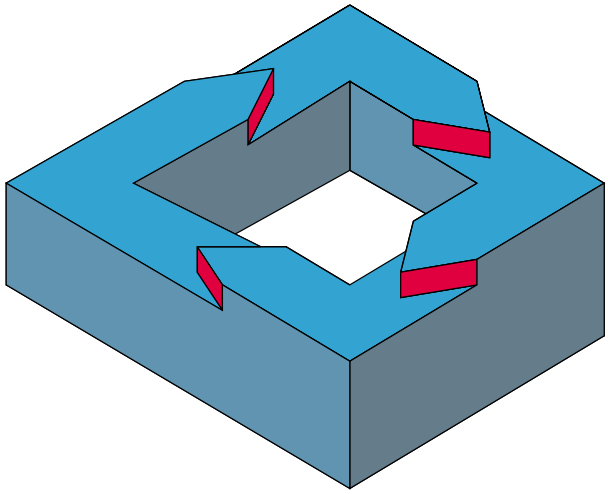
Re-Entrant Flow Lines



Dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in
Management of Manufacturing and Automation

Carlos Filipe Gomes Bispo

September 22, 1997

Committee
Professor Sridhar Tayur (Chairman)
Professor Uday Rao
Professor Art Hsu

Graduate School of Industrial Administration
and the Robotics Institute
Carnegie Mellon University, Pittsburgh, PA 15213

In memoriam

My four grandparents could not see this day, nor ever dreamt that the young boy they once called grandson with so much love would ever come this way, nor could understand a single thing of what is in here even if written in Portuguese.

Some of them could not even read in our own language. They were simple peasants who understood what is really important in this world: the value of a friendship and the value of a person's good word and honest work.

To them, Elisa and António, and Emília and Bernardo, I dedicate this work in loving memory.

Acknowledgements

I am grateful to many for having reached this milestone of my work and my life. To do justice and be fair it would probably take me many pages to acknowledge so many friends, colleagues, and teachers who, sometime or another, gave me their support, inspiration, incentive, and dealt with my mood swings from euphoria and ecstasy to despair and disbelief, never charging me the price of a lost friendship.

Also, my usual unability to express myself in writing is a strong handicap, of which I apologize up-front. I am probably going to forget many of you, either for pure temporary oblivion or for lack of space. Please be assured that you are in my heart, and my not mentioning you here does not imply I am less grateful for the journeys we have made together. So, let us get on with the business...

To Sridhar Tayur my first thanks for guiding me through the narrow paths of this work, for his outstanding ability to see through the forest of my random thoughts, and walk through the cliffs of my bad writing.

João Sentieiro is definitely the biggest responsible for this adventure, which started many years ago when I was just finishing my first degree. I found in you a caring advisor, an inspiring teacher, a warm leader, and a friend in a million. No word is enough to tell how much I treasure you.

My staying in the United States of America was sometime or another financially supported by the following Portuguese institutions: Fundação Calouste Gulbenkian, Junta Nacional de Investigação Científica e Tecnológica, INVOTAN, and Fundação Luso-Americana para o Desenvolvimento. Also, the Graduate School of Industrial Administration has financially supported my work. I could not have survived without their sponsoring and I am grateful for their trust in my potential, hoping that they may find this thesis worthy of their investments.

Many are the teachers who played an important role in my life as a student, as a researcher, as a teacher, and as a person. I recall my first teacher as well as I recall the most recent. For them I send my warmest word of gratitude for their efforts and inspiring lectures. I am happy and honored for counting some of them as my present friends. A special word of gratitude to the members of my committee and to the independent reader, Professor Bruce Krogh, for accepting this task and for their most valuable comments.

Out of the colleagues of trade I wish to bring João Paulo to the top of the list. We have come a long way, since the early days in London, freezing to death in a house with broken pipes and no water for three days to the freezing days in Pittsburgh, trying to jump start a car that had seen better days. You have been patient enough to endure my cooking experiments, while making you my private guinea pig. You are the whole package and as good as they get.

I am lucky for being able to have pleasure in the work I do. Luckier still for counting so many good friends among my colleagues at the Instituto Superior Técnico and, in particular, at the Instituto de Sistemas e Robótica. Out of these, I would like to express my particular gratitude to Carlos Pinto Ferreira, Isabel Lourtie, Isabel Ribeiro, José Alberto Vitor, Luís Custódio, Pedro Aguiar, and Pedro Lima, for the friendship, inspiration, and support. To some of these and many others I am also thankful for covering for me when I could not lecture, allowing me to have a lengthy stay abroad.

As a student at the Carnegie Mellon University, I had the privilege of interacting and suffer in arms with many outstanding persons. Ravi Anupindi, my roommate in the early time of my stay, a great chef, an inspiring colleague, and a warm friend. Marcel Becker, probably the one classmate who has endured me in my worst moments. Your patience has the size of a mountain and I will be grateful forever for your friendship, solidarity, and companionship through all the all-you-can-eat joints of Pittsburgh. A special word also to Roman Kapuscinski, big as a house and sweet as a bunny. One of the most bright classmates I met and the really smooth character without the cigarettes.

The staff of the Graduate School of Industrial Administration has always been a source of wonder to me, for their professionalism and caring. Above all, a special word to Jackie Cavendish, the granny of all Ph.D's, who always had a word of encouragement and the sweetest smile on her face.

Outside the smaller circle of classmates there are many good friends who have left an indelible mark in my heart. A first reference to the *Portuguese gang*, for the stimulating company and solidarity. In particular, a warm word of thanks to Ana Paula Pereira, Lino Santos, and Margarida Jácome.

Crossing the country boundaries, words are again insufficient and maybe what is left unsaid is more important. António Alonso from *Galiza*, the true impersonation of D. Quixote, even in his emotions. How beautiful it is to know you. Lynn Tomasits, your sensitivity and caring for others goes beyond my ability to describe it, you are a true silver friend becoming gold. Thank you for Shakespeare and many other things. A word to all the coffee friends, which includes those who served it and those who drank it. Naturally, I am referring to the *espresso* lovers. Also a word of thanks to those who have provided me with shelter during my many visits to the U.S.A. and were not mentioned before. They are my cousins Jorge and Piedade, Richard and Maria João, and P.J. My thanks for keeping my good spirits go to David Letterman and Dennis Miller. They do not know me, but I know how much I owe them.

Some other friends live in Portugal and have managed to fight for my attention while I was away. Their perseverance is the best demonstration of their friendship which I hope to return with interests. Of all these soldiers of friendship, I would like to send a warm hug to Alice Godinho and Mário Silva. Thank you, Aldina and António Carlos, and Leonor and Manuel for the Saturday night games and for your warm and cosy company.

A word to my parents for always leading me in the direction of knowledge and learning. May I do justice to your expectations and honor all the sacrifices made by you. A big hug and a truckload of kisses to my sister. Sorry for being so unavailable and so unapproachable sometimes.

Finally, my wife and son are probably the ones who have paid the highest price for my dreams. Maria do Rosário, you have stoically endured my absence, late hours, and bad humor all these years without denying me your support and love. All the words in the world are not enough to express my gratitude. João Pedro has practically grown up having his father somewhere else. I hope there is still time to forget that those days ever existed, my dearest son.

G.S.I.A., September of 1997

Carlos Filipe

Abstract

Re-entrant flow lines have attracted significant interest from the research community in recent years because of their direct applicability to semiconductor fabrication, [Kumar, 1993]. A flow line is a manufacturing system where several products are processed through a common and fixed sequence of operations. Each operation is performed on a different machine. A re-entrant flow line is a manufacturing system where this sequence of operations is repeated several times before the products are completed. Facilities that manufacture semiconductor wafers have such a structure. This thesis addresses several aspects related to managing these lines in a unified manner, provides insights into the several complex interactions that are inherent to these systems, and establishes a framework to study more complex re-entrant lines.

The class of re-entrant flow lines addressed by this thesis processes several different types of products, each subject to an external demand. The thesis makes contributions on the specific problem of production control. That is, given that there exists an external demand for each different product and given that each machine in the line has a bounded capacity, it addresses the questions on how to decide the production quantities so that the demand is tracked and the operational costs minimized.

The questions addressed in this thesis concern issues like what should be the degree of capacity sharing for these systems. Namely, should capacity be shared at the same level by all the products in their different stages of processing or should there be some rigid and static allocation of capacity to products and/or levels? If there is some degree of dynamic capacity sharing, how should the available capacity be distributed among the several products?

Inventory control in discrete time is a popular paradigm to model production systems. To the author's knowledge there is no such study for re-entrant production systems. The present thesis is a first attempt at providing a framework broad enough for such systems. It follows closely the approach proposed in [Glasserman and Tayur, 1994, Glasserman and Tayur, 1995].

In [Glasserman and Tayur, 1995], the authors consider a capacitated, multi-machine, single product flow line. They propose a multi-echelon base stock policy to control such systems. In order to determine the optimal base stock levels, they use Infinitesimal Perturbation Analysis, (IPA), [Ho and Cao, 1991]. The simulation based approach computes the optimal levels in order to minimize standard holding and penalty costs along the production line. A computational study provides

important insights.

In [Glasserman and Tayur, 1994], the authors address the issue of stability for the systems considered in [Glasserman and Tayur, 1995], establishing conditions under which their policies ensure it. The stability result itself is trivial, since the necessary and sufficient condition for their policies to ensure stability is such that the expected demand is under the smallest of the capacities along the line. However, its establishment is also necessary to validate some of the IPA results.

This thesis extends the above mentioned work in the following manner: incorporating the re-entrant structure and considering multiple products. Additionally, [Glasserman and Tayur, 1994] is extended by addressing the stability issue in the situations where random yield is present.

The approach used to study these problems is *simulation based optimization*. Simulation is used as a tool to obtain estimates of cost and estimates of the cost gradient with respect to the parameters describing the control policy. Most research dealing with issues in semiconductor manufacturing prefer simulation over simple analytic models, [Uzsoy et al., 1992]. However, in the majority of the situations, simulation is used as a performance evaluation tool. Simulation offers the flexibility to model the complexities adequately, while the gradient computation (via IPA) helps in identifying good solutions quickly. Besides using simulation as a tool for optimization, the thesis also makes use of it to compare the performance of different capacity management schemes, providing clear conclusions about their relative performances.

Contents

List of Figures

List of Tables

Part I

Prelude

Chapter 1

Introduction

Re-entrant flow lines have attracted significant interest from the research community in recent years because of their direct applicability to semiconductor fabrication, [Kumar, 1993]. A flow line is a manufacturing system where several products are processed through a common and fixed sequence of operations. Each operation is performed on a different machine. In this thesis a re-entrant flow line is a manufacturing system where a sequence of operations is repeated several times before the products are completed. Facilities that manufacture semiconductor wafers have such a structure. This thesis addresses several aspects related to managing these lines in a unified manner, provides insights into the several complex interactions that are inherent to these systems, and establishes a framework to study more complex re-entrant lines.

The class of re-entrant flow lines addressed by this thesis processes several different types of products, each subject to an external demand. The thesis makes contributions on the specific problem of production control. That is, given that there exists an external demand for each different product and given that each machine in the line has a bounded capacity, it addresses the questions on how to decide the production quantities so that the demand is satisfied and the operational costs minimized.

These same issues for non re-entrant systems producing more than one type of product are already difficult to solve. The difficulty results from the fact that when more than one product demands the utilization of common resources, these resources will have to be *allocated* to products in a satisfactory manner. There are no analytical solutions for these allocation problems. For a re-entrant system this difficulty is compounded by the fact that each product visits each machine at different stages of production. This type of material flow may induce positive feedback loops within

the line so that the overall net capacity is reduced, thus affecting its ability to effectively satisfy demand. A production control policy for re-entrant flow lines producing more than one product type will need to address how the capacity is shared among both the different product types and identical products at different processing stages.

On the specifics of the above, the questions addressed in this thesis concern issues such as what should be the degree of capacity sharing for these systems. Namely, should capacity be shared by all the products in their different stages of processing or should there be some rigid and static allocation of capacity to products and/or groups of operations? If there is some degree of dynamic capacity sharing, how should the available capacity be distributed among the several products?

It turns out that the best results are achieved when all the capacity of each machine is equally shared by the several products at their different processing stages. The total sharing of capacity reduces the frequency of capacity bounds and this allows for a faster recovery from backlog. However, there are particular dynamic capacity managing schemes for which costs increase when the sharing degree increases (Section 5.2).

Given the lack of analytical and structured results on optimal policies, which could be used to guide the search for their optimal settings, this thesis proposes a class of policies inspired by a long trail of results on inventory control theory. Although sub-optimal, this class of policies is attractive because it is easy to implement and parameterize.

Our model of a re-entrant production system has M machines in series (which will be called *stages*). Each machine feeds a downstream buffer were parts wait for next operation. Each one of the P products processed by the system has to cycle K times (each cycle termed as being a *level*) through those M machines before being completed. The framework is a *discrete time* (or *periodic review*) capacitated multiple-product production-inventory system operating under a modified *echelon base stock policy*[1]: every level and stage operates on a base stock policy for echelon inventory. More specifically, given a particular product, the decision maker adds all inventory downstream from that level and stage to determine the echelon inventory. If the echelon inventory falls below the corresponding base stock value, the decision will be to produce the difference, provided there is enough capacity and (relevant) upstream inventory. Since there are multiple products and multiple visits to each machine, it is necessary to allocate capacity dynamically to individual needs whenever

[1]Echelon base stock policies are defined in the context of uncapacitated systems. [Clark and Scarf, 1960]. Since the systems considered here are capacitated, the term *modified* expresses that the policies used are a variant of the original to incorporate such feature.

the products require more than the available capacity.

Inventory control in discrete time is a popular paradigm to model production systems. To the author's knowledge there is no study using inventory control in discrete time as a model for re-entrant production systems. The present thesis is a first attempt at providing a framework broad enough for such systems. It follows closely the approach proposed in [Glasserman and Tayur, 1994, Glasserman and Tayur, 1995].

In [Glasserman and Tayur, 1995], the authors consider a capacitated, multi-machine, single product flow line. They propose a multi-echelon base stock policy to control such systems. In order to determine the optimal base stock levels, they use Infinitesimal Perturbation Analysis, (IPA), [Ho and Cao, 1991]. The simulation based approach computes the optimal levels in order to minimize standard holding and penalty costs along the production line. A computational study provides important insights.

Given that their class of control policies is not derived from a particular optimization procedure, it is naturally necessary to show that under such class of policies the systems being controlled are stable. *Stability* in this context is characterized by the fact that the inventories sitting in between each machine have bounded expected values. This is equivalent to saying that the cumulative production of the single product trails the cumulative demand by no more than a bounded amount. In [Glasserman and Tayur, 1994], the authors address the issue of stability for the systems considered in [Glasserman and Tayur, 1995], establishing conditions under which their policies ensure it. The stability condition itself is trivial, since the necessary and sufficient condition for their policies to ensure stability is such that the expected demand is less than the smallest of the capacities along the line. However, its establishment is also necessary to validate some of the IPA results.

This thesis extends the above mentioned work in the following manner: incorporating the re-entrant structure and considering multiple products. Additionally, [Glasserman and Tayur, 1994] is extended by addressing the stability issue in the situations where random yield is present.

Due to the complexity of the models and policies, it is very difficult to compute explicitly the values of the optimal parameters, even when assuming some sort of simplifying structure on the stochastic processes involved. Resorting to alternative methods is therefore necessary. This was the strategy in [Glasserman and Tayur, 1995]: using an IPA based optimization procedure to compute the optimal base stock values. This is also the strategy of this thesis. One of the most elegant

qualities of IPA is that it does not rely on any strongly restrictive assumptions for the random processes that disturb the systems being controlled. However, its utilization requires that the used control policies be described by a set of parameters. Therefore, without possessing a knowledge of the parameters that fully describe the true optimal polices one has to rely on sub-optimal approximations.

The approach used to study these problems is *simulation based optimization*. Simulation is used as a tool to obtain estimates of cost and estimates of the cost gradient with respect to the parameters describing the control policy. Most research dealing with issues in semiconductor manufacturing prefer simulation over simple analytic models, [Uzsoy et al., 1992]. However, in the majority of the situations, simulation is used as a performance evaluation tool. Simulation offers the flexibility to model the complexities adequately, while the gradient computation (via IPA) helps in identifying good parameter values quickly. Besides using simulation as a tool for optimization, the thesis also makes use of it to compare the performance of different capacity management schemes, providing clear conclusions about their relative performances.

1.1 Organization of the Thesis

This document is composed of three parts. Part I includes this chapter and Chapter 2 — **Literature Review**, where a review of relevant literature is presented. It is also an objective of Chapter 2 to discuss the structure of the optimal policies for re-entrant systems subject to a variety of disturbances.

Part II concentrates on the analysis of a simple class of systems as a first step towards analyzing a broader family of production systems under the framework of discrete time inventory control. The class considered can be seen as the inventory model counterpart to the *Kelly-type networks* discussed in [Dai and Weiss, 1996]. These networks are characterized by having the same service time distributions across all servers and products. Under the modeling framework of this thesis, the service time distributions will be converted into deterministic processing times. This ends up translating into equal capacity needs for each unit of product — termed in the thesis as the *uniform load assumption*. The only source of uncertainty considered will be the random demand. In Chapter 3 — **Theoretical Foundation**, the basic model and theoretical validation of the IPA approach will be presented. The definition of the basic model entails establishing the dynamic

equations governing the systems, presenting the derivatives of the state variables with respect to the parameters defining the control policies, establishing the production decision mechanism given that sharing of resources occurs, and setting the performance measures.

Out of these, the production decision mechanism assumes a particular importance because it encompasses the adoption of a particular control policy and the definition of the capacity sharing procedures — *production rules*. It is necessary to define how the sharing of resources is done. The thesis will introduce some production rules which interpret different ways of doing such sharing. Sharing can be done through a fixed list of priorities, proportional to individual needs, or in a way that gives more priority to higher individual needs. The validation of the IPA approach entails the verification that the state variables possess the adequate smoothness properties so that one can exchange the order in which to apply the expected value operator and the derivative operator.

Stability will be discussed in Chapter 4 — **Stability**. This issue assumes a key role in the IPA approach when validating it to solve infinite horizon problems. Stability ensures that the stochastic processes possess the regenerative structure necessary for the validation of the IPA. Also, it is of particular relevance in re-entrant systems. The manner in which products flow in re-entrant systems may induce positive feedbacks that generate unstability, [Lu and Kumar, 1991, Bramson, 1994]. This chapter provides a sound discussion, showing under which conditions the proposed control policies ensure stability. The derived necessary and sufficient conditions will be the natural and trivial extensions of the stability condition mentioned above for a flow line producing a single product.

Part II closes with Chapter 5 — **Experimental Study**, where a relatively extensive set of experimental data is presented and some structural properties are discussed. The experiments conducted were designed with the purpose of evaluating the relative performance of the production rules proposed in Chapter 3. To accomplish such objective, the experimental studies evaluate the influence of holding costs, backlog costs, demand variance, and capacity distribution among machines and levels. Also, the influence of the degree of capacity sharing is investigated. Clear conclusions are drawn from these experiments relative both to the degree of capacity sharing that achieves the best performances, and to which of the production rules incurs lower costs.

One of the most relevant structural properties established is the equivalence between standard operational costs and Type-1 service level. The optimal base stock levels verify the newsboy property relative to the probability of not satisfying demand in the period it occurs. In many situations,

there is also a subtle relationship between machine capacities and optimal base stock levels. For balanced systems and for particular degrees of sharing, the difference between consecutive base stock levels matches the capacity of the machine that separates them.

Part III concludes the thesis with an extension of the theory of Part II. In Chapter 6 — **Theoretical Foundation Revisited**, the basic theoretical results will be reviewed to include random yield in the model and to drop the Kelly-type structure. Once again, the stability problem assumes a central role in the discussion. The first sections of Chapter 6 are entirely dedicated to the stability problem. This discussion will be made by steps as the original model of Part II gets enriched by new features: first, the results of Chapter 4 are extended for systems with non uniform loads; next, the original model is extended to incorporate random yield; and finally, both random yield and non uniform loads are considered simultaneously. At the end of this stability discussion, the chapter presents the necessary and sufficient stability conditions for re-entrant flow lines subject to a variety of random disturbances and operated under the classes of policies proposed. To establish these conditions, it will be necessary to use classes of control policies different from the one used in Part II. These classes constitute generalizations of the class used in Part II. Also, the technique used constitutes a base to establish stability of re-entrant systems with very generic material flows, not just of the flow line type.

Next in this chapter, the basic validation result of the IPA technique is presented and the chapter concludes with a brief summary of some experimental results for systems with non uniform loads and perfect yield.

The last chapter, Chapter 7 — **Conclusions and Future Research**, presents a summary of the thesis and discusses some topics for future research.

As some of the proofs will be skipped during the theoretical discussion, Appendices A and B will contain proofs which are either relatively trivial extensions of material already published or correspond exactly to results published by others. Their inclusion here is done with the objective of providing a self contained document. In Appendix C presents a summary of the optimization procedure addopted. Finally, Appendix D presents some graphics which complement the ones presented in Chapter 5 and the summary presented at the end of Chapter 6.

Before listing our contributions, it is probably important to discuss some of the limitations of the approach or issues not contemplated in the present work. Some of the limitations rely on the

modeling assumptions. For instance, although the production of parts is discrete, the thesis models the inventory variables as continuous with the purpose of extracting derivatives for the IPA. At the end of the optimization procedure the achieved values will have to be converted to integer values. The round up to nearest integers is not serious in terms of cost, since the cost functions are usually relatively flat around the optimal. However, when it comes to systems that are unstable for a particular choice of parameters and stable for others, even though the optimization procedure may converge to a stable setting, there is no guarantee that the round up values will ensure stability.

Infinitesimal Perturbation Analysis cannot be used to find optimal policies. It can only be used to find optimal parameters for a pre-specified class of policies. Therefore, the end results are only as good as the quality of the approximation with respect to the actual optimal policies.

The issue of machine reliability is not directly addressed in this thesis. The stability results and IPA validation carry through for systems with random capacity. However, there is no explicit consideration of this feature in the experiments nor in the theory. Chapter 7 presents a brief discussion on how to handle different types of machine unreliability, illustrating how the present approach deals with some of them.

Random yield is considered for the purpose of validating the IPA approach and is explicitly dealt with for stability purposes. However, the thesis does not cover this feature in the experimental studies. Also, as the literature review of Chapter 2 shows, the control policies proposed in this thesis may be less adequate for systems with random yield. Possible directions to deal with random yield in this setting are discussed at the end of the thesis.

1.2 Summary of Contributions

This section presents a brief overview of what is considered by the author as the main contributions of the thesis, the objective of which being to study the problem of production control for capacitated re-entrant flow lines producing multiple products.

The problem of determining the stability conditions for re-entrant systems is completely solved for the classes of policies proposed throughout the thesis. Stability of re-entrant systems is a complex topic, and has drawn the attention of many authors. The stability proofs presented here and the insights gained from them constitute the major contributions of this thesis. The stability results presented cover re-entrant flow lines with capacitated machines, subject to random yield,

and producing multiple products with different demand processes and different loads on each visit to any machine. The main conclusion and contribution of the stability discussion is the fact that the control policies proposed are ensured to stabilize any system for which the expected load for any machine is under its capacity. Therefore, the usual necessary stability condition is shown to be also sufficient for the policies presented.

The experimental data presented illustrates the strength of IPA as an optimization tool and provides important managerial conclusions on how to handle re-entrant systems. Instances of such conclusions are listed below.

- Performance measures:
  - operational cost based measures and service level measures are equivalent for the class of policies used in Part II;
  - the experiments show that intermediate holding costs affect the distribution of safety inventory along the line, but do not affect the Type-1 service level measure nor impact the relative performance of the production rules proposed;
  - the only cost parameters that have influence on the relative performance of the productions rules and on the Type-1 service level measure are the holding and backlog costs for end products.
- Capacity sharing:
  - the best costs are achieved when the capacity of each machine in the line is equally shared by all the products at their different levels of processing.
- Connection between theory and experiments:
  - the experimental results of Chapter 5 stress some of the limitations of the used policies and suggest directions for improvement;
  - the stability discussion of Chapter 6 resorts to policies that agree in structure with the directions pointed by the experiments;
  - there is both experimental and theoretical evidence pointing in the direction of richer classes of policies, establishing a bridge between this thesis and future research in this area.

Some other conclusions require a little more detailed explanation of the production rules and of the whole model. Therefore they are not listed here. A comprehensive summary is presented at the end of Chapter 5 (Section 5.4).

Another relevant contribution of this thesis is the definition of a framework amenable to dealing with a vast number of features, such as random demand, random yield, and machine variability. On the other hand, the basic model set forth possesses the flexibility to incorporate other characterizing features of production control policies, such as bounds in local inventories and bounds in the amounts produced at any given period.

Details of the model that are of relevance concern the proposed schemes for capacity management, both from the static and dynamic perspectives. These schemes address the question of how to use the capacities for the different levels and products.

The validation of the Infinitesimal Perturbation Analysis methodology for these systems is also a major contribution of the thesis. IPA is a technique with a great potential to explore optimal policies and parameters for large scale systems, when the policies are constrained to classes defined *a priori*.

Chapter 2

Literature Review

The literature on production control covers a very broad spectrum in general, including specific contributions to managing production systems in semiconductor manufacturing. A complete and thorough review of all the work in the area is a task of enormous proportions and would probably be tedious for the reader. Therefore, a judicious selection of the most relevant contributions to production control, which are in some sense related or inspirational to the work described here, is presented. Although there is an effort at limiting the scope of the review, it turns out to be quite extensive. This has to do with another concern of this review. Besides discussing some commonalities that different approaches have with each other and with what is proposed here, it is intended that this chapter may serve as an instrument for identifying the features that a control policy should possess. This latter objective is of particular relevance in Sections 2.3 and 2.4, justifying their length.

A classification into main clusters is proposed, although some approaches may spread over more than one of the clusters. Some of the clusters will represent research done with no specific emphasis on semiconductor manufacturing and others will be defined in the specific context of semiconductor manufacturing. To properly frame the complexities and specifics of semiconductor manufacturing, a brief and summarized description of the features that make it so unique is first presented. The main clusters of research addressed in this review will cover: Inventory Control, Random Yield, Flow Rate Control, Queueing Networks, Stability, Simulation, and Infinitesimal Perturbation Analysis.

In the cluster on *Inventory Control*, a survey of classical and recent operations research methods and results will be provided. The majority of the relevant results in this area pertain to cases where the optimal policies have been shown to be of the critical number type. Such structure of optimal

policies is a strong incentive to keep using variants supported by similar principles for more complex systems, for which no analytical solution has been obtained.

Besides its re-entrant structure, one other feature that singles out semiconductor manufacturing systems is the important role the effects of *Random Yield* play in the performance of those systems. Random yield is not unique to semiconductor manufacturing, since it is present in many other production systems. However, in no other production systems does it condition the functionality as much as in these. There are several approaches targeted to deal with random yield specifically. Yield improvement is one of them, as well as characterizing the relationships between yield and other performance measures like cycle time.

Some of the approaches in inventory control treat the problem as that of regulating production rates, rather than production quantities. On this subset of inventory control, there has been a substantial body of contributions which are of great relevance for managing manufacturing systems in general, although the specifics of semiconductor manufacturing have been driving much of the effort in this area. This feature justifies a special coverage on *Flow Rate Control* approaches, isolated from the general inventory control problem approaches.

Another popular modeling paradigm is that of *Queueing Networks*. Their virtue lies in their ability to model very complex production systems, with many interacting components. The stochastic nature of generic production systems is well captured through the arrival processes distributions, routing and connection modeling, and service time distributions. Consequently, the simulation of production systems through networks of queues can provide very accurate estimates of the performance of real life systems. By means of both closed queues or open queues, there have been many research contributions to the understanding of the structure of manufacturing systems. Although, in general, networks of queues have been used to estimate performance, some effort has been made in using queues as a way for determining control policies with some success, namely in the context of re-entrant systems.

Whenever control problems are posed, there is a need to ensure that the policies used are able to keep the system's parameters within certain bounds. This is an issue of *Stability*. In general, stability is a relatively trivial issue in production systems. However, manufacturing systems with a re-entrant structure pose the question of stability on a different level of complexity, for there are situations where policies which are relatively innocuous for non re-entrant systems may lead to instability when applied to re-entrant systems. Therefore, any attempt at controlling systems with

a re-entrant structure will have to seriously address the question of stability and stabilization of those systems. There has been important research to address this problem and the present thesis makes an important contribution to that growing body of research.

In the area of performance estimation, and comparison of alternative control policies, *Simulation* has possibly been one of the most widely used techniques. In many situations, when the dimension of the problems under consideration grows to magnitudes where analytical solutions are no longer available or when numerically derived solutions are time consuming, simulation is in fact the only practical tool available to study production systems. Also, many optimization procedures use simulation as the underlying model to collect information on performance measures and even on gradients of performance measures with respect to the control parameters. An example is the present thesis, that makes use of a simulation based procedure in order to determine the optimal parameters that characterize the control policies adopted.

A special case of simulation based techniques is that of the *Infinitesimal Perturbation Analysis* used in this thesis. Infinitesimal perturbation analysis, when applicable, is a very powerful tool to guide optimization procedures. It allows the estimation of gradient information from a single simulation run, with very mild assumptions on the nature of the random processes disturbing the systems to control. Since its early days, when its application was restricted to a very small set of problems, infinitesimal perturbation analysis has matured to the point of becoming one of the most important tools of the present when it comes to optimizing control parameters for large scale systems subject to a variety of random disturbances. Effort has been placed on formally setting the conditions under which it is a valid procedure and on determining instances that satisfy the theoretical frameworks available. This thesis looks at infinitesimal perturbation analysis from this latter point of view. A relatively good sample of both types of work for very diverse applications is provided in the literature review below.

2.1 Semiconductor Manufacturing

The process by which semiconductor devices are manufactured is composed of four basic steps: *wafer fabrication*, *wafer probe*, *assembly or packaging*, and *final testing*. Of these four, the most complex and capital intensive step is the wafer fabrication. This is the step where the complex and intricate circuitry necessary for the integrated circuit is built. A very large scale integrated

circuit is constructed out of wafers of silicon or gallium arsenide by means of creating several layers with different physical properties and diverse connections between each other. Once this process is completed, the finished wafers are sent to the wafer probe step. Each wafer may contain hundreds of individual circuits, which are individually tested in order to determine if they meet their specifications. The individual chips to be are referred to as *dice*. Chips that fail the tests are marked for future disposal when the wafers are cut into their individual circuits. After this, the good circuits are placed in plastic or ceramic packages for protection from damage caused by the outside environment. In the final testing, the circuits are subjected to a variety of automated testing operations to determine their quality. The testing operations range from performance evaluation at different temperatures to "burn in", where the devices are subjected to thermal stress in order to precipitate latent defects that would otherwise manifest themselves in operation.

The wafer fabrication step, which is done in specially dedicated facilities, usually referred to as *wafer fabs*, may involve several hundreds of operations, depending on the complexity of the devices produced. Many of these operations have to be performed in a clean-room environment to avoid contamination of the wafers by impurities present in the air. The basic set of operations performed on a wafer includes: cleaning, deposition, lithography, etching, ion implantation, photoresist strip, and inspection and measurement.

The *cleaning* operation is performed with the objective of removing particles from the surface of a wafer before a layer is produced. The *deposition* step has the objective of growing or depositing a layer of material on the cleaned surface of the wafer. The most complex operation is the *lithography*, when a photoresistant liquid is deposited over the wafer and the circuit is defined using photography. This is the operation that requires the highest precision. The photoresist is first deposited and baked. It is then exposed to ultraviolet light through a mask which contains the pattern of the circuit for the layer being built. Finally the exposed wafer is developed and baked. The material which gets exposed to the light is then etched away in the *etching* operation. Then comes the *ion implantation*, when selected impurities are introduced in a controlled fashion to change the electrical properties of the exposed portion of the layer. After this, it is necessary to remove the photoresist remaining on the wafer during the *photoresist strip* operation. The sequence of operations is concluded with an *inspection and measurement* operation to check if there are defects. This overall sequence of operations has to be repeated for each layer of the integrated circuit. A detailed and very complete description of semiconductor manufacturing technologies can be found

in [Sze, 1983], [Gise and Blanchard, 1986], or in [Runyan and Bean, 1990]. More concise descriptions can be found in [Chen et al., 1988] and [Uzsoy et al., 1992].

Many things can go wrong during each cycle, so that the wafers at the end of a series of cycles may have defective individual circuits. Impurities present in the room may fall over the wafer and change the properties of the circuit, the masks may be out of place, and many other problems can occur. Also, the fact that the resources are very expensive makes it impossible to have multiple machines to process the several layers of the wafers. Therefore, the wafers have to be sent back to the same machines time and time again to be processed on each cycle. This feature makes the wafer fabs unique among all production systems, since products flow a number of times through a series of machines, causing them to possess what is called a re-entrant structure.

Recently, a review of research on production planning and scheduling for the semiconductor manufacturing industry was published in a set of two papers, [Uzsoy et al., 1992, Uzsoy et al., 1994]. Given its outstanding quality and the fact that it was published so recently, much of this section will be based on some of the classifications proposed there. Readers are referred to those two papers for more extensive coverage on some of the topics discussed here and some of the areas analyzed in the remainder of this chapter.

According to [Uzsoy et al., 1992], the relevant research is classified into large clusters defining three problem areas, which are basically a consequence of the scope of time they address. These three problem areas, which are not exclusive of semiconductor manufacturing, are:

1. *Performance evaluation.* Models whose objective is descriptive rather than prescriptive in nature, used for understanding the behavior of a given system;

2. *Production planning.* Long-term, more aggregate production planning with a time horizon of months or weeks;

3. *Shop-floor control,* which addresses the questions of how much material to start into the facility and how to control the flow of this material.

In the performance evaluation area, the objective is to establish models that can be used to evaluate the performance of a given system. They are typically used to answer long term questions like plant layout, or machine configuration, and system capacity. The main tool to address this type of questions has been simulation. Due to the fact that most of the questions are formally

intractable by other means, and since simulation models have become richer in their ability to model manufacturing systems, simulation is easier and less expensive. See [Uzsoy et al., 1992] for a very extensive list of significant contributions on this area using simulation. Another tool that has made its way into the semiconductor manufacturing is modeling by means of queueing networks. They have the potential to represent many of the specific features of an uncertain production system and may allow the elegance of closed form analytical results for the level of detail required by this problem area.

The production planning area tries to take into account long-term goals while determining parameters that will be used as guidelines by the shop-floor control area. The level of detail is somewhat greater than that of the performance evaluation but there is still some sort of aggregation of the available data. As opposed to the performance evaluation, there is already an attempt at determining optimal parameters rather than just evaluating how the system performs.

The shop-floor control area deals basically with two types of decisions: release and scheduling. The release policies determine when, what, and how much material is allowed to enter a system. The scheduling policies determine what is the next operation to be performed by an idle resource. Of these two, the biggest amount of work has been placed on scheduling. The diverse methods range from the deterministic scheduling algorithms for job shops to knowledge-based approaches. In between there are contributions on local dispatching rules for job shops, specific applications to batch processing machines, approaches concerned with the re-entrant structure of the wafer fabs, etc. See [Uzsoy et al., 1994] for a clustered and comprehensive classification of the several approaches which deal with shop-floor control.

There are some features specific to semiconductor manufacturing that make it particularly complex as a manufacturing process, namely in the context of wafer fabs. The first unique feature is the fact that products have complex flows. That is, different end products require diverse production recipes, each involving a great number of processing steps, some of which are performed on the same machines. So, the products have to cycle through a line of machines more than once in order to be completed. This re-entrant structure of wafer fabs introduces additional complexities to the problem of controlling production, since it demands a judicious choice on how the capacity of a single machine is managed, and raises serious stability questions. Each machine is requested for service not only by different products, but also by products of the same type on different stages of their processing. It is necessary to establish how to share capacity among all those products

both from the static and dynamic perspective. The re-entrant structure of the flows may introduce positive feedback loops into the system with a poor choice of priorities, so that the cumulative production of a product or a set of products does not manage to satisfy demand even in situations where the load imposed by the demand process is under the capacity of the fab. So, when it comes to a wafer fab, we have a system producing great volumes of individual products, different end products, and complex product flows.

Yield is definitely the biggest *Achilles' heel* of wafer fabs. Process output is uncertain due to contamination of wafers by environment particles that may deposit on their surface, plus problems due to equipment and material. Although yields may be well established for mature products, the rapidly changing characteristic of the markets and technology is always requiring the introduction of new products with new engineering problems to solve. Naturally, random yield is not present only in semiconductor manufacturing, but only in semiconductor manufacturing does it have such a central relevance. This is due to the fact that random yield in semiconductor manufacturing is not due to a technological shortcoming of the processes, but is inherent to those processes. Average yields of mature processes may be as high as 90%, but recently developed products may have yields under 10%. The ability to satisfy demand is naturally complicated by the presence of this unavoidable source of uncertainty. Some of the defective products have to be scraped, but in some cases they can be reworked adding to the flow complexity.

Another distinguishing feature of wafer fabs is the fact that the development of new products has to be tested on a real life manufacturing system. Computer simulations are not enough. New products or processes are constantly being developed given the highly dynamic market of semiconductor products. In many facilities the equipment is frequently used simultaneously for *production lots* and *engineering lots*. An engineering lot corresponds to a testing lot used during the development of a new product or process. A production lot corresponds to a lot associated with a mature process that is currently being marketed. The presence of these different lots in a manufacturing system creates conflicting goals in terms of production control. It is necessary to give priority to engineering lots to shorten the development phase, but the external demand for current products has to be satisfied in a timely manner. Some companies have different facilities for R&D, but many use the same facilities for R&D and regular production.

Adding to the uncertainty caused by random yield is the fact that machines are not reliable. The equipment is extremely sophisticated and it requires extensive preventive maintenance and

calibration. Nevertheless it is subject to unpredictable failures that disrupt the normal functioning of the facility. The resources used in wafer fabs have a wide range of distinguishing features. Some machines have long and/or sequence dependent set-up times when switching between different operations. Some machines are batch processing machines while some others are not. Some consecutive operations require that the amount of time in between is bounded above by some amount, or else the second operation cannot be performed and the first is wasted. The volume of data collected and maintained in a semiconductor facility is tremendous. Data has to be stored on a per operation basis. [Uzsoy et al., 1992] quote [Sullivan and Fordyce, 1990] on this issue as saying that the transaction volume of an IBM wafer fab is of 240,000 per day.

Cycle times in wafer fabs can be of the order of months for each starting lot. Some individual steps take hours to be executed. Besides the usual problems that long cycle times create in terms of tracking demand in a precise manner, there is the issue of the correlation between cycle time and yield. A negative correlation between cycle time and yield has been established by several authors. That is, the longer the cycle time the higher the chances of getting defective end products. It is easy to understand why this is so. Given that one of the sources of random yield is the deposition of environmental particles over the wafers, the longer the time a wafer spends in a facility the higher the chance it will get contaminated. Another difficulty resulting from long cycle times is the risk of obsolescence, which is particularly relevant in a volatile market such as that faced by the semiconductor industry.

2.2 Inventory Control

In inventory control problems one is faced with decisions regarding when and how much to order or produce so that an exogenous demand process is satisfied. Issues to take in consideration for modeling are how many stages of production exist, how many different products to process, how is the demand process characterized in terms of its stochastic nature, how is the production process characterized in terms of operation times, availability, reliability, and how many and at what time are the decisions to take place.

At the bottom of the scale of simplicity is the early model of [Harris, 1913], known these days as the classic *Economic Order Quantity (EOQ)* model, which was made popular by [Wilson, 1934]. More complex models have been formally developed since, mainly after the advent of Dynamic

Programming due to [Bellman, 1957]. In many of the early models the key concept of *base stock* policies has created a niche in such a way that almost all of the literature on inventory control addresses the issue of showing its optimality or, when this is impossible, to propose heuristic policies that rely on similar principles.

Another classic is the famous *newsboy* problem for a single period, single product, and single production stage, where demand is random and there is unit holding cost charged for holding stock at the end of the period and a unit shortage penalty cost for any demand unsatisfied. Also, there exists a unit ordering cost for the amount acquired at the beginning of the period. The optimal policy was shown to be of the base stock type by [Arrow et al., 1951], that is, there is an amount to order up to, given the initial inventory is under such amount. Although very simple, this result and the previous constitute the basic building blocks of inventory control theory.

For single machine and single product systems, base stock policies have been shown to be optimal in a variety of settings. Single period problems, multiple period finite horizon problems, and multiple period infinite horizon problems all possess the same structure for the optimal control policy in situations where there is no bound for the machine capacity, [Karlin, 1960, Morton, 1978]; for deterministic capacity bounds, [Federgruen and Zipkin, 1986a, Federgruen and Zipkin, 1986b]; or even for stochastic capacity bounds, [Ciarallo et al., 1994].

For multiple machines in series and a single product, [Clark and Scarf, 1960] show that base stock policies are optimal in terms of multi-echelon inventory. That is, they define multi-echelon inventory of a given machine as the sum of inventory from that machine all the way down the production system to the end product inventory. For each of these echelon inventories there is a critical number to order up to for finite horizon problems. Their machines have no capacity bounds. This result was extended to the infinite horizon case by [Federgruen and Zipkin, 1984].

There are other models for which a base stock policy has been shown to be optimal, however for cases of multiple machines in series, with capacitated machines, producing single or multiple products, little is known about the optimal policies. The same can be said regarding re-entrant systems. See [Graves et al., 1992] for a very extensive coverage of inventory control problems and theory. A more compact survey of stochastic inventory control can be found in [Porteus, 1990].

In situations where the structure of the optimal policy is not well known, it is often the case that an approximate, or sub-optimal, policy is proposed as a heuristic. Due to their simplicity

and given they are optimal for many simpler systems, it is often the case that some variant of a base stock policy is proposed and used. An example of this approach is that of [Glasserman and Tayur, 1995]. The authors propose to control a system producing a single product, with multiple machines in series, where each machine has a fixed capacity, by means of a modified multi-echelon base stock policy. The optimal base stock values are computed with the help of an optimization procedure which relies on gradient and cost estimates obtained through infinitesimal perturbation analysis. Also for multi-echelon systems, [Graves, 1996] develops a procedure to determine the echelon base stock levels of an uncapacitated system where each location may serve more than one site, but where each site receives inventory from only one other site. For Poisson demands, the approach allows the closed form calculation of first and second moments of several state variables. It is possible to determine the optimal base stock levels for a two stage echelon system. Things get more difficult for more than two stages though.

2.2.1 Random Yield

There exist basically three areas of interest when it comes to studying systems possessing random yield: modeling yield, improving yield, and controlling systems with a given yield structure. Modeling yield is concerned with identifying how to best approximate the random yield process as a stochastic process. Improving yield deals with quality control issues and aims at improving production processes to increase yield or monitoring the output to quickly detect major disruptions that call for readjustment of the production process. The third area deals with determining production decisions given that random yield is present and unavoidable to some extent. The first two areas are outside the scope of this thesis.

However, relative to the modeling of random yield there is an important issue, which concerns the relationship between cycle time and random yield in the context of semiconductor manufacturing. [Wein, 1992b] discusses the strong correlation that exists between the length of the cycle time and the net capacity of a semiconductor fab. Longer cycle times tend to reduce yield and, therefore, reduce the overall capacity. The importance of this observation is that strategies that concern keeping cycle times low should be given particular attention. Since cycle times are correlated with the amount of work in process, care should be taken in terms of the amount of material that is allowed to enter such a facility. For an analysis of random yield models and their evaluation, see [Cunningham, 1990].

The presence of random yield in production systems has a very destructive impact over many of the nice structural properties of inventory control policies. Moving from a model with unlimited capacity and deterministic demand to a model with stochastic demand, or from this one to another model where capacity is fixed, or even to a model with random capacity, the base stock structure of the optimal policies is retained. However, as soon as random yield is added to the model the base stock structure is typically lost.

A distinction should be made here between *order point* policies and *order up to* policies. Many of the base stock policies mentioned above are simultaneously order point and order up to policies. Order point policies are those where there is a point (or set of points) in terms of initial inventory, above which it is optimal not to order. Order up to policies are those where, when it is optimal to order, the optimal ordered amount is such that the ending inventory is a particular inventory point (or set of points). A base stock policy satisfies these two properties, and in many situations both points coincide.

When random yield is present, many systems are such that they retain the order point feature but lose the order up to feature. It could be expected that the order up to feature would be retained for the expected production, that is, the amount produced multiplied by the expected random yield. However, this is not the case.

[Gerchak et al., 1988] discuss the structure of optimal inventory control problems for a single machine system with random yield. One of the most interesting conclusions is the fact that the order point remains unchanged relatively to the perfect yield counterpart, that is, it does not depend on the yield distribution. They analyze the single period and the two period problems in some detail to show that the amount to order depends not only on the expected random yield but also on the second moment of the random yield. Actually, the optimal order grows as the expected yield decreases but it decreases as the variance increases. If there is a high yield variance, big production amounts have a significant probability of wasting a big amount of material whereas smaller orders incur smaller absolute waste. Regarding the multiple period problem, they observe that the solution is non myopic since the sufficient condition for dynamic decisions to have myopic solutions is not satisfied — given the action, the current state has no impact on next state. It is clear that, after deciding the amount to order, the end inventory is a function of the starting state when random yield is present.

On a subsequent paper, [Henig and Gerchak, 1990] discuss the structure of periodic review

policies in the presence of random yield. The paper starts with a very good literature review on random yield problems covering continuous and periodic review models, to which the interested reader is referred. They also analyze a single machine case and make a complete analysis of single period, finite-horizon and infinite horizon models with general production, holding and shortage cost structures. For the single product case, they confirm that the order point structure exists and that it is equal to the value for perfect yield models. They provide an approximation for the optimal ordering quantity, obtained through the Taylor series. For multiple period problems, it can be shown that there exist a critical number per period, S_n, such that nothing is ordered if and only if the starting inventory for the period is above that number.

Given the complexity of the inventory control problem when random yield is present, many authors have proposed heuristic approximations to the order quantity. Examples of such a strategy can be found in [Akella et al., 1992] for the multiple product, capacitated, two machine system and in [Bollapragada and Morton, 1994] for a single item periodic review inventory problem. The first paper proposes a linear decision rule derived from approximating costs by means of a quadratic function. The second paper proposes a stepwise linear heuristic that approximates the optimal production decision. One common first approximation heuristic is to produce the difference between the present inventory and a target level inflated by the expected random yield. The problem with this is the fact that when the difference is too big the order is naturally big and the amount lost due to the random yield is significant, so that to improve costs it is of advantage to impose some sort of upper bound on the amount ordered. More sophisticated heuristics also take variance into account.

An outstanding and very recent overview of research in the context of optimal inventory control in the presence of random yield can be found in [Yano and Lee, 1995]. Readers are referred to it for more details on the structure of the optimal policies and on heuristics.

2.3 Flow Rate Control

Some authors approach the problem of production control as one of regulating the rates of production, rather than the production quantities. The most important line of research on this area was initiated with the Ph.D. thesis [Kimemia, 1982], out of which a summarized version was published in 1983, [Kimemia and Gershwin, 1983]. The problem addressed was that of controlling

a production system with multiple machines and multiple part types each subject to deterministic demand rates, where the machines were prone to failures. A multilevel hierarchical control algorithm was proposed, involving a stochastic optimal control problem at the top level.

The hierarchy was proposed taking into account the several time scales at which several classes of events take place. The shortest time period is that of the setup when switching among the family of operations for which a machine is configured. It is assumed that these times are short when compared with the remaining times and are ignored from the analysis. The next time period is that of the typical operation, assumed to be several orders of magnitude above the setup times. If the operation times are random, they are replaced by their means in the formulation. Next in the time scale come events like machine failures and machine repair times, assumed to be exponentially distributed and defined in terms of *Mean Time Between Failures — MTBF* and *Mean Time To Repair — MTTR*. The longest time period is the planning horizon for the problem under consideration. It is assumed that demand is known and constant for periods of time larger than the typical MTBF or MTTR.

The hierarchical controller was defined as having three levels, each level dealing with events of a particular time scale. Therefore, at the lowest level is a *Sequence Controller*, which schedules times at which to dispatch parts already in the system, and a *Routing Controller*, which calculates route splits for parts having more than one alternative processing path. At the next level of the hierarchy lies the *Flow Controller*, which determines the short-term production rates for each part type that are feasible for the current machine status. At the highest level, the off-line calculation of the control policies to be used in the flow and routing levels is executed, thus generating the *Decision Tables*.

An optimal control problem is formulated at the top level. One component of the system state is defined as the production surplus, i.e., the difference between the cumulative output of end products and the cumulative demand. Since the demand is defined as a deterministic rate, d, the system dynamics is described by a continuous time differential equation of the type

$$\dot{x}(t) = u(t) - d, \tag{2.1}$$

where $x(t) \in R^N$ denotes the production surplus, N is the number of different product types, and $u(t) \in R^N$ is the production rate which controls the system. The system is composed of M

workstations, each subject to random failures and repair times. The other state component, $\alpha(t)$, describes the available capacity in terms of the state of the workstations. This particular state component is modeled by means of a continuous time finite state *Markov chain*. At any given moment there is a set of feasible production rates defined as $\Omega[\alpha(t)]$. The cost function to be minimized is of the form

$$J(x, \alpha, t_0) = \mathbf{E}\left\{ \int_{t_0}^{t_f} g[x(t)]dt | x(t_0) = x, \alpha(t_0) = \alpha \right\}. \tag{2.2}$$

The objective is to determine the optimal production rates so that (2.2) is minimized and such that the production rates are feasible for the system's capacity at all times in the interval $[t_0, t_f]$, i.e., $u(t) \in \Omega[\alpha(t)]$. The expected value is taken over all possible trajectories of the Markov chain. To solve the problem, a *Dynamic Programming* algorithm is formulated and the *Hamilton-Jacobi-Bellman (HJB)* equation is derived. For general references on continuous time dynamic programming and on the derivation of the HJB equation see [Bellman, 1957, Bryson and Ho, 1969].

The HJB equation generates a *Linear Programming* problem on the production rates. The coefficients of the linear program are the derivatives of the optimal cost function with respect to the x components. If these coefficients are calculated, then the problem of determining the optimal production rates is very simple. The underlying Markov chain divides the state space into regions where the optimal production rates are constant. These rates will correspond to the extreme points of $\Omega(\alpha)$ on each region.

Next the authors note that, for each feasible region and for the stationary problem, there is a fixed buffer level x_α^H above which the optimal production rates are zero. A feasible region is defined as a region of the state space for which $u(t) = d$ is a feasible value for the control. This point (x_α^H) is denominated as the *hedging point* and is such that, if the Markov chain remains in a given state long enough, the buffer value will converge to its corresponding hedging point and will remain there as long as there are no changes on $\alpha(t)$.

The interpretation of the hedging point value is that of a base stock. For each feasible region, x_α^H is the value to hedge against machine failures, to compensate for the periods where the demand rate is not feasible. Note that, since the formulation only takes into account the production surplus for end products, nothing can be said about the values of the internal buffers. Although the authors are dealing with a multiple machine system, their conclusion is the same as it would be for a case

with single machine. This feature, the existence of a base stock in terms of end products, is also present for systems where the cost of inventory on internal buffers is accounted for, as will be discussed ahead.

The difficulty of this approach lies in the determination of the coefficients for the linear program. Their calculation requires the solution of a coupled set of differential equations, which can be done for very small problems, but is not recommended for realistic size problems, as the authors mention in [Kimemia and Gershwin, 1983]. So, they propose to estimate the value function and use the estimated function to generate the coefficients of the linear program. An upper and a lower bound on the exact value function is derived. The methodology proposed consists of determining and storing off-line the lower and upper bounds on the value function for each value of the production surplus and machine status. On-line, whenever a machine state change occurs, the control $u(t)$ is determined by a linear program using the stored values of the estimates.

The general control policy is such that, for feasible states the optimal control is to produce at the maximum possible rate for all products below their hedging points, to produce nothing when above the hedging points, and to produce at the demand rate when the production surplus equals the hedging point. For infeasible states it is not possible to recover to the hedging points, so the optimal control is to produce at the maximum possible rate while giving preference to products which are more important in terms of their individual costs. This by no means signifies that a fixed priority list is defined.

An explanation in terms of the general principles of push and pull for production systems is immediately available for this policy. The system is operated in a push mode when under the production target and on a pull mode when matching it.

Once the production rates are determined for each state, it is up to the lower level to decide the release of new parts into the system, in a manner compatible with such rates. Since the cost coefficients are functions of the surplus, the linear program is solved periodically, at each time step, to generate the adequate production rates.

Immediately following the work just described, there was a significant body of research developed to improve over the basic principles of [Kimemia, 1982]. As the main shortcoming of the approach has to do with the calculation of the value function estimates, some heuristic approximations were proposed in order to simplify the off-line computation. Others have treated richer models that take

the internal buffers into account.

Dealing with the exact same problem of [Kimemia and Gershwin, 1983], [Gershwin et al., 1985] propose a heuristic procedure to estimate the value of the optimal cost function. The authors take a quadratic approximation of the optimal cost function, or the *cost-to-go function*, and propose its structure to be

$$J(x,\alpha) = \frac{1}{2}x^{\mathrm{T}}A(\alpha)x + b(\alpha)^{\mathrm{T}}x + c(\alpha), \tag{2.3}$$

where $A(\alpha)$ is a positive definite diagonal matrix, $b(\alpha)$ is a vector, and $c(\alpha)$ is a scalar. With this cost-to-go function, the determination of the hedging points is simple, since it is enough to take derivatives of (2.3) with respect to x and set them to zero. So, the main issue is the determination of the values for the matrix, $A(\alpha)$, and the vector, $b(\alpha)$. To obtain such values, the authors analyze a sample path on the surplus trajectory using the MTBF and the MTTR values and derive the optimal hedging point for each product type, on regions for which demand is feasible. The derivation of the hedging points is made by means of cost considerations, like holding cost for positive surplus and backlog cost for negative surplus. This procedure determines the matrix and the vector of (2.3).

In [Kimemia and Gershwin, 1983], the linear program to determine the optimal production rates was solved every time step, which was assumed to be of one minute, because the cost coefficients of the linear program are functions of x. As observed in [Kimemia, 1982] this induces *chattering* when crossing certain boundaries between regions, since the change in optimal rates may be bigger than the actual loading of parts into the system, thus even leading the system to fail to meet demand when demands are close to capacity. In [Gershwin et al., 1985] it is observed that, although the cost coefficients change with x, there are well defined regions for which the *optimal basic solution* of the linear program does not change. Therefore, it is possible to compute the future behavior of the surplus variables, *projected trajectory*, once the boundaries that produce a change in the optimal basic solution are determined. The low level release is done through comparison between the projected trajectory and the realized trajectory.

By doing this the authors avoid the generation of a solution for the linear program every time period, and avoid the chattering behavior. These same ideas are discussed in [Akella et al., 1984], where a simulation study for a real life flexible manufacturing system is presented, comparing

the hierarchical policy with other heuristic based policies. In [Akella et al., 1984], an intuitive description of the meaning of the linear program cost coefficients is made, which illustrates how to obtain them. The linear program cost coefficients are defined to be of the form

$$c_j(x_j) = A_j(\alpha)(x_j - H_j(\alpha)), \tag{2.4}$$

where $H_j(\alpha)$ is the hedging point for product type $j = 1, 2, \ldots N$ when the machine state is α, computed as described in [Gershwin et al., 1985], and $A_j(\alpha)$ is a positive quantity that reflects the relative value and vulnerability of each part type. A measure of the vulnerability is the number of machines a part type visits during production. Also, the smaller the MTBF of the visited machines, the more vulnerable a part type is. To simplify the analysis the authors in [Akella et al., 1984] propose to make $A_j(\alpha)$ equal to the number of machines a part type j visits.

Although the arguments for the simplifications are intuitively acceptable and they have been shown to produce very competitive performances, it should be stressed that the control parameters determined overlook many aspects of a system's performance. For instance, the procedure to compute the hedging points relies solely on the average values of the disturbances and uses a nominal failure/repair cycle. No effects of process variance are taken into account. It should be expected that hedging points are sensitive to the variance of the disturbances affecting the production system. It would also be desirable to have more sound procedures to determine the linear program cost coefficients, given that these calculations are done off-line.

In [Akella and Kumar, 1986], a version of the problem addressed in [Kimemia and Gershwin, 1983] is considered. There, a single machine, single product type system is studied. The authors manage to analytically determine the closed form expression for the cost-to-go and for the hedging point. The optimal policy for this system is naturally one of the critical number type, so that the system should produce at maximum rate when the production surplus is under the critical number, should produce at the demand rate when the production surplus matches exactly the critical number, and should produce nothing if the critical number is exceeded by the production surplus. There are no such elegant results for systems with a little more complexity, like more than one machine or more than one product type, on the same continuous time setting with deterministic demand. For single product single machine, where machines may have more than two states, it has been shown that the optimal policy is also of the hedging point type with multiple hedging

points, [Sharifnia, 1988]. The author derives equations for the steady state probability distribution of the surplus level. With these equations it is possible to compute the cost-to-go as a function of arbitrarily chosen hedging points and then use that as a tool for determining the optimal hedging points.

In summary, the problem of controlling a flexible manufacturing system with unreliable machines was formulated as a stochastic control problem in [Kimemia, 1982]. A multiple machine, multiple part type system was considered, but internal buffers were not treated explicitly. Both demand and processing times were assumed to be deterministic and fixed. Heuristic algorithms based on approximations of the value function were subsequently proposed in [Akella et al., 1984, Gershwin et al., 1985]. Exact solutions for a one machine, one part type system were derived in [Akella and Kumar, 1986] and [Bielecki and Kumar, 1988]. This latter work establishes and discusses conditions for which the optimal hedging point has zero value, even though there is uncertainty in the production system through the failures of the single machine.

After the work of [Kimemia, 1982], including the research just described, it can be said that two main branches have emerged. On one hand, some authors adopted the approach of trying to determine approximations for the optimal policies that are simple and exhibit competitive performances. On the other hand, some others continued on the path of approximating the cost function and extended the results to more complex and richer systems.

2.3.1 Approximating the control policies

In the area of determining approximations for the control policies, the work described in [van Ryzin, 1987, van Ryzin et al., 1993] has been the inspiration for many further developments. There, the internal buffers were first considered after [Kimemia, 1982]. As a first approach to include internal buffers into the formulation, a system with two machines in tandem producing a single part type was studied. Since, even for such simple model, the analytical solution is impossible for practical reasons, the authors proposed to numerically solve the discrete time version of the original problem. They discretize the differential equation of the system and use the *Value Iteration Algorithm* of Dynamic Programming as described in [Bertsekas, 1987].

The production surplus is defined as a vector in R^2, where the first component is the production surplus of the first machine and the second component is the production surplus of the second machine. Naturally $x_1(t) \geq x_2(t)$, since the inventory sitting in the buffer between machine one

and machine two is always non negative. This system's dynamic equation in continuous time is of the same type of (2.1) for both surplus variables. The discrete time counterpart is of the form

$$\begin{aligned} x_1[t+1] &= x_1[t] + u_1[t] - d \\ x_2[t+1] &= x_2[t] + u_2[t] - d \end{aligned} \tag{2.5}$$

where $u_2[t]$, besides being feasible for the machine status at time t, cannot exceed $x_1[t]$. They consider infinite horizon discounted costs. The inventory between machine one and machine two ($x_1[t] - x_2[t]$) incurs a holding cost and the inventory of finished goods ($x_2[t]$) incurs holding cost when positive and backlog cost when negative. Each of the two machines may be in one of two states: up or down. This particular dynamic behavior is modeled through a Markov chain as before.

The numerical results show that the state space, in terms of the surplus variables, is divided into regions. Fig. 2.1 displays the typical switching curves obtained for the situation where both machines are operational.

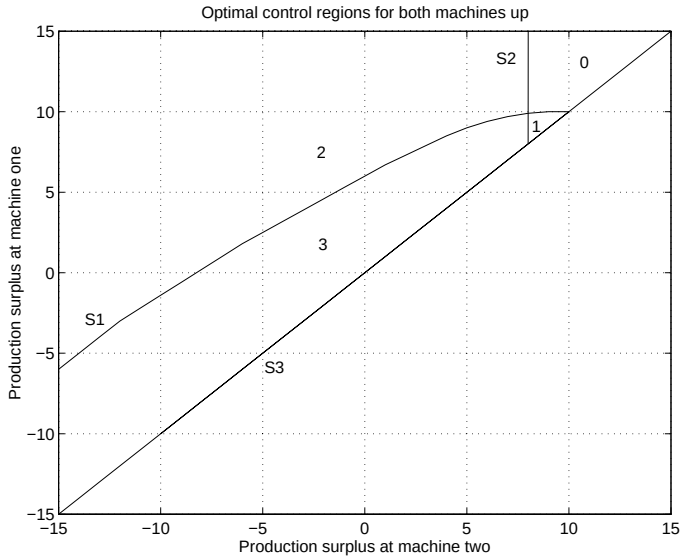


Figure 2.1: Typical optimal control regions.

Curve S3 marks the frontier of feasible states. The variable $x_1[t]$ is above $x_2[t]$ for points above S3. The two other switching curves divide the state space in four regions: 0, 1, 2, and 3. In region 0 it is optimal not to produce; in region 1 only machine one should produce at its maximum rate;

machine two produces at maximum rate alone in region 2; and both machines produce at maximum rate in region 3. The intersection of curves S1 and S2 determines the optimal hedging point for this system. That is, it is the point to which the state space converges if the machines remain up enough time. If machine one is down and machine two is up, the surplus space is divided into two regions only by a curve like S2. If machine two is down and machine one is up, the space is also divided into two regions by a curve like S1. Naturally, there will be no switching curve when both machines are down. Also, for the cases where only one machine is down the curve present does not match the corresponding curve when both machines are up. There is a slight sliding downwards of S1 closer to S3, for instance. If the state lies on S1 or S2, which are called attractive boundaries, the optimal control is such that the system should remain on the curves and move along them towards the hedging point. This is when the optimal control rate matches the demand rate.

Although their general shape does not change with the system parameters, the curves will change as a function of the specific holding and backlog costs used, as well as they are influenced by the MTBF and MTTR, and depend on the discount rate used for each fixed demand rate under the system's capacity. Changing the state representation from surplus values to inventory values, Fig. 2.1 assumes the, perhaps more familiar, form shown in Fig. 2.2.

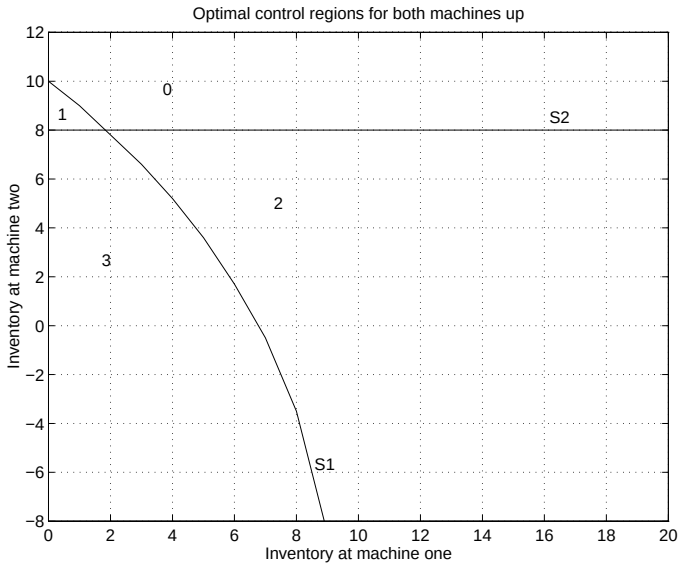


Figure 2.2: Typical optimal control regions.

These switching curves, as displayed in Fig. 2.2 do not differ much from the ones obtained in

[Veatch and Wein, 1994] and in [Bispo, 1992]. In [Veatch and Wein, 1994], a two station in tandem system was considered also. Their model assumed reliable machines and an exogenous Poisson demand process. The service times were assumed to be random and exponentially distributed. They computed the optimal machine rates using Dynamic Programming, and characterized some general properties of the switching curves. Namely, the slope of S1 is always below -1 and converges to $-\infty$; S2 is not constant for all values of the inventory at machine one, although it rapidly stabilizes at a constant value; S2 has a non negative slope; and both curves intersect for non negative values of the state variables. The discrepancies between these findings on S2 and those displayed may have to do with the fact that in [van Ryzin et al., 1993] the discount factor used was very close to zero. Using a discount factor closer to one will allow the conclusion that S2 is not constant for all values of the inventory at machine one. In [Bispo, 1992], the same two station system of [Veatch and Wein, 1994] was used. The main differences were the inclusion of machine failures into the model and the last machine had batch capabilities. The same switching curves were found by means of Dynamic Programming, and it was shown they obeyed the same generic properties as those of [Veatch and Wein, 1994].

Although obtained for systems subjected to different types of random disturbances, the optimal policies possess the same structural properties. Namely, the existence of a hedging point, or optimal base stock value for both variables, to which the state space converges after some time and where it attempts to remain. Given that the switching curves, besides their structural properties, are not amenable to be used for practical reasons, [van Ryzin et al., 1993] proposes the utilization of approximate switching curves. The basis of the approximation is to observe that curve S1 tends to be parallel to S3 (Fig. 2.1) as the production surplus of machine two approaches $-\infty$ and that it tends to be flat as it approaches curve S3 for positive values of the surplus at machine two.

Fig. 2.3 displays the structure of the approximation proposed, for both alternative representations of the state variables. Curve S2 is constant, defining a base stock for the surplus of machine two, and curve S1 is composed of two linear segments. On the first segment (closer to the hedging point) there is a base stock for the production surplus of machine one and on the second segment there is a base stock for the inventory between machine one and machine two. The intuitive reason for such approximation, that also explains the structure of the optimal switching curves, is that when the surplus of machine one is negative with high absolute value, there is no point in producing more than enough to keep machine two working. Any excess inventory will have to sit waiting for

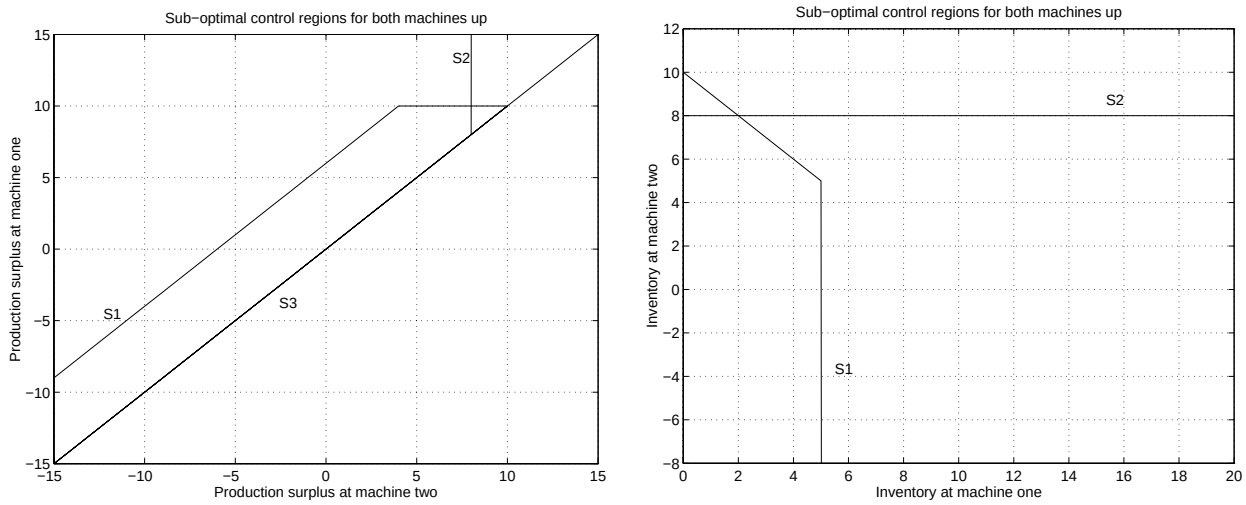


Figure 2.3: Sub-optimal approximation for the switching curves.

its turn to get processed and incurring needless holding cost while waiting. Once the state is close enough to the hedging point, the need for inventory between the two machines decreases and the system works towards the hedging point by exchanging inventory from one machine to the other. Although the optimal switching curves change for different combinations of machine status, they use the same approximation curves for all cases.

According to [van Ryzin et al., 1993], the control policy is such that machine two follows a pure surplus-level control and machine one follows a surplus-level control in the region where x_2 is close to zero and a buffer-level control where x_2 is very negative. The overall control policy is termed *two-boundary* control policy, because it assumes that the optimal control regions are divided by two piecewise linear boundaries.

In order to determine the optimal parameters that characterize the two-boundary control policy the following set of approximations is proposed:

1. Replacing hedging points that are machine-state-dependent by those that are independent of the machine states;

2. Using time averaged values of functions in place of the functions themselves;

3. Incorporating the effects of an empty buffer as a modified failure rate.

The above simplifications are then used in combination with probabilistic arguments to derive the parameters that fully define the control policy.

Following this work, a system with three machines in tandem was analyzed and the two-boundary control policy extended to a system with one part type and N machines in tandem, [Lou and van Ryzin, 1989]. The essentials of the extended policy are:

1. For each machine define an inventory hedging point h_i, and a surplus hedging point h_{s_i};
2. At any given instant of time compute the actual machine inventory x_i, as the sum of the number of parts in its buffer and the number of parts being processed;
3. Compute the actual machine surplus s_i, as the difference between cumulative production and cumulative demand, or the sum of the inventories of all the downstream machines;
4. If $x_i \leq h_i$, $s_i \leq h_{s_i}$, and the machine is operational, load machine i at full rate. Otherwise do not load.

Note that such policy takes care of two types of decisions: *release decisions* — loading of new parts into the system, and *scheduling decisions* — loading of existing parts into the machines.

This policy is used to control a re-entrant system with multiple machines and a single part type in [Lou and Kager, 1989]. Although there is a re-entrant structure, the authors consider the system as a series of machines in tandem and determine the policy parameters for such structure. It is not clear from their analysis and discussion how the capacity of the re-entrant machines is allocated to each visit for the purpose of calculating the parameters. In terms of the dynamic management of capacity, they award capacity to parts closer to being completed, thus implementing a priority based capacity management. A simulation study is presented comparing the extended two-boundary control policy with an open loop control policy that loads a fixed number of parts into the system periodically. Their results show a clear advantage of the two-boundary policy over the open loop policy. However, their study is biased by the fact that they start collecting statistics from time zero, and start with an empty system. The open loop policy loads parts into the system at a rate that matches demand. Therefore, it will always lag demand since it starts with zero inventories at all buffers. Once the system becomes fully loaded, the open loop policy stabilizes its behavior, but remains with negative surplus for end products. The two-boundary policy manages to recover to positive surplus because it is an intrinsically closed loop policy. Naturally, a closed

loop policy is always expected to perform at least as weel as an open loop policy. But, when the loading rate has not been optimized for the open loop policy and when the system is started from zero, the performance differences between the two policies will be higher than they truly are.

Immediately following [Lou and Kager, 1989], a semiconductor manufacturing system with re-entrancy, random yield, and multiple parts was addressed in [Lou et al., 1990]. A more extensive description of it was later published in [Yan et al., 1996]. The authors used the two-boundary control policy, with the parameters determined for a so called "virtual flow line". That is, the re-entrant system is converted into a flow line with no sharing of capacity for the purpose of determining the optimal parameters of the control policy.

On the model of [Yan et al., 1996], the authors do not consider the fact that each product type on different visits imposes different loads on the machines. Also they deal with random yield by scaling down the machine capacities by the average yield and simulate the system as having perfect yield for those capacities. This way, all products that are loaded into the system will eventually get out. Although this scaling may represent the average of the output process with some accuracy, it does not seem appropriate to model random yield this way since there is no accounting for lost parts and the effects they have on the process variance, and consequently on the actual performances.

Given that [Yan et al., 1996] deal with multiple part types, it is necessary to address the problem of allocating capacity to the various products that compete for the same resources. To do that the authors assign a set of weights to each part type and a set of additional weights to each entry of each part type on the re-entrant structure. This second set of weights is non decreasing as a part type approaches the last processing steps. The combination of the two sets of weights determines the procedure to order the different part types and different entries. The overall procedure is a mix of priority based allocation and a shortfall based allocation. Priority for the release of new parts into the system is given to the highest weighted global surplus deviation. The first set of weights alone is used to produce the ordering. Priority for the local dispatch into the machines is given to the highest weighted local surplus deviation. The two sets of weights are used for this ordering. The procedure to determine the specific weights is not discussed. The policies are shown, through simulation, to be relatively robust to changes in the nominal values of the system's parameters, such as the mean up times of the machines.

Many authors have proposed different methodologies to determine the optimal parameters of the two-boundary control policy. [van Ryzin et al., 1993] refers a probabilistic approach, which

makes use of the steady state probability distribution of the surplus, and a heuristic approach, that produces values close to the actual optimal parameter values computed in [van Ryzin et al., 1993] (see references therein).

Of particular interest to the present thesis are the IPA based approaches. [Yan et al., 1992] validate a continuous time perturbation analysis based optimization procedure to determine the optimal hedging points for systems with one and two machines, no re-entrancy, and a single part type. A more extensive report has been presented in [Yan et al., 1994], where a simulation study is also presented. The authors use perturbation analysis to compute the two-boundary control parameters and to compute the thresholds for a Kanban policy. A very similar idea is the work reported in [Caramanis and Liberopoulos, 1992]. The authors present and validate an IPA algorithm to determine the optimal hedging points for the model of [Kimemia and Gershwin, 1983], that is, only final buffers are taken into account to compute cost. They approximate the cost by a quadratic function of the hedging points. The paper reports on an application for two machines and two products. The authors discuss a framework to deal with more products by means of aggregation to two-product problems.

Along the lines of the hierarchical controller of [Kimemia and Gershwin, 1983] there has been some work on the design of controllers that can skip the linear program calculation, [Custódio et al., 1994]. There, the hedging points are determined through a procedure similar to that of [Gershwin et al., 1985]. The differences lie in the middle and lower level controllers. The middle level controller determines the loading rates through a fuzzy controller that attempts to track the cumulative demand, while keeping a low work-in-process. The lower level controls the flow of parts among the resources, using a modified version of Yager's fuzzy decision method. For general references on fuzzy sets, fuzzy logic, fuzzy controllers, and Yager's method see [Zadeh, 1965, Yager, 1978, Berenji, 1992].

2.3.2 Approximating the value function

The alternative approach to directly finding approximations to the optimal control policies, is to approximate them through an estimate of the optimal cost function. That has been done in several instances. The work reported in [Akella et al., 1984, Gershwin et al., 1985] and described above is one such case. In [Bai and Gershwin, 1995] a tandem system with N machines, $N-1$ buffers, and a single part type is analyzed. The multiple part type version of this tandom system

is discussed in [Bai and Gershwin, 1994], and a re-entrant system is studied in [Bai and Gershwin, 1996].

A policy similar to the two-boundary control policy is developed for each machine and each part type. Three parameters for each machine and each part type characterize the policy: the local surplus, the buffer level, and the buffer size. Following the procedure of [Akella et al., 1984] and [Gershwin et al., 1985], they approximate the value function of the dynamic programming problem with a quadratic. This, and the upper bound on the buffer level, lead to regions whose boundaries are straight lines. By further assuming that production rate decisions depend only on the local machine state, local surplus, and the buffer level, the boundaries will be parallel to the coordinate axes in surplus space, and there will be only three unknown parameters for the two-machine problem: the hedging points for each surplus, and the maximum buffer level.

The determination of the hedging points is selected in [Gershwin et al., 1985] to approximately minimize the integral of a penalty over a typical repair-failure cycle, in which different positive costs are assessed for positive and negative values of the surplus. Whereas in [Gershwin et al., 1985] the surplus was only conditioned by the repair states of the machines, for limited buffers it will be also affected by whether the machine is blocked, starved, or neither. Therefore, they have to estimate the fraction of time that the buffers are full or empty. The approximately optimal hedging points and buffer size are functions of these fractions through an optimization procedure.

The extension of these results to a system with N machines and a single part type is done by decomposing the original problem into a series of two-machine problems, [Bai and Gershwin, 1995]. Again, they must calculate the control parameters: one buffer size per buffer and one hedging point per machine. The calculations are similar to those of the two-machine case, with the following exception: the fraction of time the machine $i+1$ is starved is expressed as a function of the probability that machine i is down due to its own failures and the fraction of time that machine i is starved (among other things). Blockage probabilities propagate similarly, but in the opposite direction. This leads to a set of nonlinear equations which are used as constraints in an optimization problem. The decision variables are one buffer size per buffer and one hedging point per machine.

When considering systems with multiple part types, the extension of results is done by evaluating the blockage and starvation probabilities in approximately equivalent systems, [Bai and Gershwin, 1994]. These are N-machine systems with only one part type, whose parameters are related to those

of the original system. The blockage and starvation probabilities are used to calculate the optimal buffer sizes and hedging points. Finally, when dealing with re-entrant systems, this procedure is extended to systems with multiple machines, multiple part types, and arbitrary, different, but fixed routing for each part type, [Bai and Gershwin, 1996]. Again, an approximately equivalent system is developed for the purpose of estimating blockage and starvation probabilities, which will then enable the computation of the optimal buffer sizes and hedging points.

2.3.3 Summary

The area of flow rate control has produced one of the most comprehensive and complete methodologies to deal with production planning and control for manufacturing systems. It involves long term decisions, such as determining safety levels for the production, based on long term data like demand rate and machine statistical behavior. It involves middle term decisions, such as the release of new parts into the system, based on current system status described in terms of actual production and availability of machines. And finally, it involves short term decisions, such as scheduling parts into the available machines, based on current production evolution.

The approach relies on the sound theory of optimal control and benefits from its elegant results, generating control policies that are functions of the system state, operating in closed loop. The main advantage of closed loop policies is their robustness to small changes and disturbances that may occur in the nominal parameters of the system under control.

There are however some questions which remain to be dealt with in a more satisfactory way. The only source of randomness formally considered is machine breakdowns. Demand variance is not taken into account, as well as the effects of random yield and processing time uncertainty, in order to determine the long term safety levels. The calculation of the safety levels is typically a very complex procedure, or the heuristic procedures to simplify their computation rely on somewhat simplistic approximations to smaller dimension systems. The possibility for, and different means of, sharing capacity is not taken into account to determine the safety levels. Also, there is no effort made to determine how capacity should be shared, nor on studying the effects that different dynamic capacity schemes have on the performance of the manufacturing systems. When it comes to sharing capacity, typically the question is answered by means of proposing priority schemes of greater or lesser complexity, which are neither derived from, nor a consequence of, the theoretical model.

The attempts to solve the higher level problem without resorting to the explicit integration of the HJB equation but through simulation based approaches have always been done for very small systems. The other methodologies rely on very detailed and elaborate statistical arguments based on the particular statistical distributions assumed.

2.4 Queueing Networks

Queueing networks have long been one of the most used tools to model systems possessing discrete event dynamics. Systems with discrete event dynamics are characterized by the fact that their state does not change continuously as time evolves, but rather it jumps from one point to another at a particular instant of time, when it is said that some event occurred. The instants of time at which the state changes, or at which events occur, are random. Events that trigger the state change are as various as the type of systems one can consider, but usually they are associated with the fact that some sort of activity ends. For instance, and in the context of manufacturing, events can be a machine finishing an operation on a particular part, a machine breaking down, a machine becoming available after being repaired, or a new part arriving into the system.

Given that the occurrence of events is associated with some activity coming to the end, it is natural to say that there will be some sort of entity responsible for the execution of the activity. Also, an activity involves the cooperation of some entities within the system. Entities involved in one activity will become free to initiate other activities with some other entities, after the current activity terminates. Some of the entities are mobile, some are fixed, and some other may be virtual.

Queueing networks capture this type of dynamics by the definition of two types of entities: servers and customers. Customers are assumed to be mobile entities in the sense that they are external to the system. They arrive into the system to get some service done in one or more servers, wait for their turn, and then depart upon completion of their service requirements. A rich variety of discrete event dynamic systems is modeled by means of the parts arrival processes, service time distributions, and routing of parts along the system. Servers are assumed to be fixed entities in the sense that they belong to the system and provide the customers with some sort of service, according to the customers' requirements. The discrete event dynamic systems are modeled by means of the number of servers in the system, how they interact with other servers, and how they operate over the customers waiting to be served. Besides these features, other features of the

queueing networks help in the modeling of very generic systems such as the queue discipline, the existence or absence of limited buffers for queues, and how the input of new customers into the system is done.

Consequently, the simulation of production systems through networks of queues can provide very accurate estimates of the performance of real life systems. By means of both closed or open queues, many contributions have been made to the understanding of the structure of manufacturing systems. Although, in general, networks of queues have been used to estimate performance, some effort has been made in the direction of using queueing models as a way to determine control policies with some success, namely in the context of re-entrant systems.

Typically, queueing theory has been centered on determining long term measures like average waiting time for the customers, average server utilization, departure rate, etc. In the context of manufacturing systems, this type of tool has been used mainly for performance evaluation, for instance, to evaluate specific configurations or to compare the performance of alternative service disciplines. In many instances simulation goes hand in hand with this modeling tool, in order to provide this type of answer for very complex systems.

The main contributions to manufacturing control, in general, and to semiconductor manufacturing control, in particular, can be classified in two groups depending on how the population of the underlying networks is assumed to evolve. There are authors who address manufacturing control by modeling the systems as open queueing networks, where there is no control over the amount of customers that get into the system. Others impose some sort of loading control and model their systems through closed queueing networks, where the number of customers in the system is either fixed and constant or bounded above by some value.

If the model used is that of open queueing systems, the questions typically posed can be in terms of determining the number of servers to be used so that some performance can be achieved, determining the rate of service of the servers, evaluating the performance of a system, comparing the performance of alternative routing schemes and/or queueing disciplines. When using closed queueing networks the questions addressed also deal with the evaluation of input regulating policies. In this thesis' perspective, it is assumed that questions related to the design of the production system have been answered. Therefore, the interest resides in the evaluation of performance for alternative queueing disciplines and input regulating policies.

For general references on queueing networks theory, see [Kelly, 1979, Assmussen, 1987, Walrand, 1988]. An excellent text on discrete event dynamic systems is [Cassandras, 1993].

2.4.1 Closed Queuing Networks

One good example of the use of queueing networks theory to model the performance of semiconductor manufacturing is the work of [Chen et al., 1988]. More specifically, they concentrated on analyzing the quality of queueing networks theory as a tool to evaluate the performance of the wafer fabrication stage. There, the authors say that

> "The most obvious means to generate such performance predictors is the Monte Carlo simulation, but experience in other areas suggests that mathematically tractable queueing network models, although less flexible than simulation models and based on apparently restrictive assumptions, are far easier to use, generate more qualitative insight with respect to essential system relationships, and are accurate enough to provide quantitative guidance to system designers."

By making use of the classical theory of product form queueing networks, [Kelly, 1979, Baskett et al., 1975], they use past history data of a particular wafer fabrication system to estimate some of the parameters of the queueing network. They adopt a mixed network model, which is closed with respect to *nonmonitor lots* and open with respect to *monitor lots*. The past operating data is used to estimate parameters such as the population size of the nonmonitor lots, the fraction of nonmonitor lots for each type, the average input rate for monitor lots, the effective service rate at each station, etc.

With these estimated parameters they are able to predict performance measures by means of queueing network theory. The measures predicted are the overall average throughput for nonmonitor lots and the average cycle time for the lots. Cross checking predicted measures against past performance the authors show the adequacy of the queueing network model. The errors between predicted and actual performances have a maximal deviation of 14%. These results are quite remarkable given the crude approximations made on the model, like assuming that the service times are described by exponential distributions and arrivals are Poisson, for instance. There are many features of the manufacturing process not addressed in their model and the authors discuss these limitations. Among the missing features are equipment failures, scheduled off-periods, and the restrictive distributional assumptions.

Of particular relevance for this thesis is the research on closed queueing that goes beyond the simple evaluation of performance or validation of queueing networks as a modeling tool, to address the issue of controlling the input of new customers into the system. It is a known fact that the higher the variance of the several processes that characterize a production system the poorer the performance. One of the main sources of randomness is the arrival of new customers into the system. Open queueing networks assume that there is no control over the arrival process, or assume that such control is done somewhere else. Thus, the performance of an open queueing system subject to a random arrival process is highly conditioned by it. In real life production systems, there is usually the possibility for regulating the input of new parts into the system, which may have the effect of reducing the overall variance of the production system and therefore reducing costs.

Regulating the input of new parts into a production system, while keeping the amount of material in the system within relatively moderate bounds reduces cycle time, given that the queues in front of each server are small and the waiting time for service is reduced as established by Little's law, [Little, 1961]. The importance of low cycle times in the context of wafer fabrication has already been stressed, due to the correlation between high cycle times and low yields.

The work of [Resende, 1987] deals explicitly with the issue of regulating the input of new material into a production system. The objective is to release as little as possible so as to keep a low inventory while maintaining high throughput. The approach relies on the definition of a bottleneck machine as the machine with the highest load in the production line. The policy, called *starvation avoidance* is a closed loop release control policy. A compact version of the approach is presented in [Glassey and Resende, 1988].

The goal is to minimize cycle time subject to achieving some throughput level on wafer fabs. Their approach concentrates on releasing just enough material into the system so that the existing bottleneck machine will not be idle. The model includes machine failures as the source of randomness and the demand rate is assumed constant. Simulation in a single product type setting is the mode of study.

They compute the expected time to exhaust the existing stock and the expected time to replenish stock. Whenever this latter is larger than the former, a stock out is expected to occur. Therefore, to avoid stock outs they reorder whenever the existing inventory is below some threshold. The determination of this threshold is somewhat heuristic, but it relies on the basic principle of defining a safety level to cover for the uncertainties. The best level is found through trial and error by

simulation.

In order to determine when to release new parts into the system they compute the *virtual inventory*, W, defined as the work content of all parts/jobs either at the bottleneck machine or expected to arrive to the bottleneck within a given lead time. The lead time to replenish, L, is defined as the total processing time before the first visit to the bottleneck machine. A decision to release a new part is made when $W < \alpha L$, where $\alpha > 0$ is a safety factor. To attain an efficient tradeoff frontier between idle time at the bottleneck and the inventory level is the objective. Each chosen α corresponds to a particular point of this tradeoff curve.

The starvation avoidance approach is tested against some other release rules. The starvation avoidance methodology defines the release policy but does not provide any answers relative to the local scheduling for each machine. The study made compares several release policies operating with several local scheduling rules. The importance of release rules as compared to local scheduling is demonstrated by the experimental results. The choice of release policy has a bigger impact on the performance than the choice of scheduling rule. Regarding the scheduling rules, the two best performers are: give higher priority to the parts which are closer to their conclusion and a more complex rule that weights this priority and priority given to parts that are heading to the bottleneck machine. The operational objective is to get end products out of the system as fast as possible, making the best use of the bottleneck machine, i.e., maximizing throughput. Note that this concept of starvation avoidance is in some measure closely related to the concept of a base stock.

In [Lozinski and Glassey, 1988], a graphical package designed to help monitor the implementation of starvation avoidance is presented together with some simulation results.

Still in the context of release rules is the work on Brownian motion control. Whereas the prior approach uses queueing networks to model a manufacturing system, and simulates such a network to get estimates on relative performances, the Brownian motion models attempt to determine optimal release and scheduling policies. A good reference on the development of the Brownian motion model can be found in [Harrison, 1988].

The problem of determining the optimal policies for a two-station network processing multiple classes of customers is discussed in [Wein, 1990b]. The general methodology is as follows:

1. approximate a queueing network scheduling problem by a dynamic control problem for a Brownian network;

2. reformulate the Brownian network control problem in terms of workloads;

3. solve the workload formulation;

4. interpret the solution of the workload formulation in terms of the queueing system to obtain an effective scheduling policy for the original queueing network scheduling problem.

The workload formulation is formally solved in [Wein, 1990a] and the remaining steps are discussed in [Wein, 1990b]. The scheduling policy derived from this procedure consists of a *sequencing rule* and an *input rule*. The sequencing rule addresses the choices of customers in the queue when a server becomes idle, and the input rule addresses the choices of instants to load new customers into the system. The sequencing rule is termed *dynamic reduced cost policy* and the input rule is termed *workload regulating policy*.

Prior to this work, the literature on input control for queueing networks considered decisions of whether to accept or reject Poisson arrivals, as presented in the extensive survey of [Stidham, Jr., 1985]. According to [Wein, 1990b], the emphasis should not be placed on whether to accept or reject a customer but rather on when to allow a customer to enter the system. The line of research started by [Kimemia, 1982] is a good example of this latter concern, although they do not make use of queueing networks as the modeling paradigm.

The sequencing rule computes dynamic reduced costs from a linear program for each class of customers and at all times gives priority at each station to the class with the largest reduced cost. The input rule releases a new customer into the system whenever either server appears to be threatened with idleness and there is not too much work already in the system.

The dynamic reduced cost policy is reminiscent of the *workload balancing sequencing policy* derived in [Harrison and Wein, 1990]. Also in a two-station multiple customer classes setting, the authors follow a methodology similar to that of [Wein, 1990b] to address the problem of determining the optimal sequencing rule, given that the population of the queueing network is fixed at some constant level. In [Wein, 1990b] the population size can vary over time, although it remains bounded due to the structure of the input rule. The workload balancing policy is shown to be asymptotically optimal in heavy traffic for the problem of maximizing server utilization, and thus it maximizes

throughput. Basically, it awards priority to customers so that the high priority choices favor the balance of the work content for each station's queue. By means of a small experimental study, this sequencing rule is shown to achieve lower cycle times for the same throughput than *First Come First Out — FIFO*, *Shortest Processing Time — SPT*, and *Shortest Remaining Processing Time — SRPT*. That is, it achieves the same throughput as that of those sequencing rules with a smaller population size.

With the above as the theoretical foundation for input and sequencing rules, [Wein, 1988] reports an excellent statistical study in the specific context of semiconductor manufacturing. There, several input and sequencing rules are compared for the control of wafer fabs. The system under study produces a single product that requires 172 total operations at 24 different single or multi-server stations. A queueing model is assumed and the main objective is to determine which combination of input and sequencing rules incur lower average cycle time, where cycle time is measured as the time a lot takes from the moment it is released into the system until its completion. Also, the paper assumes the existence of stable bottleneck machines.

A major conclusion of the paper is that proper choice of input rules have more impact on the system's performance than the local sequencing rules, agreeing with the findings of [Glassey and Resende, 1988]. Moreover, the relative performance of the local sequencing rules seems to be dependent on the input rule utilized and also dependent on the number of bottleneck machines. The input policy that performs best in the study is the workload regulating policy derived in [Wein, 1990b]. The other input rules considered are random Poisson inputs, deterministic input, and closed loop input. The deterministic input releases customers into the system at a fixed constant rate matching the demand rate. The closed loop input releases a new customer whenever some customer leaves the system, thus maintaining a fixed population.

To take care of more than one bottleneck machine, the author constructs variants of the workload regulating policy that monitor the amount of work in the system for the bottlenecks. If there is a single bottleneck, a new lot is released into the fab when the expected amount of work for the bottleneck machine drops below some fixed threshold. If there are two bottleneck machines, there are two thresholds that determine the release of new lots, one per machine. In the case of four bottleneck machines, there is an overall threshold for the expected amount of work for all of the bottlenecks. These are the cases considered.

[Ou and Wein, 1995] consider random yield, by-production, and several stages of service to extend the earlier work of [Wein, 1992a] for a single queue system. By using Brownian motion control under heavy traffic conditions, the authors derive a priority order for the several types of products and different visits to the single server. Due to the heavy traffic assumption the approach focuses on dispatch decisions for the bottleneck machine. The system is assumed to produce K types of products which are ordered by quality. The K product types are produced according to K different production processes. Process k, with $k = 1, \ldots K$, is used primarily to produce product type k. However, process k can also produce lower quality products due to the presence of random yield. The server is kept idle as long as the *optimal aggregate base stock level* has been reached and no product is backlogged. If there is at least one product backlogged then there is an ordering of products based on their backlog cost and machine consumption. That is, priority is given to products with high backlog costs and low machine consumption.

According to [Ou and Wein, 1995], inventory should be held only in the place that achieves the smallest holding cost per unit of expected machine time already invested — *minimum index location*. If there is no product backlogged, but the optimal aggregate base stock level has not been reached, the scheduler should build up inventory in the minimum index location. If such location is one of WIP inventory, priority should be given to the feeding levels, giving higher priority to levels closer to the minimum index location. If there is no feeding WIP inventory, a new part should be released into the system. If the location is of finished goods inventory, all the WIP inventory is to be cleared and converted into finished goods inventory, following the same priority scheme.

Since the approach is non anticipating, it only allows the triggering of production to be made when products are backordered, the system is not as responsive as it should. To circumvent that, the authors propose to trigger production when a product is in *danger of being backordered*, to add an anticipating feature to the policy.

Although motivated by different interests, the similarities between the starvation avoidance approach and the Brownian motion based approach are highly visible in the structure of their proposed release policies. The concept of safety stock is present, either coming naturally from the formulation or *a priori* imposed to ensure some degree of quality for the resulting policies. On a critical evaluation of the workload regulating policy and the starvation avoidance policy, [Glassey and Resende, 1988] make two interesting observations which are worthy of reference and further discussion.

Observation 1. "Imagine a breakdown of the last workstation. Then no work leaves, so under Fixed-WIP no new work starts, and, if the breakdown lasts long enough, the bottleneck starves. But both Starvation Avoidance and Workload Regulating ignore all lots that have passed the bottleneck for the last time and continue to feed work into the bottleneck. Similarly, if the bottleneck station breaks down, Fixed-WIP will continue to pile up the inventory of new work in front of the bottleneck (until everything after the bottleneck has left the shop), but both Starvation Avoidance and Workload Regulating will stop releases."

Observation 2. "In reentrant flows, Workload Regulating counts all the work remaining at the bottleneck on each lot, not just the next bottleneck operation. Let $L = 0.9$h[1] and consider two cases of an almost empty shop. In case 1, two jobs are in queue at the bottleneck, each with 0.5h of work at the next bottleneck operation. Starvation Avoidance would not start a new lot (assuming $\alpha = 1$). Now suppose only one job is in queue with two bottleneck operations to be performed, each taking 0.5h, and 1h of processing on some other machine between them. Workload Regulating would not distinguish between these cases, but Starvation Avoidance would start a new lot in the second case."

Regarding observation 1, if a machine breaks down long enough so that a bottleneck is starved, the inference should be that the machine broken down becomes a bottleneck. That is, there has been a dynamic shift on which machine is the bottleneck. What makes sense to do is to ensure that every machine stops after some time. A release policy should not want to keep downloading new material into the shop for the sake of the *assumed* fixed bottleneck, since there is a new bottleneck. There should be a point after which all machines should stop processing, if there is a breakdown lasting long enough on the line. Neither policy seems to ensure that, unless the breakdown occurs for the fixed bottleneck or for some upstream machine.

On observation 2, the authors have indeed a good point given that workload regulating ignores work between visits. However, their example is strange in some sense. Workload regulating is derived under heavy traffic assumptions. In heavy traffic there is almost always some job whose next operation is on the bottleneck machine. Their example is for an almost empty shop. An almost empty shop is not operating under heavy traffic, and in such a situation one should expect the bottleneck machine to be naturally and desirably idle. Given that the input of new material

[1] This means hours in the original paper.

is controlled in a closed loop fashion to maximize utilization of the machines, only in a very odd situation would the system become so empty. Besides this, in their second case the authors say that there could be a job with two operations for the bottleneck machine that require a smaller amount of processing time each than the single non-bottleneck operation. One could, therefore, question if this is a normal situation when defining bottleneck machines. The definition of a bottleneck machine assumes that in fact the processing times on the non-bottleneck machines are much smaller than those at the bottleneck.

In any case, these two observations are revealing of both approaches' limitations for more general settings. Although there may exist one or a handful of long term bottleneck machines, there may be a dynamic short term shift of the bottlenecks in the presence of machine failures, as the flow rate control approach was so clear in considering. The presence of random yield and demand are also sources of short term shifting of the bottlenecks. A sound release policy has to be sensitive to these short term bottleneck shifts.

The modeling of manufacturing systems through closed queueing networks, while placing the emphasis on maximizing throughput (or minimizing cycle time), ignores the existence of an exogenous demand process. In [Dessouky and Leachman, 1994], an attempt is made at determining release policies for wafer fabs taking into account a demand process. There, it is argued that "notwithstanding the fact that minimizing the work in process inventory is a worthy objective, it is imperative to develop a schedule that can meet demand in a timely manner."

The authors formulate an integer programming model to determine optimal release dates and rates. The source of uncertainty is failure of the machines. A simulation study is presented comparing the policy generated by the integer program versus workload regulating and uniform release. The integer programming approach is shown to exhibit better performances in terms of operational costs at the expense of additional work in process inventory, when compared to the workload regulating approach. This should not come as a big surprise, since neither workload regulating nor uniform release attempt to track demand explicitly.

The proposed optimization model is to be resolved periodically or whenever a major change in factory status occurs, such as a major machine failure. The production schedule is thus updated to account for the factory status and demand. They quote [Bechte, 1988] as saying that

> "... changing the priorities at the queues in front of machines does not change their capacity and therefore can neither accelerate nor delay the manufacturing flow as a whole."

while placing emphasis on the importance of release over local scheduling. The experimental studies in [Glassey and Resende, 1988] and [Wein, 1988] seem to agree with this general statement. However, as noted in the flow rate control approach, local scheduling has an impact over the specific contents of local queues. While the quote above may be acceptable in the single server context, it is not adequate for multiple server systems, since some machines may be starved or blocked, and thus affected in their capacity, if local choices on other machines are not sensitive to their impact on the total system. This is of special importance in the context of re-entrant production systems, as Section 2.5 will show.

The model assumptions in [Dessouky and Leachman, 1994] are that products are processed in lots, each with a specific sequence of operations, the processing times are deterministic, set-up times are negligible or sequence independent, and that all operations requiring the same resource have the same processing time. They present a set of constraints that characterize their formulation of the release problem and argue that any objective function can be used depending on the particular problem to be solved. For the experimental results, the authors use the standard operational costs objective, where holding and backlog costs are incurred depending on the value of inventory.

2.4.2 Open Queuing Networks

Some authors approach the issue of controlling manufacturing systems by means of open queueing networks. By assuming they have no control over the release of new materials into the system they proceed to determine how to assign capacity to the diverse products requesting processing. That is, which scheduling rules to use. It is assumed that demand manifests itself through the arrival of customers to a system of queues modeling the production system.

In [Perkins and Kumar, 1989], the manufacturing systems process several types of products, each with a different routing through the available machines and subject to a constant demand rate. Demand is realized through the arrival of new parts into the system. The part types are characterized by their processing times' requirements and sequence of machines to visit. The load imposed by the different part types on a given machine, m is defined as

$$\sum_{(p,i)} \tau_{pm}^{(i)} d_p, \tag{2.6}$$

where d_p is the demand rate of type p and $\tau_{pm}^{(i)}$ is the processing time of the i-th visit of type p to machine m. Each part may incur variable transportation delays when moving from one machine to the other and set-up times are required when a machine changes from running a part type to another. Some of the types may need assembly and disassembly operations and there may be alternative routes for a part type upon exiting from a machine, induced by poor quality of the parts or other reasons. Given the rich model for the flow of parts along the system, there may be re-entrancy for some of the part types.

The paper presents scheduling policies which guarantee that, whenever the workload imposed by demand is below capacity, the manufacturing system is stable in the sense that the cumulative production for every part-type trails the cumulative demand by no more than a constant. Moreover, upper bounds on the amount of parts waiting for each machine are obtained.

The transportation times are assumed bounded by some constant for every pair of machines. Given that the set-up times are relevant in their model, one of the concerns is to make the production runs as long as possible for each part type, since there is an actual loss of capacity during set-ups. The policies proposed can be regarded as distributed feedback based policies to control the frequency of set-ups. Their policies are based solely on the values of the queues feeding each machine. Given that, the authors first discuss some classes of policies for single server systems. The first class of policies is the *Clear-a-Fraction* class — *CAF* —, where upon becoming idle the server sets-up for the part type which occupies at least some fraction of the buffer and then works on parts of that type until the buffer has no more of them. An instance of this class may be the *Clear-the-Largest-Buffer-Level* policy — *CLB* —, where the part type selected is the one with more parts in the queue. Another class of policies considered is the *Clear-the-Largest-Work* class — *CLW* —, where the server switches to the part type which has the highest amount of work in terms of the load it imposes on the server. Both these classes are shown to ensure bounded buffers and shown to track the cumulative demand by no more than a finite constant. Lower bounds on the performance of stable policies are derived also.

Next, the authors analyze systems with more than one server and no re-entrant structure, denominated *acyclic systems*. They show that the *distributed* CAF policies stabilize acyclic manu-

facturing systems when the load imposed by the demand processes is under the systems' capacity, i.e. when

$$\sum_{\text{p: part } p \text{ visits machine } m} \tau_{pm} d_p < 1 \quad \text{for every } m, \tag{2.7}$$

is satisfied. Since the system is acyclic, there is at most one visit per type to each machine. A distributed CAF policy is defined as a scheduling policy where each server uses a CAF policy for its own queue.

When it comes to nonacyclic systems, the authors are not able to prove stability for distributed CAF policies. A CAF policy, when there is re-entrancy, distinguishes between parts of the same type waiting for different visits at a given machine. Therefore, from the perspective of control, different visits are treated as different part types in order to establish which part type to select from the ones queued up for service. The necessary stability condition of becomes

$$\sum_{\{\text{(p,i): } i-\text{th visit of part } p \text{ to machine } m\}} \tau_{pm}^{(i)} d_p < 1 \quad \text{for every } m, \tag{2.8}$$

for nonacyclic systems. Although they cannot show that distributed CAF policies stabilize those systems when (2.8) is satisfied, they are able to show that there is a sequence of times $t_n \to \infty$ such that $x_{pm}^{(i)}(t_n) = 0$ for every (p, m, i), where $x_{pm}^{(i)}(t)$ is the queue length of part type p on its i-th visit to machine m at time t. Also, they show that every part entering the system eventually leaves it.

To ensure stability for nonacyclic systems they proceed by defining an alternative class of policies, which is an extension of the distributed CAF policies: *distributed CAF policies with Backoff*. The basic idea of this extended class is to allocate time slots to each part type, in order to ensure that the queues of part types not being processed will not grow excessively. Essentially, the policy produces a part type only when it would have produced the same part type if it had been in isolation, and its inputs had been strictly linear functions of time. By *strictly linear inputs* the authors mean situations where

$$td_p + \gamma \geq u_{pm}(t) \geq td_p - \gamma \quad \text{for all } p \text{ and } t \geq 0, \tag{2.9}$$

where $\gamma > 0$ is some constant and $u_{pm}(t)$ is the cumulative arrival of parts of type p to machine m. So, they allow for a machine to be allocated to a part type even if there are no products waiting,

thus departing from the *work conserving* policies earlier proposed. This has the effect of defining some bounds on the length of the production runs, preventing eventual starvation of other machines that may have empty queues simply because the types they process are held in a particular machine which is serving some other type. For this policy, the authors are able to prove stability and further establish that it stabilizes systems with assembly, disassembly, and proportional rerouting, as long as (2.8) holds.

It is interesting to observe that for nonacyclic systems, wasting capacity may be one way to ensure stability. This is counterintuitive, since keeping the servers idle is usually perceived as an unrecoverable loss of capacity and, therefore, a liability in terms of tracking an external and uncontrollable demand process.

Following this paper is [Kumar and Seidman, 1990], which uses the same modeling paradigm to provide an analysis of the dynamics of nonacyclic manufacturing systems. Some previously unresolved stability issues are reviewed and examples of systems and scheduling rules that lead to unstability are discussed. They further establish a sufficient condition under which distributed CAF policies are able to stabilize any nonacyclic production system. This stability condition has the disadvantage of being more stringent than condition (2.8). Because of that, they move along the direction of proposing classes of policies that are ensured to stabilize nonacyclic systems when condition (2.8) alone holds. They propose a *supervisory mechanism* that controls the length of each production run and keeps a close track of part types whose queue size grows beyond some threshold, defining a special hot list of the larger queues. These larger queues are given priority over the others, but each run on parts in the special queue is also bounded in time. If the special queue is empty, then the policy used is a distributed CAF. They show this generic mechanism to ensure stabilization for any nonacyclic system.

In [Lu and Kumar, 1991], an extensive study of local scheduling policies for wafer fabs is provided. The system is modeled as a pure re-entrant network of queues with deterministic routing, no control over release dates, no control over the due dates assigned to the parts, and there is a single part type. Set-ups are not considered in the model so that there is no need to make production runs as long as possible to decrease the set-up frequency. The objective is to reduce mean cycle time and its variance. The source of uncertainty is only due to the arrival process randomness.

To control the scheduling of parts along the system, they propose a series of local scheduling rules that choose parts based solely on the local queues' contents. The rules are *First Buffer First*

Serve — FBFS —, *Last Buffer First Serve — LBFS —*, *Earliest Due Date — EDD —*, and *Least Slack — LS*. It is assumed that parts waiting for a given machine are placed in different buffers depending on their stage of processing. So, each part has a buffer per visit. The FBFS rule assigns decreasing priority by increasing number of visits already made; LBFS assigns decreasing priority by decreasing number of visits already made; EDD assigns decreasing priority by increasing order of the due dates; and LS assigns decreasing priority by increasing order of the time remaining to the due date, discounted by the total amount of processing time required for the parts to be completed and exit the system.

All of the above four priority rules are shown to be stable as long as the necessary stability condition holds. An example of a buffer priority rule which is unstable is discussed. This particular example is the same as the one discussed in [Kumar and Seidman, 1990]. Curiously, given that [Kumar and Seidman, 1990] showed that non work conserving policies may well be the key to ensure stability, they ([Lu and Kumar, 1991]) do not consider non work conserving policies for the production scheduling of the particular example discussed.

After establishing the theoretical stability results, the authors provide and extensive simulation study to evaluate the relative performance of these rules. The study shows that giving priority to parts closer to completion (LBFS) attains the best score in terms of mean cycle time. However, the best performances in terms of variance for the cycle time are accomplished when priority is given to products with a smaller slack (LS). They conclude by proposing a convex combination of the two best rules, if the concern is to control both performance measures: mean cycle time and variance of cycle time.

An excellent and comprehensive survey of the line of research just described can be found in [Kumar, 1993], where some other related work omited here is described and some open problems listed. The overall approach is placed into perspective with other research in the area.

In a more recent work, [Kumar and Kumar, 1994], the authors are able to compute lower and upper bounds for the performance of the optimal scheduling rule, assumed to be *non-idling* (work conserving) and stationary. For each system they formulate a pair of linear programming problems that constitute an upper and a lower bound on the mean number of parts in the system. They use the stationary balance equations to specify equality constraints and their inequality constraints are a consequence of using non-idling control policies. They observe that, if the objective is to

minimize a weighted sum of the parts in the system with different weights for each buffer[2], or at least one of the weights differs from the others, every optimal policy may require some idling.

After establishing the generic linear programming problems, they apply the method to compute the bounds for a series of systems. The first example is an open re-entrant line with two machines for which they compute the bounds explicitly. Then they show that the bounds obtained are tighter than those of [Ou and Wein, 1992][3]. Next, they use their methodology to compute bounds for the first example assuming the control policy to be the LBFS and FBFS. The conclusions are that the LBFS policy is nearly optimal in light traffic and it is always better than the FBFS policy for the system considered.

They depart from the open queueing framework to the closed queueing framework in order to address the performance of release policies in networks of queues. For this, they formulate another set of linear programming problems to establish upper and lower bounds on the throughput of a closed queueing network. They fix the population size to some values and compute bounds for the optimal throughput of the optimal policy (assumed to be a nonidling policy), the LBFS, the FBFS, the *balanced policy* of [Harrison and Wein, 1990], and an unbalanced policy. The bounds for the balanced policy exclude the unbalanced policy since the lower bound of the former is higher than the upper bound of the latter. Excluding the unbalanced policy, all the bounds for the other three policies are non conclusive about their relative behavior.

Also in the context of local scheduling rules for queueing networks, an outstanding statistical study was recently published in [Lu et al., 1994]. The authors propose a new class of scheduling rules, termed generically as *Fluctuation Smoothing* policies — *(FS)*. These policies are a subclass of the Least Slack policies mentioned above. One of the first surprises of the study is the fact that their results shed new light on the relative importance of local scheduling rules versus release rules. Prior authors claimed that release is more important than local scheduling, [Glassey and Resende, 1988, Wein, 1988], in the sense that the proper choice of the release policy has more impact on the performance than the choice of local scheduling policy. The findings of [Lu et al., 1994] are that the proper choice of local scheduling rules may have similar impact as that of the release policy. As an example, using the workload regulating policy described in Section 2.4.1, their local scheduling policies achieve a reduction of 22.4% in the mean queueing time and a reduction of 52.0% in the

[2] As with non decreasing holding costs, for instance.
[3] Applying their method to the systems discussed in [Ou and Wein, 1992].

standard deviation of cycle time, when compared with the baseline FIFO policy.

The class of policies proposed was designed with the objective of reducing mean and variance of cycle time. Their three fluctuation smoothing policies are:

- *Fluctuation Smoothing for Variance of Lateness — FSVL*, where each lot arriving to the plant carries a due date and there is, for each buffer, an estimate of the remaining cycle time for the lot to be completed. Higher priority is awarded to the job in the queue that has the smallest difference between the due date and the estimated cycle time to completion.

- *Fluctuation Smoothing for Variance of Cycle Time — FSVCT*, which is exactly the same as the previous with the assigned due date made equal to the arrival time.

- *Fluctuation Smoothing for Mean Cycle Time — FSMCT*, where there is a fictitious due date assigned to each lot that incorporates the notion that there should be some sort of periodicity in the output process of each server. That is, since delays experienced by lots at a server are caused by the burstiness of their arrivals, the scheduling rule attempts to uniformize the interarrival times to each buffer. The due date assigned to the n-th lot entering the system is given by n/λ, where λ is the arrival rate. Once the due date is defined the policy is like the previous two.

The experimental study was conducted in the context of re-entrant production systems operating a single product. The methodology to determine the estimates of the cycle time to completion are done in a similar way to that of [Vepsalainen and Morton, 1988]. They simulate the system with an initial set of estimates, determine the statistics of the actual cycle times, and repeat the procedure until the estimates used closely match the actual cycle times. They test several alternative release policies and several alternative scheduling rules. The release policies are random Poisson arrivals, constant interarrival times, fixed number of customers in the system, and the three variants of the workload regulating policy as presented in [Wein, 1988]. The scheduling rules considered were all of the ones used in [Wein, 1988] plus EDD, SRPT++[4], and the last two FS policies.

In all the tests for the alternative release policies considered, the FS policies ranked first in terms of mean queue time and standard deviation of cycle time. Among these two policies, the

[4]This is a variant of Shortest Remaining Processing Time that gives priority according to the size of the immediate queue. To all the lots in the queue that are headed to a queue with small size, choose according to SRPT. This is an attempt to avoid starvation of downstream machines.

FSMCT achieves the lowest mean queue time almost all of the times, and the FSVCT achieves the lowest standard deviation almost all of the times. The final recommendation for the best pair release/scheduling rule is to use workload regulating releases together with FSMCT for scheduling.

In their closing remarks, [Lu et al., 1994] mention that it would be desirable to have cycle time estimates dependent of the system state, since they use fixed estimates in their study. Also, they only consider a single process flow. Things get a little more complex for other settings.

Many of the scheduling rules discussed so far had their origin in the context of deterministic scheduling for job-shops. There is an enormous body of literature on job-shop scheduling, the specifics of which are not of central importance for this thesis, and consequently not relevant for this review. However, the reader interested in the specifics of job-shop scheduling and in the derivation and motivation of some of the release rules discussed above is referred to [Conway et al., 1967, Baker, 1974, French, 1982] for classical introductory level text books on scheduling theory. A more updated text book in scheduling theory and heuristics is [Morton and Pentico, 1993]. There are also some fundamental surveys on production scheduling like [Panwalker and Iskander, 1977, Graves, 1981, Lawler et al., 1982].

2.5 Stability

Often in the background of some of the above mentioned studies the issue of stability plays an important role. Although the issue of stability is relatively trivial for production systems which are not re-entrant or do not have flows in opposite directions[5], such is not the case when it comes to re-entrant systems or non re-entrant systems where some part types may flow in opposite directions. Although the load of the system may be below capacity, a poor choice of release and scheduling policies for non acyclic systems may lead to instability, even when such choice is made from a subset of work conserving policies.

When approaching the problem of production control by means of formulating an optimal control problem, usually stability questions are trivially answered in the sense that if there is one policy that makes the system stable, the optimal policy will also be stable and it will be found through the optimization procedure. However, when the approach is to choose a particular policy that is not derived in such a way, then stability has to be addressed explicitly in order to determine

[5] As long as demand is below capacity, systems are stable.

if the given policy ensures stability or not.

Therefore, in the context of flow rate control, there is no explicit consideration of stability. Even though the approach is applied to re-entrant systems or to generic job shops, the necessary stability condition is also sufficient. In fact, stability was never addressed explicitly by the authors who have done work on flow rate control, since the necessary stability condition was always taken as sufficient. The exception to this is [Caramanis and Liberopoulos, 1992]. There, the authors explicitly stated and proved that for failure prone systems, with deterministic demand rates, and deterministic processing times, as long as the demand vector is an interior point of the expected capacity set, then there exists a flow control policy that results in a stable system. Stability is taken in the sense that the expected end product inventory is finite for all products. Their proof does not rely on any specific assumption on the flow patterns inside the production system.

However, when modeling production systems by means of open queueing networks and proposing specific scheduling rules, it often has been the case that stability becomes a hard question to answer. Also, to establish the heavy traffic limit theorems of [Harrison, 1988, Harrison and Wein, 1990], it is necessary to establish the stability of the queueing networks considered. Examples of networks for which the Brownian approximation does not hold have been presented, as [Dai and Wang, 1993] is one example.

Usually, the issue of stability in networks of queues is established by explicitly determining an invariant distribution. The classes of queueing networks for which such invariant distribution is known are very limited. Typically, networks of queues operated under local scheduling policies are among those for which little is known about their invariant distribution or even if one exists. They fall outside the classes for which there are product form solutions. Product form solutions exist for the *generalized Jackson networks*: single class networks with exponential interarrival and service times, where queues are served in a first come first serve order, [Jackson, 1975]. For some scheduling disciplines in multiclass networks, with special distributional assumptions on interarrival and service times, the stationary distributions were explicitly determined in [Baskett et al., 1975, Kelly, 1979].

One example of open queueing networks where addressing the stability problem is highly relevant is the work of [Perkins and Kumar, 1989], already discussed in Section 2.4.2. Their distributed CAF policies for local scheduling were ensured to be stable for acyclic systems as long as the necessary stability condition holds. However, the authors were unable to show similar properties for nonacyclic systems and had to propose a modification of the original policies to stabilize any such system. One

of the central statements of the present thesis is that the structure of the modification proposed holds the key to the problem of stability.

Their original distributed CAF policies are non-idling, or work conserving as some authors prefer to call them. The central feature of a non-idling policy is that no server should be kept idle as long as there is at least a customer in its queue. There is an intuitively clear reason for preferring non-idling policies: keeping a server idle, while there is work in front of it, is a clear waste of capacity and should be avoided when the objective is to maximize throughput, or tracking a demand process which imposes a load very close to the system's maximum capacity. Recall that the problem addressed in [Perkins and Kumar, 1989] was one of reducing the number of set-ups as much as possible, since a set-up time is a waste of capacity.

Nevertheless, when the authors found themselves unable to prove the stability of their policies for non-acyclic systems, they proposed a modification that basically increased the number of set-ups, thus incurring more waste of capacity, which is the exact opposite of what should be expected in intuitive terms. Besides that, they also allowed each server to remain idle when not in a set-up (distributed CAF policies with backoff), even if there would be other jobs in the queue.

Note that, from the perspective of queueing networks, it is simpler to consider non-idling policies. An idling policy is very complex to model since it carries the inclusion of an additional option to consider whenever a server becomes idle upon completion of a service. Besides that, this additional option also entails the decision of how much time should the server remain idle. The answer to this question, being so difficult to determine, has kept idling policies outside the classes of policies usually considered for queueing networks. In the context of flow rate control, it is easier to include idling policies. Recall that the two-boundary policy of [van Ryzin et al., 1993] is an idling policy. For a system with two machines in tandem, when the surplus is negative with high values, the first machine does not produce at its maximum possible rate because the policy imposes a bound on the amount of inventory between machine 1 and machine 2 (see Fig. 2.3). Some generalizations of the two-boundary policy to more than two machines in tandem, with or without re-entrancy, include this feature by bounding the maximum amount of inventory in each of the buffers, even when the overall surplus is lagging demand by a big amount. Therefore, in some situations a particular machine remains idle while there is inventory in its feeding buffer and the cumulative production lags the cumulative demand. A lag in cumulative demand for flow rate control models is in some sense equivalent to non empty queues for queueing network models. It could be argued that, in the

context of queueing networks, similar ideas could be used by including a blocking feature in the model. However valid and feasible, this approach has not been as visible in the literature as one would expect. The work of [Ou and Wein, 1995] can be seen as a contribution along the lines of considering idling policies, since they allow for idleness of the servers.

Note also that the choice of policies is dependent of the production objectives. If the objective is to minimize cycle time, or maximize machine utilization, it seems natural to consider non-idling policies. However, if the objective is to track demand, it makes some sense to add a little slack to the cycle time or the machine utilization. After all, the machine utilization cannot be above the load imposed by demand. Therefore, allowing for some idleness should not necessarily mean that there is a waste of capacity, but rather that there is an effort to distribute the load more evenly. Another point in favor of idling policies is the fact that they enforce a reduction of variance. Having a server that reacts automatically to its feeding queue implies that the server utilization is almost[6] as stochastic as the arrival process. This randomness carries through to the following servers in terms of their queues and consequently of their utilizations. Allowing for some idleness of the servers has the effect of filtering out some of that randomness.

In [Perkins and Kumar, 1989], the inability to prove stability for the original policies could still be thought of as a problem that could be solved in due time, since the authors did not show that in fact instability could occur. This question ended up being answered through an example not much later.

In [Kumar and Seidman, 1990], also in the context of open queueing systems, the authors discuss several issues relative to the stability and stabilization of production systems, and introduce an example of a re-entrant system for which there exists a non-idling control policy that yields unbounded trajectories for the buffer sizes, although the workload imposed by demand is below the available capacity. That same example is revisited in [Lu and Kumar, 1991] as a special case of poor priority assignment for the buffers feeding the two-machine re-entrant system. However, more recently, some authors managed to generate examples of queueing networks that can become unstable even for *reasonable* scheduling policies. One striking example has been the case of the First Come First Serve rule, which was shown to lead to instability, even when the necessary stability condition is satisfied: [Bramson, 1994] on a two machine re-entrant system, with exponential interarrival and service times; and [Seidman, 1994] on a more general network (not re-entrant) with

[6] The fact that service times are non zero introduces some filtering on the machine utilization process.

multiple customer arrivals.

The control policy that yields the unbounded trajectories for the example of [Lu and Kumar, 1991] is a non-idling policy, as is the case for FCFS. Rather than considering idling policies for the local scheduling decisions, much of the effort has been put either on determining non-idling policies that are stable for the necessary stability conditions, or on determining additional conditions that have to be satisfied by the production systems so that all or a given non-idling policy ensure stability.

In [Kumar and Meyn, 1995], a methodology is proposed to evaluate networks of queues and scheduling policies in terms of their stability properties. They use linear and nonlinear programming to determine an appropriate quadratic functional to be used as a Lyapunov function. Provided the underlying system is Markovian, the method establishes the existence of a steady-state probability distribution by showing that there is a negative drift in the sizes of the queues. The authors concentrate on non-idling scheduling policies and are able to deal with a single product following a fixed path or multiple products following random paths and with multiple entry points. For any given system, the answers obtained are either that all non-idling policies stabilize the system or they determine regions on the parameter space for which all non-idling policies induce stability. The parameter space used is that of the load imposed by the demand processes on each of the servers. So, when they apply their methodology to a given system it may be the case that the answer is: as long as the load imposed by demand is below capacity for all servers, every non-idling scheduling policy is stable. When such a result is not possible the methodology is only conclusive for the range of parameters for which it establishes stability. That is, the test determines regions for which all non-idling policies ensure stability. Outside those regions we do not know if non-idling policies lead to instability and we do not know if there is at least one non-idling policy that yields the system unstable.

The stability of re-entrant production systems is discussed by means of analyzing the stability of equivalent fluid models in [Dai and Weiss, 1996]. They establish generic properties on the fluid model in order to ensure that systems are stable when demand is below capacity, and show specific systems and buffer priorities that satisfy those properties. They also make use of Lyapunov functions, but theirs are piecewise linear in the queues' contents. The overall strategy is to show that these possess negative drift as well. They discuss the "Lu-Kumar" example and present conditions, additional to the trivial load below capacity conditions, under which the system is stable for that

choice of priorities and for all non-idling policies. Also, they introduce some new examples of *Kelly-type networks*, [Kelly, 1979], which are unstable for a given choice of buffer priorities.

Kelly networks are characterized by the fact that different visits to any given server obey the same service distribution. If the distributions are exponential and the scheduling rule is FCFS, [Kelly, 1979] proved that such networks are stable as long as load is under capacity. A Kelly-type network follows the same assumption that the mean service times are the same in different visits to any given server, but the local scheduling decisions are not made with the FCFS policy, nor are the service distributions exponential. The authors manage to prove stability for all non-idling policies on any Kelly-type network with a unidirectional ring structure processing multiple customers and with possibly many entry points.

Along similar lines of [Dai and Weiss, 1996] is the more recent work reported in [Bertsimas et al., 1996]. The authors also make use of fluid queueing networks to discuss stability conditions. They present a stability test based on linear programming for non-idling policies. The linear program is shown to be a necessary and sufficient condition for the stability of all non-idling policies for multiclass fluid queueing networks with two stations. Also, they present new sufficient conditions for the stability of multiclass queueing networks involving any number of stations.

The validation of the Infinitesimal Perturbation Analysis approach in [Glasserman and Tayur, 1995] for infinite horizon costs relies on the proof of stability for the single-product, multiple-machine, and non re-entrant system presented in [Glasserman and Tayur, 1994]. The authors show that it suffices to have the expected demand below the capacity of the machine with the lowest output in order to ensure their control policies to be stable.

Although they are dealing with a non re-entrant system, for which the stability issue is trivial, the discussion on stability is useful and necessary to identify renewal points of the stochastic processes considered, which has implications on the validation of the approach to estimate values and gradients of infinite horizon performance measures.

From the perspective of the present thesis, the concern regarding stability is twofold. On the one hand, the thesis addresses the problem of production control of re-entrant systems for which, as discussed above, answers are not necessarily trivial, and there is the need to ensure that the control policies proposed ensure stability. On the other hand, there is a need similar to that of [Glasserman and Tayur, 1994] regarding the validation of the Infinitesimal Perturbation Analysis

approach.

In Chapter 6, it will be shown that it is necessary to depart from non-idling policies in order to make the trivial necessary stability conditions sufficient for all re-entrant systems considered. Given the modeling paradigm used in this thesis, it will be easy to define a class of policies that include simultaneously idling and non-idling policies as connected subclasses. That is, it will be possible for the optimization procedure to move from a non-idling policy to an idling policy and back, if such is the need of the particular production system under study. Therefore, the methodology proposed can be used as a tool to identify which production systems can be stabilized by means of a non-idling policy and which are the production systems that need an idling policy to be stabilized.

The emphasis will be put on determining a class of policies that can stabilize all re-entrant systems, rather than on determining what are the systems that can be stabilized by a particular class of policies. This emphasis seems to be more in line with the traditional approach of Control Theory.

Moreover, it will be shown in Chapter 5 that idling policies may bring the benefit of incurring lower operational costs, even when their use is not required by stability considerations. That is, it may be the case that, although any non-idling policy stabilizes a particular system, the optimal policy is an idling policy. Naturally, for such cases it is of particular importance to be able to define a compact class of policies where both subclasses are connected.

2.6 Simulation

Simulation is a very powerful tool for the analysis and evaluation of complex systems. In many circumstances it is used when analytical models are not available nor easy to obtain. Traditionally, simulation is used to compare configurations, different policies, validate models, and many other qualitative features with the purpose of answering *what if* questions. In almost all the sections above there was at one time or another a reference to this type of methodology, typically when some sort of performance evaluation was one of the issues to determine.

When intending to study the relative performance of a particular scheduling policy over another, usually simulation is the tool to use, providing relevant statistical information. Some examples of previously mentioned such cases are [Wein, 1988, Glassey and Resende, 1988, Lu and Kumar, 1991], just to name a few.

Another previously mentioned simulation study, which is particularly representative of the use simulation can have, is that of [Chen et al., 1988]. In their work, the authors used simulation as a validation tool for queueing networks to serve as an accurate model for a particular manufacturing system.

Others have used simulation just to identify macroscopic behavior of manufacturing systems, so that some sort of structural characteristics can be extracted. Such is the case of the *signature analysis* concept discussed in [Atherton and Dayhoff, 1986]. Their simulation based method characterized a particular wafer fab dynamics in terms of curves for inventory, cycle time, and throughput as functions of wafer start rates. Their claim was that any given manufacturing system possesses some sort of *finger print — signature —* which is unique. The simulations carried out help to aggregate diverse information into a small set of plots characterizing such a signature. The knowledge of a signature may help in determining better control policies for the system under study.

Along the lines of identifying a particular system's behavior when some parameters change is the work of [Ehteshami et al., 1992]. They used simulation to understand the impact of *hot lots* on the cycle time of other lots in a semiconductor fab. Their simulation runs provide approximations for the statistical distributions of cycle time for various work-in-process loads and different percentages of hot lots in the system.

The issue of simulating manufacturing systems naturally raises the questions of the simulation capabilities in terms of being able to represent accurately the important features of a given system. Some effort has been put in the area of designing simulation tools that emulate as closely as possible the true dynamics of manufacturing systems in general and of semiconductor manufacturing systems in particular. See [Prasad, 1991] for a report on such a development.

There are many other situations where simulation is the only available tool to answer the types of questions generically identified as *what if* questions. In the review of [Uzsoy et al., 1992] there are many other references to the type of questions answered by simulation and to particular instances of research contributions aimed at answering those questions through simulation.

There are other questions, that can generically be defined as the *how to* questions, that have been successfully answered via simulation in recent years. For instance, these questions refer to the comparison between two different realizations of a given policy, differing only through a small set of parameters characterizing it. The issue is how to use simulation as an optimization tool.

Significative advances in this area were made in the last twenty years.

A good introductory level tutorial to the area of *simulation optimization* can be found in [Azadivar, 1992]. Using the author's own words about simulation optimization:

> "A simulation optimization problem is an optimization problem where the objective function, constraints, or both are responses that can only be evaluated by computer simulation. As such, these functions are only implicit functions of decision parameters of the system. In addition, these functions are often stochastic in nature as well."

The author also presents a cluster classification in terms of approaches to solve the above described type of problems. Out of these, the ones of interest for this thesis are the *gradient based search methods*. Generically, these methods use simulation to generate estimates of the objective function derivatives with respect to some parameters that characterize diverse features or control policies of the systems. The gradient based methods can further be classified into: finite difference estimation, frequency domain analysis, likelihood ratio estimators, and infinitesimal perturbation analysis.

Excluding the finite difference estimation as being a crude technique and the frequency domain analysis for its complexity, one is left with two main choices when both techniques are applicable[7].

In [Strikland, 1993], both techniques (likelihood ratio and infinitesimal perturbation analysis) are compared in terms of the quality of results they produce. The main conclusion is that, in general, the likelihood ratio technique produces higher variance estimates than infinitesimal perturbation analysis. Moreover, the likelihood ratio estimates tend to have a variance that grows with the simulation run size, whereas such is not the case with infinitesimal perturbation analysis. Depending on the values to estimate by simulation there are techniques that can be used with the purpose of reducing variance. Some of those techniques are discussed thoroughly in [L'Ecuyer, 1994]. Other relevant issue in simulation based optimization has to do with convergence. [Gürkan et al., 1994] discusses some of the state of the art regarding this problem.

The great advantage of infinitesimal perturbation analysis over other techniques is the fact that it generates gradient information from a single simulation run. That is, variables are set at some

[7]Neither likelihood ratio nor infinitesimal perturbation analysis is a panacea. Also, the set of problems for which each one is valid does not contain nor is contained in the set of the other. Comparison between these two can only be done for those cases where both can be used.

nominal value and simulation is performed; once the simulation run terminates, both the objective function and its gradient, relative to some set of parameters, will be available without disturbing the nominal path.

In the next section, some specifics of the infinitesimal perturbation analysis will be reviewed and placed into context for this thesis.

2.6.1 Infinitesimal Perturbation Analysis

The perturbation analysis approach had its origins in [Ho et al., 1979], dealing with buffer storage optimization in a production line. Since those early days IPA has undergone an astonishing growth both in terms of its applications and its theoretical maturity.

The essential feature of perturbation analysis is the realization that a single simulation run contains more information than just first and second order statistics. The technical issue is the fact that expected value and differentiation may be permutable operators, so that the derivative of an expected value can be computed as the average of individual derivatives along the simulation run. In the beginning, many contributions were along the lines of establishing the validity of such permutation on a case by case basis.

There are two main categories of perturbation analysis: *finite perturbation analysis* and *infinitesimal perturbation analysis*. Up until the late 80's the finite perturbation analysis had some experimental results but there was still a lack of theoretical understanding of the algorithms. On the contrary, the theory of infinitesimal perturbation analysis had a faster development and by the mid 80's there was already a considerable body of theory supporting it.

For good summaries about the developments on perturbation analysis see [Suri and Zazanis, 1988], which provides not only a substantial review of significant literature up to that point, but also contains an introductory level summary of the essential aspects of perturbation analysis. Another good tutorial on perturbation analysis can be found in [Ho, 1992]. Actually, there are many outstanding papers by Yu-Chi Ho on the basis of the approach, its developments, and with substantial literature reviews.

In the early 90's some essential theoretic foundations started to fit into place so that rather than establishing the validity for individual problems it was possible to establish general properties on the systems and problems to which the approach is valid. See [Glasserman, 1991, Glasserman,

1992, Chong and Ramadge, 1994] for such contributions. The two relevant books that establish the body of theory for perturbation analysis are [Glasserman, 1990, Ho and Cao, 1991].

Although there are still open questions regarding the use of the approach in many settings and further research on ways to extend the basic principles to those cases goes on (see [Shi, 1996] dealing with discontinuities of the sample performance function), there is a significant body of literature on applications of perturbation analysis to particular optimization problems. In many cases, one first issue is to establish that the problem at hand does fit the general properties that allow the use of PA. Then it is possible to concentrate on the specific issues of the optimization itself and on the results achieved.

In the specific context of interest for this thesis, some applications of PA were already mentioned in previous sections. Such was the case of [Song et al., 1992, Caramanis and Liberopoulos, 1992] in the context of flow rate control approaches, although their validation results were limited to small systems. [Song et al., 1992] address a single product and two machines system for which they validate the IPA and present some experimental results. [Caramanis and Liberopoulos, 1992] validate the IPA to general size systems but their derivative calculation has to be established for each particular system due to their parameterization of the policies. They present results for two machines, either with two part types and three part types. They propose to decompose a P part type problem into P two part type problems where each two part type problem approximates a part type's interaction with the remaining ones by aggregating them into a single proxy part type.

More recently, and also in the area of flow rate control, an IPA based approach was validated in [Brémaud and Malhamé, 1997] for a single machine with multiple machine states and a single part type. Their approach relies on the existence of a regenerative structure on the process. To establish this regenerative structure the authors present a stability analysis, where they identify the necessary and sufficient condition for stability. Their model and control policy are inspired by the work of [Sharifnia, 1988]. The objective of the authors is to compute the several hedging points that characterize the optimal control policy for this problem.

In the context of inventory control, [Glasserman and Tayur, 1995] verify the validity of IPA and use it to compute the optimal parameters of a multi-echelon base stock policy for a single product flow line composed of a number of capacitated machines. Their modeling framework is sufficiently general so that any dimension can be tackled straightforwardly. This thesis follows their methodology in terms of the modeling framework. The control policies proposed here are of

the same type and the objective is to deal with systems that produce more than a part type and have a re-entrant structure. Later in the thesis, based on the experimental results of Chapter 5 and motivated by the stability discussion of Chapter 6, an extension of the control policies will be proposed.

Regarding previous work proposing IPA based methodologies to compute parameters of control policies for re-entrant systems, this thesis stands out for not restricting the formal scope to small dimension problems, for explicitly dealing with capacity sharing mechanisms, and for providing a syntactically rich framework.

Part II

Uniform Loads and Perfect Yield

Chapter 3

Theoretical foundation

The model of re-entrant production systems considered in this thesis has M machines in series (*stages*). Each one of the P products processed by the system has to cycle K times (*levels*) through those M machines before being completed. The framework used is a discrete time (or periodic review) capacitated multiple-product production-inventory system operating under an *echelon base stock policy*: every level and stage operates on a base stock policy for echelon inventory. That is, given a particular product, the decision maker adds all inventory downstream from that level and stage to determine the echelon inventory. If the echelon inventory falls below the corresponding base stock value the decision will be to produce the difference, provided there is enough capacity and (relevant) upstream inventory.

The present analysis concentrates on a simple class of systems as a first step towards analyzing a broader family of production systems under the framework of discrete time inventory control. In this chapter the scope is limited to studying what can be seen as the inventory model counterpart to the *Kelly type networks* mentioned in Chapter 2. That is, each product requires the same amount of capacity per unit (or processing time per unit) processed no matter what is its processing stage and level. This will be referred ahead as the *uniform load assumption*. The class of systems now addressed is also characterized as having perfect yield and having reliable machines with deterministic processing times.

In any period, demand for the products occurs. Production decisions that are made are constrained by available inventory and capacity. Several different production and capacity allocation rules will be analyzed and a procedure to *jointly optimize* these decisions along with inventory levels to minimize operating costs will be developed. Both the finite horizon and the stationary infinite

horizon versions of this model will be considered.

Discrete time is a valid assumption for this problem, because decisions are made every shift or every half-shift, inducing a periodic review framework on the production decisions. A base stock class of policies is assumed because it is a large class of easily implementable policies, and because for certain simple models it is optimal among all possibilities.

The rest of the present chapter is organized as follows. The basic recursion equations for the state variables followed by the recursion equations of the derivatives will be introduced first in Sections 3.1 and 3.2. Section 3.3 will then address the production decisions and their derivatives. The validation of the IPA in a finite horizon setting will be provided in Section 3.4. The infinite horizon estimates are validated in Section 3.5. The chapter concludes in Section 3.6 with a brief summary and with the presentation of derivatives with respect to the capacity slots.

3.1 Basic Model

Consider controlling a re-entrant, multi-stage, multi-product, capacitated production system facing random demand. The simplest model of a re-entrant production system is one where there are M machines in series (*stages*). Each one of the P different products processed by the system has to cycle K times (*levels*) through those M machines before being completed. The discrete time model considered is a generalization of [Glasserman and Tayur, 1994, Glasserman and Tayur, 1995] in two ways: multiple products and re-entrancy.

In any period, each machine can process different parts belonging to different levels; the total production is only limited by its capacity. After being processed by a machine, parts are placed in intermediate buffers where they wait their turn to be processed by the next machine or until they are depleted by external demand if the previous operation was the last of the parts' requirements. The capacity of the buffers is assumed to be infinite.

The following is the list of notation for this chapter:

- P products (indexed by p);
- M stages (indexed by m) in each level;
- K levels (indexed by k);

- $(km)^+$: denotes the level and stage immediately before level k and stage m;
- $(km)^-$: denotes the level and stage immediately after level k and stage m.
- d_n^p: demand for product p in period n (at the last stage and last level only).
- z^{kmp}: echelon base stock level (in echelon terms, i.e., $z^{kmp} \geq z^{(km)^- p}$);
- Δ^{kmp}: alternative set of variables accounting for the inventory between stages;
- I_n^{kmp}: inventory in time period n for product p at stage (machine) m in level k;
- E_n^{kmp}: echelon inventory in time period n (it is the sum downstream of all relevant I_n^{**p});
- Y_n^{kmp}: shortfall in time period n for product p at stage m in level k, $Y_n^{kmp} = z^{kmp} - E_n^{kmp} + d_n$;
- P_n^{kmp}: production amount in period n for product p at stage m in level k;
- C^m: capacity of stage (machine) m;
- C^{km}: capacity of stage (machine) m allocated to level k (PS), $C^m = \sum_{k=1}^{K} C^{km}$;
- C^{kmp}: capacity of stage (machine) m allocated to product p at level k (NS), $C^m = \sum_{k,p} C^{kmp}$;

The first machine to be encountered in the series is machine M and the last is machine 1. Also, the first cycle that the parts must undergo is level K and the last cycle is level 1. Therefore, I_n^{11p} denotes the inventory of product p at the last buffer at the beginning of period n, from which demand is satisfied directly or backlogged (most upstream closest to raw material).

There are, at least, two possible ways the cost can be computed as sugested by [Tayur, 1992]:

- Costs incurred after demand

  At the beginning of each period the production quantities are set for each level, stage and product. At the end of the period the inventory quantities are updated according to the achieved production and the demand. Demand is assumed to occur after production has been decided. If there is enough inventory demand is satisfied immediately; otherwise it is backlogged and satisfied with production from future periods.

- Costs incurred after production

  At the beginning of each period demand for each product occurs at the final stage and level of the system, that is for $m = 1$ and $k = 1$. If there is inventory enough for a given product, its corresponding demand is immediately satisfied, otherwise demand is backlogged. Next the production quantities for each level, stage and product are determined. At the end of the period, the inventory levels are updated according to the achieved production.

The second setting will be used for simplicity of exposition. However, everything follows in the first setting in a straightforward manner and the experimental data of Chapter 5 are obtained for this setting, following the traditional costs accounting procedure of inventory control theory. The study is restricted to the following class of inventory control, capacity allocation and production rules. Demands are assumed to be continuous, independent across products, and i.i.d. for each product in time (stationary).

1. Inventory Control. Every level and stage operates on a base stock policy for echelon inventory. That is, given a particular product, the decision maker adds all inventory downstream from that level and stage to determine the echelon inventory. If the echelon inventory falls below the corresponding base stock value the decision will be to produce the difference, provided there is enough capacity and (relevant) upstream inventory.

2. Capacity Allocation. Each machine m, with $m = 1, \ldots, M$, has a fixed total capacity C^m. This total capacity can be divided into $K \times P$ single slots and a slot can be assigned to each product and level (C^{kmp}). Alternatively, the total capacity can be divided into only K slots, each assigned to each level (C^{km}), and so shared by P products within each level k. Finally, the total capacity of a single machine can be shared by all products and levels. Thus, this capacity allocation is *static*. Let the first choice be called as the *No Sharing* mode (NS), the second as the *Partial Sharing* mode (PS), and the last as the *Total Sharing* mode (TS).

3. Production Rules. Whenever there is some degree of capacity sharing it is necessary to establish a capacity management scheme. That is, it is necessary to know how the available capacity is to be distributed among the several products whenever production requirements are bound by capacity. To take care of this dynamic decision making, three *production rules* are proposed: Linear Scaling, Priority, and Equalize Shortfall. The Linear Scaling Rule

(LSR), scales down all production needs to fit capacity. The Priority Rule (PR), assigns capacity according to a fixed priority list. (Within PR, note that there are several choices — Section 5.4.1.) The Equalize Shortfall Rule (ESR), assigns capacity to the products whose echelon inventories are more distant from their target base stock levels.

Note that the NS mode corresponds to a situation where the original system would be converted into P different production systems each with a single product and non re-entrant flow lines.

The model detailed below is still relatively simplistic but it possesses the potential to be extended quite easily. The only source of uncertainty presently considered lies in the demand process. Semiconductor manufacturing in general, and wafer fabrication in particular, are known for possessing many other sources of uncertainty. The main sources of uncertainty lie in the production process itself: machines fail, processing times are not deterministic, and yields are random. In Part III, some of these additional sources of uncertainty will be explicitly included and the inclusion of some others will be discussed in some detail in Chapter 7.

It is difficult to manage complex systems without understanding simplified versions of them. Managing re-entrant systems with deterministic capacity, deterministic processing times, and perfect yield in a multi-product setting with random demands is not well understood yet. For this reason, I have chosen to present a model stripped of the production randomness. Furthermore, almost all the results and the essential details are similar in a more complex model with production uncertainties (handled via random capacity and random yield), as will be discussed in Chapters 6 and 7. What may change for these more complex systems is the proper class of control policies to consider in order to attain the best performance.

3.1.1 Basic Recursions

The basic recursion equations governing a re-entrant flow shop will now be presented. Before that, a small notational detail needs to be clarified. For the re-entrant system previously defined, let

$$(km)^{-} = \begin{cases} (k, m-1) & m \neq 1 \\ (k-1, M) & k \neq 1 \text{ and } m = 1 \\ \text{undefined} & k = m = 1 \end{cases}$$

$$(km)^{+} = \begin{cases} (k, m+1) & m \neq M \\ (k+1, 1) & k \neq K \text{ and } m = M \\ \text{undefined} & k = K \text{ and } m = M \end{cases}$$

So, $(km)^{-}$ designates the stage and level a particular product at stage m and level k moves to after one single operation, whereas $(km)^{+}$ is the stage and level that feeds stage m, level k.

Inventory Dynamic Equation

The inventory equations are given by

$$I_{n+1}^{kmp} = \begin{cases} I_n^{11p} - d_n^p + P_n^{11p} & m = 1 \; and \; k = 1 \\ I_n^{kmp} - P_n^{(km)^- p} + P_n^{kmp} & \text{otherwise} \end{cases} \tag{3.1}$$

The first line refers to the depletion of inventory by the external demand at the last stage and level of production. The second line of (3.1) describes the standard evolution of an intermediate level and stage: inventory of a given level and stage is depleted by the amount engaged in production by the downstream stage and level, and is increased by the amount actually produced at the corresponding level and stage.

Echelon Inventory Equation

The echelon inventory is governed by the following equations

$$E_n^{kmp} = \begin{cases} I_n^{11p} & m = 1 \text{ and } k = 1 \\ I_n^{kmp} + E_n^{(km)^- p} & \text{otherwise.} \end{cases} \tag{3.2}$$

It should be easy to verify that the above corresponds to the sum of inventory downstream for each product starting at a given level k and at a given stage m. The first line refers to the last level and stage where the echelon inventory is nothing more than the local inventory. The second line defines the echelon inventory recursively for all the other stages and levels.

Shortfall Dynamic Equation

The shortfall is defined as the difference between the echelon base stock and the echelon inventory.

$$Y_n^{kmp} = z^{kmp} - E_n^{kmp}, \tag{3.3}$$

where z^{kmp} denotes the echelon base stock level for product p at stage m, level k. So the shortfall measures the distance relative to the target echelon inventory and is by construction always non-negative.

Therefore, it is possible to write a dynamic equation for the shortfalls similar to the one for inventories as

$$Y_{n+1}^{kmp} = Y_n^{kmp} + d_n^p - P_n^{kmp}. \tag{3.4}$$

That is, demand moves the echelon inventory away from the target. The net production attempts to restore the echelon inventory to its target (so attempts to make shortfall equal to zero).

Production Net Needs

The production decision is influenced by the production rule that becomes active whenever available capacity is exceeded and by the way capacity is pre-assigned. In any case, it is possible to define the *production net needs* for a given product, level and stage as:

$$f_n^{kmp} = \begin{cases} (z^{KMp} + d_n^p - E_n^{KMp})^+ & m = M \text{ and } k = K \\ \min\left\{(z^{kmp} + d_n^p - E_n^{kmp})^+, I_n^{(km)^+p}\right\} & \text{otherwise,} \end{cases} \tag{3.5}$$

where $(x)^+ = \max\{0, x\}$.

The term f_n^{kmp} denotes the production decision if there would not be any capacity constraints. Basically it states that the production decision should be given by the difference between the echelon inventory after demand is realized and the base stock level, or less if there is a limitation on the inventory available at the previous buffer. Since raw material is assumed to be always available, for $k = K$ and $m = M$ there is no explicit limitation on inventory.

The above equation may be alternatively expressed in terms of the shortfalls as

$$f_n^{kmp} = \min\left\{Y_n^{kmp} + d_n^p, (z^{(km)^+p} - z^{kmp}) - (Y_n^{kmp} - Y_n^{(km)^+p})\right\}, \tag{3.6}$$

by making use of Eqs. (3.2) and (3.3) to express $I_n^{(km)^+p}$ solely as a function of the shortfalls. Note that to keep consistency it is assumed that $z^{(KM)^+p} = +\infty$ and $Y_n^{(KM)^+p} = 0$ for all n.

Initial Conditions

In order to implement the simulation of such model, the initial conditions have to be defined. The state variables will be set at their base stock levels. That is, $I_0^{11p} = z^{11p}$ and $I_0^{kmp} = z^{kmp} - z^{(km)^-p}$ for k and m not simultaneously equal to 1. The echelon inventories will be set according to (3.2), that is $E_0^{kmp} = z^{kmp}$ because the state variables are set to their base stock levels. All other initial variables are set to zero.

Alternative Variables

It is possible to define the control policy by an alternative set of control variables that are directly related to the multi-echelon base stock variables. They are defined as

$$\Delta^{kmp} = \begin{cases} z^{11p} & k = 1 \text{ and } m = 1 \\ z^{kmp} - z^{(km)^-p} & \text{otherwise} \end{cases} \tag{3.7}$$

There are situations where it will be of advantage to consider this alternative set of control variables. The base stock variables have to be ordered, that is $z^{kmp} \geq z^{(km)^-p}$ to keep consistency. To enforce such ordering during simulation may be difficult. On the other hand, to enforce such ordering on this alternative set of variables if suffices to impose a non negativity constraint on all Δ^{kmp}. This is much easier to take care of during optimization. Also, the base stock variables impose an implicit coupling among the stages and levels such that the cost function may have contradictory gradient information for echelon base stock variables associated with consecutive levels and stages. That is, the gradient information may force a base stock variable to be bellow its successor. On those situations it is not at all clear where to move. On the contrary, the Δ variables are not coupled.

Remark. In this chapter I will always make use of the base stock variables for ease of explanation and because their use makes the equations shorter. The validation of the theoretical results do not depend on the set of control variables used. It is the actual numerical simulation that may benefit from one set of variables more than from the other. This will be clarified in Chapter 5.

3.1.2 The Performance Measures

The performance measures considered can be broadly classified in two classes: operational cost based and service level based. The operational cost based measures refer to the traditional assignment of cost to inventories and backlogs. The service level based measures relate to the fraction of times where a stockout does not occur or to the stockout size itself. Of this latter class, this thesis addresses only the occurrence of stockouts.

Operational Cost Based Measures

Let

$$\begin{aligned} h^{kmp} &= \text{holding cost rate for level } k, \text{ stage } m, \text{ and product } p; \\ b^{p} &= \text{backloging cost rate for level 1, stage 1 and product } p \end{aligned} \tag{3.8}$$

and let the single stage cost be defined as

$$C_n = \sum_{p=1}^{P} C_n^p \tag{3.9}$$

where C_n^p is given by

$$C_n^p = (I_n^{11p})^{-} b^p + (I_n^{11p})^{+} h^{11p} + \sum_{m=2}^{M} I_n^{1mp} h^{1mp} + \sum_{k=2}^{K} \sum_{m=1}^{M} I_n^{kmp} h^{kmp}. \tag{3.10}$$

Therefore, the finite horizon average cost is

$$c_N = \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}[C_n], \tag{3.11}$$

the infinite horizon discounted cost is

$$c_{\alpha,\infty} = \mathbf{E}\left[\sum_{n=1}^{\infty} \alpha^n C_n\right], \tag{3.12}$$

where $\alpha \in [0,1]$ is the discount factor. The infinite horizon average cost is

$$c_{\infty} = \lim_{N\to\infty} \mathbf{E}\left[\frac{1}{N}\sum_{n=1}^{N} C_n\right]. \tag{3.13}$$

Service Level Based Measures

For service level measures, we consider only Type-1 service level. That is, we measure the system's performance in terms of serving demand in the same period that it occurs. Let

$$V_N^p = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{I_n^{11p} \geq d_n^p \text{ or } d_n^p = 0\}, \tag{3.14}$$

be the fraction of periods in which demands for product p are filled immediately. Using this measure one can define the following finite horizon measure

$$\bar{V}_N = \frac{1}{P}\sum_{p=1}^{P} V_N^p. \tag{3.15}$$

Based on (3.15), the average Type-1 service level is defined as $\bar{v}_N = \mathbf{E}[V_N]$. The infinite horizon average service level is:

$$\bar{v}_{\infty} = \lim_{N\to\infty} \bar{v}_N. \tag{3.16}$$

3.2 The Derivatives of the Basic Model

The purpose of the IPA derivatives will be primarily to find the optimal echelon base stock levels. However, it is also possible to find the optimal allocation of each stage's capacity to each of the K levels when the system is operated on a partial sharing mode. Therefore, derivatives of the state variables (and other variables for that matter) will have to be taken with respect to all z^{kmp}

and all C^{km} if the purpose is also to find the optimal capacity allocation. Also, it is necessary to take derivatives with respect to the alternative set of variables introduced in Section 3.1.

The base stock levels have to remain ordered during simulation, even under arbitrarily small perturbations, therefore assume that for each p

$$0 < z^{kmp} < z^{(km)^+p} \qquad \text{for all } k, m, \tag{3.17}$$

or alternatively $\Delta^{kmp} > 0$[1]. For partial sharing, the capacity allocation among levels is constrained by the available capacity at any stage, that is

$$\sum_{k=1}^{k=K} C^{km} \quad = \quad C^m \text{ for all } m = 1, 2, \ldots, M. \tag{3.18}$$

Capacity will be assumed unconstrained in the derivatives calculation and, at the end of each simulation run, the computed gradient will be projected on the hyper-plane defined by (3.18), which defines the capacity constraint.

Consider first the derivatives with respect to all z^{kmp}. Let $z^* = z^{k^*m^*p^*}$ denote the variable with respect to which the derivatives are taken for some $k^* = 1, \ldots, K; m^* = 1, \ldots, M; p^* = 1, \ldots, P$ (the derivatives with respect to the Δ variables are similar). The derivative recursions with respect to capacity are presented in Section 3.6.1. The sub-index (z) denotes that the derivatives are taken with respect to some base stock variable. The sub-indices (c) and (Δ) will denote that the differentiation variable belongs to the capacity allocation set and the delta variables, respectively.

3.2.1 Derivatives for the State Variables

- Inventory derivatives

$$I'_{(z)_{n+1}}{}^{kmp} \quad = \quad \begin{cases} I'_{(z)_n}{}^{11p} + P'_{(z)_n}{}^{11p} & m = 1 \text{ and } k = 1 \\ I'_{(z)_n}{}^{kmp} - P'_{(z)_n}{}^{(km)^-p} + P'_{(z)_n}{}^{kmp} & \text{otherwise} \end{cases} \tag{3.19}$$

[1] These inequalities need to be strict to ensure differentiability

- Echelon inventory derivatives

$$E'_{(z)_n}{}^{kmp} = \begin{cases} I'_{(z)_n}{}^{11p} & m = 1 \text{ and } k = 1 \\ I'_{(z)_n}{}^{kmp} + E'_{(z)_n}{}^{(km)^- p} & \end{cases} \tag{3.20}$$

- Shortfall derivatives

$$Y'_{(z)_{n+1}}{}^{kmp} = Y'_{(z)_n}{}^{kmp} - P'_{(z)_n}{}^{kmp} \tag{3.21}$$

- Production net needs derivatives

$$f'_{(z)_n}{}^{kmp} = \begin{cases} \begin{cases} \mathbf{1}\{z^* = z^{KMp}\} - E'_{(z)_n}{}^{KMp} & \text{bound by demand} \\ 0 & \text{if } f_n^{KMp} = 0 \end{cases} & \begin{cases} m = M \\ k = K \end{cases} \\ \begin{cases} \mathbf{1}\{z^* = z^{kmp}\} - E'_{(z)_n}{}^{kmp} & \text{bound by demand} \\ I'_{(z)_n}{}^{(km)^+ p} & \text{bound by supply} \\ 0 & \text{if } f_n^{kmp} = 0 \end{cases} & \text{otherwise} \end{cases} \tag{3.22}$$

- Initial conditions derivatives

$$\begin{aligned} I'_{(z)_0}{}^{11p} &= \mathbf{1}\{z^* = z^{11p}\} \\ I'_{(z)_0}{}^{kmp} &= \mathbf{1}\{z^* = z^{kmp}\} - \mathbf{1}\{z^* = z^{k(m-1)p}\} \quad m \neq 1 \\ I'_{(z)_0}{}^{k1p} &= \mathbf{1}\{z^* = z^{k1p}\} - \mathbf{1}\{z^* = z^{(k-1)Mp}\} \quad \text{otherwise} \end{aligned}$$

for the echelon inventory the derivatives are trivial, that is, $E'_{(z)_0}{}^{kmp} = \mathbf{1}\{z^* = z^{kmp}\}$. The initial shortfall derivatives are zero.

3.2.2 Derivatives of the Performance Measures

In this section I will limit the discussion to the presentation of the single stage performance derivatives. Later, when validating the method, the complete analysis will be done. The derivative of the single stage cost is:

$$C'_n = \sum_{p=1}^{P} C'^p_n. \tag{3.23}$$

The expression for C'^p_n is given by

$$\begin{aligned} C'^p_n &= -\mathbf{1}\{I_n^{11p} < 0\}(I')_n^{11p} b^p + \\ &\quad +\mathbf{1}\{I_n^{11p} > 0\}(I')_n^{11p} h^{11p} + \sum_{m=2}^{M} I'^{1mp}_n h^{1mp} + \sum_{k=2}^{K}\sum_{m=1}^{M} I'^{kmp}_n h^{kmp}. \end{aligned} \tag{3.24}$$

The finite-horizon average Type-1 service level measure, given by equation (3.15) defining $\bar{V}_N$, is not differentiable because it is not even continuous, as pointed in [Glasserman and Tayur, 1995]. The strategy to obtain a differentiable representation is to replace the indicator function in (3.15) with a conditional expectation. Let

$$\Phi_n^p(x) = \int_0^x \phi_n^p(t)dt, \tag{3.25}$$

where $\phi_n^p(.)$ is the probability density function of the demand of product p on period n. It is possible to show that (see Appendix A),

$$\bar{v}_N = \mathbf{E}[\bar{V}_N] = P^{-1}\sum_{p=1}^{P}\left(N^{-1}\sum_{n=1}^{N} Pr(d_n^p = 0) + \mathbf{E}[N^{-1}\sum_{n=1}^{N}\Phi_n^p(I_n^{11p})]\right).$$

Since $Pr(d_n^p = 0)$ does not depend on z^* or C^*, working only with

$$\tilde{\bar{V}}_N = P^{-1}\sum_{p=1}^{P} N^{-1}\sum_{n=1}^{N}\Phi_n^p(I_n^{11p}), \tag{3.26}$$

will suffice. As a function of z^* or C^*, $\tilde{\bar{V}}_N$ is differentiable, except possibly on the zero-probability event that some I_n^{11p} equals zero, [Glasserman and Tayur, 1995]. For all other non zero probability events, it holds that

$$\tilde{\bar{V}}'_N = P^{-1} \sum_{p=1}^{P} N^{-1} \sum_{n=1}^{N} \mathbf{1}\{I_n^{11p} > 0\} \phi_n^p (I_n^{11p})(I_n^{11p})'. \tag{3.27}$$

Before moving on to the specifics of the production rules, a brief discussion on the static and dynamic capacity management is in order, starting with the static part. Each machine m, for $m = 1, \ldots, M$, on the re-entrant line has some capacity C^m. If that total capacity is divided in $K \times P$ single slots, C^{kmp}, and assigned to each product and level, the system is run in a no sharing mode (NS). By dividing the total capacity in K slots, C^{km}, assigned to each level to be shared by all products at that level, the system is being run in a partial sharing mode (PS). Finally, if the total capacity is equally shared by all products at any level, the system operates in a total sharing mode (TS).

Whenever there is some capacity sharing (so for PS and TS modes), it is necessary to establish a dynamic capacity management scheme. That is, it is necessary to decide, in each period, how the available capacity is going to be distributed among the several competing products, if there is a shortage of capacity regarding the total production need. This is what each one of the three production rules proposed in Section 3.4 does. They are the Linear Scaling Rule (LSR), which proportionally scales down all needs to fit capacity; the Priority Rule (PR), which assigns capacity to products in decreasing order of their priority, which is established up-front; and the Equalize Shortfall Rule, which assigns capacity to products in decreasing order of their present distance to target levels.

3.3 Production Decisions and their Derivatives

In what follows we will present the production rules together with the corresponding derivative expressions. The production rules are designed to take care of the dynamic capacity allocation. That is, they are activated whenever the sum of the production net needs for a given machine and/or level is above the available capacity.

3.3.1 Linear Scaling Rule with Partial Sharing

For the Linear Scaling Rule (LSR), and assuming there is partial sharing (PS) of capacity, the production decision is defined as:

$$P_n^{kmp} = f_n^{kmp} g_n^{km}, \tag{3.28}$$

where f_n^{kmp} is given by (3.5), its derivative is given by (3.22), and

$$g_n^{km} = \min\left\{\frac{C^{km}}{\sum_p f_n^{kmp}}, 1\right\}. \tag{3.29}$$

The term g_n^{km} considers the existence of a capacity constraint. Whenever the unconstrained production requirements fall below capacity the system behaves as if it were uncapacitated. If capacity is less than the requirements, then all requirements are linearly scaled down to fit the system's capacity. Note that, as mentioned earlier, the diverse products are assumed to impose the same load on the level and stage per unit processed. This is what we have been referring as the discrete time inventory control counterpart of a Kelly type network, or the *uniform load assumption.*

The derivative expression for g_n^{km} is given by

$$g'_{(z)_n}{}^{km} = \begin{cases} \frac{-C^{km} \sum_p f'_{(z)_n}{}^{kmp}}{(\sum_p f_n^{kmp})^2} & \text{bound by capacity} \\ 0 & \text{no bound in capacity} \end{cases} \tag{3.30}$$

Therefore, the derivative expression for P_n^{kmp} is

$$P'_{(z)_n}{}^{kmp} = f'_{(z)_n}{}^{kmp} g_n^{km} + f_n^{kmp} g'_{(z)_n}{}^{km}. \tag{3.31}$$

3.3.2 Linear Scaling Rule with Total Sharing

For the Linear Scaling Rule (LSR), and assuming there is total sharing (TS) of capacity, the production decision is defined as

$$P_n^{kmp} = f_n^{kmp} g_n^{m}, \tag{3.32}$$

where g_n^m assumes the obvious extension

$$g_n^m = \min\left\{\frac{C^m}{\sum_{k=1}^{K}\sum_{p=1}^{P} f_n^{kmp}}, 1\right\}. \tag{3.33}$$

Naturally, the derivative for P_n^{kmp} is given as

$$P'_{(z)}{}_n^{kmp} = f'_{(z)}{}_n^{kmp} g_n^m + f_n^{kmp} g'_{(z)}{}_n^m, \tag{3.34}$$

with

$$g'_{(z)}{}_n^m = \begin{cases} \dfrac{-C^m \sum_{k=1}^{K}\sum_{p=1}^{P} f'_{(z)}{}_n^{kmp}}{\left(\sum_{k=1}^{K}\sum_{p=1}^{P} f_n^{kmp}\right)^2} & \text{bound by capacity} \\ 0 & \text{no bound in capacity} \end{cases} \tag{3.35}$$

3.3.3 Priority Rule with Partial Sharing

Assume that we assign capacity according to a priority for the products. Assume that $p(i)$, for $i = 1, \ldots, P$, is the product that comes in the ith position on the priority list, that is, product $p(1)$ is the product with the highest priority and product $p(P)$ has the lowest priority. The production decision will be

$$\begin{aligned} P_n^{kmp(1)} &= \min\{f_n^{kmp(1)}, C^{km}\} \\ P_n^{kmp(2)} &= \min\{f_n^{kmp(2)}, C^{km} - P_n^{kmp(1)}\} \\ &\vdots \\ P_n^{kmp(i)} &= \min\{f_n^{kmp(i)}, C^{km} - \sum_{j=1}^{i-1} P_n^{kmp(j)}\} \end{aligned} \tag{3.36}$$

where $i \leq \min\{i^*, P\}$.

If product $p(i^*)$ fills capacity, we have that for all $p = p(1), \ldots, p(i^* - 1)$, all net needs are satisfied, whereas for all $p = p(i^* + 1), \ldots, p(P)$ nothing is produced. For the particular case of $p = p(i^*)$, only the available capacity defined by $C^{km} - \sum_{j=1}^{i^*-1} P_n^{kmp(j)}$ will be used to approach its base stock.

According to (3.36) the production derivatives for the PR are

$$
\begin{aligned}
P'_{(z)_n}{}^{kmp(1)} &= \begin{cases} f'_{(z)_n}{}^{kmp(1)} & \text{bound by the production net needs} \\ 0 & \text{bound by capacity} \end{cases} \\
P'_{(z)_n}{}^{kmp(i)} &= \begin{cases} f'_{(z)_n}{}^{kmp(i)} & \text{bound by the production net needs} \\ -\sum_{j=1}^{i-1} P'_{(z)_n}{}^{kmp(j)} & \text{bound by capacity} \end{cases}
\end{aligned}
\tag{3.37}
$$

Note that for those products whose production was zero due to capacity constraints, the derivatives are zero as well.

3.3.4 Priority Rule with Total Sharing

Assume that we assign capacity according to a priority for the products and levels. Assume that $k(i)$ and $p(i)$, for $i = 1, \ldots, K \times P$ are the level and product with the ith position on the priority list, that is, the value of $P_n^{k(1)mp(1)}$ is the first to be decided and the value of $P_n^{k(K\times P)mp(K\times P)}$ is the last. The production decisions and their derivatives are naturally similar to the ones on the previous section and their presentation is omitted for the sake of brevity.

3.3.5 Equalize Shortfall Rule with Partial Sharing

Another way to dynamically allocate capacity is by trying to equalize the *shortfall* for every product. The shortfall is defined as the difference between the target echelon base stock and the current echelon inventory. If at a given instant of time the net production needs are below capacity the production decision is given as before by:

$$
f_n^{kmp} = \min\{Y_n^{kmp} + d_n^p, I_n^{(km)^+p}\}. \tag{3.38}
$$

In contrast to the previous two production rules, the production decision is obtained iteratively when the net production needs exceed the available capacity. The intuition behind the algorithm below is that one should start by allocating capacity to the product with the highest shortfall, that is, the product that is most away from its target level, until it reaches the level of the product with the second highest shortfall. After, capacity will be assigned in equal parts to both products until their shortfalls match that of the third highest shortfall, and so forth. Notice that, at any point, it may be the case that the shortfalls are not made equal at the end of the production decision because of insufficient inventory for some products or just because capacity is exhausted.

Under the framework of partial sharing and uniform loads the following algorithm is applied for each $k = 1, \ldots, K$ and $m = 1, \ldots, M$,

Equalize Shortfall Procedure

Step 0. For all $p = 1, \ldots, P$ set $\mathcal{Y}^{kmp} = Y_n^{kmp} + d_n^p$, $\mathcal{Y}'^{kmp} = Y'^{kmp}_n$, $P_n^{kmp} = P'^{kmp}_n = 0$, $\mathcal{I}^{kmp} = I_n^{kmp}$, and $\mathcal{I}'^{kmp} = I'^{kmp}_n$.

Also, set $\mathcal{C}^{km} = C^{km}$, $\mathcal{C}'^{km} = C'^{km}$, and $j = P$.

Step 1. Order the products by decreasing order of their shortfall after demand is realized. Let $p(1), \ldots, p(j)$ denote that ordering, that is $\mathcal{Y}^{kmp(1)}$ is the maximum value and $\mathcal{Y}^{kmp(j)}$ is the minimum.

Set $l = 1$ and $\mathcal{Y}^{km(j+1)} = \mathcal{Y}'^{km(j+1)} = 0$.

Step 2. Let $\Delta = \mathcal{Y}^{kmp(l)} - \mathcal{Y}^{kmp(l+1)}$. If $\Delta \neq 0$, set $\Delta' = \mathcal{Y}'^{kmp(l)} - \mathcal{Y}'^{kmp(l+1)}$ and go to Step 4. Otherwise, continue.

Step 3. If $l < j$, set $l = l + 1$ and go to Step 2. Otherwise, STOP.

Step 4. The first l products are tied. Therefore the production decision and its derivative are updated as follows:

$$P_n^{kmp(i)} = P_n^{kmp(i)} + \mathcal{P}^{kmp(i)} \quad \text{for } i = 1, \ldots, l. \tag{3.39}$$

$$P'^{kmp(i)}_n = P'^{kmp(i)}_n + \mathcal{P}'^{kmp(i)} \quad \text{for } i = 1, \ldots, l. \tag{3.40}$$

where

$$\mathcal{P}^{kmp} = \min\{\Delta, \mathcal{I}^{(km)^+p}, \mathcal{C}^{km}/l\} \tag{3.41}$$

and

$$\mathcal{P}'^{kmp} = \begin{cases} \Delta' & \text{if bound by the jump size} \\ \mathcal{I}'^{(km)^+p}_n & \text{if bound by inventory} \\ \mathcal{C}'^{km}/l & \text{if bound by capacity} \end{cases} \tag{3.42}$$

Step 5. Update the shortfalls, inventories, and available capacity.

$$\begin{aligned} \mathcal{Y}^{kmp(i)} &= \mathcal{Y}^{kmp(i)} - \mathcal{P}^{kmp(i)} \\ \mathcal{I}^{(km)^+ p(i)} &= \mathcal{I}^{(km)^+ p(i)} - \mathcal{P}^{kmp(i)} \text{ for } i = 1, \ldots, l, \\ \mathcal{C}^{km} &= \mathcal{C}^{km} - \sum_{i=1}^{l} \mathcal{P}^{kmp(i)} \end{aligned} \tag{3.43}$$

Similarly the derivatives are

$$\begin{aligned} \mathcal{Y}'^{kmp(i)} &= \mathcal{Y}'^{kmp(i)} - \mathcal{P}'^{kmp(i)} \\ \mathcal{I}'^{(km)^+ p(i)}_{n} &= \mathcal{I}'^{(km)^+ p(i)}_{n} - \mathcal{P}'^{kmp(i)} \text{ for } i = 1, \ldots, l. \\ \mathcal{C}'^{km} &= \mathcal{C}'^{km} - \sum_{i=1}^{l} \mathcal{P}'^{kmp(i)} \end{aligned} \tag{3.44}$$

Step 6. If $\mathcal{C}^{km} = 0$, STOP. The total production for level k and stage m is bound by capacity. Otherwise, continue.

Step 7. For each $i = 1, \ldots, l$, if $\mathcal{I}^{(km)^+ p(i)} = 0$ remove product $p(i)$ from the list and set $j = j - 1$. If $j = 0$, STOP. The total production for level k and stage m does not use up all capacity. Otherwise, go to Step 1.

Notice that the algorithm produces (3.5) if there is no bound on capacity. Also note that the algorithm generates the production decision and its derivative at the same time. At the end of the Equalize Shortfall Procedure, P_n^{kmp} will contain the production decision and P'^{kmp}_n will contain its derivative.

3.3.6 Equalize Shortfall Rule with Total Sharing

For the setting of total sharing and uniform loads we will skip the repetitious and lengthy presentation of the Equalize Shortfall Procedure. With minor changes of the algorithm above it is possible to obtain the procedure that has to be repeated for every $m = 1, \ldots, M$.

3.4 Finite Horizon Validation

The basic procedure to establish the validity of the IPA approach is introduced here for the state variables. Almost all the necessary formal results for IPA follow a similar scheme. To prove

validity one has to first show that all variables are differentiable, then indicate what are their values, and finally prove that the expected value and the derivative are permutable operators.

To obtain the derivatives of the state variables, simply differentiate the equations governing the system, as shown before. The problem is that some of those equations contain non-differentiable terms. The non-differentiable terms are due to the existence of min and max operators. Differentiability is ensured only if either ties occur with zero probability or if the derivatives of the tying terms in the operators are equal when ties occur with nonzero probability.

The concept of Lipschitz function and a technical lemma will be used in the validation proofs.

Definition 3.4.1 A function ϕ mapping $S \in \mathbf{R}$ into $\mathbf{R}$ is *Lipschitz* if there exists a constant k_ϕ, called the modulus, for which

$$|\phi(x) - \phi(y)| \leq k_\phi |x - y|. \tag{3.45}$$

Definition 3.4.2 A random function is *Lipschitz with probability one* if there exists a random variable K that serves as a path-wise modulus.

It is convenient to state here Lemma 3.2 of [Glasserman and Tayur, 1995]:

Lemma 3.4.3 Let $\{X(s), s \in S\}$ be a random function with S an open subset of $\mathbf{R}$. Suppose that $\mathbf{E}[X(s)] < \infty$ for all $s \in S$. Suppose, further, that X is differentiable at $s_0 \in S$ with probability one, and that X is almost surely Lipschitz with modulus K_X satisfying $\mathbf{E}[K_X] < \infty$. Then $\mathbf{E}[X(s_0)]'$ exists and equals $\mathbf{E}[X'(s_0)]$.

3.4.1 Preliminary Results on the Equalize Shortfall Algorithm

In order to validate the IPA methodology, there are two results specific to the ESR that are needed so that Eq. (3.42) is correct.

Proposition 3.4.4 The ordering generated in Step 1 of the Equalize Shortfall Procedure at the first iteration remains unchanged with probability one in a neighborhood of the base stock levels.

Proof: Assume first that for a given vector z there are no ties in the shortfall quantities after demand is realized at the beginning of period n. Under this assumption, there exists a $\delta =$

$\min_i\{\mathcal{Y}^{kmp(i)} - \mathcal{Y}^{kmp(i+1)}\} > 0$. Therefore, there exists an $\epsilon > 0$ such that a change smaller than ϵ in any of the components of z will produce a change in shortfall variables which is smaller than δ. Thus, the ordering remains unchanged.

Now consider the case where at least two shortfall quantities are tied at period n. A tie in $\mathcal{Y}$ can only occur if they were equal at the end of period $n-1$ and demand in period n was zero for at least those two products. However, if at least two shortfall quantities were made equal in period $n-1$ for a given vector z, this means that with probability one they will also be made equal in period $n-1$ in a neighborhood of z.

The reason is that, with probability one either capacity was exhausted but there was some inventory not used up for at least those two products on period $n-1$ with vector z, or capacity was not exhausted and the shortfalls were made zero on period $n-1$.

Thus, the result follows.

□

With this result and having the proof in mind it is now easy to establish the following stronger result, which is not as intuitively obvious but crucial to validate Eq. (3.42).

Proposition 3.4.5 At each iteration of the Equalize Shortfall Procedure, the number of tied products l, at the beginning of Step 4 remains unchanged with probability one in a neighborhood of z.

Proof: The proof goes by induction on the number of iterations of the above procedure. Denote by $l(r)$ the number of products tied at the beginning of Step 4 during iteration r. Proposition 3.4.4 establishes the result for the first iteration. Now assume that at some iteration $r-1$, $l(r-1)$ is invariant relative to sufficiently small changes in the base stock variables. We want to see what happens to $l(r)$.

From the $l(r-1)$ products, let us assume that some $\lambda(r-1) \leq l(r-1)$ remain tied upon application of Step 4. With probability one there are only two ways under which this can happen. The first is for the $\lambda(r-1)$ to have inventory enough for the amount Δ to be assigned to each of them together with available capacity to do so. In this situation for a sufficiently small neighborhood of z, the inventory and the capacity available will still allow the same $\lambda(r-1)$ products to remain tied.

The second way may occur if there is a bound in capacity for all $\lambda(r-1)$ products. However, a bound in capacity will remain for a sufficiently small neighborhood of z, w.p.o. Also, $\lambda(r-1) < l(r-1)$ iff some products have their share reduced due to insufficient inventory, which again will remain as so w.p.o. in a neighborhood of z.

Thus, $\lambda(r-1)$ is invariant in a neighborhood of z. The value of $l(r)$ depends on $\lambda(r-1)$ and on what happens on Step 4. If the production decision for the $\lambda(r-1)$ products is bound by Δ, then there is more capacity available to execute a new iteration. Also, a new iteration will take place if there are still products with nonzero shortfalls. In this case $l(r)$ will be the sum of $\lambda(r-1)$ with the products that were tied for second place before iteration $r-1$. This is because the shortfall of the $\lambda(r-1)$ products were brought down to the same levels as those. By similar arguments, the number of products tied for second place does not change for sufficiently small changes of any base stock variable.

There is also the possibility that the tie of the $\lambda(r-1)$ products occurred due to a bound in capacity. If $\lambda(r-1) = l(r-1)$, there will be no more iterations, since the available capacity at the beginning of iteration $r-1$ will be exhausted during this iteration. It can happen that $\lambda(r-1) < l(r-1)$, since some of the products may have been bounded by inventory. In this second situation, there will be capacity available to perform at least one more iteration, and it turns out that $l(r) = \lambda(r-1)$.

Therefore, the result of the proposition follows.

□

Propositions 3.4.4 and 3.4.5 are valid for the TS mode without change. The proofs are exactly the same. What these two propositions establish is that the derivative of l in Eq. (3.42) with respect to the base stock variables is zero, and thus the equation is correct.

3.4.2 Validation of the State Variables

Theorem 3.4.6 establishes the main validation result for the state variables and their derivatives with respect to the base stock for a system operated under any rule and any capacity sharing mode.

Theorem 3.4.6 If $\{d_n^p, n = 1, 2, \ldots, p = 1, 2, \ldots, P\}$ are independent and each d_n^p has a density on $(0, \infty)$, then the following hold:

1. For $k = 1, \ldots, K, m = 1, \ldots, M, p = 1, \ldots, P$, and $n = 1, 2, \ldots$, each I_n^{kmp} and E_n^{kmp}, as given by (3.1) and (3.2) respectively is, w.p.o., differentiable at $(z^{111}, \ldots, z^{KMP})$ with respect to each z^{qrs}, $q = 1, \ldots, K, r = 1, \ldots, M$, and $s = 1, \ldots, P$. Moreover, these derivatives satisfy (3.19) and (3.20), respectively. Also for

   a. *Linear Scaling Rule on Partial Sharing*

      P_n^{kmp} as given by (3.28) is also differentiable w.p.o. and its derivative satisfies (3.31);

   b. *Linear Scaling Rule on Total Sharing*

      P_n^{kmp} as given by (3.32) is also differentiable w.p.o. and its derivative satisfies (3.34);

   c. *Priority Rule on Partial Sharing*

      P_n^{kmp} as given by (3.36) is also differentiable w.p.o. and its derivative satisfies (3.37);

   d. *Priority Rule on Total Sharing*

      P_n^{kmp} as given by the natural extension of (3.36) for Total Sharing is also differentiable w.p.o. and its derivative satisfies the natural extension of (3.37) for this mode;

   e. *Equalize Shortfall Rule on Partial Sharing*

      P_n^{kmp} as given by (3.39) is also differentiable w.p.o. and its derivative satisfies (3.40);

   f. *Equalize Shortfall Rule on Total Sharing*

      P_n^{kmp} as given by the natural extension of (3.39) for Total Sharing is also differentiable w.p.o. and its derivative satisfies the natural extension of (3.40) for this mode.

2. *For all production rules and capacity sharing modes*

   If in addition $\mathbf{E}[d_n^p] < \infty$ for all n, then $\mathbf{E}[I_n^{kmp}]'_{(z)}$, $\mathbf{E}[E_n^{kmp}]'_{(z)}$, and $\mathbf{E}[P_n^{kmp}]'_{(z)}$ exist and equal $\mathbf{E}[(I'_{(z)})_n^{kmp}]$, $\mathbf{E}[(E'_{(z)})_n^{kmp}]$, and $\mathbf{E}[(P'_{(z)})_n^{kmp}]$.

Proof: I will only present the proof for the LSR in the PS mode. The reasoning presented carries through for the other rules and capacity sharing schemes in a straightforward manner.

For part (1.a) it is the case that the differentiability of the state variables relies on the structure of the recursive equations defining them. Due to the structure of (3.28) one has only to check if (3.5) and (3.29) are differentiable. The remaining equations being linear combinations of state variables do not pose any problem relative to their differentiability. So, let us concentrate on (3.5) and (3.29).

- Equation (3.5)

  The following reasoning applies to the second line of (3.5). (The specifics of the first line are trivial.) A tie between the two terms of (3.5) may induce non-differentiability. Also, the term $(z^{kmp} + d_n^p - E_n^{kmp})^+$ may induce non-differentiability when $z^{kmp} + d_n^p - E_n^{kmp} = 0$. Since demands are continuous, ties between the two terms of (3.5) have zero probability, except for the case where both terms are zero.

  $z^{kmp} + d_n^p - E_n^{kmp} = 0$ only if $d_n^p = 0$[2] (the echelon inventory never exceeds z^{kmp}), in which case the inventory reached its base stock level in the previous period. Thus, since demands are continuous, if the inventory reaches its base stock level at some value z^*, then w.p.o. it does so through a neighborhood of z^*. Therefore, the first term of (3.5) remains zero throughout a neighborhood and introduces no non-differentiability. A similar reasoning is valid for the second term of (3.5)[3], i.e., if the second term is zero, then it remains zero throughout a neighborhood of z^* w.p.o..

- Equation (3.29)

  Again, a tie between the two terms of (3.29) may induce non-differentiability. However, since f_n^{kmp} is a continuous random variable, $\sum_{p=1}^{P} f_n^{kmp}$ is also a continuous random variable. Therefore, the event $\frac{C^{km}}{\sum_p f_n^{kmp}} = 1$ has zero probability *in general* (see remark below). Due to the fact that $Pr\{f_n^{kmp} = 0\} > 0$, P is finite, and f_n^{kmp} is always positive, a non-differentiability could be induced when $\sum_{p=1}^{P} f_n^{kmp} = 0$. However, under that case (3.29) is trivially equal to 1 for which the derivative exists and equals zero.

Therefore, w.p.o., differentiability is preserved at each period.

Remark: There may be exceptions to this situation.

- If for some stage and level $C^{km} = C^{(km)^+}$, then $Pr\{\sum_p f_n^{kmp} = C^{km}\} \neq 0$, no matter what the values of the control variables are. In fact, since $Pr\{I_n^{kmp} = 0\} \neq 0$, there is the possibility that at some period all the inventory sitting in front of machine m at level k is zero. Simultaneously it may happen that the feeding stage, $(km)^+$, is bound by capacity implying that on the next period we will have a tie

[2] Demands are only assumed continuous in $(0, \infty)$. It is possible that $Pr\{d_n^p = 0\} > 0$, otherwise the event would have zero probability and therefore would not induce non-differentiability.

[3] The second term on (3.5) has a point mass at zero.

between C^{km} and the total production net needs. Note, however, that such bound in capacity will occur with probability one in a neighborhood of z^*. Therefore, the derivative will be zero for both terms of (3.29).

- Let us take for example the case where for some stage and level the value of $z^{(km)^+p} - z^{kmp} = C^{km}$. For this situation non-differentiability is induced because, if for example we take derivatives with respect to $z^{(km)^+p}$, the first term of (3.29) will have a nonzero derivative, the second term has zero derivative, and there is a nonzero probability that the tie will occur. One could argue that there is a slim chance that $z^{(km)^+p} - z^{kmp} = C^{km}$. In fact that is the case in general. However, in many circumstances it turns out that the optimal values of the base stock variables are exactly given by the above equality (see Chapter 5). Therefore, it is not unlikely that the simulation will have to be run close to such configurations and in those cases this theorem does not apply[4].

  In order to still be able to apply IPA to those configurations we have to slightly change the formulation in order to obtain a fully differentiable model. In fact, if we replace all occurrences of $z^{(km)^+p}$ by $z^{kmp} + C^{km}$ we obtain a fully differentiable model with one less variable. For such a modified model the present theorem will then be applicable. There are many other choices for the values of the base stock variables that induce similar types of non-differentiabilities at optimality. In any such cases, it is always possible to reduce the original problem to another with less variables that is differentiable with probability one. These situations will be examined in the numerical study (Chapter 5).

Regarding part 2 of Theorem 3.4.6, according to Lemma 3.4.3 we only have to show that the system variables are, with probability one, Lipschitz functions of the base stock levels having integrable moduli. The proof goes by induction on n. Since the state variables at time zero are linear on the base stock levels, they are Lipschitz. Since the operations min, max, addition and multiplication preserve that property, it follows that each I_n^{kmp}, E_n^{kmp}, and P_n^{kmp} is a composition of Lipschitz functions, and therefore is Lipschitz.

Remark: Division does not preserve the Lipschitz property in general due to the possibility of the term in the denominator being zero for finite values of the variable.

[4]The same problem occurs for zero values of the Δ variables.

This would imply the resulting function to be unbounded for finite values of the variable which is not consistent with the Lipschitz property. Therefore the term g_n^{km} deserves special attention due to the existence of a division. However, since whenever the term $\sum_p f_n^{kmp}$ drops below C'^{km}, g_n^{km} will be equal to 1 and since the transition from the first argument in the min{.} to the second is done smoothly, g_n^{km} preserves the Lipschitz property.

Since $\mathbf{E}[d_n^p] < \infty$ for all n, then every I_n^{kmp} has finite expectation. Consequently, each E_n^{kmp} has finite expectation. Also, each P_n^{kmp} is integrable because they are bounded. In the context of single product, non re-entrant systems it has been shown, [Glasserman and Tayur, 1995], that $|I'_{(z)_n}{}^{kmp}|$, $|E'_{(z)_n}{}^{kmp}|$, and $|P'_{(z)_n}{}^{kmp}|$ are bounded by unity. Such is not the case in a multiple product re-entrant setting, as is this case, or even in a single product setting, as long as there is re-entrance. If some base stock level changes by a small amount δ then each state variable or production quantity may change by more than δ.

To show that the derivatives are bounded I will use the Δ variables to make exposition clearer. If we establish boundedness on the derivatives with respect to these, it is straightforward to establish the same for the base stock variables.

Irrespective of the production rule used, changing $\Delta^{k^*m^*p^*}$ by a small amount δ implies that the echelon inventory for level K and stage M changes by δ only for product p^*. Notice that the production decisions for level K and stage M, whatever the product, are only bound by capacity or by demand. Neither of these depends on the control variables. Thus, the sum of the inventories' derivatives along the production line (the top echelon derivatives) is zero for all products, except for p^* where it is equal to 1.0 for all products. Note that such is also the case for the derivatives of all echelons from (KM) down to $(k^*m^*)^+$. Things change only from (k^*m^*) onwards. However, no matter what, the conclusion for (KM) tells us that the sum of the individual inventory derivatives for each product with respect to $\Delta^{k^*m^*p^*}$ is bounded.

One could argue that, although their sum is bounded, each individual derivative could be unbounded. Therefore the above does not suffice. Take two sample paths and let us compare how they evolve. Let the first one be called the *nominal path* and the second the *disturbed path*. For the nominal path make $\Delta^{kmp} = \Delta_N^{kmp}$, for the disturbed path make $\Delta^{k^*m^*p^*} = \Delta_N^{k^*m^*p^*} + \delta$, and for all other values of k, m, and p make $\Delta^{kmp} = \Delta_N^{kmp}$. The disturbed path is obtained from the

nominal path by changing only one of the control variables by δ.

As long as the production of level/stage $(k^*m^*)^-$ is not bound by inventory for product p^* it will be the case that the two paths only differ at the inventory amount $I_n^{k^*m^*p^*}$ which is in the disturbed path δ above the value of the nominal path. Therefore, all the state derivatives with respect to $\Delta^{k^*m^*p^*}$ are zero, except for the inventory just mentioned, which has a derivative of 1.0.

Things change the first time $I_n^{k^*m^*p^*}$ bounds production. There are two possibilities here — either there is also a bound in capacity for level/stage $(k^*m^*)^-$ or there is no such bound.

1. *There is also a bound in capacity.* The total amount produced on both paths is the same, which implies that $||I_{n+1}^{(k^*m^*)^-}||$ (defined as the sum over all p) does not change from the nominal to the disturbed path. Also, $||I_{n+1}^{k^*m^*}||$ (sum over all p) remains in the disturbed path above that of the nominal path by exactly δ. What changes are the individual values of the inventories on these consecutive levels/stages. In any case, the derivatives of the individual inventories are all under 1.0 in modulus. The extra δ has the effect of reducing the production amounts of the products other than p^*, implying that, in the disturbed path, the inventories $I_{n+1}^{k^*m^p}$ will all be at least equal to those of the nominal path. Moreover, the inventories $I_{n+1}^{(k^*m^*)^-p}$ in the disturbed path will at most be equal to those of the nominal path, except for product p^* which will be above in the disturbed path. How the δ is distributed depends on the particular production rule used.

2. *There is no bound in capacity.* The extra δ existing at level/stage (k^*m^*) for product p^* will be entirely moved to level $(k^*m^*)^-$ for that product, implying that the sum $||I_{n+1}^{k^*m^*}||$ is the same for both paths and the sum $||I_{n+1}^{(k^*m^*)^-}||$ will be higher by δ on the disturbed path, relative to the nominal path.

So, a simultaneous bound in inventory and capacity retains the δ amount where it originally lies, whereas just a bound in inventory moves it downwards one step. This is in terms of the sum of the inventories for all products at each pair (km).

In terms of the individual components of inventory, each one changes by no more than δ and their sum of changes is either zero or exactly δ, depending on the pair (km) being monitored. In any case, the disturbed path only differs from the nominal at most in two consecutive levels/stages after such bound occurs.

Now it remains to see how these changes evolve for the next periods. One thing should be made clear before moving on. A bound in capacity, with or without inventory bounds, retains the overall sum change where it was before the period, but it may carry the individual changes downwards. A bound in the changed inventories without capacity bound moves the individual changes down a step and may carry the overall change too. If there is no bound in capacity nor a bound in inventories (renewal) the changes move backwards. They will keep moving backwards while there are no bounds in capacity and inventory until a complete renewal is attained, where the only difference between the disturbed path and the nominal path is due to the changed $\Delta^{k^* m^* p^*}$.

Since bounds in capacity retain the overall change, and bounds in inventory move it downwards, only when two consecutive levels/stages have the right combination of bounds, that is the upper most has a bound in inventory with no bound in capacity and the lower one has a bound in capacity, will we have a situation where the changes may add up to values that overall are above δ.

It is when this happens that derivatives above 1.0 may show up. However, it should be clear that the changes add up (do not multiply), what implies that there will always be an adding of fractions of δ. Since these occurrences are finite on a finite horizon setting it follows now that the overall derivatives will at most be $q\delta$ in modulus for some finite, positive, and undetermined $q \in R$.

So, $|I'_{(z)_n}{}^{kmp}|$, $|E'_{(z)_n}{}^{kmp}|$, and $|P'_{(z)_n}{}^{kmp}|$ are bounded as functions of the base stock variables because of the finite horizon setting.

Thus, Lemma 3.4.3 is applicable and the result follows.

□

3.4.3 Validation of the Performance Measures

Turning to the validation of the derivatives of the performance measures, for the operational cost based measure, the following result holds.

Theorem 3.4.7 If, for $n = 1, 2, \ldots, N$, $\mathbf{E}[d_n^p] < \infty$ for all $p = 1, 2, \ldots, P$, the d_n^p are independent, and each d_n^p has a density on $(0, \infty)$, then C'_n exists with probability one and

$$E\left[\frac{1}{N}\sum_{n=1}^{N} C'_n\right] = c'_N. \tag{3.46}$$

Proof: The proof follows the reasoning described at the beginning of Section 3.2. First, the variables I_n^{11p} take the value zero with probability zero. For other inventory variables, if one of them takes the value zero then, with probability one, it remains zero throughout a neighborhood of z^*, because demands are continuous. So, with probability one, C'_n is as stated in Equations (3.23) and (3.24).

Since C_n is Lipschitz with modulus $K_I(\sum_{p=1}^{P} b^p + \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{p=1}^{P} h^{kmp})$, where

$$K_I = \max_{k,m,p,n}\{|{I'_{(z)}}_n^{kmp}|\}, \tag{3.47}$$

the result of (3.46) follows by Lemma 3.4.3 because K_I is finite recall the proof of Theorem 3.4.6).

□

Validation of the Type-1 service level measure follows from [Glasserman and Tayur, 1995]. The result is stated without proof.

Theorem 3.4.8 If, in addition to the conditions of Theorem 3.4.7, ϕ_n^p is bounded for all n and p, then $\mathbf{E}[\tilde{\bar{V}}'_N] = \bar{v}'_N$.

3.5 Infinite Horizon Validation

Consider first the infinite horizon α-discounted cost $c_{\alpha,\infty}$, as defined in (3.12), with $0 < \alpha < 1$. The echelon inventory associated with each given product at any particular stage and level cannot exceed the corresponding echelon base stock level. Also, the backlog for any product cannot exceed the corresponding cumulative demand. Therefore, it follows that

$$\mathbf{E}[C_n^p] \leq \left(b^p + \sum_{k=1}^{K}\sum_{m=1}^{M} h^{kmp}\right)\left(\sum_{k=1}^{K}\sum_{m=1}^{M} z^{kmp} + K \times M \sum_{i=1}^{n}\mathbf{E}[D_i^p]\right). \tag{3.48}$$

Consequently, with

$$\sup_{n\geq 1}\mathbf{E}[D_n^p] < \infty \text{ for all } p = 1,\ldots,P, \tag{3.49}$$

it is insured that $c_{\alpha,\infty}$ is finite.

Computing a derivative estimate for $c_{\alpha,\infty}$ from its infinite series representation is impractical. Instead, one can use the method of [Fox and Glynn, 1989] that replaces the infinite horizon with a random finite horizon. The following preliminary (known) result is necessary:

Lemma 3.5.1 Suppose $c_{\alpha,\infty} < \infty$. Let L be a geometric random variable with $P(L = n) = \alpha^n(1-\alpha)$, independent of the demands and of the random yield. Define $\tilde{C}_{\alpha,L} = \sum_{n=1}^{L} C_n$. Then $\mathbf{E}[\tilde{C}_{\alpha,L}] = c_{\alpha,\infty}$.

Proof: See Appendix A.

Theorem 3.5.2 In addition to conditions of Theorem 3.4.6 if demands satisfy (3.49), then $c'_{\alpha,\infty} = \mathbf{E}[\tilde{C}'_{\alpha,L}]$, where L is a geometric random variable with $P(L = n) = \alpha^n(1-\alpha)$ and

$$\tilde{C}'_{\alpha,L} = \sum_{n=1}^{L} C'_n, \tag{3.50}$$

with C'_n as given by (3.23).

Proof: Lemma 3.5.1 provides an estimator of the infinite horizon discounted cost from a finite number of transitions. The same idea leads to an unbiased estimator of $c'_{\alpha,\infty}$ from L transitions. With probability one, $\tilde{C}_{\alpha,L}$ is differentiable at any z^* and

$$\tilde{C}'_{\alpha,L} = \sum_{n=1}^{L} C'_n, \tag{3.51}$$

with C'_n as given in (3.23). Moreover, $\tilde{C}'_{\alpha,L}$ is Lipschitz and

$$LK_I\Big(\sum_{p=1}^{P} b^p + \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{p=1}^{P} h^{kmp}\Big) \tag{3.52}$$

is an integrable modulus. Combining these observations with Lemma 3.4.3 the result follows.

□

The analysis of infinite horizon average costs relies on the notion of a *Harris recurrent* Markov chain; see [Assmussen, 1987] and [Nummelin, 1984] for an extensive coverage of key definitions

and results. The treatment and background of [Thorisson, 1983] and [Sigman, 1988] is particularly relevant to this application, namely in what concerns the connection with coupling arguments. A Markov chain that is *positive* Harris recurrent has a unique stationary distribution. It will be shown, in Chapter 4, conditions under which a process $\{X_n, n \geq 1\}$ is positive Harris recurrent. See the (natural) expressions below for the PS and the TS cases respectively.

$$\mathbf{E}[\sum_{p=1}^{P} D_0^p] < \min_m \{C^{km}\} \tag{3.53}$$

$$K\mathbf{E}[\sum_{p=1}^{P} D_0^p] < \min_m \{C^m\} \tag{3.54}$$

In Section 3.4 it was shown that the derivative recursions are bounded for finite horizon. If a system is stable, regeneration occurs in finite time with probability one. Therefore, in an infinite horizon setting the derivative recursions are also bounded.

For PS and TS modes it is necessary first to establish the following. Let Y_n^{kmp} be the (appropriate) shortfall, and W_n^{kmp} be the derivative of shortfall Y_n^{kmp} with respect to the base stock levels. Let $\boldsymbol{Y}_n$ and $\boldsymbol{W}_n$ be the appropriate vectors for period n. Let $\boldsymbol{Y}_n^{km} = [Y_n^{km1}\ Y_n^{km2}\ \ldots\ Y_n^{kmP}]$ and $||\boldsymbol{Y}_n^{km}|| = \sum_{p=1}^{P} Y_n^{kmp}$.

Lemma 3.5.3 If $\{\boldsymbol{Y}_n, n \geq 1\}$ is positive Harris recurrent, then so is $\{(\boldsymbol{Y}_n, \boldsymbol{W}_n), n \geq 1\}$.

Proof: The derivative process admits coupling when the process is stable (Chapter 4). Also, the shortfall process admits coupling under the appropriate stability condition for the PS mode. Note that when the shortfall of level K and stage M hits zero (all values Y_n^{KMp} are zero for this n), at some finite time N_{KM}, the derivatives of $Y_{N_{KM}}^{KMp}$ are also zero for all $p = 1, \ldots, P$. Therefore, for any $n \geq N_{KM}$ not only the process coincides with a copy started at zero, but all the derivatives of that process coincide with a copy of the derivatives of a process started at zero. Suppose now that for all $n \geq N_{km}$, $(||\boldsymbol{Y}_n^{km}||, \ldots, ||\boldsymbol{Y}_n^{KM}||)$ together with the derivative processes coincide with the corresponding components started at zero. After, when $||\boldsymbol{Y}_n^{(km)^-}||$ couples so it will be the case for its corresponding derivative processes, since $||\boldsymbol{Y}_n^{(km)^-}||$ only depends explicitly of $\boldsymbol{Y}_n^{kmp}$. By induction, it is possible to conclude that the coupling of the whole shortfall process will imply the coupling of the derivative processes.

The result for the TS mode is trivially derived from the above due to the stochastic dominance argument used in Chapter 4. Since the derivative processes for the shortfall process admit coupling it is the case that the pair is positive Harris recurrent.

□

By imposing (3.53) for the PS case or (3.54) for the TS case, stability for the derivatives is ensured along with stability of the state variables. Following the arguments similar to those in [Glasserman and Tayur, 1995], it follows that

Theorem 3.5.4 Suppose $\{D_n^p, n = 1, 2, \ldots\}$ are i.i.d. for each p with finite expectation and that the adequate stability condition holds for the PS mode. Then $N^{-1}\sum_{n=1}^{N} C'_n \to c'_\infty$, with probability one, at almost every z^*.

If, in addition, $\sup_x f^p(x) < \infty$, then $\tilde{\bar{V}}'_N \to \bar{v}'_\infty$ and with probability one, at almost every z^*.

Proof: See Appendix A.

3.6 Conclusions

This chapter presented a model for multi-product re-entrant flow shops subject to random demand, using a discrete time, capacitated, production-inventory framework. As a first attempt at dealing with complex re-entrant production systems it was assumed a cyclic re-entrant structure with all products following the same path from entry point to exit. Moreover, it was assumed that each product unit at any given level of production imposes the same load on the machines visited (*uniform load*) and that yield is perfect. Assuming a multi-echelon base stock policy as the backbone to decide on production levels, the IPA derivatives for several capacity allocation and production rules were developed. These derivatives can be used to minimize operation costs. The discussion on stability, due to its particularity and to avoid excessive cluttering, is postponed to Chapter 4. A numerical study to gain insights to help manage multi-product re-entrant lines is presented in Chapter 5.

Later, in Part III, the issue of stability will again be discussed for systems with random yield and non uniform loads, in an attempt at moving to more complex systems. Both random yield and

non uniform loads make the analysis more difficult and require the use of broader classes of policies than the one presented here.

3.6.1 Optimizing the Capacity Slots

Before closing this chapter, it remains to consider taking the derivatives with respect to all C^{km} on a PS model. Let $C^* = C^{k^*m^*}$ denote the generic variable with respect to which the derivatives are taken for some $k^* = 1, \ldots, K; m^* = 1, \ldots, M$. Note, that these derivatives are taken as if any C^{km} is a non constrained variable.

The derivative recursions of the production decisions for the Linear Scaling Rule with Partial Sharing are given by

$$P'_{(c)_n}{}^{kmp} = f'_{(c)_n}{}^{kmp} g_n^{km} + f_n^{kmp} g'_{(c)_n}{}^{km}. \tag{3.55}$$

The derivatives of f_n^{kmp} and g_n^{km} with respect to the capacity allocation are given by:

$$f'_{(c)_n}{}^{kmp} = \begin{cases} \left\{ \begin{array}{ll} -E'_{(c)_n}{}^{KMp} & \text{bound by demand} \\ 0 & \text{if } f_n^{KMp} = 0 \end{array} \right\} & m = M \text{ and } k = K \\ \left\{ \begin{array}{ll} -E'_{(c)_n}{}^{kmp} & \text{bound by demand} \\ I'_{(c)_n}{}^{(km)^+p} & \text{bound by supply} \\ 0 & \text{if } f_n^{kmp} = 0 \end{array} \right\} & \text{otherwise} \end{cases} \tag{3.56}$$

$$g'_{(c)_n}{}^{km} = \begin{cases} \frac{1\{C^*=C^{km}\}\sum_p f_n^{kmp} - C^{km}\sum_p f'_{(c)_n}{}^{kmp}}{(\sum_p f_n^{kmp})^2} & \text{bound by capacity} \\ 0 & \text{no bound in capacity} \end{cases} \tag{3.57}$$

For the Priority Rule with Partial Sharing,

$$\begin{aligned} P'_{(c)_n}{}^{kmp(1)} &= \begin{cases} 1\{C^* = C^{km}\} & \text{bound by capacity} \\ f'_{(c)_n}{}^{kmp(1)} & \text{otherwise} \end{cases} \\ P'_{(c)_n}{}^{kmp(i)} &= \begin{cases} 1\{C^* = C^{km}\} - \sum_{j=1}^{p(i)-1} P'_{(c)_n}{}^{kmp(j)} & \text{bound by capacity} \\ f'_{(c)_n}{}^{kmp(i)} & \text{otherwise} \end{cases} \end{aligned} \tag{3.58}$$

Finally, for the Equalize Shortfall Rule with Partial Sharing note that on Section 3.4, when the Equalize Shortfall Procedure is described, no explicit assumption is made as to which variable the derivatives are taken. Therefore the derivatives w.r.t. capacity for this rule are the same as were then.

Regarding the initial conditions derivatives, all variables have zero initial conditions for the derivatives with respect to the capacity allocation, except for ${C'_{(c)}}^{km} = 1\{C^* = C^{km}\}$ (see Equalize Shortfall Procedure).

It is relatively trivial to extend the validation presented earlier for base stock variables to this set of variables. For the sake of brevity, it will be skipped.

Chapter 4

Stability

To validate the infinite horizon measures and derivatives in Chapter 3, the stability conditions for the systems considered have to be established rigorously. While validating those measures, this stability discussion was postponed to this chapter. The objective of this present chapter is to address exclusively the topic of stability for re-entrant flow lines. The discussion of stability will cover systems with uniform loads, perfect yield, and deterministic capacities.

There are two settings for which to investigate the stability conditions. The first refers to the case where capacity is partially shared (PS). The second refers to the case where the capacity is totally shared (TS). The no sharing (NS) case has been established in [Glasserman and Tayur, 1994] for a single product.

To study the conditions under which the system is stable it is necessary to resort to the technique used in [Glasserman and Tayur, 1994]. There, the stability conditions were established for a single product system and no re-entrant structure. Under adequate changes, the same technique will be used to prove stability for the PS case. Then, using a stochastic dominance argument, the conditions under which the TS case is stable will be established.

At first it will be assumed that the production decisions are taken with the use of the Linear Scaling Rule. At the end, it will be shown that the results discussed here remain valid for any of the other two production rules.

First, le us review the model discussed in Chapter 3 using shortfall variables to replace inventory variables. Recalling (3.3), when yield is perfect, the shortfall variables satisfy the following dynamic equation

$$Y_{n+1}^{kmp} = Y_n^{kmp} + d_n^p - P_n^{kmp} \text{ for all } k, m, p. \tag{4.1}$$

The order of presentation will be:

1. Establishing conditions for the stability of the shortfall echelon process when demands are stationary and ergodic.

2. Examining the regenerative structure of $\{Y_n, n \geq 0\}$ when $\{D_n, n \geq 0\}$ is an i.i.d. sequence. The regenerative properties are valuable in establishing convergence of costs and also simulation estimators.

It will be shown that the stability condition suffices to ensure that $\{Y_n, n \geq 0\}$ possesses the regenerative structure of a *Harris ergodic* Markov chain. Under a stronger condition, it will be established that the vector of shortfalls returns to the origin infinitely often, with probability one.

Many of the attractive properties of classical regenerative processes have been shown to hold for the somewhat weaker regenerative structure of Harris recurrent Markov chains. An extensive coverage of key definitions and results of this framework can be found in [Assmussen, 1987] and [Nummelin, 1984]; the treatment in [Sigman, 1988] is particularly relevant to this application.

Harris Recurrence and Explicit Regeneration Points

The general setting for Harris recurrence is a Markov chain $X = \{X_n, n \geq 0\}$ on a state space $\mathbf{S}$ with Borel sets $\mathcal{B}$. Let P_x denote the law of X when $X_0 = x$. Then X is Harris recurrent if there exists a σ-finite measure ψ on $(\mathbf{S}, \mathcal{B})$, not identically zero, such that, for all $A \in \mathcal{B}$,

$$\psi(A) > 0 \Rightarrow P_x \left(\sum_{n=0}^{\infty} \mathbf{1}\{X_n \in A\} = \infty \right) = 1, \quad \text{for all } x \in \mathbf{S}. \tag{4.2}$$

Thus, every set of positive ψ-measure is visited infinitely often from all initial states. Every Harris recurrent Markov chain has an invariant measure π that is unique up to a multiplication by a constant. The sets of positive π-measure are precisely those that are visited infinitely often from all initial states. If π is finite (hence a probability, without loss of generality), then X is called *positive* Harris recurrent. If, in addition, X is aperiodic, then it is Harris ergodic.

The connection with regeneration enters as follows. If X is Harris recurrent, then there exists a (discrete-time) renewal process $\{\tau_k, k \geq 1\}$ and an integer $r \geq 1$ such that

$$\{(X_{\tau_k+n}, n \geq 0), (\tau_{n+k+1} - \tau_{n+k}, n \geq 0)\} \tag{4.3}$$

has the same distribution for all $k \geq 1$ and is independent of

$$\{\tau_1, \ldots, \tau_k, (X_n, 0 \leq n \leq \tau_k - r)\}. \tag{4.4}$$

When $r > 1$, there may be dependence between consecutive cycles $\{X_n, \tau_{k-1} \leq n \leq \tau_k\}$, in contrast to the classical case of independent cycles (and this is indeed the case in this model). However, if X is positive Harris recurrent and if $f : \mathbf{S} \to \mathbf{R}$ is π-integrable, then the regenerative ratio formula

$$\mathbf{E}_\pi[f(X_0)] = \frac{\mathbf{E}[\sum_{n=\tau_{k-1}}^{\tau_k - 1} f(X_n)]}{\mathbf{E}[\tau_k - \tau_{k-1}]} \tag{4.5}$$

remains valid, as does the associated central limit theorem (under second-moment assumptions). Moreover, if X is Harris ergodic then for all initial conditions the distribution of X_n converges to π in *total variation*; that is,

$$\sup_{A \in \mathcal{B}} |P_x(X_n \in A) - \pi(A)| \to 0 \tag{4.6}$$

as $n \to \infty$, for all $x \in \mathbf{S}$. Indeed, this total variation convergence to a probability measure completely characterizes Harris ergodicity.

A powerful tool in the analysis of Harris ergodic Markov chains is a connection with coupling; see for example [Thorisson, 1983] and [Sigman, 1988] for background. The main result is this: a Markov chain with an invariant probability measure admits coupling if and only if it is Harris ergodic. I will be using a coupling argument for Y while establishing the stability conditions, which will by itself render the Harris ergodicity of the shortfall process.

While Harris recurrence ensures the existence of (wide-sense) regeneration times $\{\tau_k, k \geq 1\}$, it does not provide a means of identifying these times. Explicit regeneration times are not needed for convergence results, but they are useful in, for example, computing confidence intervals from

simulation estimators. At the end of each section I will give a sufficient condition for $\{\boldsymbol{Y}_n, n \geq 0\}$ to have readily identifiable regeneration times.

4.1 Stability and Regeneration for Partially Shared Systems

For this setting I will first define a dynamic equation for a linear combination of shortfalls such that the theoretical framework of [Glasserman and Tayur, 1994] is readily applicable. Although the proofs are not different from those, they will be presented for the sake of completeness of this thesis. Recall that for the PS case under the LSR the production decision is defined as

$$P_n^{kmp} = f_n^{kmp} g_n^{km} \tag{4.7}$$

where f_n^{kmp} denotes the net production vector if no capacity constraint is present and g_n^{km} enforces the capacity constraint as described earlier.

As discussed above the dynamic equation for the shortfall quantities is given by (4.1). To simplify the analysis we can define a vectorial dynamic equation for each stage and level by defining $\boldsymbol{D}_n = [d_n^1\ d_n^2\ \ldots\ d_n^P]^T$, $\boldsymbol{P}_n^{km} = [P_n^{km1}\ P_n^{km2}\ \ldots\ P_n^{kmP}]^T$, and $\boldsymbol{Y}_n^{km} = [Y_n^{km1}\ Y_n^{km2}\ \ldots\ Y_n^{kmP}]^T$. Therefore, the dynamic equation will assume the form

$$\boldsymbol{Y}_{n+1}^{km} = \boldsymbol{Y}_n^{km} + \boldsymbol{D}_n - \boldsymbol{P}_n^{km} \text{ for all } k, m. \tag{4.8}$$

Let $||\boldsymbol{x}||$, be defined as the sum of all components of $\boldsymbol{x}$. Note that $||\boldsymbol{x}||$ is not a norm and it verifies the following

$$\begin{aligned} ||\boldsymbol{x} + \boldsymbol{y}|| &= ||\boldsymbol{x}|| + ||\boldsymbol{y}|| \\ ||a\boldsymbol{x}|| &= a||\boldsymbol{x}|| \end{aligned} \tag{4.9}$$

Now, since level and stage (K, M) draws raw material from an infinite source, we have

$$\boldsymbol{Y}_{n+1}^{KM} = \max\{\vec{0}, (\boldsymbol{Y}_n^{KM} + \boldsymbol{D}_n)(1 - \frac{C^{KM}}{||\boldsymbol{Y}_n^{KM} + \boldsymbol{D}_n||})\} \tag{4.10}$$

Due to the structure of the above the following operation is valid

$$||\boldsymbol{Y}_{n+1}^{KM}|| \quad = \quad \max\{0, ||\boldsymbol{Y}_{n}^{KM}|| + ||\boldsymbol{D}_n|| - C^{KM}\} \tag{4.11}$$

which is a Lindley equation. Note the use of (4.9).

For the remaining cases we will have

$$\begin{aligned} \boldsymbol{Y}_{n+1}^{km} \quad = \quad & \max\{\boldsymbol{Y}_{n}^{km} + \boldsymbol{D}_n - \frac{C^{km} \min\{\boldsymbol{Y}_{n}^{km} + \boldsymbol{D}_n, \boldsymbol{I}_{n}^{(km)^+}\}}{||\min\{\boldsymbol{Y}_{n}^{km} + \boldsymbol{D}_n, \boldsymbol{I}_{n}^{(km)^+}\}||}, \\ & \vec{0}, \boldsymbol{Y}_{n}^{km} + \boldsymbol{D}_n - \boldsymbol{I}_{n}^{(km)^+}\} \end{aligned} \tag{4.12}$$

From this equation it is possible to compute $||\boldsymbol{Y}_{n+1}^{km}||$ as follows

$$\begin{aligned} ||\boldsymbol{Y}_{n+1}^{km}|| \quad = \quad & \max\Big\{0, ||\boldsymbol{Y}_{n}^{km}|| + ||\boldsymbol{D}_n|| - C^{km}, \\ & \sum_{p=1}^{P} \left(Y_{n}^{(km)^+ p} + d_{n}^{p} - (z^{(km)^+ p} - z^{kmp})\right)^+ \Big\}. \end{aligned} \tag{4.13}$$

The scalar equations (4.11) and (4.13) are the multiple product generalizations of the dynamic equations for shortfalls presented in [Glasserman and Tayur, 1994] for single product systems.

4.1.1 The Stationary Regime

Let us now introduce the framework and notation corresponding to Lemmas 1 and 2 of [Glasserman and Tayur, 1994] which help establishing the stability conditions.

Lemma 4.1.1 The echelon shortfalls satisfy $\boldsymbol{Y}_{n+1} = \Phi(\boldsymbol{Y}_n, \boldsymbol{D}_n)$ where $\Phi : R_+^{KMP} \times R^P \to R_+^{KMP}$ is defined by (4.10, 4.12). Also, the total shortfall satisfy $||\boldsymbol{Y}_{n+1}|| = \phi(\boldsymbol{Y}_n, \boldsymbol{D}_n) = ||\Phi(\boldsymbol{Y}_n, \boldsymbol{D}_n)||$ where $\phi : R_+^{KMP} \times R^P \to R_+^{KM}$ is defined by (4.11, 4.13). In particular, ϕ is increasing and continuous.

Supposing that the demands form a stationary process, without loss of generality, we can assume that $\boldsymbol{D}_n$ is defined for all integer n with $\{\boldsymbol{D}_n, -\infty < n < \infty\}$ stationary. In what follows I will use $\Rightarrow$ to denote convergence in distribution. Some of the proofs will be omited here to avoid excessive clutter. Some of them are relatively trivial extensions of similar results published. Some others are exactly the same. Some of the former will be presented in Appendix B for the sake of completeness of the present document.

Lemma 4.1.2 Let $\{\boldsymbol{D}_n, -\infty < n < \infty\}$ be stationary. There exists a (possibly infinite) stationary process $\{\tilde{\boldsymbol{Y}}_n, -\infty < n < \infty\}$ satisfying $||\tilde{\boldsymbol{Y}}_{n+1}|| = \phi(\tilde{\boldsymbol{Y}}_n, \boldsymbol{D}_n)$ for all n, such that if $||\boldsymbol{Y}_0|| = 0$, a.s., then $||\boldsymbol{Y}_n|| \Rightarrow ||\tilde{\boldsymbol{Y}}_0||$.

Proof: See Appendix B.

With the support of the above two Lemmas it is now easy to establish the stability condition for this model.

Theorem 4.1.3 Suppose the demands $\{\boldsymbol{D}_n, -\infty < n < \infty\}$ are ergodic as well as stationary. If

$$\mathbf{E}[||\boldsymbol{D}_0||] = \sum_{p=1}^{P} \mathbf{E}[d_0^p] < \min\{C^{km} : k = 1, \ldots, K; m = 1, \ldots, M\}, \tag{4.14}$$

then $||\tilde{\boldsymbol{Y}}_0||$ is almost surely finite. If for some (k, m), $\mathbf{E}[||\boldsymbol{D}_0||] > C^{km}$, then $||\tilde{\boldsymbol{Y}}_0^{qr}|| = \infty$, a.s., for all (q, r) corresponding to levels and stages coming after (k, m).

Proof: See Appendix B.

This result for the scalar dynamic equations implies the stability of the vectorial process.

Corollary 4.1.4 Under the assumptions of Theorem 4.1.3 $\tilde{Y}_0^{kmp}$ is almost surely finite for all p, where $\tilde{Y}_0^{kmp}$ denotes component p of $\tilde{\boldsymbol{Y}}_0^{km}$.

Proof: The result follows trivially due to the non negativity of the shortfalls.

□

The above results show that the process $\{\boldsymbol{Y}_n, n \geq 0\}$ converges to a stationary distribution only if $\boldsymbol{Y}_0 = 0$. The following theorem establishes that the convergence occurs for any initial point, that is, the process admits coupling.

Theorem 4.1.5 Under the stability condition $\mathbf{E}[||\boldsymbol{D}_0||] < \min_{k,m}\{C^{km}\}$, the echelon shortfall process admits coupling. Consequently, its stationary distribution is unique, and $\boldsymbol{Y}_n \Rightarrow \tilde{\boldsymbol{Y}}_0$ for all $\boldsymbol{Y}_0$.

Proof: See Appendix B.

4.1.2 Regeneration and Explicit Regeneration Times

Recall that a Markov chain with an invariant probability measure admits coupling if and only if it is Harris ergodic. In the previous subsection we used a coupling argument for $\boldsymbol{Y}$, therefore it is now easy to show that,

Theorem 4.1.6 Let demands $\{\boldsymbol{D}_n, n \geq 0\}$ be i.i.d. with $\mathbf{E}[||\boldsymbol{D}_0||] < \min_{k,m}\{C^{km}\}$. Then $\{\boldsymbol{Y}_n, n \geq 0\}$ is a Harris ergodic Markov chain.

Proof: Since $\boldsymbol{Y}_{n+1} = \Phi(\boldsymbol{Y}_n, \boldsymbol{D}_n)$, $n \geq 0$, $\boldsymbol{Y}$ is a Markov chain when $\boldsymbol{D}$ is i.i.d. We established in Theorem 4.1.3 and Corollary 4.1.4 that $\boldsymbol{Y}$ has an invariant (i.e., stationary) distribution and in Theorem 4.1.5 that $\boldsymbol{Y}$ admits coupling. Thus, $\boldsymbol{Y}$ is Harris ergodic.

□

As a result of Theorem 4.1.6, $\boldsymbol{Y}$ inherits the regenerative structure of Harris ergodic Markov chains, the attendant ratio formula, and convergence results. The same holds for the inventory levels:

Corollary 4.1.7 The inventory process $\{(\boldsymbol{I}_n^{11}, \ldots, \boldsymbol{I}_n^{KM}), n \geq 0\}$, under the conditions of Theorem 4.1.6, is a Harris ergodic Markov chain.

Proof: There is a one-to-one correspondence between shortfalls and inventories for all n as defined by

$$\begin{aligned} I_n^{11p} &= z^{11p} - Y_n^{11p} \\ I_n^{kmp} &= (z^{kmp} - z^{(km)^- p}) + (Y_n^{(km)^- p} - Y_n^{kmp}). \end{aligned} \tag{4.15}$$

Consequently, $\boldsymbol{I} = \{\boldsymbol{I}_n, n \geq 0\}$ is Markov if $\boldsymbol{Y}$ is, and $\boldsymbol{I}$ is Harris ergodic if $\boldsymbol{Y}$ is.

□

It is now possible to give the characterization of the regeneration times.

Theorem 4.1.8 Let demands be i.i.d. with $\mathbf{E}[||\boldsymbol{D}_0||] < \min_{k,m}\{C^{km}\}$. Define $\boldsymbol{z}^{(11)^-} \equiv \vec{0}$ and suppose that

$$P(d_0^p \leq z^{kmp} - z^{(km)^- p}) > 0, \;\; k = 1, \ldots, K; \; m = 1, \ldots, M; \; p = 1, \ldots, P. \tag{4.16}$$

Then $\boldsymbol{Y}$ returns to the origin infinitely often, with probability one.

Proof: See Appendix B.

Corollary 4.1.9 The inventory process $\{(\boldsymbol{I}_n^{11}, \ldots, \boldsymbol{I}_n^{KM}), n \geq 0\}$, under the conditions of Theorem 4.1.8, returns to $(\boldsymbol{z}^{11}, \boldsymbol{z}^{(11)^+} - \boldsymbol{z}^{11}, \ldots, \boldsymbol{z}^{KM} - \boldsymbol{z}^{(KM)^-})$ infinitely often, with probability one.

Proof: Consequence of the relationship between shortfalls and inventories.

□

The conclusion of Theorem 4.1.8 is not in general true without (4.16) or further distributional assumptions on demands. This is particularly clear when $z^{(km)^+ p} = z^{kmp}$ for some value of k, m, and p; that is, stage $(km)^+$ keeps no safety stock for product p. In this case, the total shortfall $||\boldsymbol{Y}^{km}||$ can never reach zero unless $d_0^p = 0$ with positive probability.

4.2 Stability and Regeneration for Totally Shared Systems

Operating on a TS mode with the LSR, the production decision is given by

$$P_n^{kmp} = f_n^{kmp} g_n^m, \tag{4.17}$$

capacity is shared among all products and levels for each machine, and g is given by (3.33).

For this model the shortfall process is governed by the following

$$\begin{aligned} \boldsymbol{Y}_{n+1}^{km} &= \max\{\boldsymbol{Y}_n^{km} + \boldsymbol{D}_n - \frac{C^m \min\{\boldsymbol{Y}_n^{km} + \boldsymbol{D}_n, \boldsymbol{I}_n^{(km)+}\}}{\sum_{k=1}^K ||\min\{\boldsymbol{Y}_n^{km} + \boldsymbol{D}_n, \boldsymbol{I}_n^{(km)^+}\}||}, \\ &\quad \vec{0}, \boldsymbol{Y}_n^{km} + \boldsymbol{D}_n - \boldsymbol{I}_n^{(km)^+}\} \end{aligned} \tag{4.18}$$

Due to the structure of the decision process, it is the case that

$$\begin{aligned} ||\boldsymbol{Y}_{n+1}^{km}|| &= \max\{0, ||\boldsymbol{Y}_n^{km}|| + ||\boldsymbol{D}_n|| - \frac{C^m||\min\{\boldsymbol{Y}_n^{km} + \boldsymbol{D}_n, \boldsymbol{I}_n^{(km)^+}\}||}{\sum_{k=1}^K ||\min\{\boldsymbol{Y}_n^{km} + \boldsymbol{D}_n, \boldsymbol{I}_n^{(km)^+}\}||}, \\ &\quad \sum_{p=1}^P \left(Y_n^{(km)^+p} + d_n^p - (z^{(km)^+p} - z^{kmp})\right)^+\}, \end{aligned} \tag{4.19}$$

where it is assumed that $\boldsymbol{I}_n^{(KM)+} = \infty$ for all n.

As in the previous model we were interested on the total shortfall per level and stage, we will now be interested in the *Total Shortfall per Stage*. Thus, the following expression is of importance in what follows

$$\begin{aligned} \sum_{k=1}^K ||\boldsymbol{Y}_{n+1}^{km}|| &= \max\{0, \sum_{k=1}^K ||\boldsymbol{Y}_n^{km}|| + K||\boldsymbol{D}_n|| - C^m, \\ &\quad \sum_{k=1}^K \sum_{p=1}^P \left(Y_n^{(km)^+p} + d_n^p - (z^{(km)^+p} - z^{kmp})\right)^+\} \end{aligned} \tag{4.20}$$

4.2.1 The Stationary Regime

Lemma 4.2.1 The total echelon shortfall per stage satisfies

$$\sum_{k=1}^K ||\boldsymbol{Y}_{n+1}^{km}|| = \psi(\boldsymbol{Y}_n, \boldsymbol{D}_n) = \sum_{k=1}^K \phi(\boldsymbol{Y}_n, \boldsymbol{D}_n), \tag{4.21}$$

where $\phi : R_+^{KMP} \times R^P \to R_+^{KM}$ is defined by (4.19) and $\psi : R_+^{KMP} \times R^P \to R_+^M$ is defined by (4.20). In particular, ψ is continuous and increasing.

It is easily possible to establish a result similar to that of Lemma 4.1.2 for this second model and to prove stability we will make use of Theorem 4.1.3.

Theorem 4.2.2 Under the assumptions of Theorem 4.1.3, the system operated under the TS mode is stable, in the sense that the shortfalls are almost surely finite, if

$$K\mathbf{E}[||\boldsymbol{D}_0||] < \min\{C^m : m = 1, \ldots, M\}. \tag{4.22}$$

Proof: Assume that the capacity of each machine is divided into slots of equal size, that is $C^{km} = \frac{C^m}{K}$. Assume the system operates as if capacity was not shared. Then, according to Theorem 4.1.3, the system would be stable iff

$$\begin{aligned} \mathbf{E}[||\boldsymbol{D}_0||] &< \min\{C^{km} : k = 1, \ldots, K, m = 1, \ldots, M\} \\ &= \frac{1}{K} \min\{C^m : m = 1, \ldots, M\} \end{aligned}$$

Now we have to evaluate how does the system behave under the TS case when compared with its performance under the PS case. To show that stability of the PS case implies stability of the TS case I will investigate a sample path.

Assume we have two identical systems subject to the same sample path. One is operated under the PS mode with $C^{km} = \frac{C^m}{K}$ and the other is operated under the TS mode. In particular, one is interested in the process defined by the total shortfall for each stage. Assume that both systems start from the origin, that is

$$\sum_{k=1}^{K} ||\boldsymbol{Y}_0^{km}||^1 = \sum_{k=1}^{K} ||\boldsymbol{Y}_0^{km}||^2 = 0 \text{ for all } k, m. \tag{4.23}$$

Comparing equation (4.19) with (4.11) and (4.13) it is obvious that while there is no bound in capacity for any of the two systems they remain coupled. Let n^* denote the first period for which at least one of the two systems has a bound in capacity for some level and/or stage. Therefore, we have for all k and m

$$||\boldsymbol{Y}_n^{km}||^1 = ||\boldsymbol{Y}_n^{km}||^2 \text{ for all } n = 0, \ldots, n^*, \tag{4.24}$$

which implies that

$$\sum_{k=1}^{K} ||\boldsymbol{Y}_n^{km}||^1 = \sum_{k=1}^{K} ||\boldsymbol{Y}_n^{km}||^2 \text{ for all } n = 0, \ldots, n^*. \tag{4.25}$$

The first time one of these two systems has at least one production decision bounded by capacity there is a possibility for decoupling. Let us take system 1 as the reference. Whenever there is at least a level and stage for which system 1 is bound by capacity, one of two things can happen to system 2:

i) Bound in capacity for system 1 and no bound for system 2.

For this case there exists at least a k^* and an m^* such that

$$||\min\{\boldsymbol{Y}_{n^*}^{k^*m^*} + \boldsymbol{D}_{n^*}, \boldsymbol{I}_{n^*}^{(k^*m^*)^+}\}|| > \frac{C^{m^*}}{K}, \tag{4.26}$$

but

$$\sum_{k=1}^{K} ||\min\{Y_{n^*}^{km^*} + D_{n^*}, I_{n^*}^{(km^*)^+}\}|| < C^{m^*}. \tag{4.27}$$

ii) Bound in capacity for both systems

In this case we have at least a k^* and an m^* such that

$$||\min\{\boldsymbol{Y}_{n^*}^{k^*m^*} + \boldsymbol{D}_{n^*}, \boldsymbol{I}_{n^*}^{(k^*m^*)^+}\}|| > \frac{C^{m^*}}{K}, \tag{4.28}$$

and

$$\sum_{k=1}^{K} ||\min\{\boldsymbol{Y}_{n^*}^{km^*} + \boldsymbol{D}_{n^*}, \boldsymbol{I}_{n^*}^{(km^*)^+}\}|| > C^{m^*}. \tag{4.29}$$

I am interested on knowing how does $\sum_{k=1}^{K} ||\boldsymbol{Y}_n^{km}||$ (the total shortfall for stage m) behave for both cases. In case i) since system 2 has no bound in capacity it must be the case that not all the levels of stage m^* have a bound in capacity for system 1. Therefore,

$$\begin{aligned}\sum_{k=1}^{K}||\boldsymbol{Y}_{n^*+1}^{km^*}||^1 &= \sum_{k\neq k^*}\sum_{p=1}^{P}\left(Y_{n^*}^{(km^*)^+p}+d_{n^*}^{p}-(z^{(km^*)^+p}-z^{km^*p})\right)^+ + \\ &\quad + \sum_{k=k^*}(||\boldsymbol{Y}_{n^*}^{km^*}||^1+||\boldsymbol{D}_{n^*}||-\frac{C^{m^*}}{K}) \\ &> \sum_{k=1}^{K}\sum_{p=1}^{P}\left(Y_{n^*}^{(km^*)^+p}+d_{n^*}^{p}-(z^{(km^*)^+p}-z^{km^*p})\right)^+ \\ &= \sum_{k=1}^{K}||\boldsymbol{Y}_{n^*+1}^{km^*}||^2 \end{aligned} \tag{4.30}$$

because there is a bound in capacity for all levels k^* in system 1 and there is no such bound in system 2 and using equations (4.13) and (4.19)[1].

For case ii), when both systems are capacity bounded for some stage m^*, there are two possibilities: there is a bound in capacity for all levels of stage m^* in system 1; not all levels of stage m^* are capacity bounded for system 1.

For the first situation it will be the case that

$$\begin{aligned}\sum_{k=1}^{K}||\boldsymbol{Y}_{n^*+1}^{km^*}||^1 &= \sum_{k=1}^{K}(||\boldsymbol{Y}_{n^*}^{km^*}||^1+||\boldsymbol{D}_{n^*}||-\frac{C^{m^*}}{K}) \\ &= \sum_{k=1}^{K}||\boldsymbol{Y}_{n^*+1}^{km^*}||^2, \end{aligned} \tag{4.31}$$

because $\sum_{k=1}^{K}||\boldsymbol{Y}_{n^*}^{km*}||^1=\sum_{k=1}^{K}||\boldsymbol{Y}_{n^*}^{km*}||^2$.

In the second situation we will have

$$\begin{aligned}\sum_{k=1}^{K}||\boldsymbol{Y}_{n^*+1}^{km^*}||^1 &= \sum_{k\neq k^*}\sum_{p=1}^{P}\left(Y_{n^*}^{(km^*)^+p}+d_{n^*}^{p}-(z^{(km^*)^+p}-z^{km^*p})\right)^+ + \\ &\quad + \sum_{k=k^*}(||\boldsymbol{Y}_{n^*}^{km^*}||^1+||\boldsymbol{D}_{n^*}||-\frac{C^{m^*}}{K}) \\ &> \sum_{k=1}^{K}||\boldsymbol{Y}_{n^*}^{km^*}||^1+K||\boldsymbol{D}_{n^*}||-C^{m^*}\end{aligned}$$

[1]Note that equation (4.11) can be made equal to (4.13) by defining $I_n^{(KM)^+}=\vec{\infty}$ and making the adequate change for $z^{(KM)^+}$.

$$= \sum_{k=1}^{K} ||\boldsymbol{Y}_{n^*+1}^{km^*}||^2 \tag{4.32}$$

because the change in total shortfall for stage m^* in system 1 is smaller than C^{m^*}.

Thus, we have that for period $n^* + 1$ the total shortfall for each stage of system 2 is bounded above by the total shortfall for each stage of system 1, with probability one.

Now it remains to see what happens after $n^* + 1$ (the first decoupling period). Assume, there is a third system that starts operating in a TS mode as system 2 with the state variables of system 1, that is, coupled to system 1. System 1 and system 3 will remain coupled until a capacity bound occurs at some other period. By the above discussion we know that a bound in capacity is favorable to system 3, when compared with system 1. Due to Lemma 4.2.1 the total shortfall per stage of system 2 will remain dominated by that of system 3. So we have that until the first decoupling between system 1 and system 3, system 1 will dominate system 2, due to transitivity. If we force system 3 to receive the state of system 1 whenever there is a decoupling between the two the process repeats itself whenever there is a new bound in capacity and it follows then that the total shortfall per stage for system 2 will remain dominated by that of system 1, with probability one.

□

In order to establish the uniqueness of the distribution it is also possible to show that the total shortfall per stage process admits coupling.

Theorem 4.2.3 Under the stability condition $K\mathbf{E}[||\boldsymbol{D}_0||] < \min_m\{C^m\}$, the total shortfall per stage admits coupling and so does the shortfall process as a consequence. Therefore, its stationary distribution is unique and $\boldsymbol{Y}_n \Rightarrow \tilde{\boldsymbol{Y}}_0$ for all $\boldsymbol{Y}_0$.

Proof: According to the proof of Theorem 4.2.2 the total shortfall process per stage of the PS case dominates that of the TS case. Therefore, if the first admits coupling, so does the second because the shortfalls are always non negative. By Theorem 4.1.5, it is the case that the first admits coupling.

Thus, the result follows.

□

4.2.2 Regeneration and Explicit Regeneration Times

Since in the previous subsection a coupling argument was used for Y, it is now easy to show the following.

Theorem 4.2.4 Let demands $\{\boldsymbol{D}_n, n \geq 0\}$ be i.i.d. with $K\mathbf{E}[||\boldsymbol{D}_0||] < \min_m\{C^m\}$. Then $\{\boldsymbol{Y}_n, n \geq 0\}$ is a Harris ergodic Markov chain.

Proof: Since $\sum_{k=1}^{K} ||\boldsymbol{Y}_{n+1}^{km}|| = \psi(\boldsymbol{Y}_n, \boldsymbol{D}_n)$, $n \geq 0$, $\boldsymbol{Y}$ is a Markov chain when $\boldsymbol{D}$ is i.i.d. Theorem 4.2.2 established that $\boldsymbol{Y}$ has an invariant (i.e., stationary) distribution and Theorem 4.2.3 established that $\boldsymbol{Y}$ admits coupling. Thus, $\boldsymbol{Y}$ is Harris ergodic.

□

Corollary 4.2.5 The inventory process $\{(\boldsymbol{I}_n^{11}, \ldots, \boldsymbol{I}_n^{KM}), n \geq 0\}$, under the conditions of Theorem 4.1.6, is a Harris ergodic Markov chain.

Proof: There is a one-to-one correspondence between shortfalls and inventories for all n. Consequently, $\boldsymbol{I} = \{\boldsymbol{I}_n, n \geq 0\}$ is Markov if $\boldsymbol{Y}$ is, and $\boldsymbol{I}$ is Harris ergodic if $\boldsymbol{Y}$ is.

□

The regeneration times can now be characterized.

Theorem 4.2.6 Let demands be i.i.d. with $K\mathbf{E}[||\boldsymbol{D}_0||] < \min_m\{C^m\}$. Define $\boldsymbol{z}^{(11)^-} \equiv \vec{0}$ and suppose that

$$P(d_0^p \leq z^{kmp} - z^{(km)^- p}) > 0, \;\; k = 1, \ldots, K; \; m = 1, \ldots, M; \; p = 1, \ldots, P. \tag{4.33}$$

Then $\boldsymbol{Y}$ returns to the origin infinitely often, with probability one.

Proof: The proof follows from the fact that the same system operated under a PS mode with $C^{km} = C^m/K$ will have a shortfall process that dominates that of a system operated on a TS mode. Since for the PS mode Theorem 4.1.8 is applicable it is the case that if $\boldsymbol{Y}$ returns to origin infinitely often under the PS mode so it does for the TS mode due to the dominance earlier discussed.

□

Corollary 4.2.7 The inventory process $\{(\boldsymbol{I}_n^{11},\ldots,\boldsymbol{I}_n^{KM}), n \geq 0\}$, under the conditions of Theorem 4.1.8, returns to $(\boldsymbol{z}^{11}, \boldsymbol{z}^{(11)^+} - \boldsymbol{z}^{11}, \ldots, \boldsymbol{z}^{KM} - \boldsymbol{z}^{(KM)^-})$ infinitely often, with probability one.

Proof: Consequence of the relationship between shortfall variables and inventories.

□

4.3 Total Shortfall Dynamic Equation for PR and ESR

In this section I will show that both rules satisfy dynamic equations similar to (4.10) and (4.12) for the PS mode and (4.20) for the TS mode.

Let us consider the dynamic equation for $||\boldsymbol{Y}_n^{km}||$ for the PS case. In order to compute the dynamic equation for $||\boldsymbol{Y}_n^{km}||$ note that under the Strict Priority Rule if product $p(i^*)$ fills capacity we have that for all $p = p(1),\ldots,p(i^*-1)$, all net needs are filled whereas for all $p = p(i^*+1),\ldots,p(P)$ nothing is produced. For the particular case of $p = p(i^*)$, only the available capacity defined by $C^{km} - \sum_{i=1}^{i^*-1} P_n^{kmp(i)}$ will be used to reduce its shortfall. In any case, as long as one product fills capacity the change in $||\boldsymbol{Y}_n^{km}||$ is positive due to the total demand, $||\boldsymbol{D}_n||$, and negative due to capacity, C^{km}. If no product fills capacity, then the total change in shortfall may be only bounded by available inventory, otherwise the total shortfall drops to zero. Therefore, the total shortfall for each stage and level, under strict priority, follows the dynamic equation given by

$$||\boldsymbol{Y}_{n+1}^{km}|| = \begin{cases} 0 & \text{if no bound in capacity and inventory} \\ \sum_{p=1}^{P}(Y_n^{kmp} + d_n^p - I_n^{(km)^+p})^+ & \text{if no bound in capacity} \\ ||\boldsymbol{Y}_n^{km}|| + ||\boldsymbol{D}_n|| - C^{km} & \text{if bound by capacity} \end{cases} \tag{4.34}$$

which is exactly as (4.12). If $k = K$ and $m = M$ there will never be a bound in inventory, so that the above equation will simplify to (4.10). It should not be difficult to understand that the Equalize Shortfall Rule yields the same dynamic equations for $||\boldsymbol{Y}_n^{km}||$.

It should be evident that both rules yield (4.20) for the dynamic equation of $\sum_k ||Y_n^{km}||$ when the system is operated on a TS mode.

Therefore, all results discussed eralier for the LSR are trivially applicable to both remaining production rules. Although the stability results remain untouched by the different production rules,

it should be clear that the application of each rule impacts on the way each individual shortfall evolves, which has implications only at the attained costs.

Chapter 5

Experimental Study

This chapter continues the previous two by analyzing some experimental results (obtained by simulation based optimization) for a family of re-entrant systems. The analysis will concentrate only on random demand, since otherwise it would be difficult to have a minimally focused experimental study.

A series of computational studies will be presented – for single product (Section 5.2) as well as multi-product (Section 5.3) settings – that provide insights into the properties of the optimal solutions within the class of capacity management, production rules, and inventory control proposed. The experiments will cover the infinite horizon average cost setting.

The chapter concludes, in Section 5.4, with a summary of the main insights obtained and some comments on future research directions, bridging to Part III. A subsection of this one justifies the fact that we restrict the study in Sections 5.2 and 5.3 to a specific priority assignment within the Priority Rule.

The experimental evidence that will be presented here on how to set the priorities for the several products and buffers along the re-entrant line agrees with previously published experimental results, [Glassey and Resende, 1988, Lu and Kumar, 1991]. On multiple product, we will provide a set of data showing that keeping the same list of priorities along the production line has advantages over different priority lists on each production stage. The studies of [Glassey and Resende, 1988, Lu and Kumar, 1991] were concerned with single product, so that such conclusions were not possible to attain.

In Appendix C we detail some key features of the optimization procedure and of the experiments conducted.

5.1 Optimality Condition

Before turning to the analysis of the data obtained with the simulator, there is a structural result that often helps identify optimal solutions.

Proposition 5.1.1 If $\{d_n^p, n = 1, 2, \ldots; p = 1, \ldots, P\}$ are independent and stationary, where each d_n^p is drawn from a density on $(0, \infty)$, the optimal base stock levels for the average cost measure for any production rule and any capacity sharing mode are such that

$$Pr(d_0^p \leq I^{11p}) = \frac{b^p}{b^p + h^{11p}} \quad \text{for all } p = 1, \ldots, P. \tag{5.1}$$

Proof: We prove the result by considering that the optimization is made relative to the delta variables. Let us fix the values of $(\Delta^{KM_p}, \ldots, \Delta^{(11)^+p})$ for all $p = 1, \ldots, P$, and fix the values of Δ^{11p} for all $p \neq \hat{p}$. Consider the optimization made exclusively with respect to $\Delta^{11\hat{p}}$.

Whatever the value of $\Delta^{11\hat{p}}$ the production decisions will only depend on the other delta variables and on demand because the production decisions are solely dependent on shortfalls (see Chap. 3). Therefore, if we disturb $\Delta^{11\hat{p}}$ by a small amount δ, the sample path will only differ from the original undisturbed sample path for the values of $I_n^{11\hat{p}}$. Thus, the cost function only depends on $\Delta^{11\hat{p}}$ through $I_n^{11\hat{p}}$. Therefore, the equation defining the single stage cost derivative only depends on product $\hat{p}$ and will assume the simpler form

$$C'_n = C'^{\hat{p}}_n,$$

because the other components of the single stage cost have zero derivatives. $C'^{\hat{p}}_n$ is given by

$$\begin{aligned} C'^{\hat{p}}_n &= -\mathbf{1}\{I_n^{11\hat{p}} - d_n^{\hat{p}} < 0\}(I')_n^{11\hat{p}} b^{\hat{p}} + \mathbf{1}\{I_n^{11\hat{p}} - d_n^{\hat{p}} > 0\}(I')_n^{11\hat{p}} h^{11\hat{p}} \\ &= -\mathbf{1}\{I_n^{11\hat{p}} - d_n^{\hat{p}} < 0\} b^{\hat{p}} + \mathbf{1}\{I_n^{11\hat{p}} - d_n^{\hat{p}} > 0\} h^{11\hat{p}}, \end{aligned}$$

because $(I')_n^{11\hat{p}} = 1$ for all n when the derivative is taken with respect to $\Delta^{11\hat{p}}$ and zero otherwise.

The average of the derivatives of a single run is obtained by summing for all periods up to N and dividing by N

$$\frac{1}{N}\sum_{n=1}^{N}\left[-\mathbf{1}\{I_n^{11\hat{p}}-d_n^{\hat{p}}<0\}b^{\hat{p}}+\mathbf{1}\{I_n^{11\hat{p}}-d_n^{\hat{p}}\geq 0\}h^{11\hat{p}}\right]. \tag{5.2}$$

Note that $\sum_{n=1}^{N}\mathbf{1}\{I_n^{11\hat{p}}-d_n^{\hat{p}}<0\}$ counts the number of times $I_n^{11\hat{p}}-d_n^{\hat{p}}<0$ during N periods. When divided by N it measures the relative frequency of the event.

Since we can interchange expected value with taking derivatives, it follows that

$$\begin{aligned}\lim_{N\to\infty} c'_N &= \lim_{N\to\infty}\mathbf{E}\left[\frac{1}{N}\sum_{n=1}^{N}C'_n\right]\\ &= -Pr(I^{11\hat{p}}-d^{\hat{p}}<0)b^{\hat{p}}+Pr(I^{11\hat{p}}-d^{\hat{p}}\geq 0)h^{11\hat{p}},\end{aligned} \tag{5.3}$$

given the stationarity of the demand process.

By setting the derivative equal to zero and because $Pr(I_n^{11\hat{p}}-d_n^{\hat{p}}<0)=1-Pr(I_n^{11\hat{p}}-d_n^{\hat{p}}\geq 0)$, it follows that the optimal $\Delta^{11\hat{p}}$ is such that (5.1) holds.

Now, since one can choose arbitrary values for the other delta variables, it is the case that the above holds also for the optimal values of those variables. Naturally, the above reasoning can be repeated for any product.

□

In what follows, we will refer to the above result as the *optimality condition*[1]. There are cases where the optimization algorithm stops short of achieving a set of optimal variables where the optimality condition is satisfied. These are the cases where the cost function is non differentiable as indicated in Chapter 3 (see Section 3.4 therein); we will be discussing such situations in Section 5.2.2 below. This optimality condition establishes a clear equivalence between (average) operational costs and Type-1 service level.

A comprehensive study of re-entrant systems is truly an enormous task to perform, given the number of parameters to be taken into account: average demand, demand variance, holding costs, penalty costs, number of machines, number of levels, capacity of the machines, allocation rules and

[1] It is a necessary condition.

production rules. We limited the study to one and two products and assumed re-entrant structure on a single machine. Even in this simplified setting the range of parameters is very wide. We start by investigating, in a single product setting, the optimal allocation of capacity to the different levels, when operating on the PS mode. After establishing simple rules for the optimal allocation of capacity in PS mode, we study the relative performance of the several production rules. For single product we also study the effect of changing holding costs along the line for different machine loads. One conclusion of this study is that changing both costs and loads, while affecting the absolute value of costs, does not change significantly the relative performance of different modes of operation. It will be possible to see a very subtle relationship between base stock values and capacity.

Next we will move to a two product setting, first confirming that many results from the single product case continue to hold. The few cases where they do not, will be briefly discussed. Then, we investigate the effect of penalty costs, mean demand and demand variance on the relative performances of the capacity allocation and production rules.

5.2 Single Product

For re-entrant systems producing a single product we are interested first on determining how one should allocate capacity to the different levels when not operating on total sharing of capacity. The experimental results of [Glasserman and Tayur, 1995] show that capacity should be nondecreasing along the flow line. The main conclusions here are that for a wide range of holding costs along the line, *the optimal capacity allocation is obtained by giving the same share to each level on a uniform load setting.* This conclusion also carries through for multiple products. Therefore, it is possible to cut out one order of complexity in the problem by always using equal capacity slots on the Partial Sharing mode.

Next we investigate the impact of holding costs and machine load on the relative performance of the production rules. For single product I find that, while affecting the values of the optimal cost and of the optimal control parameters, *the relative performance of the different rules is not affected by the machine load nor by the particular holding cost structure used.*

This conclusions will also carry through for multiple products as will be shown. Therefore, we proceed the experimental study with any choice on these parameters (holding cost pattern and machine load), given that a particular choice is not establishing a specific preference to a rule over

the others. There are exceptions to this, which will be discussed latter.

5.2.1 On the capacity allocation to levels for partial sharing

We are faced with the problem of deciding how to allocate slots of capacity to each one of the K levels, out of a global available capacity for each stage (machine). Note that this PS case degenerates to the NS case when dealing with only one product. This subsection uses Figures 5.1-5.7 as the basis of discussion.

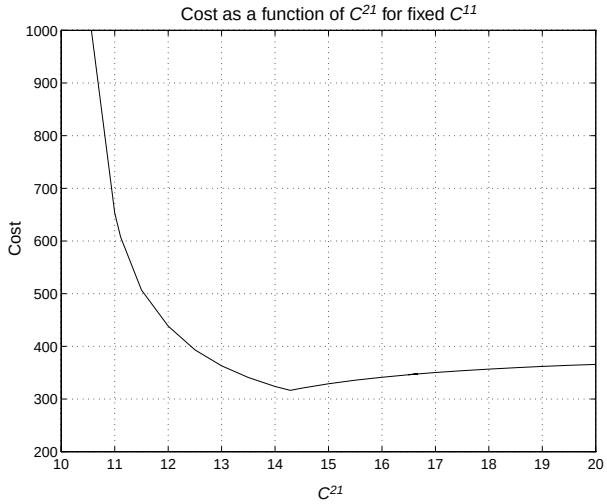


Figure 5.1: Re-entrant system operated in the NS mode. $C^{11} = \mathbf{E}[d_0^1]/0.7 = 14.29$.

Figure 5.1 refers to a situation where the value of C^{11} was kept fixed and C^{21} was varied. C^{11} was fixed at 14.29 for a $K = 2, M = 1, P = 1$ system. Note that cost decreases as C^{21} approaches C^{11} both from higher and lower values. When $C^{21} < C^{11}$, level 2 is the bottleneck and builds up inventory to avoid starvation of level 1, thus incurring high costs. When level 2 has its capacity above that of level 1, this extra capacity only increases the speed at which inventory enters the system but does not affect the speed at which it reaches the finished goods inventory, since level 1 is now the bottleneck. Therefore the extra capacity at level 2 can only increase the cost as also observed in [Glasserman and Tayur, 1995].

The behavior displayed in Fig. 5.2 refers to the situation where C^{21} is kept fixed and C^{11} is now varied. Low values for C^{11} incur high costs because level one is the bottleneck and the load imposed is high. Once, C^{11} goes above C^{21}, the cost *may* not change because, although there is

more capacity available at the last stage, it is not used since the input of inventory is bound by the output of level 2 which becomes the bottleneck.

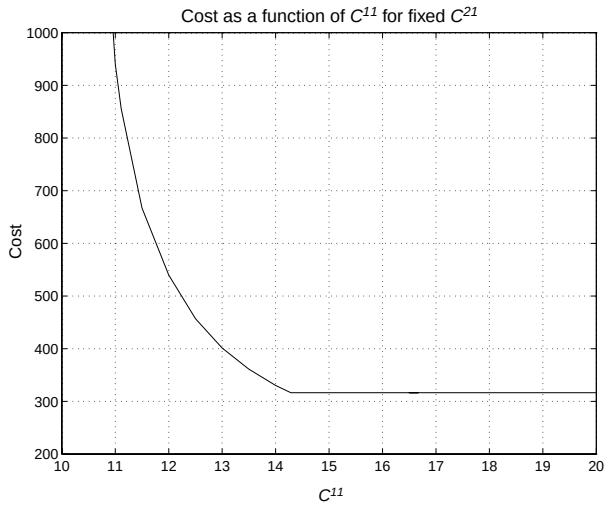


Figure 5.2: Re-entrant system operated in the NS mode. $C^{21} = \mathbf{E}[d_0^1]/0.7 = 14.29$.

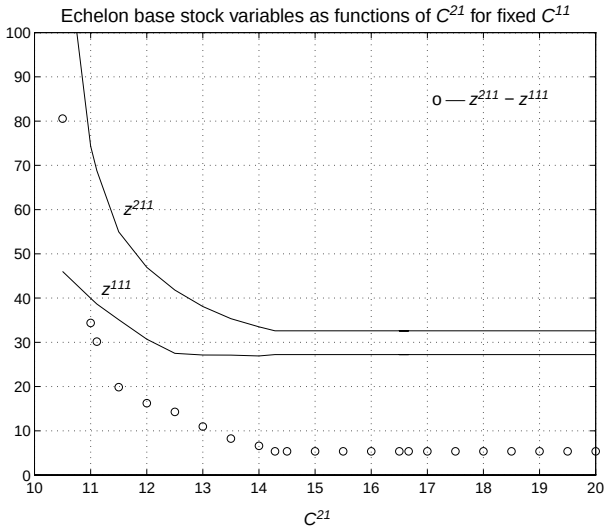


Figure 5.3: Re-entrant system operated in the NS mode. $C^{11} = \mathbf{E}[d_0^1]/0.7 = 14.29$.

The plots of the optimal base stock variables for the above two studies are presented in Figs. 5.3-5.4). They contribute to understanding why the costs have the behavior observed in Figures 5.1-5.2.

Both Figs. 5.3 and 5.4 show that the difference in consecutive base stock levels remains constant after the changing variable becomes higher than the fixed. This explains why costs increase in

Fig. 5.1 and remain constant in Fig. 5.2. Note also, that whenever level 1 imposes a strong bottleneck on the system, the quantity $z^{211} - z^{111}$ equals C^{11}. Basically, there is no need to have inventory above the capacity of level one if it is not going to move forward in less than a period. Observe that same behavior in Fig. 5.5.

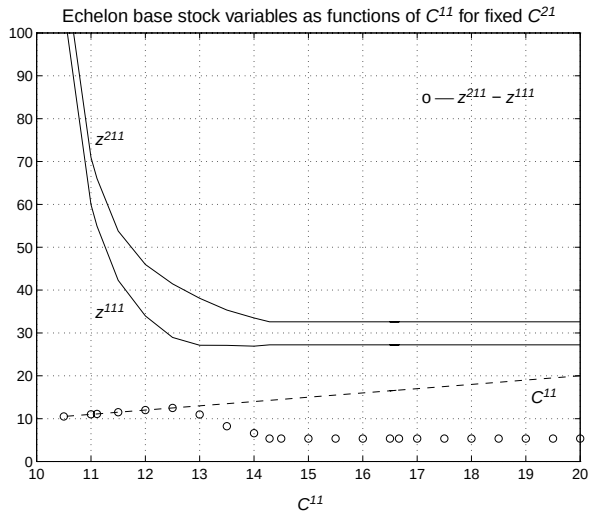


Figure 5.4: Re-entrant system operated in the NS mode. $C^{21} = \mathbf{E}[d_0^1]/0.7 = 14.29$.

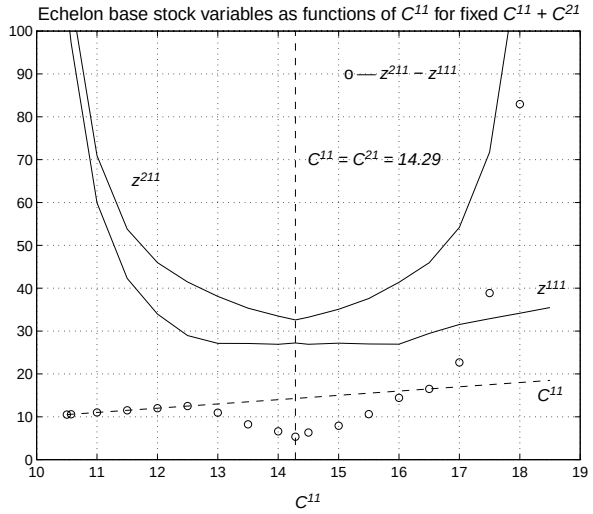


Figure 5.5: Re-entrant system operated in the NS mode. $C^{11} + C^{21} = 2\mathbf{E}[d_0^1]/0.7 = 28.57$.

One can also take derivatives with respect to the capacity slots in order to determine their optimal values (with the additional feature that there is a bound for the sum of the capacity slots).

In this scenario, there is a budget capacity on the machine, C^1, and it is necessary to decide what share to assign to each level. One can expect the optimal allocation of capacity to levels to depend on the holding costs along the production line. Surprisingly, in the majority of cases, the optimal allocation is to divide capacity equally among levels. This is observed in most of the experiments with different values of capacity, holding and penalty costs, number of levels (K), number of machines (M) and different demand distributions. Figure 5.6 displays the optimal cost from a typical experiment, and Figure 5.5 displays the corresponding optimal base stock variables.

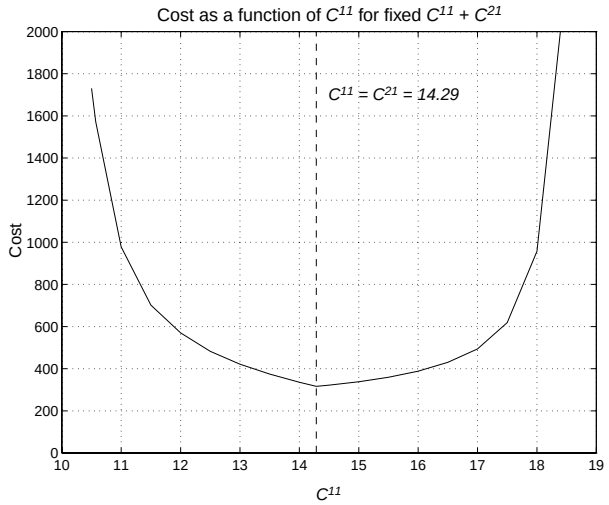


Figure 5.6: Re-entrant system operated in the NS mode. $C^{11} + C^{21} = 2\mathbf{E}[d_0^1]/0.7 = 28.57$.

However, as the holding costs of early levels decrease to low values, the optimal allocation of capacity changes. To evaluate the effects of holding costs on the optimal allocation of capacity we ran a set of experiments for the above system keeping the values of $h^{111} = 10$ and $b^1 = 20$ constant while the value of h^{211} was changed from 0 to 10. For each case, we computed the optimal cost for the optimal capacity allocation and the optimal cost with $C^{21} = C^{11} = 12.5$. The sum $C^{21} + C^{11}$ was kept constant and equal to 25, corresponding to a load of 80%.

Note that one can fix h^{111} at any value and the conclusions do not change. Once h^{111} is fixed, what matters is the relative value of h^{211} and b^1 with respect to h^{111}. We can take $\hat{h}^{211} = h^{211}/h^{111}$, $\hat{h}^{111} = 1$, and $\hat{b}^1 = b^1/h^{111}$ instead, and achieve the same conclusions, given that the total costs will only be changed by a multiplicative constant and the optimal base stock values will be the same.

Fig. 5.7 displays the results of such study. The graph on the right is a zoom of the graph on

the left for low values of h^{211}/h^{111}. The line marked with "o" refers to the percent deviation of the cost achieved with the equal capacity case relative to the absolute optimal cost. The line marked with "+" refers to the absolute optimal cost. The line marked with "×" refers to the ratio of the optimal value of C^{21} over the total capacity.

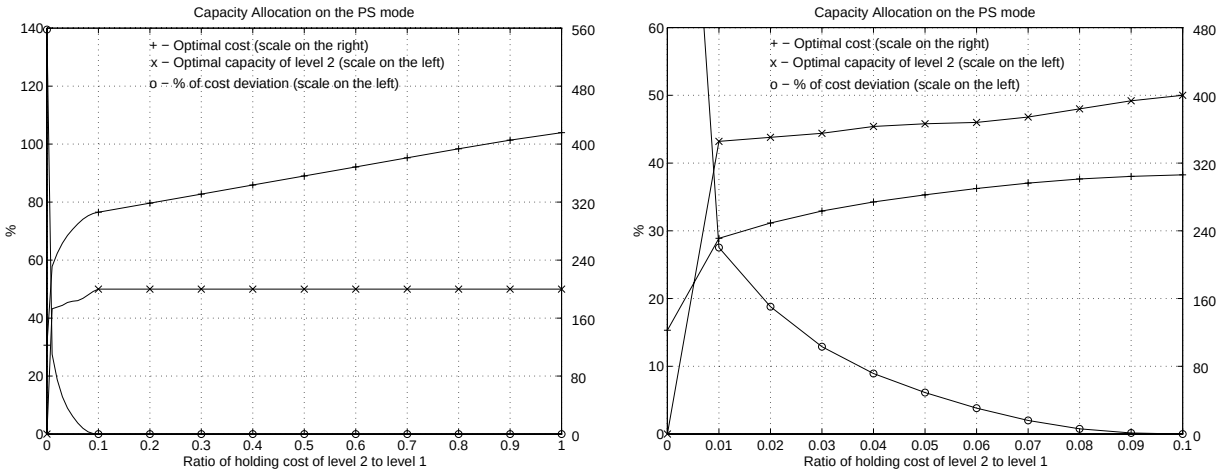


Figure 5.7: Capacity allocation as a function of holding costs.

For values of h^{211} above 10% of h^{111}, the optimal capacity allocation is achieved by dividing C^1 into two equal slots. Below that value, the optimal capacity allocation is achieved with $C^{11} > C^{21}$. The limit situation is that of $h^{211} = 0$, where it is optimal not to give any capacity to level 2. In this case, the optimal base stock levels and capacity are such that the earlier machines get less capacity at the expense of building up an enormous amount of inventory at their output buffer. Since inventory at earlier stages does not have a significant cost we can force them to be (nearly) "infinite" so creating a production system with less production stages or levels. The capacity of later machines gets the extra amount taken from the earlier machines. It is as if we would have the overall length of the production line reduced by some stages. In some cases, the difference in cost between the optimal and equal capacity allocation is as high as 40%.

The value of the penalty cost also affects the optimal allocation of capacity. The higher the value of b^1, the higher the value of h^{211}/h^{111} above which it is optimal to have $C^{21} = C^{11}$. As instances, take a case with $b^1 = 50$ where such solution is optimal for $h^{211}/h^{111} \geq 0.15$, with $b^1 = 100$ it is optimal for $h^{211}/h^{111} \geq 0.18$, and with $b^1 = 1000$ it is optimal for $h^{211}/h^{111} \geq 0.21$.

For convenience and because the starting holding costs are usually fairly large, meaning that

space does indeed cost or value is added to the products once they complete the first operation, in the remaining of this chapter the allocation of capacity to levels is done by dividing the total amount available into equal slots. Thus, some of the results obtained for very low values of the holding costs should be interpreted carefully. This is specially relevant when comparing the performance of the priority rule with others on the PS mode. In the TS mode the priority rule is the only one which is able to build up inventory at early levels in order to emulate the behavior described here.

5.2.2 Total Sharing

This subsection uses the results obtained for a system with $K = 3$, $M = 1$, and $P = 1$, with the optimization done with respect to the base stock variables, as the basis for discussion. The study is intended at understanding the effect of different holding cost structures for the levels and at comparing the three production rules on the TS mode and the NS mode. The study also encompasses an analysis for different loads. We will be presenting only a representative sample of the type of data obtained. The intermediate holding costs are changed for fixed values of $h^{111} = 10$ and $b^1 = 20$.

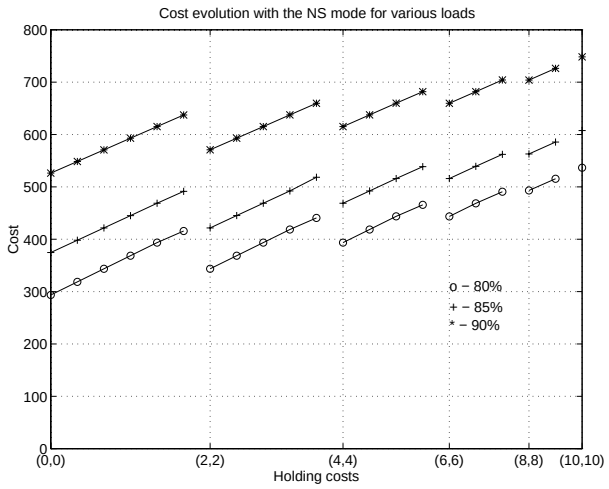


Figure 5.8: NS mode with different system loads: optimal cost.

Figure 5.8 displays the optimal *cost* for the NS mode. The results for the three production rules under the TS mode for three different production loads as a function of the intermediate holding costs are shown in figures 5.9-5.11. The product cost structure ranges from $[0, 0, 10, 20]$ to $[10, 10, 10, 20]$. On the coordinate axis the entry $(4, 4)$ represents a system with $h^{311} = h^{211} = 4$,

$h^{111} = 10$, and $b^1 = 20$, that is, $[4, 4, 10, 20]$. Between label $(2, 2)$ and label $(4, 4)$ lie labels $(2, 4), (2, 6), (2, 8)$, and $(2, 10)$ in that order, which correspond to the cost structures $[2, 4, 10, 20]$, $[2, 6, 10, 20]$, $[2, 8, 10, 20]$, and $[2, 10, 10, 20]$ respectively. All data presented for the Priority Rule refers to priority given to levels closer to completion. Any other priority assignment, in the single product setting, achieves higher costs (See Section 5.4).

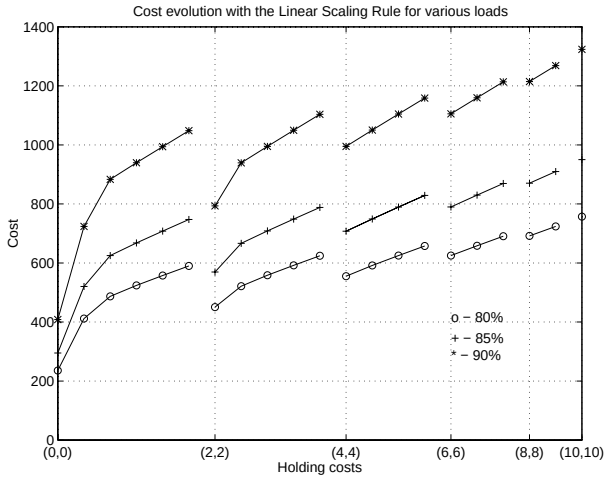


Figure 5.9: LSR with different system loads: optimal cost.

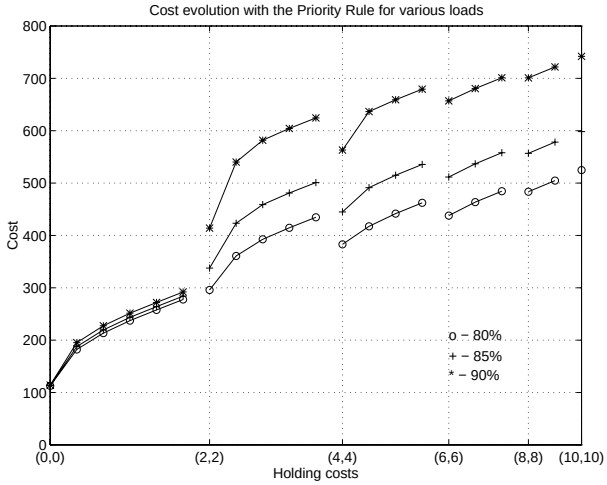


Figure 5.10: PR with different system loads: optimal cost.

From the analysis of these figures (see Figure 5.12 at 90 % load that compares all four rules in one figure), it is easy to see that *the Priority Rule outperforms the other two rules in the TS*

mode as well as the NS mode. The Equalize Shortfall Rule has only a very slight advantage over the NS mode and converges to the same levels of performance as those of the Priority Rule. The LSR performs terribly in the TS mode.

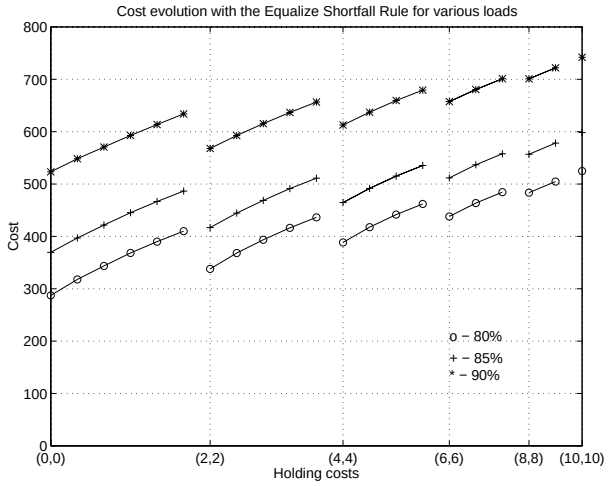


Figure 5.11: ESR with different system loads: optimal cost.

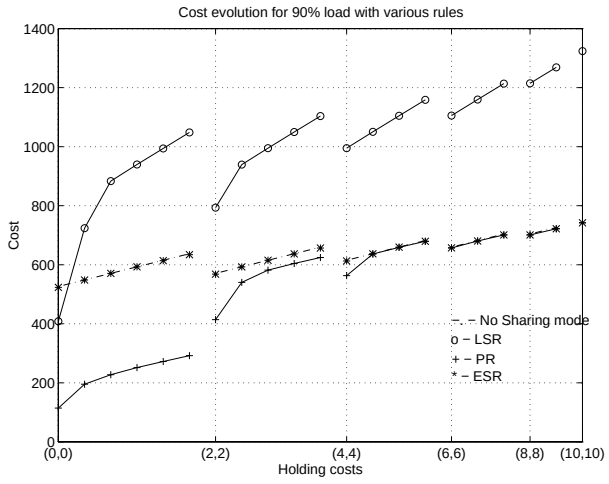


Figure 5.12: Optimal cost for 90% load comparing the four different capacity management schemes.

A closer inspection reveals that this is due to the way the scaling of production net needs is done. All levels except the entering level (level K) may be bound by feeding inventory. Level K is never bound by feeding inventory because this is assumed to be infinite. Therefore, if there is a large shortfall, the production net needs of level K match the shortfall, but all other levels may

be bound by inventory. Since the scaling (for capacity allocation) is done in terms of production net needs, it turns out that level K gets a higher share, thus affecting the lower levels. So, it is as if we are giving a higher priority to level K in terms of the dynamic capacity allocation. Thus, this preference for new material makes it more difficult for the products to move fast down the production line. It is known, from the studies on the PR partially shown in the appendix, that giving priority to level K over the others incurs always high costs. The LSR in the TS mode does just that. As it will be seen for multiple products, this behavior is not as relevant in the PS mode. That is due to the fact that the total available capacity for each level is smaller in the PS mode that it is in the TS mode. Therefore, in the PS mode the distortion of level K over the others has a smaller impact.

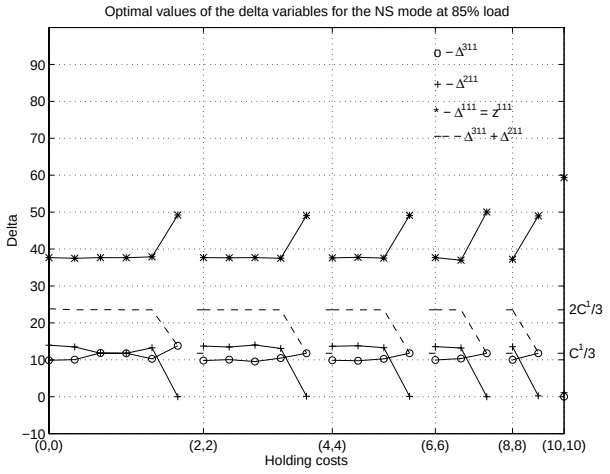


Figure 5.13: Optimal Δ variables for the NS mode under an 85% load.

Figures 5.13-5.16 show how the *optimal delta levels* behave under the four rules at 85 % load and various holding cost settings. Optimization can either be done with respect to the base stock variables or with respect to the delta variables. These plots refer to the later, although they could be generated from the former. The reason for the choice of these plots is because they help make some of the structural properties of the solutions more evident[2]. Note that, besides the convergence in optimal costs, there is also an interesting convergence for the optimal variables between the ESR, the PR, and the NS mode for single product. The convergence between the PR and the other two schemes occurs only when the holding cost structure becomes balanced (or even) along the

[2]Plots of the base stock variables can be found in Appendix D.

production line. In the cases where the initial stage incurs no costs or very little cost, the PR takes advantage of that by building up inventory and therefore emulating a system with one or two less stages. Whenever that is no longer possible due to cost considerations then ESR or the NS mode achieve the same costs as the PR.

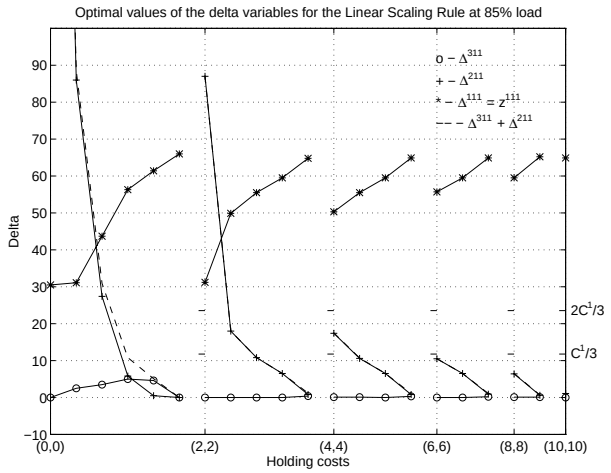


Figure 5.14: Optimal Δ variables for the LSR under an 85% load.

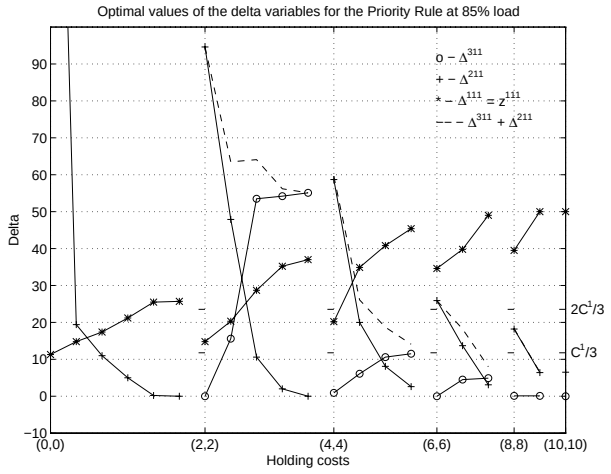


Figure 5.15: Optimal Δ variables for the PR under an 85% load.

Referring still to Figures 5.13-5.16, note the *almost constant* behavior of $z^{311} = \Delta^{311} + \Delta^{211} + \Delta^{111}$ in the NS mode and the ESR across the different holding costs. Simply add the dashed line

with the starred line on the plots[3]. This also occurs for the other rules once the intermediate holding costs become significant when compared with the terminal holding cost. This seems to imply that the topmost base stock is more sensitive to the terminal holding and penalty costs than it is to the intermediate values, if the intermediate values are not too small. For all the rules it is also evident that the different distribution of holding costs along the the system has the effect of distributing the inventory on the different levels. Note that $\Delta^{311} = z^{311} - z^{211}$ approaches zero when $h^{311} - h^{211}$ approaches zero and that $\Delta^{211} = z^{211} - z^{111}$ also approaches zero when $h^{211} - h^{111}$ approaches zero.

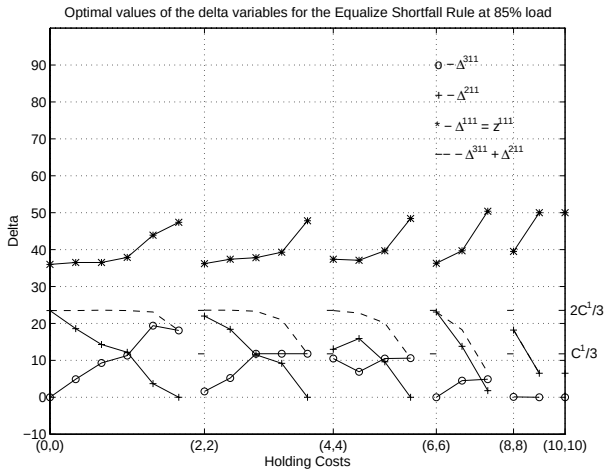


Figure 5.16: Optimal Δ variables for the ESR under an 85% load.

It is interesting to note that for some instances the sum $\Delta^{311} + \Delta^{211}$ is constant (Figs. 5.13 and 5.16). The value achieved for those instances equals $\frac{2}{3}C^1$. In some other instances, the value Δ^{311} or Δ^{211} is constant and equals $\frac{1}{3}C^1$ (Fig. 5.16). This shows the interaction between capacity and inventory levels is subtle, and in some sense, very exact.

Recall the proposition in Section 5.1 referring to the *optimality condition*. Most of the simulation runs satisfy this condition, but not all. Consider Figures 5.17-5.18 that show the cost function around the optimal. These plots were taken along the gradient direction. The first figure of the set was taken for the LSR and the second for the ESR. In the first case, the optimality condition is satisfied for the five values of $a = h^{211}$. In the second case, the system failed to achieve the optimality condition for one value of $a = h^{311}$ (starred on the figure). Note that it naturally

[3]Or check Appendix D.

coincides with a situation where the cost function is non differentiable. The failure to achieve the optimality condition coincides with a case where either $\Delta^{311} + \Delta^{211} = \frac{2}{3}C^1$, $\Delta^{311} = \frac{1}{3}C^1$, $\Delta^{211} = \frac{1}{3}C^1$, or some $\Delta^{k11} = 0$. In these cases the cost function is not differentiable around the optimum as these graphs show; recall the remarks made in Section 3.4.

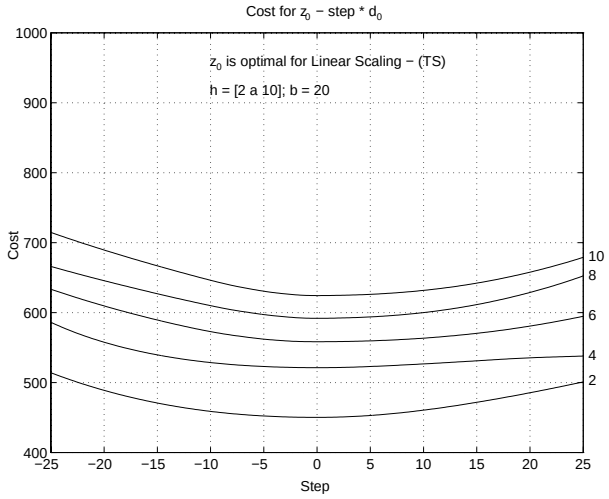


Figure 5.17: Cost along the gradient direction: summary for $h^{311} = 2$.

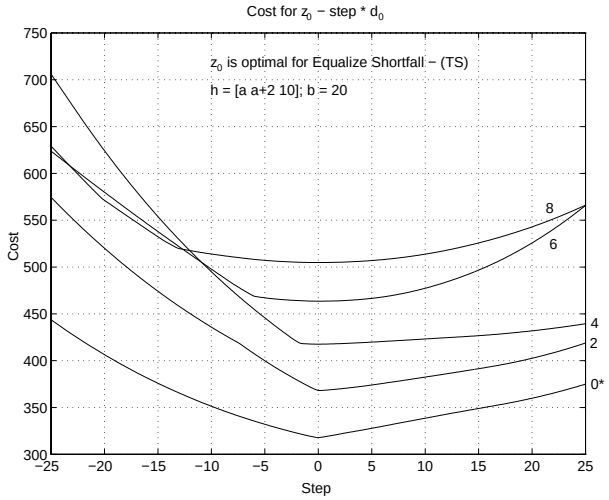


Figure 5.18: Cost along the gradient direction: summary for $h^{311} - h^{211} = 2$.

Naturally, to converge to the absolute optimal values, these couplings between successive variables would have to be explicitly included, reducing the number of variables to consider[4]. The

[4]See Appendix C for a more detailed discussion on this issue.

results here displayed were obtained without this consideration. Whenever a non differentiable point occurs during simulation, the system takes one of the subgradients. It turns out that, as the cost function is relatively flat around the optimal, the deviation to the optimal values is not very significative when doing this. That is, when trying to improve over the values obtained with the first optimization, by explicitly forcing the coupling observed at the end of it, the gain in terms of cost was marginal in all the cases tested.

5.3 Multiple Products

Turning to systems with multiple products, the objective is to understand the performance of the capacity allocation and production rules in this setting. As in the single product case, *the best production rule for the TS mode always achieves better costs than the best production rule for the PS mode*, although in some cases the difference may be very small. This reinforces the generally held belief that the bigger the flexibility, the best advantage one can make of the available resources. A notable fact is that *the LSR is the only rule that degrades its performance in the TS mode relative to the PS mode*. We briefly explained the reasons for this before. Specific observations and insights gained will be mentioned in the following subsections.

First, we will analyze the effect of holding costs on the relative performance of the rules. The main finding is similar to that of the single product case. In general, the relative performance of the rules is not affected by the particular holding cost structure for each product. The exception occurs for negligible holding costs at earlier levels for reasons already discussed.

Once this is established, one can fix the holding costs arbitrarily for each product and move on to analyze the effects of the penalty costs and demand variance.

This study considers systems with a single machine and two products. Each product is required to visit the machine three times before completion. So, we will have $K = 3$, $M = 1$, and $P = 2$. The average demand for both products is fixed in all the experiments conducted. Their values are $\mathbf{E}[d_0^1] = 8$ and $\mathbf{E}[d_0^2] = 12$. The total capacity of the single machine is fixed to an average load of 80%, that is, $C^1 = (3 \times 8 + 3 \times 12)/0.8 = 75$.

The presence of multiple products introduces additional options in the way priority can be assigned within the PR. It was mentioned before, in the single product setting, that within the PR, the best performances are achieved when higher priority is given to the levels closer to completion.

For that reason we limited the studies to such situations. In the multiple product setting, even if one decides to maintain such priority scheme, it is necessary to establish if the multiple products are intertwined or if one product is always preferred to the other no matter the level. There is also the option of establishing different priorities for the products on each level. During the experiments conducted, it was found that *the best performance is obtained by keeping the same priority across levels for all the products*. Also, it was found that, for total sharing of capacity the best performance is achieved when products are intertwined. See Section 5.4.1 for description of this concept. Therefore, we will be using the following priority scheme on the studies reported here.

A global priority for the products is defined. For the partial sharing case that is all needed; for the total sharing case the global priority for products is applied level by level, starting from the level closer to completion. So, for $K = 3$, $M = 1$, $P = 2$ on total sharing mode with priority given to product 1, the production decisions are taken by the order: P_n^{111}, P_n^{112}, P_n^{211}, P_n^{212}, P_n^{311}, and P_n^{312}. Note that for this type of systems there are a total of $(K \times P)! = 6! = 720$ different priority assignments.

5.3.1 Same holding cost structure for products

The first study was designed to establish a means of comparison for all the other studies. The basic features of the systems considered are as defined above. Additionally we kept the coefficient of variation for both products fixed at 1. The costs h^{11p} and b^p were fixed at 10 and 20, respectively for $p = 1, 2$. Twenty one different systems were generated, by changing the holding costs of level 3 and level 2, that is h^{31p} and h^{21p} for $p = 1, 2$. All of the 21 systems have the same cost structure for both products, that is, $h^{311} = h^{312}$ and $h^{211} = h^{212}$. The cost structure of the first system is given by $[0, 0, 10, 20]$ for both products and the cost structure of system number 21 is $[10, 10, 10, 20]$. For each one of the 21 systems was obtained the optimal solution for all three production rules each with PS and TS. For the case of the PR, there are two choices as discussed earlier. All combined, there are 8 solutions per system what totals 168 solutions.

Figure 5.19 on the left displays the optimal costs for the PS mode. The first observation that should be made is that the change in holding costs does not affect the relative performance for the several rules. The LSR and the ESR achieve practically the same costs and perform better than any of the two possible priority assignments. Within PR, priority should be given to product 1 over product 2 to achieve the best performances. In general, all things being equal, priority should

be given to products with the lower average demand within PR.

For the TS mode it is also the case that priority should be given to the product with the lowest demand to achieve better performances, as shown in Figure 5.19 on the right. There is no best rule across all costs in this setting, in contrast to the PS setting. Priority to either product achieve practically the same best costs for the situations where $h^{31p} = 0$ and also for the case where $h^{31p} = h^{21p} = 2$. ESR performs best in any situation with higher holding costs. The advantage of the PR in the low holding cost cases is due to the build up of inventory since the costs are so low. Recall Figure 5.7 and the associated discussion on the case with low holding cost at early levels.

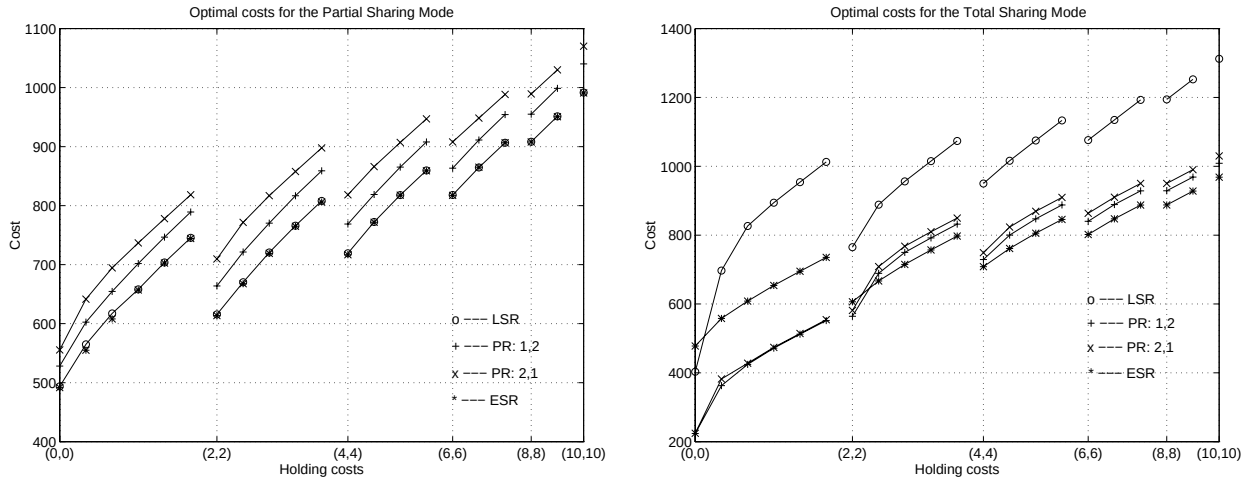


Figure 5.19: Optimal cost for the PS and TS modes as a function of the same holding costs for both products.

5.3.2 Different holding cost structure for both products

The second study concerns investigating the relative performances of the different rules when products have different holding cost structures. The study is composed of two sets of experiments. In the first set, the cost structure of product one was kept constant at $[2, 6, 10, 20]$, and the cost structure of product two was changed from $[0, 0, 10, 20]$ to $[10, 10, 10, 20]$ thus generating 21 different systems. In the second set we exchanged the positions of product one and product two, generating another set of 21 different systems. For each one of the two sets of experiments there were 8 solutions generated, as before. This study is comprised of 336 solutions.

Figures 5.20 and 5.21 on the left display the optimal costs for the PS case for both sets. The

ESR rule continues to achieve the best performance irrespective of costs for the PS mode. The LSR is a very close second. Regarding the priority assignment, the conclusion of the previous sub-section remains valid: priority should be given to the product with the lowest demand, when the coefficient of variance is the same. Note the only exception to the conclusion over priorities, which occurs in the second set with the cost structure of product 1 equal to $[0, 0, 10, 20]$. This is the only case, in both sets, where the best performance for PR occurs by giving priority to product 2 over product 1. Note also the convergence in performance for the two choices of priority in the first set (Fig. 5.20). It is noticeable that as the holding cost of product 2 increases and is higher than the holding costs of product 1, giving priority to product 1 over product 2 achieves a lesser gain. For the second set the behavior is naturally opposite. As the holding costs of product 1 rise so does the net gain of switching from higher priority to product 2 to higher priority to product 1.

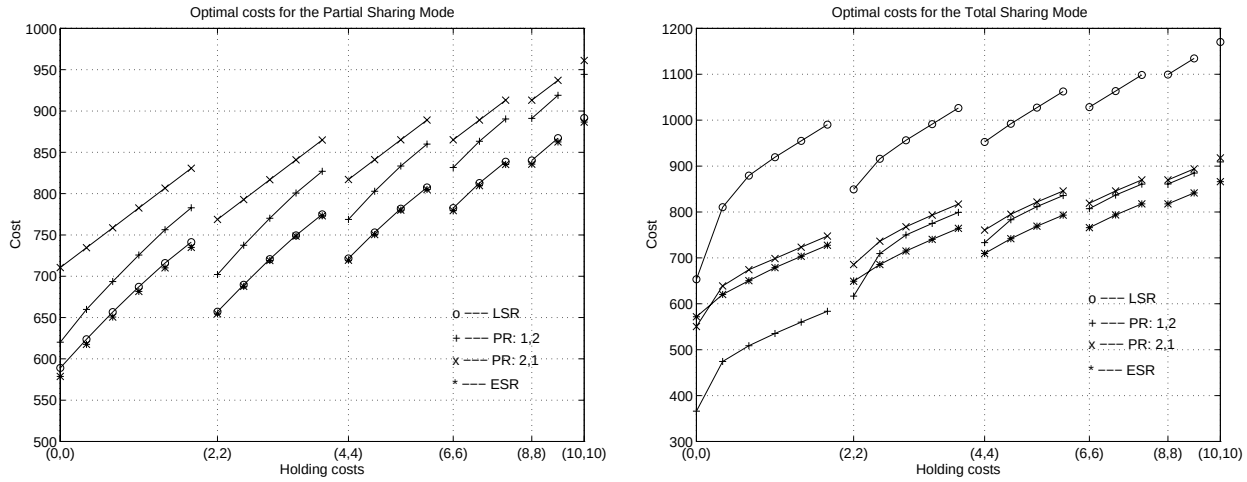


Figure 5.20: Optimal cost for the PS and TS mode as a function of the holding costs for product 2.

When it comes to the TS mode (Figs. 5.20 and 5.21 on the right), things change a little for the PR. One observation continues to hold: the ESR is still the best rule for moderate to high holding costs. Also for moderate to high costs, priority to product 1 outperforms priority to product 2. Only for very low holding costs the PR achieves the overall best performance. Moreover, priority to product 2 is better than priority for product 1 only for low holding costs of the second set, when the holding cost of product 1 is very close to zero. Also, the performance of the two choices for the PR converge when the holding costs of product 2 rise above those of product 1, and they diverge in the opposite situation.

This and the previous sub-sections have shown that the average demand seems to be a determinant factor in deciding to which product one should give higher priority. It is now necessary to see the effect of changing the penalty cost for the products, which were kept constant and equal for both products. This should probably justify why the ESR achieves an overall best score of all the rules. Since the penalty for backlog is the same for both products, as are the terminal holding costs, it is as if the shortfall has the same price (or cost), and therefore trying to equalize it should be a good strategy.

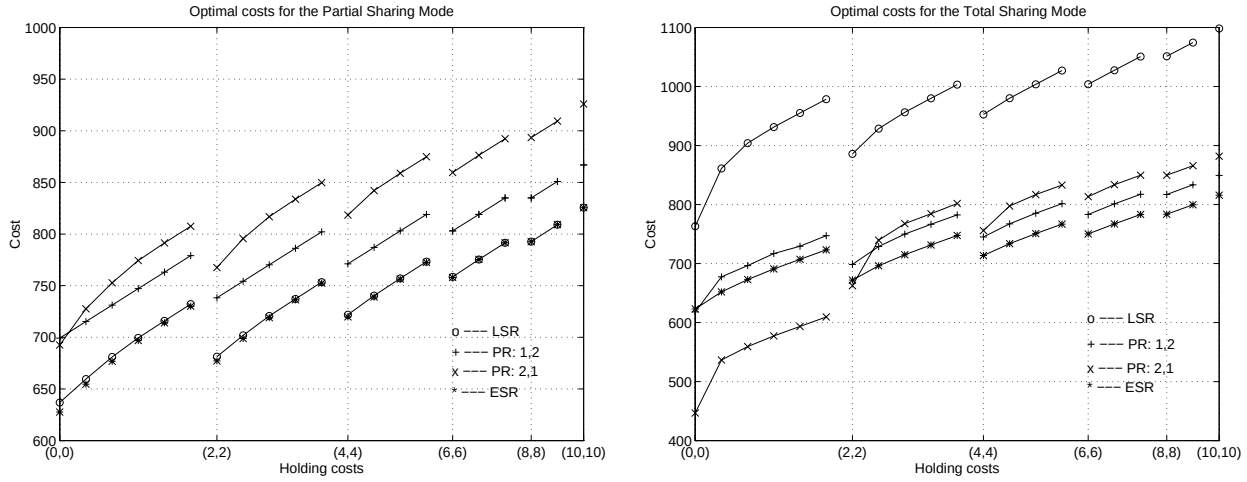


Figure 5.21: Optimal cost for the PS and TS mode as a function of the holding costs for product 1.

5.3.3 Changing the penalty costs

In studies described so far, although changing the holding cost patterns for the products, the penalty costs were kept constant. This third study analyses the effect of changing the penalty costs. So, the basic features remain the same as those of the two earlier studies. Two sets of experiments were run. The first set is characterized for having the cost structure for product one fixed at $[6, 8, 10, 20]$ and the cost structure of product two is $[2, 6, 10, b^2]$, with $b^2 \in [10, 50]$. The second set of experiments has the cost structure of product one at $[6, 8, 10, b^1]$, with $b^1 \in [10, 50]$ and the cost structure of product two fixed at $[2, 6, 10, 20]$. Each of the two sets comprises 21 different systems, leading to the generation of 336 different solutions.

Let us analyze the results obtained set by set. The optimal costs of the first set in the PS mode

are displayed in Figure 5.22 on the left. As before, both the ESR and the LSR tie for the first place being very close to each other. It turns out that the ESR wins for high values of b^2 and loses to the LSR for low values. For the PR, the penalty cost variation introduces a more interesting behavior. There is a value of b^2 above which the best performance is achieved by the PR when priority is given to product 2. However, only for very low values of b^2 is possible for the PR with priority given to product 1 to approach the performance of the ESR and the LSR.

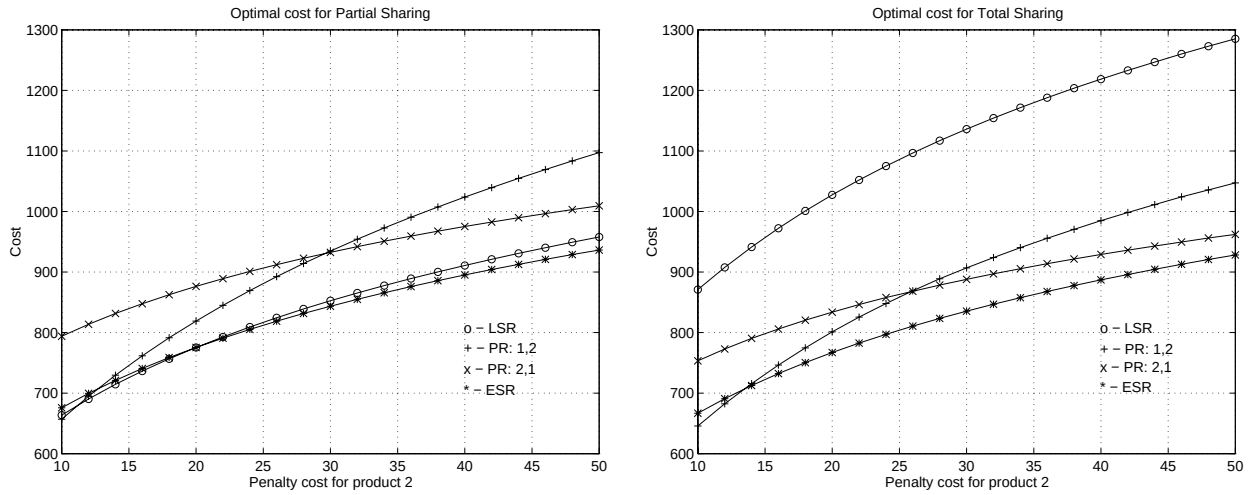


Figure 5.22: Optimal cost for the PS and TS mode as a function of the penalty cost for product 2.

The ESR is still the winner for the TS mode for a wide range of values. Off those cases, the PR with priority given to product 1 achieves the best performance for low values of b^2. As in the PS mode, there is a value for b^2 above which priority should be given to product 2. Observing the slope of the curves in Figure 5.22 on the right one can argue that eventually there will also be a value for b^2 above which the best production rule is the PR, with priority given to product 2. To confirm this we ran a case with $b^2 = 200$. The optimal cost achieved with the ESR was 1204.4 and the PR, giving priority to product 2, achieved a cost of 1176.5, thus confirming the hypothesis.

Turning to the second set of this study, where the value of b^1 was changed, in the PS mode there is still a competition between the LSR and the ESR, and for high values of b^1, the PR with priority given to product 1 achieves the best costs. See Figure 5.23. The ESR wins for low values of b^1 and the LSR wins for intermediate values. Priority to product 2 is completely out of contest, except that there is a range of values where it beats priority to product 1. As should be expected, this occurs for values of b^1 much lower than those of b^2.

In the TS mode, the only rules that are worth mentioning are the ESR and the PR with priority to product 1. In fact, there is a value of b^1 below which the ESR should be used and above which the PR should be used. This observation and those made earlier on this sub-section re-enforce the observation we set forth earlier that the ESR performs best when the shortfalls have similar prices.

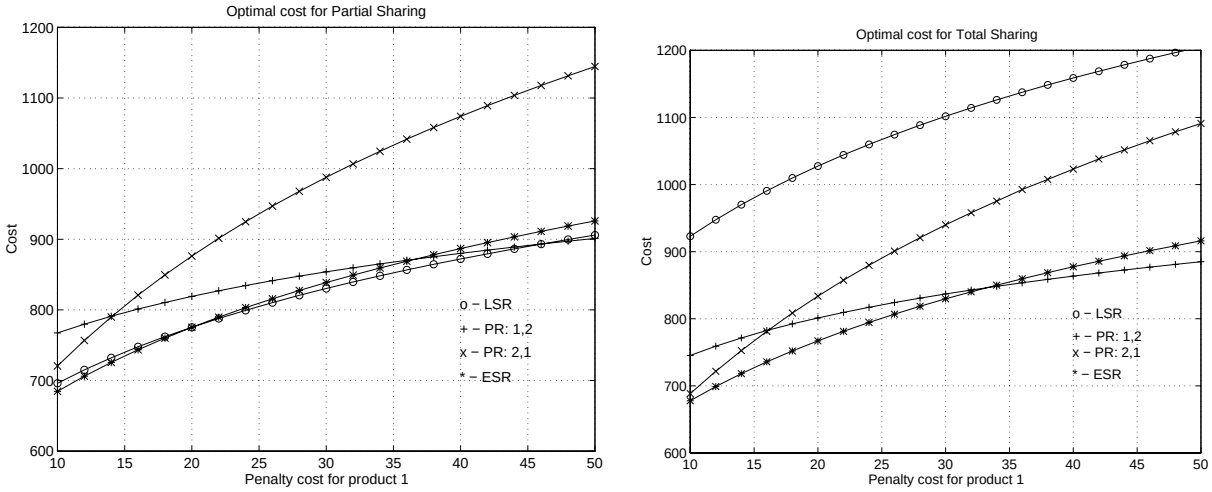


Figure 5.23: Optimal cost for the PS and TS mode as a function of the penalty cost for product 1.

5.3.4 Changing the coefficient of variation for the demand

In this next study, the objective is to investigate the effect of variance on the relative performance of the production rules.

Table 5.1: Parameters for the experiments.

Set	Cost structure for product 1	Cost structure for product 2	Variance range	Product
1	[6, 8, 10, 20]	[2, 6, 10, 14]	0.1 – 1.0	2
2	[6, 8, 10, 20]	[2, 6, 10, 14]	0.1 – 1.0	1
3	[6, 8, 10, 20]	[2, 6, 10, 50]	0.1 – 1.0	2
4	[6, 8, 10, 20]	[2, 6, 10, 50]	0.1 – 1.0	1
5	[6, 8, 10, 14]	[2, 6, 10, 20]	0.1 – 1.0	2
6	[6, 8, 10, 14]	[2, 6, 10, 20]	0.1 – 1.0	1
7	[6, 8, 10, 50]	[2, 6, 10, 20]	0.1 – 1.0	2
8	[6, 8, 10, 50]	[2, 6, 10, 20]	0.1 – 1.0	1

The dimension of the system, average demand for the products, and load were kept the same

as before. Eight sets of experiments were run. Their characteristics are described in Table 5.1. Each of the eight sets has 10 different systems for which 8 different solutions were generated, which totals 640 solutions.

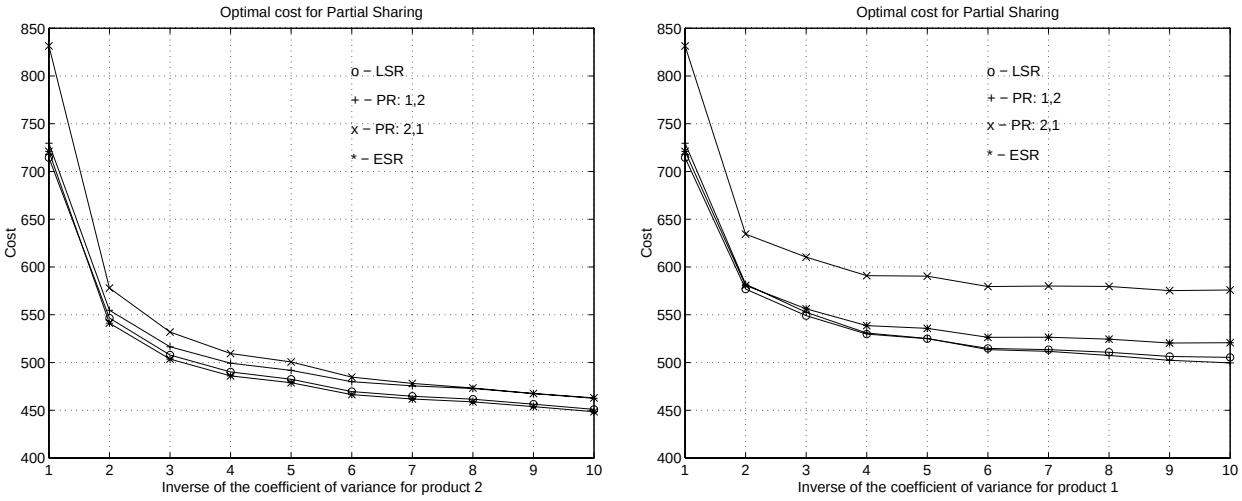


Figure 5.24: Optimal cost for the PS mode of set number 1 and number 2.

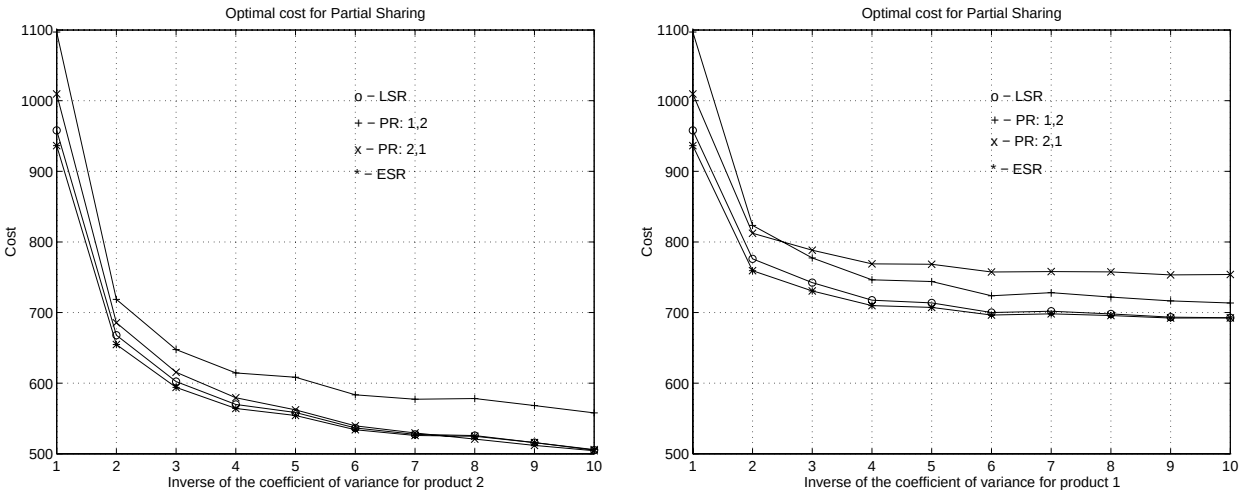


Figure 5.25: Optimal cost for the PS mode of set number 3 and number 4.

Figures 5.24 and 5.25 display the optimal costs for the first four sets in the PS mode. In all but the second set, the ESR is the best rule with the LSR as a close second. For the second set, we have the PR as the best rule and the LSR as second. In the second set, note that all elements are combined in the same direction to favor the PR: the product to which is given priority has the

lowest demand, the lowest variance, and the highest penalty cost.

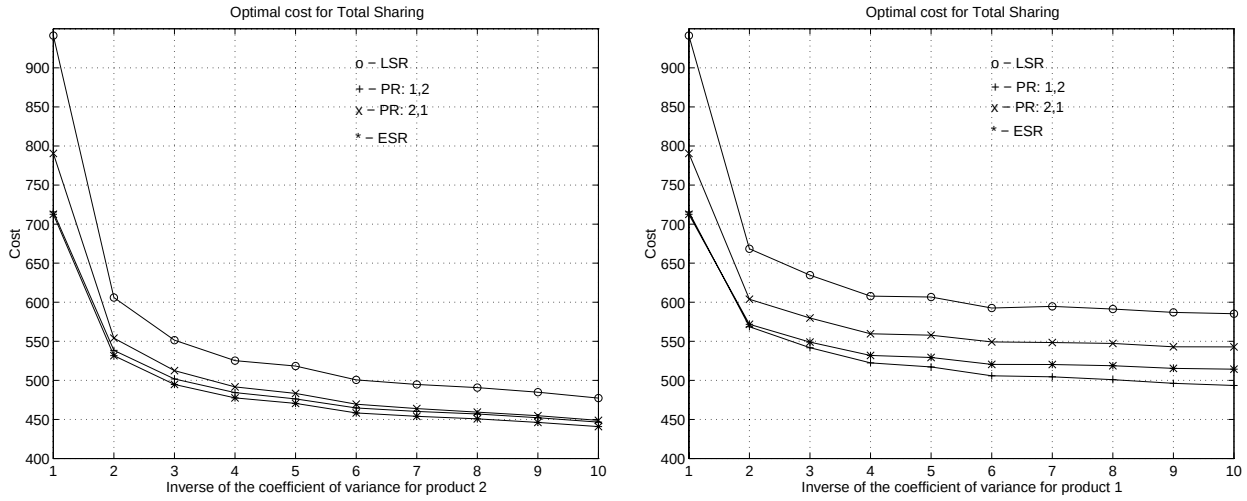


Figure 5.26: Optimal cost for the TS mode of set number 1 and number 2.

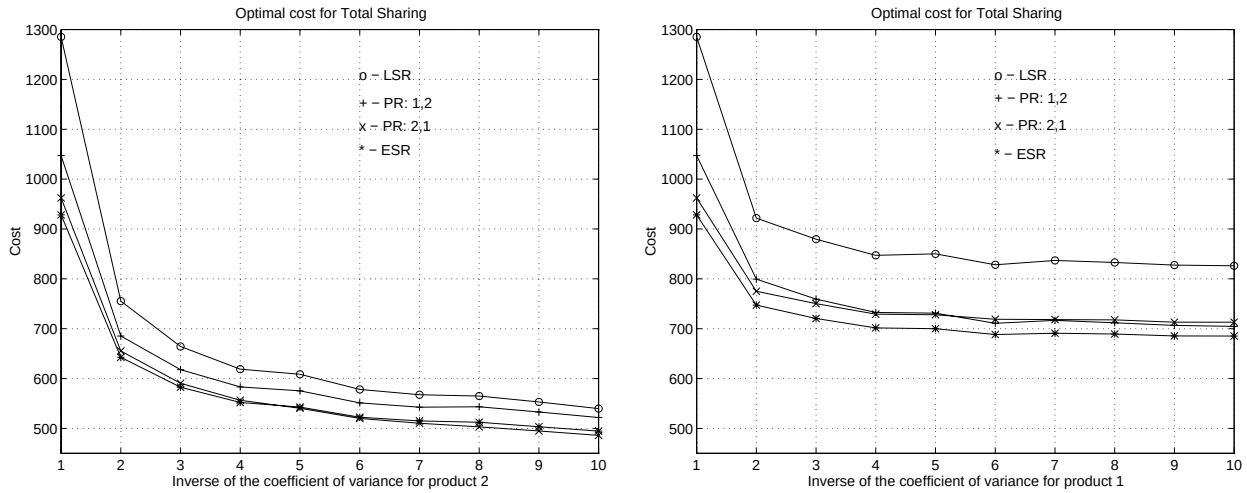


Figure 5.27: Optimal cost for the TS mode of set number 3 and number 4.

These results allow the conclusion that variance has a very strong effect on the relative performance of the PR. In the first set, as the variance of product 2 decreases, giving priority to product 1 (with the lowest demand) does not translate in such a high gain over priority to product 2. Set number 2 also shows a clear preference for the product with the lowest demand. In set three the decrease in variance for product 2, combined with its higher penalty cost, translates into an increasing gap between product 2 and product 1. In set number 4, at first the winner is priority to

product 2 which has the higher penalty cost but, as the variance of product 1 decreases, priority should then be given to product 1 despite its lower penalty cost.

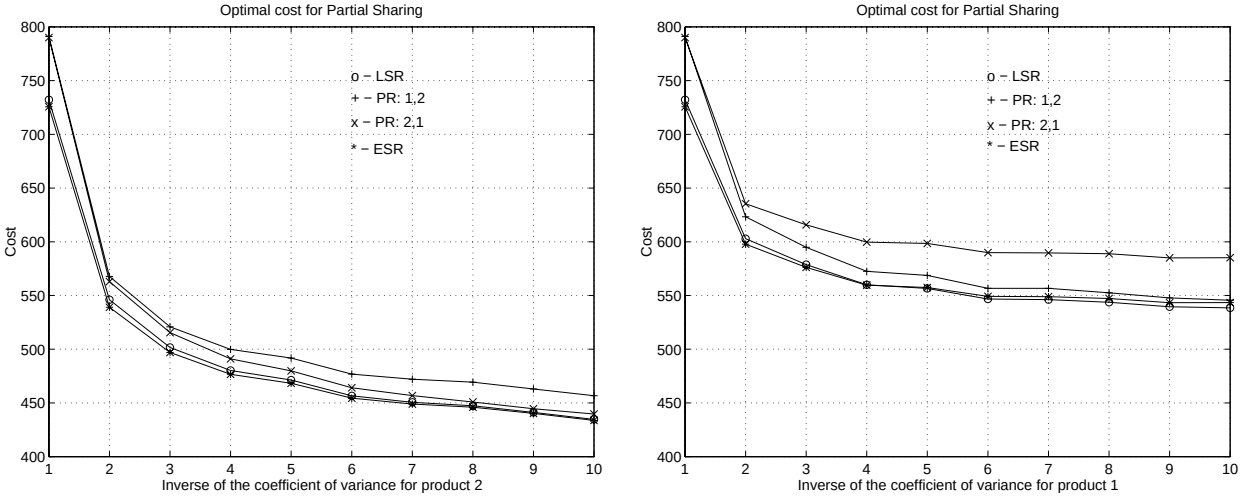


Figure 5.28: Optimal cost for the PS mode of set number 5 and number 6.

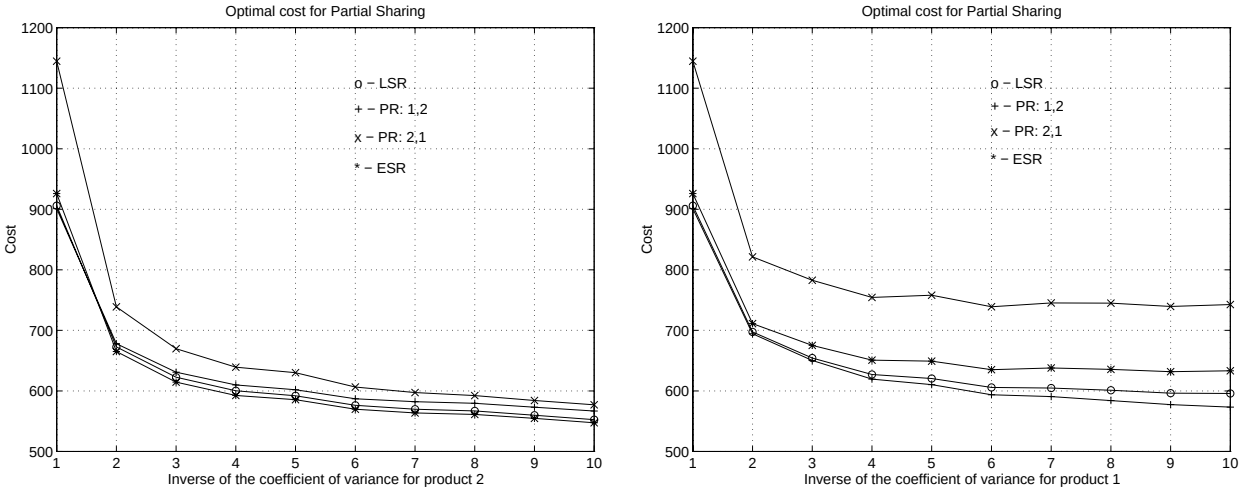


Figure 5.29: Optimal cost for the PS mode of set number 7 and number 8.

The optimal costs for the first four sets when capacity is run on a TS mode are shown in Figures 5.26 and 5.27. The ESR achieves the overall best performance in sets number 1 and 4. In set number 2, priority to the product with lowest mean demand, lowest variance and highest penalty cost (product 1) is the best choice. In set number 3, the ESR and priority to product 2 (lowest variance and highest penalty cost) share the first place. The PR outperforms the ESR when

variance for product 2 is sufficiently low. However, note that the differences between these two for set number 3 are very small.

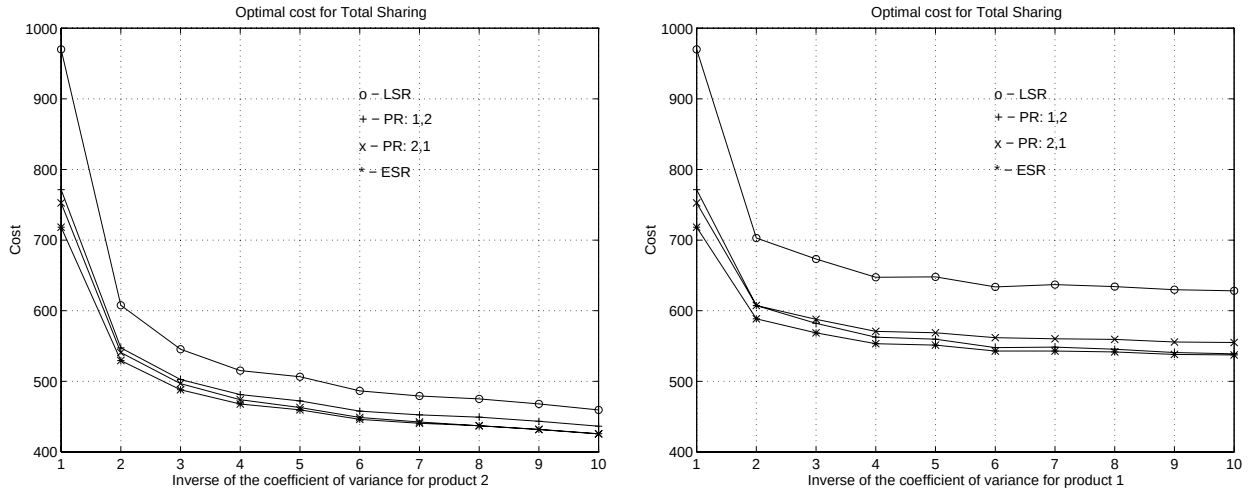


Figure 5.30: Optimal cost for the TS mode of set number 5 and number 6.

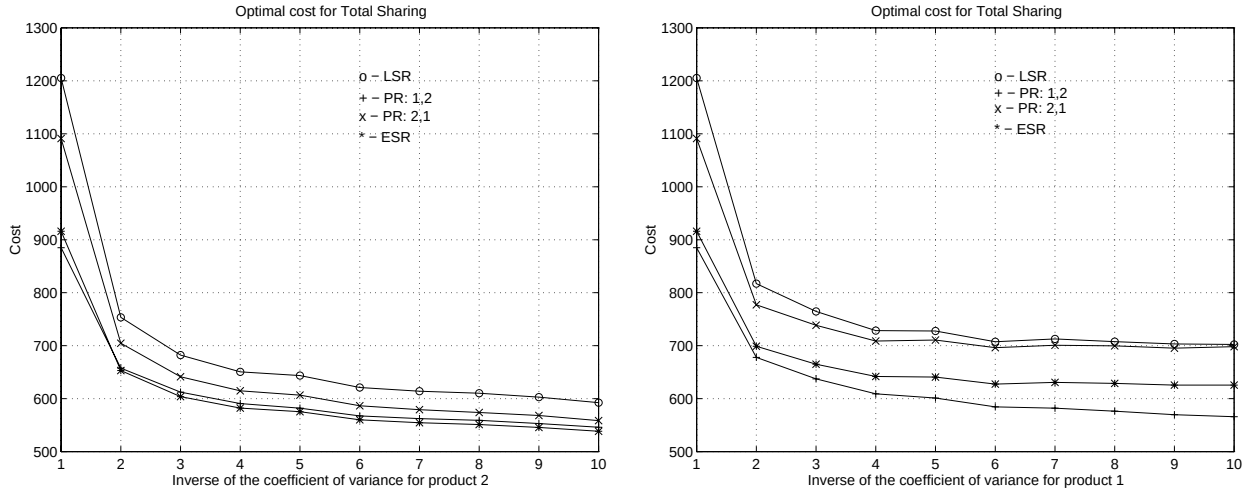


Figure 5.31: Optimal cost for the TS mode of set number 7 and number 8.

The last four sets of experiments do not change significantly the main observations made for the earlier sets. The LSR and the ESR share the first place on the PS mode, except for set number 8 (Fig. 5.29, right). For this set, the first place is shared between the LSR and priority to product 1, which is the product that simultaneously has the highest penalty cost, lowest mean demand, and lowest variance. The advantage of the PR over the LSR grows as the variance for product

1 decreases. The relative performance of the two choices of priority does not change significantly from those observed earlier in the first four sets.

In summary, for the TS mode, the winner is again the ESR, except for the cases where there is a strong favorable combination of parameters towards a single product being given priority over the other.

5.3.5 Effects of Capacity for the TS Mode

Another aspect of interest is the evaluation of the effect of capacity distribution along a production line on the different machines. That is, given that some machines may have different loads, or different capacities when loads are uniform, is there a real advantage in using all of the available capacities on the TS mode?

To investigate this aspect a series of systems were evaluated. The dimensions are $K = 2$, $M = 3$, and $P = 2$. The average demand for product 1 is 12 and for product 2 is 8, both with coefficient of variation set at 1. The cost structure for product 1 is [5 6 7 8 9 10 20] and for product 2 is [2 2 2 10 10 10 50]. For each of the rules two sets of studies were performed. In the first set, each one of the capacities was varied from 50 to 100 while keeping the other two fixed at the value of 50. In the second set, each one of the capacities was kept fixed at 50 while the other two were varied from 50 to 100.

In Fig. 5.32 is the plot of optimal costs for the LSR operated on the TS mode. The plot on the left corresponds to the cases where only one of the capacities is changed from 50 to 100. The plot on the right corresponds to the cases where only one capacity is fixed at 50 while the other two change from 50 to 100 with equal values.

When only one capacity is increased (plot in the left), cost increases as C^3 increases, for reasons similar to those discussed earlier when referring to the input of new material into the system. In fact, during the busy periods, given that C^3 is higher it allows the input of more material than the downstream machines can process it. Therefore, this extra material only increases holding costs. When C^2 increases, cost does not change given that whatever gets through machine C^3 is always under the capacity of machine C^2. Therefore, since there is no change in the way inventories move forward there should not be any change in costs. Things change when it is C^1 to change. There is some advantage in terms of having the capacity of the last machine a little over the previous two.

This is easily explainable, since this extra capacity allows for faster recoveries during busy periods. However, note that the improvement is not only marginal but it also stabilizes after some value of capacity. In fact, the advantage on having extra capacity at the last machine is only relevant for the last operation, since all the other levels continue to be constrained by the other machines in the same fashion as with $C^1 = 50$. There should be a point for C^1 after which the edge provided by the extra amount of capacity becomes irrelevant and the cost should stabilize as observed.

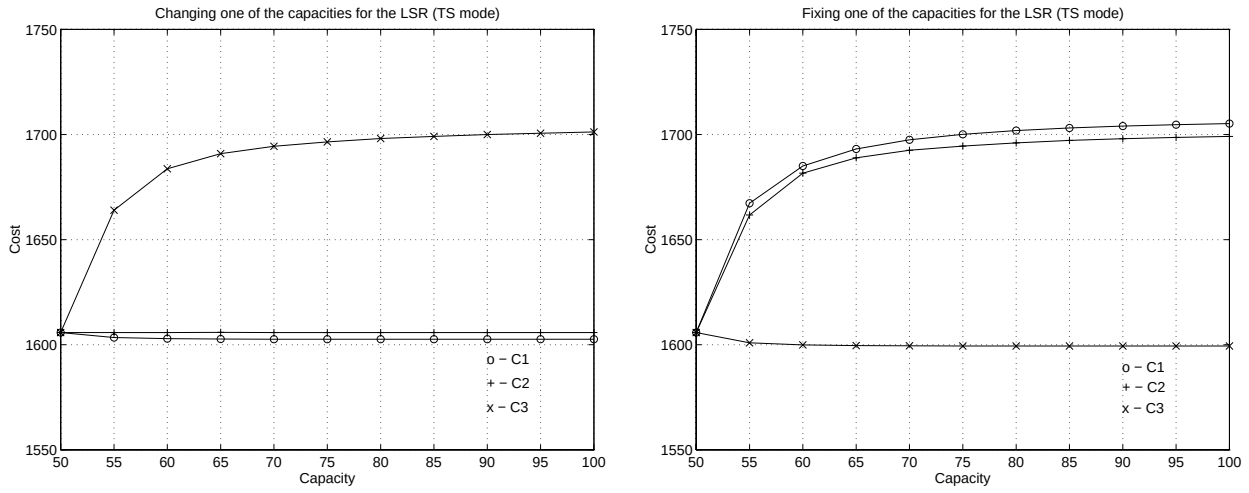


Figure 5.32: Effect of capacity along the line for the LSR.

When it comes to changing two capacities at the same time, the results are in line with the first study. If there should be an advantage for having an unbalanced production system, it occurs for the last two machines. Any other situation, where machine C^3 has higher capacity, increases costs relatively to the balanced case. That is why the only curve displaying a decrease in costs is the one where C^3 is kept fixed.

In Appendix D there are similar plots for the other two production rules. They are no different in structure relatively to those of Fig 5.32.

It is also interesting to compare the performance of the production rules to see if this type of change in capacity may change the relative performance of them. In Fig. 5.33 there are plots comparing the rules for the best configurations of capacity. The plot on the left refers to the case where only C^1 is varied and the plot on the right refers to the case where only C^3 is kept fixed.

Again, as in previous studies when using the PR, priority should be given to the product with

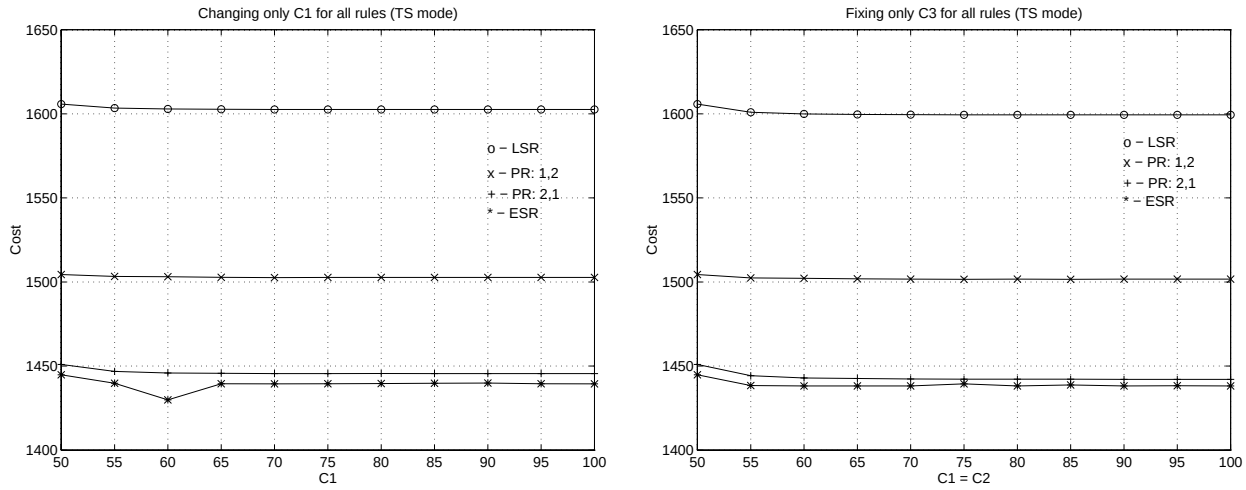


Figure 5.33: Comparison among the rules.

the lowest average demand and simultaneously with the highest penalty cost. The ESR continues to be the best rule across the range of values considered. The change in capacity for the different machines does not affect in any way the relative performance of the rules.

5.4 Conclusions

This chapter concentrated on studying a multi-product re-entrant system with capacity constraints facing random demands via simulation based optimization. It is hoped that the insights gained in Sections 5.2 and 5.3 provide guidelines into managing these systems effectively. The evidence displayed in these past sections seems to point out a clear set of conclusions that can be drawn:

- *Service versus cost performance.* There is a clear connection between optimal echelon base stock values and the service level obtained as given by Proposition 5.1.1. They are equivalent measures of performance.
- *Allocation of capacity to levels.* Unless the holding costs of earlier stages are very low, equal allocation of capacity to levels is optimal (see Fig. 5.7).
- *PS mode vs. TS mode* The best rule in TS mode outperforms the best rule in PS mode. However, if LSR is used in both modes, PS performance is better (see Fig. 5.12 and 5.23).

- *Load and holding costs impact on performance.*

  - The holding costs along the production line have a limited influence on the relative performance of the several production rules; only the cases with zero (or very low values of) holding costs introduce distortions on the best rule to apply, because the PR rule manages to raise inventory to infinite levels, thus reducing the apparent length of the production system (see Fig. 5.12).

  - The load has no effect on either the qualitative properties of the rules or in the relative performance (see Fig. 5.8)

  - As loads decrease to very low values (uniformly across products), and the variances of demand are also uniform, all production rules and capacity allocations behave similarly. For low loads, capacity bounds are less frequent so that the different dynamic and static capacity schemes play no significant role.

- *Inventory levels.* There is a subtle connection between capacity and inventory levels. For balanced systems (or when capacity of lower level is below capacity of higher level) with high loads, the optimal Δ variables tend to match the capacity of the downstream machine/level as long as the value added per operation is not too low (see Fig. 5.4 and 5.13). When the value added of one operation approaches zero, the associated optimal Δ variable tends to zero (see Fig. 5.16).

- *Performance of the PR.* From the experimental results in the Section 5.4.1 we may restrict ourselves to the following rules: (1) priority should be given to items closer to completion; (2) priority to products should be intertwined, i.e., there is a global priority to individual product types used at each level, but each level is taken care of completely before moving to the next in the list; and (3) priority at each level and stage across all products should be the same.

  - The average demand of products influences strongly the relative performance of the PR when deciding which product to assign priority. All other parameters being equal (variance of demand and penalty costs), priority should be given to products with the lowest average demand (see Fig. 5.19 on the left).

  - The penalty cost of products influences strongly the relative performance of the PR. All other parameters being equal (expected demand and demand variance), priority should be given to products with the highest penalty cost (see Fig. 5.22).
  - The variance of the demand process influences strongly the relative performance of the PR. All other parameters being equal (expected demand and penalty costs), priority should be given to products with lowest variance. Overall cost is more significantly reduced when decreasing the demand variance of products with higher expected demand (see Fig. 5.25).
  - The products that should be given priority, in general, are the ones that exhibit a combination of the following features: low mean demand, high penalty costs, low variance (see Fig. 5.31 on the right). Trade-offs have to be made in case there is a conflict and simulation can be used to evaluate the best among a set of dominant alternatives.

- *In the PS mode.*
  - When there is no strong case for a product to be given priority over the other, both the ESR and the LSR perform very well (see Fig. 5.20 on the left).
  - When there is a strong case for a product to be given priority over the other, the best choice is the PR and the LSR is a very close second (see Fig. 5.29 on the right).

- *In the TS mode.*
  - When there is no strong case for a product to be given priority over the other, the ESR is the best performer (see Fig. 5.22 on the right).
  - When there is a strong case for a product to be given priority over the other, the PR is the best performer and the ESR is a close second, but in some cases (low variance or high penalty costs) its costs may be significantly higher than those of the PR (see Fig. 5.31 on the right).

Future work will have to address non-uniform loads, yield losses, random capacity, and different control policies. The influence of some of these features will be discussed in Part III. Before closing this chapter, we will present a sample of data relative to the determination of the priority scheme when using the PR.

5.4.1 Alternate Choices within PR

The previous studies assumed a particular priority scheme for the PR, based on the evidence collected in many experiments conducted. This section presents the essentials of those studies in an attempt at justifying those choices. There is a need to differentiate between systems processing a single product and systems processing multiple products.

Single product

In a single product setting, only one priority assignment to levels was considered when operating the system in the TS mode (Section 5.2). It was then claimed that giving priority to levels closer to demand is the best among the $K!$ options. Let us look at some data, obtained for a system with $K = 3$, $M = P = 1$, and 80% load, to justify this statement. There are $3! = 6$ different priority assignments for the levels. Table 5.2 displays the optimal costs for each one of the priority assignments. The leftmost column lists the levels by decreasing order of their priority. Thus 1-2-3 stands for priority to level 1, then to level 2, and finally to level 3; similarly, 3-1-2 stands for priority to level 3, then to level 1, and finally to level 2. There is a tie for the optimal assignment, that occurs in both situations for which priority to level 3 is lowest. Any other priority assignment, where level 3 is not the last in priority, incurs a higher cost, and the highest among these costs is over three times higher than the lowest (optimal) cost. The behavior here displayed is typical of all systems for which this comparison, between different priority assignment of levels on the TS mode, was made.

Table 5.2: Optimal costs for single product.

Priority of levels	Optimal Cost
1 - 2 - 3	463.57
1 - 3 - 2	677.69
2 - 1 - 3	463.57
2 - 3 - 1	1110.14
3 - 1 - 2	701.75
3 - 2 - 1	1390.86

In general, the range of costs is usually very wide and the lowest costs are achieved whenever level K is the lowest in priority. For these situations (level K having the absolute lowest priority), if there is ever a difference in costs, the case where priority increases as the products are closer to

demand always incurs the lowest. Therefore, in all experiments in a single product setting, priority was assigned in this manner.

Multiple product

In a multiple product setting there are more options within the PR. Several experimental studies were performed. The first study serves as a benchmark and it is similar to the one just presented for the single product setting. Therefore, consider a case with $K = 3$, $M = 1$, $P = 2$, and 80% load. The average demand of product 1 is 8 and for product 2 is 12. Also, their holding and backlog costs are different. Both product demands have the same coefficient of variance.

Table 5.3: Optimal costs for multiple products.

Priority of levels	Priority of products	Optimal Cost
1 - 2 - 3	1 - 2	771.41
1 - 3 - 2	1 - 2	1003.93
2 - 1 - 3	1 - 2	771.41
2 - 3 - 1	1 - 2	1644.92
3 - 1 - 2	1 - 2	1033.68
3 - 2 - 1	1 - 2	1654.78
1 - 2 - 3	2 - 1	752.62
1 - 3 - 2	2 - 1	1031.05
2 - 1 - 3	2 - 1	752.73
2 - 3 - 1	2 - 1	1613.78
3 - 1 - 2	2 - 1	1066.89
3 - 2 - 1	2 - 1	1635.31

The overall product priority was fixed and the priority to levels was changed in the same manner as in the single product case. Therefore, a situation where the priority for levels is $\{2, 1, 3\}$ and the priority of products is $\{2, 1\}$ signifies that, in the TS mode, production decisions are taken in the order: $\{P^{212}, P^{211}, P^{112}, P^{111}, P^{312}, P^{311}\}$. That is, one decides the production amounts level by level, according to the priority to levels and on each level one uses the global product priority. Table 5.3 displays the optimal costs for the 12 different combinations. There are six different priority lists for each possible product priority.

The best assignment of priority to levels is $\{1, 2, 3\}$ no matter what is the priority of the products. There is again a tie with $\{2, 1, 3\}$, when priority is given to product one over product two. However, such a tie is no longer present for the opposite case of product priority; nevertheless,

the difference is minor. This is in line with what was observed in the single product setting.

Table 5.4: Comparison between method 1 and method 2.

Priority of levels	Priority of products	Method	Optimal Cost
1 - 2 - 3	1 - 2	1	771.41
1 - 2 - 3	1 - 2	2	786.38
1 - 2 - 3	2 - 1	1	752.62
1 - 2 - 3	2 - 1	2	783.04

There is still another priority scheme to be considered in the multiple product setting. So far, the data presented concerns studying priorities primarily between levels. Alternately, one may choose to assign priority primarily by product. That is, if priority to levels is $\{2,1,3\}$ and to products is $\{2,1\}$ the production decisions may be taken by the order $\{P^{212}, P^{112}, P^{312}, P^{211}, P^{111}, P^{311}\}$. Table 5.4 presents the comparison of this method (which will be called *method 2*) with the previous (which will be designated as *method 1*). Given the previous observation on the priorities to levels, the systems were run only for the best choice of priority for the levels.

Table 5.5: Comparison of method 1 and method 2 as the penalty cost changes.

Priority of products	b^1	b^2	Method	Optimal Cost
1 - 2	14	20	1	771.41
1 - 2	14	20	2	786.38
1 - 2	1000	20	1	1186.13
1 - 2	1000	20	2	1194.68
2 - 1	14	20	1	752.62
2 - 1	14	20	2	783.04
2 - 1	14	1000	1	1329.18
2 - 1	14	1000	2	1353.34

From this table, one concludes that method 2 is worse than method 1 for the parameters presented. We need to investigate if changing variance and penalty costs produces any qualitative change to the above conclusions. All the data presented from now on was obtained by using the list $\{1,2,3\}$ for priority to levels. Table 5.5 presents the comparison in costs for both methods as the penalty cost of either product 1 or product 2 is changed. If there should be any change, that would have to occur when the product which has highest priority also has the highest penalty cost. For this reason we have only changed the penalty cost of the product with the higher priority. The

conclusion is that the penalty cost change does not affect the previous conclusion: it is better to solve level by level rather than product by product.

Table 5.6: Comparison of method 1 and method 2 as the demand variance changes.

Priority of products	b^1	b^2	$E[d_0^1]/\sigma^1$	$E[d_0^2]/\sigma^2$	Method	Optimal Cost
1 - 2	1000	20	5	1	1	697.33
1 - 2	1000	20	5	1	2	697.77
1 - 2	2000	20	5	1	1	717.08
1 - 2	2000	20	5	1	2	718.07
2 - 1	14	1000	1	5	1	643.80
2 - 1	14	1000	1	5	2	638.57
2 - 1	14	2000	1	5	1	677.30
2 - 1	14	2000	1	5	2	669.86

Finally, Table 5.6 presents the results of changing the coefficient of variance for the demand. The systems were run by giving priority to the product with the lowest demand variance and highest penalty cost simultaneously. For product 1, the one with the lowest average demand, the previous qualitative conclusions do not change; that is, it is better to use method 1, even when the variance is lower. However, for product 2, the one with the highest average demand, method 2 is better than method 1 when the variance decreases. Note also that the differences for product 1 – first four rows of the table – are almost negligible. The differences for product 2, last four rows of the table, are a little higher, but under 1.2%. We do not present results for priority to products with lowest penalty cost or highest variance since those are always worse than the ones displayed here. If we keep decreasing the variance of product 1 we will eventually come to a point where method 2 is preferable to method 1. We ran a test case for $E[d_0^1]/\sigma^1 = 10$ and $b^1 = 1000$ where method 2 beats method 1 by a slight margin.

The data here presented, as a sample of many other experiments conducted, constitutes a strong case in favor of method 1, despite the results of this last table. This is why the data of the early sections of this chapter were obtained with method 1 and by giving higher priority to levels closer to demand, when experimenting with the PR for multiple products in the TS mode.

Part III

Non Uniform Loads and Random Yield

Chapter 6

Theoretical Foundation Revisited

This chapter deals with the extension of some of the theory presented in Chapters 3 and 4 to systems where the loads are not uniform and where random yield is present. Many of the validation results carry through trivially. However, there are some exceptions which will be discussed in detail below. A notable one being the stability issue, as the technique to establish it is drastically different from the one presented earlier. Consequently, we open this chapter with a complete stability analysis for re-entrant systems, subject to random demand, random yield, and with non uniform loads. The stability discussion will be done by phases. First, the model of Part II will be changed to accommodate non uniform loads while assuming perfect yield (Section 6.1). In the second phase, the model of Part II will be changed to accommodate random yield while keeping the uniform load assumption (Section 6.2). Finally, both random yield and non uniform loads will be pieced together, thus concluding the stability analysis (Section 6.3). After this, in Section 6.4, the validation results will be reviewed and the significant changes to the earlier results will be underlined. Section 6.5 will discuss the impact of non uniform loads into the main conclusions drawn in Chapter 5.

6.1 Non Uniform Loads and Perfect Yield

We review the models presented in Section 3.1 to accommodate this extra feature. The recursions for inventory, echelon inventory, and shortfall do not change. What changes are the specifics of the production decisions. Recall that the production expression for the LSR operated in the PS mode is

$$P_n^{kmp} = f_n^{kmp} g_n^{km}. \tag{6.1}$$

Since the production net needs, f_n^{kmp}, only depend on shortfalls and feeding inventories their expressions and derivatives do not change to include the non uniform loads. What changes is the expression for g_n^{km}, because it accounts for the impact of the net needs over the available capacity. Let us assume that every product p on level k and stage m needs τ^{kmp} units of capacity per unit of material produced. In the analysis so far it was assumed that $\tau^{kmp} = 1$ for all k, m, p. Given the inclusion of the τ^{kmp} constants, not necessarily all equal to 1, the expression for g_n^{km} becomes:

$$g_n^{km} = \min\left\{\frac{C^{km}}{\sum_p \tau^{kmp} f_n^{kmp}}, 1\right\} \tag{6.2}$$

Whereas f_n^{kmp} expresses the production net needs in terms of parts, the term $\tau^{kmp} f_n^{kmp}$ expresses those needs in terms of machine capacity. According to (6.2), its derivative will be

$$g'_{(z)}{}_n^{km} = \begin{cases} \dfrac{-C^{km} \sum_p \tau^{kmp} f'_{(z)}{}_n^{kmp}}{(\sum_p \tau^{kmp} f_n^{kmp})^2} & \text{bound by capacity} \\ 0 & \text{no bound in capacity} \end{cases} \tag{6.3}$$

This is the only change we need to take care of in order to model the existence of non uniform loads for the LSR. We skip the presentation of the changes needed for the other two production rules given the fact that, as in Chapter 4, it suffices to show stability for the LSR. Recall that the stability arguments were solely made in terms of the overall dynamic equation for shortfalls and it is possible to show that all rules follow similar such equations. Also, the model extension to the TS mode follows trivially from the discussion above.

As before, let us first address the discussion of stability for the PS mode.

6.1.1 Stability and Regeneration for Partially Shared Systems

For this setting there is no change relative to the partially shared systems with perfect yield and uniform loads by replacing $||\boldsymbol{Y}_n^{km}||$ with $||\boldsymbol{Y}_n^{km}||_\tau$, defined as

$$||\boldsymbol{Y}_n^{km}||_\tau = \sum_{p=1}^{P} \tau^{kmp} Y_n^{kmp}. \tag{6.4}$$

With this change, equation (4.11) becomes

$$\begin{aligned} ||\boldsymbol{Y}_{n+1}^{KM}||_\tau &= \max\{0, ||\boldsymbol{Y}_n^{KM} + \boldsymbol{D}_n||_\tau (1 - \frac{C^{KM}}{||\boldsymbol{Y}_n^{KM} + \boldsymbol{D}_n||_\tau})\} \\ &= \max\{0, ||\boldsymbol{Y}_n^{KM}||_\tau + ||\boldsymbol{D}_n||_\tau - C^{KM}\} \end{aligned} \tag{6.5}$$

and equation (4.13) becomes

$$\begin{aligned} ||\boldsymbol{Y}_{n+1}^{km}||_\tau &= \max\Big\{0, ||\boldsymbol{Y}_n^{km}||_\tau + ||\boldsymbol{D}_n||_\tau - C^{km}, \\ &\qquad \sum_{p=1}^{P} \tau^{kmp} \left(Y_n^{(km)^+ p} + d_n^p - (z^{(km)^+ p} - z^{kmp}) \right)^+ \Big\}. \end{aligned} \tag{6.6}$$

These dynamic equations, for the *weighted shortfall sums*, fall exactly into the framework described in Section 4.1. Therefore, the adequate stability condition becomes the following.

Theorem 6.1.1 Suppose the demands $\{\boldsymbol{D}_n, -\infty < n < \infty\}$ are ergodic as well as stationary. If

$$\mathbf{E}[||\boldsymbol{D}_0||_\tau] = \sum_{p=1}^{P} \tau^{kmp} \mathbf{E}[d_0^p] < C^{km} \quad \text{for all k, m.} \tag{6.7}$$

then the shortfall process is stable when the system is operated in the PS mode.

Proof: After performing the changes above indicated, the proof is the same as that of Theorem 4.1.3.

□

All the results presented for the PS mode in Section 4.1 are valid for this setting without change.

6.1.2 Stability and Regeneration for Totally Shared Systems

A simple observation of equations (6.5) and (6.6) helps to understand why we cannot resort to the technique used in Section 4.1, when proving stability for totally shared systems with perfect yield and uniform loads. Note that the stochastic dominance may be destroyed when production is bound by inventory. When loads are uniform, all values of $\tau^{kmp} = 1$ and stochastic dominance

follows trivially. This dominance would be maintained if the value $Y_n^{(km)^+p}$ would be multiplied by $\tau^{(km)^+p}$ in the expressions above, but it is multiplied by τ^{kmp}. In general $\tau^{kmp} \neq \tau^{(km)^+p}$.

For totally shared capacity systems we are unable to present a strong stability proof for all production rules as before. There is a need to introduce some changes on the structure of the control policies. The stability will be established by presenting a particular choice of parameters for the new control policy that yields a stable system. Given the proposed choice of parameters is feasible and induces stability, it necessarily constitutes an upper bound on the cost. The optimal parameters will have to incur lower costs. Therefore, by providing an upper bound which is stable the stability of the system will be asserted.

The main structural change on the control policies proposed is the addition of an input bound. That is, there is a need to impose a maximum amount of new material entering the production system for each product per period. Although some bound exists already, given that machines have finite capacity, this is not enough to establish stability. It is necessary to define tighter bounds. Recall that in Chapter 5 the PR performs quite poorly when the entering level has priority over the others. Recall also that the degradation of the LSR when switching from the PS mode to the TS mode was due to the fact that the potential input of new material jumped from C^{KM} to a total of C^M per period, distorting the proportions between the several levels in favor of the input of new material. This preference is given at the expense of a slower travel speed along the production line. Moreover, it was shown in Chapter 5 that in the PS mode a system with $K = 2$, $M = P = 1$, and $C^{21} > C^{11}$ improves its performance if we chop the excess capacity of level 2, making $C^{21} = C^{11}$. Then it was argued that having a higher capacity on level 2 only increases the speed at which inventory moves to the buffer feeding level 1, but does not make it move faster towards the output buffer, since level 1 is the bottleneck.

Although stability is not at risk for the cases discussed in Chapter 5, the fact that we could benefit from the existence of an input bound in such cases constitutes strong evidence favoring the definition of this richer class of control policies. Besides having the base stock variables as the control parameters, we can have the input bound as an additional control variable, thus defining a wider class of multi-echelon base stock policies. The existence of such bounds is crucial to establish stability.

With this new class of policies in mind, it is now easy to define a set of parameters that stabilizes

any of our re-entrant systems for any of the proposed production rules. Let

$$\Delta^p = \min_{k,m}\{\frac{\mathbf{E}[d_0^p]}{\sum_{i=1}^{K}\sum_{j=1}^{P}\tau^{imj}\mathbf{E}[d_0^j]}C^m\} \quad \text{for all } p = 1,\ldots,P, \tag{6.8}$$

and define $I^{(KM)^+p} = \Delta^p$ as the bound for the input of product p into the system. That is, $I^{(KM)^+p}$ is the feeding inventory of stage M and level K. Set $\Delta^{kmp} = \Delta^p$ for all k and m, except for Δ^{11p} that may assume any positive value.

Assume that the system is operated using any production rule in the TS mode. With this set of delta variables all inventory variables, except I_n^{11p}, will always be Δ^p for each product. At any level and stage, the amount

$$\sum_{k=1}^{K}\sum_{p=1}^{P}\tau^{kmp}\Delta^p \leq C^m, \tag{6.9}$$

by the definition of Δ^p. Therefore, there is never a bound in capacity and the system behaves as if there is no capacity sharing, thus being operated as if there exist P different and decoupled production systems with no re-entrance. The only bound in capacity occurs for the equality between production needs and capacity which can be seen as a no capacity bound situation, since the match is perfect.

We know that for no sharing of capacity a system is stable as long as $\tau^{kmp}\mathbf{E}[d_0^p] < C^{kmp}$. This conclusion is easily derived from the stability result for partially shared systems with single product, discussed before (Theorem 6.1.1). Slicing the capacity of machine m into $k \times p$ slots and calling each one C^{kmp}, for $k = 1,\ldots,K$ and $p = 1,\ldots,P$ and adding over all products and levels we get

$$\sum_{k=1}^{K}\sum_{p=1}^{P}\tau^{kmp}\mathbf{E}[d_0^p] < \sum_{k=1}^{K}\sum_{p=1}^{P}C^{kmp} = C^m, \tag{6.10}$$

which is the stability condition for totally shared systems with non uniform loads and perfect yield. This condition holds iff $\mathbf{E}[d_0^p] < \Delta^p$ for all $p = 1,\ldots,P$.

Having provided a set of parameters which stabilizes the production system for any production rule in the TS mode it should now be evident that the optimal set of parameters will have to incur lower costs than the costs incurred by the parameters just defined. The optimal set of parameters cannot, therefore, induce an unstable system as long as $\Delta^p > \mathbf{E}[d_0^p]$. Define the original class of

base stock policies as Π_0 and the new class introduced as Π_1. Clearly, $\Pi_0 \subset \Pi_1$ and the following result has been proven.

Theorem 6.1.2 Suppose the demands $\{\boldsymbol{D}_n, -\infty < n < \infty\}$ are ergodic as well as stationary. If (6.10) holds, then the shortfall process is stable when the system is operated in the TS mode, using class Π_1.

The regeneration and explicit regeneration times discussed earlier carry through trivially for this setting.

Remarks on the Class Π_1

For the system to be stable, the minimum amount of each product that can get through the system at any period has to be above the average demand. This is the same as saying that the bottleneck machine, the machine for which (6.9) holds in the equality, has capacity above the load imposed by the demand process.

Note that one can use any of the production rules and, in the particular case of the priority rule, one can use any arbitrary priority list without risking stability. This constitutes a strength of the class of policies introduced (recall the literature review on stability).

Moreover, the argument here used for stability allows us to drop one of the main constraints of the present model: the re-entrant structure adopted. This technique extends easily to more complex re-entrant systems where not all the products are processed by the same number of levels and not all the products visit all the machines in the same order. Such was not the case of the stability proof for systems with uniform loads, since the stochastic dominance argument relies on the fact that the shortfalls added belong to the output buffers of the same machines.

The optimal policy does not necessarily have the above bound for the entering inventory. It may be the case that, during the optimization, the solution converges to values of $I^{(KM)^+p}$ which are equal or above C^M for all $p = 1, \ldots, P$. If such is the case we may drop the explicit bound on input inventory, since being above C^M has no physical significance. The cases where the optimization procedure converges to values of $I^{(KM)^+p}$ below C^M, can be clearly identified as systems that may need such bound for the input inventory in order to remain stable. Naturally, it is not necessarily the case that all the systems for which the optimal $I^{(KM)^+p}$ is under C^M are only stabilized by

policies from class Π_1, since cost considerations are taken into account when determining such values. Note also that while policies in Π_0 are nonidling in terms of the shortfalls, such is no longer the case for policies in Π_1.

6.2 Uniform Loads and Random Yield

To accomodate random yield we simply change the dynamic equations for inventories and for shortfalls. The inventory dynamic equations now assume the form

$$I_{n+1}^{kmp} = \begin{cases} I_n^{11p} - d_n^p + \alpha_n^{11p} P_n^{11p} & m = 1 \text{ and } k = 1 \\ I_n^{kmp} - P_n^{(km)^- p} + \alpha_n^{kmp} P_n^{kmp} & \text{otherwise} \end{cases} \tag{6.11}$$

The first line refers to the depletion of inventory by the external demand at the last stage and level of production. The second line of (6.11) describes the standard evolution of an intermediate level and stage: inventory of a given level and stage is depleted by the amount engaged in production by the downstream stage and level, and is increased by the amount actually produced at the corresponding level and stage.

The multiplicative random yield, α_n^{kmp}, is assumed to be independent for each level, stage, and product. Also, it is assumed that the random yield is continuous and i.i.d. for each period taking values in the set $[0, 1]$. Demands are assumed continuous, independent across products, and i.i.d. for each product along time. Both sets of random variables, demand and yield, are assumed independent.

The shortfall dynamic equation in the presence of random yield assumes the following form:

$$Y_{n+1}^{kmp} = Y_n^{kmp} + d_n^p - \alpha_n^{kmp} P_n^{kmp} + \sum_{qr=(km)^-}^{q,r=1,1} (1 - \alpha_n^{qrp}) P_n^{qrp}, \tag{6.12}$$

where the additional summation accounts for the parts lost in the downstream machines due to the presence of random yield.

For the random yield case it is easy to show stability for single product, NS mode, with uniform or non uniform loads. To prove stability for the multiple product cases and other sharing schemes we follow the approach of Section 6.1.

6.2.1 Stability and Regeneration for Partially Shared Systems

The presence of random yield in the context of uniform loads does not change the basics of the formal result. The main difference is the explicit stability condition. Aside from that, we can repeat the same steps as in Section 4.1. Therefore, the stability condition proof will be presented and the natural extension of previous results to this situation will be listed.

Assume a system operating in the PS mode with the LSR and replace P_n^{kmp} in the dynamic equation for the shortfall variables.

$$\begin{aligned} Y_{n+1}^{kmp} &= Y_n^{kmp} + d_n^p + \sum_{qr=(km)^-}^{q,r=1,1} (1-\alpha_n^{qrp}) P_n^{qrp} - \alpha_n^{kmp} \min\{f_n^{kmp}, f_n^{kmp} \frac{C^{km}}{\sum_{p=1}^P f_n^{kmp}}\} \\ &= \max\{Y_n^{kmp} + d_n^p + \sum_{qr=(km)^-}^{q,r=1,1} (1-\alpha_n^{qrp}) P_n^{qrp} - \alpha_n^{kmp} f_n^{kmp}, \\ &\quad Y_n^{kmp} + d_n^p + \sum_{qr=(km)^-}^{q,r=1,1} (1-\alpha_n^{qrp}) P_n^{qrp} - \alpha_n^{kmp} f_n^{kmp} \frac{C^{km}}{\sum_{p=1}^P f_n^{kmp}}\} \end{aligned} \tag{6.13}$$

where $f_n^{kmp} = \min\{Y_n^{kmp} + d_n^p, I_n^{(km)^+ p}\}$.

The above dynamic equation for the shortfall variables is not as easy to deal with as it was for previous settings. Because of this, one has to proceed differently. First, the stability condition for single product systems with no re-entrance is established. Later, by the approach of Section 6.1, stability for the PS mode for multiple products will be defined. We show that the stability condition for the PS mode is

$$\sum_{p=1}^P \frac{\mathbf{E}[d_0^p]}{\prod_{q,r=1,1}^{q,r=k,m} \mathbf{E}[\alpha_0^{qrp}]} < C^{km} \quad \text{for} \begin{cases} m = 1, \ldots, M \\ \\ k = 1, \ldots, K \end{cases} \tag{6.14}$$

The indexes in $\prod_{q,r=1,1}^{q,r=k,m} h(q,r)$ signify that the factors are taken up the production line from $h(1,1)$ to $h(k,m)$. It does not mean that the iteration is taken from 1 to k and from 1 to m independently of each other.

To simplify the notation, consider a system with single product and no re-entrance in the presence of random yield and composed of M machines. Except for random yield, this is addressed by [Glasserman and Tayur, 1994, Glasserman and Tayur, 1995]; we add random yield here. For

this simplified version, we have

Theorem 6.2.1 Suppose the demand $\{d_n, -\infty < n < \infty\}$ is ergodic as well as stationary. Additionally, suppose the random yield $\{\alpha_n^m, -\infty < n < \infty\}$ is ergodic and stationary. The shortfall process is stable iff

$$\frac{\mathbf{E}[d_0]}{\prod_{i=1}^{i=m} \mathbf{E}[\alpha_0^i]} < C^m, \quad \text{for all } m = 1, \ldots, M, \tag{6.15}$$

holds for the single product system.

Proof: The dynamic equation for shortfalls will be

$$\begin{aligned} Y_{n+1}^m &= \max\{Y_n^m + d_n + \sum_{i=m-1}^{1} (1-\alpha_n^i) P_n^i - \alpha_n^m f_n^m, \\ &\qquad Y_n^m + d_n + \sum_{i=m-1}^{1} (1-\alpha_n^i) P_n^i - \alpha_n^m C^m\} \end{aligned} \tag{6.16}$$

which, by direct comparison with the equation for perfect yield leads to the following necessary and sufficient stability condition

$$\mathbf{E}[d_0 + \sum_{i=m-1}^{1} (1-\alpha_0^i) P^i] < \mathbf{E}[\alpha_0^m C^m]. \tag{6.17}$$

In the perfect yield situation, it holds that $\alpha_n^{kmp} = 1$ and it is the case that the system is stable iff $\mathbf{E}[d_n - C^m] < 0$. A similar reasoning is applied here to propose the above condition: this condition ensures the existence of a negative drift when production is bound by capacity. We need only to establish a connection between (6.15) and (6.17). To do so, we first establish a relationship between production amounts in consecutive machines.

The production of machine i is conditioned by what is effectively produced by machine $(i+1)$. What is effectively produced by machine $(i+1)$ during period n is $\alpha_n^{i+1} P_n^{i+1}$. If production starts at a point where $I_0^{i+1} = \Delta^{i+1} = z^{i+1} - z^i$ it turns out that

$$\sum_{n=1}^{N} P_n^i \leq \Delta^{i+1} + \sum_{n=1}^{N} \alpha_n^{i+1} P_n^{i+1}, \tag{6.18}$$

since machine i cannot engage more material in production than the available inventory.

Dividing the above by N and taking the limit as $N \to \infty$ we get

$$\mathbf{E}[P^i] \leq \mathbf{E}[\alpha_0^{i+1}]\mathbf{E}[P^{i+1}]. \tag{6.19}$$

Given that α_n^{i+1} is independent of P_n^{i+1}, the yield process is i.i.d., and the machines are capacitated, the limit exists and equals the expected value.

Assume now that the inequality above holds strictly. If that is the case, the inventory sitting in front of machine i, I^{i+1}, grows to infinity because

$$I_n^{i+1} = \Delta^{i+1} + \sum_{j=1}^{n}(\alpha_j^{i+1}P_j^{i+1} - P_j^i), \tag{6.20}$$

and taking the limit as $n \to \infty$ we get

$$\begin{aligned} I_\infty^{i+1} &= \Delta^{i+1} + \lim_{n\to\infty}\sum_{j=1}^{n}(\alpha_j^{i+1}P_j^{i+1} - P_j^i) \\ &= \Delta^{i+1} + \lim_{n\to\infty}\sum_{j=1}^{n}\alpha_j^{i+1}P_j^{i+1} - \lim_{n\to\infty}\sum_{j=1}^{n}P_j^i \\ &= \Delta^{i+1} + \lim_{n\to\infty} n\mathbf{E}[\alpha^{i+1}]\mathbf{E}[P^{i+1}] - \lim_{n\to\infty} n\mathbf{E}[P^i] \\ &= \infty, \end{aligned} \tag{6.21}$$

by the law of large numbers and because of the assumption on the strict inequality.

If the value of the feeding inventory for any machine grows to infinity the system is unstable. Also, if the value of the feeding inventory grows to infinity, it must be the case that production of that machine is being bound by capacity in the long run. Thus, it is established that on a stable system it must be the case that (6.19) holds at equality for all machines. It is also easy to show that

$$\mathbf{E}[P^i] \leq C^i \quad \text{for } i = 1, \ldots, M. \tag{6.22}$$

The expected production of any machine is either bounded by the expected production of the preceding machine as presented in (6.19) or is bounded by the available capacity as presented in

(6.22). That is, only one of these inequalities will hold at equality. If at least for one machine the bound occurs due to capacity, then the system is unstable, implying $\mathbf{E}[\alpha_0^1]\mathbf{E}[P^1] < \mathbf{E}[d_0]$ and the value of I^1 grows to $-\infty$. For a system to track demand, $\mathbf{E}[\alpha_0^1]\mathbf{E}[P^1]$ has to be equal to $\mathbf{E}[d_0]$.

Now, observe that if all $m-1$ stages are stable, each $\mathbf{E}[P^i]$, for $i = 1, \ldots, m-1$, can be written as a function of $\mathbf{E}[P^1]$ as follows[1]:

$$\mathbf{E}[P^i] = \frac{\mathbf{E}[P^1]}{\prod_{j=2}^{i} \mathbf{E}[\alpha_0^j]}, \tag{6.23}$$

and $\mathbf{E}[P^1] = \frac{\mathbf{E}[d_0]}{\mathbf{E}[\alpha_0^1]}$.

Proceeding by backward induction, consider first the case of $m = 1$. Expression (6.17) will reduce to

$$\mathbf{E}[d_0] < \mathbf{E}[\alpha_0^1]C^1, \tag{6.24}$$

which is exactly the same as (6.15) for $m = 1$. To prove the stability condition for stage m, let us assume that all $m-1$ downstream stages are stable. That is, assume that unstability cannot be caused by the last $m-1$ machines. Therefore, (6.17) becomes

$$\begin{aligned}
\mathbf{E}[\alpha_0^m]C^m &> \mathbf{E}[d] + \mathbf{E}[1-\alpha^{m-1}]\frac{\mathbf{E}[P^1]}{\prod_{j=2}^{m-1}\mathbf{E}[\alpha^j]} + \ldots \mathbf{E}[1-\alpha^1]\mathbf{E}[P^1] \\
&= \mathbf{E}[d] + \mathbf{E}[P^1]\left(\frac{1}{\prod_{j=2}^{m-1}\mathbf{E}[\alpha^j]} - \mathbf{E}[\alpha^1]\right) \\
&= \mathbf{E}[d] + \mathbf{E}[d]\left(\frac{1}{\prod_{j=1}^{m-1}\mathbf{E}[\alpha^j]} - 1\right) \\
&= \mathbf{E}[d]\frac{1}{\prod_{j=1}^{m-1}\mathbf{E}[\alpha^j]}
\end{aligned} \tag{6.25}$$

showing that if (6.17) holds, so does (6.15). It remains to see what happens when (6.17) does not hold.

Let us assume that (6.17) does not hold for at least one machine. Given that this is a necessary and sufficient condition for stability, it follows that the system is unstable. Therefore, it must be the case that $\mathbf{E}[P^1] < \mathbf{E}[d_0]/\mathbf{E}[\alpha_0^1]$.

[1] Using (6.19) with the equality sign.

Given that there is at least one machine violating (6.17), let m^* be the bottleneck machine of the line. That is, the machine that is furthest away from the stability region. For this machine it is the case that $\mathbf{E}[P^{m^*}] = C^{m^*}$ and for all the machines downstream it is the case that (6.19) holds at equality. Therefore, for $i = 1, \ldots, m^*$, $\mathbf{E}[P^i]$ can be expressed as a function of $\mathbf{E}[P^1]$ as described in (6.23), since there is no unstability caused by machines following the bottleneck. Inequality (6.17) for machine m^* does not hold, so

$$\begin{aligned}
\mathbf{E}[\alpha_0^{m^*}]C^{m^*} &< \mathbf{E}[d_0 + \sum_{i=m^*-1}^{1} (1-\alpha_0^i)P^i] \\
&= \mathbf{E}[d_0] + \sum_{i=1}^{m^*-1} \mathbf{E}[(1-\alpha_0^i)]\mathbf{E}[P^i] \\
&= \mathbf{E}[d_0] + \mathbf{E}[P^1] \sum_{i=1}^{m^*-1} \frac{\mathbf{E}[1-\alpha_0^i]}{\prod_{j=2}^{i} \mathbf{E}[\alpha_0^j]} \\
&< \mathbf{E}[d_0] \left(1 + \sum_{i=1}^{m^*-1} \frac{\mathbf{E}[1-\alpha_0^i]}{\prod_{j=1}^{i} \mathbf{E}[\alpha_0^j]}\right) \\
&= \mathbf{E}[d_0] \frac{1}{\prod_{j=1}^{m^*-1} \mathbf{E}[\alpha_0^j]}, \qquad (6.26)
\end{aligned}$$

showing that (6.15) does not hold for machine m^*. Thus, the equivalence between (6.15) and (6.17) is established and the result for single product follows.

□

It remains to generalize the above to the multiple product situation. By using a class of policies that imposes *bounds on production quantities* it will be possible to provide a set of parameters that ensure no sharing of capacity when the system is operated in the PS mode.

To simplify the notation, assume we are dealing with a flow line constituted by $\hat{M}$ machines and with no re-entrance. In the PS mode, set $\hat{M} = KM$. Define Ω^m as the long run expected amount of work imposed on machine m by all products. This amount is given by

$$\Omega^m = \sum_{p=1}^{P} \frac{\mathbf{E}[d_0^p]}{\prod_{j=1}^{m} \mathbf{E}[\alpha_0^{jp}]} \quad \text{for all } m = 1, \ldots, \hat{M}. \qquad (6.27)$$

Define the long run average load of machine m, for all $m = 1, \ldots, \hat{M}$, as

$$\Lambda^m = \frac{\Omega^m}{C^m}. \tag{6.28}$$

It is not difficult to see that it is necessary for all values of Λ^m to be below unity in order for the system to be stable.

Now, define as the long run bottleneck machine the one which has the highest long run average load. So, we have m^* as the machine for which

$$\Lambda^* = \frac{\Omega^{m^*}}{C^{m^*}} = \max_m\{\Lambda^m\}. \tag{6.29}$$

Define the share of each machine that can be used by each product in the long run as

$$C^{mp} = \frac{\mathbf{E}[d_0^p]/\prod_{j=1}^m \mathbf{E}[\alpha_0^{jp}]}{\Lambda^m}, \tag{6.30}$$

and set the values for Δ^{mp} that constitute the control variables for this problem as

$$\Delta^{m^+p} = \begin{cases} C^{m^*p}\prod_{j=m+1}^{m^*}\mathbf{E}[\alpha_0^{jp}] & \text{if } 1 \le m \le m^*, \\ C^{m^*p} & \text{if } m = m^*, \\ C^{m^*p}/\prod_{j=m^*+1}^{m}\mathbf{E}[\alpha_0^{jp}] & \text{if } \hat{M} > m \ge m^*. \end{cases} \tag{6.31}$$

Note that Δ^{m^+p} is the nominal inventory of product p that sits in front of machine m. That is why there is no need to define Δ^{1p}, which remains free as before[2]. The other control variables are the bounds on the input of new material per period for each product, which are

$$I^{\hat{M}^+p} = C^{m^*p}/\prod_{j=m^*+1}^{\hat{M}}\mathbf{E}[\alpha_0^{jp}]. \tag{6.32}$$

Given the fact that each value of $\Delta^{m^+p} \le C^{mp}$, it is the case that, as long as $I_n^{m^+p} \le \Delta^{m^+p}$, there is never a situation where the capacity of machine m has to be shared in the PS mode. This would always be the case if yield would be deterministic and exactly equal to its average value for all periods. Since in general $Pr(\alpha_n^{mp} > \mathbf{E}[\alpha_0^{mp}]) > 0$, we cannot ensure that the available inventory

[2]The same is true of the non negativity constraint.

for all products sitting in front of a given machine is always such that its summation is below the machine's capacity. Thus, in the PS mode, there will be periods where sharing does indeed occur and equation (6.13) would have to be used explicitly to establish stability. It was said earlier that such dynamic equation is too cumbersome to be tackled. This implies that it is not possible to derive stability just by imposing a bound on the new material entering the system as it was done in Section 6.1. It is necessary to add further features to the control policies in order to obtain an instance that ensures no sharing in the PS mode and which can constitute an upper bound on the optimal cost, while maintaining stability.

The natural extension of class Π_1, furthers the extension proposed in Section 6.1 by adding a new set of variables. These new variables impose bounds on the amount of material allowed to enter production for each product at every machine on any given period. In this way one imposes a maximum share that each product can take from each machine, even if there is available inventory to produce more. This class of control policies, which will be called Π_2, turns out to be the sensible thing to do from the practitioners' point of view as well[3].

With this broader class of base stock policies in mind, the obvious instance which ensures stability and constitutes an upper bound on the cost of the optimal solution is such that all the new variables are equal to Δ^{m^+p} as well. That is, the additional bound for machine m to produce product p is the nominal value of the associated delta variable.

As was remarked at the end of Section 6.1, it may also be the case here that the optimal values for those bounds are such that sharing will eventually occur. It should be clear that there is no intention of running these systems as P independent production lines. Doing that would signify losing the flexibility allowed by the sharing of resources. For instance, it was discussed in Chapter 5 that the best performance in the TS mode was always better than the best performance in the PS mode. The greater the flexibility the better potential use one can make of the available resources. However, it may be the case that such flexibility may need a minimum amount of restraint to ensure *fairness* for all the products. Again, the use of the bounds is only essential to establish stability for infinite horizon systems.

Thus, the above discussion established the following theorem.

Theorem 6.2.2 Suppose the demand $\{d_n^p, -\infty < n < \infty\}$ is ergodic as well as stationary. Addi-

[3]Clearly, $\Pi_0 \subset \Pi_1 \subset \Pi_2$.

tionally, suppose the random yield $\{\alpha_n^{kmp}, -\infty < n < \infty\}$ is ergodic and stationary. If equation (6.14) holds, then the shortfall process is stable for multiple product systems operated in the PS mode, using class Π_2.

We argued in terms of a flow line composed of $\hat{M}$ machines. When a re-entrant system with K levels and M machines is operated in the PS mode it is transformed into a flow line with no re-entrance, where it is possible to map each pair (km) into a global ordering for $\hat{M} = K \times M$ machines.

Once the stability condition has been established, all the other results discussed in Section 4.1 are trivially derived. Theorem 4.1.8 and the associated corollary are the exceptions. In order to characterize the regeneration times we need one additional assumption, due to the presence of random yield. Additionally to condition (4.16), the following condition has to hold so that the shortfall process returns to the origin infinitely often, with probability one

$$Pr(\alpha^{kmp} = 1) > 0, \quad k = 1, \ldots, K\ ; m = 1, \ldots, M\ ; p = 1, \ldots, P. \tag{6.33}$$

If this does not hold, then the convergence of the shortfalls to zero can only occur in infinite time, since it will be accomplished through a geometric series.

6.2.2 Stability and Regeneration for Totally Shared Systems

The stability condition for the TS mode is the natural extension of the previous condition for random yield in the PS mode.

Theorem 6.2.3 Suppose the demand $\{d_n^p, -\infty < n < \infty\}$ is ergodic as well as stationary. Additionally, suppose the random yield $\{\alpha_n^{kmp}, -\infty < n < \infty\}$ is ergodic and stationary. If

$$\sum_{k=1}^{K} \sum_{p=1}^{P} \frac{\mathbf{E}[d_0^p]}{\prod_{q,r=1,1}^{q,r=k,m} \mathbf{E}[\alpha_0^{qrp}]} < C^m \quad \text{for } m = 1, \ldots, M, \tag{6.34}$$

then the shortfall process is stable for multiple product systems operated in the TS mode, using class Π_2.

Proof: To establish this result we only need to produce an instance of class Π_2, defined in the earlier subsection. The instantiated parameters of Π_2 follow the same reasoning just presented at

the end of the previous subsection. That is, compute the average work on each machine; define the machine with the highest average load; determine the average share that each product at each level demands from the bottleneck machine; and use that share to determine the values of Δ^{kmp} and the values for the bounds on the production for all the levels, stages, and products. Given those, the system operated in the TS mode never shares capacity across products and levels. Also, every share allocated is never below the average work imposed. This implies that the P decoupled systems are all stable and the cost incurred by such control variables constitutes an upper bound on the performance of the optimal control variables.

Therefore, the optimal values of these same control variables will have to incur a lower cost and have to necessarily maintain stability. Also, the optimal values of the control variables may be such that sharing of capacity does indeed occur and the TS mode really allows a flexible use of all the available capacity as intended.

□

Taking into account the discussion on the regeneration times made at the end of the previous subsection, all the results discussed for the TS mode in Section 4.2 carry through trivially for this setting.

Remarks on the Class Π_2

The class Π_2 of modified base stock policies constitutes a similar qualitative step from Π_1 as this latter constituted from Π_0. It may be the case that while optimizing relative to the base stock levels and production bounds the optimal values are such that no sharing really occurs either in the PS or the TS mode. This only means that such is the optimal thing to do and may have no direct relation with the fact that policies from Π_1 or Π_0 may induce unstability.

Recalling the discussion of Section 2.3, and in particular the plots of the optimal switching curves, the existence of production bounds other than the net capacity may be beneficial in terms of minimizing operational costs, independent of the stability issue.

Modeling the production system by means of a periodic review inventory control turns out to allow the definition of a broad class of policies that can incorporate nonidling features in a very natural way. The lack of this feature was one of the drawbacks of other approaches, as queueing networks is one paradigmatic example.

Other modifications can be added to these policies, namely the need to impose upper bounds on the amount of inventory sitting at each buffer, which could be of advantage due to cost considerations and also to tackle the existence of machine failures. However, the modifications introduced to Π_0 to generate Π_2 are the minimal needed to establish stability.

Note also that, when controlling systems with random yield, deciding to produce the exact difference between a target value and the present value of inventory is known to be non-optimal. Other classes of policies would have to be proposed in order to eventually achieve better performances. Namely, inflating each current production decision by the reciprocal of the expected random yield would be a good candidate for a first approximation, although this is also known to be non-optimal. This type of generalizations are outside the scope of the present thesis and are only here referred to clarify that there is no substantial claim on the class Π_2 other than it may allow lower costs than Π_0, it ensures stability for the re-entrant systems addressed by this thesis, and even ensures stability for more complex re-entrant systems as mentioned in Section 6.1.

6.3 Non Uniform Loads and Random Yield

Given the discussion of the previous two sections, the stability results for this setting are

Theorem 6.3.1 Suppose the demand $\{d_n^p, -\infty < n < \infty\}$ is ergodic as well as stationary. Additionally, suppose the random yield $\{\alpha_n^{kmp}, -\infty < n < \infty\}$ is ergodic and stationary. If

$$\sum_{p=1}^{P} \tau^{kmp} \frac{\mathbf{E}[d_0^p]}{\prod_{q,r=1,1}^{q,r=k,m} \mathbf{E}[\alpha_0^{qrp}]} < C^{km} \quad \text{for} \begin{cases} m = 1, \ldots, M \\ \\ k = 1, \ldots, K \end{cases} \tag{6.35}$$

then the shortfall process is stable for multiple product systems operated in the PS mode, using class Π_2.

Theorem 6.3.2 Suppose the demand $\{d_n^p, -\infty < n < \infty\}$ is ergodic as well as stationary. Additionally, suppose the random yield $\{\alpha_n^{kmp}, -\infty < n < \infty\}$ is ergodic and stationary. If

$$\sum_{k=1}^{K} \sum_{p=1}^{P} \tau^{kmp} \frac{\mathbf{E}[d_0^p]}{\prod_{q,r=1,1}^{q,r=k,m} \mathbf{E}[\alpha_0^{qrp}]} < C^{m} \quad \text{for } m = 1, \ldots, M, \tag{6.36}$$

then the shortfall process is stable for multiple product systems operated in the TS mode, using class Π_2.

6.4 Validation for Non Uniform Loads and Random Yield

This section discusses the validation of the IPA algorithm for the general case with random yield and non-uniform loads. For simplicity on the presentation, only the LSR case will be formally presented for the PS mode, skipping the details of the other production rules and other modes of operation.

Unlike the simpler case of uniform loads, the IPA validation does not carry through for all production rules, once the LSR in the PS mode is validated. For example, it is not valid for the PR in the TS mode. So, after discussing the result for the specifics of the LSR, the details of the PR operating in the TS mode will be described. Later, the result for the LSR in the PS mode will be extended to the other rules and to the TS mode. The cases for which the extension is not valid will be discussed.

The production decision for the LSR in the PS mode when random yield is present and the loads are non uniform is as presented in (6.1). The production net needs are as in (3.5) and their derivatives as in (3.22). The capacity constraint is expressed as in (6.2) and its derivative as in (6.3).

With this in mind we have the following result for finite horizon and using class Π_0.

Theorem 6.4.1 If $\{d_n^p, n = 1,2,\ldots, p = 1,2,\ldots,P\}$ and $\{\alpha_n^{kmp}, n = 1,2,\ldots, k = 1, 2, \ldots, K, m = 1,2,\ldots,M, p = 1,2,\ldots,P\}$ are independent, each d_n^p has a density on $(0,\infty)$, and each α_n^{kmp} has a density on $(0,1)$ then the following hold:

1. For $k = 1,\ldots,K, m = 1,\ldots,M, p = 1,\ldots,P$, and $n = 1,2,\ldots$, each I_n^{kmp} and E_n^{kmp}, as given by (6.11) and (6.12) respectively is, w.p.o., differentiable at $(z^{111},\ldots,z^{KMP})$ with respect to each z^{qrs}, $q = 1,\ldots,K, r = 1,\ldots,M$, and $s = 1,\ldots,P$. Moreover, these derivatives satisfy the obvious extensions for non uniform loads and random yield of (3.19) and (3.20), respectively. Also for

   a. *Linear Scaling Rule with Partial Sharing*

      P_n^{kmp} as given by (6.1), where the individual factors are given by (3.5) and (6.2), is also differentiable w.p.o. and its derivative satisfies (3.31), where the individual factors are given by (3.22) and (6.3), respectively;

2. If in addition $\mathbf{E}[d_n^p] < \infty$ for all n, then $\mathbf{E}[I_n^{kmp}]'_{(z)}$, $\mathbf{E}[E_n^{kmp}]'_{(z)}$, and $\mathbf{E}[P_n^{kmp}]'_{(z)}$ exist and equal $\mathbf{E}[(I'_{(z)})_n^{kmp}]$, $\mathbf{E}[(E'_{(z)})_n^{kmp}]$, and $\mathbf{E}[(P'_{(z)})_n^{kmp}]$.

Proof: It should not be difficult to see that the inclusion of the constants τ^{kmp} in the inventory and echelon inventory equations, together with the random yield effects does not change the substance of the reasoning presented for the proof of Theorem 3.4.6. The only qualitative change refers to the situations under which the theorem is not valid due to the occurrence of non differentiable points with non zero probability. Taking one of the examples used for the above referenced proof, a non differentiable point will now be attained on $\tau^{kmp}(z^{(km)^+p} - z^{kmp}) = C^{km}$ (see the remark below). It should be obvious how to reduce this case to a fully differentiable one.

□

Remark: In the presence of random yield, the optimal base stock variables do not satisfy relations such as $(z^{(km)^+p} - z^{kmp}) = C^{km}$, that hold in the perfect yield case. In the experiments conducted for random yield, the optimal base stock variables are such that $(z^{(km)^+p} - z^{kmp}) > C^{km}$. The system tends to *prefer* that the difference between consecutive base stock values is a little over the capacity of the machine, attempting, we believe, to hedge against the uncertainty caused by the random yield.

The above result extends easily for the other two production rules in the PS mode as well as for the LSR in the TS mode.

6.4.1 The Singularity of the PR in the TS Mode

The production decision with non uniform loads will assume the following form for the PR in the TS mode.

$$\begin{aligned}
P_n^{k(1)mp(1)} &= \min\{f_n^{k(1)mp(1)}, \frac{C^m}{\tau^{k(1)mp(1)}}\} \\
P_n^{k(2)mp(2)} &= \min\{f_n^{k(2)mp(2)}, \frac{C^m - \tau^{k(1)mp(1)} P_n^{k(1)mp(1)}}{\tau^{k(2)mp(2)}}\} \\
&\vdots \\
P_n^{k(K\times P)mp(K\times P)} &= \min\{f_n^{k(K\times P)mp(K\times P)}, \frac{C^m - \sum_{i=1}^{K\times P-1} \tau^{k(i)mp(i)} P_n^{k(i)mp(i)}}{\tau^{k(K\times P)mp(K\times P)}}\}
\end{aligned} \tag{6.37}$$

The derivatives of the above production decisions are

$$
\begin{aligned}
P'_{(z)_n}{}^{k(1)mp(1)} &= \begin{cases} 0 & \text{bound by capacity} \\ f'_{(z)_n}{}^{k(1)mp(1)} & \text{otherwise} \end{cases} \\
P'_{(z)_n}{}^{k(i)mp(i)} &= \begin{cases} -\sum_{j=1}^{i-1} \frac{\tau^{k(j)mp(j)}}{\tau^{k(i)mp(i)}} P'_{(z)_n}{}^{k(j)mp(j)} & \text{bound by capacity} \\ f'_{(z)_n}{}^{k(i)mp(i)} & \text{otherwise,} \end{cases}
\end{aligned}
\tag{6.38}
$$

for $i = 1, \ldots, \min\{i^*, K \times P\}$, with i^* the level and stage above which all decisions are zero, due to a possible capacity bound.

It turns out that the above recursions for the derivatives, although they seem innocuous, cannot be used as they are shown in equation (6.38). In fact, when loads are non uniform, irrespective of yield, it may be the case that the above expression generates values that grow exponentially as the simulation evolves.

In what follows we will explain why and how an exponential growth can be generated with the above expressions, when that growth does not take place for the other production rules, and why we cannot use the gradient information so generated.

To understand why we can get exponential growth with the above expressions note first that, for any particular machine with total sharing of capacity, a bound in capacity for a particular product at any given level depends linearly on the derivatives of higher priority levels of the same product. Moreover, such dependence is proportional to the load ratio between the higher priority levels and the level that is bound by capacity. Assume that such ratio is above one for at least one higher priority level and let that ratio be $(1 + a)$, with $a > 0$. Assume also that the derivative of the production decision for that higher priority level is $b \neq 0$, possibly by a bound in inventory. The amount produced by the level bound by capacity will have one term on its derivative given by $-(1 + a)b$ and it will move down the line until it gets to be the input of the higher priority level. At that moment, the production of this higher priority level, if it is not bound by capacity, will necessarily be bound by inventory and therefore its derivative will be that of the inventory itself, that is, one of its terms will be $-(1 + a)b$. If again the same lower priority level gets bound by capacity, one of the terms of its production derivative will be $(1 + a)^2 b$. If this process goes on like this for a set of consecutive periods the derivative will grow in modulus at a rate of $(1 + a)$ and therefore it will grow exponentially.

Example 6.4.2 To better illustrate the concept take a simple system with $K = 2$ and $M = P = 1$, such that $\tau^{211} = 1$ and $\tau^{111} = 2$. Suppose capacity is 15, $z^{211} = 20$ and $z^{111} = 15$. Suppose also the system starts from the base stock values[4], priority is given to P_n^{111}, and that demand is fixed at 10 each period during the first $L/2$ periods, is 1 on period $L/2+1$, and zero after that. On the first period, the value of P_1^{111} will be bound by inventory, I_0^{211}, and its derivative with respect to z^{211} is 1. The value of P_1^{211} will be bound by capacity and therefore its derivative is going to be $-P'^{111}_1\tau^{111}/\tau^{211} = -2$ (from (6.38)).

On the second period, the inventory levels will be $[5, 10]$. Again the value of P_2^{111} will be bound by inventory, but the derivative of this inventory with respect to z^{211} is now -2 because the available inventory on period two is exactly the value of P_1^{211}. The decision P_2^{211} will be bound by capacity and, as a consequence, its derivative will be $-P'^{111}_2\tau^{111}/\tau^{211} = 4$. The inventory entering period 3 is $[5, 5]$.

At the end of the $L/2$ periods we will have the following inventory levels entering period $L/2 + 1$: $[5, 15 - 5L/2]$. The derivative of $I_{L/2}^{211}$ with respect to z^{211} is $(-2)^{L/2}$. It will take the system $L/2 + 1$ more periods[5] to bring the shortfalls to zero and at the end of the L periods the derivative for I_L^{211} will be $(-2)^L$. The moment the shortfall reaches zero, at the end of period $L + 1$ the derivatives of the inventory variables will be reset due to the regeneration of the inventories themselves.

Defining *busy period* as the time that elapses between two regeneration points, the example above shows that a busy period will yield derivatives for some variables that grow exponentially with the period size. This is not to say that those derivatives are not *correct*. They are correct only for very small perturbations of the base stock variables. In what follows we specify what is meant by a derivative being correct.

In the formulation presented so far we have only imposed physical meaning on the recursions of the state variables and trusted that such meaning would be preserved for the derivative recursions. It was imposed that the intermediate inventories are always positive variables, and consequently that the production decisions are always positive and bounded by the available inventories. The derivative recursions also have a physical meaning. If we change any of the base stock variables by

[4] $[I_0^{211}, I_0^{111}] = [5, 15]$.
[5] Counting with period $L/2 + 1$.

an amount δ, the state variables or production decisions will change proportionally to δ according to the value specified by the derivatives. Since there are bounds of variation for state and decision variables (non negativity, for one), it is the case that those derivative recursions are only valid for values of $\delta \in [-\bar{\delta}, \bar{\delta}]$ such that, when changing one base stock variable by any amount with modulus above $\bar{\delta}$, we force some state variable to go out of its physically meaningful bounds. Regarding the derivative recursions generated by expression (6.38), it is the case that the value of $\bar{\delta}$, being inversely proportional to the modulus of the derivatives, may be negligible. This creates numerical difficulties for the optimization procedure because it will be impossible to move out of the initial solution given that a feasible step along the gradient direction may be too small to be represented in a computer.

Moreover, it turns out that a slight change in one of the base stock levels may incur a radical change in the cost. As a consequence, being the case that the derivatives are so high, and being the case that for finite choices of the base stock variables the cost function is always finite (for stable systems and/or finite horizon), it must be the case that the cost function is not smooth as a function of the base stock variables. Additionally to this, the size of the busy period is highly sensitive to very small changes of the base stock variables.

Experimental evidence

To better illustrate the above ideas, let us look at some experimental results that fully describe the behavior obtained with the PR in the TS mode. The setting is that of non uniform loads and perfect yield.

The data displayed here concerns the usage of the PR in the TS mode for a system with $K = 3$, $M = P = 1$. The capacity is $C^1 = 50$, the average demand is $\mathbf{E}[d_0^1] = 10$, and $\tau^{111} = 2$, $\tau^{211} = \tau^{311} = 1$.

All the plots displayed in Figs. 6.1–6.3 were taken around the same nominal point and along the same direction. The direction used was an estimate of the gradient obtained through simulation. The plots on the right are zooms of the ones on the left. The difference in the three figures has to do with the different step size used on the successive estimates of the cost.

The explanation for this behavior is simple. For any finite choice of base stock variables, or Δ variables, the overall cost is finite. This is due to the fact that demand has a bounded expected

value. It is the case that the derivatives at any given point have high absolute values, many orders of magnitude above the cost itself. Therefore, the only way to ensure bounded costs with derivatives many orders of magnitude above them is by having a cost function that oscillates.

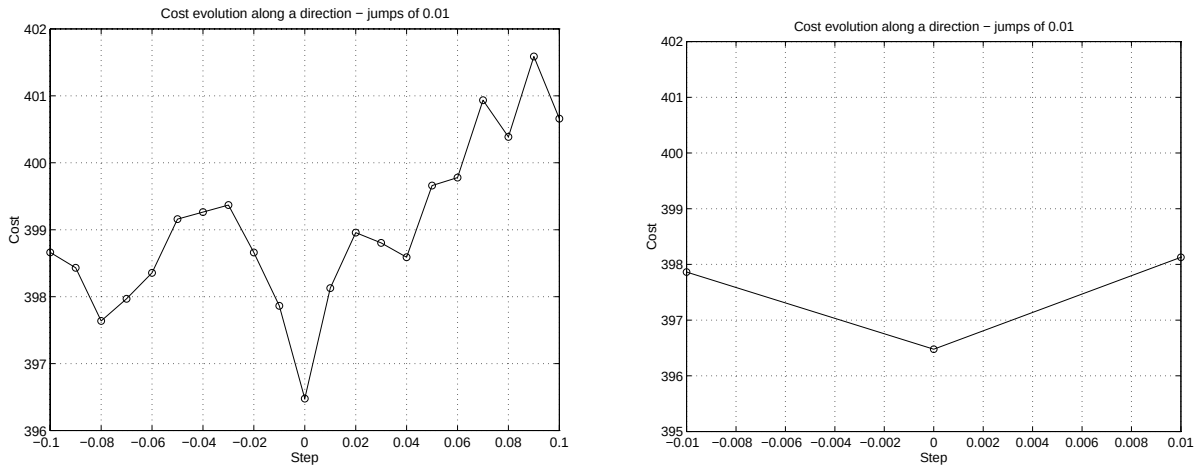


Figure 6.1: Plot of cost using a step size of 0.01.

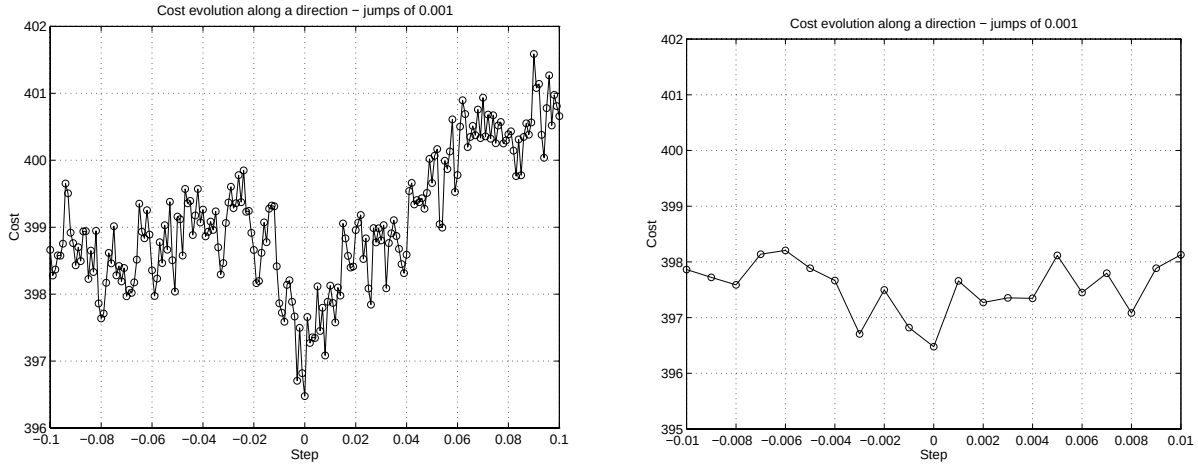


Figure 6.2: Plot of cost using a step size of 0.001.

This is not to say that the cost function is not differentiable. At any given finite choice of the control variables there is a sufficiently small neighborhood for which the derivatives are valid. The point is that once we get out of the neighborhood the derivatives will change very rapidly and will change signs equally fast, thus ensuring a bounded cost, but working against the possibility of using

a gradient based optimization procedure.

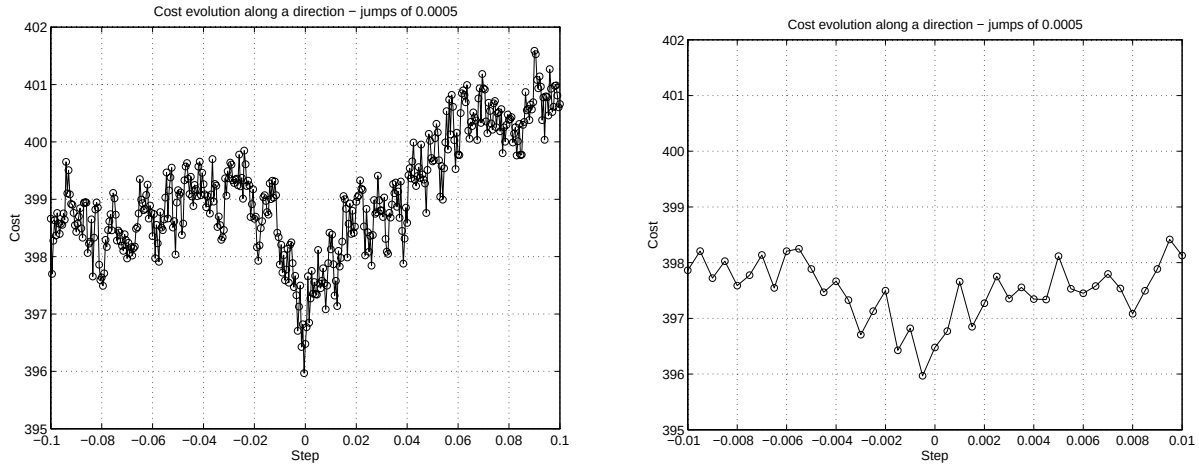


Figure 6.3: Plot of cost using a step size of 0.0005.

If for a given busy period the derivative at the end of it is say x, for a busy period with one more period in size it will be $-ax$, with $a > 1$. So, the change in signs occurs in very small neighborhoods of the control variables, due to the fact that the size of the busy period is very sensitive to those changes.

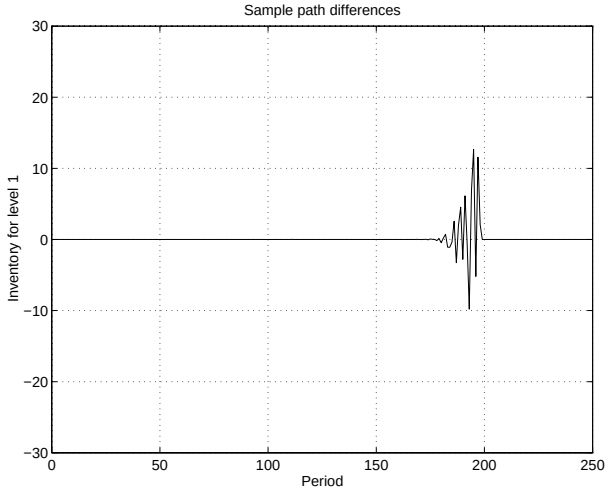


Figure 6.4: Change on I_n^{111} from sample path #1 to sample path #2.

The next set of figures (Figs. 6.4–6.6) displays the comparison between two sample paths. The sample paths have a length of 250 periods and were taken with base stock variables that differ

from each other by 10^{-5}, that is, if z_i is the base stock vector for sample path $i = 1, 2$, then $||z_1 - z_2|| = 10^{-5}$.

Each figure shows the actual difference between the inventories for both paths. So, they should be seen as a confirmation on the correctness of the derivative estimates.

Note also that this behavior of the PR in the TS mode is due to the non uniform loads. It did not occur before because all $\tau^{kmp} = 1$.

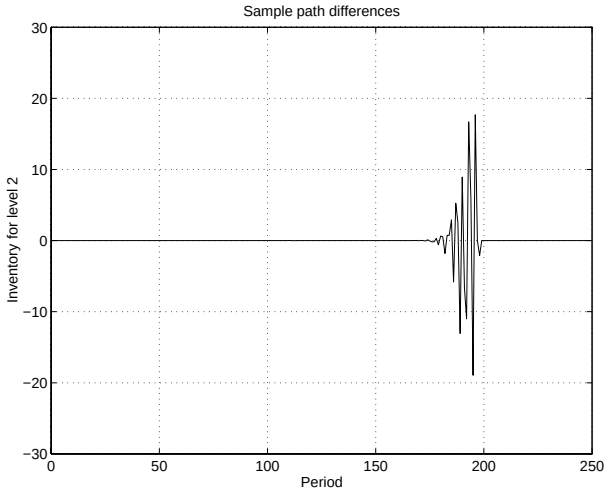


Figure 6.5: Change on I_n^{211} from sample path #1 to sample path #2.

6.4.2 The Other Production Rules

Let us now address the issue of exponentially growing derivatives for the other production rules and other capacity sharing modes. The question is to investigate if whether or not the behavior just described in Example 6.4.2 and illustrated in the above figures may happen in any of the other production rules. It turns out that such behavior is not exclusive of the PR in the TS mode. Recall that the exponential growth is due to the fact that a particular upstream level of production is bounded by capacity during a string of periods and that the downstream levels of the same product will as a consequence be bounded by inventory and retroactively impose the capacity bound.

Such effect cannot take place in the partial sharing mode, no matter what production rule is being used. The reasoning is simple: in the partial sharing mode, downstream levels cannot induce capacity bounds on upstream levels because they do not share the same capacity slots;

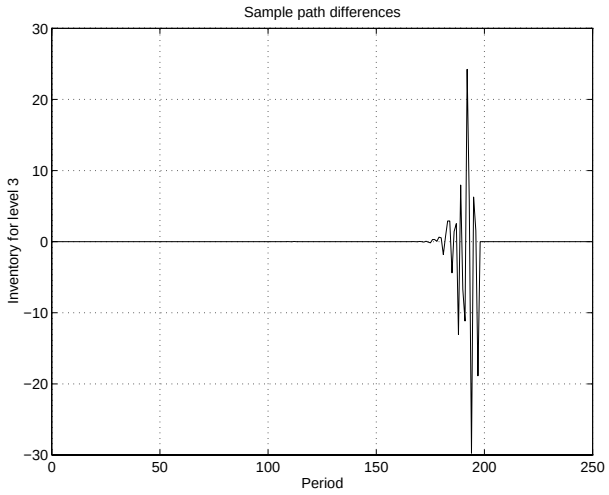


Figure 6.6: Change on I_n^{311} from sample path #1 to sample path #2.

therefore, although an upstream level that is bound by capacity may induce bounds in inventory for downstream levels, those downstream level decisions have no effect on the upper levels; for partial sharing of capacity, the production decisions on a given level only depend on the level itself and on past decisions for upper levels. So, we are left with the total sharing cases.

> **Remark:** The above reasoning is only valid for perfect yield. The presence of random yield induces feedback effects, since parts lost downstream will have to be produced upstream again. However, they are not of the same type as the ones described above.

Take the LSR in the TS mode. It is impossible for any level to be bound by capacity alone. As the production decision is taken, either all of the decisions are bound by capacity or none is bound by capacity. The importance of this is that if a bound in capacity for an upstream level later induces a bound in inventory for a downstream level, this bound in inventory cannot occur at the same time as a bound in capacity for that upstream level due to the scaling. So the coupling does not occur.

Regarding the Equalize Shortfall Rule, it is possible that a bound in capacity will occur for a single level and product, unlike the LSR. However, for the exponential growth to occur it is necessary that during a busy period a downstream level is always bound by inventory and the upstream is bound by capacity. For this to happen in the ESR it is necessary that the downstream

level has a shortfall consistently higher than that of the upstream level, and that it is always bounded by inventory.

This can happen in the ESR and in fact there are situations were the derivatives grown exponentially. We have been able to produce one such case for $K = 5$, $M = P = 1$, with $\tau^{111} = 10$ and all the others equal to 1. However, there is a striking difference between the ESR and the PR. While for the PR, no matter what is the starting point, the optimization locks in a close neighborhood of the starting point, for the ESR this behavior seems to depend on the starting point.

The system above described converges to the optimal solution if the starting delta variables are sufficiently high (around $C^1/\sum_k \tau^{k11}$). However, if the starting delta variables are very small there is no convergence.

Given that the exponential growth is due to bounds in inventory that keep the ordering of the shortfalls the same, when the starting point is very low on the delta variables those bounds in inventory occur and lead to exponentially growing derivatives.

Although IPA is not valid, we have been able to use ESR without any problems. That is, the theoretical possibility exists but the occurrences of systems and/or parameters for which the exponential growth occurs appears to be very rare. Some systems have the right structure to produce exponential derivatives, but during optimization it is very rare to encounter such numerical difficulties. For the PR, such growth is always present as long as the parameters are chosen according to the structure described earlier.

6.5 Experimental Studies

Experiments similar in scope to the ones of Chapter 5 were conducted for systems with non uniform loads and, in some cases, random yield was also included. Although the class of policies that should be used is Π_2, all the experiments were done with class Π_0. This thesis does not present the validation of the IPA approach for class Π_2. Also, in none of the systems studied was it ever necessary to use any sort of bounds to ensure stability. The exceptions were the two machine unstable system of [Lu and Kumar, 1991] and the three machine unstable system of [Dai and Weiss, 1996]. For these systems it was impossible to obtain bounded costs using class Π_0, but the costs became bounded simply by means of using policies from class Π_1 and keeping the same priority lists as in the original papers. However, as these systems do not possess the re-entrant structure

adopted here and neither does this thesis address the validation of class Π_1, their discussion will be limited to these observations.

The optimality condition continues to hold when the loads are non uniform, random yield is present, and class Π_0 is used.

The main conclusions presented in Section 5.4 do not change much when moving from uniform loads to non uniform loads. *The optimal allocation of capacity to levels in the PS mode becomes proportional to the loads imposed on each level*, rather than being equally divided. This is a natural and expected extension. The optimal capacity slots, when the holding costs of early stages are very close to zero, exhibit a behavior similar to the one discussed in Section 5.2.1.

The structure of the holding costs along the production system and the different loads continue to have no significant impact over the relative performance of the the production rules.

Regarding the influence of the average demand on the performance of the PR, its influence is now replaced by the concept of *expected load*. That is, whereas before priority should be given to products with lower expected demand (all other things being equal), now priority should be given to products with lower expected load. This particular property is very difficult to identify for generic systems with non uniform loads. One given product may impose a lower expected load in late stages of production and impose a higher expected load in early stages. Additionally to this, we could not get any results on the PR for the TS mode, for reasons already explained.

Although, the ESR possesses the same theoretical shortcomings as the PR for the TS mode, we were able to get some results with it. In the TS mode, the ESR performs better than the LSR. There should be cases for which the PR performs better than the ESR, but we found these impossible to determine.

Given the fact that the experimental data is so similar for non uniform loads and given the fact that class Π_0 may be less adequate for systems with random yield, we will skip repeating figures and tables that are not substantially different from the ones presented in Chapter 5. In Appendix D we present graphics relative to some of the studies conducted.

Chapter 7

Conclusions and Future Research

This thesis proposed a framework to manage re-entrant flow lines producing multiple products. It concentrated the analysis on a simple (and implementable) set of capacity management schemes and production rules as a first step towards understanding broader classes of systems. The re-entrant lines were modeled as discrete time capacitated multi-product production/inventory systems, operating under modified multi-echelon base stock policies. At the beginning of any period, production decisions have to be made constrained by available inventory and capacity. Several capacity sharing mechanisms were discussed and some production rules to manage capacity both from dynamic and static points of view were proposed.

Since these systems are too complex to handle analytically, the study used simulation optimization. To study the properties of optimal policies within the classes proposed, an Infinitesimal Perturbation Analysis approach was validated in Chapter 3. A set of recursions that describe the dynamics of the state variables and production decisions was provided and their derivative recursions validated. The IPA approach was validated for finite horizon performance measures such as operation costs and Type-1 service level and their respective derivatives. To validate the infinite horizon measures and their derivatives, stability conditions were rigorously established in Chapter 4. By doing so, the thesis provided a framework where simulation can work as an optimization tool to derive the optimal parameters of the control policies proposed.

Once the general framework that supports the utilization of IPA was established, the thesis presented a series of computational studies that allowed the extraction of insights about how to manage re-entrant systems and provided a series of interesting structural properties. Moreover, the study provided substantial hints about the structure of the true optimal polices, which are useful

for determining better sub-optimal approximations. See Section 5.4 for a detailed summary of the main conclusions drawn from the experimental data.

One key advantage of the IPA technique, together with the modeling paradigm proposed in this thesis, is the fact that it does provide a good basis for studying more general control policies, as long as their description can be done by a well determined and unambiguous set of parameters.

The stability discussion made in Chapter 6 places the emphasis on determining stable policies rather than determining conditions under which a given policy induces stability. One of the elegances of the stability discussion is that it agrees with some of the insights produced by the experimental data and works concurrently with them. Therefore, the classes of policies that ensure stability provide an important contribution of this thesis for further research. Although some of the features of the richer policies are not particularly new nor unexpected their study is still relatively insignificant. That has to do with the complexity of those policies in terms of analytical analysis. However, as long as a general tool like IPA can be used, their study becomes an easier task to undertake.

Although this thesis does not present any type of formal validation for the policies proposed in Chapter 6, nor presents any computational study on them, the validation of Chapter 3 provides the strong intuition that their validation can be obtained straightforwardly. There should not be any major technical shortcomings on future developments, given the robustness of the approach and the simplicity of the extensions needed, regarding the original model (see discussion below).

The thesis concentrates on simple re-entrant flow lines, but we can state *a set of very general conclusions useful to the wider problem of production control in wafer fabs.* Production control involves two types of decisions: input control and flow control. *Input control* concerns deciding the type and amount of new material that should be allowed into the system at any given moment. *Flow control* concerns the type and amount of material allowed into the next operation.

The main insight provided by the stability discussion together with the experimental data analysis is the fact that there is a definite advantage in controlling production with idling policies. Even when backlogs are high there should be some restraint on the amounts of new material entering the system and on the amounts of material allowed to move to the next operation. Much of the research in the past has concentrated on nonidling policies for intuitive reasons: "a machine should not be kept idle when there is work to do". This thesis clearly states that this intuition is not

rigorous and *nonidling policies should be avoided for non acyclic systems, multiple products, non uniform loads, and random yield.*

Regarding the dynamic capacity management schemes, the main conclusion for *uniform loads* is that the *Equalize Shortfall Rule achieves the best performances across a wide range of parameters.* The option to use the Priority Rule should only be made if there is a clear hierarchy of products in terms of expected demand, demand variance, and penalty costs. The *Priority Rule outperforms the Equalize Shortfall Rule for sure when it is possible to unambiguously order the products with these three combined parameters*, i.e., when the product with the lowest expected demand has lowest variance and highest penalty cost.

In terms of static capacity allocation, *the best performances are achieved with the greater flexibility*, i.e., when the overall capacity is shared by all products in different processing stages. Also, if a system is unbalanced it should be run as a fully balanced system. By unbalanced we mean a system where the machines are subject to different loads. *The excess capacity of upstream machines should be ignored*, whereas there may be a marginal advantage on having increasing capacities as we move from the first machine in the line to the last one.

Setting the adequate holding and penalty costs is usually a difficult task. In general, it is hard to measure the exact value added by a given operation. It is also hard to measure the exact impact of backlog in terms of cost. However, noting that the significant influence of intermediate holding costs has to do with inventory distribution along the line, it will be possible to *assign those holding costs so that the inventory levels along the line are relatively moderate and balanced.* If early holding costs are very close to zero, it is likely that some early buffers will have high inventory quantities.

Moreover, noting the equivalence between penalty costs and service level estalished by Proposition 5.1.1 and knowing that *what matters for the performance is the relative proportion between holding and penalty costs*, it is easy to assign an overal set of holding and backlog costs. Defining a target service level is an easier task than measuring exactly the value added by each operation.

Given the theoretical difficulties of the PR and the ESR (described in Section 6.4.1), *the natural choice for systems with non uniform loads will have to be the Linear Scaling Rule.* The bad performances observed for this rule in the TS mode can be easily improved with idling versions of it, as the experimental results seem to support. The question is the fact that the LSR does not treat products differently from each other as much as one would desire in some circumstances.

Namely, what if there is a clear reason to prefer a product over the other? Are we to scale their needs linearly and simply hope that the right setting of penalty cost parameters will force the right inventory levels? Or is there something else that should be done[1]?

7.1 Future research

There is a broad set of issues that are worthy of further investigation. Probably, the main one has to do with considering control policies from class Π_2, i.e., a multiechelon base stock policy with bounds on production amounts for each operation. A first priority coming out of the discussion along the thesis would be to validate the IPA approach for this broad class of policies. It would be interesting to investigate the impact it has in terms of costs even for systems where its utilization is not required by stability considerations. The experimental evidence from this thesis and some theoretical and experimental work done by other authors seems to point in this direction.

Eventually, the LSR in the TS mode may see its performance significantly improved to the point of being competitive with the other two rules. This is of special importance due to the formal shortcomings of these two production rules, as pointed at the end of Chapter 6.

The shortcomings of the PR and the ESR seem to point out that it is not a good idea to dynamically assign capacity in ways that may take care of a product at a time. The LSR takes care of all products simultaneously and has no theoretical problems. Therefore, we need a production rule that takes care of all products simultaneously but takes into consideration that some of those products should be preferred over the others.

With these considerations in mind, the natural next step would be to convert the desired priority list into a set of weights and dynamically allocate capacity with a *Weighted Scaling Rule*. Assume that those weights are real numbers and can be optimized through an IPA procedure (or some other optimization tool). Then, the combinatorial problem of setting a priority list would be reduced to the non linear programming problem of finding a set of real valued weights.

Besides class Π_2 there are other classes of policies that require attention. As the discussion of Chapter 2 illustrated, there are other desirable features on a production control policy. Namely, the ability to deal with machine uncertainty. Any of the three policy classes discussed along the thesis is able to deal with a certain type of machine variability, but none of them is sufficiently sound to

[1] Section 7.1 provides more details on this.

deal with any type of machine variability. In what follows we present a brief characterization of this.

It can be said that there are two basic types of machine uncertainty to consider: the *short term variation* and the *long term variation*.

Short term variation of machines refers to machine uncertainty in terms of processing times, which translates into random capacity. We modeled the capacity of each machine as being a constant value. It would be possible to model it as being a random variable, in line with the approach of [Ciarallo et al., 1994]. It should be noted that the inclusion of this feature does not pose any major difficulty for class Π_0. Both the validation of the IPA and the stability discussion of Part II carry through trivially. Note that the stability condition derived in Theorem 4.1.3 is equivalent to

$$\mathbf{E}[||\boldsymbol{D}_0|| - C^{km}] < 0 \quad k = 1, \ldots, K; m = 1, \ldots, M. \tag{7.1}$$

The requirement for the existence of a negative drift when the capacity is a random variable translates into the expected demand to be below the expected capacity. Recall that we used this same argument in Section 6.2 when presenting inequality (6.17).

One could argue that by considering random capacities we could take care of any sort of machine variability. However, such an approach may not apply to long term variation, which results from machine breakdowns. This type of variation occurs less frequently in time and when it does it remains for relatively long periods. A machine may be down for several periods, whereas when it is up its capacity may oscillate from period to period, as discussed above.

Incorporating both these types of variation at the same level would imply a very high variance process for the machines' capacity. The effects of such a high variance process, due to the inclusion of failures, would impact the costs of the attained optimal controls since the safety stocks would naturally be higher than they really need be. Therefore, a two level approach, in line with [Kimemia, 1982], would have to be pursued. The optimization procedure would have to be run for each possible state of the machines in order to determine the optimal parameters. Each calculation would take into account the higher frequency disturbances caused by the short term variation of the machines.

Besides this hierarchical framework, it should be obvious that none of the three classes of policies addressed in this thesis is good enough. Given that the three classes concentrate on shortfalls and

they possess no bounds for local inventories, it would follow that in case a given machine would fail for a long period, all the machines upstream would still be producing to reduce shortfall and restore the echelon base stock. The three classes of policies lack the existence of a blocking feature to prevent this occurrence, in line with [Wein, 1988] or [Glassey and Resende, 1988]. To ensure that local inventories will not grow needlessly once a particular machine is down for a long enough period it is necessary to impose bounds for local inventories.

This calls for another class of control policies that differs from Π_2 by the existence of maximum values for the local inventories. Note that the approach of [Wein, 1988] defines this type of bound only in terms of the amount of work headed to the bottleneck machine. We are proposing here the definition of such bounds for all the machines, products, and levels. From the modeling perspective, this extra bound is easily incorporated and the validation of the IPA should not pose any major difficulties. However, the moment one considers this new class of policies the questions of stability will have to be re-evaluated.

One other aspect that was not fully addressed in this thesis was the design of control policies more adequate to deal with systems subject random yield. The general strategy has to incorporate some sort of *order amplification* to compensate for yield losses. The IPA technique is nothing but a good tool to investigate the possible benefits of policies that will order more than effectively needed to compensate for random yield, as long as a control policy is described by a simple set of parameters and the overall model ensures the adequate smoothness properties.

Appendix A

Validation Related Proofs

This appendix includes some of the proofs on auxiliary results, skipped in Chapter 3 that were omited then to avoid loosing sight of the essentials and because some of them are relatively trivial extensions of other results published.

Their inclusion here is intented at making this document the more self contained possible.

Derivation of Equation 3.26

$$\begin{aligned}
\bar{v}_N = \mathbf{E}[\bar{V}_N] &= E\left[P^{-1}\sum_{p=1}^{P}\left(N^{-1}\sum_{n=1}^{N}1\{d_n^p = 0 \text{ or } d_n^p \leq I_n^{11p}\}\right)\right] \\
&= P^{-1}\sum_{p=1}^{P}E\left[N^{-1}\sum_{n=1}^{N}1\{d_n^p = 0 \text{ or } d_n^p \leq I_n^{11p}\}\right] \\
&= P^{-1}\sum_{p=1}^{P}E\left[N^{-1}\sum_{n=1}^{N}\mathbf{E}[1\{d_n^p = 0 \text{ or } d_n^p \leq I_n^{11p}\}|I_n^{11p}]\right] \\
&= P^{-1}\sum_{p=1}^{P}\left(N^{-1}\sum_{n=1}^{N}Pr(d_n^p = 0) + \mathbf{E}[N^{-1}\sum_{n=1}^{N}\Phi_n^p(I_n^{11p})]\right)
\end{aligned}$$

Proof of Lemma 3.5.1

Taking expectations, we get

$$\begin{aligned}
\mathbf{E}[\tilde{C}_{\alpha,L}] &= \mathbf{E}[\sum_{n=1}^{\infty}C_n\mathbf{1}\{L \geq n\}] \\
&= \mathbf{E}[\sum_{n=1}^{\infty}C_nP(L \geq n)]
\end{aligned}$$

$$= \quad \mathbf{E}[\sum_{n=1}^{\infty} C_n \alpha^n] = c_{\alpha,\infty}.$$

□

Proof of Theorem 3.5.4

We detail the case of c_∞, the argument for $\bar{v}_\infty$ being exactly the same. It follows from the Harris recurrence of $\{X_n, n \geq 1\}$ that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}[C_n] = c_\infty, \tag{A.1}$$

for all initial states. Similarly, via Lemma 3.5.3, we get the existence of

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}[C'_n] = c'_\infty, \tag{A.2}$$

For the rest of the proof, we incorporate the value of z^* as an explicit argument; e.g., we write $C_n(z)$ for the value of C_n when $z^* = z$. From Theorem 3.4.7 we know that, for all z_1, z_2,

$$\mathbf{E}[C_n(z_2)] - \mathbf{E}[C_n(z_1)] = \int_{z_1}^{z_2} \mathbf{E}[C'_n(z)]dz; \tag{A.3}$$

the function $\mathbf{E}[C_n(.)]$ is the integral of its derivative because it is Lipschitz (since $C_n(.)$ is). Now take infinite horizon time averages of both sides. The left side converges to $c_\infty(z_2) - c_\infty(z_1)$. On the right side we get

$$\lim_{N\to\infty} N^{-1} \sum_{n=1}^{N} \int_{z_1}^{z_2} \mathbf{E}[C'_n(z)]dz = \int_{z_1}^{z_2} \lim_{N\to\infty} N^{-1} \sum_{n=1}^{N} \mathbf{E}[C'_n(z)]dz \tag{A.4}$$

for all z_1, z_2. But this means that, at almost every z, $c'_\infty(z)$ exists and is given by the integrand on the right.

□

Appendix B

Stability Related Proofs

This appendix includes some of the proofs on auxiliary results, skipped in Chapter 4 that were omited then to avoid loosing sight of the essentials and because some of them are relatively trivial extensions of other results published.

Their inclusion here is intented at making this document the more self contained possible.

Proof of Lemma 4.1.2

Define

$$\begin{aligned}
&\Phi_1 = \Phi, \\
&\Phi_n(\boldsymbol{Y}, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_n) = \Phi_{n-1}\left(\Phi(\boldsymbol{Y}, \boldsymbol{D}_1), \boldsymbol{D}_2, \ldots, \boldsymbol{D}_n\right), \\
&\\
&\phi_1 = \phi, \\
&\phi_n(\boldsymbol{Y}, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_n) = ||\Phi_n(\boldsymbol{Y}, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_n)|| = \phi_{n-1}\left(\Phi(\boldsymbol{Y}, \boldsymbol{D}_1), \boldsymbol{D}_2, \ldots, \boldsymbol{D}_n\right),
\end{aligned}$$

$n = 2, 3, \ldots$, with Φ and ϕ as in Lemma 4.1.1. Then

$$||\boldsymbol{Y}_n|| = \phi_n(\boldsymbol{Y}_0, \boldsymbol{D}_0, \ldots, \boldsymbol{D}_{n-1}), \text{ a.s.} \tag{B.1}$$

Each ϕ_n is increasing and continuous.

For integer i, define ${}^i\boldsymbol{Y}_0$ such that $||{}^i\boldsymbol{Y}_0|| = 0$ and

$$||{}^i\boldsymbol{Y}_n|| = \phi_n(\vec{0}, \boldsymbol{D}_{i-n}, \ldots, \boldsymbol{D}_{i-1}), \quad n \geq 1. \tag{B.2}$$

That is, $||{}^i\boldsymbol{Y}_n||$ is the nth-period total shortfall for a process starting at the origin a time $i - n$. Therefore, if $||\boldsymbol{Y}_0|| = 0$, then $||{}^i\boldsymbol{Y}_n||$ has the distribution of $||\boldsymbol{Y}_n||$, due to the stationarity of $\{\boldsymbol{D}_n\}$.

Moreover, since ϕ is increasing,

$$\begin{aligned} ||^i\boldsymbol{Y}_{n+1}|| &= \phi_{n+1}(0, \boldsymbol{D}_{i-n-1}, \ldots, \boldsymbol{D}_{i-1}) \\ &= \phi_n(\phi(0, \boldsymbol{D}_{i-n-1}), \boldsymbol{D}_{i-n}, \ldots, \boldsymbol{D}_{i-1}) \\ &\geq \phi_n(0, \boldsymbol{D}_{i-n}, \ldots, \boldsymbol{D}_{i-1}) \\ &= ||^i\boldsymbol{Y}_n||. \end{aligned} \tag{B.3}$$

This means that, for each i, $||^i\boldsymbol{Y}_n||$ increases almost surely to a limit as $n \to \infty$. Denote this limit by $||\tilde{\boldsymbol{Y}}_i||$. Notice that

$$\begin{aligned} ||^{i+1}\boldsymbol{Y}_n|| &= \phi\left(\Phi_{n-1}(\vec{0}, \boldsymbol{D}_{i-n+1}, \ldots, \boldsymbol{D}_{i-1}), \boldsymbol{D}_i\right) \\ &= \phi(^i\boldsymbol{Y}_{n-1}, \boldsymbol{D}_i). \end{aligned} \tag{B.4}$$

Letting n increase and using the continuity of ϕ, we conclude that

$$||\tilde{\boldsymbol{Y}}_{i+1}|| = \phi(\tilde{\boldsymbol{Y}}_i, \boldsymbol{D}_i) \tag{B.5}$$

for all i. For the last assertion in the lemma, notice (as above) that $||^0\boldsymbol{Y}_n||$ has the same distribution as $||\boldsymbol{Y}_n||$ if $||\boldsymbol{Y}_0|| = 0$, so that if $\{||^0\boldsymbol{Y}_n||, n \geq 0\}$ increases almost surely to $||\tilde{\boldsymbol{Y}}_0||$, then the distribution of $\{||\boldsymbol{Y}_n||, n \geq 0\}$ increases to that of $||\tilde{\boldsymbol{Y}}_0||$.

□

Proof of Theorem 4.1.3

The proof follows a reasoning similar to the one used in [Glasserman and Tayur, 1994] to prove their Theorem 1 by using here equations (4.11) and (4.13).

For level K and stage M the total shortfall process $\{||\boldsymbol{Y}_n^{KM}||, n \geq 0\}$ follows a Lindley recursion, (4.11). It follows from Loynes' analysis of the single-server queue that if $\mathbf{E}[||\boldsymbol{D}_0||] < C^{KM}$ then $||\tilde{\boldsymbol{Y}}_0^{KM}|| < \infty$, a.s., whereas if $\mathbf{E}[||\boldsymbol{D}_0||] > C^{KM}$ then $||\tilde{\boldsymbol{Y}}_0^{KM}|| = \infty$, a.s..

The proof proceeds by induction on the levels and stages from (K, M) down to 11, assuming that (4.14) holds. Suppose $||\tilde{\boldsymbol{Y}}_0^{km}||$ is finite, a.s.. To show that the same must be true of $||\tilde{\boldsymbol{Y}}_0^{(km)^-}||$, we argue that if $||\tilde{\boldsymbol{Y}}_0^{(km)^-}|| = \infty$, then we would have $\mathbf{E}[||\boldsymbol{D}_0||] \geq C^{(km)^-}$. Observe, first, that if $||\tilde{\boldsymbol{Y}}_n^{(km)^-}|| = \infty$, then so is $||\tilde{\boldsymbol{Y}}_{n+1}^{(km)^-}||$. In other words, the event $\{||\tilde{\boldsymbol{Y}}_n^{(km)^-}|| = \infty\}$ is invariant under a shift in the time index and must therefore have probability zero or one (by the ergodicity of demands).

Now we use the random variables $||^i\boldsymbol{Y}_n||$ defined in Lemma 4.1.2. As shown there, $||^i\boldsymbol{Y}_{n+1}|| \geq ||^i\boldsymbol{Y}_n||$, a.s., for all n and i. Moreover, $||^i\boldsymbol{Y}_{n+1}||$ has the same distribution as $||^{i+1}\boldsymbol{Y}_{n+1}||$, so $\mathbf{E}[||^{i+1}\boldsymbol{Y}_{n+1}|| - ||^i\boldsymbol{Y}_n||] \geq 0$; this holds, in particular, for the $(km)^-$-th component:

$$\mathbf{E}[||^{i+1}\boldsymbol{Y}_{n+1}^{(km)^-}|| - ||^i\boldsymbol{Y}_n^{(km)^-}||] \geq 0. \tag{B.6}$$

From (B.4) we know that $||^{i+1}\boldsymbol{Y}_{n+1}|| = \phi(^i\boldsymbol{Y}_n, \boldsymbol{D}_i)$. So, $||^{i+1}\boldsymbol{Y}_{n+1}^{(km)^-}|| - ||^i\boldsymbol{Y}_n^{(km)^-}||$ is the increase in the echelon-$(km)^-$ total shortfall due to demand D_i, and therefore cannot exceed $||\boldsymbol{D}_i||$. Thus,

$$||^{i+1}\boldsymbol{Y}_{n+1}^{(km)^-}|| - ||^i\boldsymbol{Y}_n^{(km)^-}|| \leq ||\boldsymbol{D}_i||, \text{ for all } n \geq 0. \tag{B.7}$$

If every C^{km} is infinite, then the conclusion of the Theorem is immediate; suppose then that some C^{km} is finite. Then $\mathbf{E}[||\boldsymbol{D}_i||] < \infty$, so a consequence of Fatou's lemma and (B.7) is

$$\mathbf{E}[\limsup_{n\to\infty} ||^{i+1}\boldsymbol{Y}_{n+1}^{(km)^-}|| - ||^i\boldsymbol{Y}_n^{(km)^-}||] \geq \limsup_{n\to\infty} \mathbf{E}[||^{i+1}\boldsymbol{Y}_{n+1}^{(km)^-}|| - ||^i\boldsymbol{Y}_n^{(km)^-}||], \tag{B.8}$$

and, by (B.6), this is non negative. Now if $||\tilde{\boldsymbol{Y}}_0^{(km)^-}||$ is infinite while $||\tilde{\boldsymbol{Y}}_0^{km}||$ is finite, then

$$\begin{aligned}
\limsup_{n\to\infty}\{||^{i+1}\boldsymbol{Y}_{n+1}^{(km)^-}|| - ||^i\boldsymbol{Y}_n^{(km)^-}||\} &= \limsup_{n\to\infty} \max\Big\{0, ||^i\boldsymbol{Y}_n^{(km)^-}|| + ||\boldsymbol{D}_i|| - C^{(km)^-}, \\
&\qquad \sum_{p=1}^{P}\Big(^iY_n^{kmp} + d_i^p - (z^{kmp} - z^{(km)^-p})\Big)^+ \Big\} - ||^i\boldsymbol{Y}_n^{(km)^-}|| \\
&= ||\boldsymbol{D}_i|| - C^{(km)^-},
\end{aligned}$$

implying $\mathbf{E}[||\boldsymbol{D}_i||] - C^{(km)^-} \geq 0$. Thus, if in fact $\mathbf{E}[||\boldsymbol{D}_i||] < C^{(km)^-}$, then $||\tilde{\boldsymbol{Y}}_0^{(km)^-}||$ must be finite with probability one.

Conversely, suppose that $\mathbf{E}[||\boldsymbol{D}_0||] > C^{km}$ and let (k,m) be the earliest level and stage for which this holds. From (4.13) we see that $||\boldsymbol{Y}_{n+1}^{km}|| \geq ||\boldsymbol{Y}_n^{km}|| + ||\boldsymbol{D}_n|| - C^{km}$, and similarly

$$||^0\boldsymbol{Y}_{n+1}^{km}|| \geq \sum_{r=1}^{n+1}(||\boldsymbol{D}_{-r}|| - C^{km}). \tag{B.9}$$

Hence, letting n increase to ∞,

$$||\tilde{\boldsymbol{Y}}_0^{km}|| \geq \limsup_{n\to\infty} \sum_{r=1}^{n+1}(||\boldsymbol{D}_{-r}|| - C^{km}), \tag{B.10}$$

and this is ∞ when $\mathbf{E}[||\boldsymbol{D}_0||] - C^{km} > 0$. For qr such that level q and stage r occur after level k and stage m, notice that

$$\begin{aligned}
\boldsymbol{Y}_n^{qr} &= \boldsymbol{Y}_n^{(qr)^+} - (\boldsymbol{z}^{(qr)^+} - \boldsymbol{z}^{qr}) + \boldsymbol{I}_n^{qr} \\
&= \boldsymbol{Y}_n^{((qr)^+)^+} - (\boldsymbol{z}^{((qr)^+)^+} - \boldsymbol{z}^{(qr)^+}) + \boldsymbol{I}_n^{(qr)^+} - (\boldsymbol{z}^{(qr)^+} - \boldsymbol{z}^{qr}) + \boldsymbol{I}_n^{qr} \\
&= \boldsymbol{Y}_n^{((qr)^+)^+} - (\boldsymbol{z}^{((qr)^+)^+} - \boldsymbol{z}^{qr}) + \boldsymbol{I}_n^{(qr)^+} + \boldsymbol{I}_n^{qr} \\
&\vdots \quad \vdots \quad \vdots \\
&= \boldsymbol{Y}_n^{km} - (\boldsymbol{z}^{km} - \boldsymbol{z}^{qr}) + \sum_{s,t=q,r}^{(km)^-} \boldsymbol{I}_n^{st}
\end{aligned}$$

for all n, which leads to

$$\begin{aligned}
||\boldsymbol{Y}_n^{qr}|| &= ||\boldsymbol{Y}_n^{km}|| - (||\boldsymbol{z}^{km}|| - ||\boldsymbol{z}^{qr}||) + \sum_{s,t=q,r}^{(km)^-} ||\boldsymbol{I}_n^{st}|| \\
&\geq ||\boldsymbol{Y}_n^{km}|| - (||\boldsymbol{z}^{km}|| - ||\boldsymbol{z}^{qr}||)
\end{aligned}$$

because $||\boldsymbol{I}_n^{st}|| \geq 0$ for all n, s, and t.

From this we can conclude that $||\tilde{\boldsymbol{Y}}_0^{qr}|| = \infty$ if $||\tilde{\boldsymbol{Y}}_0^{km}|| = \infty$.

□

Proof of Theorem 4.1.5

The proof follows exactly the same reasoning as that of Theorem 2 in [Glasserman and Tayur, 1994].

It suffices to show that for all $||\boldsymbol{Y}_0||$, the process $\{||\boldsymbol{Y}_n||, n \geq 0\}$ eventually coincides with a copy started at zero when both are driven by the same demands. Notice that $||\boldsymbol{Y}_n^{km}||$ is always at least as large as the corresponding component of a copy started at zero. Since $||\boldsymbol{Y}_n^{KM}||$ follows a Lindley recursion withe negative drift, it hits zero at a finite time N_{KM}. Subsequently, it coincides with the $(KM)-$th component of the process started at zero. Suppose now that for all $n \geq N_{km}$, $(||\boldsymbol{Y}_n^{km}||, \ldots, ||\boldsymbol{Y}_n^{KM}||)$ coincides with the corresponding components started at zero. We claim that for some almost-surely finite $N_{(km)^-} \geq N_{km}$,

$$||\boldsymbol{Y}_{N_{(km)^-}}^{(km)^-}|| = \max\left\{0, \sum_{p=1}^{P}\left(Y_n^{kmp} + d_n^p - (z^{kmp} - z^{(km)^- p})\right)^+\right\}; \tag{B.11}$$

this will provide the coupling time for $||\boldsymbol{Y}^{(km)^-}||$ since $||\boldsymbol{Y}^{km}||$ has already coupled. Suppose there is no such $N_{(km)^-}$. Then

$$||\boldsymbol{Y}_{n+1}^{(km)^-}|| = ||\boldsymbol{Y}_n^{(km)^-}|| + ||\boldsymbol{D}_n|| - C^{(km)^-} \tag{B.12}$$

for all $n \geq N_{km}$, implying that $\liminf_n ||\boldsymbol{Y}_n^{(km)^-}|| = -\infty$, since $\mathbf{E}[||\boldsymbol{D}_0||] < C^{(km)^-}$. This is impossible, because the shortfalls are always non negative, so (B.11) must indeed occur in finite time. Subsequently, $||\boldsymbol{Y}^{(km)^-}||$ coincides with the copy started at zero. We conclude by induction that there is an N_{11}, finite a.s., such that the entire vector $||\boldsymbol{Y}_n||$ couples with the initially zero process at time N_{11}. From this it follows that $||\boldsymbol{Y}_n|| \Rightarrow ||\tilde{\boldsymbol{Y}}_0||$ since $||\tilde{\boldsymbol{Y}}_0||$ is the limit in distribution when $||\boldsymbol{Y}_0|| = 0$.

Uniqueness follows. If $||\hat{\boldsymbol{Y}}_0||$ is stationary then $||\hat{\boldsymbol{Y}}_n||$ couples with $||\tilde{\boldsymbol{Y}}_n||$ in finite time, implying that they must have the same distribution.

□

Proof of Theorem 4.1.8

If $\mathbf{E}[||\boldsymbol{D}_0||] < C^{km}$, then $P(||D_0|| < C^{11}) > 0$. Consequently, under the conditions of the theorem there exists an ϵ with $\epsilon < \min_{k,m} C^{km}$ and $\epsilon/P \leq \min_{k,m,p}(z^{kmp} - z^{(km)^- p})$ such that

$\delta \stackrel{\Delta}{=} P(d_0^1 \leq \epsilon/P, \ldots, d_0^P \leq \epsilon/P) > 0$. Since $\boldsymbol{Y}$ has a finite stationary distribution, there exists a constant $b > 0$ such that the set $B_b \subseteq \mathrm{R}^{K \times M}$ defined by

$$B_b = \{(y^{11}, \ldots, y^{KM}) : 0 \leq y^{kmp} \leq b/P, k = 1, \ldots, K; m = 1, \ldots, M; p = 1, \ldots, P\} \tag{B.13}$$

is visited infinitely often by $\boldsymbol{Y}$. We will show that there exists an integer $r \geq 0$ and a real q such that

$$P_x(||\boldsymbol{Y}_r|| = 0) \geq q > 0 \quad \text{for all } x \in B_b, \tag{B.14}$$

from which it follows that $\boldsymbol{Y}$ visits 0 infinitely often.

If $d_0^p \leq \epsilon/P$, then either $||\boldsymbol{Y}_1^{KM}|| = 0$ or $||\boldsymbol{Y}_1^{KM}|| \leq ||\boldsymbol{Y}_0^{KM}|| - (C^{KM} - \epsilon)$. Thus, every time a demand for all products falls in $[0, \epsilon/P]$, the echelon-KM shortfall is decreased by at least $C^{KM} - \epsilon$, until it reaches zero. Starting in B_b, it takes at most $r_{KM} = \lceil b/(C^{KM} - \epsilon) \rceil$ consecutive such demands to drive that shortfall to zero. Thus, with $q_{KM} = \delta^{r_{KM}}$, we have $P_x(||\boldsymbol{Y}_{r_{KM}}^{KM}|| = 0) \geq q_{KM}$ for all $x \in B_b$.

Suppose now that $||\boldsymbol{Y}_0^{(km)^+}||, \ldots, ||\boldsymbol{Y}_0^{KM}|| = 0$ for some (k, m) and that $\boldsymbol{Y}_0^{kmp}|| \leq b/P$, for all $p = 1, \ldots, P$. With probability at least δ^n, shortfalls $(km)^+, \ldots, (K, M)$ will remain at zero for the next n transitions. Moreover, for any n, if $||\boldsymbol{Y}_n^{(km)^+}|| = 0$ and $||\boldsymbol{Y}_n^{km}|| > 0$, then the inventory $I_n^{(km)^+ p}$ available for use by stage (k, m) is greater or equal to $(z^{(km)^+ p} - z^{kmp})$, for all $p = 1, \ldots, P$, being it the case that the inequality holds for at least one product, because of (4.15). Thus, if $||d_n^p|| \leq \epsilon/P$, stage (k, m) cannot be constrained by inventory, and either $||\boldsymbol{Y}_{n+1}^{km}|| = 0$ or $||\boldsymbol{Y}_{n+1}^{km}|| \leq ||\boldsymbol{Y}_n^{km}|| - (C^{km} - \epsilon)$. If we set $r_{km} = \lceil b/(C^{km} - \epsilon) \rceil$ then, with probability at least $q_{km} = \delta^{r_{km}}$. $||\boldsymbol{Y}^{km}||$ is driven to zero in r_{km} steps. We conclude that with probability at least $q = q_{11} \cdots q_{KM}$, $||\boldsymbol{Y}_{r_{11} + \cdots + r_{KM}}|| = 0$ for any $\boldsymbol{Y}_0 \in B_b$.

□

Appendix C

Optimization Procedure and Experiments

This appendix describes the optimization procedure and details the experiments conducted. Each simulation run provides a cost estimate for the present setting of parameters and an estimate of the cost gradient with respect to the parameters describing the control policy used.

C.1 Optimization Procedure

One simulation run only provides first order derivatives. The use of first order derivatives in large scale non linear programming problems has a very slow convergence. In order to speed up the convergence we used a discrete step version of the BFGS algorithm. See [Bazaraa and Shetty, 1979] for a description of the BFGS algorithm and discrete step optimization algorithms.

This algorithm uses gradient and cost information to build an estimate of the inverse of the Hessian matrix so that a measure of local curvature can improve the search during the optimization. We detail the specifics of the our implementation for the Δ variables.

Let Δ_i denote the value of the Δ variables for iteration i, $J(\Delta_{i-1})$ denote the cost estimate at the end of a simulation run for the previous iteration, $\nabla J(\Delta_{i-1})$ denote the cost gradient, $B(\Delta_{i-1})$ the estimate of the inverse Hessian, and δ_i the step size for iteration i. Therefore,

$$\Delta_i = \Delta_{i-1} - \delta_i \frac{\nabla J(\Delta_{i-1}) B(\Delta_{i-1})}{||\nabla J(\Delta_{i-1}) B(\Delta_{i-1})||}. \tag{C.1}$$

If $J(\Delta_i) < J(\Delta_{i-1})$ the iteration is called a *success*. Otherwise, it is called a *failure*. In the case

of a success, a new estimate of the inverse Hessian is computed and the step size for next iteration is increased by 5%. In the case of a failure, the simulation is run for

$$\Delta_{i+1} = \Delta_{i-1} - \delta_{i+1} \frac{\nabla J(\Delta_{i-1}) B(\Delta_{i-1})}{||\nabla J(\Delta_{i-1}) B(\Delta_{i-1})||}, \tag{C.2}$$

where $\delta_{i+1} = \delta_i/2$.

As long as we keep getting better costs the step size increases slightly at each iteration, as it is increasing its confidence that it is moving in the right direction. The 5% increase was chosen through experimentation. We do not want a step size which grows too fast nor we want a step size that grows too slow.

If we have a sequence of 4 failures in a row we drop the second order information contained in matrix B, and perform a simple steepest descent step from the point reached in the last success using the step size of the first failure of the sequence. The second order estimation is resumed as soon as we get the first success iteration after this. For the steepest descent step there is no limit in the number of failures. So we keep cutting the step size in half until the first success.

The reason to have this decoupling step after a given number of failures has to do with the fact that for some choices of parameters, the cost function is non-differentiable at optimality. Close to these points the curvature information provided by matrix B is meaningless. See [Luenberger, 1973] for details on this steepest descent step strategy in the context of fully differentiable optimization.

The stopping criteria is based on the value of the step size. When the step size becomes lower than a pre-specified amount the optimization algorithm stops. We experimented using as stopping criteria the cost function variation between two successes. Given that the cost function is usually very flat around optimality we found that for a given stopping error we would get more precise information for service level measures if the stopping is based on the step size. The service level measure is more sensitive to step size than is the operational cost measure.

If the search direction combined with the current step size is such that one of the Δ variables becomes negative, the procedure replaces that negative value by a small positive value. This value is proportional to the current step size to ensure that the given variable is able to converge to zero but to prevent it to be strictly equal to zero. A zero value for a Δ variable is a situation were the cost function is non-differentiable with nonzero probability. For this case, Lemma 3.4.3 does not hold and IPA cannot be carried out.

C.2 Details of the Experiments

We tested several simulation run lengths to determine what would be the a good number of periods simulated trading off the variance of the estimates and the computational time. All the simulation runs correspond to a horizon of 21,000 periods. Each system is simulated for 1,000 periods before starting the collection of data to ensure that stationarity has been achieved.

The stopping step size was set at 10^{-6}. To improve the efficiency of the simulation, all random numbers are generated up-front and stored for further use in each iteration of the optimization procedure. Using the same sample path in simulation based optimization ensures lower variance and faster convergence for the optimal parameters. See [L'Ecuyer, 1994] for a discussion on variance reduction procedures.

The information contained in an input file describes the dimensions of the system, which rule to use, which capacity sharing mode, expected demand, coefficient of variance, priority list (used if needed), units of capacity needed per unit of product at each operation, capacities, starting control parameters, holding and backlog costs, stopping error, initial step size, number of periods simulated, number of periods not used for computing cost and derivatives, which variables to optimize, maximum string of failures after a success in order to drop second order information, and yield parameters.

The output file contains the same format of the input file plus additional information about number of iterations performed, initial cost, final cost, optimal variables, and service level estimate. This output file can be used as an input file for another optimization. This feature is of particular importance given that an input file may contain a smaller subset of the information needed and the program will generate the remaining parameters randomly, like holding and backlog costs as an example. So, it is convenient to store the randomly generated features for future use.

C.3 Non-differentiability

It was observed before that the cost function is non-differentiable at some singular surfaces. In the experiments conducted there were cases were the optimization procedure would stop in one of those surfaces. In all of these cases there was some Δ variable (or a set) that was either close to zero or close to a capacity slot (or a multiple of a capacity slot). In all of these cases, the values

achieved were correct in the sense that there was a sound explanation for those variables to have such singular values.

However, in some of those cases the optimization procedure stopped short of satisfying the optimality condition. As discussed in Chapter 3, problems with this optimal structure can be reduced to fully differentiable problems by imposing the singular values *a priori* and taking derivatives with respect to the remaining variables.

This was done in many of those cases and the findings were that the gain in cost was relatively marginal, regarding the original optimization cycle. That is, although designed for differentiable problems, the version of the BFGS algorithm used was able to get very close to the actual optimal solutions in cases were there was non-differentiability at optimality.

Naturally, future work in these problems will have to address the issue of non-differentiability optimization explicitly. *Bundle methods* are an example of some of the methods available for this type of optimization problems. See [Lemaréchal, 1989] for a review on optimization methods for non-differentiable performance measures.

Another issue that may be raised on this discussion is the validity of Lemma 3.4.3 for these cases. The case were some Δ variable converges to zero was already discussed. The situation that remains is for the cases were it converges to the value of a capacity slot. It should be noted that each variable is free to be above and below a capacity slot, unlike the case of zero value. Therefore, during optimization it is the case with probability one that the actual value of each given Δ variable is never exactly equal to a capacity slot although it may converge to it. Therefore, at any given iteration differentiability is preserved almost always and the lemma supporting the IPA procedure holds.

Appendix D

Complementary Plots

This appendix includes some plots that complement the data discussed in Chapter 5 and some others that support the brief summary of Section 6.5.

D.1 Optimal Base Stock Values

Figures D.1-D.4 display the optimal base stock variables for the single product case discussed in Section 5.2.2. They were obtained for the same parameters of Figs. 5.13-5.16, with the derivatives being taken with respect to the base stock variables during the optimization.

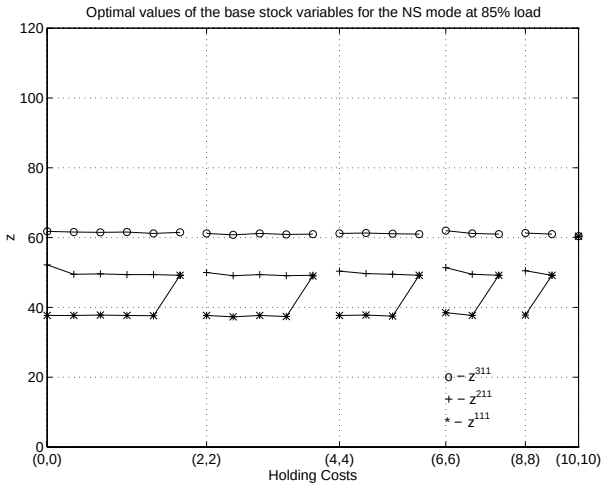


Figure D.1: Optimal base stock levels for the NS mode under an 85% load.

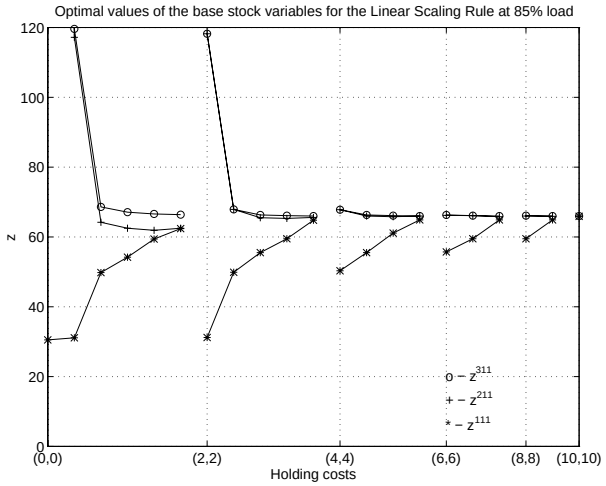


Figure D.2: Optimal base stock levels for the LSR under an 85% load.

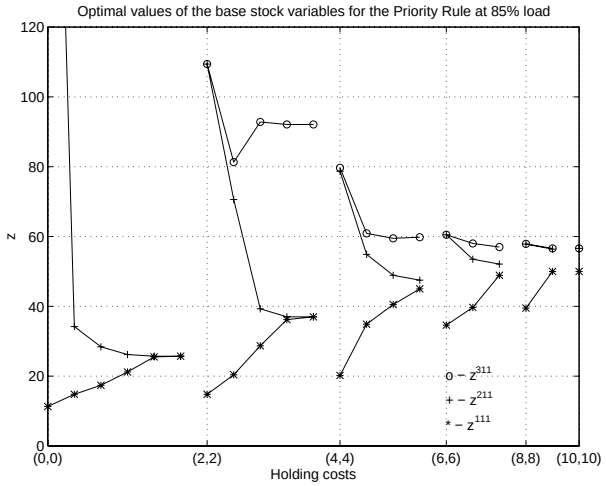


Figure D.3: Optimal base stock levels for the PR under an 85% load.

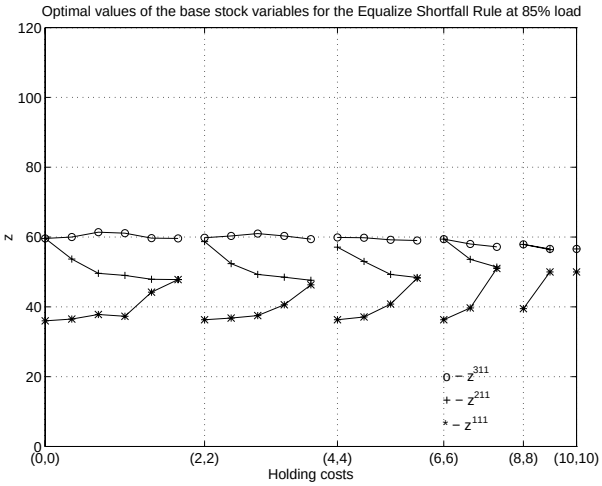


Figure D.4: Optimal base stock levels for the ESR under an 85% load.

D.2 Effect of Capacity Along the Line for the TS Mode

This next set of figures refers to the study of Section 5.3.5. They correspond to the utilization of the PR for the two alternatives of priority and to the utilization of the ESR.

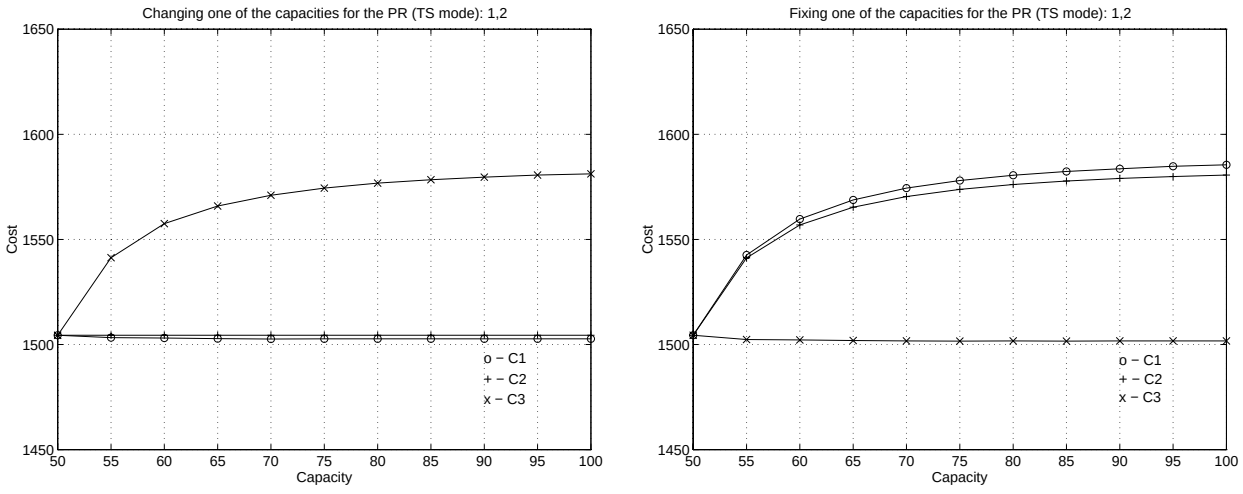


Figure D.5: Effect of capacity along the line for the PR with priority to product 1.

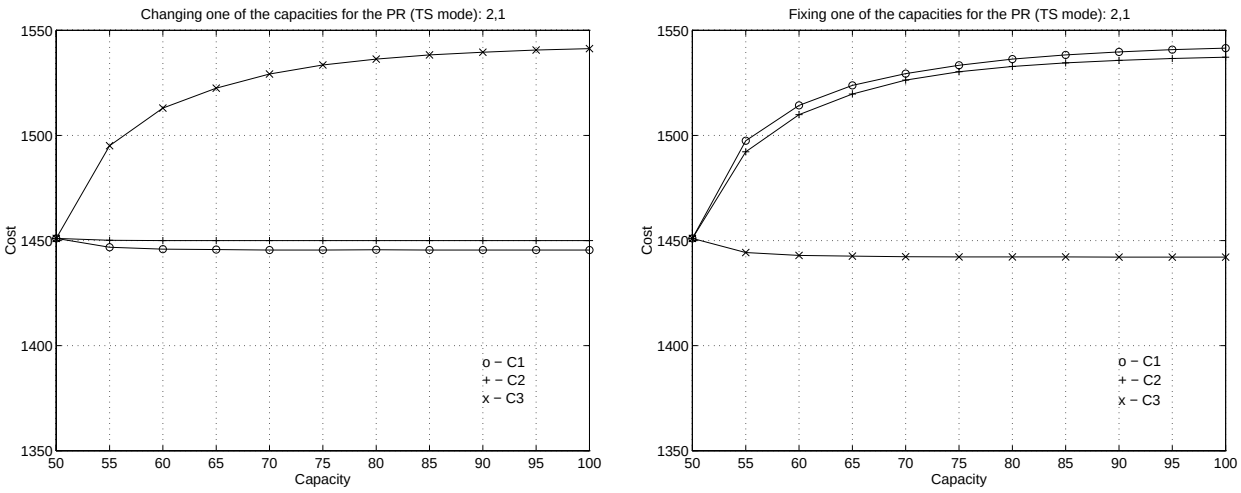


Figure D.6: Effect of capacity along the line for the PR with priority to product 2.

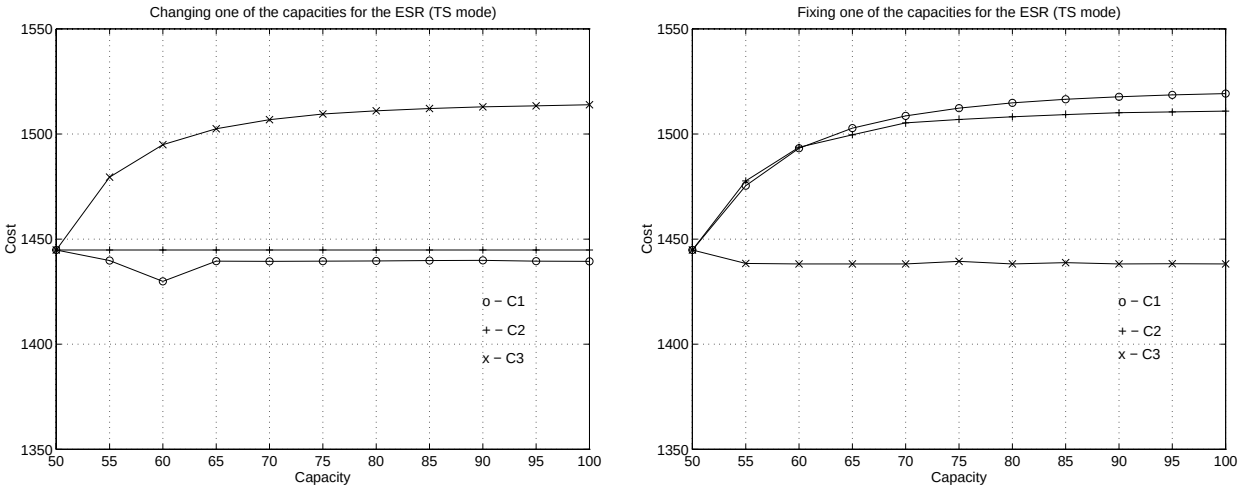


Figure D.7: Effect of capacity along the line for the ESR.

D.3 Experimental data for systems with non uniform loads

This section presents a sample of results obtained for systems with non uniform loads, and it illustrates that the global structure of the experimental data does not change significantly relatively to the case of uniform loads.

Figure D.8 corresponds to a system with $K = 3$, $M = 1$, and $P = 2$. The expected demand is 12 and 8 for product 1 and 2, respectively. The loads imposed by each product are $\tau^1 = [1\ 3\ 5]$ and $\tau^2 = [3\ 2\ 1]$. The leftmost number corresponds to τ^{31i} and the rightmost number corresponds to τ^{11i}, with $i = 1, 2$.

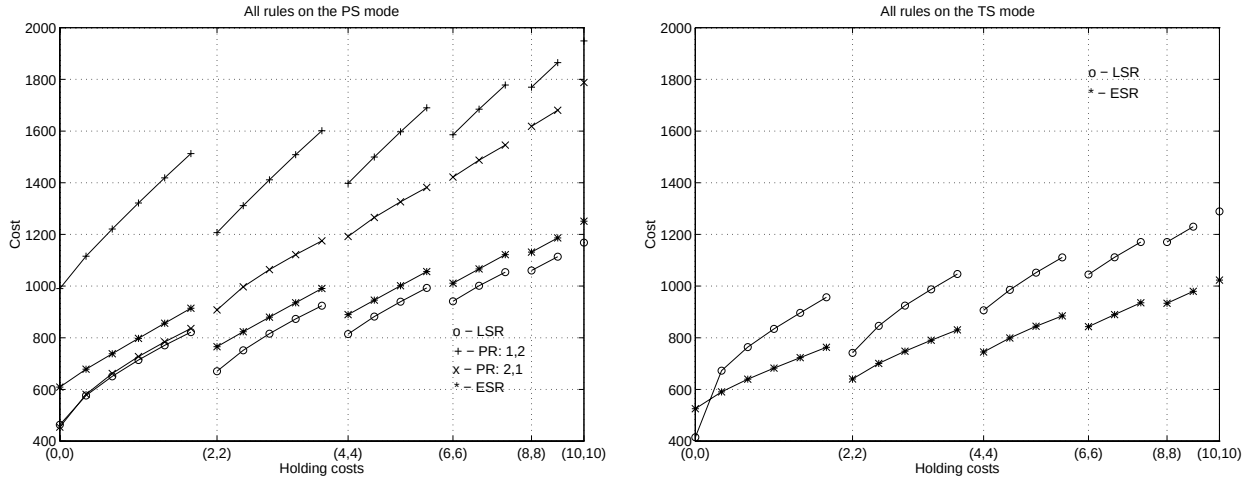


Figure D.8: Optimal cost as function of the holding costs for non uniform loads.

The capacity of the machine was set at 80% load, i.e., $C^1 = [12\times(1+3+5)+8\times(3+2+1)]/0.8$. The holding costs were changed simultaneously for both products, while keeping the backlog costs fixed. This is similar to the study presented in Section 5.2.2.

The graph on the left corresponds to the PS mode and the graph on the right to the TS mode. Naturally, there are no results for the PR in the TS mode. The load imposed by product 2 is below the load imposed by product 1, justifying why giving priority to the former achieves the best performances for the PR in the PS mode. The ESR in the TS mode achieves the overall best performance.

Next figure, (Fig. D.9), shows the comparison between both priority choices in the PS mode for a system with expected demand of 9 and 11 for product 1 and 2, respectively. The loads are

$\tau^1 = [4\ 3\ 5]$ and $\tau^2 = [3\ 4\ 2]$. The overall load imposed by product 2 is below the load imposed by product 1 and, as shown in the figure, priority should be given to product 2, although its expected demand is higher that that of product 1.

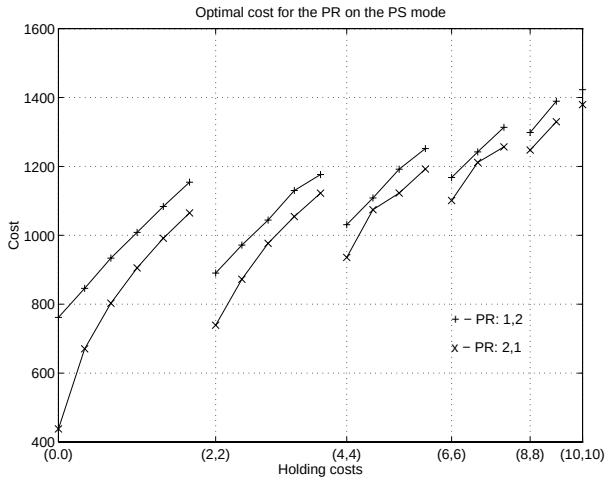


Figure D.9: Comparison between priority choices for non uniform loads.

Bibliography

[Akella et al., 1984] Akella, R., Choong, Y. F., and Gershwin, S. B. (1984). Performance of Hierarchical Production Scheduling Policy. *IEEE Trans. on Components, Hybrids, and Manuf. Technol.*, CHMT-7(3):225–240.

[Akella and Kumar, 1986] Akella, R. and Kumar, P. R. (1986). Optimal Control of Production Rate in a Failure Prone Manufacturing System. *IEEE Trans. on Automatic Control*, AC-31(2):116–126.

[Akella et al., 1992] Akella, R., Rajagopalan, S., and Singh, M. R. (1992). Part Dispatch in Random Yield Multi-Stage Flexible Test Systems for Printed Circuit Boards. *Operations Research*, 40(4):776–789.

[Arrow et al., 1951] Arrow, K., Harris, T., and Marschak, J. (1951). Optimal Inventory Policy. *Econometrica*, 19:250–272.

[Assmussen, 1987] Assmussen, S. (1987). *Applied Probability and Queues*. John Wiley & Sons, New York.

[Atherton and Dayhoff, 1986] Atherton, R. W. and Dayhoff, J. E. (1986). Signature Analysis: Simulation of Inventory, Cycle Time, and Throughput Trade-Offs in Wafer Fabrication. *IEEE Trans. on Components, Hybrids, and Manufacturing Technology*, CHMT-9(4):498–507.

[Azadivar, 1992] Azadivar, F. (1992). A Tutorial on Simulation Optimization. In *1992 Winter Simulation Conference*, pages 198–204.

[Bai and Gershwin, 1994] Bai, S. X. and Gershwin, S. B. (1994). Scheduling Manufacturing Systems with Work-In-Process Inventory Control: Multiple-part-type Systems. *International Journal of Production Research*, 32(2):365–385.

[Bai and Gershwin, 1995] Bai, S. X. and Gershwin, S. B. (1995). Scheduling Manufacturing Systems with Work-In-Process Inventory Control: Single-part-type Systems. *IIE Transactions*, 27(5):599–617.

[Bai and Gershwin, 1996] Bai, S. X. and Gershwin, S. B. (1996). Scheduling Manufacturing Systems with Work-In-Process Inventory Control: Reentrant Systems. *OR Spectrum*, 18(4):187–195.

[Baker, 1974] Baker, K. R. (1974). *Introduction to Sequencing and Scheduling.* John Wiley & Sons, New York.

[Baskett et al., 1975] Baskett, F., Chandy, K. M., Muntz, R. R., and Palacios, F. G. (1975). Open, CLosed and Mixed Networks of Queues with Different Classes of Customers. *J. Assoc. Comput. Mach.*, 22:248–260.

[Bazaraa and Shetty, 1979] Bazaraa, M. S. and Shetty, C. M. (1979). *Nonlinear Programming: Theory and Algorithms.* John Willey & Sons, New York.

[Bechte, 1988] Bechte, W. (1988). Theory and Practice of Load-oriented Manufacturing Control. *International Journal of Production Research*, 26(3):375–395.

[Bellman, 1957] Bellman, R. (1957). *Dynamic Programming.* Princeton University Press, Princeton, N.J.

[Berenji, 1992] Berenji, H. R. (1992). Fuzzy logic controllers. In Yager, R. R. and Zadeh, L. A., editors, *An Introduction to Fuzzy Logic Applications in Intelligent Systems*, pages 69–96. Kluwer Academic Publishers.

[Bertsekas, 1987] Bertsekas, D. P. (1987). *Dynamic Programming: Deterministic and Stochastic Models.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[Bertsimas et al., 1996] Bertsimas, D., Gamarnik, D., and Tsitsiklis, J. (1996). Stability Conditions for Multiclass Fluid Queueing Networks. *IEEE Trans. on Automatic Control*, 41(11):1618–1631.

[Bielecki and Kumar, 1988] Bielecki, T. and Kumar, P. R. (1988). Optimality of Zero-Inventory Policies for Unreliable Manufacturing Systems. *Operations Research*, 29(4):392–400.

[Bispo, 1992] Bispo, C. F. G. (1992). Modeling a Two-Station Tandem System to Accommodate Batch and Machine Failures. GSIA Working Paper #1992-E872096960.

[Bollapragada and Morton, 1994] Bollapragada, S. and Morton, T. E. (1994). Myopic Heuristics for the Random Yield Problem. Technical Report WP #1993-11, Graduate School of Industrial Administration and The Robotics Institute, Carnegie Mellon Univ., Pittsburgh, PA.

[Bramson, 1994] Bramson, M. (1994). Instability of FIFO Queueuing Networks. *Annals of Applied Probability*, 4:414–431.

[Brémaud and Malhamé, 1997] Brémaud, P. and Malhamé, R. P. (1997). A Manufacturing System with General Stationary Failure Process: Stability and IPA of Hedging Control Policies. *IEEE Trans. on Automatic Control*, 42(2):155–170.

[Bryson and Ho, 1969] Bryson, A. E. and Ho, Y.-C. (1969). *Applied Optimal Control*. Gynn and Company.

[Caramanis and Liberopoulos, 1992] Caramanis, M. and Liberopoulos, G. (1992). Perturbation Analysis for the Design of Flexible Manufacturing System Flow Controllers. *Operation Research*, 40(6):1107–1125.

[Cassandras, 1993] Cassandras, C. (1993). *Discrete Event Systems. Modeling and Performance Analysis*. Aksen Associates Inc.

[Chen et al., 1988] Chen, H., Harrison, L. M., Mandelbaum, A., Ackere, A. V., and Wein, L. M. (1988). Empirical Evaluation of a Queueing Network Model for Semiconductor Wafer Fabrication. *Operations Research*, 36(2):202–215.

[Chong and Ramadge, 1994] Chong, E. K. P. and Ramadge, P. J. (1994). Stochastic Optimization of Regenerative Systems Using Infinitesimal Perturbation Analysis. *IEEE Trans. on Automatic Control*, 39(7):1400–1410.

[Ciarallo et al., 1994] Ciarallo, F. W., Akella, R., and Morton, T. E. (1994). A Periodic Review, Production Planning Model with Uncertain Capacity and Uncertain Demand – Optimality of Extended Myopic Policies. *Management Science*, 40(3):320–332.

[Clark and Scarf, 1960] Clark, A. J. and Scarf, H. (1960). Optimal Policies for a Multi-Echelon Inventory Problem. *Management Science*, 6:475–490.

[Conway et al., 1967] Conway, R. W., Maxwell, W. L., and Miller, L. W. (1967). *Theory of Scheduling*. Addison-Wesley, Reading, MA.

[Cunningham, 1990] Cunningham, J. A. (1990). The Use and Evaluation of Yield Models in Integrated Circuit Manufacturing. *IEEE Trans. on Semiconductor Mannufacturing*, 3(2):60–71.

[Custódio et al., 1994] Custódio, L. M. M., Sentieiro, J. J. S., and Bispo, C. F. G. (1994). Production Planning and Scheduling Using a Fuzzy Decision System. *IEEE Trans. on Robotics and Automation – Special Issue on Computer Integrated Manufacturing*, 10(2):160–168.

[Dai and Wang, 1993] Dai, J. G. and Wang, Y. (1993). Nonexistence of Brownian Models of Certain Multiclass Queueing Networks. *Queueing Systems: Theory and Applications: Special Issue on Queueing Networks*, 13:41–46.

[Dai and Weiss, 1996] Dai, J. G. and Weiss, G. (1996). Stability and Instability of Fluid Models for Reentrant Lines. *Mathematics of Operations Research*, 21(1):115–134.

[Dessouky and Leachman, 1994] Dessouky, M. M. and Leachman, R. C. (1994). An Optimization-Based Methodology for Release Scheduling. *Production and Operations Management*, 3(4):276–295.

[Ehteshami et al., 1992] Ehteshami, B., Pétrakian, R. G., and Shabe, P. M. (1992). Trade-Offs in Cycle Time Management: Hot Lots. *IEEE Trans. on Semiconductor Manufacturing*, 5(2):101–106.

[Federgruen and Zipkin, 1984] Federgruen, A. and Zipkin, P. (1984). Approximations of Dynamic, Multilocation Production and Inventory Problems. *Management Science*, 30(1):69–84.

[Federgruen and Zipkin, 1986a] Federgruen, A. and Zipkin, P. (1986a). An Inventory Model with Limited Production Capacity and Uncertain Demands. I. The Average-Cost Criterion. *Mathematics of Operations Research*, 11(2):193–207.

[Federgruen and Zipkin, 1986b] Federgruen, A. and Zipkin, P. (1986b). An Inventory Model with Limited Production Capacity and Uncertain Demands, II. The Discounted-Cost Criterion. *Mathematics of Operations Research*, 11(2):208–215.

[Fox and Glynn, 1989] Fox, B. L. and Glynn, P. W. (1989). Simulating Discounted Costs. *Management Science*, 35(11):1297–1315.

[French, 1982] French, S. (1982). *Sequencing and Scheduling: an introduction to the mathematics of the Job-Shop*. John Wiley & Sons, New York.

[Gerchak et al., 1988] Gerchak, Y., Vickson, R. G., and Parlar, M. (1988). Periodic Review Production Models with Variable Yield and Uncertain Demand. *IEE Transactions*, 20(2):144–150.

[Gershwin et al., 1985] Gershwin, S. B., Akella, R., and Choong, Y. F. (1985). Short-term Production of an Automated Manufacturing Facility. *IBM J. Res. Develop.*, 29(4):392–400.

[Gise and Blanchard, 1986] Gise, P. and Blanchard, R. (1986). *Modern Semiconductor Fabrication Technology*. Prentice-Hall, Englewood Cliffs, N.J.

[Glasserman, 1990] Glasserman, P. (1990). *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers.

[Glasserman, 1991] Glasserman, P. (1991). Structural Conditions for Perturbations Analysis Derivative Estimation: Finite-Time Performance Indices. *Operations Research*, 39(5):724–738?

[Glasserman, 1992] Glasserman, P. (1992). Derivative Estimates from Simulation of Continuous-Time Markov Chains. *Operations Research*, 40(2):292–308.

[Glasserman and Tayur, 1994] Glasserman, P. and Tayur, S. (1994). The Stability of a Capacitated, Multi-Echelon Production-Inventory System under a Base-Stock Policy. *Operations Research*, 42(5):913–925.

[Glasserman and Tayur, 1995] Glasserman, P. and Tayur, S. (1995). Sensitivity Analysis for Base-Stock Levels in Multi-Echelon Production-Inventory Systems. *Management Science*, 41(2):263–281.

[Glassey and Resende, 1988] Glassey, C. R. and Resende, M. G. C. (1988). Closed-Loop Job Release Control for VLSI Circuit Manufacturing. *IEEE Trans. on Semiconductor Manufacturing*, 1(1):36–46.

[Graves, 1981] Graves, S. (1981). A Review of Production Scheduling. *Operations Research*, 29:646–675.

[Graves, 1996] Graves, S. (1996). A Multiechelon Inventory Model with Fixed Replenishment Intervals. *Management Science*, 42(1):1–18.

[Graves et al., 1992] Graves, S., Rinnooy Kan, A. H. G., and Zipkin, P. (1992). Logistics of Production and Inventory. In *Handbooks in Operations Research and Management Science*, volume 4. Elsevier (North-Holland), Amsterdam.

[Gürkan et al., 1994] Gürkan, G., Özge, A. Y., and Robinson, S. M. (1994). Sample-Path Optimization in Simulation. In *1994 Winter Simulation Conference*, pages 247–253.

[Harris, 1913] Harris, F. (1913). How many parts to make at once. *Factory, The Magazine of Management*, 10:135–136, 152.

[Harrison, 1988] Harrison, J. M. (1988). Brownian models of queueing networks with heterogeneous costumer populations. In Flemming, W. and Lions, P., editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, volume IMA 10, pages 147–186, New-York. Springer-Verlag.

[Harrison and Wein, 1990] Harrison, J. M. and Wein, L. M. (1990). Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Closed Network. *Operations Research*, 38(6):1052–1064.

[Henig and Gerchak, 1990] Henig, M. and Gerchak, Y. (1990). The Structure of Periodic Review Policies in the Presence of Random Yield. *Operations Research*, 38(4):634–643.

[Ho, 1992] Ho, Y.-C. (1992). Perturbation Analysis: Concepts and Algorithms. In *1992 Winter Simulation Conference*, pages 231–240.

[Ho and Cao, 1991] Ho, Y.-C. and Cao, X. R. (1991). *Perturbation Analysis of Discrete Event Dynamic Systems*. Kluwer Academic Publishers.

[Ho et al., 1979] Ho, Y. C., Eyler, M. A., and Chien, T. T. (1979). A Gradient Technique for General Buffer Storage Design in a Serial Production Line. *International Journal of Production Research*, 17(6):557–580.

[Jackson, 1975] Jackson, J. R. (1975). Networks of Waiting Lines. *Operations Research*, 5:518–521.

[Karlin, 1960] Karlin, S. (1960). Dynamic Inventory Policy with Varying Stochastic Demands. *Management Science*, 6:231–258.

[Kelly, 1979] Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. John Willey & Sons, New York.

[Kimemia, 1982] Kimemia, J. G. (1982). *Hierarchical Control of Production in Flexible Manufacturing Systems*. PhD thesis, Mass. Inst. Technol., Cambridge, MA.

[Kimemia and Gershwin, 1983] Kimemia, J. G. and Gershwin, S. B. (1983). An Algorithm for the Computer Control of a Flexible Manufacturing System. *IIE Transactions*, 15(4):353–362.

[Kumar, 1993] Kumar, P. R. (1993). Re-entrant lines. *Queueing Systems*, 13:87–110.

[Kumar and Meyn, 1995] Kumar, P. R. and Meyn, S. P. (1995). Stability of Queueing Networks and Scheduling Policies. *IEEE Trans. on Automatic Control*, 40(2):251–260.

[Kumar and Seidman, 1990] Kumar, P. R. and Seidman, T. I. (1990). Dynamic Instabilities and Stabilization Methods in Distributed Real-Time Scheduling of Manufacturing Systems. *IEEE Trans. on Automatic Control*, 35(3):289–298.

[Kumar and Kumar, 1994] Kumar, S. and Kumar, P. R. (1994). Performance Bounds for Queueing Networks and Scheduling Policies. *IEEE Trans. on Automatic Control*, 39(8):1600–1611.

[Lawler et al., 1982] Lawler, E. L., Lenstra, J. K., and Rinnooy Kan, A. H. G. (1982). Recent Developments in Deterministic Sequencing and Scheduling. In *Deterministic and Stochastic Scheduling*, Dordrecht, The Netherlands. Reidel.

[L'Ecuyer, 1994] L'Ecuyer, P. (1994). Efficiency Improvement and Variance Reduction. In *1994 Winter Simulation Conference*, pages 122–132.

[Lemaréchal, 1989] Lemaréchal, C. (1989). Nondifferentiable optimization. In Nemhauser, G. L. et al., editors, *Hanbooks in OR and MS*, volume 1, pages 529–572. Elsevier Science Publishers.

[Little, 1961] Little, J. D. C. (1961). A Proof for the Queueing Formula: $L = \lambda W$. *Operations Research*, 9(3):383–387.

[Lou and Kager, 1989] Lou, S. X. C. and Kager, P. W. (1989). A Robust Production Control Policy for VLSI Wafer Fabrication. *IEEE Trans. on Semiconductor Manufacturing*, 2(4):159–164.

[Lou and van Ryzin, 1989] Lou, S. X. C. and van Ryzin, G. (1989). Optimal Control Rules for Scheduling Job Shops. *Annals of Operations Research*, 17:233–248.

[Lou et al., 1990] Lou, S. X. C., Yan, H., Sethi, S., Gardel, A., and Deosthali, P. (1990). Hub-centered Production Control of Wafer Fabrication. In *Advanced Semi-Conductor Manufacturing Conf. and Workshop*, Danvers, MA.

[Lozinski and Glassey, 1988] Lozinski, C. and Glassey, C. R. (1988). Bottleneck Starvation Indicators for Shop Floor Control. *IEEE Trans. on Semiconductor Manufacturing*, 1(4):147–153.

[Lu et al., 1994] Lu, S. C. H., Ramaswamy, D., and Kumar, P. R. (1994). Efficient Scheduling Policies to Reduce Mean and Variance of Cycle-Time in Semiconductor Manufacturing Plants. *IEEE Trans. on Semiconductor Manufacturing*, 7(3):374–388.

[Lu and Kumar, 1991] Lu, S. H. and Kumar, P. R. (1991). Distributed Scheduling Based on Due-Dates and Buffer Priorities. *IEEE Trans. on Automatic Control*, 36(12):1406–1416.

[Luenberger, 1973] Luenberger, D. G. (1973). *Introduction to Linear and Nonlinear Programming*. Addison-Wesley.

[Morton, 1978] Morton, T. E. (1978). The Nonstationary Infinite Horizon Inventory Problem. *Management Science*, 24(14):1474–1482.

[Morton and Pentico, 1993] Morton, T. E. and Pentico, D. (1993). *Heuristic Scheduling Systems*. John Wiley & Sons, New York.

[Nummelin, 1984] Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, London, UK.

[Ou and Wein, 1992] Ou, J. and Wein, L. M. (1992). Performance Bounds for Scheduling Queueing Networks. *Annals of Applied Probability*, 2:460–480.

[Ou and Wein, 1995] Ou, J. and Wein, L. M. (1995). Dynamic Scheduling of a Production/Inventory System with By-products and Random Yield. *Management Science*, 41(6):1000–1017.

[Panwalker and Iskander, 1977] Panwalker, S. S. and Iskander, W. (1977). A Survey of Scheduling Rules. *Operations Research*, 25:45–61.

[Perkins and Kumar, 1989] Perkins, J. R. and Kumar, P. R. (1989). Stable, Distributed, Real-Time Scheduling of Flexible Manufacturing/Assembly/Disassembly Systems. *IEEE Trans. on Automatic Control*, 34(2):139–148.

[Porteus, 1990] Porteus, E. L. (1990). Stochastic Inventory Theory. In Heyman, D. P. and Sobel, M. J., editors, *Handbooks in Operations Research and Management Science*, volume 2, chapter 12. Elsevier (North-Holland), Amsterdam.

[Prasad, 1991] Prasad, K. (1991). A Generic Computer Simulation Model to Characterize Photolithography Manufacturing Area in an IC FAB Facility. *IEEE Trans. on Components, Hybrids, and Manufacturing Technology*, 14(3):483–487.

[Resende, 1987] Resende, M. G. C. (1987). *Shop Floor Scheduling of Semiconductor Wafer Manufacturing.* PhD thesis, Dept. Industrial Engineering Operations Research, Univ. California, Berkeley, CA.

[Runyan and Bean, 1990] Runyan, W. R. and Bean, K. E. (1990). *Semiconductor Integrated Circuit Processing Technology.* Addison-Wesley.

[Seidman, 1994] Seidman, T. I. (1994). 'First Come First Served' can be Unstable! *IEEE Trans. on Automatic Control*, 39:2166–2171.

[Sharifnia, 1988] Sharifnia, A. (1988). Production Control of a Manufacturing System with Multiple Machine States. *IEEE Trans. on Automatic Control*, AC-33(7):620–625.

[Shi, 1996] Shi, L. (1996). Discontinuous Perturbation Analysis of Discrete-Event Dynamic Systems. *IEEE Trans. on Automatic Control*, 41(11):1676–1681.

[Sigman, 1988] Sigman, K. (1988). Queues as Harris Recurrent Markov Chains. *Queueing Systems*, 3:179–198.

[Song et al., 1992] Song, D., Tu, F., and Lou, S. X. C. (1992). Parameter Optimization of a Control Policy for Unreliable Manufacturing Systems. In *31st Conference on Decision and Control*, pages 1655–1656.

[Stidham, Jr., 1985] Stidham, Jr., S. (1985). Optimal Control of Admission to a Queueing System. *IEEE Trans. on Automatic Control*, AC-30:705–713.

[Strikland, 1993] Strikland, S. G. (1993). Gradient/Sensitivity Estimation in Discrete-Event Simulation. In *1993 Winter Simulation Conference*, pages 97–105.

[Sullivan and Fordyce, 1990] Sullivan, G. and Fordyce, K. (1990). IBM Burlington's Logistics Management System. *Interfaces*, 20(1):43–64.

[Suri and Zazanis, 1988] Suri, R. and Zazanis, M. (1988). Perturbation Analysis Gives Strongly Consistent Sensitivity Estimates for the $M/D/1$ Queue. *Management Science*, 34(1):39–64.

[Sze, 1983] Sze, S. M. (1983). *VLSI Technology*. McGraw-Hill, New York.

[Tayur, 1992] Tayur, S. (1992). Computing the Optimal Policy for Capacitated Inventory Models. *Comm. Statistics – Stochastic Models*, 0(0):585–598.

[Thorisson, 1983] Thorisson, H. (1983). The Coupling of Regenerative Processes. *Adv. Appl. Prob.*, 15:531–561.

[Uzsoy et al., 1992] Uzsoy, R., Lee, C.-Y., and Martin-Vega, L. A. (1992). A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part I: System Characteristics, Performance Evaluation and Production Planning. *IIE Transactions*, 24(4):47–60.

[Uzsoy et al., 1994] Uzsoy, R., Lee, C.-Y., and Martin-Vega, L. A. (1994). A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part II: Shop-Floor Control. *IIE Transactions*, 26(5):44–55.

[van Ryzin, 1987] van Ryzin, G. (1987). Control of Manufacturing Systems with Delays. Master's thesis, Mass. Inst. Technol., Cambridge, MA.

[van Ryzin et al., 1993] van Ryzin, G., Lou, S. X. C., and Gershwin, S. B. (1993). Production Control for a Tandem Two-Machine System. *IIE Transactions*, 25(5):5–20.

[Veatch and Wein, 1994] Veatch, M. H. and Wein, L. M. (1994). Optimal Control of a Two-Station Tandem Production/Inventory System. *Operations Research*, 42(2):337–350.

[Vepsalainen and Morton, 1988] Vepsalainen, A. P. J. and Morton, T. E. (1988). Improving Local Priority Rules with Global Lead-Time Estimates: A simulation study. *Journal of Manufacturing and Operations Management*, 1:102–118.

[Walrand, 1988] Walrand, J. (1988). *An Introduction to Queueing Networks*. Prentice-Hall, New Jersey.

[Wein, 1988] Wein, L. M. (1988). Scheduling Semiconductor Wafer Fabrication. *IEEE Trans. on Semiconductor Manufacturing*, 1(3):115–130.

[Wein, 1990a] Wein, L. M. (1990a). Optimal Control of a Two-Station Brownian Network. *Math. Opns. Res.*, 15:215–242.

[Wein, 1990b] Wein, L. M. (1990b). Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Network with Controllable Inputs. *Operations Research*, 38(6):1065–1078.

[Wein, 1992a] Wein, L. M. (1992a). Dynamic Scheduling of a Multiclass Make-to-Stock Queue. *Operations Research*, 40(4):724–735.

[Wein, 1992b] Wein, L. M. (1992b). On the Relationship Between Yield and Cycle Time in Semiconductor Wafer Fabrication. *IEEE Trans. on Semiconductor Manufacturing*, 5(2):156–158.

[Wilson, 1934] Wilson, R. (1934). A scientific routine for stock control. *Harvard Business Review*, 13:116–128.

[Yager, 1978] Yager, R. R. (1978). Fuzzy Decision Making Including Unequal Objectives. *Fuzzy Sets and Systems*, 1:87–95.

[Yan et al., 1996] Yan, H., Lou, S., Sethi, S., Gardel, A., and Deosthali, P. (1996). Testing the Robustness of Two-Boundary Control Policies in Semiconductor Manufacturing. *IEEE Trans. on Semiconductor Manufacturing*, 9(2):285–288.

[Yan et al., 1992] Yan, H., Yin, G., and Lou, S. X. C. (1992). Approximating Optimal Threshold Values for Unreliable Manufacturing Systems via Stochastic Optimization. In *31st IEEE Conference on Decision and Control*, pages 1657–58.

[Yan et al., 1994] Yan, H., Yin, G., and Lou, S. X. C. (1994). Using Stochastic Optimization to Determine Threshold Values for the Control of Unreliable Manufacturing Systems. *Journal of Optimization Theory and Applications*, 83(3):511–39.

[Yano and Lee, 1995] Yano, C. A. and Lee, H. L. (1995). Lot Sizing with Random Yields: A Review. *Operations Research*, 43(2):311–334.

[Zadeh, 1965] Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*, 8:338–353.