision variable. For example, the tooling cost is a nonlinear function of the processing speed since the speed affects the tool wear. The objective in their approach is to choose the processing rates so as to minimize the total operating cost subject to a certain minimum throughput rate constraint.

There has also been a considerable amount of work done in the area of optimization in queueing networks. If one had to allocate capacity among the different workstations

[1] However, since these workstations are highly automated, the total labor costs might still be lower in an FMS.

and a certain performance measure is a function of this capacity allocation, how does one assign this capacity so that this measure is optimized? The square root formula of Kleinrock [Klei64] is perhaps the most well known. In this case, the objective is to minimize the average number of customers in the system in equilibrium subject to a linear constraint on the processing rates. More recent work includes that of Yao and Buzacott [Yao85], and Vinod and Solberg [Vin85]. Wein [Wein89] considers a Brownian network with a product form density and shows that the optimal capacity allocation has a form that is comparable to the square root formula.

In the discussion so far, we have discussed some of the modeling approaches for flexible manufacturing systems. The importance of these models lies in their ability to identify bottlenecks and indicate the operating strategies that would perform better in a given system. If there are two available technologies, how does one evaluate which one is more flexible? To answer this question, one needs a good measure for the flexibility of a system. This is difficult to do because there are many attributes in a system that make a total ordering almost impossible. In addition, in a multi-attributed system, it is very difficult to map the set of attributes into a set of rewards because there are many intangible entities that cannot be quantified.

While it is hard to define flexibility in a manner that encompasses all its attributes, nevertheless we can describe some of its intuitive properties. The main property (or axiom) is that if there are two technologies, one of them is said to be more flexible than the other if as the operating environment becomes more diverse, the first technology performs better than the second. (For a nice discussion, see de Groote [deGr90].) This increase in diversity may assume the form of an increase in the variability of the demand process or the product mix, an increase in the frequency of design changes or product volume etc. In some sense, the notion is related to the idea of the number of available choices. That is, how many alternatives does the present configuration afford? For example, suppose there are n machines at a station that can be used to perform a certain task. Let p_i be the processing capacity of machine i as a fraction of the total processing capacity at the station, $i = 1, \ldots, n$, and $\sum_{i=1}^{n} p_i = 1$. If

the utilization at the station is high, then on average, the fraction of work that is processed by machine i is p_i. A system where the largest fraction p_i is 0.9 is less flexible than one where the largest fraction is 0.5, all other things being the same. The system is more dependent on the machine that processes 90% of the jobs because it leaves the system vulnerable to breakdowns that may occur at that machine. One measure of flexibility that has been proposed is the negative of the system entropy, i.e., $-\sum_{i=1}^{n} p_i \ln p_i$. This measure is the same as the entropy function in thermodynamics. It was used by Shannon [Shan48] in communication theory and has since been used in many other fields (see Kumar [Kum87]). If we wish to maximize this function subject to the constraint $\sum_{i=1}^{n} p_i = 1$, the optimum solution is $p_i^* = \frac{1}{n}$ for all $i = 1, \ldots, n$. According to this measure, the most flexible system is the one where the workload is evenly distributed among the n machines.

In the discussion above, we have tried to illustrate how the flexibility is related to the availability of choices and described a measure that captures this effect. In other fields, there exist similar measures that try to capture this idea of flexibility (although they do not explicitly refer to it as such). Consider, for example, a market in which n firms compete. Let p_i be the market share of firm i, $i = 1, \ldots, n$. The Herfindahl-Hirshmann index (HHI) is defined as the sum of the squares of the market shares of the n firms, i.e.,

$$\text{HHI} = \sum_{i=1}^{n} p_i^2$$

where $\sum_{i=1}^{n} p_i = 1$. Basically, this measures the degree of competition in a given market with lower values corresponding to a higher levels of competition. (If one firm had a 100% share of the market, the value of the index is 1.) The HHI can take on values between $\frac{1}{n}$ and 1 and is minimum when all firms have an equal market share, i.e., $p_i = \frac{1}{n}$ for all $i = 1, \ldots, n$. In this example, the vector $(p_1, \ldots, p_n)$ corresponds to the choices available to the consumer. In fact, the U. S. Department of Justice uses this as one measure to decide whether a merger between two firms would significantly reduce the level of competition (See Rose [Rose89]).

In this chapter, we consider a problem of capacity allocation that is different from those that have been studied before. To illustrate the problem, let us describe it in its simplest form. Suppose there exists a workstation that consists of one or more parallel machines and there are two products, A and B, that are produced at this workstation. Should separate machines be dedicated for the production of item A and separate machines be dedicated for the production of item B? Or should there exist machines that are capable of processing both items?

Earlier, we had mentioned that the diversity in product characteristics and processing requirements in an FMS adds another level of complexity to the task of information management. The question that we consider is the following: Suppose there were two products being manufactured at a certain facility. Is it better to have separate lines for product A and B, or to merge the flows and have a common line that processes both items? We are examining the question solely from the operational standpoint. That is, either configuration does not entail a significantly higher cost in terms of the initial investment or the effort required to reach operational readiness. The broader question that we are trying to address is that of dedicated versus flexible factories. Which among these two alternatives would reduce the total amount of work-in-process inventory?

In this case, it is not at all clear that having a single facility that processes both items will reduce the amount of work-in-process inventory. Suppose the demand for product A is well known and stable. Then it may not be a good idea to make both products in a single facility. If the two streams are combined and the demand for product B is more variable, then the output stream for product A could become much more variable than it would have been if there were a dedicated line for each product. In addition, the total work-in-process inventory could be higher because of the instability that has been introduced in the flow stream of product A.

It is possible to model this scenario as a two server queue with two types of customers. In the situation where we have two dedicated machines, we can model this as two queues, each with a single server. Customers of type A join the first

queue and customers of type B join the second queue. (The customer cannot choose which queue to join — the queue that each customer enters is determined by the customer type.) In the case where we have a single facility that produces both items, we can model it as a single queue with two servers. Customers join the queue when they arrive and they are processed in a FCFS order. If one were to use this model to choose between the two alternatives using the average number of customers as a yardstick to measure performance, clearly the preferable alternative is to have a single queue. In the case when there are two queues, customers from one queue cannot join the other even if the server is idle. When there is a single queue, a server will never be idle if there is a customer waiting for service. Therefore, the weighted average waiting time for the two queues will be higher than that of a single queue with two servers.

Yet, as we had argued earlier, it is not evident why this alternative is better. Besides, if the demand processes are positively correlated, the benefits of merging the two streams would be lower. This is because if the customer arrivals for the two streams are positively correlated and the two queue configuration is maintained, when one server is busy, it is more likely that the other server is busy as well. We are not trying to be critical of queueing models here — all we are saying is that in this particular scenario, they may not provide the right alternatives.

The structure of this chapter is as follows: In section 3.2, we describe the basic model. In sections 3.3 and 3.4, we look at the two cases of dedicated machines and a single flexible machine. In section 3.4.1, we compare the two alternatives and try to provide some insights on why a given alternative is preferable under a certain set of conditions. In section 3.6, we show how the conditions change when one considers setup times for the single flexible machine.

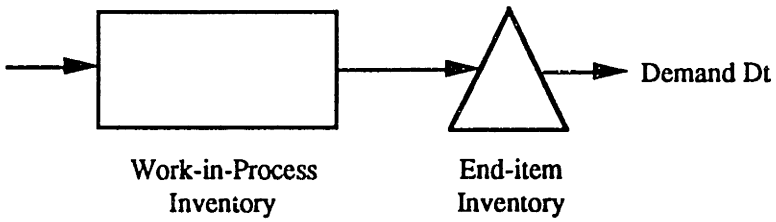

Figure 3.1: A single stage in a manufacturing system

3.2 The Model

The model that we use is one proposed by Graves [Gra88b]. We shall describe the model for one product, but the extension to the multi-product case is straightforward. Consider a single stage in a manufacturing system (See Figure 3.1). The notion of a stage, as we describe it, is fairly general. If the system is a factory or an assembly line, a stage could mean a single machine or a group of similar machines. On the other hand, if the system was a network of manufacturing plants in a region, an entire plant in this network could be modeled as a stage. The inventory at a stage is of two types: work-in-process inventory and end-item inventory. The end-item inventory, also known as safety stock, is used to satisfy demand and any excess demand is assumed to be backlogged. The demand in period t is D_t and this occurs at the beginning of the period. Any of the work-in-process inventory that has been processed completely becomes a part of the end-item inventory at the end of the period in which it is processed. Whenever a certain amount of inventory is taken from the safety stock buffer to meet the demand, an identical amount of *raw material* is added to the work-in-process inventory. Thus the total of the work-in-process inventory and end-item inventory is always kept constant.

Let I_t be the end-item inventory at the start of period t after the demand for period t has been satisfied. Let X_t be the work-in-process inventory at the start of period t after the amount of raw material R_t has been released into the system. (Note

that we have set $R_t = D_t$ for all t.) The flow balance equations for the work-in-process inventory are

$$X_t = X_{t-1} + R_t - P_{t-1} \tag{3.1}$$

Similarly, the flow balance equations for the end-item inventory are

$$I_t = I_{t-1} + P_{t-1} - D_t \tag{3.2}$$

Using the fact that $R_t = D_t$ and adding equations (3.1) and (3.2), we get

$$\begin{aligned} X_t + I_t &= X_{t-1} + I_{t-1} \qquad \text{for all } t. \\ &= B \text{ (constant)} \end{aligned}$$

In this model, the control rule for production is a linear function of the work-in-process inventory. The merits of this rule have been discussed extensively in Graves [Gra88b] and in Chapter 4 of this thesis. We shall state them here very briefly. First, the resulting system becomes linear and this simplifies the analysis considerably. Second, a rule of this form can be shown to be optimal when the production and inventory holding costs are quadratic (see Holt, Modigliani and Simon [Holt55]). Third, this rule is an asymptotic case of a more general class of rules proposed in Chapter 4. Let

$$P_t = \alpha X_t \tag{3.3}$$

In each time period, we process a fraction $\frac{1}{n}$ of the work-in-process inventory. Therefore, from equations (3.1) and (3.3), we get

$$\begin{aligned} X_t &= X_{t-1} + R_t - \alpha X_{t-1} \\ &= (1-\alpha)X_t + D_t \qquad\qquad \text{since } R_t = D_t. \\ &= \sum_{s=0}^{\infty} (1-\alpha)^s D_{t-s} \end{aligned} \tag{3.4}$$

assuming that the process has an infinite history. Using the fact that $\{D_t\}$ are independent and normally distributed with mean μ and variance σ^2, we get

$$E[X_t] = \frac{\mu}{\alpha} \quad \text{and} \quad Var[X_t] = \frac{\sigma^2}{2\alpha - \alpha^2}$$

We shall now consider the case where there is more than one item being produced at the stage. Suppose that there is a single machine in a job shop and a two product mix. When we refer to a single machine, we actually mean that there is a single cluster which may consist of one or more than one machine, all of the same type. Each of the items require processing at exactly one machine in this cluster, and any of the machines in the cluster could be used to process either item. After processing, an item enters its final product inventory and the items are assumed to be nonsubstitutable. In other words, the inventory for item 1 cannot be used to satisfy the demand for item 2, and vice versa. In this model, time is assumed to be discrete.

We assume that the demand for both items is satisfied at the beginning of each time period and any excess demand is backlogged. Let the demand for item 1 in period t be D_{1t}. We assume that $\{D_{1t}\}_t$ are independent and normally distributed random variables with

$$E[D_{1t}] = \mu_1 \quad \text{and} \quad Var[D_{1t}] = \sigma_1^2 \tag{3.5}$$

In a similar manner, let $\{D_{2t}\}_t$ be the demand for item 2 in period t. We assume that the demands for the two products are independent[2], so $\text{Cov}(D_{it}, D_{2t}) = 0$ for all t. $\{D_{2t}\}_t$ are independent and Gaussian random variables with

$$E[D_{2t}] = \mu_2 \quad \text{and} \quad Var[D_{2t}] = \sigma_2^2 \tag{3.6}$$

Consider a scenario where the demand for item 1 is relatively stable, whereas the demand for item 2 is more variable. The question that we would like to address is the following: Should we process both items on a single machine, or should we split the capacities between two machines and have dedicated lines for each product? There are tradeoffs that have to be considered when we make these decisions. There are advantages to processing both items on a single machine. The benefits of having both items on a single machine is that the resources are pooled and the amount of time that a job would require for processing would be lower.

[2]Later, we shall relax the assumption that the demands for the two items are independent in a single period. However, we would still require that the inter-temporal demands be independent, i.e. $\text{Cov}(D_{is}, D_{jt}) = 0 \; \forall \; s \neq t$ and $i, j \in \{1, 2\}$.

There are advantages to dedicating machines for individual products. First, since the demand for item 1 is relatively stable, we preserve the stability of the production process for item 1. As a result, we reduce the amount of work in process inventory, or safety stock that is required of item 1. If the cost of carrying inventory for item 1 is higher, the benefits would be even greater. If the demand for item 1 is relatively stable while the demand for item 2 is highly variable, if we process both items on a single machine, the production rate for item 1 would be more variable .

The issues that we try and address are the following: How does the overall machine flexibility affect the decision to merge flows, or dedicate lines? And how does the mean volume and the coefficient of variation for the two products affect this decision?

3.3 Dedicated versus Flexible Machines

3.3.1 Case I: A Single Flexible Machine

Suppose we decide to produce both items on a single machine. We do not assume any setup costs and do not concern ourselves with detailed issues like scheduling of the items on the machines, setup times etc.

Let X_{it} be the work-in-process inventory for item i, $i = 1, 2$, at the start of period t, and let I_{it} be the end product for item i at the start of period t. Let P_{it} be the amount of item i produced in period t, D_{it} be the demand for item i (which is satisified at the start of period t), and R_{it} be the amount of item i released into the system (as WIP) at the start of period t. (Note that $R_{it} = D_{it}$ for $i = 1, 2 \; \forall \; t$.)

The flow balance equations for the WIP and end product inventory for item i, $i = 1, 2$, are given below:

$$X_{it} = X_{i,t-1} + R_{it} - P_{i,t-1} \qquad i = 1, 2. \tag{3.7}$$

and

$$I_{it} = I_{i,t-1} + P_{i,t-1} - D_{it} \qquad i = 1, 2. \tag{3.8}$$

Suppose, as in the single product case, the amount that we release into the system in a period is equal to the demand for the product in that period. In other words, this is a pull type system. The inventory level for item i determines the amount of item i to be released into the system. Letting $R_{it} = D_{it}$, and adding equations (3.7) and (3.8), we get

$$\begin{aligned} X_{it} + I_{it} &= X_{i,t-1} + I_{i,t-1} \qquad i = 1, 2. \\ &= B_i \qquad \text{(Base Stock for item } i\text{)} \end{aligned} \tag{3.9}$$

Upto this point, the analysis is the same as in section 3.2. We would now like to specify a control rule for the production of item i.

$$P_{it} = \alpha X_{it} \qquad i = 1, 2. \tag{3.10}$$

According to this rule, the amount that we process in period t is a fraction α of the WIP inventory for that item in period t. Since a fraction α of the WIP inventory is processed in each period, on average, it will take $\frac{1}{\alpha}$ periods for inventory to be processed. Therefore, $\frac{1}{\alpha}$ is the lead time for the stage. By choosing a smaller value of α, the effect is that the production for the stage becomes more smooth. But smaller values of α also result in higher levels of WIP inventory and safety stock. Using equations (3.7) and (3.10) and the fact that $R_{it} = D_{it}$, we get

$$\begin{aligned} P_{it} &= \alpha D_{it} + (1 - \alpha) P_{i,t-1} \\ &= \alpha \sum_{s=0}^{\infty} (1 - \alpha)^s D_{i,t-s} \end{aligned} \tag{3.11}$$

Taking the expected value and the variance of both sides in equation (3.11) and assuming that the process has an infinite history, we get

$$E[P_{it}] = \mu_i \quad \text{and} \quad Var[P_{it}] = \frac{\alpha^2 \sigma_i^2}{2\alpha - \alpha^2} \qquad i = 1, 2 \tag{3.12}$$

In other words, this says that the mean production rate for item i must equal the mean demand μ_i. As σ_i^2, the variance of the demand for item i increases, the variance of production $Var[P_{it}]$ for item i also increases. However, as α decreases, the production becomes more smooth and this reduces the variance of production.

If $D_t = D_{1t} + D_{2t}$ is the total demand in period t, then D_t is normally distributed with mean $\mu := \mu_1 + \mu_2$ and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$. Let P_t be the total production of items 1 and 2 in period t. Since the time series of demand for the two items, $\{D_{1t}\}_t$ and $\{D_{2t}\}_t$, are independent, from equation (3.11), $\{P_{1t}\}_t$ and $\{P_{2t}\}_t$ are independent. From the definition of P_t,

$$P_t = P_{1t} + P_{2t} \tag{3.13}$$

Therefore, taking the expected value and variance of both sides in equation (3.13), we get

$$E[P_t] = E[P_{1t}] + E[P_{2t}] \tag{3.14}$$

$$= \mu_1 + \mu_2 \tag{3.15}$$

$$= \mu \qquad \text{by the definition of } \mu. \tag{3.16}$$

and

$$\begin{aligned} Var[P_t] &= Var[P_{1t}] + Var[P_{2t}] \\ &= \frac{\alpha^2\sigma_1^2}{2\alpha - \alpha^2} + \frac{\alpha^2\sigma_2^2}{2\alpha - \alpha^2} \qquad \text{since} \quad \mathrm{Cov}(P_{1t}, P_{2t}) = 0. \\ &= \frac{\alpha^2\sigma^2}{2\alpha - \alpha^2} \qquad \text{by the definition of } \sigma. \end{aligned} \tag{3.17}$$

Let C be the capacity of the machine and let us define the excess capacity by $\chi = C - \mu$. Let k_a be a factor that corresponds to a certain level of service. A factor of service k_a means that one is able to meet the demand for items 1 and 2 for $100a\%$ of the time. For example, $k_{0.95} = 1.645$ corresponds to the 95% service level. In other words, 95% of the time, the production rate dictated by the model is less than the processing capacity. Through the rest of this chapter, we shall denote this factor by k. We would like to set the value of α such that the probability that the production requirement exceeds the available capacity is no greater than 0.05. Since P_{it} is the sum of independent Gaussian random variables (see equation (3.11)), the production rate for item i as dictated by the model has a Gaussian distribution. Therefore, the mean $E[P_t]$ and the variance $Var[P_t]$ completely characterize the distribution of P_t.

From equation (3.17), we see that $Var[P_t]$ is a function of α. Clearly, we would like the value of the lead time α to be as high as possible. On the other hand, as α increases, the variance of the production increases. Since the production rule is linear, it is unbounded. Therefore, it is important that the production rate dictated by the model is consistent with the actual processing capacity at the stage. The idea is to choose the smallest value of n is chosen such that

$$E[P_t] + k\sqrt{Var[P_t]} = C$$

Since $E[P_t] = \mu$, we get

$$\begin{aligned} \chi &= k\sqrt{Var[P_t]} \\ &= \frac{k\alpha\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{2\alpha - \alpha^2}} \\ &= \frac{k\alpha\sigma}{\sqrt{2\alpha - \alpha^2}} \end{aligned} \tag{3.18}$$

We can now define the flexibility of a the machine in terms of a dimensionless constant. Let

$$\begin{aligned} F &= \text{Flexibility of the machine} \\ &:= \frac{\chi}{k\sigma} \end{aligned} \tag{3.19}$$

The flexibility is the ratio of the excess capacity over the mean to the standard deviation of the demand times the service factor. This is a measure of the ratio of the ability of the system to respond, χ, to the amount of the response required to maintain a certain level of service, $k\sigma$. Intuitively, this says that as the excess capacity increases, the system is able to respond to variability in the operating environment better and so the flexibility increases. When the standard deviation σ of the demand increases, the variability of the demand increases and so the flexibility would decrease. From equations (3.18) and (3.19), we get

$$F = \frac{\alpha}{\sqrt{2\alpha - \alpha^2}} \tag{3.20}$$

Since $0 \leq \alpha \leq 1$, the flexibility of the machine is a number between 0 and 1. It follows from equation (3.20) that $\alpha = \frac{2F^2}{1+F^2}$. The plot in Figure 3.2 shows how the lead time varies with machine flexibility. As the machine flexibility decreases, the lead time $\frac{1}{\alpha}$ increases dramatically. When the flexibility of the machine is 1, the leadtime is 1. But why can't the flexibility of the machine be greater than one? If the flexibility of the machine is greater than 1, the lead time is less than 1. Although it is possible that χ could be greater than $k\sigma$, in terms of this model, it means that the choice of the duration of a time period is inappropriate. One should then choose a time period of a shorter duration so that any WIP entering the system in a given period leaves the system at least one period later. Since $P_{it} = \alpha X_{it}$ from equation (3.10), we have that

$$E[X_{it}] = \frac{\mu_i}{\alpha} \quad \text{and} \quad Var[X_{it}] = \frac{1}{\alpha^2} Var[P_{it}]$$
$$= \frac{\sigma_i^2}{2\alpha - \alpha^2} \qquad i = 1, 2$$

Upto this point, we know $E[X_{it}]$ and $Var[X_{it}]$, the mean and the variance of the WIP inventory level and this is sufficient to characterize the distribution of X_{it} since it is Gaussian. We also know that I_{it} has a Gaussian distribution and $Var[I_{it}] = Var[X_{it}]$ since $X_{it} + I_{it} = B_i$ for all t, $i = 1, 2$. So the only thing that remains is to set the value of $E[I_{it}]$, the mean level of safety stock for item i. The idea here is to set the mean level so that a certain percentage of the time, say 95%, of the time, the demand for item i can be met from the safety stock. Another way of saying this is

$$P[I_{it} > 0] = 0.95$$

The mean level of safety stock for item i that satisfies this is

$$\begin{aligned} E[I_{it}] &= k\sqrt{Var[I_{it}]} \\ &= k\frac{\sigma}{\sqrt{2\alpha - \alpha^2}} \end{aligned} \tag{3.21}$$

The base stock for item i is given by

$$B_i = E[X_{it}] + E[I_{it}]$$

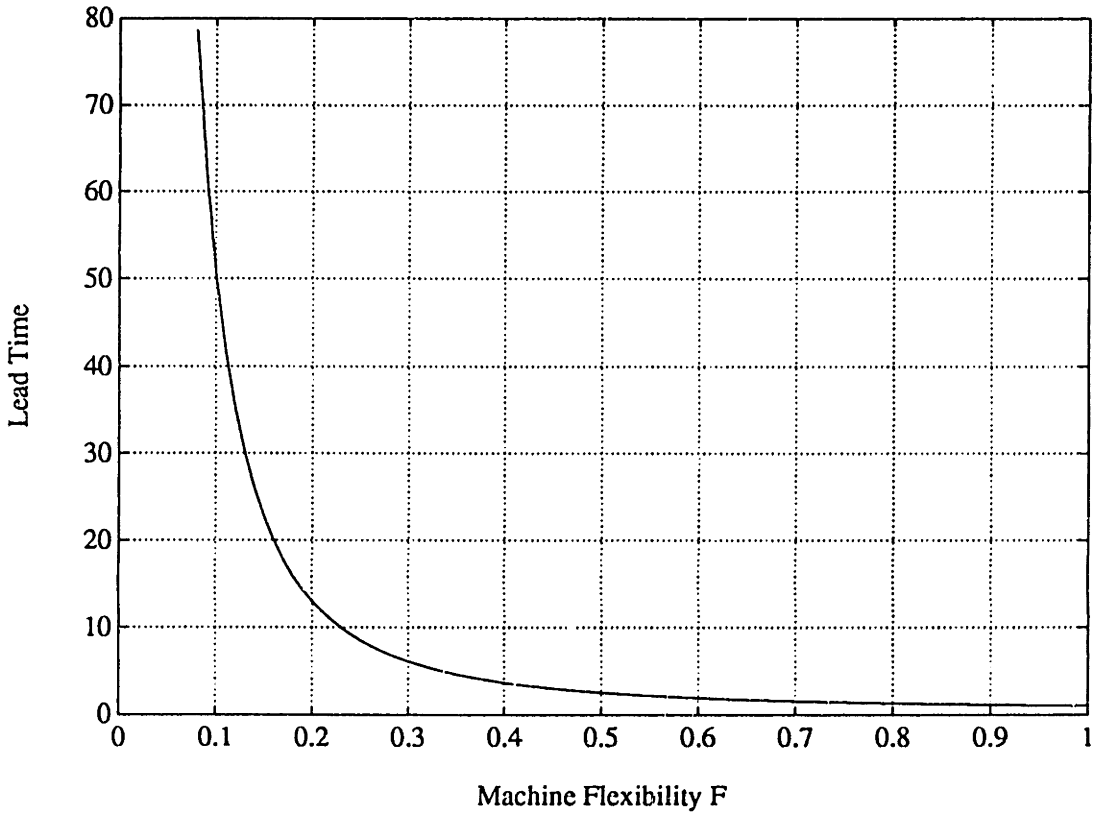

Figure 3.2: Plot of Leadtime vs. Machine Flexibility

$$
\begin{aligned}
&= \frac{\mu_i}{\alpha} + k\frac{\sigma}{\sqrt{2\alpha - \alpha^2}} \\
&= \frac{1}{\alpha}\left[\mu_i + k\frac{\alpha}{\sqrt{2\alpha - \alpha^2}}\sigma\right] \\
&= \frac{1}{\alpha}\left[\mu_i + kF\sigma\right] \qquad i = 1,2 \quad \text{from equation 3.20}
\end{aligned}
$$

Therefore, the total base stock is given by

$$
\begin{aligned}
B &= B_1 + B_2 \\
&= \frac{1}{\alpha}\left[\sum_i \mu_i + kF\sum_i \sigma_i\right] \\
&= \frac{1+F^2}{2F^2}\left[\sum_i \mu_i + kF\sum_i \sigma_i\right] \qquad (3.22)
\end{aligned}
$$

3.3.2 Case II: Splitting Capacity between Two Machines

In the previous subsection, we had looked at the case where both the items were processed on a single machine and both items have the same lead times. In this case, we had obtained the expression for the total base stock at the stage. In this subsection, we model the case where there are dedicated machines for each product.

Suppose we decide to dedicate machines according to product. The question then remains: How should the capacity be distributed between the two machines? If χ is the total excess capacity on both machines, let χ_1 be the excess capacity assigned to machine 1 and χ_2 be the capacity assigned to machine 2, i.e, the capacities on machines 1 and 2 are $\mu_1 + \chi_1$ and $\mu_2 + \chi_2$ respectively. The sum of the two excess capacities must equal the total excess capacity. Therefore,

$$\chi_1 + \chi_2 = \chi$$

Rewriting this in terms of machine flexibility , we have

$$\sigma_1\frac{\chi_1}{k\sigma_1} + \sigma_2\frac{\chi_2}{k\sigma_2} = \sigma\frac{\chi}{k\sigma}$$

Using the definition of machine flexibility (for the definition, see equation (3.19)), we get

$$\sigma_1 F_1 + \sigma_2 F_2 = \sigma F \qquad (3.23)$$

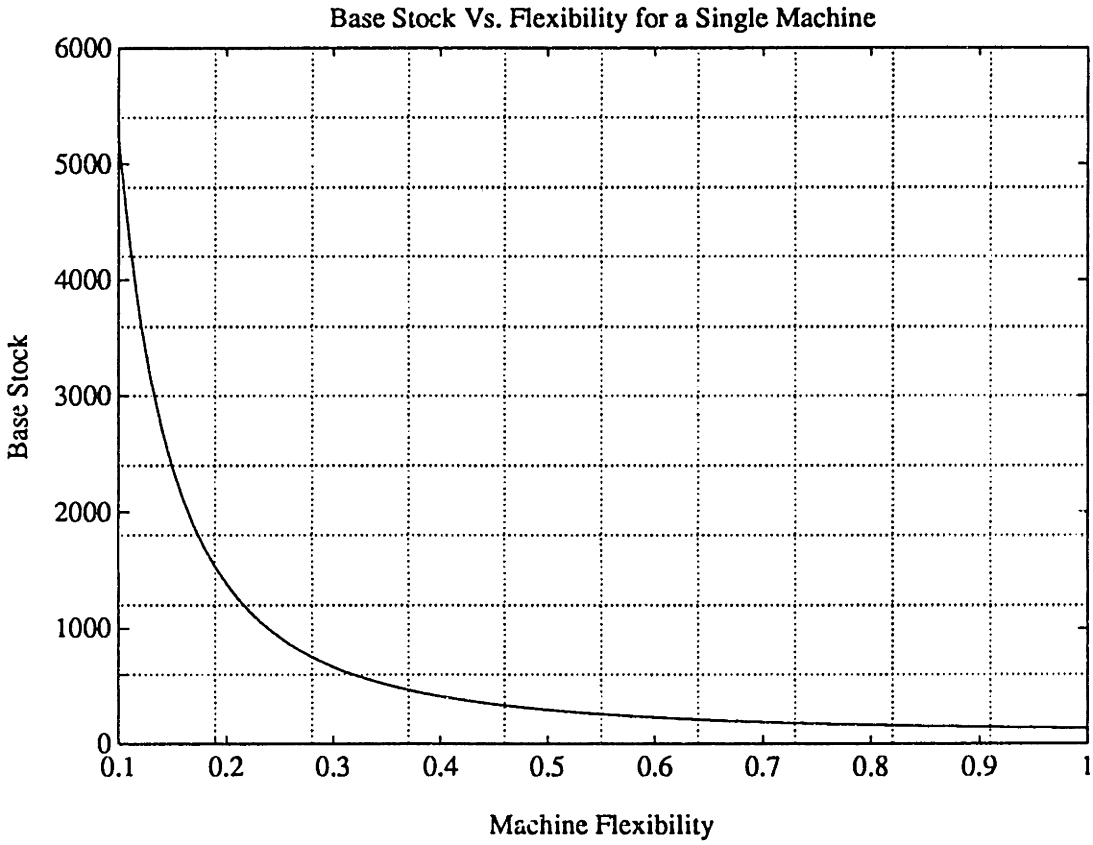

Figure 3.3: Plot of Base Stock vs. Machine Flexibility for a Single Machine

$\mu = 100, \ \sigma = 20$

If F is fixed, then the RHS of equation (3.23) is constant. If one were to plot F_2,the flexibility of machine 2 versus F_1, the flexibility of machine 1, the relationship between the two flexibilities is a straight line (See Figure 3.4).

We now consider each machine separately. Let α_1 be the fraction of WIP inventory processed in each period for machine 1. The control rule for production for machine 1 is identical to the linear rule that we had described earlier.

$$P_{1t} = \alpha_1 X_{1t}$$

Using an analysis similar to the one that we had described earlier, we get $E[P_{it}] = \mu_1$ and $Var[P_{1t}] = \frac{\alpha_1^2\sigma_1^2}{2\alpha_1 - \alpha_1^2}$. Since $P_{1t} = \alpha_1 X_{1t}$, the mean and variance of the inventory level for item 1 is

$$E[X_{1t}] = \frac{\mu_1}{\alpha_1} \qquad and \qquad Var[X_{1t}] = \frac{\sigma_1^2}{2\alpha_1 - \alpha_1^2} \tag{3.24}$$

Therefore, the level of base stock for item 1 is given by

$$\begin{aligned} B_1 &= E[X_{1t}] + k\sqrt{Var[X_{1t}]} \\ &= \frac{\mu_1}{\alpha_1} + k\frac{\sigma_1}{\sqrt{2\alpha_1 - \alpha_1^2}} \\ &= \frac{1}{\alpha_1}\Big[\mu_1 + kF_1\sigma_1\Big] \\ &= \frac{1+F_1^2}{2F_1^2}\Big[\mu_1 + kF_1\sigma_1\Big] \end{aligned}$$

Upto this point, we have said the following: Suppose we knew the value of F_1, the flexibility of machine 1. Then the lead time $\frac{1}{\alpha_1}$ is known and the base stock for item 1 can be determined. Once F_1 is known, F_2 can be determined from equation (3.23) since the flexibilities must satisfy this relationship. Using the value of F_2, the lead time for item 2 and the base stock for item 2 can be determined.

$$B_2 = \frac{1}{\alpha_2}\Big[\mu_2 + kF_2\sigma_2\Big]$$

Therefore, the total base stock is given by

$$\begin{aligned} B &= B_1 + B_2 \\ &= \frac{1+F_1^2}{2F_1^2}\Big[\mu_1 + kF_1\sigma_1\Big] + \frac{1+F_2^2}{2F_2^2}\Big[\mu_2 + kF_2\sigma_2\Big] \end{aligned} \tag{3.25}$$

How would one choose the value of F_1 so that the total base stock at the two stages in minimized? As F_1 increases, the base stock for item 1 decreases. But the value of F_2 decreases so the base stock for item 2 would increase. A plot of the total base stock versus the flexibility of machine 1 is shown in Figure 3.5. There will be an optimal value of F_1 for which the total base stock is minimized. While it is difficult to obtain a closed form expression of the optimal value of F_1, the minimum can easily be computed using a line search procedure (the Fibonacci search procedure works well in this case). The optimization problem can be written as

$$\begin{aligned} \underset{F_1,F_2}{\text{Minimize}} \quad & B_1 + B_2 \\ \text{s.t.} \quad & \sigma_1 F_1 + \sigma_2 F_2 = \sigma F \\ & 0 \leq F_1 \leq 1 \\ & 0 \leq F_2 \leq 1 \end{aligned}$$

where the value of F is fixed.

3.4 Flexibility or Dedicated Machines

The question now arises as to why it would be better to dedicate machines for the different products. After all, if we consider machine to be a server, from the queueing standpoint, the average waiting time in steady state, and hence by Little's law, the average number of customers in the system would be lower if the two streams were merged. To get some insight, let us go back to equation (3.22). In the case of one machine, the total base stock is

$$B = \frac{1+F^2}{2F^2}\left[\sum_i \mu_i + kF\sum_i \sigma_i\right] \tag{3.26}$$

In the case of two machines, the total base stock is given by

$$\begin{aligned} B &= B_1 + B_2 \\ &= \frac{1+F_1^2}{2F_1^2}[\mu_1 + kF_1\sigma_1] \\ &\quad + \frac{1+F_2^2}{2F_2^2}[\mu_2 + kF_2\sigma_2] \end{aligned} \tag{3.27}$$

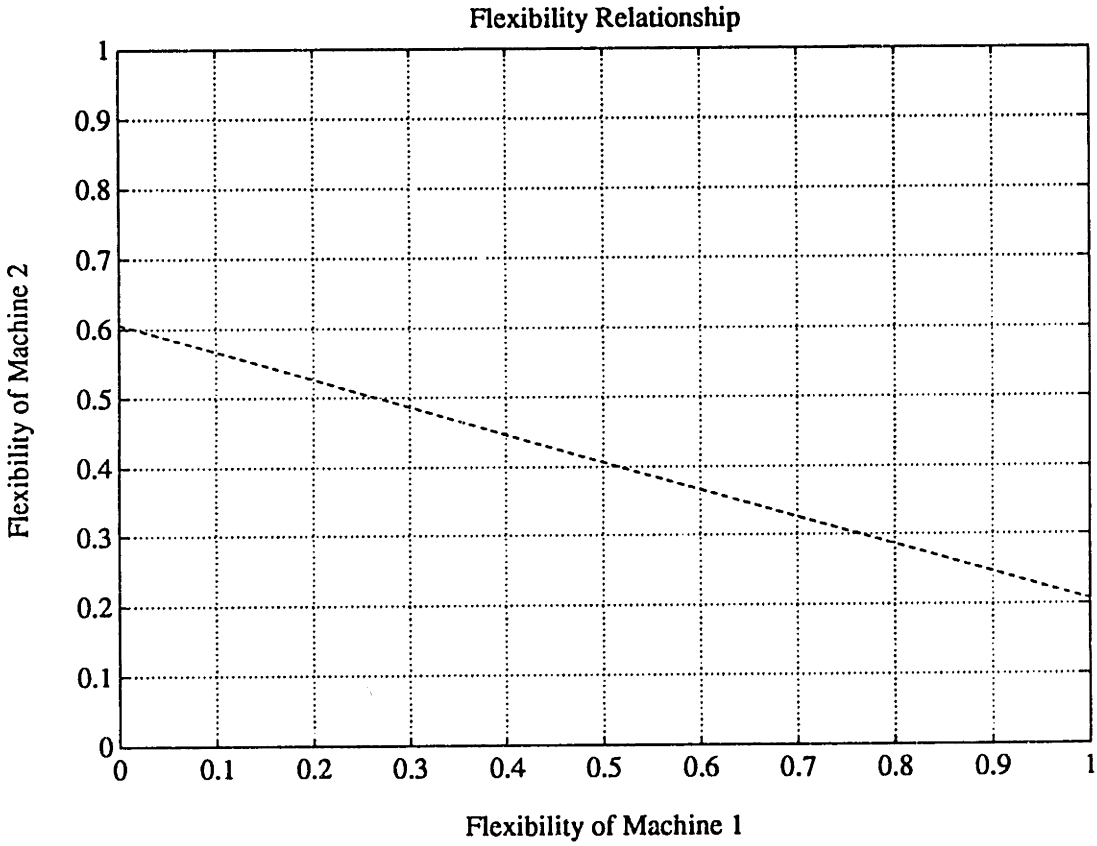

Figure 3.4: Relation between the flexibility of the two machines

$$2F_1 + 5F_2 = 3$$

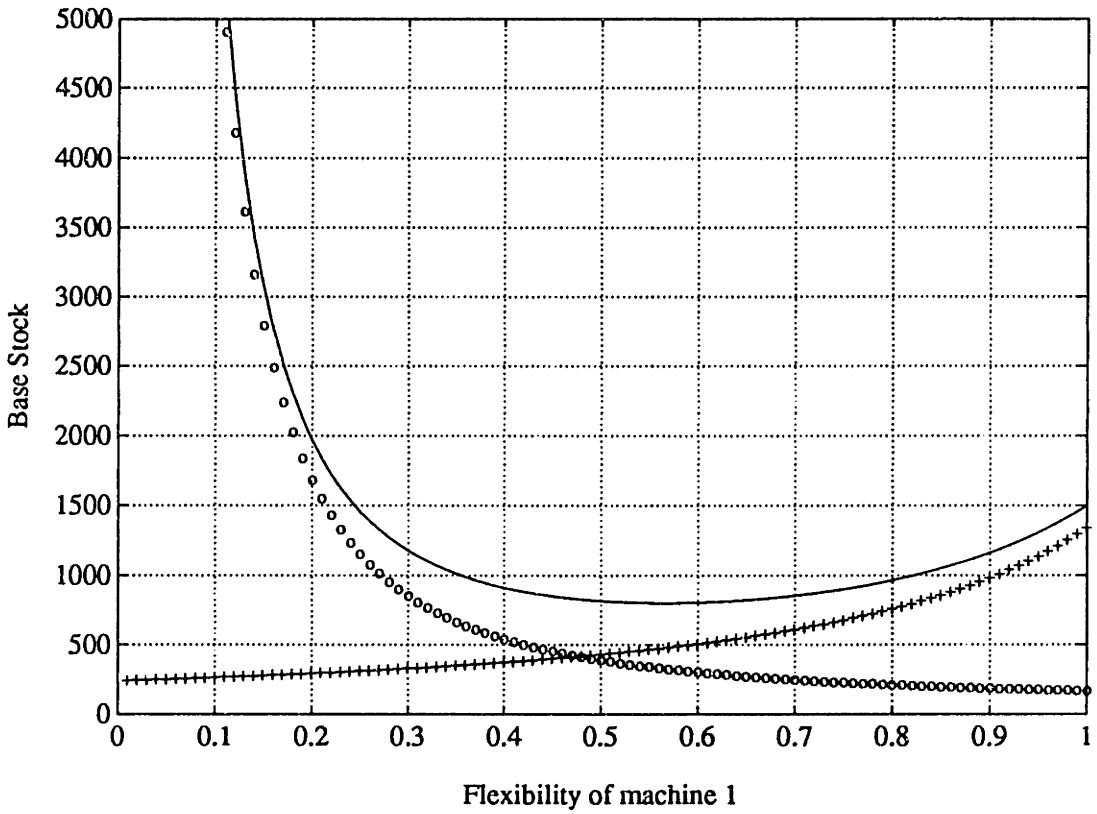

Figure 3.5: Base stock levels as a function of machine flexibility.

$$2F_1 + 5F_2 = 3$$

++++++++++ Base stock for item 1, $\mu_1 = 100$, $\sigma_1 = 40, k = 1.645$.
oooooooooo Base stock for item 2, $\mu_2 = 80, \sigma_2 = 30, k = 1.645$.
---------- Total base stock for the two items.

3.4.1 Correlation between the demands for the two items

In the analysis so far, we had assumed that the demands between the two products are independent. Suppose we now assume that the demands between the two items in a single period are correlated. We shall still assume that the inter-temporal demands for each item are independent, that is, for $s \neq t$, $Cov(D_{is}, D_{jt}) = 0$, $j = 1, 2$.

The analysis would be slightly different for the case where we process both items on a single machine. If we process the two items on separate machines, the correlation between the demands does not affect the analysis because we consider each machine as an isolated unit. Let η be the coefficient of correlation between the demands for the two items, i.e., $\eta = \dfrac{Cov(D_{1t}, D_{2t})}{\sigma_1 \sigma_2}$. The variance of the total demand in a period becomes

$$\begin{aligned} Var[D_t] &= Var[D_{1t} + D_{2t}] && (3.28) \\ &= \sigma_1^2 + \sigma_2^2 + 2\eta\sigma_1\sigma_2 \\ &:= \sigma^2 && (3.29) \end{aligned}$$

When the demands were independent, σ^2 was equal to the sum of the variances of the demands but in this case, there is a covariance term. The rest of the analysis and the results stay the same. From equation (3.29) and the definition of flexibility in equation (3.19), we find that for a fixed excess capacity, χ, as the coefficient of correlation, η, becomes more negative, the variance of the total demand σ^2 decreases. Therefore, the flexibility of the single machine F increases. On the other hand, when η becomes positive, σ increases and so the flexibility of the machine F decreases. When η is negative, it means that as the demand for one item increases, the demand for the other item is going to be lower. In that case, it might be advantageous to process both items on a single machine.

When the demands are positively correlated, it is less clear which choice would be preferable. In the case of a single machine, the flexibility is going to decrease. Therefore, the lead time, $\frac{1}{\alpha}$, is going to increase, and the total base stock would

increase. We would expect that given the values of μ_1, μ_2 and χ, the values of σ_1 and σ_2 for which dedicated machines would be preferable, would consitute a larger set as the correlation increases.

Would it be better to have two dedicated machines instead of a single machine? This is best discussed by means of an example. Consider a situation where the coefficient of variation of the demand for item 1 is relatively low whereas the coefficient of variation for the demand for item 2 is high. In this example, the mean and standard deviation of the demand for item 1 is 100 and 5 respectively, while the corresponding numbers for item 2 are 100 and 40 respectively. The coefficient of variation for the demands for items 1 and 2 are 0.05 and 0.40 respectively. Let a total of 40 units of excess capacity be available. For this problem, $\mu_1 = 100$, $\mu_2 = 100$, $\sigma_1 = 5$, $\sigma_2 = 40$, $\chi = 40$ and $\sigma^2 = \sqrt{5^2 + 40^2} = 40.31$.

	Item 1	Item 2
Mean Demand	100	100
Standard Deviation	5	40
Excess Capacity	40	
Flexibility	0.6031	
Lead Time	1.88	
Base Stock	197.32	262.60

Table 3.1 : Base stock for items processed on a single machine: an example

When there is one machine present, the lead time n, machine flexibility F and the levels of base stock for items 1 and 2 can be determined. See Table 3.1. When there are two machines present, one has to decide how to assign capacity between the two machines. Once the capacity has been assigned, the machine flexibilities F_1 and F_2 can be determined but they must satisfy the relation $\sigma_1 F_1 + \sigma_2 F_2 = \sigma F$. In this example, this corresponds to the equation $5F_1 + 40F_2 = 40.31F$. Suppose we assign

8.25 units of capacity to machine 1. Then the machine flexibilities, lead times and levels of base stock for the two items are shown in Table 3.2.

	Item 1	Item 2
Mean Demand	100	100
Standard Deviation	5	40
Excess Capacity	7.86	32.14
Flexibility	0.955	0.489
Lead Time	1.05	2.60
Base Stock	113.04	342.94

Table 3.2 : Base stock for items processed on two separate machines: an example

If we look at Tables 3.1 and 3.2, we see that in the case of a single machine, the lead time is 1.88. Suppose we have two machines. Since item 1 has a low coefficient of variation, we could try and reduce the lead time for item 1 without assigning a large amount of capacity there. By allocating 7.86 units of capacity to machine 1, the lead time is 1 and the base stock for item 1 is 113.04. The result is that the lead time for item 2 is 2.60. The base stock for item 2 increases, but a decrease in the level of one might more than offset an increase in the level of the other. The total base stock when there are two dedicated machines is 455.98, while the total base stock for the case of a single machine is 459.93. The difference in this example is minor but a few points need to be mentioned. First, in this example, the demands were assumed to be independent. If the demands were positively correlated, the total base stock would be the same for the two machine case, but it would increase when there is a single machine. Second, if the objective function, instead of being $B_1 + B_2$, was $c_1 B_1 + c_2 B_2$ where c_1 and c_2 are the costs per unit of inventory for items 1 and 2 respectively, the differences could be greater or smaller depending on the values of c_1 and c_2.

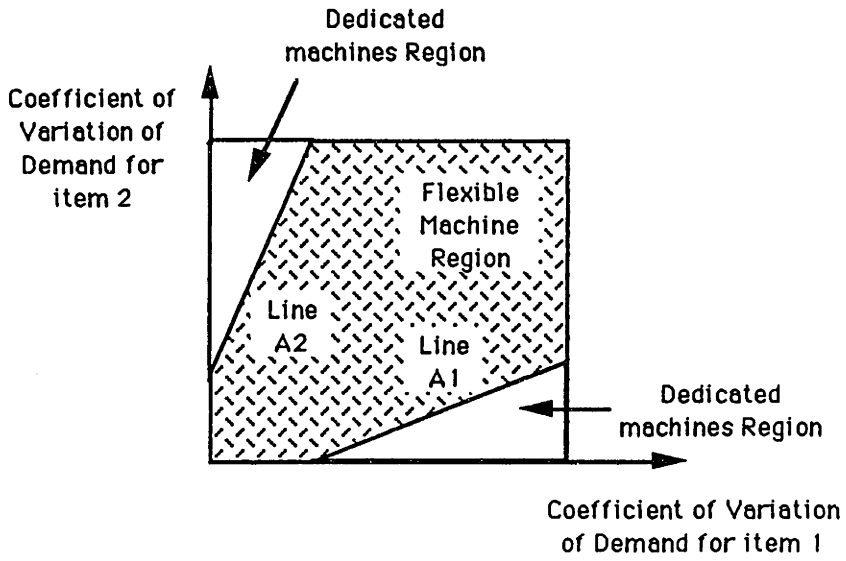

Figure 3.6: Plot of the Separating Regions

3.5 Analysis of the Results

In Figure 3.6, the variable on the x-axis is the coefficient of variation of the demand for item 1 (CV_1) and the variable on the y-axis is the coefficient of variation of the demand for item 2 (CV_2). For each ordered pair (CV_1, CV_2), we would like to know whether the preferable configuration, that corresponds to those values of CV_1 and CV_2, is the flexible machine or the dedicated machine configuration. The way the issue was decided was the following: The values of the mean demands μ_1 and μ_2, the correlation between the demands η, the excess capacity χ and the service factor k are known. Given the values of σ_1 and σ_2, the standard deviation of the total demand, σ, can be calculated. Therefore, the flexibility of the single machine, F, and the total base stock for the single machine, B, can be determined. Suppose we decide to go in for two machines. In the case of two machines, the optimal capacity allocation that would minimize the total safety stock is determined (See Figure 3.5). If the safety stock in the latter case is lower, the decision is made to go for two dedicated machines. Otherwise, one decides to process both items on a single flexible machine.

If one were to do that for each ordered pair (CV_1, CV_2) and demarcate the regions on the plane, the plot would look like the one shown in Figure 3.6. There are two regions that correspond to the dedicated machine case: when CV_1 is low and CV_2 is high and vice versa. When both the variables are approximately the same level, the single machine case is preferable (which corresponds to the central band in the graph).

As the correlation between the demands increases, we would expect the dedicated machine set to be larger. The line $A1$ would rotate in the counterclockwise direction and the line $A2$ would rotate in the clockwise direction. In Figures 3.7, 3.8, 3.9 and 3.10, we provide plots that indicate how one should go about deciding the issue of flexibile versus dedicated machines. For example, in Figure 3.7, the mean values of the demand, μ_1 and μ_2 are both equal to 100. The excess capacity χ is 40 and the service level is 95% (Service factor $k_{0.95}$ =1.645). Suppose the coefficient of correlation, $\eta = -0.5$. We find that for most values of σ_1 and σ_2, one would choose a single flexible machine.

As η increases, the set of values of CV_1 and CV_2, for which the choice of dedicated machines would be preferable, increases. In all cases, we find that when both coefficients of variation are low, a single machine is more appropriate. As we would expect, when the value of σ_1 is low and σ_2 is high, or vice versa, the decision would be to dedicate lines for each item. In the examples that we consider, we allow the values of η to range from -0.5 to 0.8.

In Figure 3.7, the variables on the x-axis and y-axis are the coefficients of variation of the demands for item 1 and 2 respectively. In Figure 3.7, $\mu_1 = 100$, $\mu_2 = 100$, $\chi = 40$ and $k = 1.65$. Since μ_1 and μ_2 are known, given the values of the coefficients of variation, σ_1 and σ_2 can be determined.

In Figure 3.7, we had taken a look at two demand processes where both the demands had the same mean, i.e., $\mu_1 = \mu_2$. Without loss of generality, suppose that the demand for item 2 had a higher mean. In Figure 3.8, we examine the case

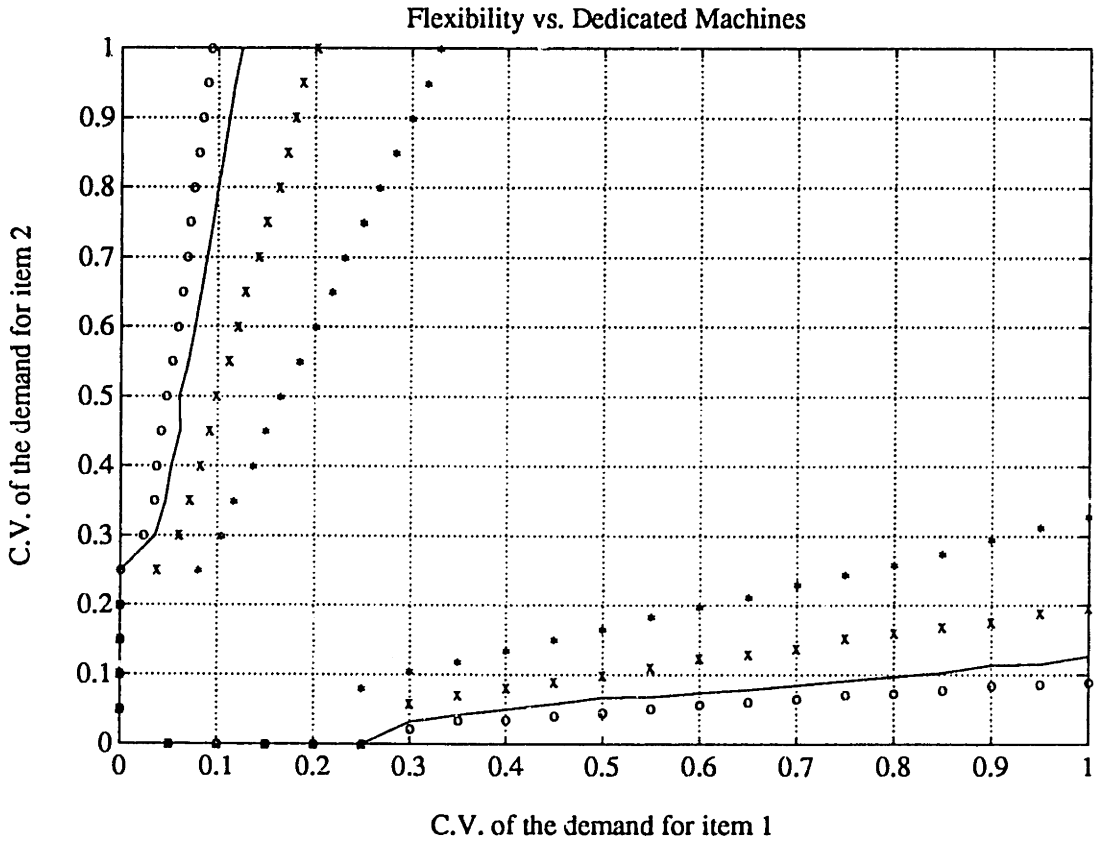

Figure 3.7: Flexible or Dedicated Machines

$\mu_1 = 100$, $\mu_2 = 100$, $\chi = 40$, Service level $= 0.95$

oooooooooo Service Level $= 0.95$, $\eta = -0.5$
---------- Service Level $= 0.95$, $\eta = 0.0$
xxxxxxxxxx Service Level $= 0.99$, $\eta = 0.5$
********** Service Level $= 0.99$, $\eta = 0.8$

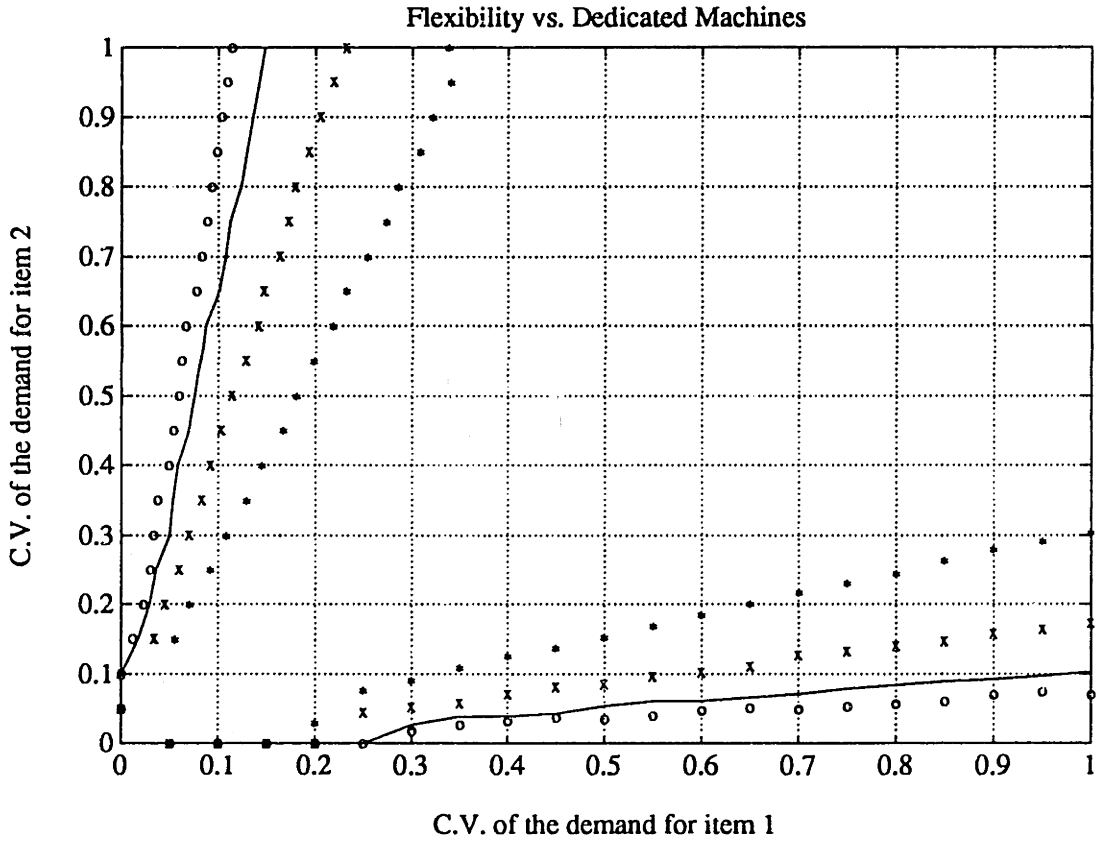

Figure 3.8: Flexible or Dedicated Machines

$\mu_1 = 100$, $\mu_2 = 200$, $\chi = 40$, Service level $= 0.95$

oooooooooo $\eta = -0.5$
---------- $\eta = 0.0$
xxxxxxxxxx $\eta = 0.5$
********** $\eta = 0.8$

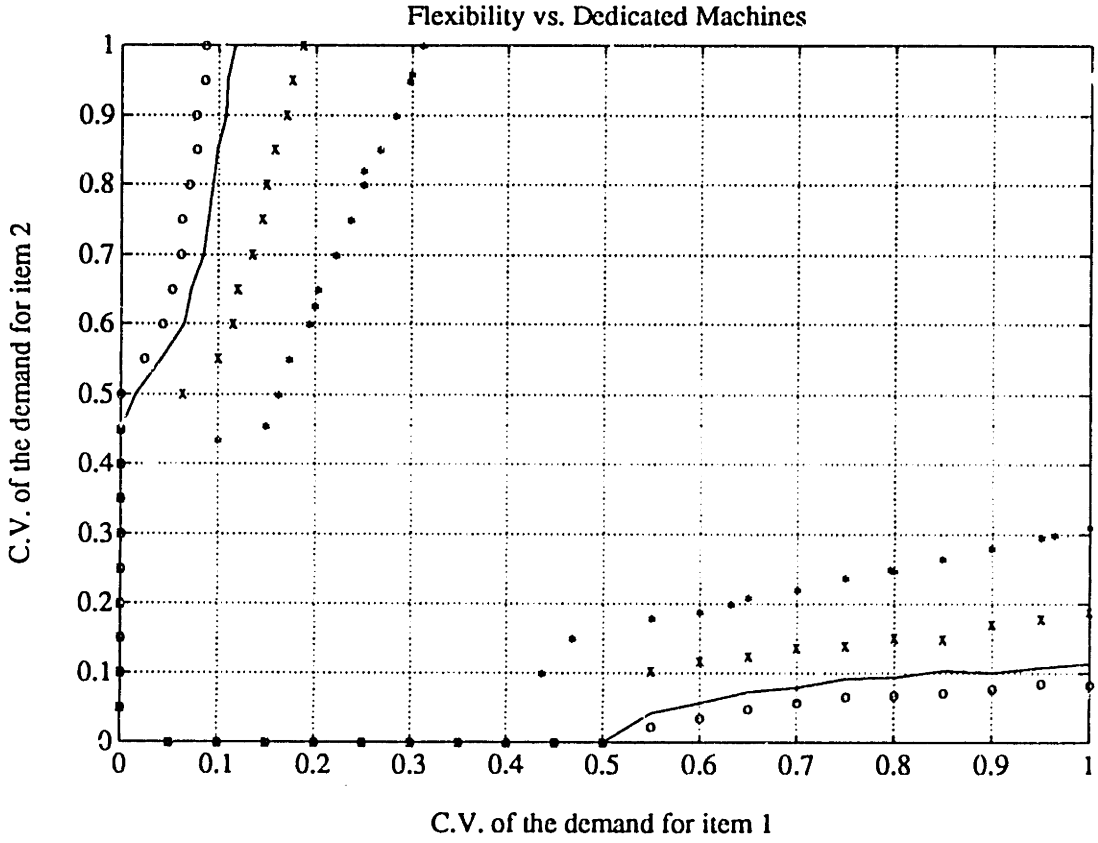

Figure 3.9: Flexible or Dedicated Machines

$\mu_1 = 100$, $\mu_2 = 100$, $\chi = 80$, Service level $= 0.95$

oooooooooo $\eta = -0.5$
---------- $\eta = 0.0$
xxxxxxxxxx $\eta = 0.5$
********** $\eta = 0.8$

when $\mu_1 = 100$ and $\mu_2 = 200$. The excess capacity is 40 and the service level factor, $k = 1.65$.

Comparing Figures 3.7 and 3.8, we see that one would prefer dedicated machines for slightly higher values of σ_1. The values of σ_2 would have to be scaled down by a corresponding factor to account for the higher mean demand rate of item 2.

In Figures 3.7 and 3.8, we had taken a look at how the relative values of the mean demands affect the decision to allocate fixed capacity, when the excess capacity is fixed. The next question that we could ask is the following: How does the availability of excess capacity affect the decision to allocate capacity? In Figure 3.9, the values of μ_1, μ_2, χ and k are 100, 100, 80 and 1.65 respectively. (Recall that in Figure 3.7, the corresponding values were 100, 100, 40 and 1.65 respectively). Comparing the two plots, we see that a single machine would be flexible over a larger set of values of σ_1 and σ_2. In retrospect, the reason for this seems apparent. To see why this is the case, let us take a look at the values of $\sigma_1 = 3$, $\sigma_2 = 50$ and $\eta = -0.5$. The standard deviation of the total demand, $\sigma = 48.57$. In the former case, when $\chi = 40$, the flexibility of a single machine $F = 0.499$, and the lead time $n = 2.54$. If one decides to split the capacity between two machines, the optimal excess capacity allocation $\chi_1 = 4.93$ and $\chi_2 = 35.07$. The flexibilities for the two machines are 0.9963 and 0.425 respectively, and the lead times are 1.003 and 3.27 respectively. The choice of two machines is preferable because one is able to reduce the lead time for item 1 and reduce its safety stock. In case 2, the flexibility is 0.99. Hence, there is not much potential for reduction in the lead time by splitting the capacity across two machines.

In Figure 3.10, we had compared plots for different values of μ_1, μ_2, χ and η, but for a fixed service factor, $k = 1.65$. One might ask the following question: Does the service level affect the decision to allocate capacity. In Figure 3.11, we see this for two different service levels, a 95% service level with $k = 1.65$, and a 99% service level with $k = 2.33$. In this plot, we find that the demarcating lines are identical. We could not find any evidence that setting the service level affects the decision making

process.

3.6 Capacity Allocation in the presence of setups

In the model that we have described, we have not considered the effect of setups and how their presence would alter the configuration. In this section, we would like to describe how we model setups and how the plots that we had shown earlier would change as a result.

In the case of a single flexible machine, two items are being processed on the same machine and they are processed in batches. Depending on the nature of the machine and how these batches are sequenced, the setup time could vary. On the other hand, when there are dedicated machines for each item, the configuration of each machine could be customized so that the setup times are considerably lower. As a result, the total processing capabilities when there are two dedicated machines could be higher than when there is a single flexible machine. In this case, in addition to the advantage of preserving the stability of the streams for the two products, dedicated machines have a higher processing rate. Intuitively, we would expect that the domain of the set (where the two machine alternative is preferable) to be larger when the setup times are considered and as these times increase, we would expect this set to increase.

Going back to section 3.3, let C be the capacity when there is a single machine and let C_1 and C_2 be the capacities when there are two machines. Since $C_1 + C_2 \geq C$ and $\mu_1 + \mu_2 = \mu$ (the mean production rate must be the same), it follows that

$$\chi_1 + \chi_2 \geq \chi$$

Therefore, the constraint $\sigma_1 F_1 + \sigma_2 F_2 = \sigma F$ is no longer valid. This is because the flexibility of the single machine F is lower than what it would have been if there were no setups because we have less excess capacity. Therefore the optimization problem for the two machine case is simply

$$\underset{F_1,F_2}{\text{Minimize}} \quad B_1 + B_2$$

$$\text{s.t.} \quad \sigma_1 F_1 + \sigma_2 F_2 = \frac{\chi_1 + \chi_2}{k}$$
$$F_1 \in (0,1], F_2 \in (0,1]$$

If we look at Figure 3.12, we see a plot for the case when $\mu_1 = 100$, $\mu_2 = 100, \chi = 40$ and $k = 1.65$. A capacity loss equal to zero simply means that no capacity is lost due to setups. When the capacity loss equals 5, it means that in the case when there are two machines, the total processing capacity is $100 + 100 + 40 = 240$ units, but when there is a single machine, the processing capacity is $240 - 5 = 235$ units in each period. Five units of processing capacity are lost due to setups.

We see that as the capacity loss increases, the set of values for which the two machine alternative is preferable becomes larger. In this example, we have looked at three values of the setups: 0, 5 and 10 units. We have assumed in this example that the correlation between the demands is -0.5.

When the correlation is increased to zero (see Figure 3.13), we see the same behavior. Moreover, if we compare the two figures for the same value of capacity loss and observe the behavior, we see that for a given capacity loss, as the correlation coefficient increases the domain for the dedicated machine set increases.

3.7 Multiple items at a single stage

In the analysis so far, we have looked at the case where two items are processed at a stage. We have seen when it is preferable to process the items separately and when it is preferable to combine the flows of the two items. We have considered the case when there are no setups, and have also seen how the presence of setups alters the decision to combine or separate the flows.

Suppose there are more than two items being processed at the stage. The demand for some of the items is highly variable, while the demand for the other items is extremely stable. If there are two machines and there is a fixed capacity that can

be allocated between these two machines, how would the items be grouped so that the total base stock in the system was minimized? Would the items be grouped so that at each machine, a mixture of high variable and stable items are processed and the overall coefficient of variation of the demand is reduced? Or would the items be grouped so that the ones with high variability are processed at one machine, and the low variability items are processed at the other machine?

This is best explained by means of an example. Suppose there are four items being produced at the stage. The mean demand for each item in a single period is 100 units. The standard deviations of the demand for items 1, 2, 3 and 4 are 5, 40, 5 and 40 respectively. Items 1 and 3 are the ones which have a stable demand and items 2 and 4 have a highly variable demand. Assume that the demands for the four items are independent. Suppose we processed items 1 and 2 at one machine, and items 3 and 4 are processed at the second machine. The combined coefficient of variation of the demand for the items processed at machine 1 is $\frac{\sqrt{5^2+40^2}}{100+100} = 0.202$, and the combined coefficient of variation of the demand for the items processed at machine 2 is also $\frac{\sqrt{5^2+40^2}}{100+100} = 0.202$. In this scenario, we are mixing the highly variable items with the stable items at each machine so that the coefficient of variation for the combined stream is lower than the coefficient of variation for the high variable items.

In the second case, suppose items 1 and 3 are processed at machine 1, and items 2 and 4 are processed at machine 2. In this scenario, we are processing the stable items at machine 1 and the high variable items are being processed at machine 2. Then the combined coefficient of variation of the demand for machine 1 is $\frac{\sqrt{5^2+5^2}}{100+100} = 0.035$, and the corresponding figure for machine 2 is $\frac{\sqrt{40^2+40^2}}{100+100} = 0.283$. If the sole criterion is to minimize the total base stock in the system, which of these two configurations is better? It is not apparent why either configuration would be preferable. If items 1 and 2 were grouped together, the coefficient of variation of the demand for each machine is 0.202. If items 1 and 3 were grouped together, the coefficients of variation for the two machines are 0.035 and 0.283 respectively. Is the decrease in the coefficient of variation of the demand at one machine sufficient to compensate for the increase

in the corresponding parameter for the other machine? It is not clear whether a coefficient of variation of 0.283 is substantially higher than 0.202, nor whether 0.035 is substantially lower than 0.202.

In the above discussion, we have tried to provide some motivation for looking at the problem of how to group multiple items at a single stage. Let us now formulate the problem. Let n items be processed at a single stage and let μ_i be the mean demand in a single period for item i, $i = 1, \ldots, n$. Let σ_i^2 be the variance of the demand for item i in a single period, and furthermore, let the demands be independent, i.e., $Cov(D_{it}, D_{jt}) = 0$ for all t where D_{it} = demand for item i in period t, $i = 1, \ldots, n$. Let $A = \{i \mid \text{item } i \text{ is processed at machine 1}\}$. That is , A is a set that contains the group of items that are processed at machine 1. Let χ be the total excess capacity at the stage, i.e., $\chi = C - \sum_i \mu_i$. We assume that this capacity can be allocated to either machine 1 or machine 2. For each grouping of items, one could allocate this capacity optimally between the two machines to minimize the total base stock in the system. We would like to determine which grouping would have the lowest total base stock in the system.

The total base stock for machine 1 is

$$B_1 = \frac{1 + F_1^2}{2F_1^2}\left[\sum_{i \in A} \mu_i + kF_1 \sum_{i \in A} \sigma_i \right]$$

where $F_1 = \dfrac{\chi_1}{k\sqrt{\sum_{i \in A} \sigma_i^2}}$ and χ_1 = excess capacity assigned to machine 1.

Similarly, the base stock for machine 2 is

$$B_2 = \frac{1 + F_2^2}{2F_2^2}\left[\sum_{i \notin A} \mu_i + kF_2 \sum_{i \notin A} \sigma_i \right]$$

where $F_2 = \dfrac{\chi_2}{k\sqrt{\sum_{i \notin A} \sigma_i^2}}$ and χ_2 = excess capacity assigned to machine 2.

The total excess capacity at the stage χ must be equal to $\chi_1 + \chi_2$. Let us consider the following example: Suppose there are 8 items being produced at a stage. The mean demand for each item in a single period is 100 units. The standard deviation

for each item is given by σ_s or σ_v depending on whether the item is stable or variable. For example, if $\sigma_s = 3$ and $\sigma_v = 40$, and we say that there are three stable items, then the mean demand vector is (100 100 100 100 100 100 100 100) and the standard deviation vector is (3 3 3 40 40 40 40 40). We also assume in this example that the demands are independent. Suppose there was a total of 60 units of excess capacity at the stage that could be allocated between machine 1 and 2. How would the items be grouped? Would all the items be processed on a single machine, or on two machines? If they were processed on two machines, we would like to determine the grouping of the items. We would like to know whether stable items will be processed at machine 1 and the remaining variable items will be processed at the other machines, or whether there will be a mix of items at each machine.

	Number of stable items							
Std. dev.	1	2	3	4	5	6	7	8
$\sigma_s = 3$	One	Two	Two	Two	Two	Two	Two	One
$\sigma_v = 40$	Mach	Mach	Mach	Mach	Mach	Mach	Mach	Mach
$\sigma_s = 4$	One	One	Two	Two	Two	Two	Two	One
$\sigma_v = 40$	Mach	Mach	Mach	Mach	Mach	Mach	Mach	Mach
$\sigma_s = 5$	One	One	One	Two	Two	Two	Two	One
$\sigma_v = 40$	Mach	Mach	Mach	Mach	Mach	Mach	Mach	Mach

Table 3.3: Flexibility versus dedicated machines for multiple items.

$$\mu = (100\ 100\ 100\ 100\ 100\ 100\ 100\ 100\ ),\ \chi = 60.$$

We see from the above table that depending on the values of σ_s, σ_v and the number of stable items, one can determine the optimal machine configuration. According to the table, whenever we choose the two machine configuration, the stable items are

processed on one machine and the high variable items are processed on the second machine.

For example, when $\sigma_s = 3$ and $\sigma_v = 40$, and the number of stable items equals one, then the optimal configuration is one single machine where all eight items are processed. But when the number of stable items increases to two, then the optimal configuration is the two machine choice where the stable items are processed on machine 1 and the other items are processed on machine 2. But why is the one machine choice optimal when there is only one stable item? Can we find a good reason for this *anomaly*? The reason is that although the coefficient of variation of the demand for this single item is $\frac{3}{100} = 0.03$, it is not low enough to ensure the optimality of the two machine configuration. When there are two stable items, the coefficient of variation of the demand for the combined streams is $\frac{\sqrt{3^2+3^2}}{100+100} = 0.021$ which is low enough to ensure the optimality of the two machine configuration. This reinforces the point that when we look at the variability of the demand, it is not only important to look at the standard deviation, but also value of the standard deviation relative to the mean, i.e., the coefficient of variation.

When σ_s is increased to 4, the one machine case is optimal even there are two stable items. Only when the number of stable items is increased to three is it optimal to have two machines, one for the stable items and one for the variable items. When all the eight items have stable demands, then the one machine choice becomes optimal again.

In the above analysis, we have assumed that the demands are independent. Things become much more complicated when the demands are correlated. Suppose the demands for the stable items are negatively correlated with the demands for the high variable flows. In that case, it may be advantageous to merge the stable and high variable streams to take advantage of the correlation effects. We choose not to address this problem here, but leave it to future research to address these issues that have been raised.

3.8 Concluding Remarks

In this chapter, we have looked at the problem of allocating capacity for a single stage in a manufacturing system where two items are produced. We looked at the conditions under which the choices of a single machine or separate machines are preferable. We extended these results to the case where there are setups and where multiple items are processed at a single stage. It remains to study these results for the case of multistage networks and examine the overall system behavior when there is more than one stage.

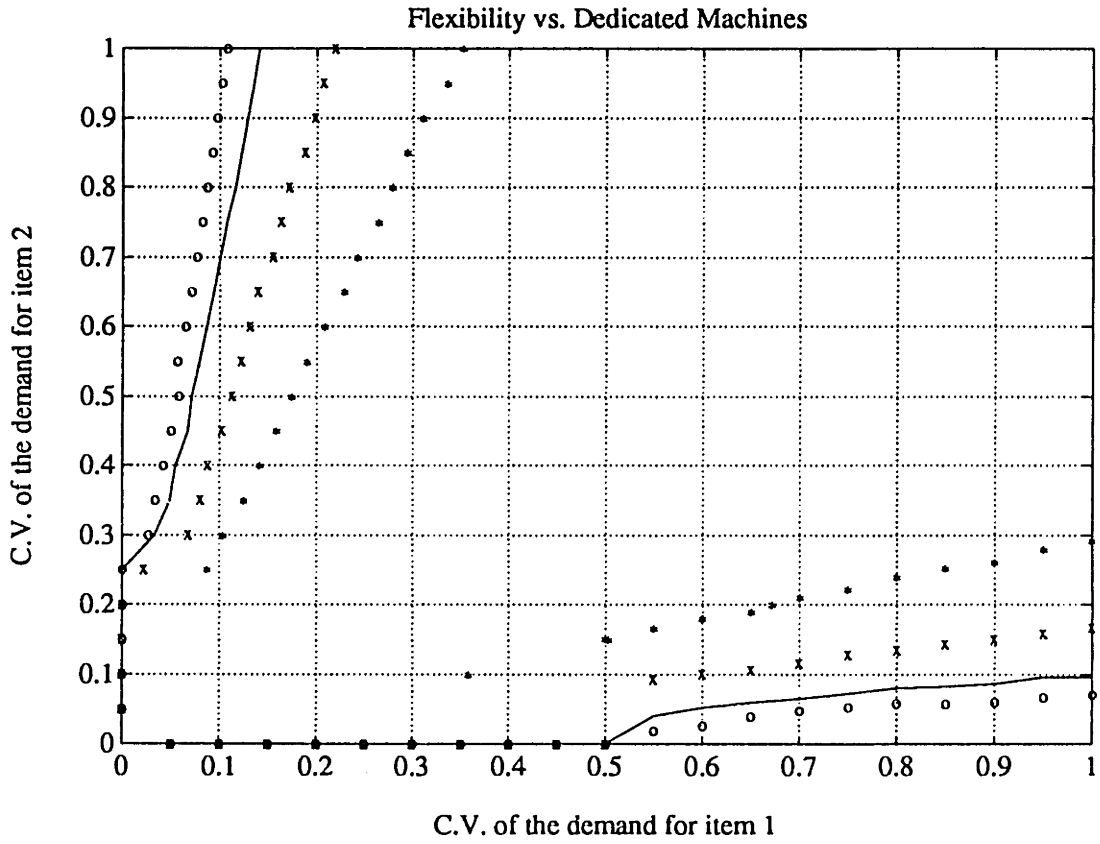

Figure 3.10: Flexible or Dedicated Machines

$\mu_1 = 100$, $\mu_2 = 200$, $\chi = 80$, Service level $= 0.95$

oooooooooo $\eta = -0.5$
---------- $\eta = 0.0$
xxxxxxxxxx $\eta = 0.5$
********** $\eta = 0.8$

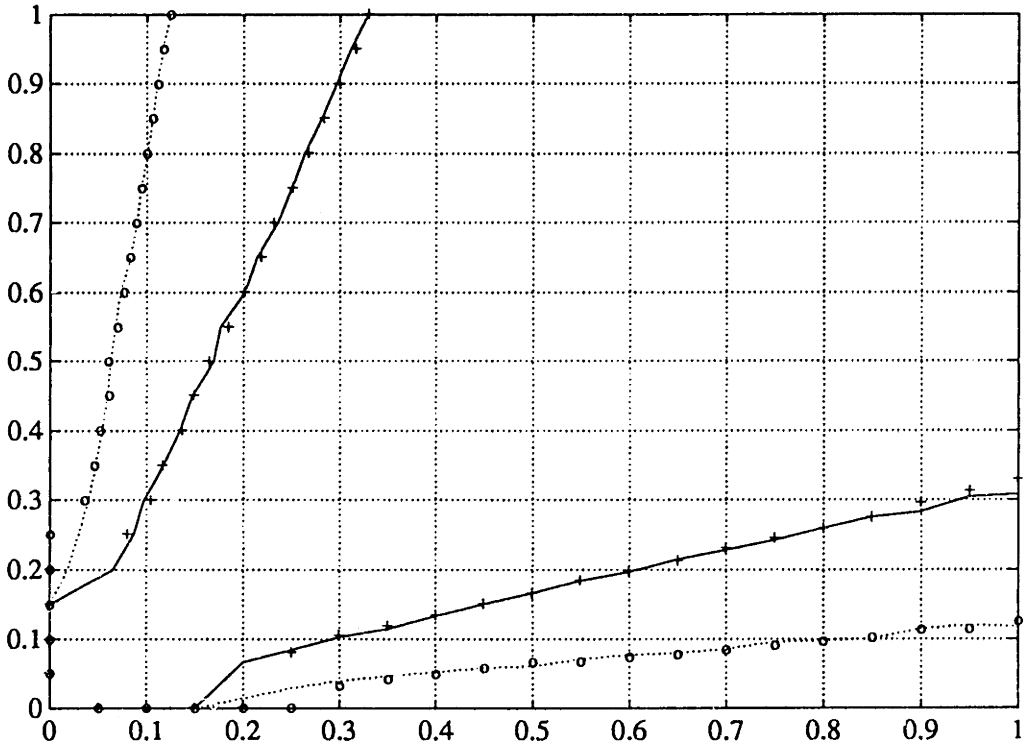

Figure 3.11: Flexibility or Dedicated Machines for different service levels

$\mu_1 = 100$, $\mu_2 = 100$, $\chi = 40$

oooooooooo Service Level = 0.95, $\eta = 0.0$

..... Service Level = 0.99, $\eta = 0.0$

++++++++ Service Level = 0.95, $\eta = 0.8$

xxxxxxxxxx Service Level = 0.99, $\eta = 0.8$

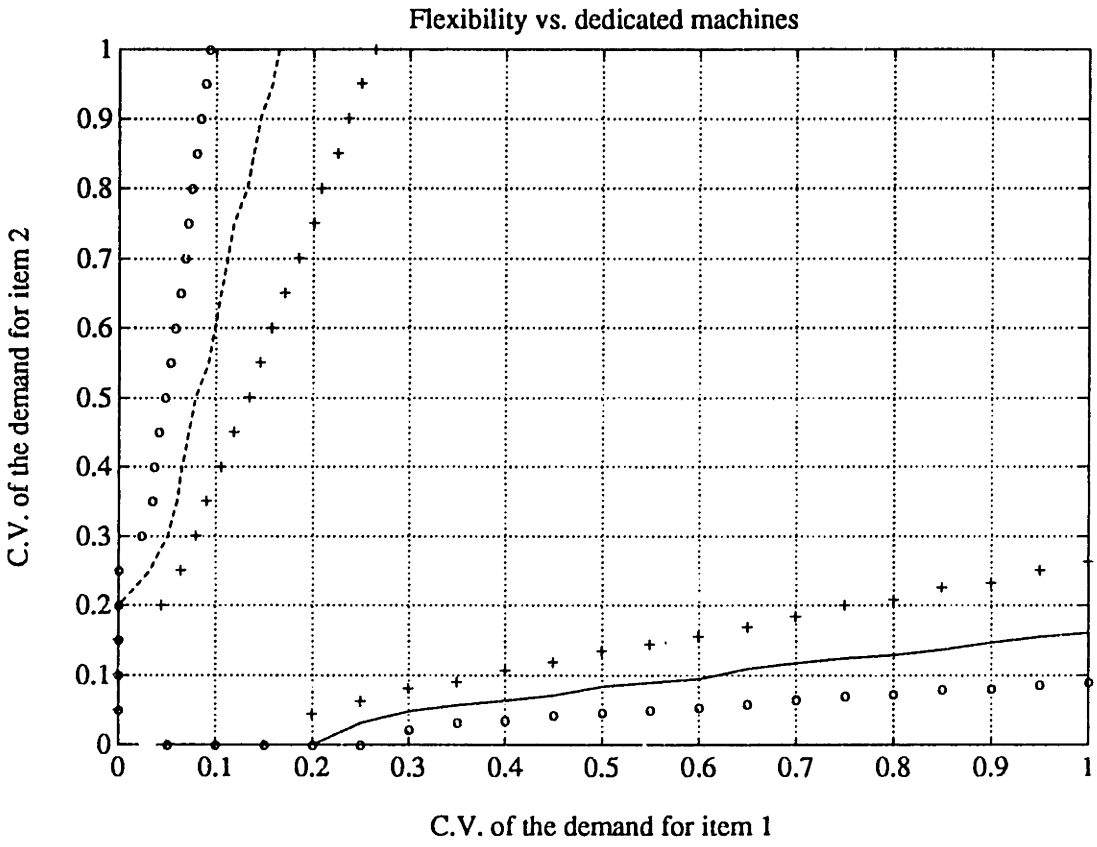

Figure 3.12: Flexibility or Dedicated Machines in the presence of setups

$\mu_1 = 100$, $\mu_2 = 100$, $\chi = 40$
$\eta = -0.5$, Service level $= 0.95$

oooooooooo Capacity Loss = 0 units
---------- Capacity Loss = 5 units
++++++++ Capacity Loss = 10 units

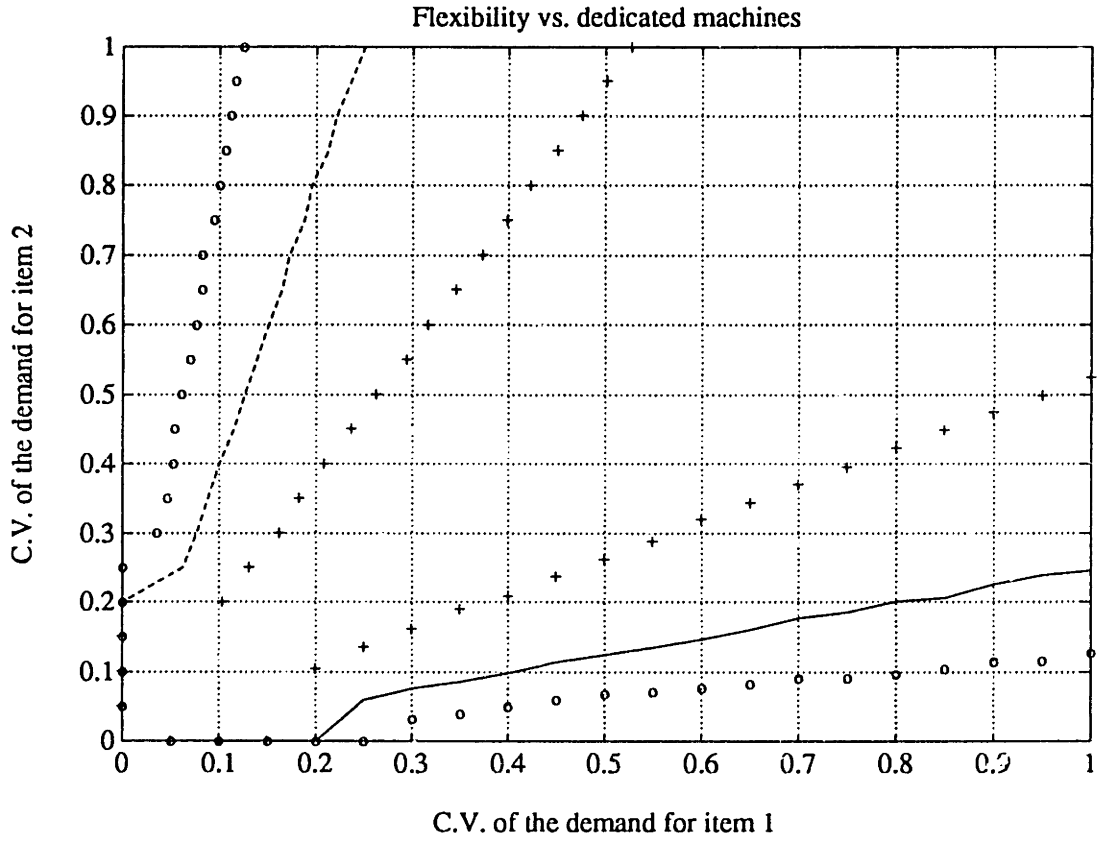

Figure 3.13: Flexibility or Dedicated Machines in the presence of setups

$\mu_1 = 100$, $\mu_2 = 100$, $\chi = 40$
$\eta = 0$, Service level $= 0.95$

oooooooooo Capacity Loss = 0 units
---------- Capacity Loss = 5 units
++++++++ Capacity Loss = 10 units

Chapter 4

The Gamma Model

4.1 Introduction and Motivation

In recent years, there has been a considerable amount of work on modeling multistage manufacturing systems and characterizing the relationship between different stages. At one end, there are aggregate capacity planning models that deal with issues that include the design of production systems and allocating capacity with a view to optimizing some measure of system performance. The focus in these models is to look at the system at a macroscopic level and ignore details such as machine breakdowns, tool changeovers and other elements that actually characterize the typical day-to-day operation of the system. At the other end, work has been done at the microscopic level where the problems of interest may include scheduling of jobs in order to minimize the total make time, determining the effect of setup costs and setup times on lot sizes etc.

Queueing network models have been widely used for capacity planning and for providing descriptions of the behavior of production systems like jobshops. These models are extremely useful because they are able to identify bottlenecks in the manufacturing system accurately. They are also able to capture the sources of variability well. Under the assumption of exponential interarrival and service times, the steady state distribution of the queue lengths has a nice product form (Jackson [Jack57,Jack63], Kelly [Kell79]). More recently, Harrison and Williams [Har87] and Harrison, Williams and Chen [Har90] have used the Brownian approximation to model networks of GI/G/1

queues.

One inherent difficulty of the queueing model is the lack of control on the part of the server. The capacities are assumed to be rigid and fixed, and the server works at a constant rate whenever the queue length is strictly positive. Unless the service rate is exponential, we cannot analyze models in which the server is able to increase the production rate when the queue length increases. It has been argued that in many manufacturing systems, it is the level of work-in-process inventories which determine the rate at which the work in the system is processed. Even though the system has a maximum processing rate, at lower WIP levels, the system is operating at well below the maximum rate. As the level of WIP inventories increase, the processing rate also increases. It is only at higher levels of WIP that the system operates at close to the maximum rate and efforts are made to reduce the WIP levels.

One can provide many examples where the processing rate varies according to the amount of work that is to be processed. In supermarket checkout counters and bank teller windows, the rate at which the queue is served depends on the number of customers in the queue. Although the maximum processing rate is limited by the total number of checkout lanes or teller windows, the number of lanes or windows that are open at any time depends on the number of customers waiting for service. The system does not process the queue at its maximum rate whenever the queue length is positive. Rather, as the number of people waiting increases, employees are pre-empted from other tasks and more lanes or windows are opened in order to reduce the waiting time. Fine and Graves [Fine89] have provided an example in the context of a semiconductor manufacturing facility. In this facility, there is a bottleneck station that operates at its maximum capacity whenever the workload is high, but when the workload is reduced, the processing capability is reduced to allow for routine maintenance tasks. One can also cite the example of aircraft manufacture. In this case, the process here can be thought off as a multistage system where at each stage, a sequence of operations is performed before the aircraft moves over to the next stage. Here, the lead times at the stages are fairly rigid even though there is some uncertainty in the rate at which

the parts or subassemblies arrive into the stage. Since the arrival rate of these parts or subassemblies is variable, the rate of production must vary in order to ensure that these lead times are maintained.

Assuming that the system is always processing work at the maximum rate could lead to significant underestimates in the levels of WIP inventory. Many authors have attempted to address this problem of queue control. Graves ([Gra86,Gra88a]) has proposed a model that uses a simple rule for production at each workcenter. In this model, the production rate is set equal to a fraction α of the WIP inventory at the workcenter, $0 \leq \alpha \leq 1$. We shall present this model in the following section. However, the production capacity is assumed to be unbounded because in this model, the workload could theoretically become very large. This can be remedied by an appropriate choice for the parameter α; more will be said about this when we present the model in the following section.

The linear production rule is extremely useful for several reasons. It is extremely simple and one can interpret $\frac{1}{\alpha}$ as being the lead time for the stage. These lead time are very predictable and this is helpful when one wants to extend the model to multiple stages. The fact that the lead times are predictable permits a decomposition of the multistage system into single stages with the total lead times given by the sum of the lead times.

The linear production rule makes the analysis tractable, although as we had stated earlier, the production capacity is assumed to be unbounded. If we wish to overcome this difficulty with a bounded production rule, the nonlinearity introduced in the model complicates the analysis considerably. Graves [Gra88a] has proposed a truncated production rule of the form $P_t' = \min[\alpha X_t, C]$ where X_t is the WIP level at time t, C is the maximum capacity, P_t' is the production rate at time t and $0 \leq \alpha \leq 1$. In other words, P_t' is set equal to either the rate proposed by the linear production rule αX_t, or the maximum rate C, whichever one is lower. Althought it is possible to obtain an expression for the shape of the density function, one cannot obtain expressions for the mean and variance of the WIP levels.

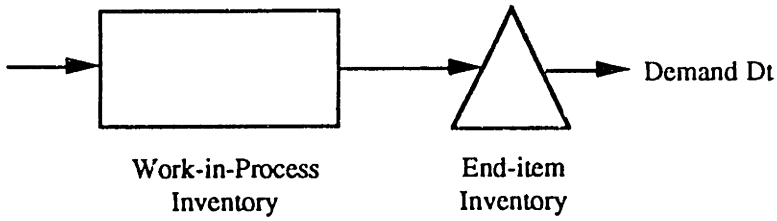

Figure 4.1: A single stage in a manufacturing system

Karmarkar [Karm89] has proposed a model that considers the capacity constraint explicitly while setting the production rate as a function of the workload. We shall also discuss this rule in greater detail in subsequent sections. In this paper, we present a model for a single stage of a manufacturing system that is closely related to the linear production rule described by Graves [Gra86,Gra88b]. Although the model that we describe is a single stage model, it is possible that this could be used as a building block for a multistage system. The model is tactical in the sense that we do not concern ourselves with detailed issues such as scheduling of items, machine breakdowns etc. The model is a descriptive model of how production systems operate and how they would operate if some of the input parameters of the system were changed. Although we present the model in the context of a production system, it is fairly descriptive of other operating entities as well.

The intent of this work is to study a production rule that encompasses a number of production rules that have already been studied. The rule that we consider incorporates a capacity constraint on production and is similar to the one studied by Karmarkar. However, the rule that we study is parameterized and more general. We are able to show that the linear production rule and the queueing model are both limiting cases of this model.

The structure of this chapter is as follows: In section 4.2, we present the linear production rule proposed by Graves in discrete time and in continuous time. In section 4.3, we introduce our model and obtain the flow balance equations for the

workcenter. In section 4.4, we obtain an expression for the mean and variance of the WIP in steady state. We choose not to present the proofs leading to the result in this section. Rather, these proofs are presented in the appendix and the interested reader can refer to them there. In section 4.5, we show how the linear production rule and the queueing model can be obtained as limiting cases. We comment on the model and its assumptions in section 4.6.

4.2 The Linear Production Rule

Graves [Gra86,Gra88a] has proposed a model for production systems that uses a simple production rule. In the next two subsections, we present the original model in discrete time and its formulation in continuous time.

4.2.1 The discrete time model

Consider a single stage in a manufacturing system as shown in Figure 4.1. This is a discrete time model and the stage processes a single item or commodity. The inventory at the stage is of two types: (i) work-in-process inventory and (ii) end-item inventory (also known as *safety stock*). At the end of each time period, the work-in-process inventory that has been processed is transferred to the end-item inventory. The demand for the item in period t is D_t and this demand is satisfied from the end-item inventory at the beginning of period t. The demand process for the item is assumed to be independent and normally distributed with mean $\mathrm{E}[D_t] = \mu$ and variance $\mathrm{Var}[D_t] = \sigma^2$. Any excess demand, that cannot be met from the end-item inventory, is assumed to be backlogged.

Whenever an amount of end-item inventory is used to meet the demand D_t in period t, an identical amount R_t is released to the work-in-process inventory instantaneously, i.e., $R_t = D_t$. Thus, the total of the work-in-process inventory and end-item inventory is kept constant throughout. Let I_t be the amount of end-item inventory at the beginning of the period t, after the demand for the period t has been met. In

other words, the end-item inventory is measured just after the demand for the period has been satisfied. Similarly, let X_t be the work-in-process inventory at the start of period t just after the amount R_t is released into the system. P_t is defined as the amount of work that is processed in period t and this amount becomes available at the end of the period. At the beginning of period t, the amount of work-in-process X_t is known. Based on X_t, the production rate P_t is set and this amount becomes available at the end of the period. We now have the following flow balance conditions:

$$X_t = X_{t-1} + R_t - P_{t-1} \tag{4.1}$$

$$I_t = I_{t-1} + P_{t-1} - D_t \tag{4.2}$$

Since $R_t = D_t$, by adding equations (4.1) and (4.2), we get

$$\begin{aligned} X_t + I_t &= X_{t-1} + I_{t-1} \\ &= B \qquad \text{(constant)}. \end{aligned} \tag{4.3}$$

Now, in period t, the production quantity P_t is set equal to a fraction α of the WIP inventory, X_t, in that period, i.e.,

$$P_{t-1} = \alpha X_{t-1}, \qquad 0 \leq \alpha \leq 1. \tag{4.4}$$

For a justification of this rule, see Graves [Gra86,Gra88b]. From equations (4.1) and (4.4), and using the fact that $R_t = D_t$, we get

$$\begin{aligned} X_t &= (1-\alpha)X_{t-1} + D_t \\ &= \sum_{s=0}^{\infty} (1-\alpha)^s D_{t-s} \end{aligned} \tag{4.5}$$

assuming that the process has an infinite history. Since the demands in the different periods are iid, we get from equation (4.5),

$$E\left[X_t\right] = \frac{\mu}{\alpha} \tag{4.6}$$

and

$$\begin{aligned} Var\left[X_t\right] &= \frac{Var[D_t]}{2\alpha - \alpha^2} \\ &= \frac{\sigma^2}{2\alpha - \alpha^2} \end{aligned} \tag{4.7}$$

Since $P_t = \alpha X_t$, it follows that

$$E[P_t] = \mu \tag{4.8}$$

and

$$Var[P_t] = \frac{\alpha^2\sigma^2}{2\alpha - \alpha^2} \tag{4.9}$$

From equation (4.3), we see that

$$Var(X_t) = Var(I_t) \qquad \text{since } X_t + I_t = B.$$

From the above analysis, we see that the mean and variance of the work-in-process inventory can be determined. We also know the variance of the end-item inventory, $Var(I_t)$. The only thing that remains is to set the mean level of safety stock, $E[I_t]$. It is desired that the level of safety stock be set so that 95% of the time, the demand can be met from the safety stock. That is,

$$P[I_t > 0] = 0.95$$

Since the demand D_t is assumed to be normally distributed, from equation (4.5), we see that the work-in-process inventory has a normal distribution. Since $X_t + I_t =$ constant, the safety stock I_t also has a normal distribution. Since I_t has a normal distribution, in order to satisfy the condition in equation (4.10), we have to set

$$\begin{aligned} E[I_t] &= k_{0.95}\sqrt{\text{Var}(I_t)} \\ &= 1.645\frac{\sigma}{\sqrt{2\alpha - \alpha^2}} \end{aligned}$$

where $k_{0.95} = 1.645$ is the service factor corresponding to the 95% percentile of the normal distribution. In other words, only 5% of the distribution of I_t lies to the left of the point zero. Therefore, the base stock level, B, is set equal to

$$\begin{aligned} B &= E[X_t] + E[I_t] \\ &= \frac{\mu}{\alpha} + 1.645\frac{\sigma}{\sqrt{2\alpha - \alpha^2}} \end{aligned} \tag{4.10}$$

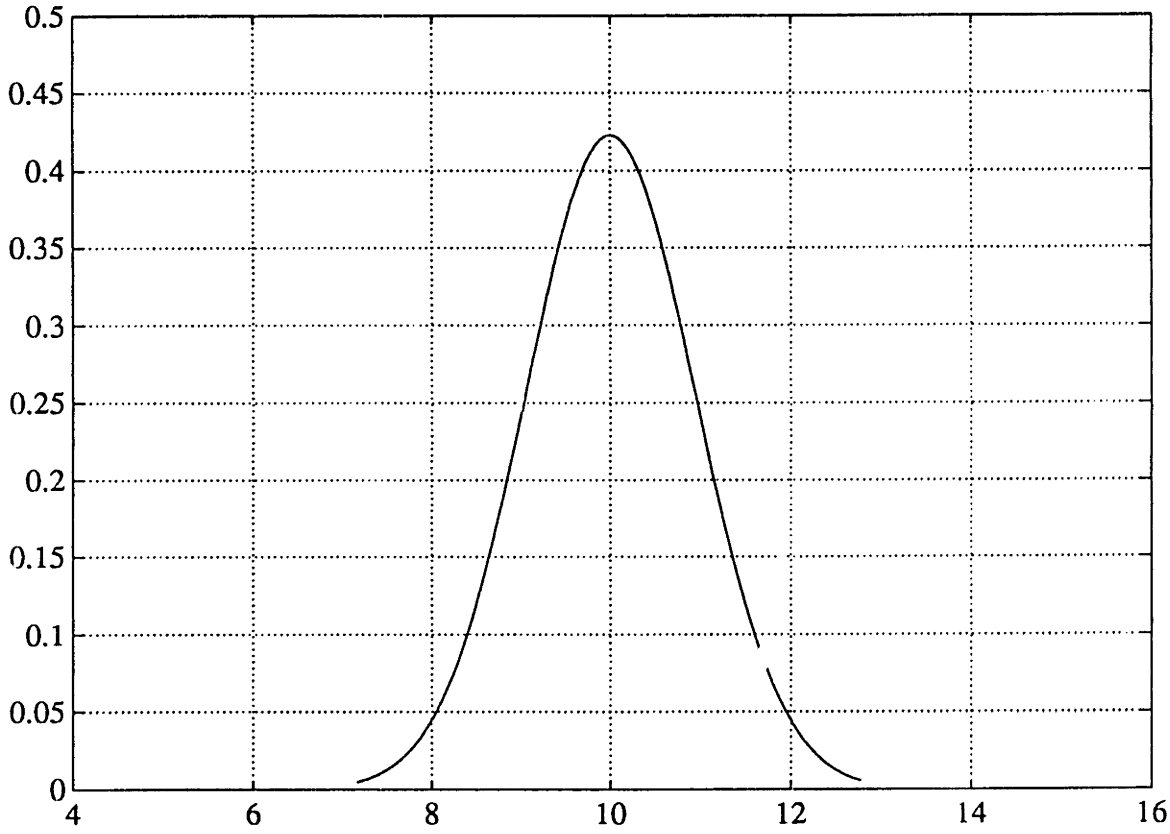

Figure 4.2: The Production Density Function for the Linear Rule

$\alpha = 0.18.$
$\mu = 10, C = 12.$

The model is extremely appealing because of its simplicity. Since in each period, the station processes a fraction α of the WIP, on average, it will take $\frac{1}{\alpha}$ periods for an item entering the WIP to be processed. Hence, the lead time at the workcenter is equal to $\frac{1}{\alpha}$. This turns out to be a useful tool for tactical planning in setting the lead times. One might argue that the production capacity at each stage should be unbounded, because theoretically, the workload could get extremely large and the production is a linear function of the workload. If C is the capacity of the stage, the idea is to set the parameter α so that the production rate is consistent with the maximum capacity rate, say, 95% of the time, i.e.,

$$E\left[P_t\right] + k_{0.95}\sqrt{\mathrm{Var}(P_t)} \;\leq\; C \tag{4.11}$$

where $k_{0.95}$ is the service factor corresponding to the 95% service level. (See Figure 4.2). Note that since $P_t = \alpha X_t$, P_t has a normal distribution. Decreasing the value of α lowers the variance of the production, but it leads to higher WIP levels. We shall now formulate the model in continuous time and compare the results with the discrete case.

4.2.2 The continuous time model: the Ornstein-Uhlenbeck process

The scenario here is identical to the case considered earlier, except that we use the continuous time framework. The cumulative demand process upto time t is assumed to be normally distributed with mean μt and variance $\sigma^2 t$ for all $t > 0$, and the demands over nonoverlapping time intervals are independent. In other words, the cumulative demand process is a diffusion process with drift $\mu > 0$ and dispersion $\sigma > 0$. Therefore, D_t satisfies the stochastic differential equation

$$\begin{aligned} dD_t &= \mu dt + \sigma dW_t \qquad\qquad D_0 = 0. \\ &= \text{Demand in the time interval } (t, t+dt) \end{aligned} \tag{4.12}$$

If X_t is the WIP inventory at time t, then the production rate at time t, $P(X_t, t)$, is set equal to αX_t, i.e.,

$$P(X_t, t) = \alpha X_t, \qquad 0 \leq \alpha \leq 1. \tag{4.13}$$

The flow balance equation in continuous time can be written as follows:

Change in WIP in $(t, t+dt)$	=	Amount released into the system in time $(t, t+dt)$	−	Amount processed in time $(t, t+dt)$

Note that $P(X_t, t)$ is a time homogeneous function since it depends only on the WIP at time t. Henceforth, we shall denote $P(X_t, t)$ by $P(X_t)$. Using equations (4.12) and (4.13), we can rewrite the flow balance equations as

$$\begin{aligned} dX_t &= dD_t - P(X_t)dt \\ &= \mu dt + \sigma dW_t - \alpha X_t dt \\ &= (\mu - \alpha X_t)\, dt + \sigma dW_t, \qquad \mu > 0,\ \sigma > 0,\ \alpha \in (0,1) \end{aligned} \tag{4.14}$$

The diffusion process in equation (4.14) is known as the Ornstein-Uhlenbeck process with drift function $a(x) = \mu - \alpha x$ and dispersion coefficient $b(x) = \sigma$. (For more details, see Karatzas and Shreve [Kara88], p. 358). As the process moves to the right of the point $\frac{\mu}{\alpha}$ on the real line, the drift of the process becomes negative and it *pulls* it towards the point $\frac{\mu}{\alpha}$. Conversely, when X_t moves to the left of $\frac{\mu}{\alpha}$, the drift becomes positive and this restores X_t towards $\frac{\mu}{\alpha}$. The state space of X_t is $(-\infty, \infty)$. It is a well known result in the theory of diffusion processes that X_t has a stationary distribution that is Gaussian with mean $\frac{\mu}{\alpha}$ and variance $\frac{\sigma^2}{2\alpha}$. Therefore, we get

$$E[X_t] = \frac{\mu}{\alpha} \tag{4.15}$$

and

$$Var[X_t] = \frac{\sigma^2}{2\alpha} \tag{4.16}$$

Using an analysis similar to the one used in the discrete time model, the base stock level B for the continuous time model is set equal to

$$B = \frac{\mu}{\alpha} + 1.645\frac{\sigma}{\sqrt{2\alpha}} \tag{4.17}$$

If we compare these results with the results obtained in the discrete case (equations (4.6) and (4.7)), we see that for the steady state WIP, the expected values are identical, but the standard deviation differs by a factor of $\frac{1}{\sqrt{1-\frac{\alpha}{2}}}$. While this factor could be as high as $\sqrt{2}$ (when $\alpha = 1$), we argue that for *typical* values of α, the values are fairly close. (If we assume that α is uniformly distributed in (0, 1), then the mean value of this factor is 1.17). Moreover, even though the factor is largest when $\alpha = 1$, the actual amounts of base level stock at this level are much smaller. If one were to plot a graph of B versus α for the two models, they are extremely close (see Figure 4.3). The point that we are trying to make is that both models provide values for the mean and standard deviation that are close and exhibit the same behavior as a function of α.

4.3 The Gamma Model: a diffusion model with capacity constraints

The models presented in the previous section do not explicity consider a capacity constraint while setting the lead time α. Rather, the idea is to select a value of α so that the production rate is less than the maximum capacity a large fraction of the time. In this section, we present *the gamma model* where the production capacity is explicitly considered while setting the production rate.

Let C be the maximum production capacity rate at the stage and let $\lambda > 0$ be a constant. If the WIP inventory at time t is X_t, then the production rate at time t, $P(X_t)$ is defined as

$$P(X_t) = \frac{\lambda C X_t}{1 + \lambda X_t} \tag{4.18}$$

$$= C - \frac{C}{1 + \lambda X_t} \tag{4.19}$$

Let us take a closer look at equation (4.19). When the system is empty, $X_t = 0$ and the production rate $P(0) = C - C = 0$. On the other hand, when X_t becomes very large, the second term becomes small and the production rate approaches the maximum rate C. The parameter λ determines how fast the production rate reaches C as a function of the workload.

By parameterizing the rule in terms of λ, we are able to subsume a broad class of rules. The plot of Figure 4.4 shows a graph of $P(x)$ versus x for different values of λ. Note that as $\lambda \to \infty$, $P(x)$ approaches a step function with a jump of C at the point 0.

4.3.1 Interpretation of λ

From Figure 4.4, it can be seen that λ determines how quickly the production rate approaches the maximum capacity C at the workcenter. As the value of λ increases, the production rate reaches the maximum rate more quickly. We call λ the startup parameter of the workcenter. From equation (4.19), we can show that

$$\lambda = -\frac{1}{2}\frac{P''(0)}{P'(0)} \tag{4.20}$$

λ is a measure of the curvature of the production function at the point 0. We shall look at special cases of the model and the corresponding values of λ in section 4.5.

4.3.2 Flow balance equations

We can write an equation similar to the one obtained in equation (4.14) using the production rule that we have proposed in equation (4.19). The flow balance equation can be written as

$$\begin{aligned} dX_t &= dD_t - P(X_t)dt \\ &= \mu dt + \sigma dW_t - \left[C - \frac{C}{1+\lambda X_t}\right] dt \qquad \text{from equation (4.19).} \\ &= \left[(\mu - C) + \frac{C}{1+\lambda X_t}\right] dt + \sigma dW_t \end{aligned} \tag{4.21}$$

Assumptions A1 : Let $X_0 = 0$, $\mu > 0$, $C > 0$, $\sigma > 0$, $\mu < C$ and $0 < \lambda < \frac{2C}{\sigma^2}$.

The behavior of the WIP inventory is governed by the stochastic differential equation in (4.21) and we would like to study the process X_t in steady state. We claim that if the parameters satisfy assumptions $A1$, then X_t has an ergodic distribution. Intuitively, this means that if the maximum processing rate C is strictly greater that the mean arrival rate μ, then the WIP inventory in the system will not explode to infinity. We need to place an upper bound on the value of λ in order to ensure that the process X_t has an ergodic distribution. (For a more detailed discussion, please refer to the appendix.)

In the next section, we state the main results of the paper. The proofs leading to the results are presented in the appendix and the interested reader can refer to them there.

4.4 The Steady State Process

In this section, we obtain expressions for the ergodic density, the ergodic mean and variance of the process X_t. We shall also obtain an expression for the ergodic mean and variance of the production process given in equation (4.18).

Theorem 8 *The stationary density function for X_t is given by*

$$f_X(x) = K(1+\lambda x)^{\frac{2C}{\lambda\sigma^2}} \, exp\left[\frac{2(\mu - C)}{\sigma^2}x\right] \qquad x \in \left(-\frac{1}{\lambda}, \infty\right). \tag{4.22}$$

where K is the normalizing constant so that $f_X(x)$ integrates to 1 in the interval $\left(-\frac{1}{\lambda}, \infty\right)$ given by

$$K = \frac{\lambda\left(\frac{2(C-\mu)}{\lambda\sigma^2}\right)^{\left(\frac{2C}{\lambda\sigma^2}+1\right)} \, exp\left(\frac{2(\mu-C)}{\lambda\sigma^2}\right)}{\Gamma\left(\frac{2C}{\lambda\sigma^2}+1\right)} \tag{4.23}$$

and Γ *is the Eulerian gamma function given by*

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$$

Proof: See the appendix. □

The density function in equation (4.22) is a shifted gamma density function (See Figure 4.5). Therefore, we can obtain a closed form expression for the mean and the variance of X_t in steady state.

Corollary 9 *The mean and variance of the diffusion process in steady state are*

$$\begin{aligned} E[X] &= \lim_{t\to\infty} E[X_t] \\ &= \frac{1}{\lambda}\frac{\mu}{C-\mu} + \frac{\sigma^2}{2(C-\mu)} \end{aligned} \tag{4.24}$$

and

$$Var[X] = \lim_{t\to\infty} Var[X_t] \tag{4.25}$$

$$= \frac{1}{\lambda}\frac{C\sigma^2}{2(C-\mu)^2} + \frac{\sigma^4}{4(C-\mu)^2} \tag{4.26}$$

Proof: Let $Z = 1+\lambda X$. By a change of variables, the density function for Z is given by

$$f_Z(z) = \frac{1}{\Gamma\left(\frac{2C}{\lambda\sigma^2}+1\right)}\left(\frac{2(C-\mu)}{\lambda\sigma^2}\right)^{\left(\frac{2C}{\lambda\sigma^2}+1\right)} z^{\frac{2C}{\lambda\sigma^2}} exp\left[\frac{2(\mu-C)}{\lambda\sigma^2}z\right] \tag{4.27}$$

Z has a gamma density function with scale parameter $\gamma = \frac{2C}{\lambda\sigma^2}+1$ and shape parameter $\beta = \frac{2(C-\mu)}{\lambda\sigma^2}$. But for a gamma density, we know that $E\left[Z\right] = \frac{\gamma}{\beta}$ and $Var\left[Z\right] = \frac{\gamma}{\beta^2}$. Using the fact that

$$E[X] = -\frac{1}{\lambda} + \frac{1}{\lambda}E[Z]$$

and

$$Var[X] = \frac{1}{\lambda^2}Var[Z]$$

the result follows. □

Equations (4.24) and (4.26) are expressions for the mean and variance, respectively, of the WIP inventory process in steady state. We would like to see how the different parameters in the model affect the mean and the variance of X. As the arrivals to the system become more unpredictable, σ increases and hence, $E[X]$ and $Var[X]$ increase. As the excess capacity $C - \mu$ increases, we would expect $E[X]$ and $Var[X]$ to decrease, and indeed, it is the case. As λ decreases, the production rate increases less quickly as a function of the workload. Since the production reacts to changes in the workload less quickly, the mean and variance of the WIP inventory increase as λ decreases.

Note that we have derived these results assuming that $\lambda < \frac{2C}{\sigma^2}$. However, if we modify equation (4.21) such that the process is reflected at the origin and the state space of the process is the nonnegative half of the real line, then we can show that this reflected process is ergodic. We contend that for higher values of λ, the ergodic density in equation (4.22) is an excellent approximation for the ergodic density of the reflected process. For further details, please refer to the appendix.

If we look at the expression for the mean WIP inventory $E[X]$ in equation (4.24), we see that there are two terms that contribute to the mean level of WIP. The first term is $\frac{1}{\lambda}\frac{\mu}{C-\mu}$. If we let $\rho = \frac{\mu}{C}$ be the utilization at the stage (ρ is the ratio of the mean production to the maximum capacity), then the first term can be rewritten as $\frac{1}{\lambda}\frac{\rho}{1-\rho}$. As the machine utilization approaches one, the WIP in the system explodes. The second term is $\frac{\sigma^2}{2(C-\mu)}$. This says that as the variability of the demand process increases, the average level of WIP in the system increases as well. In section 4.5, we compare different models and look at the answers that they provide for the mean levels of WIP inventory.

4.4.1 Setting Base Stocks in the Gamma Model

In the continuous time linear model, we had seen that the base stock level B was given by

$$B = \frac{\mu}{\alpha} + 1.645\frac{\sigma}{\sqrt{2\alpha}}$$

Since X_t was normally distributed with mean $\frac{\mu}{\alpha}$ and variance $\frac{\sigma^2}{2\alpha}$ in the linear model, the level of base stock B was set equal to the 95th percentile of X_t. In other words, if $f_X(x)$ is the limiting distribution of X_t in steady state, $\int_{-\infty}^{B} f_X(x)dx = 0.95$. We could do a similar analysis for the gamma model. Again, the idea is to set the mean level for the safety stock so that 95% of the time, the demand can be met from the safety stock. Since $X_t + I_t = B$, we have

$$\begin{aligned} P[I_t > 0] &= P[B - X_t > 0] \\ &= P[X_t < B] \\ &= 0.95 \end{aligned}$$

Therefore, B is set equal to the 95th percentile of the distribution of the WIP inventory. In the linear model, the fraction α of the WIP inventory processed at the production stage is chosen so that the production rate is consistent with the production capacity at the stage. Once the value of α is known, the level of base stock can easily be determined. However, in the gamma model, the production rule considers the capacity constraint explicitly. So what criterion does one use to determine the value of λ for the stage? If the value of λ is known, the level of base stock can be calculated. If λ is within the manager's control, the tradeoff is between higher levels of base stock and a higher variance of production. A higher value of λ means a higher variance of production but lower levels of base stock, and vice versa. A plot of the base stock level B versus λ can be obtained and the appropriate point on the graph could be chosen (See Figure 4.6).

4.5 Special Cases of the Gamma Model

In this section, we look at the linear production rule, the queueing model and the Karmarkar rule and show how all of these can be viewed as special cases of the gamma model.

4.5.1 The Linear Production Rule

The question that we ask is the following: How can we choose the values of λ and C so that the rule in equation (4.18) becomes linear? Suppose we let $\lambda \to 0$ and $C \to \infty$, but they converge at a rate such that $\lambda C = \alpha$. Then, from equation (4.18), we get

$$\begin{aligned} P(X) &= \frac{\lambda C X}{1 + \lambda X} \\ &\to \alpha x \qquad \text{as } \lambda \to 0, C \to \infty \text{ and } \lambda C = \alpha. \end{aligned} \tag{4.28}$$

Now if we look at the ergodic mean of the WIP inventory process in equation (4.24), we get

$$\begin{aligned} \mathrm{E}[X] &= \frac{1}{\lambda}\frac{\mu}{C-\mu} + \frac{\sigma^2}{2(C-\mu)} \\ &\to \frac{\mu}{\alpha} \end{aligned} \tag{4.29}$$

Since μ and σ are finite, then as $C \to \infty$, the second term of equation (4.24) goes to zero. The first term becomes $\frac{\mu}{\alpha}$ since $\lambda\mu \to 0$. We see that this expression is identical to the expression obtained in equation (4.15). Similarly, the variance in equation (4.26) can be simplified as shown below.

$$\begin{aligned} \mathrm{Var}[X] &= \frac{1}{\lambda}\frac{C\sigma^2}{2(C-\mu)^2} + \frac{\sigma^4}{4(C-\mu)^2} \\ &\to \frac{\sigma^2}{2\alpha} \end{aligned} \tag{4.30}$$

The first term becomes $\frac{\sigma^2}{2\alpha}$, while the second term goes to zero. This expression is identical to the one in equation (4.16).

Thus, we have seen that by setting the startup parameter λ very small and setting the capacity to be infinite (but keeping the product of the two quantities constant), we are able to replicate the linear production rule and obtain the exact expression as in the model in section 4.2. Essentially, what we did was to set the curvature of the production very close to zero in order to obtain a good approximation to the linear function. We shall now look at how the model relates to the diffusion approximation to the GI/G/1 queueing model.

4.5.2 The queueing model

Consider a $GI/G/1$ queue that can be approximated by a one-dimensional Brownian Motion. The state space of the Brownian Motion is the nonnegative half of the real line and it is reflected at the origin. Such a process is called a *Reflected Brownian Motion* (RBM). Assume that the drift of the process is $\mu - C < 0$ and the variance is σ^2. (For an excellent introduction to the subject of Reflected Brownian Motion, see Harrison [Har85]).

Whenever the queue length is strictly positive, the system processes the queue at a constant rate C. Work arrives to the queue at a mean rate μ, but since $\mu < C$, the drift is strictly negative. Consequently, the process does not *wander* off to $+\infty$, but eventually returns to the origin. The process is reflected at the point 0, and so, it cannot move to the left of the point 0. Suppose Y_t is the queue length at time t and $Y_0 = 0$. Then, Y_t satisfies the stochastic differential equation

$$dY_t = (\mu - C)dt + \sigma dW_t + d\zeta_t \tag{4.31}$$

where ζ_t is a continuous, nondecreasing process and ζ_t increases over the set $\{t : Y_t = 0\}$. ζ_t is adapted to the smallest σ-algebra with respect to which W_t and Y_t are measurable. It has been shown (see Harrison [Har85]) that if $\mu - C < 0$, then the steady state distribution of Y_t is exponential with mean $\frac{\sigma^2}{2(C-\mu)}$, i.e.,

$$f_Y(y) = \begin{cases} \frac{2(C-\mu)}{\sigma^2} e^{-\frac{\sigma^2}{2(C-\mu)}y} & y > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.32}$$

Therefore, we have for the exponential distribution,

$$\mathrm{E}[Y] = \frac{\sigma^2}{2(C-\mu)} \tag{4.33}$$

and

$$\mathrm{Var}[Y] = \frac{\sigma^4}{4(C-\mu)^2} \tag{4.34}$$

How can the production function in our model be chosen so that it approximates the queueing model? In the queueing model, whenever the queue length is positive, the system operates at a constant rate C which is independent of the workload.

Earlier we had stated that the results in section 4.4 were derived using the fact that $\lambda < \frac{2C}{\sigma^2}$. For values of $\lambda \geq \frac{2C}{\sigma^2}$, we claimed that the ergodic density in equation (4.22) and the ergodic mean and variance in equations (4.24) and (4.26) respectively were excellent approximations. (For a more detailed explanation, please refer to the appendix). In order to see this, suppose we allow λ to approach ∞. Then the production function approaches the maximum rate very quickly and in the limit, it looks like the function

$$P(x) = \begin{cases} C & x > 0 \\ 0 & x = 0 \end{cases} \tag{4.35}$$

Note that the state space in this case is $\lim_{\lambda \to \infty} \left(-\frac{1}{\lambda}, \infty\right) = [0, \infty)$. Letting $\lambda \to \infty$ in equations (4.24) and (4.26), we get

$$E[Y] = \frac{\sigma^2}{2(C-\mu)} \quad \text{and} \quad Var[Y] = \frac{\sigma^4}{4(C-\mu)^2} \tag{4.36}$$

which are identical to equations (4.33) and (4.34) obtained using the RBM approximation to the queueing model.

When we say that the queueing model is a special case of the gamma model, we should qualify this statement with a few remarks. The nature of this model is different from the heavy traffic approximation for queueing networks. The gamma model is tactical in nature, whereas the queueing models are more detailed in that they can be used to describe the system behavior at the operational level. Also, the time

scales for the two models are different. The gamma model directly assumes that the cumulative demand is a Brownian motion, whereas the heavy traffic approximation result is obtained only after a rescaling of the axes.

In the previous two subsections, we have seen how the queueing model and the linear production rule can be viewed as limiting cases of the gamma model. The linear production rule provides an expression for the mean and variance of the WIP inventory that are the first terms in equations (4.24) and (4.26) respectively. The expressions provided by the queueing model are the second terms in the corresponding equations. Both models have different levels of WIP inventory than the gamma model, but for different reasons. The linear production rule, by 'ignoring' the capacity constraint, overestimates the production output at higher levels of WIP inventory. On the other hand, the queueing model assumes that the system always operates at full capacity when the queue length is positive. And so it overestimates the production rate at lower WIP levels.

4.5.3 Karmarkar's Rule

Karmarkar ([Karm89]) has proposed a *clearing function* of the form

$$P(x) = \frac{x}{1+x}C \tag{4.37}$$

But when we substitute $\lambda = 1$ in equation (4.18), we get the same expression as in equation (4.37).

4.6 Comments on the Model

In this section, we would like to comment briefly on some of the features of the model.

The model assumes that for any time t, the cumulative demand at the stage upto time t is normally distributed with mean μt and variance $\sigma^2 t$. Hence, in this model, the cumulative demand process is not a nondecreasing process over time. Second, the WIP inventory at time t is equal to the difference between the cumulative amount

released into the stage upto time t and the cumulative amount processed at the stage upto time t. (The cumulative amount released into the stage upto time t is equal to the cumulative demand upto time t). Since both of these processes are nondecreasing, the WIP process is a process of bounded variation. Yet, we model the WIP process as a diffusion process whose paths are of unbounded variation. Our argument is that we are not concerned as much with the behavior of the sample paths. In this model, we are not controlling the process in real time. The gamma model is a descriptive model and we are merely setting the different parameters in the model to see how they affect the WIP inventory levels and gain some insights. In order to study the steady state properties, we need to look at the problem in continuous time. The fact that the sample paths are of unbounded variation follows from the assumption that the arrival process is a diffusion process.

The next question that may arise is the following: why are we presenting the model in continuous time? Earlier, in section 4.2.1, we had introduced the linear production rule and the Graves model in discrete time. Subsequently, we had looked at the identical model in continuous time. The reason that we looked at the linear rule in both discrete and continuous time is that this is the *only* case where we can obtain analytical results for the discrete time model. We see that the results that the two models provide are extremely close. We also use this to justify the assumption that the cumulative demand function is a Brownian motion. In the cases when the production function becomes nonlinear, we cannot solve the model in discrete time. Therefore, we need to work in the continuous time domain and use the assumption that the cumulative demand function is a diffusion process.

There is another feature of the model that we would like to discuss. The state space of the process in equation (4.21) extends to negative portions of the real line. In the analysis of the model in section 4.4, we had obtained the results using the fact that the state space of the process X_t is $(-\frac{1}{\lambda}, \infty)$. But this means that the WIP inventory is allowed to become negative. When X_t is negative, the production function $\frac{\lambda C X_t}{1+\lambda X_t}$ is also negative. However, we allow the state space to be the interval $(-\frac{1}{\lambda}, \infty)$ in order

to maintain the analytical tractability of the model, even though negative values for the WIP inventory process and the corresponding production function do not have any physical meaning.

If we restrict the state space of the process to be $[0, \infty)$, the density function of the process has the same shape as the original process over the region $[0, \infty)$, albeit the normalizing constant will be different. To be more specific, consider the process

$$d\tilde{X}_t = \left(\mu - \frac{\lambda C \tilde{X}_t}{1 + \lambda \tilde{X}_t}\right) dt + \sigma dW_t + d\zeta_t, \qquad \tilde{X}_0 \in (0, \infty) \tag{4.38}$$

The process $\tilde{X}_t$ is a Reflected diffusion with the same drift and dispersion coefficients as equation (4.21), but the process is reflected at the origin. The process stays in the nonnegative half of the real line and $\tilde{X}_t \geq 0$ for all t. W_t is a standard 1-dimensional Brownian motion and ζ_t is a continuous, nondecreasing process adapted to the smallest σ-algebra with respect to which W_t and X_t are measurable. ζ_t increases only over the set $\{t : \tilde{X}_t = 0\}$ and this is a set of Lebesgue measure zero. Then it can be shown that the ergodic density for the process $\tilde{X}_t$ is

$$f_{\tilde{X}}(x) = \bar{K}(1 + \lambda x)^{\frac{2C}{\lambda\sigma^2}} exp\left[\frac{2(\mu - C)x}{\sigma^2}\right] \qquad x \in [0, \infty) \tag{4.39}$$

where

$$\bar{K} = \int_0^\infty (1 + \lambda x)^{\frac{2C}{\lambda\sigma^2}} exp\left[\frac{2(\mu - C)x}{\sigma^2}\right] dx \tag{4.40}$$

In this case, the normalizing constant is obtained by integrating $(1 + \lambda x)^{\frac{2C}{\lambda\sigma^2}}$ $exp\left[\frac{2(\mu-C)x}{\sigma^2}\right]$ over the interval $[0, \infty)$. But now, we cannot get a nice closed form expression for $\bar{K}$, the ergodic mean $\lim_{t\to\infty} E[\tilde{X}_t]$ and the ergodic variance $\lim_{t\to\infty} Var[\tilde{X}_t]$.

For values of $\lambda \geq \frac{2C}{\sigma^2}$, we show that the process given in equation (4.38) is ergodic and then argue that the ergodic density given by equation (4.39) is well approximated by the ergodic density given by equation (4.21). For further details, see the appendix.

For the continuous time Ornstein-Uhlenbeck process discussed in section 4.2.2, the state space of the process is $(-\infty, \infty)$. Even in this process, X_t drifts into the negative real line and the corresponding production function αX_t is negative (which,

again, has no physical meaning). But we saw that the results for this model were consistent with the results obtained for the discrete time model (in section 4.2.1). In both these cases, we can easily determine the probability that X_t is less than zero in steady state and typically, this probability is very small.

The fact that the state space of the process includes portions of the negative real line permits analytical tractability of the model. This is also true of many models that assume that the demand process has a normal distribution. Yet all these models are useful, since they give good insights into the behavior of the systems being modeled.

4.7 Concluding remarks

In this paper, we have developed a model for WIP inventories for a single workcenter in a manufacturing system. The model has four parameters. We obtain an expression for the mean and variance of the WIP inventory in steady state by studying the ergodic behavior of a stochastic differential equation. We then compare the results obtained with the results obtained using other models for production systems.

Appendix

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\{\mathcal{F}_t : t \geq 0\}$ be a nondecreasing family of sub-σ-algebras of $\mathcal{F}$: $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$ for all $0 \leq s < t < \infty$. Let $\{\mathcal{F}_t\}$ be right continuous and $\mathcal{F}_0$ contain all the P-null sets. W_t is a one-dimensional Brownian motion with $W_0 = 0$ and assume that W_t is adapted to the filtration $\{\mathcal{F}_t\}$. Consider the stochastic differential equation given by equation (4.21).

$$dX_t = \left[(\mu - C) + \frac{C}{1 + \lambda X_t}\right] dt + \sigma dW_t$$

Also assume that the parameters satisfy *assumptions A1.* The drift of the diffusion process in equation (4.21) is given by

$$a(x) = (\mu - C) + \frac{C}{1 + \lambda x} \tag{4.41}$$

The function $a(x)$, shown in Figure 4.7, has a pole at $-\frac{1}{\lambda}$ and we shall consider this function in the interval $(-\frac{1}{\lambda}, \infty)$. The drift is a strictly decreasing function and since $\mu < C$, it is negative for large values of x. The process drifts back towards the origin when the value of X_t get very large since the production is close to the maximum capacity. When the value of X_t is negative, the drift of the process becomes very large and this forces the process X_t to the positive half of the real line.

The dispersion coefficient $b(x)$ is equal to σ. The scale function $h(x)$ is defined by

$$\begin{aligned} h(x) &= \int_0^x exp\left[-2\int_0^y \frac{a(z)}{b^2(z)}dz\right] dy \\ &= \int_0^x exp\left[-\frac{2}{\sigma^2}\int_0^y \left((\mu - C) + \frac{C}{1 + \lambda y}\right) dy\right] dx \end{aligned} \tag{4.42}$$

where we obtained equation (4.42) by substituting the expressions for $a(z)$ and $b(z)$ defined earlier. (For more details about the scale function and its properties, see Karlin and Taylor [Karl81].) Differentiating both sides of equation (4.42) with respect to x, we get

$$h'(x) = exp\left[-\frac{2}{\sigma^2}\int_0^x \left((\mu - C) + \frac{C}{1 + \lambda y}\right) dy\right]$$

$$
\begin{aligned}
&= exp\left[-\frac{2}{\sigma^2}(\mu - C)x - \frac{2C}{\lambda\sigma^2} ln(1+\lambda x)\right] \\
&= exp\left[-\frac{2(\mu - C)x}{\sigma^2} + ln(1+\lambda x)^{-\frac{2C}{\lambda\sigma^2}}\right] \\
&= (1+\lambda x)^{-\frac{2C}{\lambda\sigma^2}} exp\left[-\frac{2(\mu - C)}{\sigma^2}x\right] \qquad (4.43)
\end{aligned}
$$

The derivative of the scale function $h'(x)$ is continuous and strictly positive. It can easily be verified that

$$h''(x) = -\frac{2a(x)}{\sigma^2}h'(x) \qquad (4.44)$$

Under assumptions A1, if the starting point $X_0 \in (0,\infty)$, then the state space of the process X_t is $\left(-\frac{1}{\lambda},\infty\right)$ for all $t \in (0,\infty)$. The function h evaluated at the two end points, $-\frac{1}{\lambda}$ and ∞, is equal to $-\infty$ and $+\infty$ respectively, i.e.,

$$h\left(-\frac{1}{\lambda}\right) = -\infty$$

and

$$h(\infty) = \infty$$

h is a strictly increasing function and it maps the state space $\left(-\frac{1}{\lambda},\infty\right)$ into $(-\infty,\infty)$. It has a continuously differentiable inverse $q : (-\infty,\infty) \rightarrow \left(-\frac{1}{\lambda},\infty\right)$. Since $h(q(x)) = x$ (by definition), differentiating both sides using the chain rule, we get

$$h'[q(x)]\ q'(x) = 1$$

Therefore,

$$h'(q(x)) = \frac{1}{q'(x)} \qquad (4.45)$$

Now, define the function $\tilde{\sigma} : (-\infty,\infty) \rightarrow (0,\infty)$ as follows.

$$
\begin{aligned}
\tilde{\sigma}(x) &= b(q(x))\ h'(q(x)) \\
&= \sigma\ h'(q(x)) \qquad \text{since } b(x) = \sigma \text{ (constant)} \qquad (4.46) \\
&= \frac{\sigma}{q'(x)} \qquad \text{from equation (4.45).} \qquad (4.47)
\end{aligned}
$$

Let $Y_t = h(X_t)$. Since h has a well defined inverse, for any given sample path $\{Y(t,\omega), t \geq 0\}$ where $\omega \in \Omega$, the corresponding sample path $\{X(t,\omega), t \geq 0\}$ can be uniquely determined. In the analysis that follows, what we are going to do is the following: We first transform the process X_t using the scale function so as to eliminate the drift term. We then show that the transformed process $Y_t = h(X_t)$ has an ergodic distribution under certain conditions. We then obtain the ergodic distribution for X_t using the change of variable $X_t = q(Y_t)$. We shall now study the process Y_t defined above.

Lemma 10 *Let assumptions A1 be satisfied and let $\{X_t, t \geq 0\}$ be a solution to the equation*

$$dX_t = \left(\mu - C + \frac{C}{1+\lambda X_t}\right) dt + \sigma dW_t$$

Then $Y_t = h(X_t)$ satisfies the equation

$$dY_t = \tilde{\sigma}(Y_t) dW_t \tag{4.48}$$

Proof: Using the generalized Itô formula (see Karatzas and Shreve [Kara88], p. 219), we get

$$\begin{aligned}
dh(X_t) &= h'(X_t)dX_t + \frac{1}{2}\sigma^2 h''(X_t)dt \\
&= h'(X_t)dX_t - a(X_t)h'(X_t)dt && \text{from equation (4.44)} \\
&= h'(X_t)\left[a(X_t)dt + \sigma dW_t\right] - a(X_t)h'(X_t)dt && \text{from equation (4.21)} \\
&= \sigma h'(X_t)dW_t \\
&= \tilde{\sigma}(h(X_t))dW_t && \text{from equation (4.46)}
\end{aligned}$$

Since $Y_t = h(X_t)$, the result follows. □

Note that equation (4.48) does not have a drift term. We would like to show that the process Y_t defined by equation (4.48) is ergodic. If we can show that the process Y_t is ergodic, then we can use a change of variables to derive the ergodic distribution for the process X_t. To show the ergodicity of Y_t, we need the following theorem.

Theorem 11 (Skorohod) *Suppose that* $r_1 = h(-\frac{1}{\lambda}) = -\infty$, $r_2 = h(\infty) = +\infty$ *and* $\int_{-\infty}^{\infty} [\tilde{\sigma}(y)]^{-2}\, dy < \infty$. *Then the process* Y_t *is ergodic with state space* $(-\infty, \infty)$ *and has ergodic distribution*

$$\pi(A) \;=\; k' \int_A [\tilde{\sigma}(y)]^{-2}\, dy \tag{4.49}$$

where $A \subset R$ *is any Borel set.*

Proof: See Skorohod [Skor89]. □

To show that the process Y_t is ergodic, we have to show that the conditions of the above theorem are satisfied. Recall that the values of r_1 and r_2 are simply the values of the scale function h evaluated at the two endpoints of the interval $\left(-\frac{1}{\lambda}, \infty\right)$.

Theorem 12 *The process* Y_t *satisfying the equation*

$$dY_t \;=\; \tilde{\sigma}(Y_t)\, dW_t$$

is ergodic.

Proof: We have to check that the conditions of the above theorem are satisfied.

$$\begin{aligned}
r_1 &= h\left(-\frac{1}{\lambda}\right) \\
&= \int_0^{-\frac{1}{\lambda}} (1+\lambda x)^{-\frac{2C}{\lambda\sigma^2}} exp\left[-\frac{2(\mu - C)}{\sigma^2} x\right] dx \\
&= -\int_{-\frac{1}{\lambda}}^{0} (1+\lambda x)^{-\frac{2C}{\lambda\sigma^2}} exp\left[-\frac{2(\mu - C)}{\sigma^2} x\right] dx \\
&\leq -exp\left[\frac{2(\mu - C)}{\lambda\sigma^2}\right] \int_{-\frac{1}{\lambda}}^{0} (1+\lambda x)^{-\frac{2C}{\lambda\sigma^2}} dx \\
&= -\infty
\end{aligned}$$

This is the step where we need to use the fact that $\lambda < \frac{2C}{\sigma^2}$. Note that if $\lambda \geq \frac{2C}{\sigma^2}$, then $r_1 > -\infty$ and so the conditions of theorem 11 no longer hold. Therefore, $r_1 = -\infty$. Similarly, it can be shown that

$$\begin{aligned}
r_2 &= h(\infty) \\
&= \int_0^{\infty} (1+\lambda x)^{-\frac{2C}{\lambda\sigma^2}} exp\left[-\frac{2(\mu - C)}{\sigma^2} x\right] dx \\
&= \infty
\end{aligned}$$

since $\mu < C$ and so the coefficient of x in the exponent is positive. Since $\tilde{\sigma}(x) = \frac{\sigma}{q'(x)}$, we see that

$$\int_{-\infty}^{\infty} \left[q'(z)\right]^2 dz < \infty \;\Rightarrow\; \int_{-\infty}^{\infty} \left[\tilde{\sigma}(z)\right]^{-2} dz < \infty$$

Now let $v = q(z)$. Then $dv = q'(z)dz$. Therefore,

$$\begin{aligned}
\int_{-\infty}^{\infty} \left[q'(z)\right]^2 dz &= \int_{-\frac{1}{\lambda}}^{\infty} q'(q^{-1}(v))dv \\
&= \int_{-\frac{1}{\lambda}}^{\infty} \frac{1}{h'(v)} dv \qquad \text{from equation (4.34).} \\
&= \int_{-\frac{1}{\lambda}}^{\infty} (1+\lambda v)^{\frac{2C}{\lambda\sigma^2}} exp\left[\frac{2(\mu - C)}{\sigma^2} v\right] dv < \infty
\end{aligned}$$

The conditions of the above theorem are satisfied. Hence the process Y_t is ergodic. □

Now that we have shown that the process Y_t is ergodic, we can obtain the ergodic distribution for X_t using a change of variables.

Theorem 13 *The stationary density function for X_t is given by*

$$f_X(x) = K(1+\lambda x)^{\frac{2C}{\lambda\sigma^2}} exp\left[\frac{2(\mu - C)}{\sigma^2} x\right] \tag{4.50}$$

or

$$f_X(x) = \frac{\lambda \left(\frac{2(C-\mu)}{\lambda\sigma^2}\right)^{\left(\frac{2C}{\lambda\sigma^2}+1\right)} exp\left(\frac{2(\mu-C)}{\lambda\sigma^2}\right)}{\Gamma\left(\frac{2C}{\lambda\sigma^2}+1\right)} (1+\lambda x)^{\frac{2C}{\lambda\sigma^2}} exp\left[\frac{2(\mu - C)}{\sigma^2} x\right] \tag{4.51}$$

where K is the constant so that $f(x)$ integrates to 1 in the interval $(-\frac{1}{\lambda}, \infty)$.

Proof: By theorem 2, the ergodic density for the process Y_t is given by

$$\begin{aligned}
f_Y(y) &= k' \left[\tilde{\sigma}(y)\right]^{-2} \\
&= \frac{k'}{\sigma^2} \left[q'(y)\right]^2 \qquad \text{from the definition of } \tilde{\sigma}(y).
\end{aligned}$$

The ergodic density for $X_t = q(Y_t)$ is given by

$$\begin{aligned} f_X(x) &= f_Y(y)\frac{1}{\frac{\partial X}{\partial Y}}\Big|_{x=q(y)} \\ &= \frac{k'}{\sigma^2}\frac{\left[q'(y)\right]^2}{q'(y)}\Big|_{x=q(y)} \\ &= \frac{k'}{\sigma^2}q'(y)\Big|_{x=q(y)} \\ &= \frac{k'}{\sigma^2}\frac{1}{h'(q(y))}\Big|_{x=q(y)} \qquad \text{from equation (4.46).} \\ &= \frac{k'}{\sigma^2}\frac{1}{h'(x)} \\ &= K(1+\lambda x)^{\frac{2C}{\lambda\sigma^2}} exp\left[\frac{2(\mu - C)x}{\sigma^2}\right] \end{aligned}$$

Integrating $f_X(x)$ over the interval $\left(-\frac{1}{\lambda}, \infty\right)$, we get the expression for K. □

In the analysis so far, we have been able to show that the WIP inventory process is ergodic for values of $\lambda < \frac{2C}{\sigma^2}$. The intuitive reason for this is the following: As the process approaches the point $-\frac{1}{\lambda}$, the drift of the process becomes very large and consequently, the process moves towards the positive half of the real line away from the point $-\frac{1}{\lambda}$. But as the value of λ increases, the drift function $a(x)$ increases more slowly as one approaches the point $-\frac{1}{\lambda}$. Beyond a certain value $\lambda = \frac{2C}{\sigma^2}$, the drift does not increase quickly enough in order to force the process to the positive half of the real line. Therefore, the process reaches the point $-\frac{1}{\lambda}$ and stays there after a sufficiently large period of time.

So the question still remains: Is the ergodic density function given by equation (4.22) still valid for larger values of λ. We argue that this density function is an excellent approximation to the ergodic density function of a process that is reflected at the origin. To be more specific, consider the process

$$d\tilde{X}_t = \left(\mu - \frac{\lambda C\tilde{X}_t}{1+\lambda\tilde{X}_t}\right)dt + \sigma dW_t + d\zeta_t, \qquad \tilde{X}_0 \in (0,\infty) \qquad (4.52)$$

This process is reflected at the origin and hence, $\tilde{X}_t \geq 0$ for all $t \geq 0$. (We had described this process in the earlier sections). The parameters satisfy assumptions

A1 but now, there is no upper bound on the value of λ, i.e., the only restriction on λ is $\lambda > 0$. Then, it can be shown (see Skorohod [Skor89], p. 55) that the process $\tilde{X}_t$ is ergodic with ergodic density

$$f_{\tilde{X}}(x) = \bar{K}(1+\lambda x)^{\frac{2C}{\lambda\sigma^2}} exp\left[\frac{2(\mu - C)x}{\sigma^2}\right] \qquad x \in [0, \infty) \tag{4.53}$$

where

$$\bar{K} = \int_0^\infty (1+\lambda x)^{\frac{2C}{\lambda\sigma^2}} exp\left[\frac{2(\mu - C)x}{\sigma^2}\right] \tag{4.54}$$

Note that $\bar{K}$ cannot be integrated in closed form. But suppose we modify the value of K so as to include a small portion $\left(-\frac{1}{\lambda}, 0\right)$ of the real line. Then, we can integrate the resulting expression. When the value of λ is large, then the value of $-\frac{1}{\lambda}$ is closer to zero. Therefore,

$$\bar{K} \approx \int_{-\frac{1}{\lambda}}^\infty (1+\lambda x)^{\frac{2C}{\lambda\sigma^2}} exp\left[\frac{2(\mu - C)x}{\sigma^2}\right] \tag{4.55}$$

We also extend the domain of $\tilde{X}$ to include the interval $\left(-\frac{1}{\lambda}, 0\right)$. In section 4.5, we showed that this model compared very well with the standard diffusion approximation to the queueing model. In Figures 4.8 and 4.9, we compare the ergodic density functions for the processes X_t and $\tilde{X}_t$ for two values of λ, $\lambda = 1$ and $\lambda = 3$.

We can also obtain an expression for the mean and variance of the production process in steady state. The result is stated below.

Corollary 14 *The mean and variance of the production process, $P(X_t)$, defined by equation (4.18) in steady state, where X_t satisfies equation (4.21) and assumptions A1 are*

$$\begin{aligned} \lim_{t\to\infty} E[P] &= E[P_t] \\ &= \mu \end{aligned} \tag{4.56}$$

and

$$Var[P] = \lim_{t\to\infty} Var[P_t]$$

$$= \frac{(C-\mu)^2}{\left(\frac{2C}{\lambda\sigma^2}-1\right)} \qquad \lambda < \frac{2C}{\sigma^2} \tag{4.57}$$

Proof: Let $Z_t = 1 + \lambda X_t$. Then

$$\begin{aligned} P(X_t) &= C - \frac{C}{1+\lambda X_t} \\ &= C - \frac{C}{Z_t} \end{aligned}$$

Therefore, $\lim_{t\to\infty} E[P_t] = C - C\lim_{t\to\infty} E[Z_t^{-1}] = C - CE[Z^{-1}]$ and $\lim_{t\to\infty} Var[P_t] = C^2 \lim_{t\to\infty} Var[Z_t^{-1}] = C^2 Var[Z^{-1}]$. The density function for Z is given by equation (4.27). Using the fact that

$$\lim_{t\to\infty} E\left[Z_t^{-2}\right] = \frac{2(C-\mu)^2}{C(2C-\lambda\sigma^2)} \qquad \lambda < \frac{2C}{\sigma^2}$$

the result follows. □

Equations (4.56) and (4.57) are expressions for the ergodic mean and variance, respectively, of the production process. The flow balance condition implies that the mean production rate must equal the mean arrival rate, and hence, $E[P] = \mu$. But how do the other parameters of the model affect the variance of production? As the variance of the arrival process σ^2 increases, $Var[P]$ would increase since the workload process becomes more variable. We see, from equation (4.57) that as the excess capacity $C - \mu$ increases, $Var[P]$ also increases. At first, this may seem counterintuitive, but one way to see this would be as follows: For a fixed value of μ, as $C - \mu$ increases, $P(X_t)$ can take on values over a larger range. Hence, the variance of production would increase. As λ increases, the workcenter is able to react to changes in workload more quickly and so $Var[P]$ increases as λ increases. At the value $\lambda^* = \frac{2C}{\sigma^2}$, Var$[P]$ becomes infinite.

In the above analysis, we had only calculated the expression for Var$[P]$ for values of λ less than $\frac{2C}{\sigma^2}$. Earlier, we had argued that for the WIP inventory process in equation (4.21), the ergodic density in equation (4.22) is an excellent approximation for the ergodic density of the reflected process. Since we have the density function

for the steady state WIP inventory X, we could determine the steady state density function for the production function, $P(X) = \frac{\lambda C X}{1+\lambda X}$. But it turns out that for higher values of λ, the ergodic densities for the production are quite different, depending on whether $P(\cdot)$ is a function of the original process X or the reflected process $\tilde{X}$. Two points need to be mentioned here. First, unlike the case of the linear rule where we needed the expression for $Var[P]$ to choose the appropriate lead time $\frac{1}{\alpha}$, we do not need the expression for $Var[P]$ for this purpose in this model because the capacity constraint is explicitly considered while setting the production rule. Second, this rule is very accurate for lower values of λ. We saw in section 4.5, that the linear model corresponds to the case when $\lambda \to 0$. In this case, we see that the expression on the right hand side of equation (4.57) converges to $\frac{\alpha\sigma^2}{2}$ which is the variance of production for the linear model.

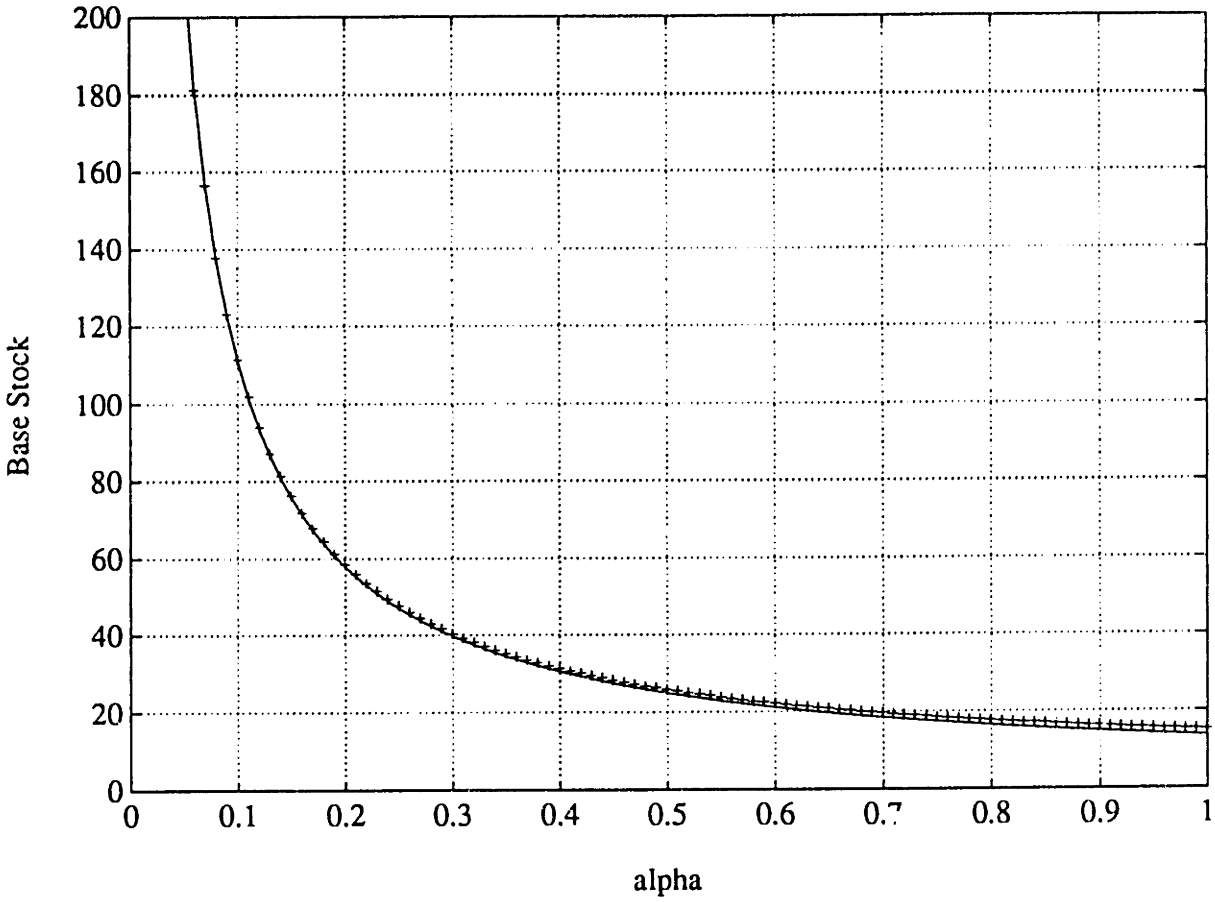

Figure 4.3: Comparing the Base Stock B versus α for the discrete and continuous time models

---------- Discrete time model.
+++++++++ Continuous time model.

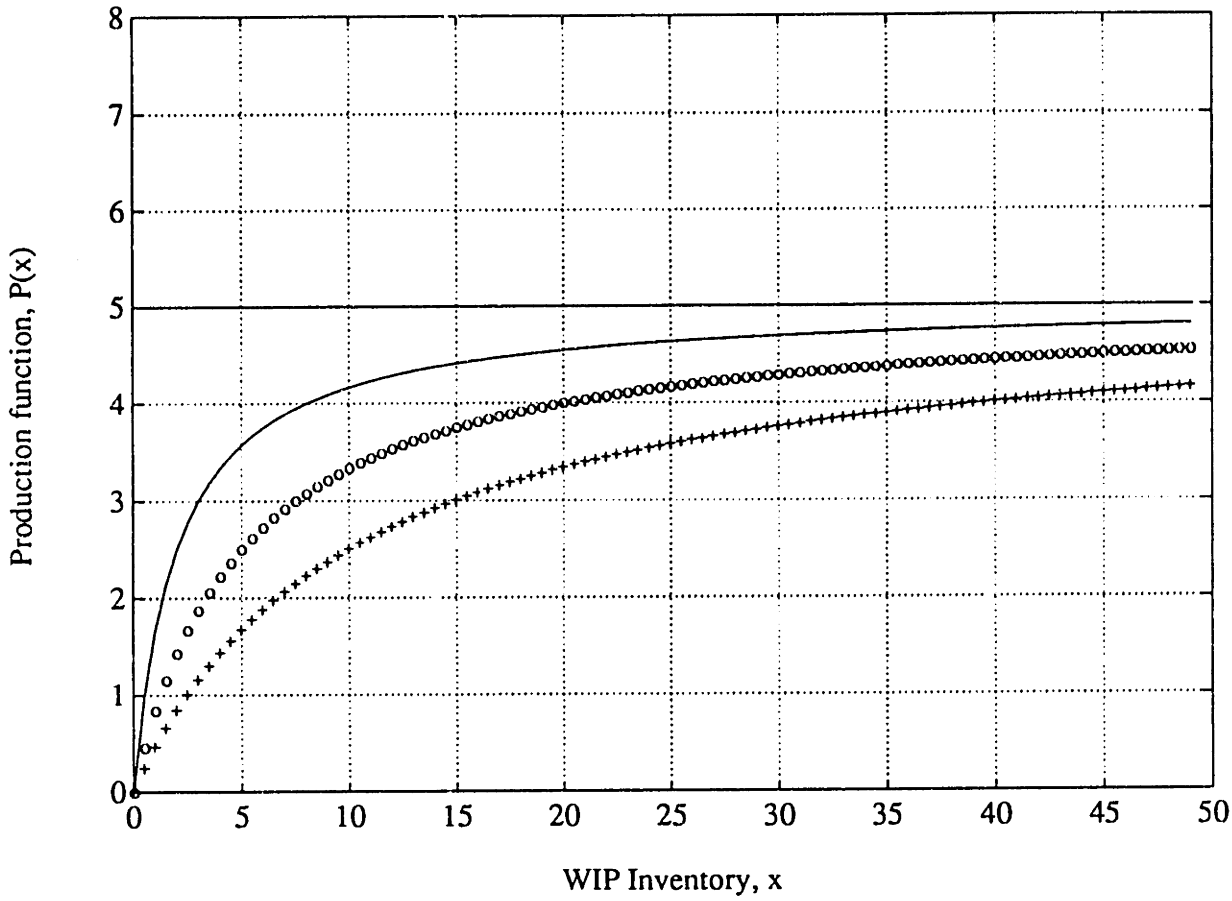

Figure 4.4: The Gamma Production Rule

++++++++++ $\lambda = 0.1$
oooooooooo $\lambda = 0.2$
---------- $\lambda = 0.5$

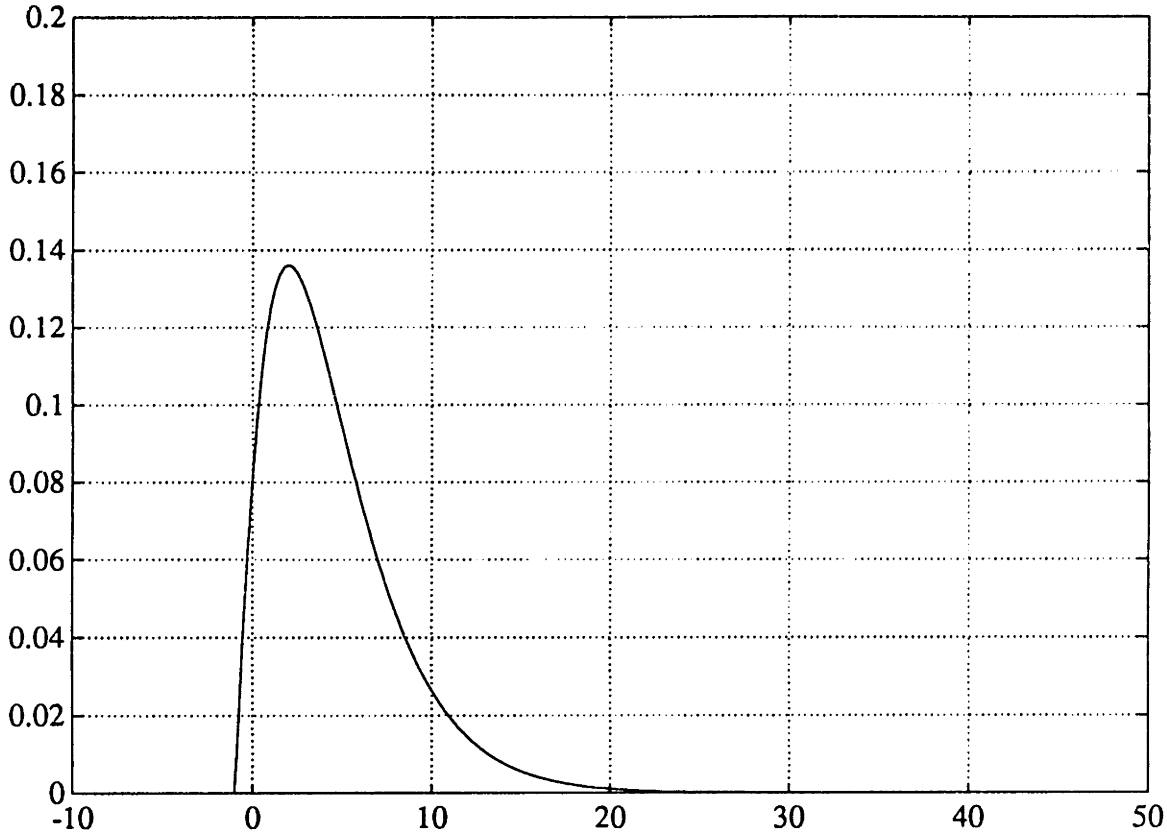

Figure 4.5: The Density Function for the Work-in-Process Inventory

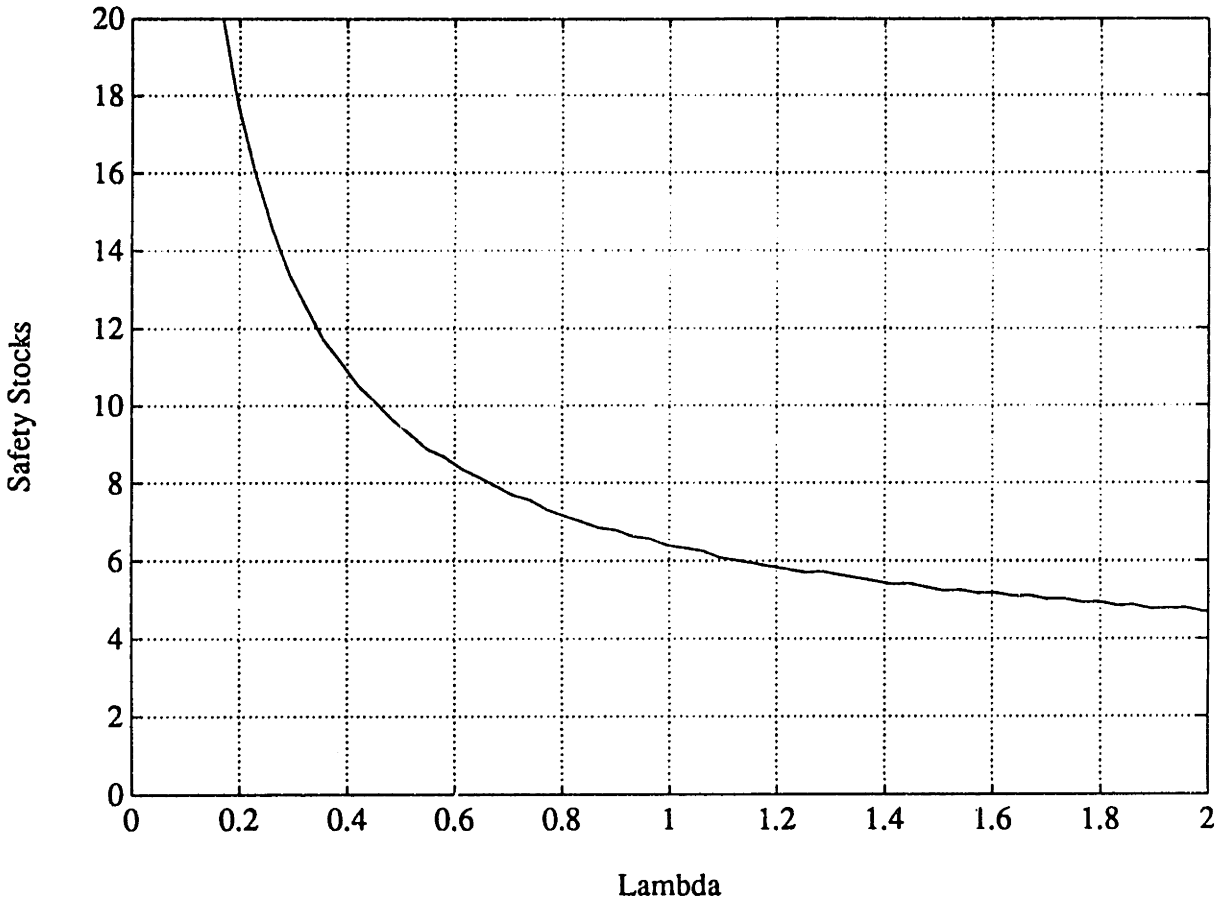

Figure 4.6: Setting Base stocks

$\mu = 10$, $\sigma = 3$, $C = 15$

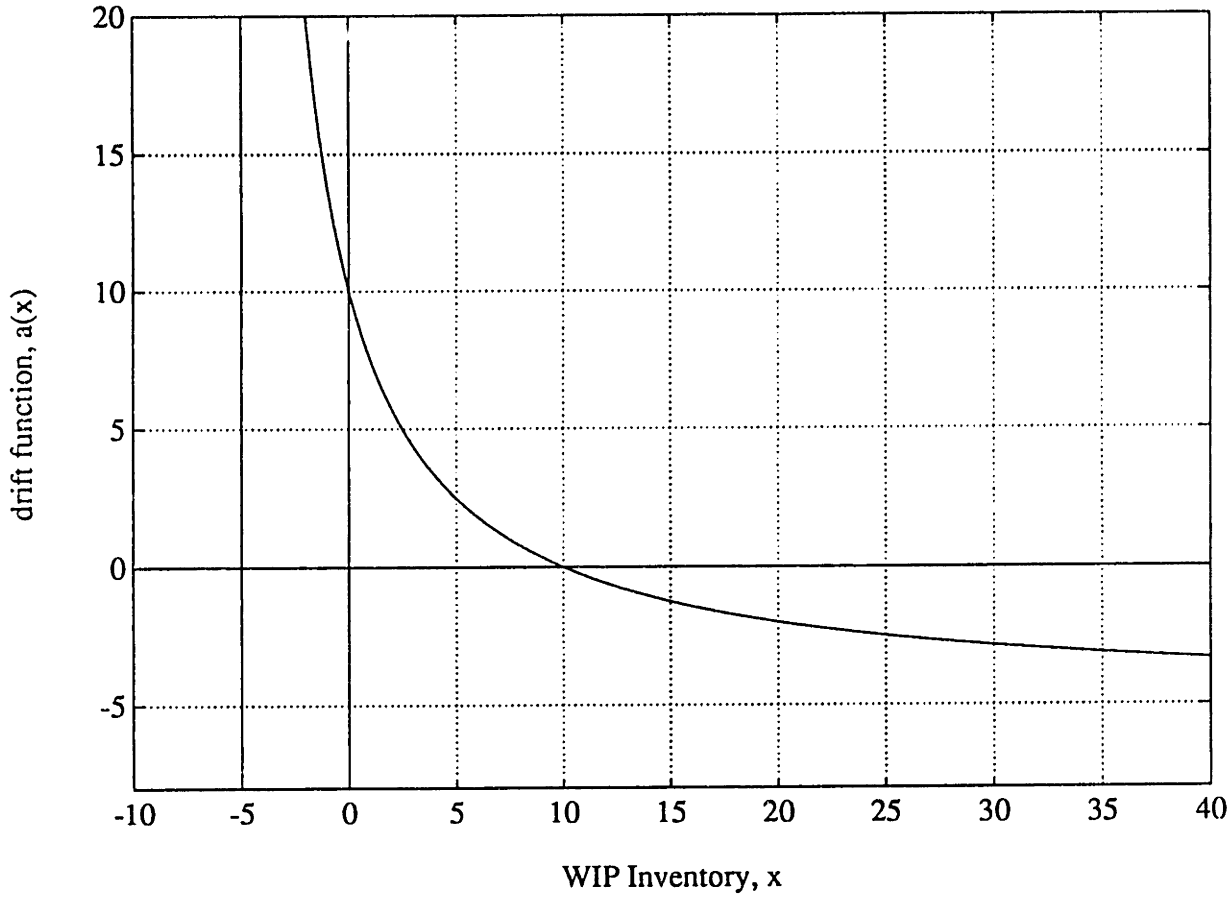

Figure 4.7: The drift function $a(x)$

$\mu = 10$, $C = 15$, and $\lambda = 0.2$

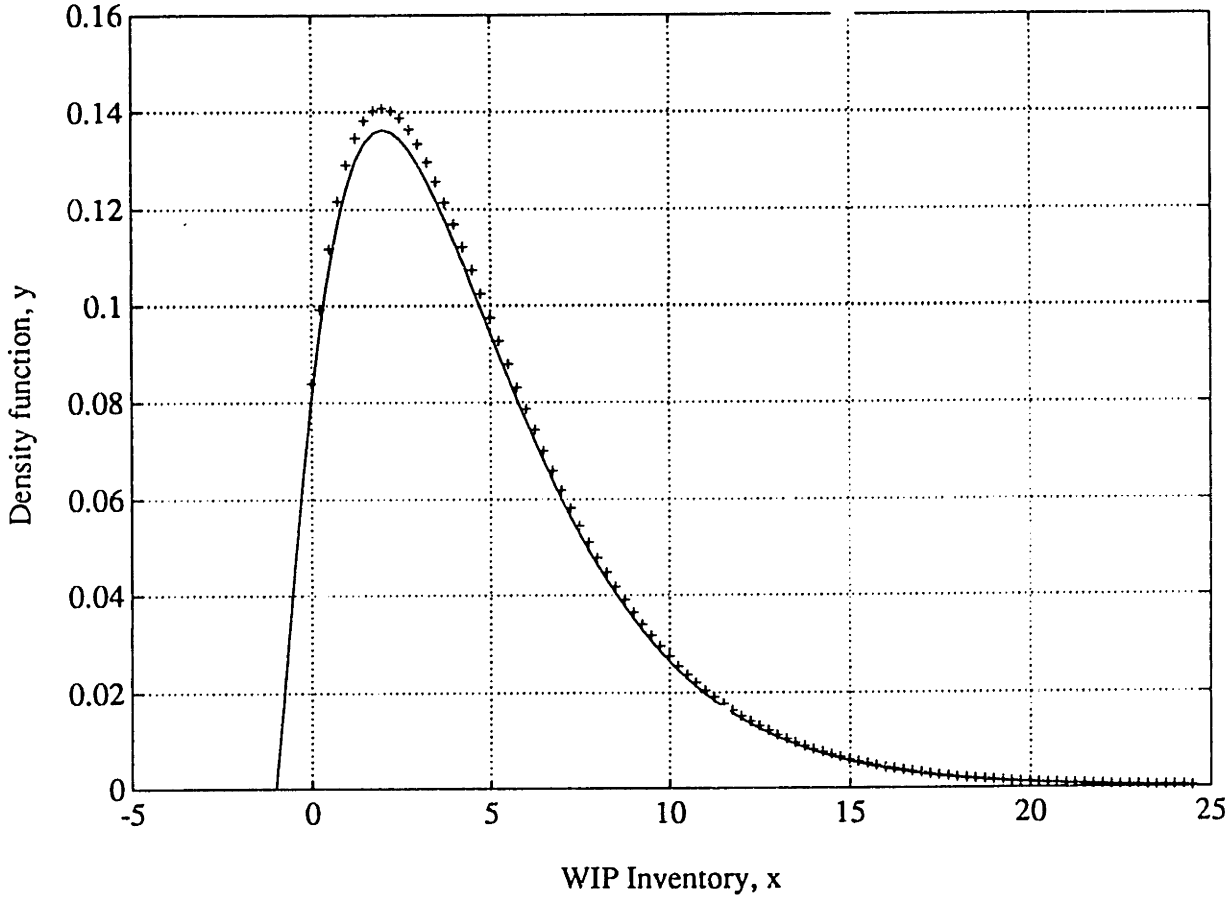

Figure 4.8: The ergodic density functions for the processes X and $\tilde{X}$

$\mu = 10$, $C = 15$, $\sigma = 5$, $\lambda = 1$.
$K = 5.6671$, $\tilde{K} = 5.4467$

---------- Density function for the process X.
+++++++++ Density function for the process $\tilde{X}$

	Process X	Process $\tilde{X}$
Expected Value	4.5	4.689
Variance	13.75	13.0140

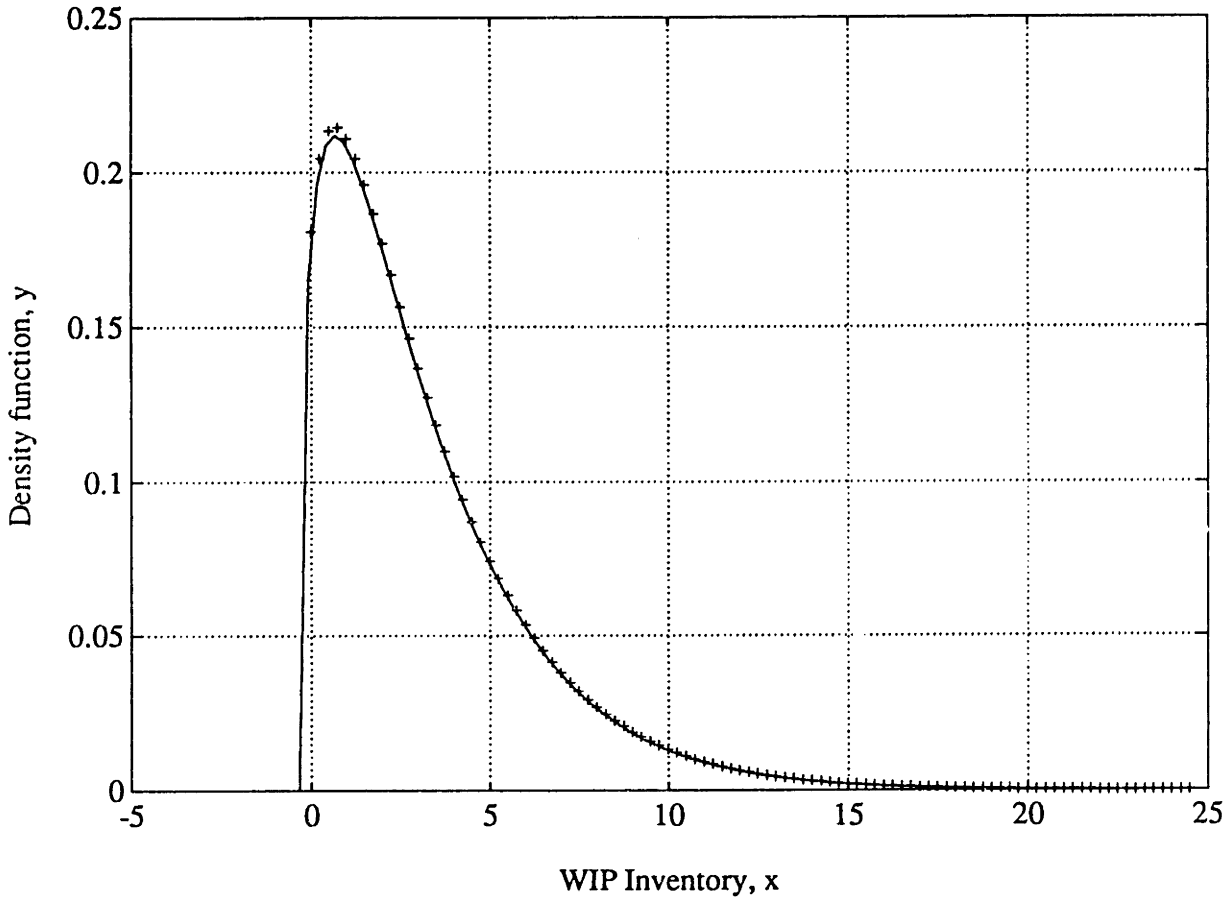

Figure 4.9: The ergodic density functions for the processes X and $\tilde{X}$

$\mu = 10$, $C = 15$, $\sigma = 5$, $\lambda = 3$.
$K = 5.6671$, $\bar{K} = 5.4467$

---------- Density function for the process X.
+++++++++ Density function for the process $\tilde{X}$

	Process X	Process $\tilde{X}$
Expected Value	3.1667	3.3021
Variance	8.75	8.5614

Chapter 5

Conclusions and Directions for Further Research

In this thesis, we have looked at three problems related to capacity allocation, work-in-process inventories, safety stocks and lead times in manufacturing systems. We had used both the discrete and continuous time framework in these models. When we used the continuous time framework, we used Brownian motion as a tool to develop and analyze the models, as in Chapters 2 and 4.

In Chapter 2, we looked at the question of how to allocate capacity among the different stages of a network. The objective was to minimize the total base stock in the system. In this model, we assumed that each stage processes one item. We obtained the results for the optimal capacity allocation and showed how they were related to the results of Kleinrock [Klei64] and Wein [Wein89] in queueing networks.

In Chapter 3, we used the discrete time model to examine the question of flexibility versus dedicated machines for a single stage. As the next step, it would be interesting to see how the system behaves when there is more than one stage in the network. We believe that in order to answer this question, some efficient algorithms and computer codes need to be developed because the standard optimization packages (e.g. MINOS 5.2) did not work well on this problem.

In Chapters 2 and 3, we assumed that the production rule was linear. In Chapter 4, we looked at a nonlinear, capacitated rule for a single stage. The key result was that the expression for WIP consisted of two terms: one term captured the effect of machine utilization and the other term captured the effect of variability in the de-

mand. We showed that the queueing model and the linear rule were asymptotic cases of this capacitated rule, and they each contributed one term of the WIP expression in the capacitated model. It remains to extend this result to the multistage case. A more ambitious task would be to examine the questions raised in Chapters 2 and 3 when the system behaves according to the capacitated rule instead of the linear rule. It would also be of interest to see what the distribution of the lead times is for the different models that we have considered.

Bibliography

[Bitr88] Bitran, G. R., and D. Tirupati, *Trade-off curves, targeting and balancing in queueing networks*, Operations Research, Vol. 37, No. 4 , 1988.

[Buza80] Buzacott, J. A., and J. G. Shanthikumar, *Models for Understanding Flexible Manufacturing Systems,* AIIE Transactions, Vol. 12, No. 4, 1980.

[deGr90] de Groote, X., *The flexibility of production processes: a generalized framework*, Working Paper, The Wharton School, University of Pennsylvania, 1990.

[Elm78] Elmaghraby, S., *The Economic Lot-Scheduling Problem,* Management Science, Vol. 24, No. 6, 1978.

[Fine89] Fine, C. H., and Graves, S. C., *A Tactical Planning Model for Manufacturing Subcomponents of Mainframe Computers*, Journal of Manufacturing and Operations Management, 2:4-34, 1989.

[Gold86] Goldratt, E., and J. Cox, *The Goal*, 1986.

[Gord67] Gordon, W. J., and G. F. Newell, *Closed queueing systems with exponential servers,* Operations Research, Vol. 15, 252–267, 1967.

[Gra86] Graves, S. C., *A Tactical Planning Model for a Job Shop*, Operations Research, July-August, 1986.

[Gra88a] Graves, S. C., *An Extension to a Tactical Planning Model for a Job Shop*, Proceedings of the 27th IEEE Conference on Decision and Control, Austin, TX, 1988a.

[Gra88b] Graves, S. C., *Safety Stocks in Manufacturing Systems*, Journal of Manufacturing and Operations Management, Vol. 1, pp. 67–101, 1988b.

[Har85] Harrison, J. M., *Brownian Motion and Stochastic Flow Systems*, John Wiley and Sons, New York, 1985.

[Har87] Harrison, J. M., and R. J. Williams, *Brownian models of open queueing networks with homogeneous customer populations*, Stochastics, 22:77–115, (1987).

[Har90] Harrison, J. M., R. J. Williams, and H. Chen, *Brownian models of closed queueing networks with homogeneous customer populations*, Stochastics, 29:37–74, (1990).

[Har90] Harrison, J. M., C. A. Holloway and J. M. Patell, *Measuring Delivery Performance: A Case Study from Semiconductor Manufacturing*, Ch. 11, Measures for Manufacturing Excellence, R. S. Kaplan, ed., HBS Press, 1990.

[Holt55] Holt, C. C., F. Modigliani and H. A. Simon, *A Linear Decision Rule for Production and Employment Scheduling,* Management Science, Vol. 2, No. 1, 1955.

[Jack57] Jackson, J. R., *Networks of Waiting Lines*, Operations Research, 5:518–521, 1957.

[Jack63] Jackson, J. R., *Jobshop-like Queueing Systems*, Management Science, 10:131-142, 1963.

[Kapl83] Kaplan, R. S., *Yesterday's accounting undermines production*, Harvard Business Review, July-August 1984.

[Kara88] Karatzas, I., and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[Karl81] Karlin, S., and H. M. Taylor, *A Second Course in Stochastic Processes*, Academic Press, Orlando, FL, 1981.

[Karm89] Karmarkar, U. S., *Capacity Loading and Release Planning with Work-in-Process and Lead Times*, Journal of Manufacturing and Operations Management, Vol. 2, pp. 105–123, 1989.

[Karm90] Karmarkar, U. S., *Lead Times and Production Management: Fleshing out the JIT Mystique*, unpublished.

[Kell79] Kelly, F. P., *Reversibility and Queueing Networks*, John Wiley and Sons, New York, 1979.

[Klei64] Kleinrock, L., *Communications Nets: Stochastic Message Flow and Delay*, Dover Publications, Inc., New York, 1964.

[Kum87] Kumar, V., *Entropic measures of manufacturing flexibility,* Int. J. Prod. Res, Vol. 25, No. 7, 957–966, 1987.

[Lam84] Lambrecht, M. R., J. A. Muckstadt and R. Luyten, *Protective stocks in multi-stage production systems,* Int. J. Prod. Res, Vol. 22, No. 6, 1001–1025, 1984.

[Rose89] Rose, P. S., *The Interstate Banking Revolution,* Quorum Books, New York, 1989.

[Schw89] Schweitzer, P., and A. Seidmann, *Optimizing processing rates for flexible manufacturing systems,* Management Science, Vol. 37, No. 4, 1991.

[Shan48] Shannon, C. E., *A Mathematical Theory of Communication,* Bell System Technical Journal, Vol. 27, 379–423, 633–659, 1948.

[Skor89] Skorohod, A. V., *Asymptotic Methods in the Theory of Stochastic Differential Equations*, American Mathematical Society, Providence, RI, 1989.

[Stec83] Stecke, K., *Formulation and solution of nonlinear integer production planning problems for flexible manufacturing systems,* Management Science, Vol. 29, No. 3, 1983.

[Stec81] Stecke, K., and J. J. Solberg, *Loading and control policies for a flexible manufacturing system,* International Journal of Production Research, Vol. 19, No. 5, 1985.

[Vin85] Vinod, B., and J. J. Solberg, *The optimal design of flexible manufacturing systems,* International Journal of Production Research, Vol. 23, No. 6, 1985.

[Von86] Von Lanzenauer, C. H., and A. Hamidi-Noori, *Fiscal-period-based service level constraints and safety stock requirements,* Int. J. Prod. Res, Vol. 24, No. 3, 483–492, 1986.

[Wein89] Wein, L. M., *Capacity Allocation in Generalized Jackson Networks*, Operations Research Letters, Vol. 8, No. 3, 1989.

[Yao85] Yao, D. D., and J. A. Buzacott, *Modeling the performance of flexible manufacturing systems,* International Journal of Production Research, Vol. 23, No. 5, 1985.