A Unified Approach to Single Machine Scheduling: Heavy Traffic Analysis of Dynamic Cyclic Policies

by

David Maxwell Markowitz

A.B./A.M., Mathematics
Harvard University (1992)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1996

Author ....

Sloan School of Management
May 1, 1996

Certified by ....

Lawrence M. Wein
Professor of Management Science
Thesis Supervisor

Accepted by ....

Thomas L. Magnanti
Co-director, Operations Research Center

A Unified Approach to Single Machine Scheduling: Heavy Traffic Analysis of Dynamic Cyclic Policies

by

David Maxwell Markowitz

Submitted to the Sloan School of Management
on May 1, 1996, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

Recent progress in heavy traffic theory has presented great opportunities in simultaneously analyzing multiple complexities facing single machine scheduling. The thesis examines how setups, due-dates and standardized-customized product mix individually and jointly effect optimal scheduling in a stochastic environment. The Heavy Traffic Averaging Principle (HTAP) is used to optimize a dynamic cyclic policy. The qualitative nature of the results provides a unified view of single machine scheduling and allows a detailed discussion of the interactions between the due-date, setup and product mix facets. A computational study is also performed on several sub-problems. The proposed dynamic cyclic policy is compared to various straw policies and an optimal policy when it can be calculated. The derived dynamic cyclic policy is shown to be robust under many circumstances.

Thesis Supervisor: Lawrence M. Wein
Title: Professor of Management Science

Acknowledgments

Tradition has put much emphasis on this page, not for its academic content, but more because its the only part of the thesis people read. Those interested in the great insights derived in this work will have been more than satisfied in the better written journal version of the material. For those whose eyes have caught this page, your curiosity is in a different area and it might remain a great mystery of the cosmos as to why you have picked this up (please see Physics dissertations for the most recent hypotheses).

These pages should contain different things for different people. For the future ORC doctoral students looking for an example of proper margin width and correct cover page administrative data, don't look here, that stuff changes every year. For Larry and Marty, may they not be looking for the equation I said I would include but has most likely slipped my mind; instead my they find a smile or two at the first mature work a student of theirs has undertaken. For my friends that helped me through my time at MIT, please see as I do that behind each formula and equation is a piece of those years: a wargame with Rich and Joe, bridge with Beril, Leon and Rodrego, a walk around the river with Cristina and those all-encompassing, ever-stimulating conversations with Stefanos.

Anyway, may you find what you seek.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The aim of this thesis is to provide a unified view of single machine scheduling. We jointly examine the influences of three facets of multiple product class manufacturing in a stochastic environment: standardized-customized product mix, setup cost and time penalties and due-dates. This is an ambitious goal as most work in the past has focused on only a few features of this general scheme. A variety of analytic methods have studied simplified systems developing well known results such as the "$c\mu$" rule. Here we apply the now maturing tools of heavy traffic theory to the whole multi-faceted problem. The main analytic contribution is the extension of a particular technique, the Heavy Traffic Averaging Principle (HTAP), to the set of dynamic cyclic policies for queueing systems with due-dates. Many of the insights we shall derive, however, draw upon the interaction of all these problem complexities. There has been no other work which has addressed them all in unison.

We consider a manufacturing system with one machine which produces multiple classes of product. Products can be either customized, which require the request of an order before production can begin, or standardized, which can be pre-stocked in a finished goods inventory (FGI). The machine is limited in capacity and can only produce one class of product at a time. Whenever the machine switches producing one class and starts another, a setup cost or setup time penalty is incurred. Orders for product arrive to the system, each requesting the product at a specified due-date. Products assigned to orders before the due-date must be held and incur an earliness (holding) cost. Products delivered late incur a tardiness (backordering) fee. Standardized products held in finished goods inventory also incur a holding cost equal to that product class' earliness fee. Demand interarrival times, service times, due-dates and setups are product class specific random variables. The machine follows a dynamic cyclic policy. The machine produces the product classes in a fixed cycle and depending on the state of the system can do one of the following three actions: 1) produce the class currently setup for, 2) begin setup of the next product class in the cycle and 3) idle. We wish to optimize the system with respect to long run average costs.

The three complexities of product mix, setups and due-dates provides a rich environment to consider many issues confronting manufacturing. The mixture of customized products and standardized ones captures the trade-offs involved with make-to-stock/make-to-order decisions and the effects of a finished goods inventory on a facility's ability to service a variety of customer demands. Setup costs and setup times each introduce different trade-offs between longer production lot sizes and setup penalties. Due-dates allow the servicing of orders early and provide a mechanism for studying the value of foreknowledge of customer demand. We attempt to optimize the system with all three of these facets and hope to address not only the impact of each one individually but also of their interactions. Parts of this thesis have already appeared in Markowitz, Reiman and Wein [30].

1.2 Literature Review

Figure 1-1 outlines the relationships between these issues. Let us step one-by-one through each of the regions in the diagram, discuss the relevant literature and highlight several qualitative trends which will be important references when discussing our own proposed policies. As we focus on the stochastic version of these problems, we do not include the vast literature on deterministic or static stochastic scheduling as their insights do not directly pertain to the work done in this thesis.

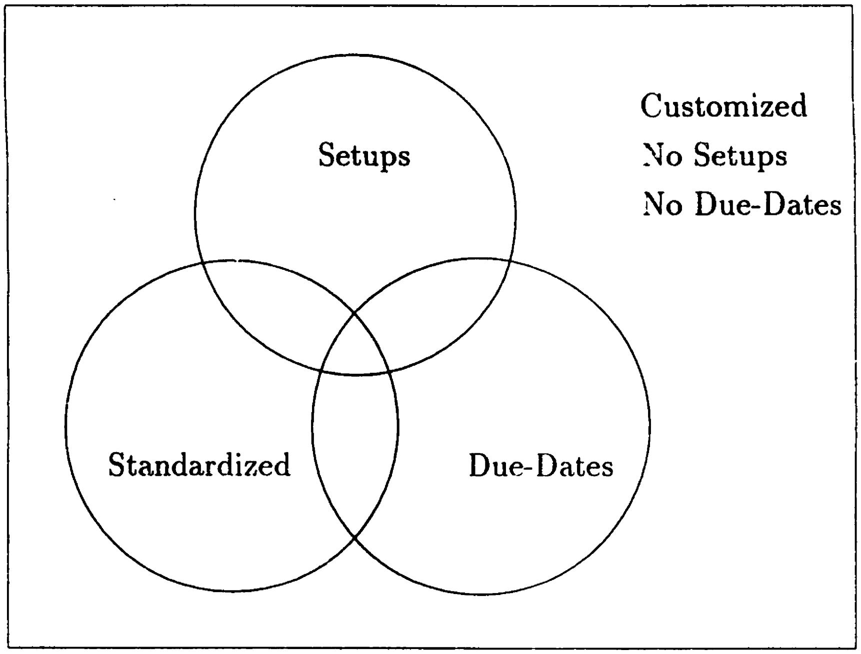

Figure 1-1: Breakdown of general problem

The outer area of Figure 1-1 corresponds to systems with only customized goods, no setup penalties and no due-dates. These systems are the simplest of the ones displayed in the diagram and have been extensively studied. Cox and Smith [10], see also Klimov [24], show the optimality of the "$c\mu$ rule," or priority service of the class which while serviced removes cost from the system at the highest rate. This simple policy has several important characteristics. First, although considered a "static" priority policy, it is dynamic in that it can instruct the machine to change setup based on any minor fluctuation in the state of the system. Second, it chooses to service high cost customers at the expense of low cost ones.

The region of the Venn diagram corresponding to customized products with setups and no due-dates has also been extensively studied. These systems are models frequently used in the design of communication and computer networks. In addition, they represent make-to-order manufacturing facilities with non-trivial penalties when switching production between product classes. The amount of work in this area is quite large; we briefly name just a few of its many contributors. The works of Takagi

[43] have studied the performance of these systems under a variety of polling policies. Hofri and Ross [20] examine a two-product case and develop a double threshold policy which is optimal when the two products have the same "$c\mu$." In this policy, the server works on a product class to exhaustion and only switches when the other class has reached a minimal level. This policy is dynamic as it depends on the current levels of inventory and the setup of the machine. The policy also gives insight on the trade-offs involved when one adds setups to a customized product system. Since the two product classes are nearly identical, there is no preference for working on one class over another as seen in the "$c\mu$" rule. Instead, the policy balances machine removal of orders from the system and machine setups: by implementing an idling threshold based on a minimum level of inventory before switching, minimum lot-sizes are guaranteed and excessive setups are avoided. This is performed at the expense of longer queues and their associated costs. Reiman and Wein [39] use the HTAP to create policies for two product classes with asymmetric cost parameters. Their policies are again dynamic and exhibit this trade-off between queuing costs and setup penalties. They implement a state-dependent switching threshold which not only controls the frequency of setups but also regulates the amount of orders in each queue. Thus, they create a state-dependent lot size which simultaneously controls the ratio of costs incured by each queue and balances these holding costs with the setups penalties. They conjecture their policies are asymptotically optimal. Systems with more than two classes of products are more difficult to analyze: there is not only the problem of determining when to switch to another class but also of finding which class to setup. Boxma, Levy and Westrate [5] create polling tables to minimize mean waiting time. Browne and Yechiali [6] create a quasi-dynamic index policy to chose sequences of classes to service at the start of each cycle. Van Oyen and Duenyas [34] construct a dynamic policy based on a myopic "look ahead," and in [12] they examine the problem with setup costs.

The region of Figure 1-1 corresponding to customized products with due-dates and without setups has also been studied, but not to the extent of either of the two previous regions. These systems reflect manufacturing facilities which service customer orders on a make-to-order basis with the additional aspect that customers do not want the goods immediately but at some future time. This problem, however, is more complex than the previous ones because the state-space of the system has exploded in dimensionality: each order potentially must pass through a continuum of due-date lead time states before it exits the system. Baccelli, Liu and Towsley [2] consider ordering policies to minimize job lateness. Pandelis and Teneketzis [35] look at earliness and tardiness penalties and examine properties of an optimal policy. Righter [40] uses stochastic ordering to further characterize aspects of an optimal policy. Several simplification schemes can also be used. Van Mieghem [44] studies a system with generalized product class costs based upon each order's age in the system. This can be interpreted as a class dependent deterministic due-date cost structure. Using heavy traffic analysis, Van Mieghem shows that a generalized "$c\mu$" rule is asymptotically optimal. The policy is similar to the zero due-date case in that it is dynamic and dedicates the machine to servicing orders which will remove cost from the system at the fastest rate.

The combination of setups and due-dates in a customized product system has been little studied. We know of no previous work analytically treating this problem.

Systems with standardized products are a conventional means of modeling finished goods inventories with backordering: a situation common in manufacturing problems. Their analysis, however, is generally considered to be more complex than systems with customized products because of the inherent non-linear cost structure introduced by having both holding and backorder costs. Additionally, when constructing policies for multi-class inventory systems, there are no natural switching boundaries as with exhausting a queue in customized systems. These difficulties have inhibited the analysis of even the most simple case: standardized products with no due-dates and no setups. Zheng and Zipkin [47] and Pena and Zipkin [36] look at multiclass symmetric product systems under base-stock policies. The base-stock method is an intuitive approach where one sets a base-stock level for each product high to avoid backordering. Ha [18] also studies a multiclass production system and examines the optimality of base-stock policies with switching curves. Wein [46] uses heavy traffic theory to show that asymptotically a mixture of priority policies with an idling threshold is optimal. All but the lowest holding cost, "$h\mu$," or backorder cost, "$b\mu$," product are prioritized and in the limit vanish. Only the cheapest holding and backorder products are seen in queue. This is reminiscent of the "$c\mu$" rule for the customized system: the policy is dynamic and it attempts to focus on products where cost can quickly be removed from the system while ignoring lower cost items. The results are also important because of its interpretation of the role of a finished goods inventory. The inventory acts as a reservoir of stored capacity allowing the server to dynamically allocate its limited resources on high cost orders. Instead of individual inventories hedging against back-ordering, the total reserve machine capacity stored in the cheapest product acts as a buffer against missing demand. Veatch and Wein [45] expand upon this by examining index policies in a two product Markovian setting.

Standardized goods with setups and without due-dates has long been considered the prototype for modeling manufacturing systems servicing a finished goods inventory.

The dynamic scheduling, or *lot-sizing*, of the machine is a stochastic version of the classic economic lot scheduling problem (ELSP), which is NP-hard (Hsu [21]) and has not been solved in general. Despite the vast literature devoted to the ELSP (see the survey paper by Elmaghraby [13], and Zipkin [48] for a list of more recent references), its deterministic viewpoint has probably prevented its widespread industrial use: the solution to a deterministic problem in a make-to-stock setting will not hedge against uncertainty in future service times (e.g., machine failures) and demand, resulting in many costly backorders (see the numerical results in Federgruen and Katalan [14]).

Not surprisingly, the stochastic version of the ELSP appears to be analytically intractable. When the state space is taken to be discrete, the stochastic ELSP (or SELSP) can be viewed as a make-to-stock version of the dynamic scheduling problem for a *polling system*, which is a traditional (i.e., make-to-order) multiclass queue with setups. The SELSP is more challenging than the polling scheduling problem because of the nonlinear cost structure and the lack of an imposed boundary at the origin. Despite its difficulty, this problem has been the subject of a recent flurry of activity.

Graves [17] develops a Markov decision model for a one-product problem, and uses it to develop a heuristic for the SELSP in a periodic review setting. Leachman and Gascon [27], Gallego [16] and Bourland and Yano [4] develop heuristic lot-sizing algorithms for the ELSP with stochastic demands that are rooted in the solution to the deterministic ELSP; the first of these papers considers a discrete time problem with nonstationary demand. Sharifnia, Caramanis and Gershwin [41] employ a hierarchical approach to develop heuristic policies for a stochastic fluid version of the problem. They propose a piecewise linear system of switching curves which dynamically trade-off holding and backorder costs with setups. Federgruen and Katalan [14, 15] develop accurate distributional approximations for polling systems, and use these to analyze the performance of a class of periodic base stock policies for the SELSP. Anupindi and Tayur [1] also consider a class of periodic base stock policies, and use a simulation based approach (infinitesimal perturbation analysis and gradient search) to obtain good base stock policies for a variety of performance measures. Sox and Muckstadt [42] formulate the SELSP as a stochastic program and propose a heuristic decomposition algorithm to solve it. Qiu and Loulou [37] formulate the problem as a semi-Markov decision process, and numerically compute the optimal solution in the two-product case; this is the only paper to date to gain any insight into the nature of the optimal solution to the SELSP.

The case of standardized products with due-dates and without setups has not been extensively studied. Similarly, the system with setups, the dead center of Figure 1-1, has also not been examined. Anupindi and Tayur [1] mention that this area is particularly vacant in consideration. As in the the customized cases, due-dates make this problem extremely difficult to exactly analyze. Nonetheless, this area has become increasingly important due to the interest in Just-in-Time manufacture. As cited in Federgruen and Katalan [15], increased information lead time of demand has sparked a desire to delete inventory and service orders for standardized goods in a make-to-order fashion. A contribution of this paper is to optimize this system as is and let the due-date distributions themselves decide which product classes are serviced in a make-to-stock versus make-to-order manner.

Additionally, we consider mixed systems with both standardized and customized products, which would correspond to the border of the "Standardized" circle of Figure 1-1. The distinction between customized and standardized is based entirely on product design. Much of the work in the area, however, has been aimed at answering make-to-stock/make-to-order decisions and so the partition of "customized" and "standardized" goods is a parameter to be optimized. This has led to models of hybrid systems like those of Carr *et. al.* [7] where the make-to-order goods represent low demand items that are made Just-in-Time and thus have priority over the make-to-stock ones. These intrinsic priority rules allow for a performance analysis of the MTO/MTS partition, but do not involve optimally scheduling of the product classes themselves. Nguyen [32] has looked at the performance of hybrid systems but with lost-sales instead of backordering. In a following paper [33], she examines different priority rules for the MTS and MTO products and suggests an algorithm for setting base-stock levels.

Mixed systems with setups and no due-dates have also often been used to examine MTS/MTO product partitions. The presence of setups, however, forces some sort of scheduling optimization to be performed. Federgruen and Katalan [15] examine such hybrid systems and compare several priority rules for switching from MTS goods to MTO instead of the absolute priority rule in Carr *et. al.* [7]. From scheduling techniques, they propose a heuristic for partitioning MTS and MTO items.

Lastly, there is the full problem: customized and standardized goods, setups and due-dates. All of the previous cases are subsets of this general model. It is the proper setting to ask questions of balancing inventory costs and setup penalties, of setting finished goods inventory levels and avoiding backordering, of determining due-date lead time effects and natural due-date partitions of MTS/MTO goods. In this paper, we asymptotically optimize the full problem over dynamic cyclic policies and are able to derive insights on the combined answer to these seemingly separate questions.

1.3 Approach

This thesis draws upon the methods of Reiman and Wein [39]. In that paper, the system is approximated using two limits under the HTAP (see Coffman, Puhlaskii and Reiman [9, 8] for details, Lennon Olsen [28] for refinements) which requires that the total utilization be close to one. Two sets of scalings are used, a fast one where time is sped up by a factor of $O(n)$ and a slow one where time is increased by a factor of $O(\sqrt{n})$. Under the fast scaling the total number of orders behaves like a diffusion process and is called the diffusion limit. Under the slow scaling, individual inventories act deterministically and can be approximated by a fluid limit. The HTAP couples these two processes. Individual inventories move an order of magnitude faster than the total workload, and so the fluid limit evolves for a period while the total workload remains relatively constant. For example, the total workload might change on the order of weeks, individual inventories change daily.

By isolating the variability of the system into one scaling, there is an opportunity to selectively optimize the system. A main analytic contribution of this thesis is determining how the proposed family of policies behave in the fluid limit and then optimize them in two stages: first in the fluid and then in the diffusion limits. The results which we obtain, however, are not closed form but must be numerically solved. For more insight into the problem, we reduce the general case to a class dependent deterministic due-date example which we can analyze in depth.

In an attempt to both assess the effectiveness of our proposed policies and the strength of our approximation, we perform a computational study. We start with the SELSP and compare our proposed policies to two straw policies that are closely related to those considered by Federgruen and Katalan [14] and by Sharifnia, Caramanis and Gershwin [41]. All policies are examined for a variety of two product problems and several five product cases. Simulations are additionally performed on a select number of two product due-date examples.

The explicitness of our results reveals a number of new and unexpected insights into the nature of the optimal solution for a wider range of scheduling problems. Readers who are not curious about the mathematical details but who wish to obtain a deeper understanding of the inter-relationships between setups, due-dates and product mix may find it useful to bypass the heavy traffic analysis and focus on Chapter 4.5, where the key insights and observations are collected.

The thesis itself is organized as follows: the exact problem is formulated and the HTAP outlined in Chapter 2. In Chapter 3 we show how the proposed policies behave under the fluid limit and in Chapter 4 we work through the deterministic example.

Chapter 2

The Problem and Approach

2.1 Problem Formulation

A single machine produces N classes of goods. Of these, N^c are customized and N^s are standardized. Without loss of generality, we assume that products $1, 2, \ldots, N^c$ are customized and $N^c+1, \ldots, N^c+N^s = N$ are standardized. Each product class i has its own generally distributed service time with mean μ_i^{-1} and coefficient of variation c_{is}. Orders for product arrive from an exogenous renewal demand process. For each class i, the demand interarrival time is generally distributed with mean λ_i^{-1} and coefficient of variation c_{ia}. The demand and service distributions for each class are assumed to be independent, although they need not be (see Reiman [38] for compound renewal processes). Orders arrive to the system with a specified due-date. The due-date lead time s of a class i arrival is determined by a random variable with density $\tilde{f}_i(s)$ and mean $\tilde{l}_i$ and is independent of the service and demand distributions. We assume that there is a minimum and maximum feasible due-date lead time, denoted by $\tilde{a}_i$ and $\tilde{b}_i$ respectively. With this notation, each product class has a utilization $\rho_i = \lambda_i/\mu_i$ and the total utilization is $\rho = \sum_{i=1}^{N} \rho_i$.

Orders for customized goods are immediately queued. The machine can only work on a customized class if an order is present. Once an order is serviced it is either held until the order's due-date or sent immediately if tardy. If held, the order incurs an earliness fee (holding cost) $\hat{h}_i$ per unit time until the due-date; if late, the order incurs a tardiness fee (backorder cost) of $\hat{b}_i$ per unit time late. These costs can also be stated in terms of units of work: we denote them by $h_i = \mu_i \hat{h}_i$ and $b_i = \mu_i \hat{b}_i$.

Standardized goods can be pre-stocked and can be assigned to any order. Thus, when an order for a customized item arrives, the order is queued and can be filled by an item either from the finished goods inventory or directly from the output of the machine. Once an item is assigned to an order, either an earliness fee $\hat{h}_i$ is incurred for each unit of time until the order's due-date or a tardiness fee of $\hat{b}_i$ per unit time late. Standardized goods in the finished goods inventory also accrue holding cost at a rate of $\hat{h}_i$ per item per unit time. Again, these costs workload equivalents are $h_i = \mu_i \hat{h}_i$ and $b_i = \mu_i \hat{b}_i$.

The state of the system is reflected in several processes. The "tilde" denotes a lack of scaling, later other notation will be given for the two heavy traffic scalings. Let $\tilde{I}_i(t)$ be the number of outstanding product i orders at time t, $\tilde{W}_i(t)$ be the amount of work in outstanding product i orders at time t and $\tilde{H}_i(t)$ the amount of work stored in product i finished goods inventory. Let $\tilde{I}(t)$ and $\tilde{H}(t)$ represent the N and N^s dimensional vector processes of $\tilde{I}_i(t)$ and $\tilde{H}_i(t)$ respectively. Lastly, let $\tilde{W}(t)$ be the total work in outstanding orders at time t and is equal to $\sum_{i=1}^{N} \tilde{W}_i(t)$. We do not have a process monitoring the setup of the machine as the notation will not be necessary in our analysis.

The machine follows a dynamic cyclic policy. All classes of items are serviced in a fixed cycle. At any point in time the machine has three options: 1) produce the class currently setup for (this might not be possible if setup for a customized class and there are no orders present), 2) begin setup of the next product class in the cycle, or 3) idle. Thus, the amount of time allocated to each class is decided by the current state of the system. This gives the policy a dynamic nature in that lot size and cycle length can be molded to address changing needs of the system. Every time the machine switches production to the next class in the cycle a penalty is incurred. This can be a cost, a period of down-time or both. Let the average cost per cycle be $\tilde{K}$, the average down-time per cycle s. If only one form of setup penalty is present and setups are product dependent, we assume that the cycle order minimizes the average cycle penalty (this can be considered a TSP where the product classes are nodes and the product dependent setup penalties are the costs for the connecting arc). If both setup costs and times are present, the situation is more complex. We do not consider the influence of pre-emptive service as the approximation scheme we use is too coarse to differentiate a system with a pre-emptive resume policy versus one without.

We wish to minimize the long run average cost of the system. In order to formulate this, we need additional notation. Let T_{ni} be the time that the nth class i product is assigned to an order. Let $\tilde{J}(t)$ be the cumulative number of cycle completions at time t. Let $\tilde{G}_i(s,t)$ be the amount of product i work at time t due at time $s+t$. When a product is assigned to an order, we assume that it is assigned to the order with the smallest due-date lead time in that product class as this will minimize cost. Thus, let $\tilde{L}_i(t)$ be the smallest due-date lead time in the product i unfinished orders queue. The long run average cost can now be stated as:

$$\begin{array}{c} \lim_{T\to\infty} \frac{1}{T}\left(\sum_{i=N^c+1}^{N}\left(\sum_{n|T_{ni}<T} \hat{h}_i\tilde{L}_i^+(T_{ni}) + \hat{b}_i\tilde{L}_i^-(T_{ni}) + \int_0^T \hat{h}_i\mu_i\tilde{H}_i(t)dt\right)\right. \\ \left. + \sum_{i=1}^{N^c}\left(\sum_{n|T_{ni}<T} \hat{h}_i\tilde{L}_i^+(T_{ni}) + \hat{b}_i\tilde{L}_i^-(T_{ni})\right) + K\tilde{J}(T)\right) \end{array} . \tag{2.1}$$

In some respects this formulation of cost is not standard. We have "reversed the order of integration." Normally the multiplicative product of cost and queue length (e.g. $\hat{h}_i\tilde{I}_i^+(t)$) is integrated over time. Instead, we have chosen to sum the product of cost and time (e.g. $\hat{h}_i\tilde{L}_i^+(T_{n,i})$) over orders. This is a necessary step in analyzing the effects of due-dates.

2.2 Heavy Traffic Approximations

The Heavy Traffic Averaging Principle (HTAP), formally developed in Coffman, Puhlaskii and Reiman [9, 8], has augmented the way we view single station queuing systems (see Lennon Olsen [28] for refinements for polling systems with setup times). By using two sets of limits, it distinguishes the order of magnitude of different sources of variability in a multiclass system. Although rigorously proved for a two product MTO system (with setups in [8]) following an exhaustive service policy, it has been assumed that its results hold for a wider class of systems. It has been used with success in Reiman and Wein [39]. We shall briefly motivate the HTAP results so that conclusions can be drawn for systems with due-dates, but refer to either Coffman, Puhlaskii and Reiman [9, 8] or Reiman and Wein [39] for a more detailed discussion.

The HTAP is based on two sets of limits, taken as the total utilization goes to one and synchronized by a scaling parameter n. The first limit of the HTAP states that as $\sqrt{n}(1-\rho) \rightarrow$ constant, then $\tilde{W}(nt)/\sqrt{n} \rightarrow$ a diffusion process, $W(t)$, with parameters defined by the system data and policy. It is called the diffusion limit. Without setups, this would be a standard heavy traffic limit (see Iglehart and Whitt [22]) and $W(t)$ would be an RBM. The second limit, called the fluid limit, states that for the same scaling parameter n and for a given total workload, $\tilde{W}_i(\sqrt{n}t)/\sqrt{n} \rightarrow \bar{W}_i(t)$ a fluid process. The result is related to the strong law of large numbers.

Moreover, setups are affected by the two scalings. Setup times completely disappear in the fluid limit and become incorporated into the drift of the diffusion process. As shown in Coffman, Puhlaskii and Reiman [8], the drift of the diffusion process for total workload level w is

$$\frac{s}{\tau(w)} - c \tag{2.2}$$

where $\tau(w)$ is the average cycle length. Since cycle length is effected by dynamic cyclic policies the drift of the diffusion process is policy dependent. The variance of the process is

$$\sigma^2 = \sum_{i=1}^{N} \frac{\lambda_i}{\mu_i^2}(c_{di}^2 + c_{si}^2), \tag{2.3}$$

and, thus, is only dependent on system parameters. Setup costs are also affected and only appear as a scaled version $K = \tilde{K}/n$ in the fluid limit.

Both the coupling of the diffusion and fluid processes by conditioning on the total workload and the disappearance of setups are the distinguishing features of the HTAP for applications. It's interpretation is intuitive: as utilization approaches one, work in the system is large and setups are infrequent. Since the average demand rate is equal to the average production rate, the total workload cannot change quickly, yet, by focusing on one class and setting up for another, work in individual queues can shift rapidly between product classes. The shifting occurs an order of magnitude, $O(\sqrt{n})$, more quickly than the total workload. The nature of the fluctuation in $\bar{W}_i(t)$ is determined by which class is given attention and which is neglected, namely the fluctuation is determined by the policy. Another consequence of this is that if the machine is following a cyclic policy, many cycles will be completed before there is significant change in total workload $W(t)$. As there are many cycles for a given total workload, we can find an average cost per cycle for a given workload. This allows us to restate equation 2.1 in terms of W, a total workload process, and $c(w)$, the average cost for a given total workload (for future readability, we denote by W the total workload process and by w an arbitrary feasible total workload). Suppressing the notation illustrating the dependence of the cost on our policy, this is

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T c(W(t))dt. \tag{2.4}$$

It is useful to view the control policy as consisting of two interrelated decisions: a busy/idle policy and a dynamic lot-sizing policy that specifies what the server should do while working. We begin by characterizing the busy/idle policy. The HTAP and the well known relationship between queueing systems and production/inventory systems (e.g., Morse [31]) imply that the system state of the heavy traffic control problem is the one-dimensional total workload process W, which measures the total machine time needed to complete the current orders in queue. Furthermore, the total workload process is directly affected by the server's busy/idle policy. Hence, a reasonable form of the optimal busy/idle policy is for the server to stay busy if $W(t) > w_0$ and to idle if $W(t) \leq w_0$, for the unspecified control parameter w_0. The quantity w_0 will often be referred to as the *idling threshold*, and can be viewed as an *aggregate base stock level*. The HTAP implies that the total workload process W is a diffusion process on $[w_0, \infty)$ under this busy/idle policy. Moreover, we can rewrite 2.4 in terms of the steady state distibution, dW of the diffusion process and get

$$\int_{w_0}^{\infty} c(w)dW(w). \tag{2.5}$$

The calculation of $c(w)$, the average cost per cycle as a function of total workload, is central to the optimization of equation 2.4 and 2.5. The strong law of large numbers embedded in the fluid limit allows $c(w)$ to be tractably analyzed. This is the focus of Chapter 3.

For the full problem, we wish to include due-dates in the HTAP calculation of $c(w)$. We assume that due-dates are incorporated in the fluid scaling and so $\tilde{f}_i(\sqrt{n}s) \to \bar{f}_i(s)$ is well defined and non-trivial. Thus, under a diffusion time scaling the density $\tilde{f}_i(ns)$ converges to a point mass at s equal to zero and is zero elsewhere. This implies that due-dates do not appear in the diffusion process: they have been isolated in the fluid limit.

The question remains as to how due-dates impact the arrival process. Let us initially assume that each class arrives with a deterministic due date lead time of $\tilde{f}_i$ with fluid version $\bar{f}_i = \tilde{f}_i/\sqrt{n}$. Under the fluid scalings the class i workload arrival process is deterministic and flows in at rate ρ_i. Thus at any dt instant of time $\rho_i dt$ units of work due in $\bar{f}_i$ time units are arriving to class i. We can generalize one step and have class i orders arrive with due-date lead time $\bar{f}'_i$ with probability p and $\bar{f}''_i$ with probability $1-p$. Thus, since a fraction of work arriving to the fluid system represents $\sqrt{n}$ units of work in the unscaled process, by the strong law of large numbers in the fluid scalings $p\rho_i$ units of class i work arrive with due-date lead time $\bar{f}'_i$ and $(1-p)\rho_i$ units of work arrive with due-date lead time $\bar{f}''_i$. The same reasoning holds for any PMF due-date distribution and thus we can approximate any continuous distribution as closely as we would like. Therefore, for general due-date distribution $\bar{f}_i(s) = \tilde{f}_i(\sqrt{n}s)$, $\bar{f}_i(s)\rho_i dt$ units of work due in s units of time arrive at any instant in the fluid limit: we have a deterministic due-date arrival rate. In addition to the scaling of $\tilde{f}_i$, the accounting functions $\tilde{L}_i$ and $\tilde{G}^I_i(s,t)$ are also transformed in the fluid limit. As before the bar superscript denotes fluid versions, and thus let $\bar{L}_i(t)$ equal $\tilde{L}_i(t)/\sqrt{n}$, let $\bar{G}^I_i(s,t)$ be $\tilde{G}^I_i(\sqrt{n}s, \sqrt{n}t)/\sqrt{n}$ and let $\bar{l}_i$ be $\tilde{l}_i/\sqrt{n}$. In the next Chapter we shall use this result to calculate $c(w)$ and detail equation 2.4.

Chapter 3

Dynamic Cyclic Policies

The goal of this chapter is to determine the structure of $c(w)$ and detail a method for optimizing the heavy traffic version of the dynamic cyclic policy. We shall study the structure in two steps: first we will study the behavior of $\bar{W}_i(t)$ under a dynamic cyclic policy and second we shall calculate the cost per cycle for a customized and a standardized product. We will then examine a general method for optimizing the heavy traffic policy and end with an interpretation of the original unscaled policy.

3.1 Cyclic Policies and $\bar{W}_i$

A cyclic policy is characterized by the lot size for each product (or equivalently, since the analysis is deterministic, the length of time each product is served in a cycle). Because idleness is only incurred when the total workload reaches a certain base stock level, we assume that no idleness is incurred during a cycle. When $W(t) = w$, a cyclic policy is best viewed as a closed $N-$dimensional deterministic path in the constant workload hyperplane $\sum_{i=1}^{N} w_i = w$; that is, the process traverses the same path repeatedly, once per cycle.

Although a cyclic policy can be specified in many ways, we choose a particular characterization that is convenient for analysis. A cyclic policy (or, equivalently, the closed loop generated by the policy) will be defined by $N+1$ quantities: the *cycle length* τ and the *cycle center* $x^c = (x_1^c, \ldots, x_N^c)$. These control parameters are actually functions of the total workload w, but this dependence will be suppressed for improved readability. The cycle length τ is the length of time required to perform a cycle, and x_i^c is product i's "center of fluctuation", or equivalently, the average amount of this product's inventory over the course of a cycle. Because the transient effects associated with initiating or temporarily moving a cycle vanish in the heavy traffic time scaling, the cycle center x^c can be placed anywhere in the constant workload hyperplane.

We begin by examining the deterministic behavior of the individual product workload levels $\bar{W}_i$ under a cyclic policy when $W(t) = w$. For the system to remain balanced, the amount of each product produced per cycle must equal the amount demanded, and hence each product must be produced a fraction ρ_i of the time; we assume that ρ equals one throughout this fluid analysis, so that the server is busy throughout the cycle. Thus, for an arbitrary instantaneous total workload w and cycle time τ, each product i must be serviced for $\rho_i \tau$ units of time per cycle. Therefore, when the machine is servicing product i, the work content in this product's order queue is augmented at rate ρ_i and is serviced at rate one, and so $\bar{W}_i$ decreases at the fixed rate $1 - \rho_i$ for $\rho_i \tau$ units of time per cycle. For the remaining $(1 - \rho_i)\tau$ time units in the cycle when product i is not being serviced, the workload queue is increasing at rate ρ_i. To uniquely determine the behavior of a cyclic policy, a reference starting point also needs to be specified. We use x_i^c, product i's average order level, as the reference point. Readers are referred to Figure 3-1 for a reinforcement of these notions.

Thus, class i's workload level fluctuates by $\rho_i(1 - \rho_i)\tau$ over the course of a cycle. This implies that the maximum amount of work over the cycle is $x_i^c + \rho_i(1-\rho_i)\tau/2$ and the minimum is $x_i^c - \rho_i(1-\rho_i)\tau/2$. Cycle center and cycle length are both parameters which can be set by the policy. There are some restrictions: the sum of the cycle centers must equal the total workload, i.e. $\sum_{i=1}^{N} x_i^c = w$; for a customized class $i \leq N^c$ the minimum amount of work over the cycle must be non-negative, $x_i^c - \rho_i(1-\rho_i)\tau/2 \geq 0$; and if there are setup penalties then the cycle length must be greater than zero.

The idling threshold w_0 is a policy parameter which has no effect on the fluid process and hence $c(w)$. It only impacts the diffusion process $W(t)$: it is the reflection point of W. There are restrictions on w_0 only if there are no standardized goods in

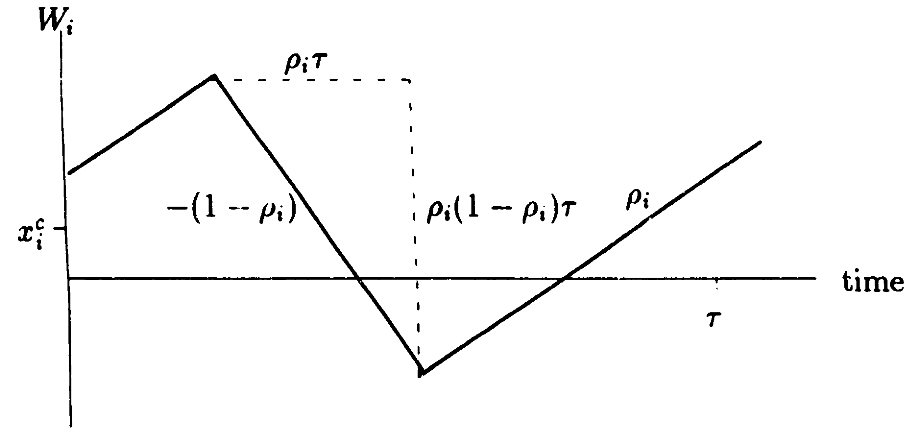

Figure 3-1: Workload fluctuation over a cycle.

the system, that is $N^s = 0$. In this case, w_0 must be non-negative.

3.2 Cost per Cycle

Given the behavior of the dynamic cyclic policy, we can express the cost of a fluid limit cycle in terms of parameters x^c and τ. As the cost per cycle is composed of individual product class costs and setup costs, we can reduce $c(w)$ to

$$c(w) = \sum_{i=1}^{N^c} c_i(x^c, \tau, w) + \sum_{i=N^c+1}^{N} c_i(x^c, \tau, w) + \frac{K}{\tau} \tag{3.1}$$

where $c_i(x^c, \tau, w)$ is the product i average cost rate for a cycle given a policy x^c and τ and total workload w.

Traditionally, the average cost per cycle is calculated by integrating the cost over a production cycle and then dividing by the cycle length. Using the notation we have developed, inventory holding and backorder costs would usually have an expression like

$$c_i(x^c, \tau, w) = \frac{1}{\tau} \int_0^\tau h_i \bar{W}_i^-(t) + b_i \bar{W}_i^+(t) dt. \tag{3.2}$$

We perform the same form of calculation for our due-date problem but "interchange the order of integrations" as previously discussed. Instead of integrating over time in the cycle, we will integrate over orders filled over the cycle. We shall first cover $c_i(x^c, \tau, w)$ for customized items and then for standardized ones.

3.2.1 Customized Goods

From the previous results, the amount of work in the system at the start of a cycle, which we denote by x_i^s, is equal to $x_i^c - \tau\rho_i(1-\rho_i)/2$. As stated before, the individual workload $\bar{W}_i(t)$ varies between $x_i^c - \tau\rho_i(1-\rho_i)/2$ and $x_i^c + \tau\rho_i(1-\rho_i)/2$ and so x_i^s is the smallest amount of orders in the system over the course of a cycle. This work must be stored in the system as orders with due-dates greater than $\bar{L}_i(0)$, by definition the earliest due-date in the system at time 0. The process $\bar{G}_i^I(s,t)$ denotes the number of goods at time t with due-date lead time s and thus is a more detailed measurement of orders than $\bar{W}_i$. Their relation at time zero is summarized by

$$x_i^s = \int_{\bar{L}_i(0)}^{\infty} \bar{G}_i^I(s,0)ds. \tag{3.3}$$

Moreover, we have

Proposition 3.1 *For $s > \bar{L}_i(0)$, $\bar{G}_i^I(s,0) = \rho_i\bar{F}_i^c(s)$ (where $\bar{F}_i^c(s)$ is defined as $1 - \bar{F}_i(s)$) and is 0 for $s < \bar{L}_i(0)$. Additionally,*

$$x_i^s = \int_{\bar{L}_i(0)}^{\infty} \rho_i\bar{F}_i^c(s)ds. \tag{3.4}$$

Proof: The orders at time t with due-date lead time s is bounded by the maximum amount of work which could have arrived with a due-date of $t+s$. At time 0, the maximum orders with lead time s is the recently arrived work plus work which arrived r units in the past with a due-date of $s+r$. Notationally, this is $\int_0^\infty \rho_i\bar{f}_i(s+r)dr$ which is equal to $\rho_i\bar{F}_i^c(s)$. Thus, $\bar{G}_i^I(s,0) \le \rho_i\bar{F}_i^c(s)$. The only way $\bar{G}_i^I(s,0)$ is strictly less than $\rho_i\bar{F}_i^c(s)$ would be if orders with due-dates higher than $\bar{L}_i(0)$ were worked on. Since an earliest due-date policy is being used and the rate of the production process is 1, which is strictly greater than $\rho_i\bar{F}_i^c(s)$, then $\bar{G}_i^I(s,t)$ at the next instant either vanishes or is untouched by service and only affected by arrivals. At time $t=0$, the machine has just switched out of producing class i. implying that the work above $\bar{L}_i(0)$ has not been touched and that there has been no opportunity for orders to arrive with due-date below $\bar{L}_i(0^-)$, the earliest due-date on hand the instant before the switch in setup. Thus, $\bar{G}_i^I(s,0) = \rho_i \bar{F}_i^c(s)$ for $s > \bar{L}_i(0)$ and is 0 below. x_i^s equal to $\int_{\bar{L}_i(0)}^{\infty} \rho_i \bar{F}_i^c(s) ds$ follows as a consequence. ∎

We can describe how $\bar{G}_i^I(s,t)$ and $\bar{L}_i(t)$ evolve over the course of the cycle. From time $t = 0$ until $t = \tau(1-\rho_i)$ no services occur and only orders arrive to the queue. Since the orders corresponding to the $\bar{G}_i^I(\bar{L}_i(0), 0)$ work age and so get closer to their due-date, the earliest due-date in class i, $\bar{L}_i(t)$, is monotonically decreasing for $t \in (0, \tau(1-\rho_i))$. From time $t = \tau(1-\rho_i)$ until τ, $\bar{L}_i(t)$ is monotonically increasing since the service rate is always greater than $\bar{G}_i^I(s,t)$, that is, orders are being filled faster than they can age.

The functions $\bar{G}_i^I(s,t)$ and $\bar{L}_i(t)$ are natural tools to understand the behavior of the system over the course of a cycle. The interaction of $\bar{G}_i^I(s,t)$ and $\bar{L}_i(t)$, however, is complex. In order to simplify our calculations, we will create another version of $\bar{G}_i^I(s,t)$ which will track its behavior. Let $\bar{G}_i(s,t)$ be the amount of work at time t with due-date lead time s *if the machine did no work on the product class from* $t = 0$ *until* τ; $\bar{G}_i(s,t)$ is only defined for $t \in (0,\tau)$. Thus, $\bar{G}_i(s,t) = \bar{G}_i^I(s,t)$ for $s > \bar{L}_i(t)$ but is not necessarily 0 for $s < \bar{L}_i(t)$. $\bar{G}_i(s,t)$ is useful for two reasons: first it is easy to describe its behavior; second $\bar{L}_i(t)$ can be found directly from it. The evolution of $\bar{G}_i(s,t)$ is described by the following proposition:

Proposition 3.2 *For* $t \in (0,\tau)$

$$\bar{G}_i(s,t) = \rho_i(\bar{F}_i^c(s) - \bar{F}_i^c(s+t)) + \bar{G}_i^I(s+t, 0). \tag{3.5}$$

Proof: By construction $\bar{G}_i(s,t)$ evolves according to the differential equation

$$\bar{G}_i(s, t+dt) = \bar{G}_i(s+dt, t) + \rho_i \bar{f}_i(s). \tag{3.6}$$

That is, at the next dt instant of time the amount of work due s units in the future is equal to the amount of work previously in the system with due-date lead time of $s+dt$ which has aged dt time units plus the amount of work with a due-date lead time of s which has just arrived. Given that $\bar{G}_i(s,t)$ equals $G_i^I(s,t)$ at $t=0$, the proposition follows as the solution to equation 3.6. ∎

The evolution of $G_i(s,t)$ has some noticeable characteristics. For a lack of terminology, we label some of the more prominent features for future reference. At time t for $s \geq \bar{L}_i(0)-t$, $\bar{G}_i(s,t)$ is the "steady state distribution" $\rho_i \bar{F}_i^c(s)$. The point $\bar{L}_i(0)-t$ marks a barrier above which there is a surge of orders and, forgive the colloquial simile, acts as a storm "front" of orders moving closer to its due-date. Below the "front" orders are gradually building-up from 0. At the point $\bar{a}_i = \tilde{a}_i/\sqrt{n}$, $\bar{G}_i(\bar{a}_i,t)$ accumulates no more orders. Let's call this the lowest due-date "line": work corresponding to $\bar{G}_i(s,t)$ below the line does not grow, but drifts at the same level toward or past its due-date.

We can calculate $\bar{L}_i(t)$ from $\bar{G}_i(s,t)$. When the machine is servicing other classes, the order with the earliest due-date corresponds either to the earliest due-date request just as the server switched out of product class i (i.e. $\bar{L}_i(0) \leq \bar{a}_i$) or the request with the due-date lead time of $\bar{a}_i$ which just arrived after the machine switched out (i.e. $\bar{L}_i(0) > \bar{a}_i$). Thus, for $t \in (0, \tau(1-\rho_i))$, $\bar{L}_i(t)$ equals $\min[\bar{a}_i - t, \bar{L}_i(0)-t]$. For $t \in (\tau(1-\rho_i), \tau)$, the server should have completed $t-\tau(1-\rho_i)$ units of work. Since the machine works on earliest due-date first, work corresponding to $\bar{G}_i(s,t)$ for s below $\bar{L}_i(t)$ has been completed. Thus we have the following proposition,

Proposition 3.3 *$\bar{L}_i(t)$ is the smallest quantity which satisfies*

$$t-\tau(1-\rho_i) = \int_{-\infty}^{\bar{L}_i(t)} \bar{G}_i(s,t)ds. \tag{3.7}$$

This uniquely determines $\bar{L}_i(t)$.

Given $\bar{L}_i(t)$ we can now state class i's average inventory costs per cycle for a given x_i^c and τ. It is equal to the average of the earliness or tardiness costs associated with orders as they are filled. Since the machine follows an earliest due-date policy, the holding or backorder time is either $\bar{L}_i(t)$ or the due-date associated with the arrival if that arrival had a due-date less than $\bar{L}_i(t)$. If there are arrivals with due-dates s less than $\bar{L}_i(t)$, then the machine spends $\rho_i \bar{f}_i(s)$ fraction of effort on them and $1-\rho_i \bar{F}_i(\bar{L}_i(t))$ fraction of effort on products with due-date $\bar{L}_i(t)$. Thus the cost for a class i good is

$$c_i(x_i^c, \tau, w) = \frac{1}{\tau} \int_{\tau(1-\rho_i)}^{\tau} \left(b_i \bar{L}_i^-(t) + h_i \left[\left(1 - \rho_i \bar{F}_i(\bar{L}_i(t))\right) \bar{L}_i^+(t) + \int_0^{L_i(t)} \rho_i \bar{f}_i(s) s ds\right]\right) dt. \tag{3.8}$$

Although this formulation is complex, the average cost per cycle is computable. The HTAP has dramatically simplified the problem. In Teneketsis, an infinite dimensional state space was needed to track the evolution of orders with due-dates over time. The fluid limit has transformed order progression through this infinite dimensional state space into $\bar{G}_i^I(s,t)$. From a functional analysis point of view, the ideas are nearly identical because $\bar{G}_i^I(\cdot,t)$ is a bounded function on a compact domain and so a point in the infinite dimensional space of square integrable functions L^2. The fluid limit thus approximates the evolution of orders in the system as a path in L^2, parameterized by the index t in $\bar{G}_i^I(\cdot,t)$. Although this relationship is abstract, the path in L^2 is easily made calculable by propositions 3.1 – 3.3. Moreover, by "reversing the order of integration," we are able to take advantage of this phenomena and translate $G_i^I(s,t)$ into average cost per cycle.

3.2.2 Standardized Goods

The difficulty in directly translating the customized goods calculation to a standardized one is that with customized goods we permanently assign an item to an order before the order's due-date is past while with standardized goods this is not necessarily the case. For example, when there are few orders in the system and the machine is about to switch out of producing a given class, the earliest due-date in the system can be quite high (notationally, $\bar{L}_i(\tau^-) >> \bar{a}_i$). In order to perform $\rho_i \tau$ units of work during the cycle, the machine must keep busy and is forced to work on orders that may be due far in the future. The customized good is then shelved and held on the shipping dock until its due-date. In the standardized case, this does not have to be done. At τ^-, the instant before switching out of the product class, we would prefer to make the product but assign it to the order that will arrive an instant later with due-date of $\bar{a}_i$. That is, at the time of an order's due-date, we like to take the item off the finished goods inventory shelf and then ship it. A standardized product is not fixed to any given order.

Therefore, we need to keep track of the finished goods inventory over the course of the cycle in addition to the number of orders (represented by $\bar{G}_i(s,t)$ and $\bar{L}_i(t)$) because we would like to have both orders and finished goods present at the same time. Let $\bar{H}_i(t)$ represent the amount of product i work in finished goods inventory at time t (time is in the fluid scaling and WLOG we assume that at time $t = 0$ the server has just switched out of product i, as was done in the customized case). An important aspect of $\bar{H}_i(t)$ is that it must be positive as it represents actual goods in inventory: backorders are in the form of unfulfilled orders in $\bar{G}_i(s,t)$. Thus, we can view total work to be done (i.e. work requested beyond our immediate capability to service) at time t as

$$\bar{W}_i(t) = \int_{L_i(t)}^{\infty} \bar{G}_i(s,t)ds - \bar{H}_i(t). \tag{3.9}$$

Thus, product i work to be done is equal to the amount of work currently being requested minus the work stored in inventory which could be assigned to the orders.

One can think of product and order flow somewhat differently for the standardized case than customized one. Orders with due-dates and products enter the system like a fluid as before. Orders enter a queue represented by $\bar{G}_i(s,t)$ and products enter the finished goods inventory represented by $\bar{H}_i(t)$. Orders "leave the system" when a product is assigned to them. Products "leave the system" when assigned to orders. If $\bar{L}_i(t)$ is greater than zero and a product is assigned to the earliest due-date then the order will sit in the shipping area and an earliness cost will be incurred. If $\bar{L}_i(t)$ is less than zero and a product is assigned, a tardiness cost is incurred. Products also accumulate holding cost by sitting in inventory. Therefore, the cost per cycle is then equation 3.8 plus the holding cost due to finished goods $\bar{H}_i(t)$ which is:

$$\int_0^T h_i \bar{H}_i(t)dt. \tag{3.10}$$

We can now make a few quick statements about the behavior of the system. Since the holding costs for goods sitting on the shipping dock are the same as those in the inventory, there is no benefit for assigning goods to orders early. More flexibility is created (and hence a better policy) by not assigning goods to orders if $\bar{L}_i(t)$ is strictly greater than zero and instead storing the products as finished goods inventory.

We then have $\bar{L}_i(t) \leq 0$ for all t. The rationale for this is simple: if for some unexplainable reason $\bar{L}_i(t) > 0$ (for instance when a rare event suddenly shifts the total workload level $W_i(t)$) then we shall assign no product to orders and place them in finished goods inventory. The earliest due-date lead time $\bar{L}_i(t)$ will then decrease to zero and never again go higher. This is a transitory effect which will be washed away after the repetition of several cycles and so can be ignored. From this we can conclude that for $s \geq 0$,

$$\bar{G}_i(s,t) = \rho_i \bar{F}_i^c(s) \tag{3.11}$$

for all t. Moreover, $\bar{L}_i(t)$ is less than zero only when there are no finished goods, i.e. $\bar{H}_i(t) = 0$. This is true because we allocate goods from inventory to prevent backorders.

From this we get the useful relation that $\bar{H}_i(t)$ is positive only if

$$\bar{W}_i(t) < \int_0^\infty \rho_i \bar{F}_i^c(s)ds$$

and so

$$\bar{H}_i(t) = \left(\bar{W}_i(t) - \int_0^\infty \rho_i \bar{F}_i^c(s)ds\right)^- . \tag{3.12}$$

It is important to notice that we already know the behavior of $\bar{W}_i(t)$ from Chapter 3.1. Thus, the holding cost portion of the cost per cycle is just a translation in terms of workload (or equivalently cycle center x_i^c) of the holding cost portion of $c_i(x^c, \tau, w)$ – a cycle center translation by $\int_0^\infty \rho_i \bar{F}_i^c(s)ds$.

The tardiness portion of the cost per cycle is also a workload translation by $\int_0^\infty \rho_i \bar{F}_i^c(s)ds$ of the SELSP costs. When there are backorder costs, the work in finished goods inventory is zero. Thus by equation (3.9) and the relation in equation (3.11), we have

$$\bar{W}_i(t) = \int_{\bar{L}_i(t)}^{\infty} \bar{G}_i(s,t)ds \tag{3.13}$$

which is

$$\bar{W}_i(t) = \int_{\bar{L}_i(t)}^{0} \rho_i ds + \int_0^\infty \rho_i \bar{F}_i^c(s)ds. \tag{3.14}$$

Noting that tardiness only occurs if $\bar{L}_i(t)$ is less than zero, we conclude that in the backorder regions

$$\rho_i \bar{L}_i^-(t) = \left(\bar{W}_i(t) - \int_0^\infty \rho_i \bar{F}_i^c(s)ds\right)^+ . \tag{3.15}$$

Thus the time average backordering cost for a given total workload is

$$\frac{b_i}{\rho_i \tau}\left[\int_{\tau(1-\rho_i)}^{\tau}\left(\bar{W}_i(t) - \int_0^\infty \rho_i \bar{F}_i^c(s)ds\right)^+ dt\right]. \tag{3.16}$$

Again, this is a translation of the SELSP cost per cycle.

Therefore, the cost per cycle for standardized goods with due-dates is exactly the same as in the SELSP with a cycle center shift by $\int_0^\infty \rho_i \bar{F}_i^c(s)ds$ which is just the average class i due-date lead time. Thus $c_i(x^c,\tau,w)$ is broken down into three regions based on if there is 1) only holding, 2) only backordering or 3) mixed costs over the cycle. The cost can be expressed as

$$c_i(x^c,\tau,w) = \begin{cases} h_i(\rho_i\bar{l}_i - x_i^c) & \text{if } \rho_i\bar{l}_i - x_i^c > \frac{\tau\rho_i(1-\rho_i)}{2} \\ (b_i+h_i)\frac{\tau\rho_i(1-\rho_i)}{8} + \frac{h_i-b_i}{2}(\rho_i\bar{l}_i - x_i^c) & \text{if } 0 \in [\rho_i\bar{l}_i - x_i^c \pm \frac{\tau\rho_i(1-\rho_i)}{2}] \\ \quad +\frac{b_i+h_i}{2\tau\rho_i(1-\rho_i)}(\rho_i\bar{l}_i - x_i^c)^2 & \\ -b_i(\rho_i\bar{l}_i - x_i^c) & \text{if } \rho_i\bar{l}_i - x_i^c < -\frac{\tau\rho_i(1-\rho_i)}{2} \end{cases} . \tag{3.17}$$

3.3 Optimization

With an expression for average cost for each level of workload we can optimize over our policy as determined by x^c, τ and w_0. The generality of the due-date distribution prevents an exact solution. A numerical method, however, is possible. Policy optimization must be performed on both the fluid and the diffusion levels. We first motivate the nature of the optimization in each scaling and then suggest an algorithm for the whole problem.

Under the fluid scalings, the cycle center x^c can be optimized with respect to a given cycle length τ and total workload level w. This is a non-linear optimization which we call the Cycle Center Optimization (CCO) program. It can be stated as follows

$$\text{(CCO)} \quad \min_{x^c \in \Re^N} \ \sum_{i=1}^N c_i(x^c, \tau, w) \tag{3.18}$$

$$\text{such that:} \quad \sum_{i=1}^N x_i^c = w \tag{3.19}$$

$$x_i^c \geq \frac{\tau \rho_i (1-\rho_i)}{2} \quad \text{for } i = 1, \ldots, N^c. \tag{3.20}$$

The highly non-linear aspects of $\bar{L}_i$ and thus $c_i(x^c, \tau, w)$ make this problem complex. Nonetheless, the solution is a function of τ and w and we denote it by $x^{c^*}(\tau, w)$.

In the diffusion scheme, the long run average cost of the entire problem is minimized. Because of the presence of setup times, the cycle length τ affects the drift of the diffusion process. Thus the minimization of equation 2.4 reduces to a diffusion control problem with parameters $\tau(w)$ and w_0 given that the optimal cycle center $x^{c^*}(\tau, w)$ is known. Although the drift of $W(t)$ is unbounded, we assume that the dynamic programming formulation applies (see Mandl [29]). Let $V(w)$ be the potential (relative value) function and g be the gain. The associated Jacobi-Hamilton-Bellman equations are

$$\min_{\tau(w)} \left\{ c(x^{c^*}(\tau, w), \tau, w) - g + \left(\frac{s}{\tau(w)} - c \right) V'(w) + \frac{\sigma^2}{2} V''(w) \right\} = 0 \quad \text{for } w \leq w_0 \tag{3.21}$$

and

$$V'(w) = 0 \quad \text{for } w \geq w_0. \tag{3.22}$$

The minimization on the left hand side of equation 3.21 is important and we call it the Cycle Length Optimization (CLO) program. Using the definition of average cost per cycle, the program can be written as

$$\text{(CLO)} \quad \min_{\tau>0} \sum_{i=1}^{N} c_i(x^{c^*}(\tau, w), \tau, w) + \frac{K + sV'(w)}{\tau} - g - cV'(w) + \frac{\sigma^2}{2}V''(w). \tag{3.23}$$

The quantity $K + sV'(w)$ represents the effective cost of setups during a cycle given a total workload level of w. As the drift, or the average rate of change, of the diffusion process is affected by the percent of time wasted setting up, the rate of change of the potential function, $V'(w)$, measures the marginal cost of time for a workload level w. We let S equal to $K + sV'(w)$ be the aggregate form of setup penalty. It is important to note that if we have a closed form solution to cost per cycle in terms of τ, w and $V'(w)$ we can solve for optimal τ by differentiating equation (3.23) with respect to τ and solving against zero.

Since problem (3.21) cannot be solved analytically, we pursue a numerical solution. The algorithm we propose implements these two levels of optimization. In particular, the *Markov chain approximation* technique developed by Kushner [25] is employed. This method systematically discretizes both time and the state space, and approximates a diffusion control problem by a control problem for a finite state Markov chain. Weak convergence methods have been developed by Kushner and his colleagues to verify that the controlled Markov chain (and its corresponding optimal cost) approximates arbitrarily closely the controlled diffusion process (and its corresponding optimal cost); we refer readers to Kushner and Dupuis [26] for an up-to-date account of this research area, and retain most of their notation for ease of reference.

Before describing the method, we introduce a slight modification to the heavy traffic analysis to account for the fact that setup times do not vanish in the original problem. The cycle length $\tau(w)$ consists of the time devoted to processing and the time allocated to setups. In the fluid scaling, $s/\sqrt{n}$ units of time are spent setting up over the course of a cycle; although this quantity vanishes in the limit, we include it in our analysis as an intended refinement. More specifically, we replace $\tau(w)$ by $\tau(w) + s/\sqrt{n}$.

Let h denote the *finite difference interval*, which dictates how finely both the state space and time are discretized. One can consider a sequence of controlled Markov chains indexed by the interval h, and as the value of h becomes smaller the resulting discrete time, finite state Markov chain described below becomes a better approximation of the controlled diffusion process.

To numerically solve (3.21), we need to confine the one-dimensional diffusion process W to a bounded region. Since W resides on a halfline, the state space of the controlled Markov chain will be $\{-M, -M+h, \ldots, M-h, M\}$, where M is a positive integer multiple of h. For now let us fix the idling threshold w_0 (this parameter will be optimized later on) such that $|w_0| \leq M$ and w_0 is an integer multiple of h. Hence, the Markov chain actually resides in $\{w_0, w_0+h, \ldots, M-h, M\}$. The approximating Markov chain has nonzero transition probabilities

$$P^h(w, w+h) = \frac{\sigma^2 + 2h\left(c - \frac{s}{\tau(w)+s/\sqrt{n}}\right)^+}{2\sigma^2 + 2h\left|c - \frac{s}{\tau(w)+s/\sqrt{n}}\right|} \tag{3.24}$$

and

$$P^h(w, w-h) = \frac{\sigma^2 + 2h\left(c - \frac{s}{\tau(w)+s/\sqrt{n}}\right)^-}{2\sigma^2 + 2h\left|c - \frac{s}{\tau(w)+s/\sqrt{n}}\right|} \tag{3.25}$$

on the interior of the state space, and the time intervals, or *interpolation intervals*, are of length

$$\Delta t^h = \frac{h^2}{\sigma^2 + h\left|c - \frac{s}{\tau(w)+s/\sqrt{n}}\right|}\ . \tag{3.26}$$

Two issues need to be addressed to obtain our approximating controlled Markov chain: (i) for an ergodic cost problem, the interpolation interval Δt^h must be independent of the state w and control $\tau(w)$ [26, page 209], and (ii) the behavior of the Markov chain at the boundary states $w = M$ and $w = w_0$. To deal with the first issue, we define $Q^h = \sigma^2 + \max_{w,\tau(w)} h\left|c - \frac{s}{\tau(w)+s/\sqrt{n}}\right|$. Since the smallest nonzero value of

$\tau(w)$ is h, we let $Q^h = \sigma^2 + |ch - s|$, and define the new nonzero interior transition probabilities

$$\bar{P}^h(w, w+h) = \frac{\sigma^2 + 2h\left(c - \frac{s}{\tau(w)+s/\sqrt{n}}\right)^+}{2Q^h} , \tag{3.27}$$

$$\bar{P}^h(w, w-h) = \frac{\sigma^2 + 2h\left(c - \frac{s}{\tau(w)+s/\sqrt{n}}\right)^-}{2Q^h} \tag{3.28}$$

and

$$\bar{P}^h(w, w) = 1 - \frac{\left(\sigma^2 + h\left|c - \frac{s}{\tau(w)+s/\sqrt{n}}\right|\right)}{Q^h} , \tag{3.29}$$

and the new interpolation interval

$$\Delta t^h = \frac{h^2}{Q^h} . \tag{3.30}$$

Now we consider the boundary states. A reflecting boundary is employed at the idling threshold. However, the Markov chain approximation method assumes that $\Delta t^h = 0$ for a reflecting boundary state. Because the interpolation interval Δt^h takes on a value different than (3.30) at w_0, this boundary state must be eliminated. We define the transition probability [26, page 212]

$$\tilde{P}^h(w_0 - h, w_0 - h) = 1 - \bar{P}^h(w_0 - h, w_0 - 2h) . \tag{3.31}$$

We also impose a reflecting boundary at state M, and define the transition probability

$$\tilde{P}^h(M - h, M - h) = 1 - \bar{P}(M - h, M - 2h) . \tag{3.32}$$

Although the reflecting barrier at M is artificial in the sense that $P(M, M - h)$ would be positive if the boundary was chosen to be larger than M, the effect of this approximation should be negligible if the value of M is sufficiently large, and consequently visited sufficiently infrequently. In our implementation, the size of the Markov chain is chosen so that a further increase in M does not change the optimal solution $(\tau(w), w_0)$. In summary, our approximating Markov chain has state space

$\{w_0 + h, w_0 + 2h, \ldots, M - 2h, M - h\}$, interpolation interval defined by (3.30), and nonzero transition probabilities $\tilde{P}^h(w, y)$ defined by (3.31)-(3.32) and $\tilde{P}^h(w, y) = \bar{P}^h(w, y)$ otherwise, where $\bar{P}^h(w, y)$ are defined in equations (3.27)-(3.29).

The dynamic programming optimality equation for the controlled Markov chain is given by [26, eqn 5.3, page 204]

$$V(w) = \sum_y \tilde{P}^h(w, y)V(y) + \left(c(x^{c^*}, \tau, w) - g\right) \Delta t^h \,, \tag{3.33}$$

and the Markov chain control problem can be solved using the following policy improvement algorithm. First, we choose the initial policy: $\tau(w) = w - w_0$ for $w \geq w_0$ and arbitrary w_0. In the evaluation step of the algorithm, a generic policy $(\tau(w), w_0)$ is evaluated (that is, $V(w)$ and g are found) recursively. Since the Markov chain is a birth-death process, the stationary probability distribution π_w for $w \in \{-M + h, ..., M - h\}$ is

$$\pi_w = \begin{cases} 0 & \text{for } w < w_0 + h \\ \pi_{w_0 - h} \prod_{k=w+1}^{w_0 - h} \frac{\tilde{P}^h(k, k-h)}{\tilde{P}^h(k-h, k)} & \text{for } w > w_0 + h \\ \left(1 + \sum_{l=-N+h}^{w_0 - h} \prod_{k=l+h}^{w_0 - h} \frac{\tilde{P}^h(k, k-h)}{\tilde{P}^h(k-h, k)}\right)^{-1} & \text{for } w = w_0 + h \end{cases} \,, \tag{3.34}$$

and the gain is

$$g = \sum_{w=-N+h}^{w_0 - h} \pi_w c(x^{c^*}, \tau(w) + s/\sqrt{n}, w). \tag{3.35}$$

We set $V(-M) = 0$, so that $V(w) = 0$ for $w \leq w_0$, by equation (3.22). For $w > w_0$, equation (3.33) implies that $V(w + h)$ can be calculated recursively by

$$V(w + h) = \frac{g - c(x^{c^*}, \tau(w), w) + (1 - \tilde{P}^h(w, w))V(w) - \tilde{P}^h(w, w - h)V(w - h)}{\tilde{P}^h(w, w + h)}. \tag{3.36}$$

In the policy improvement step we first solve for the cycle length $\tau(w)$ and then for the idling threshold w_0. The cycle length is determined by minimizing the right side of (3.33) with respect to $\tau(w) \geq 0$.

Therefore, with $V(w)$ and g determined an improvement iteration on τ can be done using equation 3.21 and on w_0 by finding the threshold which minimizes the gain g. The algorithm ends when τ and w_0 converge.

When there is no setup time, however, the diffusion control problem is trivial. Although the cycle length τ and cycle center x^c must still be optimized for each total workload level, they can be done individually without the need to refer to or update the potential function $V'(w)$. As the steady state distribution of the total workload process $W(t)$ is exponential, the idling threshold can be found by

$$w_0 = \text{argmin}_{w'} \int_{w'}^{\infty} c(w) \frac{2c}{\sigma^2} e^{-\frac{2c}{\sigma^2}(w-w')} dw \tag{3.37}$$

where c and σ^2 are the parameters for the RBM $W(t)$.

3.4 Proposed Policies

The optimal cyclic policy derived under the heavy traffic limits needs to be unscaled to be implemented. The translation to an unnormalized policy has much room for interpretation. The difficulty lies in the dimensionality of the problem. Under the fluid scalings the individual workload processes jointly form a 1-dimensional deterministic path in a constant total workload hyperplane. It is a closed loop specified by cycle center x^{c^*} and cycle length τ. The heavy traffic policy defines a 1-dimensional loop for every value of the total workload, another 1-dimensional process. This creates a 2-dimensional structure from which we need to infer how to control the machine over the N dimensions of possible queue length values in the unnormalized system.

We propose an intuitive policy. The heavy traffic theory finds a minimum level of work experienced by an individual queue over the course of a cycle for every total workload level. We suggest the the machine continue production on the current class until this minimum amount of orders is reached, then begin setup for the next class in the cycle and idle when the total workload level has fallen to the idling threshold. This can be regarded as a dynamic order up to policy. In the unscaled system parameters, the dynamic order up to level can be stated as a switch from the product class i when the unscaled order level $\tilde{I}_i$ satisfies

$$\tilde{I}_i \leq \sqrt{n}\mu_i \left[x_i^{c^*}\left(\frac{\sum_{j=1}^{N} \mu_j^{-1} \tilde{I}_j}{\sqrt{n}}\right) - \frac{\tau\left(\frac{\sum_{j=1}^{N} \mu_j^{-1} \tilde{I}_j}{\sqrt{n}}\right)\rho_i(1-\rho_i)}{2} \right] \tag{3.38}$$

where $x_i^{c^*}(w)$ and $\tau(w)$ are determined from the optimization detailed in section Chapter 3.5 and idle when

$$\sum_{i=1}^{N} \mu_i^{-1} \tilde{I}_i = w_0\sqrt{n}. \tag{3.39}$$

More complicated policies can be created. The 1-dimensional workload path in the fluid limit additionally specifies the relative evolution of the individual workload processes. For example, the heavy traffic theory predicts that when a product has reached its minimum level and the setup for the next class begins, workload of the product class being setup is $x_i^{c^*} + \tau\rho_i(1-\rho_i)/2$. Thus for any given total workload, the HTAP predicts that a machine following a dynamic cyclic policy would have a fixed range of individual workload fluctuations. This can be thought of as a workload "confidence interval." Such intervals for three workloads in a hypothetical five product case are given in Figure 3-2. This type of information can be used to determine when random events throw the cycle off course in a manner not predicted by the HTAP's $O(\sqrt{n})$ fluid scaling. The dynamic cyclic approach can thus give "red flags" or warnings for unusual surges in demand or slow-downs in machine processing. A more complicated policy using this information would be able to specify when to skip product classes in the cycle or when to jump to a product class that has unusually high number of orders. Thus, the HTAP is sufficiently robust to suggest more complicated scheduling techniques.

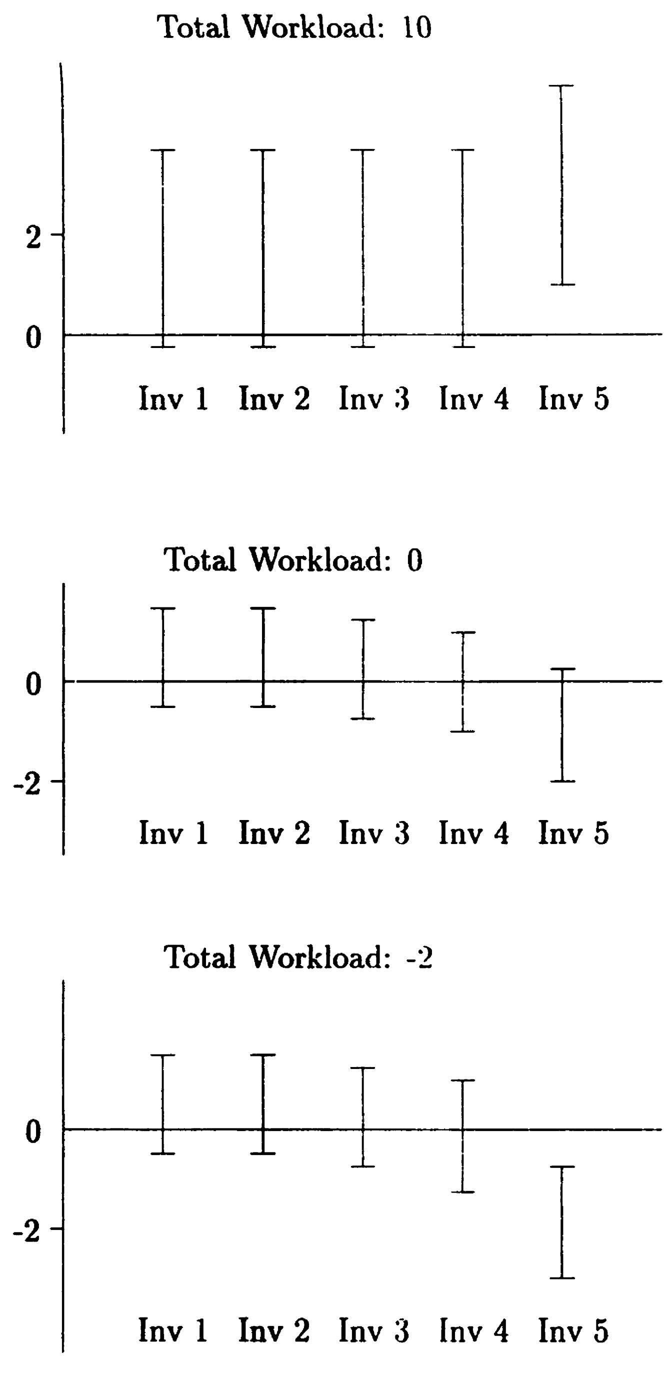

Figure 3-2: Hypothetical Workload Fluctuation "Confidence Intervals."

Chapter 4

Deterministic Due-Dates

In this chapter we investigate systems with deterministic due-dates. By explicitly stating a deterministic due-date lead time distribution, we are able to write a closed form expression for the cost per cycle (equation 3.1). In addition we are able to state a quick method for determining optimal cycle center x^{c^*} and provide a formulaic solution. With this, cycle length τ is found in terms of $V'(w)$ and total workload w. The diffusion control algorithm must still be performed. A numerical study is done to test our proposed policies. Lastly, we discuss how our proposed policy changes over the range of systems outlined in the Literature Review.

4.1 Cost per Cycle

Let each product class i have a deterministic due-date of f_i. Under the fluid scalings the due-date lead time is $\bar{f}_i = f_i/\sqrt{n}$. The cumulative distribution function is then

$$F_i^c(s) = \begin{cases} 0 & \text{if } s > f_i \\ 1 & \text{if } s \leq f_i \end{cases} . \tag{4.1}$$

4.1.1 Standardized Product Classes

From the discussion in Chapter 3.4, the average cost per cycle for a class i standardized good is not difficult to determine. According to equation (3.17), for a given cycle center

x_i^c, cycle length τ and total workload level w the cost per cycle is

$$c_i(x^c,\tau,w) = \begin{cases} h_i(\rho_i\bar{f}_i - x_i^c) & \text{if } \rho_i\bar{f}_i - x_i^c > \frac{\tau\rho_i(1-\rho_i)}{2} \\ (b_i+h_i)\frac{\tau\rho_i(1-\rho_i)}{8} + \frac{h_i-b_i}{2}(\rho_i\bar{f}_i - x_i^c) & \text{if } 0 \in [\rho_i\bar{f}_i - x_i^c \pm \frac{\tau\rho_i(1-\rho_i)}{2}] \\ \quad +\frac{b_i+h_i}{2\tau\rho_i(1-\rho_i)}(\rho_i\bar{f}_i - x_i^c)^2 & \\ -b_i(\rho_i\bar{f}_i - x_i^c) & \text{if } \rho_i\bar{f}_i - x_i^c < -\frac{\tau\rho_i(1-\rho_i)}{2} \end{cases} \tag{4.2}$$

since the mean $\bar{l}_i$ of the deterministic due-date is $\bar{f}_i$. For ease of reference, if the parameters are such that $\rho_i\bar{f}_i - x_i^c > \tau\rho_i(1-\rho_i)/2$, we call product i in *condition 1*. Similarly, we call product i in *condition 2* if $0 \in [\rho_i\bar{f}_i - x_i^c \pm \tau\rho_i(1-\rho_i)/2]$ and in *condition 3* if $\rho_i\bar{f}_i - x_i^c < -\tau\rho_i(1-\rho_i)/2$.

4.1.2 Customized Product Classes

To find the optimal cost per cycle for a customized product, we must determine the behavior of $\bar{G}_i(s,t)$, $\bar{L}_i(t)$ and $\bar{G}_i^I(s,t)$ given x_i^c and τ. By Proposition 3.1, we have

$$x_i^s = \int_{\bar{L}_i(0)}^{\bar{f}_i} \rho_i ds. \tag{4.3}$$

Solving for $\bar{L}_i(0)$ we get

$$\bar{L}_i(0) = \bar{f}_i - x_i^s/\rho_i. \tag{4.4}$$

Therefore, $\bar{L}_i(0)$ has a range of $[-\infty, \bar{f}_i]$ since for customized products x_i^s is greater than or equal to zero. This makes intuitive sense: the highest earliest due-date is $\bar{f}_i$ since no orders can arrive with larger due-dates. From Proposition 3.1 we additionally have

$$\bar{G}_i^I(s,0) = \begin{cases} \rho_i & \text{if } s \in (\bar{f}_i - x_i^s/\rho_i, \bar{f}_i) \\ 0 & \text{otherwise} \end{cases} . \tag{4.5}$$

From Proposition 3.2, we have

$$\bar{G}_i(s,t) = \begin{cases} 0 & \text{if } s \leq \bar{f}_i - x_i^s/\rho_i - t \\ \rho_i & \text{if } \bar{f}_i - x_i^s/\rho_i - t \leq s \leq \bar{f}_i \\ 0 & \text{if } \bar{f}_i < s \end{cases} . \tag{4.6}$$

We can solve for $\bar{L}_i(t)$ by using Proposition 3.3. We get for $t \in (\tau(1-\rho_i), \tau)$

$$t - \tau(1-\rho_i) = \int_{\bar{f}_i - x_i^s/\rho_i}^{\bar{L}_i(t)} \rho_i ds, \tag{4.7}$$

and so

$$\bar{L}_i(t) = \bar{f}_i - x_i^s/\rho_i - (1-\rho_i)\tau + (1-\rho_i)(t - \tau(1-\rho_i))/\rho_i. \tag{4.8}$$

As $(1-\rho_i)(t-\tau(1-\rho_i))/\rho_i - (1-\rho_i)\tau$ is less than or equal to zero, $\bar{L}_i(t)$ is bounded above by $\bar{f}_i$. This again corresponds to our intuition that the earliest due-date lead time can at most be $\bar{f}_i$. It is also important to note that $\bar{L}_i(t)$ is a function of our policy variable x_i^c in the form of x_i^s.

With $\bar{L}_i(t)$ we can find the average cost per cycle from equation 3.8. Since no orders arrive with due-date lead times $\bar{L}_i(t)$ less than $\bar{f}_i$, we can rewrite equation (3.8) as

$$c_i(x^c, \tau, w) = \frac{1}{\tau} \int_{\tau(1-\rho_i)}^{\tau} \left(b_i \bar{L}_i^-(t) + h_i \bar{L}_i^+(t) \right) dt. \tag{4.9}$$

The cost per cycle is broken down into three cases depending on if there are only late products assigned to orders, only early products or both during the course of the cycle. If $\bar{L}_i(t)$ is always greater than zero over the course of the cycle, orders are always filled early and so equation (4.9) implies

$$c_i(x^c, \tau, w) = \frac{1}{\tau} h_i \int_{(1-\rho_i)\tau}^{\tau} \left[\bar{f}_i - x_i^s/\rho_i - (1-\rho_i)\tau + (1-\rho_i)(t - \tau(1-\rho_i))/\rho_i \right] dt. \tag{4.10}$$

This simplifies to

$$c_i(x^c, \tau, w) = h_i \rho_i \left[\bar{f}_i - \frac{x_i^s + \tau\rho_i(1-\rho_i)/2}{\rho_i} \right] \tag{4.11}$$

or just

$$c_i(x_i^c, \tau, w) = h_i\rho_i\left[\bar{f}_i - \frac{x_i^c}{\rho_i}\right]. \tag{4.12}$$

This is remarkably similar to the results from standard queueing theory where one would expect the average cost to be the unit holding cost times the average queue length. This likeness also makes intuitive sense: it is an application of Little's law which is a statement about the relationship between queue length and waiting time. In our notation, Little's Law states that given the utilization ρ_i is the arrival rate, $\bar{f}_i - x_i^c/\rho_i$ is the average waiting time then their product is the number in queue. Thus, by exchanging the order of integration, we have effectively re-proved Little's Law.

Similarly, if products are always late or $\bar{L}_i(t)$ is always less than 0 for all $t \in [0, \tau]$, then the cost is

$$c_i(x^c, \tau, w) = b_i\rho_i\left[\frac{x_i^c}{\rho_i} - \bar{f}_i\right]. \tag{4.13}$$

If orders are both late and early over the cycle and so $\bar{L}_i(t)$ is both positive and negative over the cycle then the average cost per cycle is

$$\begin{aligned} c_i(x^c, \tau, w) &= b_i\rho_i\left[\frac{(\bar{f}_i - \frac{x_i^c}{\rho_i})^2}{2\tau(1-\rho_i)} - \frac{1}{2}(\bar{f}_i - \frac{x_i^c}{\rho_i}) + \frac{1}{8}\tau(1-\rho_i)\right] \\ &\quad + h_i\rho_i\left[\frac{(\bar{f}_i - \frac{x_i^c}{\rho_i})^2}{2\tau(1-\rho_i)} + \frac{1}{2}(\bar{f}_i - \frac{x_i^c}{\rho_i}) + \frac{1}{8}\tau(1-\rho_i)\right]. \end{aligned} \tag{4.14}$$

In summary, for customized products the average cost per cycle is

$$c_i(x^c, \tau, w) = \begin{cases} h_i(\rho_i\bar{f}_i - x_i^c) & \text{for } \textit{condition 1} \\ (b_i + h_i)\frac{\tau\rho_i(1-\rho_i)}{8} + \frac{h_i - b_i}{2}(\rho_i\bar{f}_i - x_i^c) & \text{for } \textit{condition 2} \\ \quad + \frac{b_i + h_i}{2\tau\rho_i(1-\rho_i)}(\rho_i\bar{f}_i - x_i^c)^2 & \\ -b_i(\rho_i\bar{f}_i - x_i^c) & \text{for } \textit{condition 3} \end{cases}. \tag{4.15}$$

This leads to a remarkable result: equations (4.2) and (4.15) imply that for the deterministic due-date case the cost structure for customized and standardized products are *exactly the same.* The only difference between the two types is the restriction that for customized goods the cycle center x_i^c cannot be less than $\rho_i(1-\rho_i)\tau/2$. Due-dates have transformed customized products into quasi-standardized ones. We shall discuss the interpretation of this result later in Chapter 4.4.

4.2 Optimization

With an explicit formulation of average cost per cycle we can optimize the system. First, we suggest a form of solution to the CCO and CLO programs and secondly we detail further an algorithm for use in the diffusion control problem.

4.2.1 Cycle Center Optimization

We begin by showing the existence of a cost-minimizing cycle center x^c for a given total workload w and cycle length τ. Note that the cost function $c(\tau, x^c, w)$ is differentiable with respect to x^c and its derivative is continuous. If one ignores the constant workload constraint $\sum_{i=i}^{N} x_i^c = w$, for fixed τ the cost function in terms of x^c is piecewise-quadratic with linear edges; its second derivative is a nonnegative step function. Thus $c(\tau, x^c, w)$ is convex and the restriction of the cost function to the constant workload hyperplane determined by w is also convex. This fact implies the existence of a solution to the constrained minimization problem.

The major complication in solving the CCO program is the boundary conditions in equation 3.20. We exploit the structure of the objective function to determine the structure of x_i^{c*} and to find which x_i^{c*} are binding with respect to the inequality constraints. As per Bertsekas [3], the Lagrangian function of the CCO program with fixed cycle length τ and total workload level w is

$$L(x^c, \lambda, \mu) = c(x^c, \tau, w) + \lambda \left(\sum_{i=1}^{N} x_i^c - w \right) + \sum_{j=1}^{N^c} \mu_j \left(\frac{\tau \rho_j (1-\rho_j)}{2} - x_j^c \right) . \qquad (4.16)$$

The Kuhn-Tucker necessary conditions state that for local minimum x^{c*} there exists lagrangian multipliers λ^* and μ_j^*, for $j = 1, \ldots, N^c$, such that

$$\nabla_{x^c} L(x^{c^*}, \lambda^*, \mu^*) = 0 \tag{4.17}$$

$$\mu_j^* \geq 0 \tag{4.18}$$

$$\mu_j^* = 0 \quad \forall j \in \Theta^* \tag{4.19}$$

where Θ^* is the set of non-binding cycle centers, i.e. products j such that $x_j^{c^*} > \tau\rho_j(1-\rho_j)/2$. We suppress the dependence of Θ^* on τ and w for increased readability. For additional ease of reference, we categorize the binding products by their *condition*. We let Θ^{*1} be the set of product classes with binding cycle centers and in *condition 1* and Θ^{*2} be the set of binding product classes in *condition 2* (as the workload must reach zero for a product class to be binding and $\bar{f}_i$ is non-negative, no binding product class can be in *condition 3*). The product classes with binding cycle centers can be considered to be "maxed-out" because no more work can be performed on that product class. Equation (4.17) implies that for each component $i \leq N^c$ of the gradient of the Lagrangian function

$$\frac{\partial}{\partial x_i^c} c_i(x^{c^*}, \tau, w) + \lambda - \mu_i = 0 \tag{4.20}$$

and for $i > N^c$

$$\frac{\partial}{\partial x_i^c} c_i(x^{c^*}, \tau, w) + \lambda = 0. \tag{4.21}$$

Thus, for $i \in \Theta^*(\tau, w)$ and $i \leq N^c$

$$\frac{\partial}{\partial x_i^c} c_i(x^{c^*}, \tau, w) + \lambda = 0. \tag{4.22}$$

Therefore the derivatives of both the average cost per cycle for non-binding standardized and customized products satisfy an identical Kuhn-Tucker equation. We use this similarity of the cost structure between non-binding customized and standardized goods to determine some properties about the optimal solution. For this discussion we need to denote what is the cheapest non-binding product as it will play an important role. Let θ_b^* be the cheapest tardiness cost product in $\Theta^*(\tau, w)$ and θ_h^* be the cheapest earliness cost product. For ease of notation let θ^* equal θ_b^* if $w > \sum_{i=1}^{N} \rho_i \bar{f}_i$ and θ_h^* otherwise.

We begin by re-deriving equations (4.21) and (4.22) so as to remove λ. The total workload constraint can be used to eliminate one of the non-binding variables and then to express the earliness and tardiness costs of the non-binding product classes as a piecewise polynomial function of $|\Theta^*|-1$ variables, where $|\cdot|$ denotes the cardinality of a set. The non-binding cycle centers must sum to $\hat{w} = w - \sum_{i \notin \Theta^*} \tau\rho_i(1-\rho_i)/2$ by definition of the binding constraints. Thus for any $j \in \Theta^*$ the cost per cycle due to non-binding product classes is

$$\sum_{i \in \Theta^*} c_i(x_i^c, \tau, w) = \sum_{i \in \{\Theta^* \setminus j\}} c_i(x_i^c, \tau, w) + c_j(\hat{w} - \sum_{i \in \{\Theta^* \setminus j\}} x_i^c, \tau, w). \tag{4.23}$$

Over this reduced hyperplane, the polynomial order of the $|\Theta^*|-1$ variables fluctuates between one and two depending on whether $|\rho_i \bar{f}_i - x_i^c| > \frac{\tau\rho_i(1-\rho_i)}{2}$ or $|\rho_i \bar{f}_i - x_i^c| \leq \frac{\tau\rho_i(1-\rho_i)}{2}$, respectively. For the gradient with respect to the non-binding cycle centers to be equal to zero, each of the $|\Theta^*|-1$ variables must be quadratic. Consequently, at the optimal x^c, at least $|\Theta^*|-1$ of the $c_i(\tau, x_i^c, w)$'s are of order two, with the remaining component possibly being linear. To see this, suppose that some of $c_i(x_i^c, \tau, w)$'s are not of order two, and let j denote the index of such a term. If we eliminate x_j, the gradient equation can then be written as

$$\nabla_{x^c}\left[\sum_{i \in \{\Theta^* \setminus j\}} c_i(x_i^c, \tau, w) + c_j\left(\hat{w} - \sum_{i \in \{\Theta^* \setminus j\}} x_i^c, \tau, w\right)\right] = 0\ .$$

This equation will have a solution only if the remaining $N-1$ $c_i(x_i^c, \tau, w)$'s are quadratic.

The following proposition greatly simplifies our analysis.

Proposition 4.1 *If there are only* $|\Theta^*|-1$ *quadratic* $c_i(x_i^c, \tau, w)$ *terms in*

$$\sum_{i \in \Theta^*} c_i(x_i^c, \tau, w)$$

at the optimal x^c, then the linear term must be $c_{\theta^*}(x^c_{\theta^*}, \tau, w)$.

Proof: This fact is most easily seen by examining the function $c_i(x^c_i, \tau, w) + c_{\theta^*}(x^c_{\theta^*}, \tau, w)$, where $i \in \Theta^*$, c_i is linear and c_{θ^*} is quadratic. By (4.2) and (4.15), the sum is

$$h_i(\rho_i \bar{f}_i - x^c_i)^+ + b_i(\rho_i \bar{f}_i - x^c_i)^- + (b_{\theta^*} + h_{\theta^*})\frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{8} + \frac{b_{\theta^*}+h_{\theta^*}}{2\tau\rho_{\theta^*}(1-\rho_{\theta^*})}(\rho_{\theta^*}\bar{f}_{\theta^*} - x^c_{\theta^*})^2 + \frac{h_{\theta^*}-b_{\theta^*}}{2}(\rho_{\theta^*}\bar{f}_{\theta^*} - x^c_{\theta^*}). \tag{4.24}$$

The tradeoff between x^c_i and $x^c_{\theta^*}$ can be examined by looking at this sum along the line $x^c_i + x^c_{\theta^*} = w'$ (with w' arbitrary). Substituting $w' - x^c_{\theta^*}$ for x^c_i into equation (4.24) and taking the derivative with respect to $x^c_{\theta^*}$ yields

$$+h_i - \frac{h_{\theta^*} - b_{\theta^*}}{2} - \frac{b_{\theta^*} + h_{\theta^*}}{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}(\rho_{\theta^*}\bar{f}_{\theta^*} - x^c_{\theta^*}) \quad \text{if } \rho_i\bar{f}_i + x^c_{\theta^*} > w', \tag{4.25}$$

$$-b_i - \frac{h_{\theta^*} - b_{\theta^*}}{2} + \frac{b_{\theta^*} + h_{\theta^*}}{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}(\rho_{\theta^*}\bar{f}_{\theta^*} - x^c_{\theta^*}) \quad \text{if } \rho_i\bar{f}_i + x^c_{\theta^*} < w'. \tag{4.26}$$

Since the quadratic region of $x^c_{\theta^*}$ is restricted to the region $|\rho_{\theta^*}\bar{f}_{\theta^*} - x^c_{\theta^*}| \le \frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2}$, it follows that the quantity in (4.25) is less than or equal to $h_{\theta^*} - h_i$, and the quantity in (4.26) is greater than or equal to $b_i - b_{\theta^*}$. Hence, neither (4.25) nor (4.26) can equal zero unless h_i equals h_{θ^*} for $\rho_i\bar{f}_i + x^c_{\theta^*} > w'$ or b_i equals b_{θ^*} for $\rho_i\bar{f}_i + x^c_{\theta^*} < w'$. If the holding or backorder costs are equal, then the optimal $x^c_{\theta^*}$ satisfies $|\rho_{\theta^*}\bar{f}_{\theta^*} - x^c_{\theta^*}| = \frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2}$, resulting in multiple optimal solutions along the line $x^c_i + x^c_{\theta^*} = w'$ with $|\rho_{\theta^*}\bar{f}_{\theta^*} - x^c_{\theta^*}| \ge \frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2}$ and $|\rho_i\bar{f}_i - x^c_i| \ge \frac{\tau\rho_i(1-\rho_i)}{2}$. Although many of these solutions lie in the region where products i and N are linear, one of the optimal solutions occurs when at least $N-1$ of the cost components are quadratic (or on the border between quadratic and linear). Hence, if there is a linear cost component at the optimal x^c, it will be $c_{\theta^*}(\tau, x^c_{\theta^*}, w)$. ∎

As a consequence, the optimal cycle center, or average amount of orders per cycle, for product $i \in \{\Theta^* \setminus \theta^*\}$ is restricted to the region $[\rho_i\bar{f}_i - \tau\rho_i(1-\rho_i)/2, \rho_i\bar{f}_i + \tau\rho_i(1-\rho_i)/2]$, whereas product θ^*'s cycle center can be arbitrarily far from $\rho_{\theta^*}\bar{f}_{\theta^*}$. Intuitively, this fact suggests that product θ^*, which is the least cost product by our indexing convention, is the product that will hold the excess or deficit amounts of work when the total workload w fluctuates far from $\sum_i \rho_i \bar{f}_i$.

We now use Proposition 4.1 to find the optimal cycle center x^c. Without loss of generality, the workload constraint is used to eliminate $x^c_{\theta^*}$ from the cost function, so that $x^c_{\theta^*} = w - \sum_{i=1}^{N-1} x^c_i$. To find the optimal center, we take the gradient of the right hand side of (4.23) and set it equal to zero:

$$\nabla_{x^c}[\sum_{i\in\{\Theta^*\backslash\theta^*\}} c_i(x^c_i, \tau, w) + c_{\theta^*}(\hat{w} - \sum_{i\in\{\Theta^*\backslash\theta^*\}} x^c_i, \tau, w) + \frac{k}{\tau}] = 0. \tag{4.27}$$

At this point, we do not know whether $c_{\theta^*}(x^c_{\theta^*}, \tau, w)$ is linear or quadratic. Let $\bar{x^c}$ be the $(|\Theta^*| - 1)$-dimensional vector that solves (4.27) under the assumption that all Θ^* of the $c_i(x^c, \tau, w)$'s are quadratic in x^c_i. Taking the $(|\Theta^*| - 1)$-dimensional gradient, we find that $\bar{x^c}$ satisfies (w and τ multiply their vectors component-wise in the analysis below)

$$\frac{1}{\tau}A(\bar{f} - \bar{x^c}) - \gamma_1 - \frac{\sum_{i\in\Theta^*} \rho_i \bar{f}_i - \hat{w}}{\tau}\gamma_2 = 0, \tag{4.28}$$

where

$$A = \begin{bmatrix} \ddots & & 0 \\ & \frac{b_i+h_i}{\rho_i(1-\rho_i)} & \\ 0 & & \ddots \end{bmatrix} + \frac{b_{\theta^*} + h_{\theta^*}}{\rho_{\theta^*}(1-\rho_{\theta^*})} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix}, \tag{4.29}$$

$$\bar{f} = \begin{bmatrix} \vdots \\ \rho_i \bar{f}_i \\ \vdots \end{bmatrix}, \tag{4.30}$$

$$\gamma_1 = \begin{bmatrix} \vdots \\ \frac{b_i-h_i}{2} - \frac{b_{\theta^*}-h_{\theta^*}}{2} \\ \vdots \end{bmatrix}, \tag{4.31}$$

$$\gamma_2 = \begin{bmatrix} \vdots \\ \frac{b_{\theta^*}+h_{\theta^*}}{\rho_{\theta^*}(1-\rho_{\theta^*})} \\ \vdots \end{bmatrix}. \tag{4.32}$$

Thus,

$$\bar{x^c} = \bar{f} - \tau A^{-1}\gamma_1 - (\sum_{i\in\Theta^*} \rho_i \bar{f}_i - \hat{w})A^{-1}\gamma_2, \tag{4.33}$$

where the matrix elements of A^{-1} are

$$\alpha_{ij} = -\frac{\frac{\rho_i(1-\rho_i)}{b_i+h_i}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}{\sum_{l\in\Theta^*}\frac{\rho_l(1-\rho_l)}{b_l+h_l}} \quad \text{for} \quad i \neq j, \quad \text{and} \tag{4.34}$$

$$\alpha_{ii} = \frac{\rho_i(1-\rho_i)}{b_i+h_i}\frac{\sum_{l\in\{\Theta^*\setminus i\}}\frac{\rho_l(1-\rho_l)}{b_l+h_l}}{\sum_{l\in\Theta^*}\frac{\rho_l(1-\rho_l)}{b_l+h_l}} \tag{4.35}$$

If $|\rho_{\theta^*}\bar{f}_{\theta^*} - (\hat{w} - \sum_{i\in\{\Theta^*\setminus\theta^*\}}\bar{x_i^c})| \leq \frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2}$ then $c_{\theta^*}(x_{\theta^*}^c, \tau, w)$ is indeed quadratic and $\bar{x_i^c}$ determines the optimal center: $x_i^c = \bar{x_i^c}$ for $i \in \{\Theta^*\setminus\theta^*\}$ and $x_{\theta^*}^c = \hat{w} - \sum_{i=1}^{N-1}\bar{x_i^c}$. If $|\rho_{\theta^*}\bar{f}_{\theta^*} - (\hat{w} - \sum_{i\in\{\Theta^*\setminus\theta^*\}}\bar{x_i^c})| > \frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2}$, we must solve the multivariate gradient equation with the linear form of $c_{\theta^*}(x_{\theta^*}^c, \tau, w)$. With this substitution, equation (4.27) decomposes into univariate expressions of the form

$$\frac{h_i - b_i}{2} + \frac{b_i + h_i}{\tau\rho_i(1-\rho_i)}(\rho_i\bar{f}_i - x_i^c) - h_{\theta^*} = 0 \tag{4.36}$$

if $\rho_{\theta^*}\bar{f}_{\theta^*} - (\hat{w} - \sum_{i\in\{\Theta^*\setminus\theta^*\}}\bar{x_i^c}) > \frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2}$ and

$$\frac{h_i - b_i}{2} + \frac{b_i + h_i}{\tau\rho_i(1-\rho_i)}(\rho_i\bar{f}_i - x_i^c) + b_{\theta^*} = 0 \tag{4.37}$$

if $\rho_{\theta^*}\bar{f}_{\theta^*} - (\hat{w} - \sum_{i\in\{\Theta^*\setminus\theta^*\}}\bar{x_i^c}) < -\frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2}$. Putting the results from (4.33) and (4.36)-(4.37) together, we obtain a complete expression for the optimal cycle center. For

$i \in \{\Theta^* \setminus \theta^*\}$ we have

$$x_i^{c^*} = \begin{cases} \rho_i \bar{f}_i - \frac{\tau\rho_i(1-\rho_i)}{b_i+h_i}\left[\frac{b_i-h_i}{2} + h_{\theta^*}\right] \\ \qquad \text{if } (\sum_{i\in\Theta^*} \rho_i\bar{f}_i - \hat{w}) - \sum_{i\in\{\Theta^*\setminus\theta^*\}} \bar{x}_i^c > \frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2} \\ \rho_i \bar{f}_i - \tau\alpha_i \cdot \gamma_1 - (\sum_{i\in\Theta^*} \rho_i \bar{f}_i - \hat{w})\alpha_i \cdot \gamma_2 \\ \qquad \text{if } |(\sum_{i\in\Theta^*} \rho_i\bar{f}_i - \hat{w}) - \sum_{i\in\{\Theta^*\setminus\theta^*\}} \bar{x}_i^c| \le \frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2} \\ \rho_i \bar{f}_i - \frac{\tau\rho_i(1-\rho_i)}{b_i+h_i}\left[\frac{b_i-h_i}{2} - b_{\theta^*}\right] \\ \qquad \text{if } (\sum_{i\in\Theta^*} \rho_i\bar{f}_i - \hat{w}) - \sum_{i\in\{\Theta^*\setminus\theta^*\}} \bar{x}_i^c < -\frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2} \end{cases} \tag{4.38}$$

where α_i is the ith row of A^{-1}. The cheapest product class cycle center is

$$x_{\theta^*}^{c^*} = \hat{w} - \sum_{i\in\Theta^*} x_i^{c^*}. \tag{4.39}$$

For $i \in \{\Theta^{*1} \cup \Theta^{*2}\}$, we have

$$x_i^{c^*} = \frac{\tau\rho_i(1-\rho_i)}{2}. \tag{4.40}$$

For completeness of notation, we denote by

$$\alpha_{\theta^*}\gamma_1 = - \sum_{i\in\{\Theta^*\setminus\theta^*\}} \alpha_i\gamma_1 , \tag{4.41}$$

and

$$\alpha_{\theta^*}\gamma_2 = 1 - \sum_{i\in\{\Theta^*\setminus\theta^*\}} \alpha_i\gamma_2 . \tag{4.42}$$

Thus, we have the following algebraic simplification for all $i \in \Theta^*$

$$\alpha_i\gamma_1 = \frac{\rho_i(1-\rho_i)}{2(b_i+h_i)}\left(b_i - h_i - \frac{\sum_{j\in\Theta^*}(b_j-h_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}}{\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}\right), \tag{4.43}$$

and

$$\alpha_i\gamma_2 = \frac{\frac{\rho_i(1-\rho_i)}{b_i+h_i}}{\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}. \tag{4.44}$$

Thus, all but the cheapest earliness or tardiness product classes are in *condition 2*. We use the *condition* of the cheapest product to categorize the total feasible workload levels by dividing them into 3 regions: region I where the cheapest product is in *condition 1*, region II where the cheapest product is in *condition 2* and region III where the cheapest product is in *condition 3*.

It is important to note that when a product class i is binding yet its Lagrangian multiplier μ_i is zero, the inclusion or absence of i in Θ^* does not effect the other product classes' cycle centers. This can be seen by setting the borderline class's cycle center as derived from equation (4.38) equal to $\tau\rho_i(1-\rho_i)/2$ and subsequent algebraic manipulations. In addition, on the border between regions II and I or III the cycle center does not change if using any of the two corresponding expressions in equation (4.38). We can therefore conclude that cycle center is a continuous function of cycle length τ.

With these expressions for cycle center, we can restate the average cost per cycle in terms of cycle length τ and total workload level w. If we let w_{Θ^*} denote $\sum_{j\in\Theta^*}\rho_j\bar{f}_j - w$ then for $i \in \Theta^* \setminus \theta^*$ cost per cycle is

$$c_i(\tau,w) = \begin{cases} (b_i+h_i)\frac{\tau\rho_i(1-\rho_i)}{8} + \frac{\tau\rho_i(1-\rho_i)}{2(b_i+h_i)}\left[\frac{b_i-h_i}{2}+h_{\theta^*}\right]\left[\frac{h_i-b_i}{2}+h_{\theta^*}\right] & \text{I} \\ (b_i+h_i)\frac{\tau\rho_i(1-\rho_i)}{8} + \frac{\tau(b_i+h_i)}{2\rho_i(1-\rho_i)}(\alpha_i\gamma_1)^2 & \\ \quad +\frac{b_i+h_i}{\rho_i(1-\rho_i)}(\alpha_i\gamma_1)(\alpha_i\gamma_2)(w_{\Theta^*}+\tau\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2}) & \text{II} \\ \quad +\frac{b_i+h_i}{2\tau\rho_i(1-\rho_i)}(w_{\Theta^*}+\tau\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2})^2(\alpha_i\gamma_2)^2 & \\ \quad +\frac{h_i-b_i}{2}(\tau\alpha_i\gamma_1+(w_{\Theta^*}+\tau\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2})\alpha_i\gamma_2) & \\ (b_i+h_i)\frac{\tau\rho_i(1-\rho_i)}{8} + \frac{\tau\rho_i(1-\rho_i)}{2(b_i+h_i)}\left[\frac{b_i-h_i}{2}-b_{\theta^*}\right]\left[\frac{h_i-b_i}{2}-b_{\theta^*}\right] & \text{III} \end{cases} \quad . \qquad (4.45)$$

For the cheapest product the cost per cycle is

$$c_{\theta^*}(\tau, w) = \begin{cases} h_{\theta^*}(w_{\Theta^*} + \tau \sum_{j \notin \Theta^*} \frac{\rho_j(1-\rho_j)}{2}) & \text{I} \\ \quad -h_{\theta^*}\tau \sum_{i \in \Theta^* \setminus \theta^*} \frac{\rho_i(1-\rho_i)}{b_i+h_i}\left[\frac{b_i-h_i}{2} + h_{\theta^*}\right] & \\ (b_{\theta^*} + h_{\theta^*})\frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{8} + \frac{\tau(b_{\theta^*}+h_{\theta^*})}{2\rho_{\theta^*}(1-\rho_{\theta^*})}(\alpha_{\theta^*}\gamma_1)^2 & \\ \quad +\frac{b_{\theta^*}+h_{\theta^*}}{\rho_{\theta^*}(1-\rho_{\theta^*})}(\alpha_{\theta^*}\gamma_1)(\alpha_{\theta^*}\gamma_2)(w_{\Theta^*} + \tau \sum_{j \notin \Theta^*} \frac{\rho_j(1-\rho_j)}{2}) & \text{II} \\ \quad +\frac{b_{\theta^*}+h_{\theta^*}}{2\tau\rho_{\theta^*}(1-\rho_{\theta^*})}(w_{\Theta^*} + \tau \sum_{j \notin \Theta^*} \frac{\rho_j(1-\rho_j)}{2})^2(\alpha_{\theta^*}\gamma_2)^2 & \\ \quad +\frac{h_{\theta^*} - b_{\theta^*}}{2}(\tau\alpha_{\theta^*}\gamma_1 + (w_{\Theta^*} + \tau \sum_{j \notin \Theta^*} \frac{\rho_j(1-\rho_j)}{2})\alpha_{\theta^*}\gamma_2) & \\ -b_{\theta^*}(w_{\Theta^*} + \tau \sum_{j \notin \Theta^*} \frac{\rho_j(1-\rho_j)}{2}) & \text{III} \\ \quad +b_{\theta^*}\tau \sum_{i \in \Theta^* \setminus \theta^*} \frac{\rho_i(1-\rho_i)}{b_i+h_i}\left[\frac{b_i-h_i}{2} - b_{\theta^*}\right]) & \end{cases} \tag{4.46}$$

The maxed-out product classes can either be in *condition 1* or in *condition 2*. Thus for $i \in \{\Theta^{*1} \cup \Theta^{*2}\}$ the cost per cycle is

$$c_i(\tau, w) = \begin{cases} h_i(\rho_i \bar{f}_i - \frac{\tau\rho_i(1-\rho_i)}{2}) & \text{for } i \in \Theta^{*1} \\ \frac{b_i+h_i}{2\tau\rho_i(1-\rho_i)}(\rho_i \bar{f}_i)^2 - b_i(\rho_i \bar{f}_i - \frac{\tau\rho_i(1-\rho_i)}{2}) & \text{for } i \in \Theta^{*2} \end{cases} \tag{4.47}$$

4.2.2 Cycle Length Optimization

Given the form of the cost per cycle in terms of τ and w, we can find an expression for the optimal cycle length τ by differentiating equation 3.23 with respect to τ and solving against zero. The optimal cycle length is expressed in terms of basic system parameters, the set of binding products and the effective setup cost per cycle $S = K - sV'(W)$. Thus Θ^*, Θ^{*1} and Θ^{*2} are all functions of S and w. Using (4.45) and (4.46), if we let

$$\xi_1^{\Theta^*} = \sum_{i \in \Theta^* \setminus \theta^*} \frac{(b_i + h_i)\rho_i(1-\rho_i)}{8} - \frac{\rho_i(1-\rho_i)}{2(b_i+h_i)}\left(\frac{b_i - h_i}{2} + h_{\theta^*}\right)^2 \\ + \sum_{i \in \Theta^{*1}} \frac{\rho_i(1-\rho_i)}{2}(h_{\theta^*} - h_i) + \sum_{i \in \Theta^{*2}} \frac{\rho_i(1-\rho_i)}{2}(h_{\theta^*} + b_i), \tag{4.48}$$

$$\xi_2^{\Theta^\bullet} = \sum_{i\in\Theta^{\bullet 2}} \frac{(\rho_i \bar{f}_i)^2(b_i+h_i)}{2\rho_i(1-\rho_i)} \tag{4.49}$$

$$\begin{aligned}\xi_3^{\Theta^\bullet} = & \sum_{i\in\Theta^\bullet}\left[\frac{(b_i+h_i)\rho_i(1-\rho_i)}{8} + \frac{b_i+h_i}{2\rho_i(1-\rho_i)}(\alpha_i\gamma_1)^2 + \frac{h_i-b_i}{2}\alpha_i\gamma_1\right. \\ & \left. + \frac{b_i+h_i}{2\rho_i(1-\rho_i)}(\alpha_i\gamma_2)^2(\sum_{j\notin\Theta^\bullet}\frac{\rho_j(1-\rho_j)}{2})^2 + \frac{h_i-b_i}{2}\alpha_i\gamma_2(\sum_{j\notin\Theta^\bullet}\frac{\rho_j(1-\rho_j)}{2})\right] \\ & - \sum_{i\in\Theta^{\bullet 1}} h_i\frac{\rho_i(1-\rho_i)}{2} + \sum_{i\in\Theta^{\bullet 2}} b_i\frac{\rho_i(1-\rho_i)}{2}\end{aligned} \tag{4.50}$$

$$\xi_4^{\Theta^\bullet} = \sum_{i\in\Theta^\bullet}\frac{b_i+h_i}{2\rho_i(1-\rho_i)}(\alpha_i\gamma_2)^2, \tag{4.51}$$

$$\begin{aligned}\xi_5^{\Theta^\bullet} = & \sum_{i\in\Theta^\bullet\setminus\theta^\bullet}\frac{(b_i+h_i)\rho_i(1-\rho_i)}{8} - \frac{\rho_i(1-\rho_i)}{2(b_i+h_i)}\left(\frac{b_i-h_i}{2}-b_{\theta^\bullet}\right)^2 \\ & - \sum_{i\in\Theta^{\bullet 1}}\frac{\rho_i(1-\rho_i)}{2}(b_{\theta^\bullet}-h_i) + \sum_{i\in\Theta^{\bullet 2}}\frac{\rho_i(1-\rho_i)}{2}(b_i-b_{\theta^\bullet})\end{aligned} \tag{4.52}$$

the optimal cycle length can be stated as

$$\tau^* = \begin{cases} \sqrt{\frac{S+\xi_2^{\Theta^\bullet}}{\xi_1^{\Theta^\bullet}}} & \text{I} \\ \sqrt{\frac{S+\xi_4^{\Theta^\bullet}(\sum_{i\in\Theta^\bullet}\rho_i\bar{f}_i-w)^2+\xi_2^{\Theta^\bullet}}{\xi_3^{\Theta^\bullet}}} & \text{II} \\ \sqrt{\frac{S+\xi_2^{\Theta^\bullet}}{\xi_5^{\Theta^\bullet}}} & \text{III}\end{cases}. \tag{4.53}$$

Using (4.43) and (4.44), we additionally have the following simplifications

$$\xi_4^{\Theta^\bullet} = \frac{1}{2\sum_{j\in\Theta^\bullet}\frac{\rho_j(1-\rho_j)}{b_j+h_j}} \tag{4.54}$$

and

$$\begin{aligned}\xi_3^{\Theta^\bullet} = & \sum_{i\in\Theta^\bullet}\left[\frac{\rho_i(1-\rho_i)}{b_i+h_i}\frac{(b_i+h_i)^2-(b_i-h_i)^2}{8}\right. \\ & \left. + \frac{\rho_i(1-\rho_i)}{8(b_i+h_i)}\left(\frac{\sum_{j\in\Theta^\bullet}(b_j-h_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}}{\sum_{j\in\Theta^\bullet}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}\right)^2\right]\end{aligned} \tag{4.55}$$

$$+\frac{\left(\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2}\right)^2}{2\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}-\frac{\sum_{j\in\Theta^*}\frac{b_j-h_j}{2}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}{\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}\left(\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2}\right)$$
$$-\sum_{i\in\Theta^{*1}}h_i\frac{\rho_i(1-\rho_i)}{2}+\sum_{i\in\Theta^{*2}}b_i\frac{\rho_i(1-\rho_i)}{2}$$

which is

$$\begin{aligned}\xi_3^{\Theta^*} &= \frac{1}{8}\sum_{i\in\Theta^*}\frac{\rho_i(1-\rho_i)}{b_i+h_i}4b_ih_i+\frac{1}{8}\frac{\left[\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}(b_j-h_j)\right]^2}{\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}} \\ &\quad+\frac{\left(\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2}\right)^2}{2\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}-\frac{\sum_{j\in\Theta^*}\frac{b_j-h_j}{2}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}{\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}\left(\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2}\right) \\ &\quad-\sum_{i\in\Theta^{*1}}h_i\frac{\rho_i(1-\rho_i)}{2}+\sum_{i\in\Theta^{*2}}b_i\frac{\rho_i(1-\rho_i)}{2} \\ &= \frac{\sum_{i\in\Theta^*}\sum_{j\in\Theta^*}\frac{\rho_i(1-\rho_i)}{b_i+h_i}\frac{\rho_j(1-\rho_j)}{b_j+h_j}(b_ib_j+h_ih_j-h_ib_j-b_ih_j+2b_ih_i+2b_jh_j)}{8\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}} \\ &\quad+\frac{\left(\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2}\right)^2}{2\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}-\frac{\sum_{j\in\Theta^*}\frac{b_j-h_j}{2}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}{\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}\left(\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2}\right) \\ &\quad-\sum_{i\in\Theta^{*1}}h_i\frac{\rho_i(1-\rho_i)}{2}+\sum_{i\in\Theta^{*2}}b_i\frac{\rho_i(1-\rho_i)}{2}.\end{aligned} \tag{4.56}$$

Although this expression looks complicated, it gives insights into how the policy behaves when either all of the earliness costs or tardiness costs are equal. That is, if $b_i = b_j$ for all i and j in Θ^* then

$$\begin{aligned}\xi_3^{\Theta^*} &= \frac{\left(\sum_{j\in\Theta^*}\rho_j(1-\rho_j)\right)^2}{8\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}-\sum_{i\in\Theta^{*1}}h_i\frac{\rho_i(1-\rho_i)}{2}+\sum_{i\in\Theta^{*2}}b_i\frac{\rho_i(1-\rho_i)}{2} \\ &\quad+\frac{\left(\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2}\right)^2}{2\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}-\frac{\sum_{j\in\Theta^*}\frac{b_j-h_j}{2}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}{\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}\left(\sum_{j\notin\Theta^*}\frac{\rho_j(1-\rho_j)}{2}\right),\end{aligned} \tag{4.57}$$

which, ignoring the terms created by the binding variables, is rather simple. Interestingly, the expression is identical if $h_i = h_j$ for all i and j in Θ^* instead of the tardiness costs.

Therefore in order to use the algorithm detailed in section 4.5, we need to be able to find τ^* and x^{c^*} for a given w and for varying $V'(w)$. Thus, it is necessary to find Θ^* as a function of $S = K - sV'(w)$ and w (for clarity we will re-introduce the notation specifying the dependence of Θ^* on S for the remainder of this section). We suggest an algorithm which constructs $\Theta^*(S)$ by finding an initial set for S equal to zero and then tracks how the set evolves as S increases. Our task is complicated by two facts: the first that as S increases the optimal solution jumps in region and the second that there is no guarantee that when a product class becomes binding and leaves $\Theta^*(S)$ it does not re-enter for larger S. We simplify our calculations of $\Theta(S)$ by including the type of region the cycle center and cycle length imply in our accounting. Let $\Theta^I(S)$ denote $\Theta(S)$ when the cycle center and cycle length satisfy the region I conditions, $\Theta^{II}(S)$ for the region II conditions and $\Theta^{III}(S)$ for the third region. The algorithm is based on the following observations:

1. Cycle center x^{c^*} and cycle length τ are continuous functions of effective setup cost per cycle S. For a given w and $V'(w)$, the objective function in the CLO program is continuous with continuous derivatives with respect to cycle length τ ($\tau > 0$), cycle center x^c and effective setup cost per cycle S. In addition, the boundary conditions are also continuous with continuous derivatives with respect to x^c, τ and S. Thus, the Kuhn-Tucker necessary optimality conditions change continuously with S. The only way the optimal x^{c^*} and τ^* could be discontinuous with respect to an increase in S is if there were multiple optimal solutions for a given S and w. This is not the case since the objective function is convex with respect to cycle center and cycle length and has the unique optimal solution presented in equations 4.38 and 4.53.

2. Optimal cycle length τ is monotonically increasing with respect to S. This is easily seen from equation 4.53 and the first observation implying the continuity of τ^* during changes of binding products Θ^* and changes in region.

3. If product class i is not in Θ^*, then for $\Theta' = \Theta^* \cup \{i\}$ the cycle center $x^{c'}$ and cycle length τ' calculated with Θ' satisfy $x_i^{c'} < \tau'\rho_i(1-\rho_i)/2$. This follows from the construction of Θ^*.

4. If S' and S'' are such that $S' < S''$ and both their respective optimal cycle lengths and cycle centers satisfy region I (III) conditions then $\Theta^{*I}(S'') \subset \Theta^{*I}(S')$

$(\Theta^{*III}(S'') \subset \Theta^{*III}(S'))$. We show this by inducting on the cheapest product classes. Consider the region I case: let θ_h^1 be the first cheapest product to become binding in region I at effective setup cost S^1 with cycle length τ^1. All other classes with smaller earliness costs must be binding in *condition 2*. The instant product θ_h^1 becomes binding the cycle center has reached the positivity boundary and so

$$w - \sum_{i\in\Theta^{*I}(S^1)\backslash\theta_h^1}\left[\rho_i\bar{f}_i - \frac{\tau^1\rho_i(1-\rho_i)}{2}\frac{b_i-h_i+2h_{\theta_h^1}}{b_i+h_i}\right] - \sum_{i\in\Theta^{*2}(S^1)}\frac{\tau^1\rho_i(1-\rho_i)}{2} - \frac{\tau^1\rho_{\theta_h^1}(1-\rho_{\theta_h^1})}{2} = 0 \; . \tag{4.58}$$

In order for the minimum work experienced over the cycle, $x_i^c - \tau\rho_i(1-\rho_i)/2$, to reach the boundary as τ increased, its derivative must be negative, that is

$$-\sum_{i\in\Theta^{*2}(S^1)}\frac{\rho_i(1-\rho_i)}{2} + \sum_{i\in\Theta^{*I}(S^1)\backslash\theta_h^1}\frac{\rho_i(1-\rho_i)}{2}\frac{b_i-h_i+2h_{\theta_h^1}}{b_i+h_i} - \frac{\rho_{\theta_h^1}(1-\rho_{\theta_h^1})}{2} < 0. \tag{4.59}$$

For larger S, the cheaper binding *condition 2* product classes will not re-enter in region 1 since they can only re-enter in *condition 2* as τ will be larger. Additional product classes might leave $\Theta^{*I}(S^1)$ but since $\frac{b_i-h_i+2h_{\theta_h^1}}{b_i+h_i} < 1$ equation 4.59 will become more negative. This implies that equation 4.58 will remain negative for larger τ, and so θ_h^1 cannot re-enter Θ^{*I}. The ith induction on the cheapest product class results in an argument identical to the previous one where equations 4.58 and 4.59 are modified by replacing θ_h^1 with θ_h^i and S^1 with S^i.

The other more expensive classes which become binding in region I also cannot re-enter Θ^{*I}. Since $h_{\theta_h^*(S'')} > h_{\theta_h^*(S')}$, from the above induction we have

$$\frac{\rho_i(1-\rho_i)}{2}\frac{b_i-h_i+2+h_{\theta_h^*(S')}}{b_i+h_i} < \frac{\rho_i(1-\rho_i)}{2}\frac{b_i-h_i+2+h_{\theta_h^*(S'')}}{b_i+h_i}. \tag{4.60}$$

Therefore, the more expensive non-binding product classes' cycle centers are lower for larger S. Since cycle length τ increases with S, if the cycle center becomes less than $\tau\rho_i(1-\rho_i)/2$, and hence binding, it will remain binding.

The same argument holds for region III.

5. *If S is such that optimal cycle length and cycle centers imply a shift from*

region II to region I (III), then $\Theta^{*I}(S^+) = \Theta^{*II}(S^-)$ $(\Theta^{*III}(S^+) = \Theta^{*II}(S^-))$. Since cycle centers and cycle length are continuous in S, this is equivalent to the statement that no binding products in region II become non-binding after a transition to region I. Binding *condition 1* product classes must have become binding in region I and remain binding by the 4th observation (In region III, *condition 1* product becoming non-binding would violate the region III definition). *Condition 2* variables in region I will have a cycle center of the form $\rho_i \bar{f}_i - \tau\alpha_i \cdot \gamma_1 - (\sum_{j\in\Theta^*} \rho_j \bar{f}_j - w')\alpha_i \cdot \gamma_2$ and so the derivative of $x_i^c - \tau\rho_i(1-\rho_i)/2$ with respect to τ is always negative. This implies that a *condition 2* binding cycle center will continue to push against the orthant boundary.

6. *If S' and S'' are such that $S' < S''$ and both their respective optimal cycle lengths and cycle centers satisfy region II conditions then* $\Theta^{*II}(S'') \subset \Theta^{*II}(S')$. We shall prove this by examining $x_i^c - \tau\rho_i(1-\rho_i)/2$, the derivative of the minimum amount of work over the course of a cycle, with respect to τ. Once the derivative is negative, we shall show that it remains so as further products are removed from Θ^{*II}. Thus when a product class becomes binding at $x_i^c - \tau\rho_i(1-\rho_i)/2 = 0$, as τ expands, it cannot re-enter: the negative derivative would continue to push the cycle center against the orthant boundary. Since no binding products in $\Theta^{*II}(S)$ can become non-binding in the other regions by observation 5, $\Theta^{*II}(S)$ is a non-increasing set with S.

The derivative of minimum workload in region II for a non-binding product class is

$$\frac{\partial}{\partial\tau}(x_i^c - \frac{\tau\rho_i(1-\rho_i)}{2}) = -\alpha_i\gamma_1 - \frac{\rho_i(1-\rho_i)}{2} - \frac{\frac{\rho_i(1-\rho_i)}{b_i+h_i}}{\Sigma} \sum_{j\notin\Theta^{*II}(S)} \frac{\rho_j(1-\rho_j)}{2}$$

$$= -\frac{\frac{\rho_i(1-\rho_i)}{2(b_i+h_i)}}{\Sigma}\left[\sum_{j\in\Theta^{*II}(S)} (2b_i + h_j - b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j} + \sum_{j\notin\Theta^{*II}(S)} \rho_j(1-\rho_j)\right] \quad (4.61)$$

where $\Sigma = \sum_{j\in\Theta^{*II}(S)} \rho_j(1-\rho_j)/(b_j+h_j)$. Therefore the derivative is negative if and only if

$$\sum_{j\in\Theta^{*II}(S)} (2b_i + h_j - b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j} + \sum_{j\notin\Theta^{*II}(S)} \rho_j(1-\rho_j) > 0. \quad (4.62)$$

Thus, at effective setup cost S', if the derivative of product class i is negative and a more expensive tardiness product class l leaves to form $\Theta^{*II}(S')$, then equation 4.62 implies

$$\sum_{j\in\Theta^{*II}(S)\cup\{l\}}(2b_i+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}+\sum_{j\notin\Theta^{*II}(S)\cup\{l\}}\rho_j(1-\rho_j)>0, \quad (4.63)$$

which is

$$\sum_{j\in\Theta^{*II}(S)}(2b_i+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}+\sum_{j\notin\Theta^{*II}(S)}\rho_j(1-\rho_j)+2\rho_l(1-\rho_l)\frac{b_i-b_l}{b_l+h_l}>0. \quad (4.64)$$

Since $b_i - b_l < 0$, we have

$$\sum_{j\in\Theta^{*II}(S)}(2b_i+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}+\sum_{j\notin\Theta^{*II}(S)}\rho_j(1-\rho_j)>0, \quad (4.65)$$

and so the derivative remains negative. If a cheaper tardiness class l leaves forming $\Theta^{*II}(S)$, the derivative of the cheaper class implies

$$\sum_{j\in\Theta^{*II}(S)\cup\{l\}}(2b_l+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}+\sum_{j\notin\Theta^{*II}(S)\cup\{l\}}\rho_j(1-\rho_j)>0. \quad (4.66)$$

The left hand side is less than

$$\sum_{j\in\Theta^{*II}(S)}(2b_i+h_j-b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}+\sum_{j\notin\Theta^{*II}(S)}\rho_j(1-\rho_j) \quad (4.67)$$

since their difference is

$$\sum_{j\in\Theta^{*II}(S)}2(b_i-b_l)\frac{\rho_j(1-\rho_j)}{b_j+h_j}. \quad (4.68)$$

Therefore the derivative of the minimum workload over a cycle for product class i remains negative after l becomes binding.

7. When a condition 1 binding product class i changes to condition 2, it remains binding. If this change occurs in regions I or III, the class i product remains maxed-out because in these regions the $\frac{\partial}{\partial\tau}(x_i^c - \frac{\tau\rho_i(1-\rho_i)}{2})$ is always negative for a *condition 2* product class. If the change occurs in region II, a *condition 1* variable in region II implies that

$$\sum_{j\in\Theta^{*II}(S)\setminus\{i\}} \frac{\rho_j(1-\rho_j)}{2(b_j+h_j)}[b_j - h_j + 2h_i] - \frac{\rho_i(1-\rho_i)}{2} < 0 \tag{4.69}$$

or

$$\rho_i(1-\rho_i) + \sum_{j\in\Theta^{*II}(S)\setminus\{i\}} \frac{\rho_j(1-\rho_j)}{b_j+h_j}[h_j - b_j - 2h_i] < 0. \tag{4.70}$$

Subtracting equation 4.62 from equation 4.69 we get

$$\sum_{j\in\Theta^{*II}(S)\setminus\{i\}} 2(b_i+h_i)\frac{\rho_j(1-\rho_j)}{b_j+h_j} \tag{4.71}$$

which is positive. As stated in observation 6, this implies that $\frac{\partial}{\partial\tau}(x_i^c - \frac{\tau\rho_i(1-\rho_i)}{2})$ is negative.

8. If S such that optimal cycle length and cycle centers imply a shift from region I or III to region II, $\Theta^{*II}(S^+)$ can be calculated by an iterative algorithm. A conclusion from the previous seven observations is that $\Theta^*(S)$ is relatively predictable with the notable exception of transitions from regions I and III to region II. At these transitions, binding variables may again enter $\Theta^*(S)$. Re-calculation of $\Theta^{*II}(S)$, however, is not difficult. At the effective setup cost S point of transition from region I or III to region II, all binding product classes not before in $\Theta^{*II}(S')$ for $S' < S$ should be re-included in $\Theta^{*II}(S)$. The cycle center can then be re-calculated. Those product classes such that either their cycle centers are infeasible or were previously binding and currently have a negative cycle center derivative (as determined by equation 4.62) can be removed from $\Theta^{*II}(S)$. By observation 6, they do not re-enter. This process can be iteratively done until $\Theta^{*II}(S)$ is found such that all binding variables with negative minimum workload per cycle derivatives are removed. The resulting $\Theta^{*II}(S)$ is equal to $\Theta^{*II}(S^+)$.

With these eight observations, the algorithm we suggest is simple. From an initial $\Theta^{*I}(0)$, $\Theta^{*II}(0)$ and $\Theta^{*III}(0)$, we track how each evolves as S is increased. Three types of events can change $\Theta^*(S)$: a shift in region, a non-binding customized class can become binding and a *condition 1* binding class can become *condition 2*. Given equations 4.38 and 4.53, we can exactly calculate the range of S before any one of these events occur.

The initial set of non-binding product classes is chosen as follows. For effective setup cost of zero, cycle length τ becomes zero as region II vanishes. A low total workload level w may imply that some product classes are binding, that is there is few orders remaining in the system and it may be advantageous to max-out some of the cheaper customized product classes. This occurs if w is less than $\sum_{i=1}^{N} \rho_i \bar{f}_i$. Let $\{e_1, e_2, \ldots, e_N\}$ be an earliness ordering of the product classes such that $h_{e_1} \leq h_{e_2} \leq \cdots \leq h_{e_N}$. We find $\Theta^{*I}(0)$ by one-by-one removing classes. A class e_j is removed if it is next in the e_i list, is customizable and is insufficient for storing the remaining work, i.e. $\sum_{i=1}^{j} \rho_i \bar{f}_i < \sum_{i=1}^{N} \rho_i \bar{f}_i - w$. The process stops when either a standardized product is reached in the e_i list or the next product e_j implies $\sum_{i=1}^{j} \rho_i \bar{f}_i > \sum_{i=1}^{N} \rho_i \bar{f}_i - w$. If w is greater than $\sum_{i=1}^{N} \rho_i \bar{f}_i$, then $\Theta^{*I}(0)$ is set to $\{1, \ldots, N\}$. Both $\Theta^{*II}(0)$ and $\Theta^{*III}(0)$ are always set to $\{1, \ldots, N\}$.

Given an initial $\Theta^*(0)$ we can find the range of effective setup cost before the set changes. The range is the minimum S such that one of the three events happen: 1) a shift in region, 2) the binding of a customized product class or 3) the change in condition of a binding class. For a given S' with non-binding set $\Theta^*(S')$ a region change occurs when S is

$$S = \begin{cases} \xi_1^{\Theta^{*I}(S')} \left(\frac{\sum_{i \in \Theta^{*I}(S')} \rho_i \bar{f}_i - w}{\Xi_1} \right)^2 - \xi_2^{\Theta^{*I}(S')} \\ \qquad \text{region I to region II shift} \\ \xi_3^{\Theta^{*I}(S')} \left(\frac{\alpha_{\theta_h^*(S')} \gamma_2 (\sum_{i \in \Theta^{*II}(S')} \rho_i \bar{f}_i - w)}{-\alpha_{\theta_h^*(S')} \gamma_1 - \alpha_{\theta_h^*(S')} \gamma_2 (\sum_{i \notin \Theta^{*II}(S')} \frac{\rho_i(1-\rho_i)}{2}) + \frac{\rho_{\theta_h^*(S')}(1-\rho_{\theta_h^*(S')})}{2}} \right)^2 \\ \quad -\xi_4^{\Theta^{*II}(S')} (\sum_{i \in \Theta^{*II}(S')} \rho_i \bar{f}_i - w)^2 - \xi_2^{\Theta^{*II}(S')} \\ \qquad \text{region II to region I shift} \\ \xi_3^{\Theta^{*I}(S')} \left(\frac{\alpha_{\theta_h^*(S')} \gamma_2 (\sum_{i \in \Theta^{*II}(S')} \rho_i \bar{f}_i - w)}{-\alpha_{\theta_b^*(S')} \gamma_1 - \alpha_{\theta_b^*(S')} \gamma_2 (\sum_{i \notin \Theta^{*II}(S')} \frac{\rho_i(1-\rho_i)}{2}) - \frac{\rho_{\theta_b^*(S')}(1-\rho_{\theta_b^*(S')})}{2}} \right)^2 \\ \quad -\xi_4^{\Theta^{*II}(S')} (\sum_{i \in \Theta^{*II}(S')} \rho_i \bar{f}_i - w)^2 - \xi_2^{\Theta^{*II}(S')} \\ \qquad \text{region II to region III shift} \\ \xi_5^{\Theta^{*III}(S')} \left(\frac{\sum_{i \in \Theta^{*III}(S')} \rho_i \bar{f}_i - w}{\Xi_2} \right)^2 - \xi_2^{\Theta^{*III}(S')} \\ \qquad \text{region III to region II shift} \end{cases} . \tag{4.72}$$

where

$$\Xi_1 = \sum_{i \in \{\Theta^{*I}(S') \setminus \theta_h^*(S')\}} \frac{\rho_i(1-\rho_i)}{b_i + h_i} \left[\frac{b_i - h_i}{2} + h_{\theta^*(S')} \right] + \frac{\rho_{\theta_h^*(S')}(1 - \rho_{\theta_h^*(S')})}{2} - \sum_{i \notin \Theta^{*I}(S')} \frac{\rho_i(1-\rho_i)}{2} \tag{4.73}$$

$$\Xi_2 = \sum_{i \in \{\Theta^{*III}(S') \setminus \theta_b^*(S')\}} \frac{\rho_i(1-\rho_i)}{b_i + h_i} \left[\frac{b_i - h_i}{2} - b_{\theta_b^*(S')} \right] - \frac{\rho_{\theta_b^*(S')}(1 - \rho_{\theta_b^*(S')})}{2} - \sum_{i \notin \Theta^{*III}(S')} \frac{\rho_i(1-\rho_i)}{2}. \tag{4.74}$$

A customized product class becomes binding when S is

$$S = \begin{cases} \xi_1^{\Theta^{\bullet I}(S')} \left(\frac{\rho_i \bar{f}_i (b_i + h_i)}{\rho_i (1-\rho_i)(b_i + h_{\theta_h^*})} \right)^2 - \xi_2^{\Theta^{\bullet I}(S')} & \text{region I, } i \neq \theta_h^* \text{ binding} \\ \xi_1^{\Theta^{\bullet I}(S')} \left(\frac{\sum_{i \in \{\Theta^{\bullet I}(S') \setminus \theta_h^*(S')\}} \rho_i \bar{f}_i - w}{\Xi_1} \right)^2 - \xi_2^{\Theta^{\bullet I}(S')} & \text{region I, } \theta_h^* \text{ binding} \\ \xi_3^{\Theta^{\bullet II}(S')} \left(\frac{\rho_i \bar{f}_i - (\sum_{j \in \{\Theta^{\bullet II}(S')\}} \rho_j \bar{f}_j - w)\alpha_i \gamma_2}{\alpha_i \gamma_1 + \frac{\rho_i(1-\rho_i)}{2} + \sum_{j \notin \Theta^{\bullet II}(S')} \frac{\rho_j(1-\rho_j)}{2}} \right)^2 - \xi_2^{\Theta^{\bullet II}(S')} & \\ \quad - \xi_4^{\Theta^{\bullet II}(S')} (\sum_{j \in \Theta^{\bullet II}(S')} \rho_j \bar{f}_j - w)^2 & \text{region II, } i \text{ binding} \\ \xi_5^{\Theta^{\bullet III}(S')} \left(\frac{\rho_i \bar{f}_i (b_i + h_i)}{\rho_i (1-\rho_i)(b_i - b_{\theta_b^*})} \right)^2 - \xi_2^{\Theta^{\bullet III}(S')} & \text{region III, } i \neq \theta_b^* \text{ binding} \\ \xi_5^{\Theta^{\bullet III}(S')} \left(\frac{\sum_{i \in \{\Theta^{\bullet III}(S') \setminus \theta_b^*(S')\}} \rho_i \bar{f}_i - w}{\Xi_2} \right)^2 - \xi_2^{\Theta^{\bullet III}(S')} & \text{region III, } \theta_b^* \text{ binding} \end{cases} \quad (4.75)$$

Lastly, *condition 1* variables change to *condition 2* when S is

$$S = \begin{cases} \xi_1^{\Theta^{\bullet I}(S')} \frac{\bar{f}_i}{1-\rho_i} - \xi_2^{\Theta^{\bullet III}(S')} & \text{region I} \\ \xi_3^{\Theta^{\bullet II}(S')} \frac{\bar{f}_i}{1-\rho_i} - \xi_4^{\Theta^{\bullet II}(S')} (\sum_{j \in \Theta^{\bullet II}(S')} \rho_j \bar{f}_j - w)^2 - \xi_2^{\Theta^{\bullet II}(S')} & \text{region II} \\ \xi_5^{\Theta^{\bullet III}(S')} \frac{\bar{f}_i}{1-\rho_i} - \xi_2^{\Theta^{\bullet III}(S')} & \text{region III} \end{cases} \quad (4.76)$$

Thus the current set of $\Theta^*(S')$, $\Theta^{*1}(S')$ and $\Theta^{*2}(S')$ is valid for effective setup cost from S' to the minimum S greater than S' in the above equations (4.72, 4.75, 4.76). At that point, $\Theta^{*I}(S)$, $\Theta^{*II}(S)$ and $\Theta^{*III}(S)$ are updated according to observations 5 through 8 and the process is re-iterated. The algorithm ends when either all customized products are binding or all but one of the customized products are binding and the cheapest is the *condition 3* product in region 3 with $\frac{\partial}{\partial\tau}(x_i^c - \frac{\tau\rho_i(1-\rho_i)}{2}) < 0$ (in the other regions, growing cycle length would imply that either the customized product will become binding or that the region would eventually shift).

Therefore for each total workload this algorithm generates the set of binding customized product classes as a function of effective cost. With this cycle length and cycle center, cycle cost can be calculated. This can then be fed into the algorithm detailed in section Chapter 3 so that $V'(w)$ can be computed and with it the proposed policy.

It is also possible to derive several structural properties about the potential function $V'(w)$ and so gain some qualitative insight about the proposed policy. We assume that as $w \to \infty$ the set of binding variables Θ^* stabilizes, that is there exists a w' such that for all $w_1, w_2 > w'$ then $\Theta^*(w_1) = \Theta^*(w_2)$. We feel that this is a natural assumption since there are only a finite number of possibilities for Θ^* and behavior evident in observations 1-8 which demonstrate how the product classes smoothly become binding with respect to S and w.

Property 1. *If average setup time s is greater than zero and as $w \to \infty$ region II conditions hold, then*

$$V'(w) = \frac{2\sqrt{\xi_3^{\Theta^*}\xi_4^{\Theta^*}} + \xi_6^{\Theta^*}}{c} w + o(w) \tag{4.77}$$

where Θ^* is the set containing the standardized goods and the cheapest backorder good (this product could either be the cheapest customized or a standardized product) and

$$\xi_6^{\Theta^*} = \frac{\sum_{j\notin\Theta^*}\rho_j(1-\rho_j) + \sum_{j\in\Theta^*}(h_j - b_j)\frac{\rho_j(1-\rho_j)}{b_j+h_j}}{2\sum_{j\in\Theta^*}\frac{\rho_j(1-\rho_j)}{b_j+h_j}}. \tag{4.78}$$

If region III conditions hold then

$$V'(w) = -\frac{b_{\theta^*}}{c} w + o(w) \tag{4.79}$$

where Θ^* is again the set containing the standardized goods and the cheapest backorder *good.*

The derivation of this property, which assumes an asymptotic monotonicity property, is nearly identical to that in the Appendix of Reiman and Wein [39]. Please see Appendix A for details of the proof.

Property 2. *If $K = 0$ and all of the products are standardized, the policy at the idling threshold w_0 satisfies the region II conditions if and only if $h_i = h_j$ for all i and j. If this condition holds then*

$$\tau^*(w_0) = \sqrt{\xi_4/\xi_3}\Big(\sum_{i\in\Theta^*} \rho_i \bar{f}_i - w_0\Big). \tag{4.80}$$

If $h_i \neq h_j$ for some i and j then the idling threshold satisfies the region I conditions and $\tau^*(w_0) = 0$.

Proof: The setup cost K equal to zero implies that the penalties associated with short cycle length are only due to the potential function $V'(w)$ and setup time s. It is not difficult to see that the idling threshold w_0 is less than or equal to $\sum_{i=1}^N \rho_i \bar{f}_i$: for any fixed cycle length τ and its associated optimal cycle center, from equations (4.45) and (4.46) the average cost per cycle is convex in total workload level with minimum at $\sum_{i=1}^N \rho_i \bar{f}_i$. Thus, the optimal idling threshold must be less than $\sum_{i=1}^N \rho_i \bar{f}_i$ since any policy with idling threshold $w_0' > \sum_{i=1}^N \rho_i \bar{f}_i$ would be improved by setting $\tau(w)$ to some trivial length for workloads levels w below w_0' up to $\sum_{i=1}^N \rho_i \bar{f}_i$. Thus, for $K = 0$ the idling threshold must be in regions I or II.

The boundary condition $p(w_0) = 0$, $K = 0$ and equation (4.53) imply that $\tau^*(w_0) = 0$ if region I conditions are satisfied and $\tau^*(w_0) = \sqrt{\xi_4/\xi_3}(\sum_{i\in\Theta^*} \rho_i \bar{f}_i - w_0)$ if region II conditions hold. As there are only standardized product classes, Θ^* is trivially $\{1, \ldots, N\}$. Thus, in order for the region II conditions to apply at total workload level w_0, $\tau^*(w_0)$ must satisfy the cycle placement conditions originally specified in equation (4.2). We can rewrite the centering condition as

$$\Big|\Big(\sum_{i\in\Theta^*} \rho_i \bar{f}_i - w\Big)\alpha_{\theta^*}\gamma_2 + \tau\alpha_{\theta^*}\gamma_1\Big| \leq \frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2} \tag{4.81}$$

Rearranging terms, this inequality is

$$\alpha_{\theta^*}\gamma_2 \leq \sqrt{\frac{\xi_4}{\xi_3}}(\frac{\rho_{\theta^*}(1-\rho_{\theta^*})}{2} - \alpha_{\theta^*}\gamma_1). \tag{4.82}$$

From the simplifications performed in (4.54) and (4.56), this inequality is false if inventory holding costs are not identical and holds at equality only if $h_i = h_j$ for all i and j. In the cost-symmetric case, the placement conditions in equation (4.81) are satisfied at equality by the idling threshold w_0, and hence w_0 is the boundary between regions I and II. ∎

Property 3. *If average setup time s is greater than zero, region II conditions hold in the limit as $w \to \infty$ if and only if the tardiness costs of all the standardized goods are equal and the cheapest tardiness cost customized good is equal to or greater than the standardized cost (or all of the customized product classes tardiness costs are equal given that there are no standardized products).*

Proof: According to Property 1, in the limit $V'(w)$ grows linearly with respect to w. Thus, for large w, τ is also increasing with respect to w and grows without bound. Therefore, if region II conditions hold, by Property 1 $\tau^*(w) \to \sqrt{\xi_4^{\Theta^*}/\xi_3^{\Theta^*}}w$ and by observation 6 Θ^* has a limit as w grows. Thus, to remain in region II, the cycle placement condition

$$|(\sum_{i\in\Theta^*}\rho_i\bar{f}_i - w)\alpha_{\theta^*}\gamma_2 + \tau\alpha_{\theta^*}\gamma_1| \leq \frac{\tau\rho_{\theta^*}(1-\rho_{\theta^*})}{2} \tag{4.83}$$

Rearranging terms, this inequality is

$$\alpha_{\theta^*}\gamma_2 \leq \sqrt{\frac{\xi_4}{\xi_3}}(\frac{\rho_{\theta^*}(1-\rho_{\theta^*})}{2} + \alpha_{\theta^*}\gamma_1). \tag{4.84}$$

Again, using equations (4.54) and (4.56), this inequality is false if the tardiness costs b_i for i in Θ^* are not equal and holds at equality only if they are identical. Thus, if the region II conditions hold in the limit, all of standardized goods goods' tardiness costs must be the same. The customized goods can have higher tardiness costs as this would imply that they are binding in the limit(that is, a non-binding, higher tardiness cost product would imply region III for large w - a class which will eventually become binding as $\tau(w)$ grows). ∎

These structural properties of the potential function $V'(w)$ give some insight on the behavior of the policy. Unfortunately, no closed form solution to $V'(w)$ is possible even if Θ^* is known for all total workload level. As in Reiman and Wein [39], we can pursue both an approximate analytical solution and a numerical solution. The approximate analytical solution aims to compute a function $V'(w)$ that nearly satisfies the optimality conditions (3.21) and (3.22). The basic idea behind the method is to use Taylor series expansions to linearize the square root terms created by the optimal form of the cycle length τ, and then to use Properties 1 through 3 to paste together the solutions of the resulting ODE's so as to ensure that the principle of smooth fit holds. However, our initial results in evaluating the scheduling policy arising from this analysis did not perform consistently well and received no further attention. What work has been done for the no due-date, standardized only case is included in Appendix B.

4.3 Proposed Policy

As stated in Chapter 3.6, the policy derived in the fluid and diffusion limits needs to be unscaled for use. The presence of deterministic due-dates, however, provides an additional method of unscaling. The policy suggested in Chapter 3.6 implements a switching rule for the machine based on $\tilde{W}_i$, the current workload present in individual orders. The equivalence between workload level and the time to the due-date of the oldest order in queue developed in Chapter 4.1 allows us to create a switching rule based on the easiest due-date in a product class queue. In some industries this approach might be easier to implement as the earliest due-date data might be more readily available.

The original proposed policy used the number of product i orders currently in the system, $\tilde{I}_i$, as a method for estimating the product i workload level, $\tilde{W}_i$, present by

$\mu_i^{-1}\tilde{I}_i \approx \tilde{W}_i$. The policy was then stated as switch when $\tilde{I}_i$ satisfies

$$\tilde{I}_i \leq \sqrt{n}\mu_i \left[x_i^{c^*}\left(\frac{\sum_{j=1}^N \mu_j^{-1}\tilde{I}_j}{\sqrt{n}}\right) - \frac{\tau\left(\frac{\sum_{j=1}^N \mu_j^{-1}\tilde{I}_j}{\sqrt{n}}\right)\rho_i(1-\rho_i)}{2} \right]. \tag{4.85}$$

By the discussion in Chapter 4.1, the HTAP states that the workload level $\tilde{W}_i$ could equally be approximated by

$$\tilde{W}_i \approx \rho_i\,(f_i - L_i) \tag{4.86}$$

where L_i is the time to the earleist due-date order in the ith product class order queue. The policy can then be stated as switch when

$$\rho_i\,(f_i - L_i) \leq \sqrt{n} \left[x_i^{c^*}\left(\frac{\sum_{j=1}^N \rho_j(f_i - L_j)}{\sqrt{n}}\right) - \frac{\tau\left(\frac{\sum_{j=1}^N \rho_j(f_i - L_j)}{\sqrt{n}}\right)\rho_i(1-\rho_i)}{2} \right]. \tag{4.87}$$

and idle when

$$\sum_{i=1}^N \rho_i\,(f_i - L_i) = w_0\sqrt{n}. \tag{4.88}$$

where n is the scale factor. Under the HTAP, either method is allowable. We further study the effectiveness of each in the second part of the computational section where we analyze a two product mixed system with due-date lead times.

4.4 Computational Study

In this section we evaluate the effectiveness of our proposed policies by conducting a series of experiments. We study a standardized only system with no due-dates and either a setup cost or setup time penalty, a mixed system with due-dates and setup times and lastly a standardized only system with due-dates. We discuss the standardized system first.

4.4.1 Standardized Products, setups and no due-dates

We look at experiments involving two-product and five-product systems which have either only setup costs or only setup time penalties. For the two-product cases, we compare the performance of our proposed policy and two "straw" policies against a numerically derived optimal policy.

For the two product cases, a dynamic programming value iteration algorithm is used to find the optimal policy and evaluate the performance of all four policies. We assume that the service times, demand inter-arrival times and setup times are exponentially distributed and that the service is pre-emptive. Thus the system can be expressed as a Markov chain where the state is given by the setup of the machine and the number of orders for each product class. With some new notation, we can state the Bellman optimality conditions for the Markov process, which form the basis for deriving the optimal policy. Let $\tilde{I}$ be the 2-vector $(\tilde{I}_1, \tilde{I}_2)$ of number of orders in queue, i the class currently set up, i^c the other product class, $\mu = \max(\mu_1, \mu_2)$, $\Lambda = \lambda_1 + \lambda_2 + \mu$, $e_1 = (1,0)$ and $e_2 = (0,1)$. For the setup cost problems, the optimal value function $V(\tilde{I}, i)$ must satisfy:

$$\begin{aligned} V(\tilde{I},i) &= \frac{1}{\Lambda}\left[\sum_{k=1}^{2}(h_k\tilde{I}_k^{-} + b_k\tilde{I}_k^{+}) + \sum_{k=1}^{2}\lambda_k V(\tilde{I}+e_k,i) \right. \\ &\left. + \min\left\{\mu_i V(\tilde{I}-e_i,i), \mu_i V(\tilde{I},i), \frac{K}{2} + \mu_{i^c} V(\tilde{I}-e_{i^c}, i^c)\right\}\right] . \end{aligned} \tag{4.89}$$

The three terms inside the minimization correspond to the three possible choices the server must make at any given time: produce the current product class, idle or switch. For the setup time problems, let μ_s be the $\max(\frac{2}{s}, \mu_1, \mu_2)$, and $\Lambda = \lambda_1 + \lambda_2 + \mu_s$. The potential function must then satisfy

$$\begin{aligned} V(\tilde{I},i) &= \frac{1}{\Lambda}\left[\sum_{k=1}^{2}(h_k\tilde{I}_k^{-} + b_k\tilde{I}_k^{+}) + \sum_{k=1}^{2}\lambda_k V(\tilde{I}+e_k,i) \right. \\ &+ \min\left\{\mu_i V(\tilde{I}-e_i,i) + (\mu_s - \mu_i)V(\tilde{I},i), \right. \\ &\left.\left. \mu_i V(\tilde{I},i), \frac{2}{s}V(\tilde{I},i^c) + (\mu_s - \frac{2}{s})V(\tilde{I},i)\right\}\right] . \end{aligned} \tag{4.90}$$

The three terms inside the minimization correspond to the three possible choices the server must make at any given time: produce the current product class, idle or begin the setup of the other class.

From this we use a value iteration algorithm to compute the *suboptimality* for the proposed and two straw policies by the formula

$$\text{policy's suboptimality} = \frac{\text{policy's cost} - \text{optimal cost}}{\text{optimal cost}} \times 100\% .$$

In implementing the value iteration algorithm, the inventory state space was truncated to $[-150, 150]$ by $[-150, 150]$ for the setup cost cases and $[-250, 250]$ by $[-250, 250]$ for the setup time cases. To achieve three-digit accuracy of the suboptimalities, 7,000 iterations of the algorithm were required for the setup cost problem and 14,000 for the setup time problem.

For the five product cases, a dynamic programming algorithm is not feasible due to the large number of inventory states, and thus no optimal policy is derived. Instead, discrete event simulation is used to evaluate the proposed policy and the two straw policies. To evaluate a policy for a particular scenario, we perform 5 independent runs of 6,000,000 time units for the setup cost problems and 10 independent runs of 6,000,000 time units for the setup time problems; each run starts with an empty system and statistics from the first 10,000 time units are discarded. We assume that the demand interarrival times, service times and setup times are exponentially distributed; unlike the two product cases, service time is non-preemptive.

For systems with two products, we consider 20 setup cost cases and 14 setup time cases; all but two cases for each type of problem assume that the products have identical parameters. Although nearly all of our cases are symmetric, the numerical results in Reiman and Wein [39] suggest that the heavy traffic analysis is equally accurate for symmetric and asymmetric problems. For systems with five products, we consider six setup cost cases and four setup time cases. We focus on the two-product setting for several reasons. The optimal solution can be numerically computed in this setting, which allows us to assess the suboptimality of our proposed policies; since the optimal policy is a dynamic cyclic policy in the two-product case (i.e., the optimal policy chooses one of the three scheduling options that we allow at each point in time), we conjecture that our proposed policies are optimal in the heavy traffic limit. Also, the graphical depictions of the various policies in two dimensions (see Figures 4-2 through 4-5) help us to understand the subtleties of the behavior of this system. The two straw policies are described in the following subsection, and the numerical results for the setup cost and setup time problems are given next. Our key observations for the SELSP are summarized last and then the due-date experiments are described.

Straw Policies

To help assess the effectiveness of the proposed policy, we consider two simpler classes of cyclic policies, and use heavy traffic analysis to optimize within these classes. One is a generalized base stock policy and the other is a fixed size corridor policy similar to one considered by Sharifnia, Caramanis and Gershwin [41]. Neither straw policy employs the $s/\sqrt{n}$ refinement that was introduced in Chapter 3.5; we discuss this issue later.

Generalized Base Stock Policy.

The generalized base stock policy can be stated as follows. If the server is set up for product i, then serve this product if the number of products in finished goods inventory is less than the base-stock level $\tilde{v}_i$, or equivalently if the work in unfilled orders $\tilde{W}_i(t)$ is greater than $\tilde{v}_i$. If $\tilde{W}_i(t) \leq \tilde{v}_i$, then idle if product j, the next product to be produced in the cycle, has a workload level $\tilde{W}_j(t) \leq \tilde{v}_j + \tilde{y}_j$; otherwise, switch to product j at this point. Hofri and Ross [20] prove that the make-to-order version of this policy is optimal in a two-product symmetric polling system. The generalized base stock policy can be thought of as a refined version of the cyclic base stock policy considered by Federgruen and Katalan [14], in the sense that their policy can only insert idleness in a state-independent manner. Although the generalized base stock policy contains $2N$ parameters, the heavy traffic behavior of this policy (see Reiman and Wein [39] for details) depends on the $\tilde{y}_i$'s only via $\max_{1\leq i\leq N} \tilde{y}_i$; let us denote this quantity by $\tilde{y}$. Hence, we set each $\tilde{y}_i$ equal to $\tilde{y}$, and optimize over the $N+1$ normalized parameters $(v_1, \ldots, v_N, y)$, where $y = \tilde{y}/\sqrt{n}$ and $v_i = \tilde{v}_i/\sqrt{n}$. In heavy traffic, this policy is equivalent to one that completes production of product i when its inventory level reaches v_i, and employs the workload idling threshold $\sum_{i=1}^{N} v_i + y$; see Figure 4-1.

To calculate the cost associated with this policy, we move from these natural parameters to those used in Chapter 4.1.1. Under the HTAP results, for a given total workload w, a generalized base stock policy has cycle center

$$x_i^c(w) = v_i + \frac{\rho_i(1-\rho_i)}{\sum_{l=1}^{N} \rho_l(1-\rho_l)}(\sum_{j=1}^{N} v_j - w) \tag{4.91}$$

and cycle length

$$\tau(w) = 2\frac{w - \sum_{i=1}^{N} v_i}{\sum_{i=1}^{N} \rho_i(1-\rho_i)}. \tag{4.92}$$

Product i's average inventory cost is obtained by substituting these parameters into equation (3.17); i.e., $c_i(v_i + \frac{\rho_i(1-\rho_i)(\sum_{j=1}^{N} v_j - w)}{\sum_{j=1}^{N} \rho_j(1-\rho_j)}, 2\frac{w-\sum_{i=1}^{N} v_i}{\sum_{i=1}^{N} \rho_i(1-\rho_i)}, w)$.

Using the HTAP results, we can derive the total average cost for the generalized base stock policy for both the setup cost and time problems. In the setup cost case, total average cost is calculated by integrating the average inventory costs plus average setup costs over the stationary distribution of the normalized total workload. Since the normalized total workload W is approximated by a RBM, the total average cost is

$$\int_{y+\sum_{i=1}^{N} v_i}^{\infty} \left(\sum_{i=1}^{N} c_i(v_i + \frac{\rho_i(1-\rho_i)(\sum_{j=1}^{N} v_j - w)}{\sum_{j=1}^{N} \rho_j(1-\rho_j)}, 2\frac{w-\sum_{i=1}^{N} v_i}{\sum_{i=1}^{N} \rho_i(1-\rho_i)}, w) + \frac{k\sum_{i=1}^{N} \rho_i(1-\rho_i)}{2(w-\sum_{i=1}^{N} v_i)} \right) \alpha e^{-\alpha(w-\sum_{i=1}^{N} v_i)} dw, \tag{4.93}$$

where $\alpha = 2\sqrt{n}(1-\rho)/\sigma^2$. For the setup time problem, the average inventory cost is similar, although the stationary distribution of W is no longer exponential, but gamma (see Coffman, Puhalskii and Reiman [8]). The total average cost is

$$\int_{y+\sum_{i=1}^{N} v_i}^{\infty} \left(\sum_{i=1}^{N} c_i(v_i + \frac{\rho_i(1-\rho_i)(\sum_{j=1}^{N} v_j - w)}{\sum_{j=1}^{N} \rho_j(1-\rho_j)}, 2\frac{w-\sum_{i=1}^{N} v_i}{\sum_{i=1}^{N} \rho_i(1-\rho_i)}, w) \right) \frac{\alpha(\alpha(w-\sum_{i=1}^{N} v_i))^{\beta}}{\Gamma(\beta+1)} e^{-\alpha(w-\sum_{i=1}^{N} v_i)} dw, \tag{4.94}$$

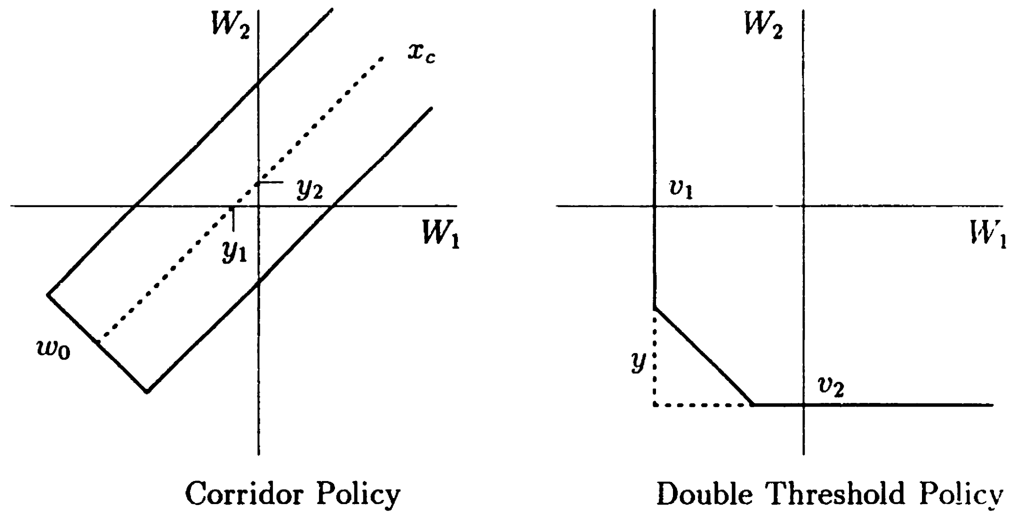

Figure 4-1: The two straw policies.

where α is as above and $\beta = s\sum_{i=1}^{N}\rho_i(1-\rho_i)/\sigma^2$. We set n equal to $(1-\rho)^{-2}$ and use a steepest descent algorithm to find the parameters $(v_1,\ldots,v_N,y)$ that minimize (4.93) and (4.94). For both the setup cost and setup time cases, we reverse the heavy traffic scaling to obtain the proposed parameter values $\tilde{v}_i = v_i/(1-\rho)$ and $\tilde{y} = y/(1-\rho)$.

The Corridor Policy

This policy can be stated in terms of switching hyperplanes in the product workload space. The hyperplanes are created to form a fixed width corridor with its long axis orthogonal to the constant workload plane (see Figure 4-1). The policy represents a natural embodiment of the "constant lot size" philosophy within a dynamic stochastic framework, and is defined by $N+2$ parameters: The cycle length τ (or corridor width), the idling threshold w_0 and the parameters $(y_1,\ldots,y_N)$, which determine the intercept of the corridor's axis. We can use these variables and the notation of the previous two sections to formulate the average inventory cost of the policy. For a given workload w, the cycle center x_i^c is equal to $w/N + y_i$ and the cycle length is τ. Product i's average inventory cost for workload w is then $c_i(w/N + y_i, \tau, w)$.

Because the cycle length is independent of workload, in both the setup cost and setup time cases the diffusion process W is approximated by a RBM on $(-\infty, w_0]$, which has an exponential steady state distribution. The drifts of the RBM are c and $c - s/\tau$ for the setup cost and time problems, respectively. Thus, the total average cost is

$$\int_{w_0}^{\infty} \left(\sum_{i=1}^{N} c_i(\frac{w}{N} + y_i, \tau, w) + \frac{k}{\tau} \right) \frac{2\sqrt{n}(1-\rho)}{\sigma^2} e^{-\frac{2\sqrt{n}(1-\rho)}{\sigma^2}(w_0 - w)} dw \tag{4.95}$$

for the setup cost problem, and

$$\int_{-\infty}^{w_0} \left(\sum_{i=1}^{N} c_i(\frac{w}{N} + y_i, \tau, w) \right) \frac{2(\sqrt{n}(1-\rho) - s/\tau)}{\sigma^2} e^{-\frac{2(\sqrt{n}(1-\rho)-s/\tau)}{\sigma^2}(w - w_0)} dw \tag{4.96}$$

for the setup time problem. The cost-minimizing parameters for (4.95) and (4.96) are determined by a steepest descent algorithm. Although we were able to use $n = (1-\rho)^{-2}$ in our computations, in the setup time problem one must be careful to choose the heavy traffic scaling factor n so that $\sqrt{n}(1-\rho)$ is greater than s/τ, thereby guaranteeing a well defined integral. This inequality is simply the stability condition that the fraction of time the server spends processing units and setting up must be less than one. Finally, the proposed parameter values are given by $\tilde{y}_i = y_i/(1-\rho)$ and $\tilde{\tau} = \tau/(1-\rho)$.

The Setup Cost Problem: Two Product Case

To standardize the two-product scenarios, we set the service rates $\mu_1 = \mu_2 = 1$ and control the utilization rates ρ_i by varying the demand rates λ_i. We also set $h_2 = 1$ and, by modifying h_1, b_1 and b_2, select product 2 as the least cost product. Inventory costs and arrival rates are identical across products in the 18 *symmetric* cases, and each case is characterized by three parameters: Backorder cost, traffic intensity and setup cost per cycle. We examine all permutations of values shown in Table I; notice that some of these scenarios grossly violate the heavy traffic conditions. The parameters for the first asymmetric case are $\lambda_1 = 0.6$, $\lambda_2 = 0.3$, $h_1 = 2$, $b_1 = 10$, $b_2 = 5$ and $K = 200$. The second asymmetric case is the same as the first, except that the backorder cost is

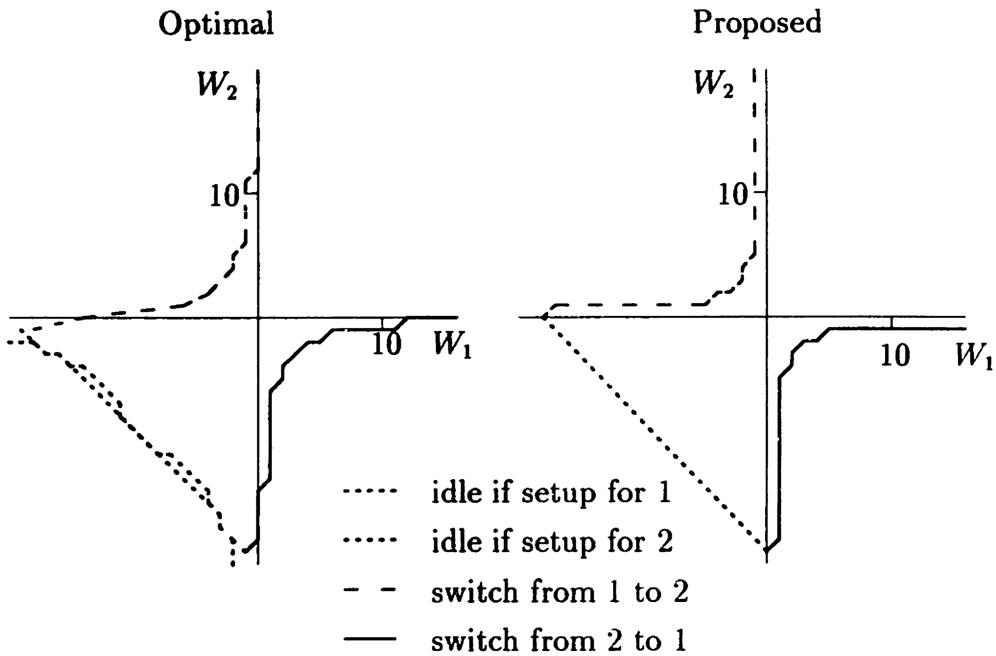

Figure 4-2: Switching curves for a symmetric setup cost case.

doubled to $b_1 = 20$ and $b_2 = 10$.

	Backorder Cost b	Setup Cost K	Traffic Intensity ρ
Low	5	20	0.5
Medium		100	0.7
High	10	200	0.9

Table I: The 18 test cases for the symmetric two-product setup cost problem.

Table II displays the results for the 20 two-product cases and Tables III to V show the averages (for the 18 symmetric cases) over individual parameters for each policy. The switching curves for the optimal and proposed policies for the ($b = 5$, $K = 200$, $\rho = 0.9$) two-product symmetric case is depicted in Figure 4-2, and corresponding curves for the $b_1 = 10$ asymmetric case are displayed in Figure 4-3.

Backorder Cost b	Setup Cost K	Traffic Intensity ρ	Optimal Policy Gain	Proposed Policy	Suboptimality Corridor Policy	Gen. Base Stock Policy
5	20	0.5	4.30	6.0%	6.5%	9.6%
5	20	0.7	6.73	8.3%	1.5%	20.1%
5	20	0.9	17.99	3.8%	7.5%	26.8%
5	100	0.5	7.00	6.2%	6.4%	18.5%
5	100	0.7	9.72	2.1%	2.6%	18.5%
5	100	0.9	20.14	1.5%	8.1%	27.4%
5	200	0.5	9.14	15.0%	7.2%	25.3%
5	200	0.7	12.24	2.6%	2.9%	21.3%
5	200	0.9	22.22	0.7%	7.5%	26.1%
10	20	0.5	5.30	14.1%	2.2%	20.0%
10	20	0.7	8.41	13.9%	4.2%	24.6%
10	20	0.9	23.58	6.0%	11.7%	33.0%
10	100	0.5	7.98	8.0%	6.1%	15.5%
10	100	0.7	11.31	4.5%	3.7%	25.5%
10	100	0.9	25.44	3.3%	9.6%	34.9%
10	200	0.5	10.21	6.4%	7.2%	18.6%
10	200	0.7	13.79	3.5%	4.7%	24.3%
10	200	0.9	27.26	2.1%	10.0%	35.0%
	Asym.	b_1=10	28.71	2.6%	13.1%	35.8%
	Asym.	b_1=20	35.76	3.4%	14.9%	45.8%

Table II: Results for the two-product setup cost cases.

Setup Cost: Five-Product Cases

We set $\lambda_i = 0.18$ and $\mu_i = 1$ for $i = 1, \ldots, 5$ for each of the six cases, resulting in a traffic intensity of 0.9. We also set $b_i = 5h_i$ for $i = 1, \ldots, 5$ for half the cases and $b_i = 10h_i$ for the other half. Each case is characterized by h_i, b_i and the setup cost. Four of the six cases are symmetric ($h_i = 1$ for $i = 1, \ldots, 5$) and two of the six cases are asymmetric ($h_i = i$ for $i = 1, \ldots, 5$). The average cost for each policy (along with 95% confidence intervals) is displayed in the first six rows of Table VI.

The Setup Time Problem

As in the two-product setup cost test cases, we assume that $\mu_1 = \mu_2 = 1$ and $h_2 = 1$. In the 12 symmetric scenarios, each product's inventory costs and service utilizations

	Backorder Cost b	Setup Cost K	Traffic Intensity ρ
Low	5.1%	8.7%	9.3%
Medium		4.3%	5.8%
High	6.9%	5.0%	2.9%

Overall Average Suboptimality = 6.0%

Table III: Average suboptimality of the proposed policy: Setup cost problem.

	Backorder Cost b	Setup Cost K	Traffic Intensity ρ
Low	5.6%	5.6%	5.9%
Medium		6.1%	3.3%
High	6.6%	6.6%	9.1%

Overall Average Suboptimality = 6.1%

Table IV: Average suboptimality of the corridor policy: Setup cost problem.

are identical and we vary only the backorder cost, the traffic intensity and the average setup time per cycle. Table VII reports all of the permutations of values analyzed. The first asymmetric scenario is defined by $\lambda_1 = 0.6$, $\lambda_2 = 0.3$, $\mu_1 = \mu_2 = 1$, $h_1 = 2$, $h_2 = 1$, $b_1 = 10$, $b_2 = 5$ and $s = 20$. The second asymmetric scenario is identical except that the backorder costs are $b_1 = 20$ and $b_2 = 10$.

The individual results for the 14 runs are displayed in Table VIII and policy summaries for the 12 symmetric runs are given in Tables IX to XI. In addition, Figures 4-4 and 4-5 provide a graphical depiction of the proposed and optimal policies for a symmetric case ($b = 5$, $s = 2$, $\rho = 0.9$) as well as the $b_1 = 10$ asymmetric case.

Results for two five-product scenarios can be found in Table VI; they are identical to the setup cost scenarios described in §3.2, except that setup times (with $s = 50$) are incurred rather than setup costs.

	Backorder Cost b	Setup Cost K	Traffic Intensity ρ
Low	21.5%	22.4%	17.9%
Medium		23.4%	22.4%
High	25.7%	25.1%	30.5%

Overall Average Suboptimality = 23.6%

Table V: Average suboptimality of the generalized base stock policy: Setup cost problem.

Back-order Cost	Setup Cost or Time	Cost Structure	Cost of Proposed Policy	Cost of Corridor Policy	Cost of Gen. Base Stock Policy
$b = 5$	$K = 50$	Symmetric	25.32(±0.46)	28.78(±0.63)	33.23(±0.45)
$b = 5$	$K = 500$	Symmetric	37.23(±0.10)	37.25(±0.32)	44.02(±0.20)
$b = 10$	$K = 50$	Symmetric	36.79(±0.73)	39.39(±0.91)	40.54(±0.68)
$b = 10$	$K = 500$	Symmetric	47.08(±0.38)	46.02(±0.27)	54.29(±0.52)
$b = 5$	$K = 500$	Asymmetric	79.91(±0.41)	86.65(±0.67)	121.11(±1.57)
$b = 10$	$K = 500$	Asymmetric	98.46(±0.77)	105.98(±1.49)	138.88(±1.14)
$b = 5$	$s = 50$	Symmetric	215.4(±4.9)	228.0(±16.1)	214.1(±2.6)
$b = 10$	$s = 50$	Symmetric	264.7(±10.4)	532.5(±136.8)	260.2(±4.7)
$b = 5$	$s = 50$	Asymmetric	610.8(±8.9)	683.9(±35.2)	661.0(±9.1)
$b = 10$	$s = 50$	Asymmetric	737.4(±18.7)	827.9(±66.5)	791.7(±16.1)

Table VI: Results for the five-product cases.

Observations

Our observations from the numerical results are summarized in this subsection. The five-product cases are discussed after the two-product cases.

Performance of the proposed policy: In the setup cost cases, the proposed policy's average suboptimality is 6.0% over the 18 symmetric scenarios. The policy performs very well when the heavy traffic conditions are satisfied; for example, the suboptimality is 0.7% when $b_1 = 5$, $K = 200$ and $\rho = 0.9$. Considering that the proposed policy was constructed via a heavy traffic approximation, it operates reasonably well over a wide range of system parameters, including a low utilization rate of 0.5. Not surprisingly,

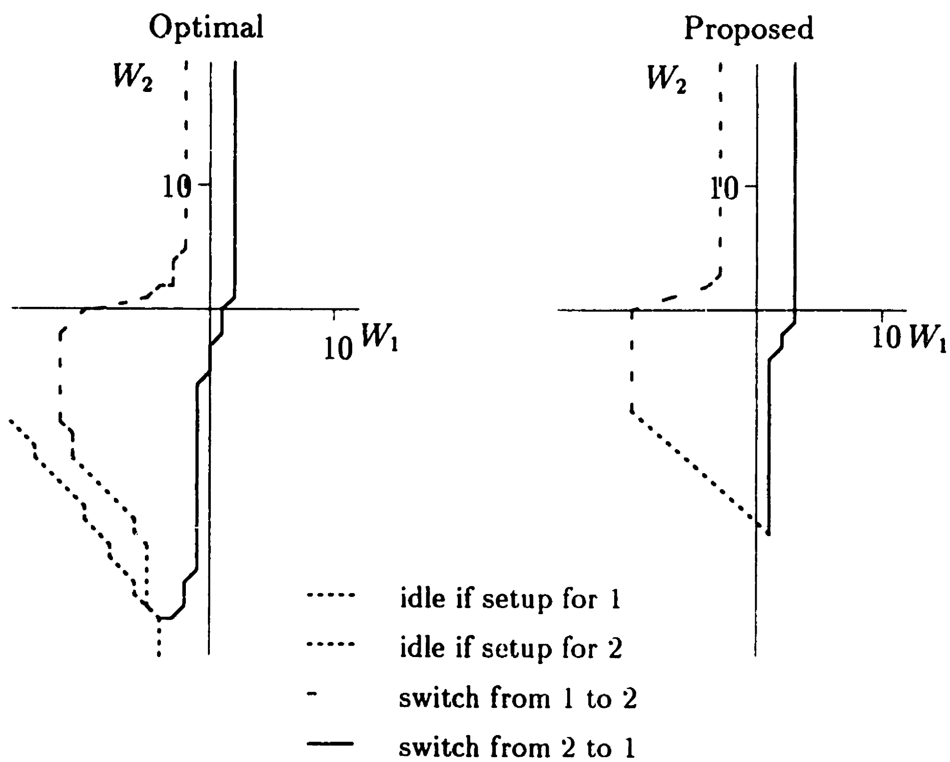

Figure 4-3: Switching curves for the $b_1 = 10$ asymmetric setup cost case.

the policy performs worst when the traffic intensity is low, the setup costs are small, and the backorder costs are high. The policy also performs well (2.6% and 3.4% suboptimalities) in the asymmetric cases.

In the setup time cases, the average suboptimality over the 12 symmetric cases is 7.2%. The policy performs very well (1.8% average suboptimality) when the traffic intensity is high, but degrades somewhat in the lighter traffic cases. It also performs well in the asymmetric cases (1.5% and 3.3% suboptimalities).

Switching curves: The switching curves of the proposed and optimal policies are remarkably similar in Figures 4-2 to 4-5 and are unlike either the corridor or generalized base stock policies. In the two symmetric problems (Figures 4-2 and 4-4), these curves have the same general shape as predicted by our heavy traffic analysis: A distinctive constant-workload idling threshold, a wide cycle time for large positive and negative inventories and a small cycle time about the zero total workload level.

	Backorder Cost b	Setup Time s	Traffic Intensity ρ
Low	5	2	0.5
Medium			0.7
High	10	20	0.9

Table VII: The 12 test cases for the symmetric setup time problem.

Backorder Cost b	Setup Time s	Traffic Intensity ρ	Optimal Policy Gain	Proposed Policy	Suboptimality Corridor Policy	Gen. Base Stock Policy
5	2	0.5	3.84	9.0%	12.8%	6.6%
5	2	0.7	7.45	7.4%	8.3%	7.8%
5	2	0.9	25.31	1.4%	6.7%	9.5%
5	20	0.5	13.58	10.1%	24.3%	19.9%
5	20	0.7	26.39	5.9%	16.0%	9.6%
5	20	0.9	79.40	1.4%	8.7%	3.5%
10	2	0.5	5.15	9.3%	11.0%	11.8%
10	2	0.7	9.85	8.2%	7.1%	9.0%
10	2	0.9	33.24	1.8%	4.5%	13.7%
10	20	0.5	17.75	18.7%	26.5%	17.9%
10	20	0.7	33.67	10.7%	16.8%	12.8%
10	20	0.9	98.93	2.6%	7.0%	5.9%
	Asym.	b_1=10	104.91	1.5%	30.6%	11.5%
	Asym.	b_1=20	129.42	3.3%	60.9%	15.8%

Table VIII: Results for the two-product setup time cases.

In the asymmetric setup cost problem in Figure 4-3, the three-region categorization predicted by the heavy traffic theory is easily recognizable in the optimal policy. Figure 4-5 confirms that lot sizes shrink as the idling threshold is approached. Finally, as the total workload $\tilde{w}$ tends to infinity, lot sizes appear to be growing roughly with $\tilde{w}$ in Figure 4-4 and with $\sqrt{\tilde{w}}$ in Figure 4-5.

Two key differences between the proposed and optimal policies emerge from studying Figures 4-2 to 4-5; numerical results (not reported here) verify that both discrepancies dissipate as the traffic intensity approaches unity, and get more severe in the lower utilization cases. First, in all four figures, the proposed heavy traffic policies

	Backorder Cost b	Setup Time s	Traffic Intensity ρ
Low	5.9%	6.2%	11.8%
Medium			8.1%
High	8.5%	8.2%	1.8%

Overall Average Suboptimality = 7.2%

Table IX: Average suboptimality of the proposed policy: Setup time problem.

	Backorder Cost b	Setup Time s	Traffic Intensity ρ
Low	12.8%	8.4%	18.7%
Medium			12.1%
High	12.1%	16.5%	6.7%

Overall Average Suboptimality = 12.5%

Table X: Average suboptimality of the corridor policy: Setup time problem.

have a tendency to backorder more than the optimal policy; this observation is most obvious in the upper right portion of Figure 4-4. Because the HTAP time scale decomposition does not hold precisely for the original stochastic system, the optimal policy hedges against backorders slightly more than the proposed heavy traffic policy, which assumes that the inventory levels respond in a deterministic fashion in the fluid limit. In terms of these figures, the cycle lengths (i.e., the distance along the total workload line between the solid and dashed curves) tend to be slightly smaller in the optimal policy; consequently, the workload process spends less time in the backorder region and some of our remarks in Chapter 4.2 regarding the inventory levels at switching epochs only hold in very heavy traffic. This limitation of the heavy traffic theory was also noted in Wein [46].

The other main discrepancy occurs near the idling threshold in the asymmetric cases: In Figure 4-5, the optimal lot sizes for product 1 decrease, rather than stay constant, as the workload idling threshold is approached, and in Figures 4-3 and 4-

	Backorder Cost b	Setup Time s	Traffic Intensity ρ
Low	9.5%	9.8%	14.0%
Medium			9.8%
High	11.8%	11.6%	8.2%

Overall Average Suboptimality = 10.7%

Table XI: Average suboptimality of the generalized base stock policy: Setup time problem.

5 there is a different idling threshold for each product. This discrepancy can be explained as follows: when the total workload is negative, the switchover cost (in Figure 4-3) or time (in Figure 4-5) makes it beneficial to be setup for product 1, so as to efficiently protect against costly product 1 backorders. If ρ is not close to one, then it is likely that the total order workload will decrease while producing product 2; that is, the decrease in product 2's order workload will exceed the increase in product 1's order workload. The optimal policy takes advantage of this imbalance by allowing product 2's inventory (i.e. negative orders) to grow beyond the product 1 idling threshold; this extra product 2 inventory allows the server to idle while setup for product 1.

Performance of the corridor policy. The corridor policy exhibits erratic behavior. The policy performs very well in the symmetric setup cost cases (it outperforms the proposed policy in six of the 18 scenarios, all of which have low or medium utilizations), but degrades slightly at high utilization. A comparison of Figures 2 and 3 leads us to believe that the parameters of the policy are being set correctly at high utilizations, and the performance degradation is due to the corridor policy's inability to employ mean lot sizes that are state-dependent.

The corridor policy does not perform as well in the symmetric setup time cases; it is not able to increase the cycle length τ for large total order workloads and so has difficulty recovering from this high backorder region. In contrast to the symmetric setup cost cases, the corridor policy's performance diminishes in light traffic; we have

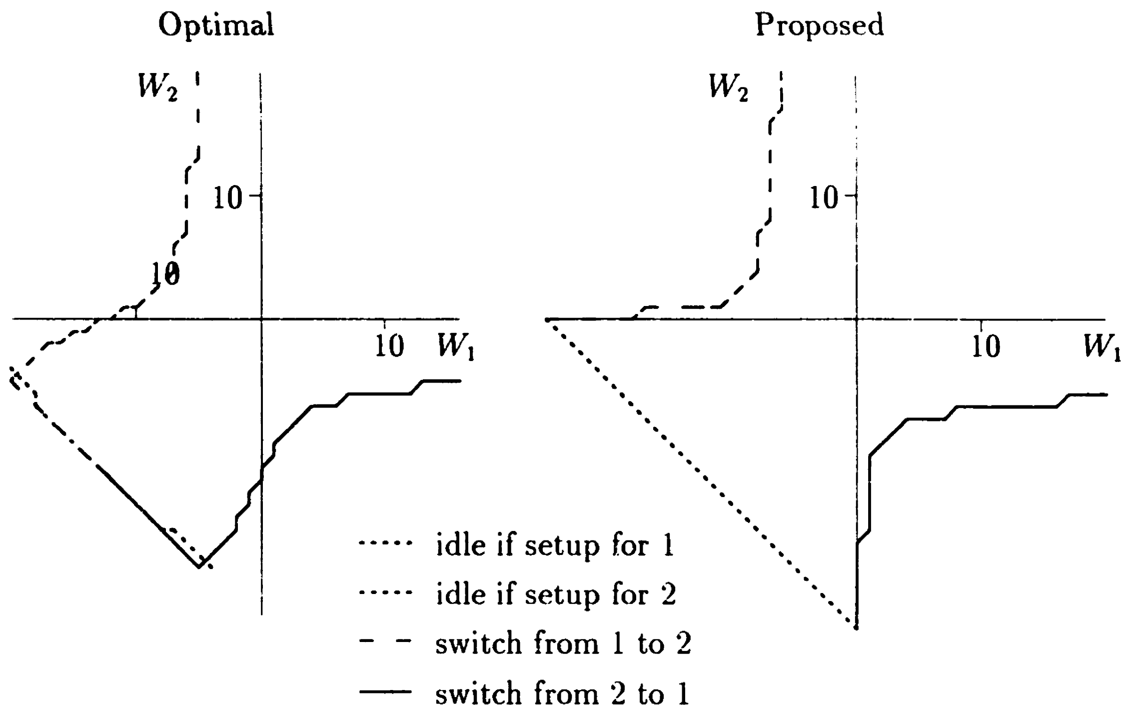

Figure 4-4: Switching curves for a symmetric setup time case.

not determined how much of this degradation is due to the inaccuracy of the heavy traffic approximation at low utilizations, and how much is intrinsic to the policy.

The corridor policy performs much worse when asymmetry is present: Its suboptimality is 13.1% and 14.9% in the setup cost cases and increases to 30.6% and 60.9% in the setup time cases. Comparing Figures 4-1, 4-3 and 4-5, it would appear that the corridor policy would never be very close to optimal for an asymmetric problem. In fact, Figure 4-5 suggests that a *hyperplane corridor* policy (see Figure 7 of Sharifnia, Caramanis and Gershwin [41]) might perform reasonably well in the asymmetric setup time problem; in the two-product case, the two lines forming the corridor in Figure 4-1 would not be parallel in the hyperplane corridor policy, but would intersect at an idling point in the lower left portion of the graph and generate a cone-shaped corridor emanating out in the northeasternly direction.

Performance of the generalized base stock policy. The generalized base stock policy performs better in the setup time cases than in the setup cost cases: Its average

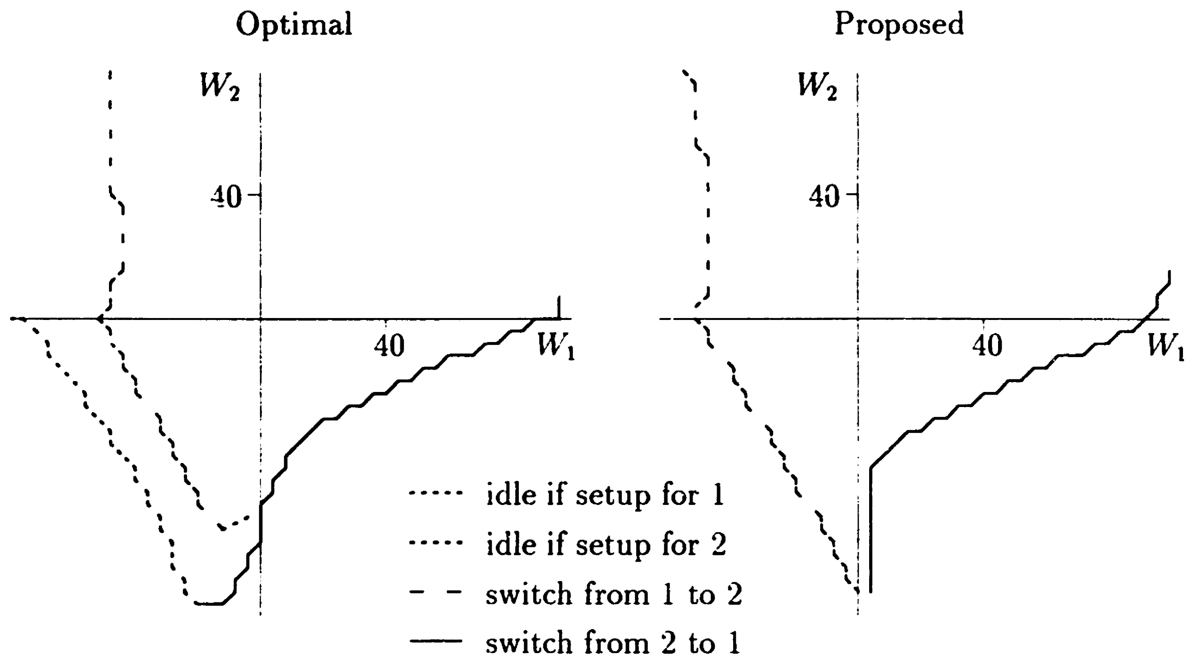

Figure 4-5: Switching curves for the $b_1 = 10$ asymmetric setup time case.

suboptimality is 23.6% for the 18 symmetric setup cost scenarios and 10.7% for the 12 symmetric setup time cases. In contrast to the corridor policy, the generalized base stock policy's use of large lot sizes when the total order workload is positive is a key reason for its ability to avoid poor performance in the symmetric setup time cases; however, these large lot sizes lead to considerable backordering in the setup cost scenarios. Like the corridor policy, the generalized base stock policy's performance deteriorates at high utilizations in the setup cost cases and at low utilizations in the setup time cases.

The generalized base stock policy's suboptimality is 35.8% and 45.8% in the asymmetric setup cost cases and 11.5% and 15.8% in the asymmetric setup time case. It is interesting to note that the generalized base stock policy handles expensive inventory in a manner opposite to that of the proposed policy. The order-up-to level of the most costly good is set larger than those of less expensive products to reduce the risk of expensive backordering; in contrast, the proposed policy minimizes the excess or deficit amounts of expensive inventory. It is clear that the generalized base stock policy is incapable of closely approximating the optimal setup cost solution in Figure 4-3.

Five-product examples: In the six setup cost cases in Table VI, the proposed and corridor policies are roughly comparable in the two $K = 500$ symmetric cases and the corridor policy is about 10% more costly in the two $K = 50$ symmetric cases and about 8% more expensive in the two asymmetric cases. The generalized base stock policy does not fare as well in the four symmetric setup cost cases, incurring an 18.7% cost increase relative to the proposed policy, on average. Once again, the generalized base stock policy performs very poorly in the asymmetric setup cost cases.

In contrast, the generalized base stock policy performs slightly better than the proposed policy in the two symmetric setup time cases in Table VI, and is about 8% more costly than the proposed policy in the two asymmetric cases. Both of these policies outperform the corridor policy in the four setup time cases. The corridor policy is about 12% more costly than the proposed policy in the two asymmetric cases, but performs extremely poorly in one of the two symmetric cases.

To compare the relative cost differences in the two-product cases and the five-product cases, we can identify the six symmetric cases in Table VI with their two-product counterparts in Tables II and VIII; for example, the first scenario in Table VI corresponds to the $b = 5$, $K = 20$, $\rho = 0.9$ case in Table II. For the four setup cost cases, the cost increases of the straw policies relative to the proposed policy are somewhat larger for the two-product cases: the generalized base stock policy's average cost increase is 5.8% for the two-product cases versus 4.7% for the five-product cases, and the corresponding quantities for the generalized base stock policy are 26.3% and 18.7%, respectively. For the two setup time cases, the average cost increase of the generalized base stock policy is 2.7% for the two-product scenarios and -1.1% for the five-product cases. Disregarding the poor performance of the corridor policy in one of the five-product symmetric setup cost scenarios, it appears that the relative cost advantage of the proposed policy degrades slightly when the number of products increases from two to five; however, further experiments are required to fully investigate this issue.

Lack of Robustness. Simulation results not reported here show that the performance of the three policies are rather sensitive to the policy parameters, particularly in the setup time problem; this is somewhat surprising, given the robustness of some simpler models (e.g., the EOQ model) that capture the tradeoff between inventory costs and setups. Because it is unable to increase its lot sizes as the total inventory decreases, the corridor policy is clearly the least robust of the three policies: if the corridor width is set too narrow (as apparently happened in the eighth row of Table VI) then stability problems can set in (notice the confidence intervals for this case).

The $s/\sqrt{n}$ Refinement. Recall that the $s/\sqrt{n}$ refinement described in Chapter 3.5 is incorporated into the proposed policy, but not the two straw policies. We tested all three policies with and without the refinement, and summarize our findings here. The refinement had a minor effect on the performance of the proposed policy in the $\rho = 0.9$ cases; however, by decreasing the cycle length, it significantly improved performance in the lower utilization cases. The refinement had a mixed influence on the generalized base stock policy, sometimes improving and sometimes degrading performance; overall, it slightly impaired performance. The refinement had a negative effect on the corridor policy, and let to a severe stability problem in the eighth row of Table VI.

Summary of SELSP Results

Although additional asymmetric cases need to be investigated before drawing definitive conclusions for the two-product problems, our observations can be summarized as follows.

The proposed policy performs very well in the 34 two-product cases: Figures 3 to 6 confirm that it captures nearly all of the complexities of the optimal policy, its suboptimality is 6.0% over the 34 cases (and 2.1% over the 12 cases that do not obviously violate the heavy traffic conditions), and it is quite robust with respect to the heavy traffic conditions, especially considering that most potential industrial applications for the SELSP are in settings with high traffic intensities. However, the relative superiority of the proposed policy appears to degrade slightly as the number of products increases, and this issue requires further investigation.

The two straw policies are not flexible enough to consistently capture the subtleties of the optimal policy. The corridor policy outperforms the generalized base stock policy in 24 of the 34 two-product examples, and its average suboptimality is 11.2% as compared to 19.5% for the generalized base stock policy. Nonetheless, in the setup time cases the corridor policy fails to use large lot sizes when the total workload is positive and can perform erratically (see Table VIII). The generalized base stock policy is never close to optimal and performs poorly in the asymmetric setup cost cases.

4.4.2 Mixed System with Due-dates

In this section we construct a series of experiments for a mixed system with due-dates. In the previous section (and Reiman and Wein [39]) it was shown that the HTAP is robust in policies for systems with either setup costs or setup times and to a limited extent multi-product systems. Here we wish to examine two further issues: the effects of due-dates on scheduling and the ability of the HTAP to derive a policy for a mixed system. Thus, we only consider a two product case with setup times. Using the same parameters as the $b_1 = 10, b_2 = 5$ two product asymmetric setup time SELSP case, we vary the length of the due-date lead time, f_i, and compare our proposed policy to several straw ones. We set $f_1 = f_2$ and test f_i for values equal to 0, 20 and 100. Since a true optimal solution cannot be found, we base our study on the results of a discrete time simulation of the systems. Again, service times, setup times and demand interarrival times are exponentially distributed and, like the two product SELSP case, the service is pre-emptive.

Straw and Proposed Policies

Since we do not have a convenient point of reference provided by an optimal policy, we use a different collection of straw policies. We consider three test policies. The first is a simple hybrid base-stock/exhaustive policy where the customized product is serviced to exhaustion and the standardized product is produced up to a base stock level equal to a $-\tilde{v}$ order workload level. The base-stock level is calculated in a fashion similar to the one described in the standardized, no due-date case. The cycle center for the customized good (or first product class) is set to $\tau(w)\rho_1(1-\rho_1)/2$. The cycle center of the standardized good is set to $\tau(w)\rho_2(1-\rho_2)/2+\tilde{v}$. Let $v=\tilde{v}/\sqrt{n}$, then the cycle length is

$$\tau(w) = 2\frac{w-v}{\rho_1(1-\rho_1)+\rho_2(1-\rho_2)}. \tag{4.97}$$

We then use the cost equations from Chapter 4.1 to find the average cost for this policy given by

$$\begin{aligned} &\int_{-v}^{\infty}\Big(c_1\big(\tfrac{w-v}{\rho_1(1-\rho_1)\sum_{i=1}^{2}\rho_i(1-\rho_i)}, 2\tfrac{w-v}{\sum_{i=1}^{2}\rho_i(1-\rho_i)}, w\big) \\ &+c_2\big(v+\tfrac{w-v}{\rho_2(1-\rho_2)\sum_{i=1}^{2}\rho_i(1-\rho_i)}, 2\tfrac{w-v}{\sum_{i=1}^{2}\rho_i(1-\rho_i)}, w\big)\Big)\tfrac{\alpha(\alpha(w-v))^{\beta}}{\Gamma(\beta+1)}e^{-\alpha(w-v)}dw, \end{aligned} \tag{4.98}$$

where α is as above and $\beta = s\sum_{i=1}^{2}\rho_i(1-\rho_i)/\sigma^2$. We set n equal to $(1-\rho)^{-2}$ and use a steepest descent algorithm to find the parameter v which minimizes (4.98).

The second policy is again a hybrid base-stock/exhaustive policy where the base-stock level is determined by exhaustive search. This is simply performed by doing multiple simulations (the results of which are not shown here). These two hybrid policies together offer the opportunity to determine the accuracy of the HTAP. The third policy is the proposed policy without due-date considerations, that is, the policy based on estimating workload by the number of orders in queue outlined in Chapter 4.3 is calculated with a due-date lead time f_i set to zero. In addition, we look at two variants of the proposed policy. They use the two methods of approximating the workload level in queue given in section 4.3. The results are given in Table XII (the abbreviation "Order Num." refers to the proposed policy where workload is approximated by the number of orders in queue, "Next D-date" refers proposed policy where the earliest due-date is used to find the workload). Switching curves for the optimal search hybrid policy and proposed policy are given in Figures 4-6 to 4-8.

Observations

Hybrid Policies. The long run average cost for both the exhaustive search and HTAP

Due-Date Lead Time	Cost of Hybrid HTAP	Cost of Hybrid Search	Cost of Proposed w/o D-date	Cost of Proposed Order Num.	Cost of Proposed Next D-date
0	290.3 (± 3.2)	290.0 (± 2.7)	274.6 (± 1.9)	274.6 (± 1.9)	295.8 (± 3.6)
20	201.8 (± 2.7)	199.5 (± 3.2)	186.3 (± 2.1)	184.1 (± 2.7)	197.0 (± 3.9)
100	122.8 (± 1.4)	121.3 (± 1.3)	158.3 (± 0.9)	109.7 (± 1.1)	109.8 (± 1.2)

Table XII: Results for the Mixed System.

hybrid policies decreases with the longer due-dates. This makes intuitive sense: with no due-date lead time orders for the high cost customized goods are immediately backordered, driving up costs. As the due-date lead times are added, orders for the customized good are less tardy and costs fall. Moreover, there is little change in the costs due to the standardized product class: its average holding and backorder costs remain approximately 23 and 14 respectively. Almost all of the savings are due to the high cost product. This is as one would expect: the base-stock level for the standardized product is chosen so as to optimally balance its earliness and tardiness costs. Due-dates alter the base-stock level necessary to achieve this optimal balance, but cannot alter the balance itself. Mathematically this result is quickly seen in the expression for the cost per cycle in Chapter 4.1.1 and is discussed in more detail in Chapter 4.5.2. Although in these three cases long run average costs were decreasing with longer due-dates, one would expect that eventually they would start to increase again as the due-date lead times become inordinately long and holding costs become excessive.

It is also interesting to compare the exhaustive search base-stock policy with the one derived by the HTAP. They are nearly identical. The base-stock levels for each are 45, 37 and 13 for the f_i = 0, 20 and 100 cases respectively. The HTAP levels are 39, 28 and 5. Although the base-stock levels for the search policy are consistently higher than those of the HTAP policy, the actual difference in long run average cost is statistically insignificant. For the exhaustive/base-stock policy the long run average cost as a function of base-stock level appears to have a shallow slope about the optimal solution. The heavy traffic theory is able to identify a set of parameters which performs

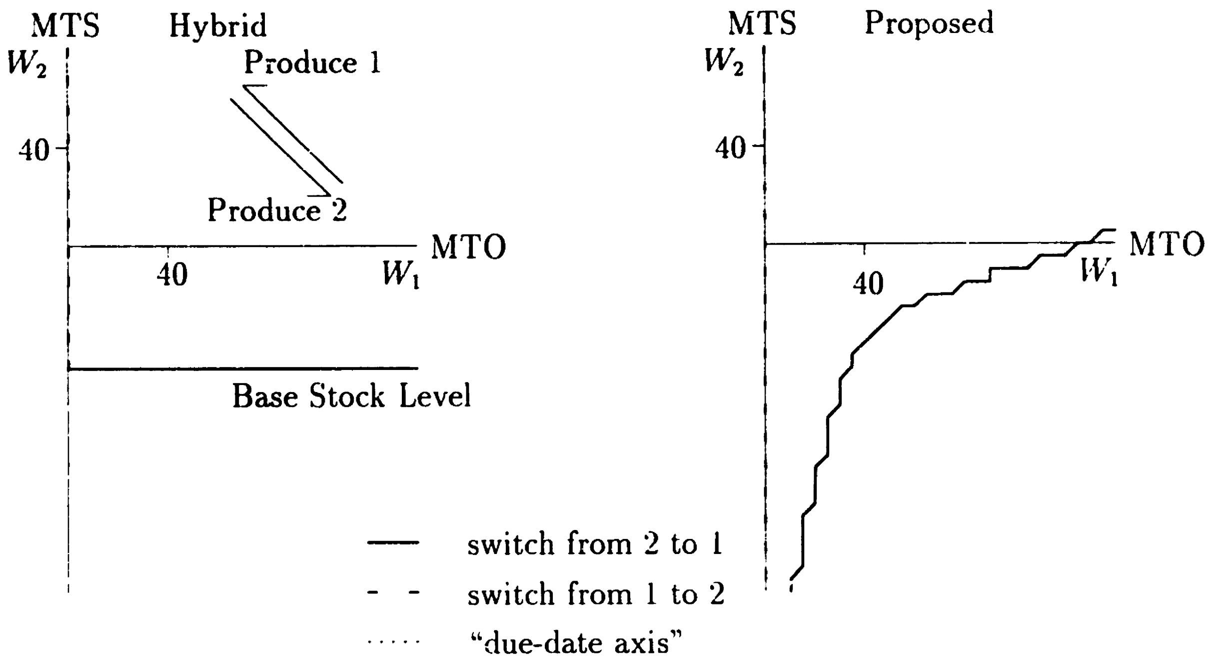

Figure 4-6: Switching curves for a mixed system with 0 due-date lead time.

quite well.

Proposed Policies. The proposed policies do quite well and, as with the hybrid policies, have decreasing long run average costs as due-dates increase. The reason for this is also the same: costs go down as there are opportunities to avoid tardiness of the high cost customized good. The proposed policies have several advantages over the hybrid policies: they are able to avoid large buildups of high cost product 1 orders both in queue and waiting to be shipped. For the $l = 0$ case, the proposed policies can avoid severe backordering by switching to setup product 1 if there is an excessive number of orders in queue. However, due to the severity of the setup penalty, cycle length τ is long and so the product 1 buildup must be large. Thus, the marginal benefit of the proposed policy over the hybrid policy is not great.

As the due-date lead time increases, however, the proposed policy has a greater opportunity to avoid excessive product 1 costs. Cycle length τ is still large but the policy is able to attain a balance between earliness and tardiness costs by cycle center placement. As seen in Figure 4-8, the amount of product 1 workload is maintained

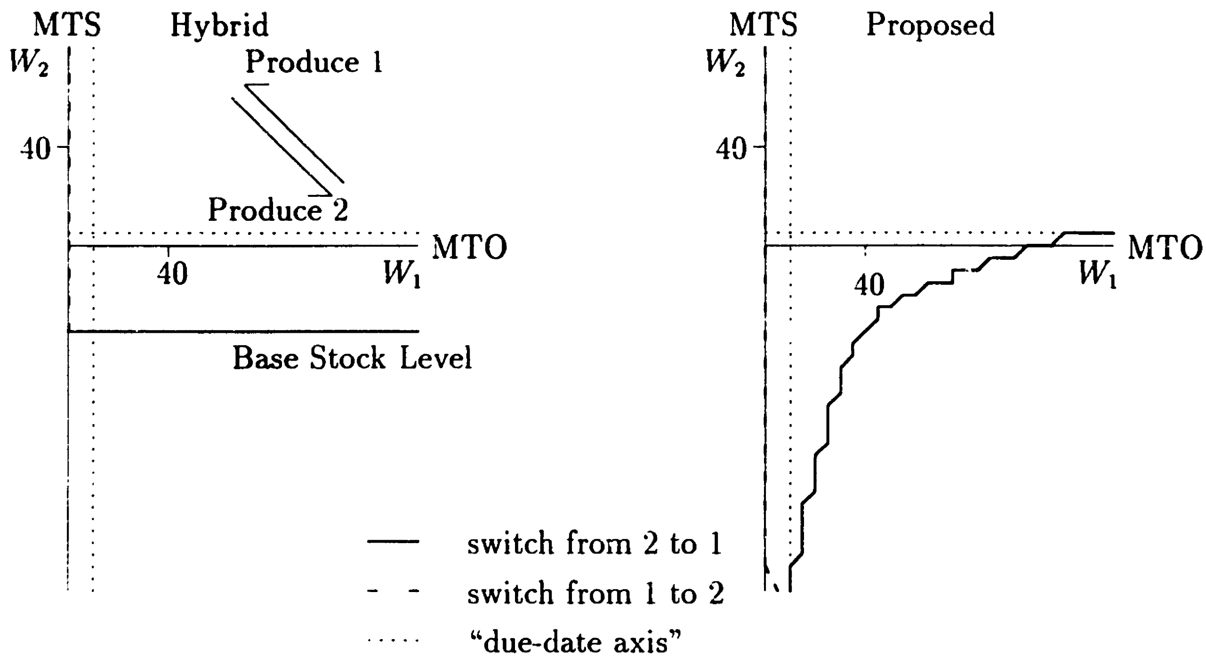

Figure 4-7: Switching curves for a mixed system with a due-date lead time of 20.

about the "due-date axis" $\rho_1 f_1$, the level of product 1 workload necessary for a new order waits in queue for an amount of time exactly equal to the due-date lead time. The hybrid policy is not able to perform this type of cost minimization.

The two methods for estimating the workload present in each queue, however, appear to have different measures of success. The one where workload is approximated by the number in queue clearly outperforms the earliest due-date approach. The difference, however, disappears as the due-date lead time length grows. This can potentially be explained by the variance of the estimate. The earliest due-date approach uses the next due-date in queue to approximate the number in queue which in turn is used to calculate the workload present. Thus, is has a higher variance while counting the number in queue is more direct. Once the due-date lead time becomes longer, there is more to average, making both estimates tighter. One of the motivations for the earliest due-date proceedure was to see if the policy was sensitive to single orders which might not account for much work but might be far overdue. The comparison between these implementations of the proposed policy implies that this factor is negligable and that

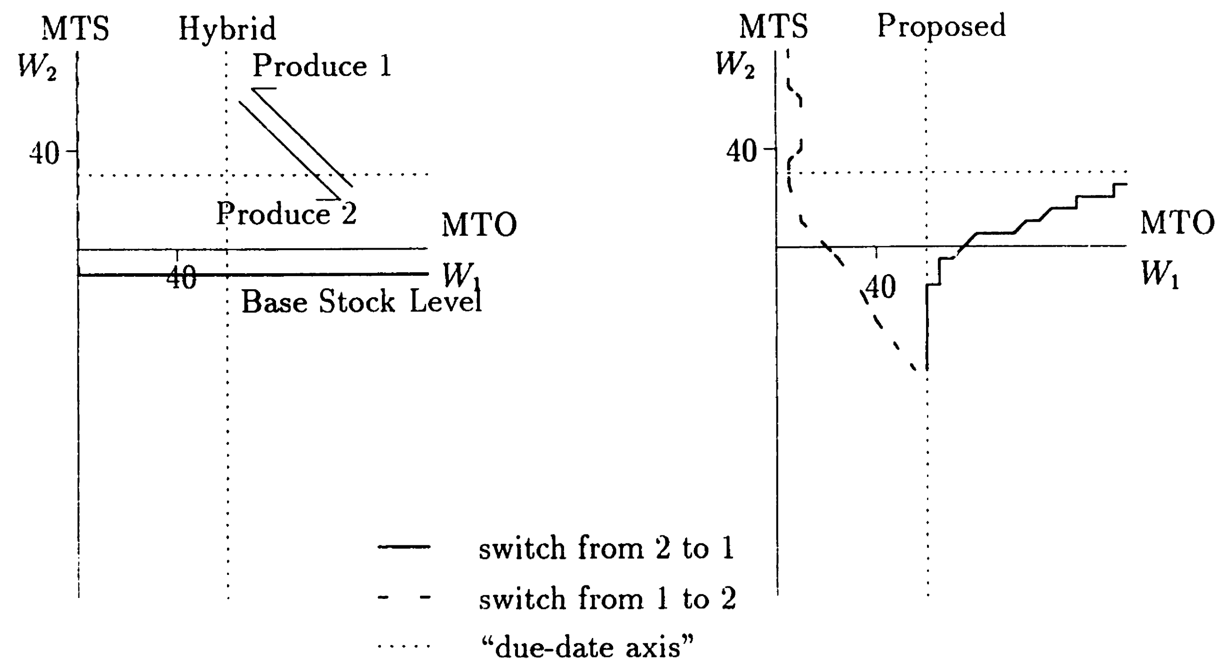

Figure 4-8: Switching curves for a mixed system with a due-date lead time of 100.

stray, overdue orders should not be used to force a setup of the machine.

MTS/MTO Partitions. It is also interesting to note the similarities between the mixed system with a hybrid policy and the two product SELSP with a generalized base-stock policy. The system parameters used in the SELSP asymmetric, setup time, $b_1 = 10$ case are identical to the ones used in the mixed system. As implied in Chapter 4.1, when the due-date lead time in the mixed system approaches v_1/ρ_1, where v_1 is the base-stock level from the SELSP case, the long run average costs should be about equal for the two systems. Slight differences may still exist because the generalized base-stock policy includes a non-trivial idling threshold not present in the hybrid system. For this SELSP scenario, v_1 was found to be 49, making the critical due-date lead time, l_1, equal to 81.6. Although we did not simulate this case, from the results shown, it is not difficult to intuit that this case would have a much lower cost than the $f_i = 0$ case.

This sheds great light on the relationship between MTS/MTO partitions and due-dates. As stated in the literature review, much of the motivation for Just-in-Time manufacture is derived from demand lead time: if demand is sufficiently forecasted then orders for goods can be filled exactly on time and so no finished goods inventory need be held. The foreknowledge of demand means that standardized goods can be serviced in a MTO fashion instead of MTS. We see this here, but in a slightly different form. Sufficient (order $\sqrt{n}$) due-date lead times can imply that a mixed system has the same long-run average cost as a purely standardized system. Thus, one is indifferent to servicing orders for that class in either a MTS or a MTO fashion. Yet, switching to a MTO service does not reduce costs as it is still optimal to fill orders early and incur the holding cost. These early orders provide a buffer against backordering as would a finished goods inventory. This is discussed in greater detail in Chapter 4.5.6.

4.4.3 Standardized System with Due-dates

In our last series of experiments, we wish to test the result that the long run average cost for standardized products is independent of the due-date lead time. We study a machine which services two identical standardized product classes, each with a holding cost of one and a tardiness cost of five, and has a setup time penalty with a cycle mean of $s = 20$. We vary due-date lead times and utilization. As with the mixed systems in the previous section, a true optimal solution cannot be found, and so we implement a discrete time simulation. Again, service times, setup times and demand interarrival times are exponentially distributed and the service is pre-emptive.

We look at three utilizations (ρ equal to 0.5, 0.7 and 0.9) and three due-date lead times (f_i equal to 0, 20 and 100). This makes for a total of nine cases. The results are given in Table XIII.

Observations

The due-date lead time f_i equal to zero corresponds well with the dynamic programming results given in the SELSP section. As due-date lead times increase, the long run average cost remains relatively constant, as described by the theory. The constant cost appears to degrade, however, in the $f_i = 100$, low utilization case. Form the

	Due-Date Lead Time		
ρ	0	20	100
0.5	14.91 (± 0.05)	13.88 (± 0.04)	18.55 (± 0.05)
0.7	27.89 (± 1.31)	27.01 (± 0.08)	29.00 (± 0.10)
0.9	80.41 (± 1.52)	79.70 (± 1.40)	79.36 (± 0.98)

Table XIII: Long run average cost of two identical standardized products with varying due-date lead times and traffic intensities.

simulation data itself, it appears that the idling threshold has been underestimated, implying that the shift of "$\rho_i f_i$" was too small. We suggest that the actual waiting time for the low utilization case is slightly different from our prediction where we assumed that the total utilization was close to one. This error is minimal for the low due-date case but becomes more significant as the due-date grows. Nonetheless, this striking result appears to hold.

4.5 Discussion

The descriptive form of the optimal cycle center and cycle length given in equations 4.38 and 4.53 specifies the qualitative form of the optimal dynamic cyclic policies. It is instructive to see how the the proposed policy changes over the range of problems outlined in Figure 1-1. In a slightly different order as presented in the Literature Review, we shall step through each region of the Venn diagram and state the proposed policy in terms of the variables created in the previous section. We shall summarize some of the results and provide a two product example for each case. We present the policy as a set of switching curves on the plane where the axes represent the amount of work to be done in each product class.

4.5.1 Customized products, no setups and no due-dates

The parameters from section 4.2 can be easily determined. The lack of setups forces region II to vanish and the cycle length τ to be zero. No due-dates indicates that

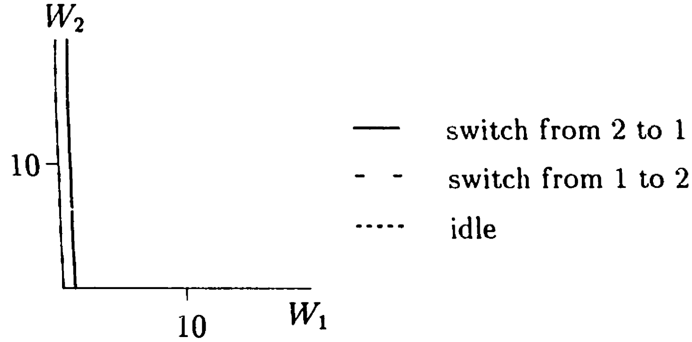

Figure 4-9: Customized products, no setups, no due-dates.

$f_i = 0$ for all product classes. This implies that only the region III conditions can apply. Thus for any total workload $w > 0$, the cycle center x_i^{c*} is set to zero for all i not equal to θ_b^*, the cheapest backorder product, which itself is set to w. In this special case, one can trivially calculate that the idling threshold w_0 is zero. The implied dynamic cyclic policy is then simple in form: service all but the least expensive product to exhaustion and switch out of producing the cheapest product if there are any other higher cost products present. The two product example given in Figure 4-9 is trivial: switch to produce product 1 whenever a product 1 order arrives, produce product 2 only when there are no product 1 orders. This causes the switching curves of the policy to nearly overlap on the product 2 axis.

This policy can be interpreted as a two level priority rule: all but the least expensive products have high priority and are serviced to exhaustion in a cyclic manner; the least expensive product has low priority. It is the dynamic cyclic version of the "$c\mu$ rule", or "b rule" in our notation. The policy minimizes the number of high cost product orders at the expense of backordering the cheapest product class. Moreover, in the heavy traffic limit the queue length of the high priority products vanish and only the lowest priority product is present. It is interesting to note that the heavy traffic limiting behavior of the proposed policy is identical to that of a strict "b" priority policy: the workload process lies along the θ_b^* axis. Thus in the heavy traffic setting nothing is lost between our proposed dynamic cyclic policy and well known $c\mu$ rule.

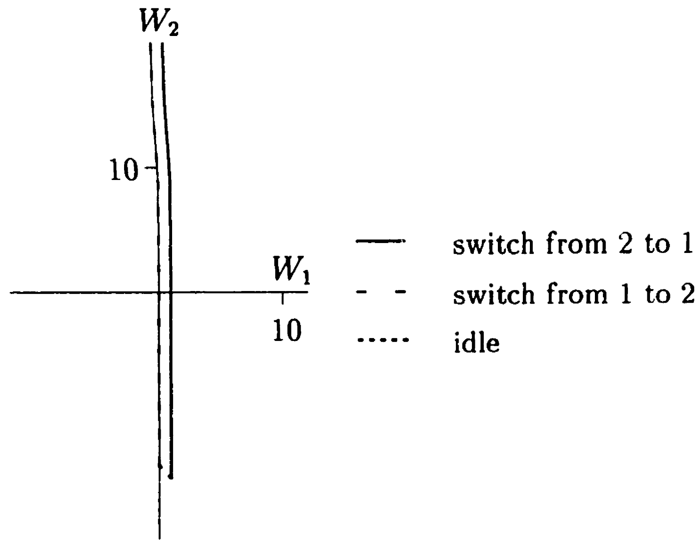

Figure 4-10: Standardized products, no setups, no due-dates.

4.5.2 Standardized products, no setups and no due-dates

From the calculations involving setup penalties and cycle length in Chapter 4.2, τ is again zero for this case and only regions I and III are possible. The cycle center of all but the cheapest product class is set to zero and $x^{c^*}_{\theta^*_b}$ is set to w for $w > 0$ and $x^{c^*}_{\theta^*_h}$ to w for $w < 0$. The idling threshold w_0 is non-trivial and negative. There is a simple interpretation of this policy. Only the cheapest holding cost product θ^*_h is stored in the finished goods inventory, all of the other classes are serviced when orders arrive. As in the previous case, the dynamic cyclic policy can be stated as a two level priority system for unfulfilled orders with the added complication of a mechanism to build-up a finished goods inventory. It has the following rules: 1) service all orders with finished goods inventory if available; 2) all backordered orders for products other than the least expensive backorder product θ^*_b have priority and are serviced in an exhaustive cyclic manner; 3) orders for the cheapest backorder product class have the lowest priority; 4) if no orders are present and the total order workload is above the idling threshold, produce the cheapest holding class product θ^*_h. An example two product policy is given in Figure 4-10 and, as in the customized case, the strict priority rule between the classes causes the switching curves to nearly overlap on the 2nd product axis.

The resulting policy is, in the heavy traffic limit, identical to the one suggested by Wein [46]. In that paper, Wein optimizes over dynamic index rules and finds that backordered products have priority in order of their relative cost per unit of work and no inventory should be held for all but the cheapest holding cost product. The heavy traffic behavior of this policy is the same as the optimal dynamic cyclic one.

Comparing the customized product and standardized product cases, we see from the proposed policy a distinct role for inventory: it hedges against the risk of backordering. According to the HTAP approximation, uncertainty in production and demand have their great impact on the total workload of the system. Finished goods inventory acts as a total workload buffer against severe backordering. Temporary backordering of high cost products is negligible compared to the costs of a high level of unfinished work. By creating a buffer of stored work, the machine can flexibly address requests for high cost products without the need of storing the products themselves and without concern for satisfying the demands for the stored product class. Thus, inventory acts as a reservoir of reserve capacity, able to absorb random fluctuations in service and demand rates. The proposed policy stores this capacity in the most economical manner possible: it places it in the cheapest holding cost product. Similarly, when the finished goods inventory is exhausted, proposed policy is still able to flexibly service requests for high cost products by neglecting the cheapest backorder product, effectively storing deficit inventory in its cheapest form.

It is also interesting to note that in several simulation runs performed in Wein [46] the heavy traffic solution can be slightly improved by a simple heuristic where small buffer stocks of expensive product are introduced. The heavy traffic approximations are not able to account for these small changes in inventory (they are absorbed in the $O(\sqrt{n})$ space scaling). We would expect the same form of improvements to hold in the policies suggested here.

4.5.3 Customized products, setups and no due-dates

As in the previous customized case the lack of due-dates and of standard goods limits the cycle center and cycle length equations to the region III formulations. Given that

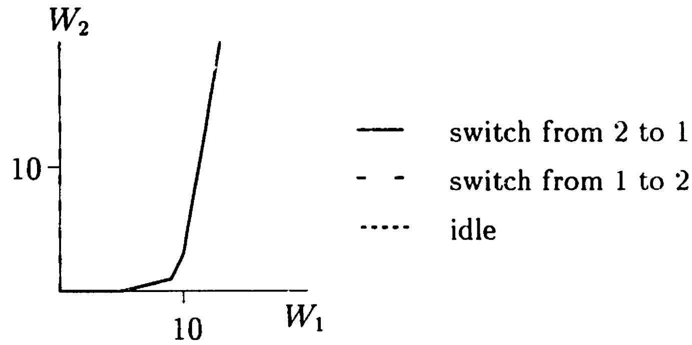

Figure 4-11: Customized products, setups, no due-dates.

f_i equals zero for all i, all but the cheapest product, θ_b^*, is binding and so $\Theta^* \subset \{\theta_b^*\}$ for all S and all w. Thus, the policy can be stated as follows: service each class in a cyclic manner. When setup for all but the cheapest backorder class, service it to exhaustion. When setup for θ_b^* work on it until its workload reaches $x_{\theta_b^*}^c(w) - \tau(w)\rho_{\theta_b^*}(1 - \rho_{\theta_b^*})/2$ where w is the current work in the system given by $\sum_{i=1}^N \mu_i^{-1} I_i$. The presence of setups eliminates the two level priority scheme seen in the previous no setup cases. This makes intuitive sense because a strict priority scheme will lead to excessive setups. The proposed policy avoids this by keeping to the cycle, yet minimizes cost by guaranteeing speedy machine resources for high cost products at the possible neglect of the θ_b^* class.

Let us be exact in specifying the tradeoffs involved: the policy cannot allocate "more" machine resources to high cost products and "less" to low cost ones as each class i will in the long run receive a ρ_i fraction of machine effort. Instead, the policy trades off on how quickly the machine can service the high cost orders. Getting to the point where the machine can service a costly item involves machine setups. Thus the essential tradeoff is between setup penalties and the degree of backordering each product class experiences. Using the insights we have gained from the HTAP, the total amount of work in the system does not change quickly and for the predicted cycle length τ we usually spend $\tau\rho_i$ units of time on each class. However, on the occasion that the total workload does change, say it increases, in the rush to the high cost products the policy neglects some of the low cost work. This can cause the gradual buildup of θ_b^* orders. Conversely, if the total workload perchance decreases, then the policy makes can make up on some of the overdue θ_b^* workload. Thus the policy can be seen as not only balancing setups and backorder costs, as directly seen in the formulation of τ and x^c, but also can be viewed as controlling the randomness of the system by isolat'ng the effects of total work fluctuation in the least cost product class. That is, the fluctuations in the inventory levels of the high cost products are relatively constrained compared to those of the low cost one θ_b^*. The degree of this isolation is limited by the size of the setup penalty.

It is interesting to note that the two product case reduces to the same policy outlined in Reiman and Wein [39]. A sample two product policy is given in Figure 4-11. The presence of setup penalties has added breadth to the switching curves of Figure 4-9. There is the same distinct feature of using the machine setup to move inventory volatility to the cheaper inventory class. This is seen by the huge range of low cost product inventory values (along the vertical axis) caused by the fluctuation of total workload in the system versus the relatively confined range of the more expensive products (along the horizontal axis).

4.5.4 Standardized products, setups and no due-dates

This case is often called the Stochastic Economic Lot Scheduling Problem (SELSP). Since all of the products are standardized, there are no orthant constraints and so no binding product classes, that is, $\Theta^* = \{1, \ldots, N\}$. All three regions are possible and the cycle length, cycle center and idling threshold are not trivial. A sample policy is given in Figure 4-12.

Our analysis reveals numerous insights into the behavior of the optimal policy in heavy traffic.

Three Workload Regions. An essential feature of the heavy traffic policy is its characterization via three workload regions. There is sufficient workload in region I, significant backorders in region III, and region II represents the intermediate case where the total workload is in an interval containing zero.

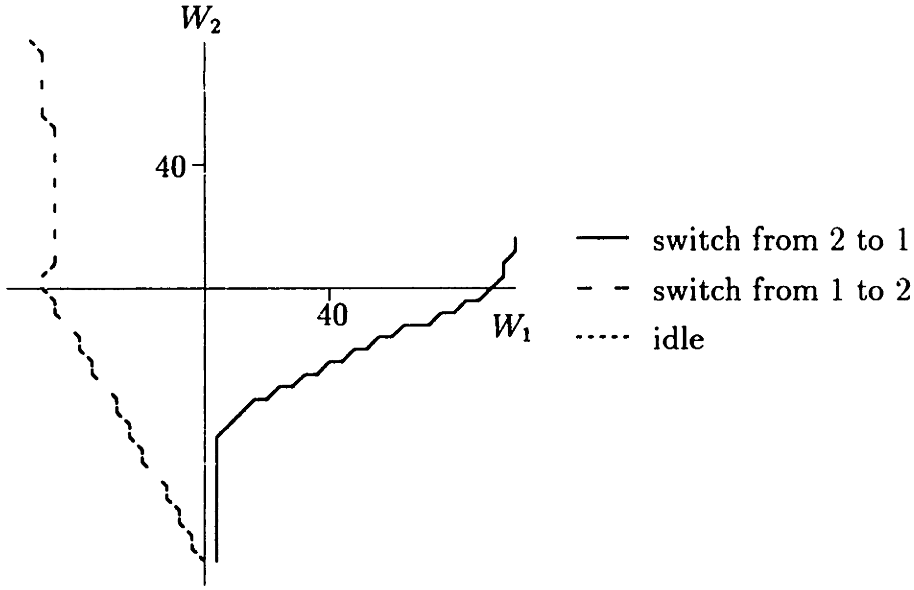

Figure 4-12: Standardized products, setups, no due-dates.

State-Dependent Lot Sizes. Because the time spent producing product i within a cycle is $\rho_i\tau$, the optimal cycle length τ determines the optimal lot sizes in heavy traffic (and determines the optimal *expected* lot sizes for the SELSP). We can observe from (4.53) that the optimal lot sizes are dependent on the potential function $V'(w)$ which is in turn a function of the state w. If there are no setups, i.e. $s = 0$, and so $V'(w)$ is trivial then we see a slightly different behavior. The lot sizes are state-dependent only when the total order workload is in region II. In contrast, the lot sizes are constant in regions I and III; in these regions, surplus or deficit inventory is unavoidable, and the trade-off between lot sizes and setup costs stabilize, thereby generating constant lot sizes. This observation and (4.38) imply that the cycle center x^c remains constant in regions I and III, and gradually shifts between these two points in the intermediate area of region II. It is worth pointing out that in nearly all of the deterministic ELSP literature (Dobson [11] is a notable exception), the analysis is restricted to policies with constant lot sizes. Yet, when there are setup times, optimal cycle length is always dynamic. Generally, as the total order workload grows, machine time is more important. Thus, it becomes optimal to increase lot-sizes so that the machine wastes less time setting up. Interestingly, when there are few orders and surplus FGI, time is cheap. Cycle lengths shorten so that the machine can quickly concentrate on high cost orders. If $s > 0$, constant lot sizes are far from optimal.

Relationship to the EOQ Model. As in the economic order quantity (EOQ) model, the lot size is proportional to the square root of the setup cost in regions I and III. In region II, the setup cost again appears in the numerator of the square root term.

Inventory is Focused in the Least Cost Products. In region I, excess inventory is built up in the product with the smallest h_i, which is a product that is inexpensive to hold (small $\hat{h}_i$) and lengthy to process (small μ_i). Similarly, in Region III, excess orders (i.e., backorders) is held in the product with the smallest backorder cost index b_i; this product is inexpensive to backorder and has a long expected processing time. In both regions, inventory is held in the least cost product so as to reduce the absolute value of the inventory of the higher (holding in region I and backorder in region III) cost products. In this regard, the dynamic lot-sizing policy derived here is similar to the standardized, no due-date case described in Chapter 4.5.2. When setup penalties are introduced, breadth is added to the normalized cycle length and, for a fixed total workload, a "corridor" of possible inventory states replaces the least cost axes.

Lot Sizes Grow with Absolute Value of Total Workload. By (4.53), we see that the optimal lot size is smallest when the total workload equals zero, and grows with the absolute value of the workload. When the total workload is near zero, costly backorders can be avoided by switching frequently between products. In contrast, when the absolute value of the workload is large, it is possible to employ large lot sizes without adversely affecting the inventory costs (because inventory tends to be held in the minimum cost product in regions I and III); in this case, it is advantageous to avoid setup penalties and produce products in large batches.

The Role of Inventory. We see from this discussion two roles for inventory. The first role is identical to that explained in Chapter 4.5.2. Finished goods inventory is used to hedge against demand and service rate uncertainty. This buffer is used on an aggregate total workload level. The second role of inventory is to avoid backordering over the course of an individual cycle. A long cycle length implies that the server

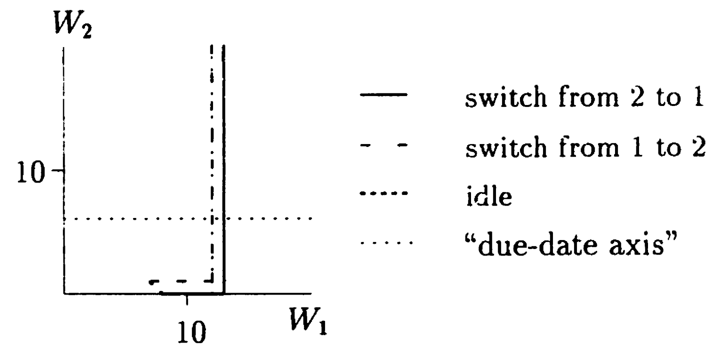

Figure 4-13: Customized products, no setups, due-dates.

will not be able to work on a product class for an extended period. The presence of finished goods absorbs demand and minimizes backordering over this wait.

4.5.5 Customized products, no setups and due-dates

As in both no setup, no due-date cases, the lack of setup penalty causes the cycle length τ and region II to vanish. The presence of due-dates, however, moves the switching curves into the orthant so that they lie on a new set of "axes." That is, by equation (4.38) the cycle centers x_i^c are shifted by $\rho_i f_i$. The proposed policy then has the following meaning: 1) if the total work in unfinished orders is less than $\sum_{i=1}^{N} \rho_i f_i$, then there is a two level priority scheme similar to the one in the standardized no due-date case; 2) orders that are almost at their due-date have priority and are serviced in a cyclic manner; 3) if there are no orders near their due-date and the total workload level is above the idling threshold, the machine works on orders for the class with the cheapest holding cost. If the total workload is greater than $\sum_{i=1}^{N} \rho_i f_i$ then again there is a two level priority rule: all product classes other than the cheapest tardiness class θ_b^* have priority and their orders are serviced in a cyclic manner just as their due-dates are reached or passed; the cheapest tardiness class θ_b^* has lowest priority and its orders are serviced only when the due-dates of the higher priority goods are distant. A two product example is given in Figure 4-13.

Again as in the no due-date, no setup examples, the switching curves for the two products nearly overlap. However, the switching curve shift onto the new due-date axes is readily transparent. The shift of the due-date axis in the workload plane represents the tolerance of the policy toward aging orders not due in the near future. The region corresponding to the "north-east" of the plane corresponds to the workload states where there is too much work and orders are completed late. The "south-west" portion of the plane is an area where there are too few orders waiting and if worked on will be completed early. The intersection of the new axes (one of them is hidden behind the switching curve) correspond to the state where the wait in queue for an order exactly equals its due-date lead-time. It is also interesting to note the "kink" in the switching curves when they come into contact with the orthant. The policy attempts to avoid backordering by staying ahead of its orders – that is completing some of them early to allow for more slack when there is an unexpected surge in demand or difficulty in production. It hedges against this uncertainty in the most economical manner possible, that is, it only completes early those products with the cheapest earliness cost. In the example case, the proposed policy attempts to store completed work in the cheapest class when possible, until no orders for that class remain, and then will further hedge against backordering by completing early the more expensive product.

Thus the proposed policy, as presented in the previous case, minimizes the inventory costs of the higher cost product classes at the expense of the cheapest product. The policy attempts to service the high cost product classes in such a manner that they are completed exactly when they are due. As in the no due-date case, excess orders are shifted to the cheapest tardiness product θ_b^*. With due-dates, however, there can now be too few orders and deficit workload is moved to the product class easiest to store. The machine can be thought of as "ahead of schedule." Moreover, customized product classes have a limit on how much "deficit" workload they can hold: the policy is forced to switch setup when a product class is exhausted of orders. This causes the "kink" on the orthant boundary in the proposed switching curves in Figure 6. In addition, given that the low total workload is costly, the idling threshold may be non-trivial so as to avoid states of high earliness costs.

It is interesting to compare these results to those of Van Mieghem. In [44], he examines a multi-class single server queuing system where the order holding costs are general convex, non-decreasing functions. This could be used to model a customized only, not setup, deterministic due-date system where the earliness costs h_i are set to zero for all of the product classes. Van Mieghem's results show that a generalized "$c\mu$" rule is optimal for this system: that is tardy orders have priority and priorities within this group are determined, in out notation, by the highest "b" rule: only when there are no late orders can early ones be filled and this can be done in any manner. There is no idling policy, the server works as long as there are orders waiting.

Our results provide the dynamic cyclic version of this policy just as was seen in the no due-date, no setups standardized and customized cases. There is a two level priority scheme for late orders with all but the cheapest tardiness cost product class θ_b^* have priority and are serviced cyclically; late θ_b^* orders are serviced only when there are no other late orders. Since all of the earliness costs are equal, the proposed policy services the orders to exhaustion when there are no late orders. The idling threshold is set at zero and so does not contribute to the policy: the machine works as long as there is work to do. Again, as in the previous no setup cases, the heavy traffic limit of the dynamic cyclic policies has the same behavior as the generalized "$c\mu$" rule. Thus our policy and Van Mieghem's are similar when they both are restricted to the one case where the models overlap.

4.5.6 Standardized products, no setups and due-dates

This problem is like the customized case without the orthant boundaries. Cycle length and region II vanish. The cycle centers are shifted onto the new "due-date" axes $\rho_i f_i$. The presence of a finished goods inventory, however, modifies the the proposed policy from the previous customized one. The policy can be interpreted as the following rules: 1) only fill orders when they are due, the order can be filled either from the finished goods inventory or directly from the machine output; 2) if the total workload present in orders minus that in finished goods inventory is above $\sum_i \rho_i f_i$ then follow a two level priority scheme: 2a) orders for high tardiness cost goods (i.e. all but θ_b^*)

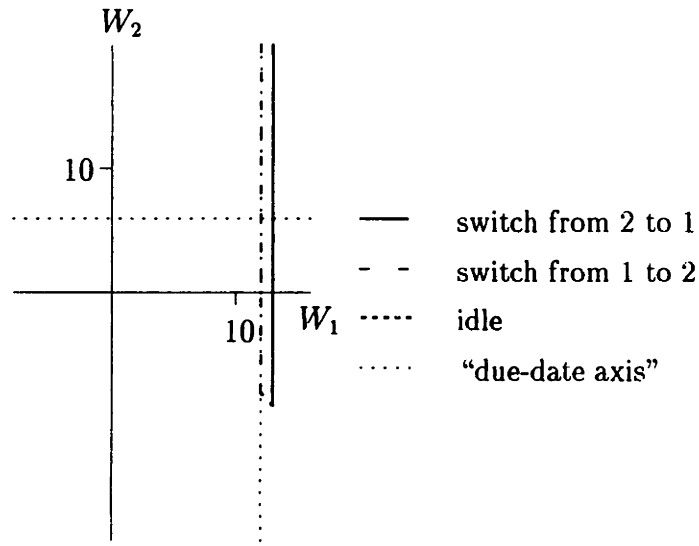

Figure 4-14: Standardized products, no setups, due-dates.

which are closer to their due-date have priority and are serviced in a cyclic manner and 2b) late θ_b^* orders have lower priority; 3) if total work is below $\sum_i \rho_i f_i$, then 3a) late orders have priority and are serviced cyclically and 3b) if there are no late orders the machine should work on the lowest holding cost product θ_h^* and store them in the finished goods inventory, up until the workload idling threshold is reached. An example two product policy is given in Figure 4-14

It is interesting to note that this policy is exactly the same as a shifted standardized, no setup, no due-date one. Moreover, in the heavy traffic limit *both have exactly the same long run average cost.* This might seem like an amazing result, but on closer inspection it makes intuitive sense. The due-dates we have considered have a special structure: they are $O(\sqrt{n})$ and so only influence the fluid limit. This implies that in our policy, orders will arrive, become late and be serviced before the total workload has an opportunity to significantly change. Moreover, the orders are continuously arriving and in this time frame the machine is not able to either get ahead or fall behind on orders. Thus, if we are servicing orders that are due today and arrived two weeks ago, tomorrow we will be servicing orders due tomorrow that arrived two weeks minus one day ago. Due-date lead times have not provided any additional flexibility to the system; orders are not serviced earlier or later than usual, only the absolute time of service has been shifted.

The difference between this case and the customized one is the ability to pre-make goods for a finished goods inventory. With standardized goods, the policy is always allowed to hedge against backordering by investing in stored work in its cheapest form. In the customized case, the inability to manufacture goods in anticipation of future orders means that the policy might be forced to store work in a more expensive class when it runs out of orders for θ_h^* products and so must finish early the next cheapest holding cost class.

It is also interesting to note the difference between our system and that of Hariharan and Zipkin [19]. In [19], they examine a facility which services requests for standardized products by re-ordering them. There are due-date lead times, L_d, on requests and lead times for the facilities orders, L_s. The facility follows a one-to-one replenishment scheme, the factor to optimize is the on-site inventory level. They find that increasing due-date lead times decreases the average cost of the system up until the due-date lead time equals the re-order lead time, when $L_d \geq L_s$ due-date lead times have no value. The reason for the difference between this result and ours is twofold: first, Hariharan and Zipkin consider an uncapacitated system, the constraining factor is re-order time; and second, the time for an order to become available is greater than the due-date lead time. Since their model has no capacity constraint, inventory in [19] plays a slightly different role: it hedges against demand uncertainty but does not act as stored workload, freeing machine resources for higher cost products. With longer due-dates there is less urgency for orders to be replenished quickly and so less inventory is needed on-site to hedge backordering risks. Yet in our model there is a capacity constraint. The same amount of stored work is needed independent of due-date lead time, as discussed above. Moreover, in their model the problem is only interesting if $L_d < L_s$, yet in our case, L_d is an order of magnitude greater than L_s by our heavy traffic scalings assumptions. By model construction, there is no conflict between these two sets of results.

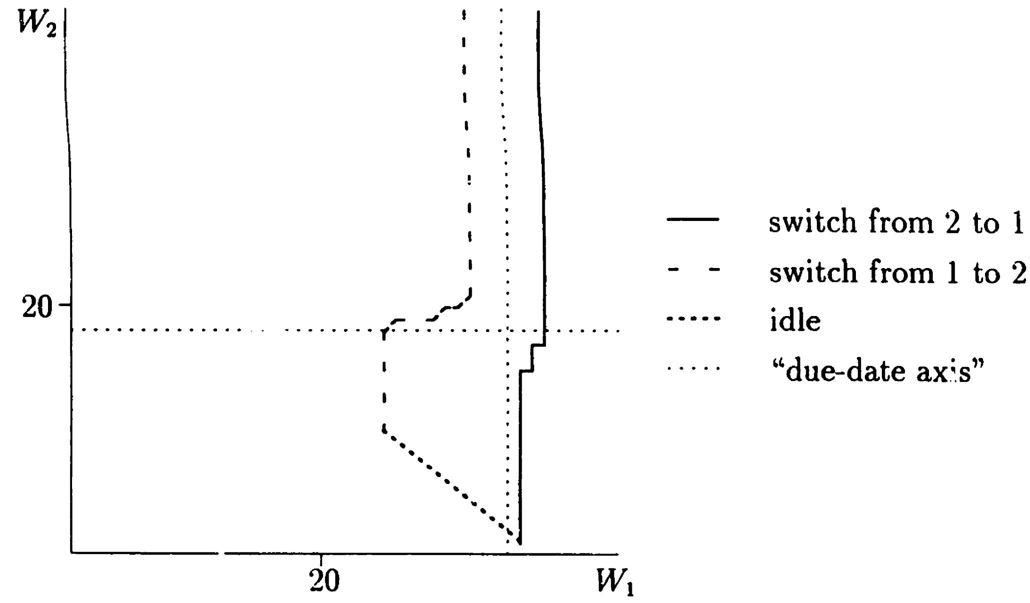

Figure 4-15: Customized products, setups, long due-dates.

4.5.7 Customized products, setups and due-dates

This case contains all of the complexity of Chapter 4.2. All three regions can be present, cycle length, cycle center and idling threshold are non-trivial and product classes can be binding. Example policies are given in Figures 4-15 and 4-16. It is interesting to note that if the due-dates are long enough, the proposed policies look like shifted SELSP switching curves (systems with setup times, however, will always be slightly different because the expanding cycle length τ will eventually hit the orthant boundary for large total workload). If the due-dates are short, the switching curves bump into the orthant and flatten out. This can be seen by comparing the long due-date time in Figure 4-15 to the shorter due-date lead time in Figure 4-16. In addition to viewing the case as a shifted SELSP policy, it can also be thought of as the customized product, no setups and due-date case with breadth added to the switching curves as was seen in the transformation between the no due-date cases without setups to the case with setups (Figures 4-9 and 4-11).

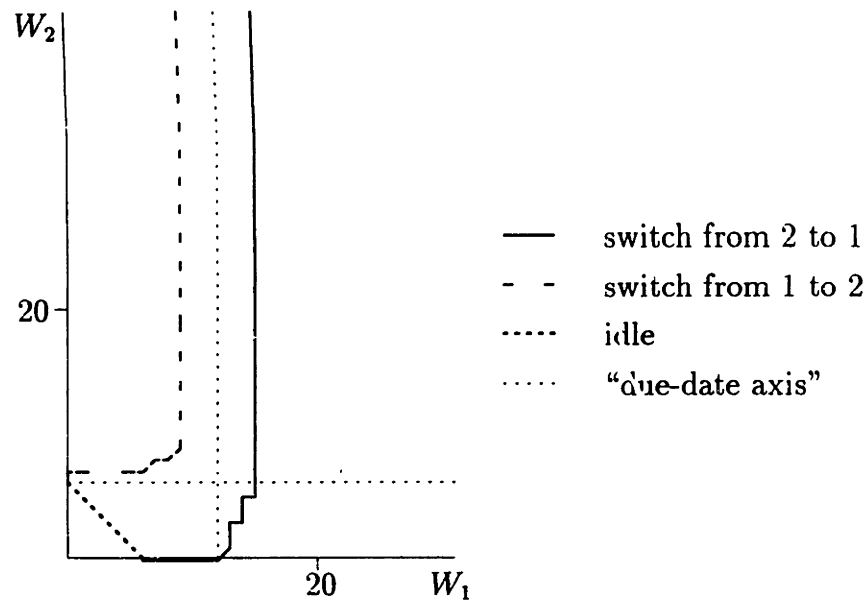

Figure 4-16: Customized products, setups, short due-dates.

4.5.8 Standardized products, setups and due-dates

The switch to standardized products from the last case simplifies the nature of the optimal dynamic cyclic policy: there are no orthant boundaries. The policy is merely the SELSP policy on the shifted set of due-date axes (see the example in Figure 4-17). It is interesting to note that as due-date lead times increase, the entire switching curves pass into the positive orthant. Standardized goods are then effectively treated as a customized set of products. Longer due-dates thus can transform a make-to-stock method of servicing demand for standardized goods into a make-to-order method. This suggests a relationship between due-date lead times and MTS/MTO partitions as briefly mentioned in Federgruen and Katalan (1994).

4.5.9 Mixed product classes

Systems with both customized and standardized goods represent the most general case considered in this paper. This mix is accounted for in our calculations by the presence of orthant constraints on some product classes and not others (see Figure 4-

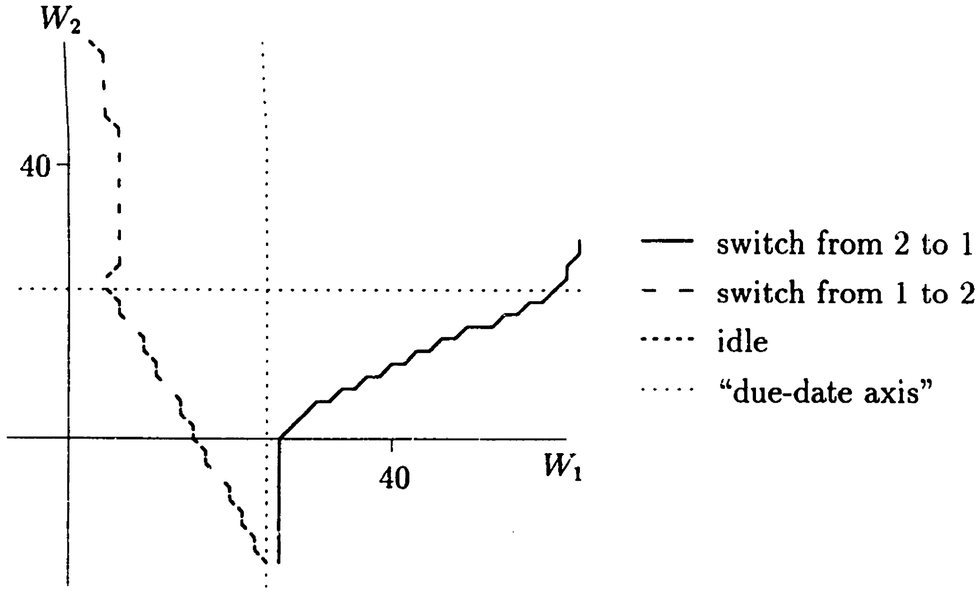

Figure 4-17: Standardized products, setups, due-dates.

18 for an example of a mixed system with one customized class, one standardized class, setup times and due-dates). This generality, however, does not complicate the system beyond the previous cases. Most of the intuitions described earlier about the interactions of setups, due-dates and orthant constraints carry over directly to the combined system.

Overall. The previous discussion has shown how our calculations have allowed us to freely move around the Venn diagram presented in Figure 1-1. Each circle adds a component to our scheduling policy. In over-simplified terms, we can label their effects on the switching curves as

$$\begin{aligned} \text{due-dates} &= \text{shifts} \\ \text{setups} &= \text{breadth} \\ \text{customized/standardized goods} &= \text{presence/absence of orthant boundaries.} \end{aligned}$$

With this three point guide we are able to qualitatively understand the nature of the deterministic due-date scheduling problem and its solution even without exactly calculating all of the solutions parameters. Although we have performed all of these calculations for the deterministic due-date case, these three effects qualitatively hold

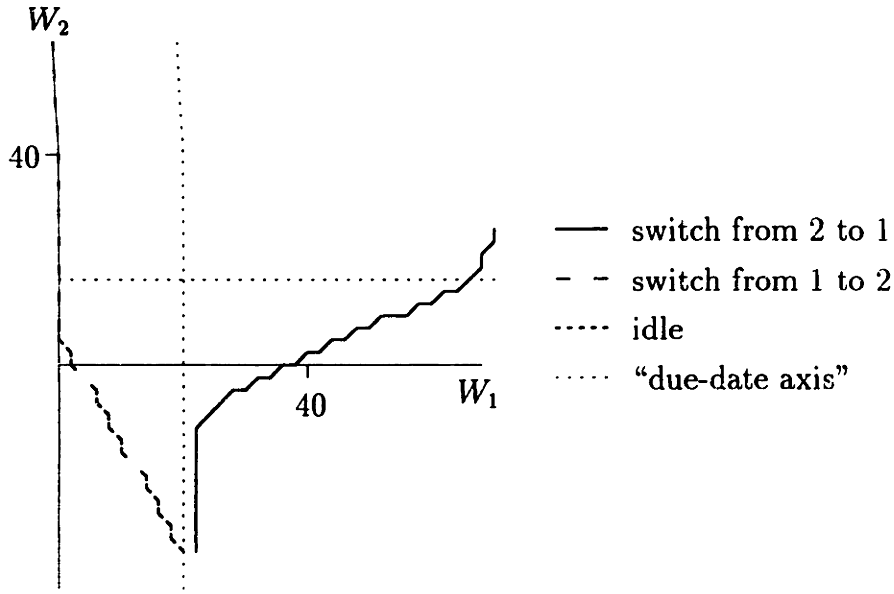

Figure 4-18: Mixed products, setups, due-dates.

for the more general case. In Appendix C, we examine the cost per cycle for a customized product with uniform due-date lead times. It would appear the same intuitions of breadth, shift and orthant boundaries are useful in scheduling this more complex problem. The insights are robust.

Chapter 5

Conclusion

Theories for scheduling a facility have examined the issues of customized/standardized product mix, setup penalties and due-dates, but have focused on each of them separately. The Heavy Traffic Averaging Principle has provided a mechanism for investigating the integrated problem. With it, we have outlined a method to numerically compute an asymptotically optimal dynamic cyclic policy. For the deterministic due-date case, we qualitatively describe the policy. The choice of dynamic cyclic policies, however, was done because of its mathematical elegance. The same heavy traffic approach could have been used for more general total workload dependent polling tables. Closed form solutions for this broader class of policies might not exist, but they should be numerically computable (although they might be intensive in nature).

Nontheless, by studying the diffusion control problem and the optimal cycle center and cycle length programs, we have not only learned about optimal dynamic cyclic policies but also about the nature of the problem. The risks inherent in the uncertainty of random demand and service rates cannot be removed. Yet, by proper scheduling, the impact of the variability can be reduced by channeling the fluctuations in inventory or order queue length into low cost areas. The presence of setups, due-dates and product mix each modify how this dampening of cost is performed. Our results give a simple interpretation of each of these facets: setups add breadth to the policy switching curves; due-dates add shifts; and customized products add orthant boundaries. These guides strengthen our understanding of the unified scheduling problem.

Appendix A

Limiting Behavior of $V'(w)$

The work done here is nearly identical to that done in Reiman and Wein [39]. Much of it is included here merely for the sake of completeness.

As referenced in Chapter 4.2.2, we wish to examine the limiting behavior of $V'(w)$. There are two claims to demonstrate: the first one is about the limit of $V'(w)$ as $w \to w$ if region II conditions hold; and the second about $V'(w)$ if region III conditions hold in the limit.

For the first part, we would like to show that

$$\lim_{w\to\infty} \frac{V'(w)}{w} = \frac{2\sqrt{\xi_3^{\Theta^*}\xi_4^{\Theta^*}} + \xi_6^{\Theta^*}}{c}, \tag{A.1}$$

which is equivalent to (4.77). Since

$$V'(w) = \lim_{\delta\to 0} \delta^{-1}[V(w+\delta) - V(w)], \tag{A.2}$$

we can express (A.1) as

$$\lim_{w\to\infty} \lim_{\delta\to 0} \frac{V(w+\delta) - V(w)}{w\delta} = \frac{2\sqrt{\xi_3^{\Theta^*}\xi_4^{\Theta^*}} + \xi_6^{\Theta^*}}{c}. \tag{A.3}$$

Thus we consider the quantity $V(w+\delta) - V(w)$. We can write

$$V(w+\delta) - V(w) = E_{w+\delta}\left[\int_0^{T_w}\left([\xi_3^{\Theta^*}\frac{\tau^*(W(t))}{W(t)} + \xi_6^{\Theta^*} + \xi_4^{\Theta^*}\frac{W(t)}{\tau^*(W(t))}]W(t) - g\right)dt\right], \tag{A.4}$$

where T_w is the first hitting time of w for the $(s/\tau^*(w) - c, \sigma^2)$ diffusion process W, and the expectation is with respect to the initial state $w+\delta$. Consider the trivial policy $\hat{\tau}(w) = Cw$ where C is a constant large enough such that the policy always satisfies the region II conditions as w grows large (this can be done since the cycle center is a linear function of w). This trivial policy is suboptimal and thus we get

$$V(w+\delta) - V(w) \le E_{w+\delta}\left[\int_0^{T_w}\left([\xi_3^{\Theta^*}C + \xi_6^{\Theta^*} + \xi_4^{\Theta^*}C^{-1}]W(t) - g\right)dt\right]. \tag{A.5}$$

Combining (A.2) and (A.5) yields

$$V'(w) \le \lim_{\delta\to 0}\frac{1}{\delta}E_{w+\delta}\left[\int_0^{T_w}\left([\xi_3^{\Theta^*}C + \xi_6^{\Theta^*} + \xi_4^{\Theta^*}C^{-1}]W(t) - g\right)dt\right]. \tag{A.6}$$

In order to show the first part about a region II limit, we need to demonstrate that both of the following are true

$$\tau^*(w) \to \infty \quad \text{as} \quad w \to \infty, \tag{A.7}$$

and

$$\frac{\tau^*(w)}{w} \to \frac{\xi_4^{\Theta^*}}{\xi_3^{\Theta^*}} \quad \text{as} \quad w \to \infty. \tag{A.8}$$

The first limit follows quickly from the region II conditions in (4.53). The second is more difficult.

Based on equation (4.53), τ^* increases with respect to w since $V'(w)$ is positive and hence S is positive. Therefore τ^* increases without bound.

We prove (A.8) by examining the region II conditions. If the region II conditions hold in the limit, then

$$\tau^*(w) = \sqrt{\frac{S + \xi_4^{\Theta^*}(\sum_{i\in\Theta^*} \rho_i \bar{f}_i - w)^2 + \xi_2^{\Theta^*}}{\xi_3^{\Theta^*}}} \tag{A.9}$$

and so for large w

$$\tau^*(w) \geq w\sqrt{\frac{\xi_4^{\Theta^*}}{\xi_3^{\Theta^*}}}. \tag{A.10}$$

Therefore the drift of $W(t)$ satisfies

$$\mu(w) = \frac{s}{\tau^*(w)} - c \to -c \quad \text{as} \quad w \to \infty\,. \tag{A.11}$$

Take w' large enough so that $\mu(w') \leq -\frac{c}{2}$. Note that $\mu(w) \leq -\frac{c}{2}$ for $w \geq w'$. Let $\tilde{W}$ denote a $\left(-\frac{c}{2}, \sigma^2\right)$ Brownian motion, and $\tilde{T}_w$ its first passage time. For $w \geq w'$, it follows that the integral in (A.6) has the bound

$$E_{w+\delta}\left[\int_0^{T_w} W(t)dt\right] \leq wE_{w+\delta}[\tilde{T}_w] + E_\delta\left[\int_0^{\tilde{T}_0} \tilde{W}(t)dt\right]\,, \tag{A.12}$$

where $\tilde{T}_0$ is the first passage time to w_0 for a $(-\frac{c}{2}, \sigma^2)$ Brownian motion.

To evaluate the last term in (A.12), let

$$h(\delta) = E_\delta\left[\int_0^{\tilde{T}_{0\wedge b}} \tilde{W}(t)dt\right], \tag{A.13}$$

where $\tilde{T}_{0\wedge b}$ denotes the first hitting time for $\tilde{W}$ to either w_0 or b. This function satisfies the ordinary differential equation (Karlin and Taylor [23])

$$-\frac{c}{2}h'(\delta) + \frac{\sigma^2}{2}h''(\delta) = -\delta, \tag{A.14}$$

subject to the boundary conditions $h(w_0) = h(b) = 0$, which yields

$$h(\delta) = \frac{2\sigma^2\delta}{c^2} + \frac{\delta^2}{c} + \frac{(2\sigma^2 b + b^2 c)(1 - e^{c\delta/\sigma^2})}{c^2(e^{cb/\sigma^2} - 1)}\,. \tag{A.15}$$

Therefore,

$$E_\delta \left[\int_0^{\tilde{T}_0} \tilde{W}(t)dt \right] = \lim_{b\to\infty} h(\delta) = \frac{2\sigma^2\delta}{c^2} + \frac{\delta^2}{c}. \tag{A.16}$$

Since

$$E_{w+\delta}[\tilde{T}_w] = \frac{2\delta}{c}\ , \tag{A.17}$$

it follows from (A.6), (A.12) and (A.16) that as $w \to \infty$,

$$\begin{aligned} V'(w) &\leq \lim_{\delta\to 0} \left(\xi_3^{\Theta^*} C + \xi_6^{\Theta^*} + \xi_4 C^{-1}\right) \left(\frac{2w\delta}{c} + \frac{2\sigma^2\delta}{c^2} + \frac{\delta^2}{c} \right) \\ &= \left(\xi_3^{\Theta^*} C + \xi_6^{\Theta^*} + \xi_4 C^{-1}\right) \left(\frac{2w}{c} + \frac{2\sigma^2}{c^2} \right). \end{aligned} \tag{A.18}$$

Since $V'(w)/w^2 \to 0$ as $w \to \infty$, by (4.53) we have $\tau^*(w)/w \to \frac{\xi_4^{\Theta^*}}{\xi_3^{\Theta^*}}$ as $w \to \infty$, which is the claim in (A.8). We are almost ready to show (A.1). Fix w and let $\tilde{W}^{(w)}(t)$ denote a Brownian motion with fixed drift $\mu(w) = s/\tau^*(w) - c$, and fixed variance σ^2. Let $\tilde{T}^{(w)}$ denote the first passage times for this process. The unbounded nature of $\tau^*(w)$ implies that

$$\begin{aligned} & wE_{w+\delta}[\tilde{T}_w^{(\infty)}] + E_\delta \left[\int_0^{\tilde{T}_0^{(\infty)}} \tilde{W}^{(\infty)}(t)dt \right] \\ & \quad \leq E_{w+\delta} \left[\int_0^{T_x} W(t)dt \right] \leq wE_{w+\delta}[\tilde{T}_w^{(w)}] + E_\delta \left[\int_0^{\tilde{T}_0^{(w)}} \tilde{W}^{(w)}(t)dt \right]\ , \end{aligned} \tag{A.19}$$

where $\tilde{W}^{(\infty)}$ is a Brownian motion with drift $-c$ and variance σ^2. By the same analysis as done in equations (A.12) to (A.18), we get

$$\frac{w\delta}{c} + \frac{\sigma^2\delta}{2c^2} + \frac{\delta^2}{2c} \leq E_{w+\delta} \left[\int_0^{T_w} W(t)dt \right] \leq \frac{w\delta}{\mu(w)} + \frac{\sigma^2\delta}{2\mu^2(w)} + \frac{\delta^2}{2\mu(w)}. \tag{A.20}$$

Since we have

$$\lim_{w\to\infty} \left[\xi_3^{\Theta^*} \frac{\tau^*(W(t))}{W(t)} + \xi_6^{\Theta^*} + \xi_4^{\Theta^*} \frac{W(t)}{\tau^*(W(t))} \right] = 2\sqrt{\xi_3^{\Theta^*}\xi_4^{\Theta^*}} + \xi_6^{\Theta^*}, \tag{A.21}$$

then

$$\lim_{w\to\infty} \lim_{\delta\to 0} \frac{V(w+\delta) - V(w)}{w\delta} = \frac{2\sqrt{\xi_3^{\Theta^*}\xi_4^{\Theta^*}} + \xi_6^{\Theta^*}}{c}. \tag{A.22}$$

This is what we wished to show for the first claim.

For the second claim we will repeat several of the same procedures used in the first part, but since the structure of $V'(w)$ is fundamentally different in the limit, the proof will be more involved and follows the spirit of Reiman and Wein's proof in [39] even more closely. We would like to show that

$$\lim_{w\to\infty} \frac{V'(w)}{w} = \frac{b_{\theta^*}}{c}, \tag{A.23}$$

which is equivalent to (4.77). As before, since

$$V'(w) = \lim_{\delta\to 0} \delta^{-1}[V(w+\delta) - V(w)], \tag{A.24}$$

we can express (A.23) as

$$\lim_{w\to\infty} \lim_{\delta\to 0} \frac{V(w+\delta) - V(w)}{w\delta} = \frac{b_{\theta^*}}{c}. \tag{A.25}$$

Thus we consider the quantity $V(w+\delta) - V(w)$. We write

$$\begin{aligned} V(w+\delta) - V(w) &= E_{w+\delta}\left[\int_0^{T_w} \left([\xi_7^{\Theta^*} \frac{\tau^*(W(t))}{W(t)} - \xi_8^{\Theta^*} \frac{1}{W(t)} + \right.\right. \\ &\qquad \left.\left. \frac{\xi_9^{\Theta^*}}{W(t)\tau^*(W(t))} + b_{\theta^*}]W(t) - g\right) dt\right], \end{aligned} \tag{A.26}$$

where

$$\begin{aligned} \xi_7^{\Theta^*} &= \sum_{i\in\{\Theta^*\backslash\theta^*\}} \left(\frac{\rho_i(1-\rho_i)}{2(b_i+h_i)}\left[\frac{b_i-h_i}{2} - b_{\theta^*}\right]\left[\frac{h_i-b_i}{2} - b_{\theta^*}\right]\right. && \text{(A.27)} \\ &\qquad \left. + \frac{\rho_i(1-\rho_i)(b_i+h_i)}{8}\right) - b_{\theta^*} \sum_{j\notin\Theta^*} \frac{\rho_j(1-\rho_j)}{2} && \text{(A.28)} \\ &\qquad + b_{\theta^*} \sum_{i\in\{\Theta^*\backslash\theta^*\}} \frac{\rho_i(1-\rho_i)}{b_i+h_i}\left[\frac{b_i-h_i}{2} - b_{\theta^*}\right] && \text{(A.29)} \\ &\qquad + \sum_{j\notin\Theta^*} b_j \frac{\rho_j(1-\rho_j)}{2} \\ \xi_8^{\Theta^*} &= \sum_{j\notin\Theta^*} b_j\rho_j\bar{f}_j + b_{\theta^*} \sum_{i\in\Theta^*} \rho_i\bar{f}_i && \text{(A.30)} \end{aligned}$$

$$\xi_9^{\Theta^*} = \sum_{j \notin \Theta^*} \frac{b_i + h_i}{2\rho_i(1-\rho_i)} \left(\rho_i \bar{f}_i\right)^2 \tag{A.31}$$

We will also make a critical assumption: that the cycle length $\tau^*(w)$ is monotonically increasing in w for region III.

As in the first part, we show two claims as an intermediary step in the proof. We first show

$$\frac{\tau^*(w)}{w} \to 0 \quad \text{as} \quad w \to \infty, \tag{A.32}$$

and

$$\tau^*(w) \to \infty \quad \text{as} \quad w \to \infty. \tag{A.33}$$

Unlike the first part, equation (A.33) is not trivial. and so instead we start with equation (A.32) and show it by contradiction. Assume that $\overline{\lim}_{w\to\infty}\tau^*(w)/w$ does not vanish. Using the same trivial policy constructed for equation (A.5) we get

$$V'(w) \le \overline{\lim}_{\delta\to 0}\frac{1}{\delta}E_{w+\delta}\left[\int_0^{T_w}\left([\xi_7^{\Theta^*}C + \frac{\xi_9^{\Theta^*}}{CW(t)} + \xi_8^{\Theta^*}C^{-1} + b_{\theta^*}]W(t) - g\right)dt\right]. \tag{A.34}$$

Assuming the monotonicity of τ^* and equation (A.33), the drift of the diffusion process goes to $-c$ as before. Thus, take w' large enough so that $\mu(w') \le -\frac{c}{2}$. Note that $\mu(w) \le -\frac{c}{2}$ for $w \ge w'$. Let $\tilde{W}$ denote a $\left(-\frac{c}{2}, \sigma^2\right)$ Brownian motion, and $\tilde{T}_w$ its first passage time. Using the same steps as done in equations (A.12) to (A.17) we get

$$\begin{aligned} V'(w) &\le \lim_{\delta\to 0}\left(\xi_7^{\Theta^*}C + \xi_8^{\Theta^*}C^{-1} + b_{\theta^*}\right)\left(\frac{2w\delta}{c} + \frac{2\sigma^2\delta}{c^2} + \frac{\delta^2}{c}\right) \\ &= \left(\xi_7^{\Theta^*}C + \xi_8^{\Theta^*}C^{-1} + b_{\theta^*}\right)\left(\frac{2w}{c} + \frac{2\sigma^2}{c^2}\right). \end{aligned} \tag{A.35}$$

Thus, by being in region III, equation (4.53) implies that $\tau^*(w)/w \to 0$, a contradiction to the assumption and so proving (A.32).

We show the second claim again by contradiction. Since we have assumed the monotonicity of τ^*, if equation (A.33) does not hold the cycle length must converge to a constant, denoted by $\tau^*(\infty)$. For large w, $W(t)$ behaves as a (μ, σ^2) Brownian motion, where $\mu = s/\tau^*(\infty) - c$ could be of either sign. From the fact that $s/\tau^*(\infty) \le s/\tau^*(w)$

we obtain

$$V'(w) \geq \lim_{\delta\to 0} \frac{b_{\theta^*}}{\delta} E_{w+\delta}\left[\int_0^{T_w} W(t)dt\right] - \lim_{\delta\to 0} \frac{g}{\delta} E_{w+\delta}[T_w] \tag{A.36}$$

$$\geq \lim_{\delta\to 0} \frac{b_{\theta^*} w - g}{\delta} E_\delta[T_0], \tag{A.37}$$

where T_0 is the first hitting time for a Brownian motion with drift μ and variance σ^2. If $\mu \geq 0$, then $E_\delta[T_0] = \infty$, and if $\mu < 0$, then $E_\delta[T_0] = -\delta/\mu$. Hence,

$$\lim_{w\to\infty} V'(w) \geq \lim_{w\to\infty} \frac{b_{\theta^*} w - g}{|\mu|} = \infty. \tag{A.38}$$

Equations (A.38) and (4.53) gives the result we wanted to show by contradiction, that $\tau^*(w) \to \infty$.

With both of these claims, we follow steps (A.19) to (A.22) from the first claim and derive the desired result that

$$\lim_{w\to\infty} \lim_{\delta\to 0} \frac{V(w+\delta) - V(w)}{w\delta} = \frac{b_{\theta^*}}{c}. \tag{A.39}$$

The structure of Θ^* follows directly from the eight observations in Chapter 4 and the fact that τ is expanding.

Appendix B

An Approximation for $V'(w)$.

In this section, we try to approximate the potential function $V'(w)$ by a closed form expression. We are interested characterizing the form of $V'(w)$ implied by equations (3.21) and (3.22) so that we can have a clearer idea about the behavior of the proposed dynamic cyclic policies. In order to perform some of our calculations, we only examine a limited case: a system with only standardized products and only setup times. Because only standardized products are considered, the cost per cycle is easy to calculate and is given by equation (3.17). In addition, there are no orthant constraints as there are no customized products and thus the set of non-binding product classes Θ^* is equal to $\{1, \ldots, N\}$. The limitation to only setup time penalties is this discussion, yet does not overly degrade its pertinence as the potential function is almost entirely dependent on the setup time penalty and not the setup cost one. Since the cost per cycle is only dependent on the first moment of the due-date lead time distribution, the expression for the optimal cycle center x^{c^*} and cycle length τ^*, given in equations (4.38) and (4.53) respectively, hold. Thus we can rewrite equation (3.21) as

$$\begin{aligned}
2\sqrt{\xi_1^{\Theta^*} sV'(w)} + h_{\theta^*} w_{\Theta^*} - g - cV'(w) + \frac{\sigma^2}{2} V''(w) &= 0 \quad \text{region I,} \qquad \text{(B.1)} \\
2\sqrt{\xi_3^{\Theta^*} \xi_4^{\Theta^*} w_{\Theta^*}^2 + \xi_3^{\Theta^*} sV'(w)} + \xi_7^{\Theta^*} w_{\Theta^*} & \\
-cV'(w) + \frac{\sigma^2}{2} V''(w) - g &= 0 \quad \text{region II,} \qquad \text{(B.2)}
\end{aligned}$$

$$2\sqrt{\xi_5^{\Theta^*} sV'(w)} - b_{\theta^*} w_{\Theta^*} - g - cV'(w) + \frac{\sigma^2}{2} V''(w) = 0 \quad \text{region III}, \quad \text{(B.3)}$$

where w_{Θ^*} is $\sum_{i\in\Theta^*} \rho_i \bar{l}_i - w$.

These ODEs have no closed form solution. We relabel our $V'(w)$ approximation $p(w)$ for ease of notation. We construct $p(w)$ in parts based on the natural region breakup of the cost function, cycle center and cycle length. In the numerical experiments, we have found that the idling threshold is critical in determining the performance of the policy and is easily underestimated. Thus we start with $p(w)$'s asymptotic limit derived in Property 3 in Chapter 4 and by using Taylor Series approximations to equation (B.1) to (B.3) attempt to push w_0 as low as the local behavior about $p(w_0)$ will allow, as one would squeeze toothpaste to the front a tube with a fixed cap.

We implement two methods depending on if the inventory costs are identical. If they are asymmetric, we approximate $p(w)$ in parts: $p_4(w)$ will represent $p(w)$ for w in region III, $p_3(w)$ for w in region II but w greater than $\sum_{i\in\Theta^*} \rho_i \bar{l}_i$, $p_2(w)$ for w in region II and w less than $\sum_{i\in\Theta^*} \rho_i \bar{l}_i$, and $p_1(w)$ for w in region I. The tail $p_4(w)$ we shall assume is the asymptotic limit. $p_3(w)$ shall solve a 1st order Taylor series approximation to the ODE about the region III – region II border, $p_2(w)$ solve one about $\sum_{i\in\Theta^*} \rho_i \bar{l}_i$, and $p_1(w)$ around w_0. Let us label the workload level at the region I – region II border as w_1 and the workload level at the region II – region III border as w_2. The three Taylor series approximations are then as follows

$$h_{\Theta^*}(w - w_0) - cp_1(w) + \frac{\sigma^2}{2} p_1'(w) = 0, \quad \text{(B.4)}$$

$$2\sqrt{\xi_3^{\Theta^*} sp(\sum_{i\in\Theta^*} \rho_i \bar{l}_i)} + \xi_7^{\Theta^*} w_{\Theta^*} - h_{\Theta^*}(\sum_{i\in\Theta^*} \rho_i \bar{l}_i - w_0)$$
$$\frac{2\xi_3^{\Theta^*} sp_3'(\sum_{i\in\Theta^*} \rho_i \bar{l}_i)}{\sqrt{\xi_3^{\Theta^*} sp_3(\sum_{i\in\Theta^*} \rho_i \bar{l}_i)}} w_{\Theta^*} - cp_2(w) + \frac{\sigma^2}{2} p_2'(w) = 0, \quad \text{(B.5)}$$

$$2\sqrt{\xi_3^{\Theta^*} \xi_4^{\Theta^*} (\sum_{i\in\Theta^*} \rho_i \bar{l}_i - w_2)^2 + \xi_3^{\Theta^*} sp(w_2)} + \xi_7^{\Theta^*} w_{\Theta^*}$$
$$+ \frac{2\xi_3^{\Theta^*} \xi_4^{\Theta^*} (\sum_{i\in\Theta^*} \rho_i \bar{l}_i - w_2) + \xi_3^{\Theta^*} sp_4'(w_2)}{\sqrt{\xi_3^{\Theta^*} \xi_4^{\Theta^*} (\sum_{i\in\Theta^*} \rho_i \bar{l}_i - w_2)^2 + \xi_3^{\Theta^*} sp_4(w_2)}} (w_2 - w)$$

$$-h_{\Theta^\bullet}(\sum_{i\in\Theta^\bullet}\rho_i\bar{l}_i - w_0) - cp_3(w) + \frac{\sigma^2}{2}p_3'(w) = 0. \qquad \text{(B.6)}$$

The proposed solutions are

$$p_1(w) = \frac{h_{\Theta^\bullet}}{c}(w-w_0) - \frac{h_{\Theta^\bullet}\sigma^2}{2c^2} + \frac{h_{\Theta^\bullet}\sigma^2}{2c^2}e^{\frac{2c}{\sigma^2}(w-w_0)}, \qquad \text{(B.7)}$$

$$p_2(w) = \phi_1 e^{2cw/\sigma^2} + \phi_2 - \frac{\xi_7^{\Theta^\bullet}}{c}w + \frac{\xi_3^{\Theta^\bullet} sp_3'(\sum_{i\in\Theta^\bullet}\rho_i\bar{l}_i)}{c\sqrt{\xi_3^{\Theta^\bullet} sp_3(\sum_{i\in\Theta^\bullet}\rho_i\bar{l}_i)}}w, \qquad \text{(B.8)}$$

$$p_3(w) = \phi_3 e^{2cw/\sigma^2} + \phi_4 - \frac{\xi_7^{\Theta^\bullet}}{c}w + \frac{2\xi_3^{\Theta^\bullet}\xi_4^{\Theta^\bullet}(\sum_{i\in\Theta^\bullet}\rho_i\bar{l}_i - w_2) + \xi_3^{\Theta^\bullet} sp_4'(w_2)}{c\sqrt{\xi_3^{\Theta^\bullet}\xi_4^{\Theta^\bullet}w_2^2 + \xi_3^{\Theta^\bullet} sp_4(w_2)}}w, \qquad \text{(B.9)}$$

$$p_4(w) = \frac{b_{\Theta^\bullet}}{c}w + \phi_5, \qquad \text{(B.10)}$$

where ϕ_1, ϕ_2, ϕ_3, ϕ_4, ϕ_5, w_0, w_1 and w_2 are unknown. By the Principle of Smooth Fit, $p'(w)$ is continuous and thus we have the following smoothness constraints on $p(w)$: $p_1(w_1) = p_2(w_1)$, $p_1'(w_1) = p_2'(w_1)$, $p_2(\sum_{i\in\Theta^\bullet}\rho_i\bar{l}_i) = p_3(\sum_{i\in\Theta^\bullet}\rho_i\bar{l}_i)$, $p_2'(\sum_{i\in\Theta^\bullet}\rho_i\bar{l}_i) = p_3'(\sum_{i\in\Theta^\bullet}\rho_i\bar{l}_i)$, $p_3(w_2) = p_4(w_2)$, $p_3'(w_2) = p_4'(w_2)$. We can additionally impose continuity constraints on $\tau(w)$. Setting $\tau(w_1^+)$ equal to $\tau(w_1^-)$ and $\tau(w_2^+)$ equal to $\tau(w_2^-)$ and solving for $p(w)$ we get

$$p_1(w_1) = \frac{\xi_4^{\Theta^\bullet}\xi_1^{\Theta^\bullet}w_1^2}{s(\xi_1^{\Theta^\bullet} - \xi_3^{\Theta^\bullet})}, \qquad \text{(B.11)}$$

$$p_3(w_2) = \frac{\xi_4^{\Theta^\bullet}\xi_5^{\Theta^\bullet}w_2^2}{s(\xi_5^{\Theta^\bullet} - \xi_3^{\Theta^\bullet})}. \qquad \text{(B.12)}$$

This makes eight equations for eight unknowns. Unfortunately, we have not been able to find a closed form solution to these unknowns and must resort to numerical methods. However, the expressions in (B.7) to (B.10) do show the general shape of the potential function: it starts off growing at an exponential rate and then settles into a linear growth rate.

If the inventory costs are symmetric, we again approximate $p(w)$ in parts: $p_1(w)$ represents $p(w)$ for $w > 0$ and $p_2(w)$ approximates $p(w)$ for $w < 0$. We shall use the same Taylor series approximation of the ODE around w_0 for $p_1(w)$. For $p_2(w)$, we shall proceed with a two step approximation. Noting that for identical costs $\xi_4^{\Theta^*} = (b+h)\frac{\sum_{i\in\Theta^*}\rho_i(1-\rho_i)}{8}$ and $\xi_3^{\Theta^*}\xi_4^{\Theta^*} = \frac{(b+h)^2}{16}$, we can first use the asymptotic limit to simplify the square root in the ODE by

$$\frac{h+b}{4}\sqrt{w_{\Theta^*}^2 - \frac{2sp(w)\sum_{i\in\Theta^*}\rho_i(1-\rho_i)}{h+b}} \approx \frac{h+b}{4}\sqrt{w_{\Theta^*}^2 - \frac{2sb\sum_{i\in\Theta^*}\rho_i(1-\rho_i)}{(h+b)c}w}. \tag{B.13}$$

We can then use a Taylor series approximation about $\sum_{i\in\Theta^*}\rho_i\bar{l}_i$ and get for $w > \sum_{i\in\Theta^*}\rho_i\bar{l}_i$

$$\sqrt{w_{\Theta^*}^2 - \frac{2sb\sum_{i=1}^N\rho_i(1-\rho_i)}{(h+b)c}w} \approx w_{\Theta^*} + \frac{sb\sum_{i\in\Theta^*}\rho_i(1-\rho_i)}{(h+b)c}. \tag{B.14}$$

Using these simplifications and noting that $\xi_7^{\Theta^*} = \frac{h-b}{2}$, we arrive at the following two ODE approximations for equation (B.2)

$$h_{\Theta^*}(w-w_0) + cp_1(w) + \frac{\sigma^2}{2}p_1'(w) = 0, \tag{B.15}$$

$$-bw_{\Theta^*} - hw_0 + \frac{sb\sum_{i=1}^N\rho_i(1-\rho_i)}{2c} + cp_2(w) + \frac{\sigma^2}{2}p_2'(w) = 0. \tag{B.16}$$

Unlike the asymmetric cost problem, we can find a closed form expression to both $p_1(w)$ and $p_2(w)$ by proposing a more general solution to the ODEs and using the Principle of Smooth Fit to find the exact answer. We propose the following generalized forms for the potential function

$$p_1(w) = \frac{h}{c}(w-w_0) + \phi_1 e^{-\phi_2(\sum_{i\in\Theta^*}\rho_i\bar{l}_i - w)} + \phi_3, \tag{B.17}$$

$$p_2(w) = \frac{h}{c}(\sum_{i\in\Theta^*}\rho_i\bar{l}_i - w_0) + \frac{b}{c}w_{\Theta^*} - \frac{\sigma^2 b}{2c^2} - \frac{sb\sum_{i\in\Theta^*}\rho_i(1-\rho_i)}{2c} + \phi_4 e^{-\frac{2c}{\sigma^2}w} \tag{B.18}$$

where ϕ_1, ϕ_2, ϕ_3, ϕ_4 and w_0 are constants which must be computed. Given the asymptotic result for $p_2(w)$, the constant ϕ_4 must be zero. Using the smoothness constraints $p_1(w_0) = 0$, $p_1'(w_0) = 0$, $p_1((\sum_{i\in\Theta^*}\rho_i\bar{l}_i)^+) = p_2((\sum_{i\in\Theta^*}\rho_i\bar{l}_i)^-)$, $p_1'((\sum_{i\in\Theta^*}\rho_i\bar{l}_i)^+) =$

$p_2'((\sum_{i\in\Theta^*}\rho_i\bar{l}_i)^-)$, we can solve for the remaining unknowns. They are

$$\begin{aligned}
w_0 &= \sum_{i\in\Theta^*}\rho_i\bar{l}_i + \frac{\ln(\frac{h}{b+h})\sigma^2 + \ln(\frac{h}{b+h})s\sum_{i\in\Theta^*}\rho_i(1-\rho_i)}{2c}, \\
\phi_1 &= -\frac{h+b}{2c^2}[\sigma^2 + s\sum_{i\in\Theta^*}\rho_i(1-\rho_i)], \\
\phi_2 &= \frac{2c}{\sigma^2 + s\sum_{i\in\Theta^*}\rho_i(1-\rho_i)}, \\
\phi_3 &= \frac{h\sigma^2 + hs\sum_{i\in\Theta^*}\rho_i(1-\rho_i))}{2c^2}.
\end{aligned}$$

With these approximations for $p(w)$ we could find both the cycle time τ and the cycle center x^c of the proposed optimal cyclic policy for the heavy traffic limit. Yet these equations for $p(w)$ are more interesting for their statement about the potential function: that is, $p(w)$ again initially grows in a highly non-linear fashion about the idling threshold and then stabilizes to a linear limit. The algorithm given in Chapter 3 is more accurate in determining the actual potential function $V'(w)$ and so was used in the numerical experiments given in Chapter 4.

Appendix C

Uniform Due-dates

In this section, we derive the cost per cycle for a customized product class with a uniform due-date lead time distribution. Let $f_i(s)$ be uniform between $\tilde{a}_i$ and $\tilde{b}_i$. The fluid limit of the distribution then is

$$\bar{f}_i(s) = \begin{cases} \frac{1}{\bar{b}_i - \bar{a}_i} & \text{if } s \in (\bar{a}_i, \bar{b}_i) \\ 0 & \text{otherwise} \end{cases}, \tag{C.1}$$

and the complement of the cumulative distribution is

$$\bar{F}_i^c(s) = \begin{cases} 1 & \text{if } s \leq \bar{a}_i \\ \frac{1}{\bar{b}_i - \bar{a}_i}(\bar{b}_i - s) & \text{if } s \in (\bar{a}_i, \bar{b}_i) \\ 0 & \text{if } s > \bar{b}_i \end{cases}. \tag{C.2}$$

For a given cycle center and cycle length, we have the product i work at the cycle start x_i^s. Proposition 3.1 implies that $\bar{x}_i^s = \int_{\bar{L}_i(0)}^{\infty} \rho_i \bar{F}_i^c(s) ds$ and so $\bar{L}_i(0)$ is defined by

$$\bar{L}_i(0) = \begin{cases} \bar{b}_i - \sqrt{2(\bar{b}_i - \bar{a}_i)\frac{x_i^s}{\rho_i}} & \text{for } x_i^s \leq \frac{\bar{b}_i - \bar{a}_i}{2}\rho_i \\ \frac{\bar{b}_i + \bar{a}_i}{2} - \frac{x_i^s}{\rho_i} & \text{for } x_i^s \geq \frac{\bar{b}_i - \bar{a}_i}{2}\rho_i \end{cases}. \tag{C.3}$$

We can find $\bar{G}_i(s,t)$ from Proposition 3.2. The function breaks down into cases depending on if the steady-state distribution (the "storm front" from before) has reached the minimum order due-date $\bar{a}_i$. Thus if $\bar{L}_i(0) - t > \bar{a}_i$ then

$$\bar{G}_i(s,t) = \begin{cases} 0 & \text{for } s \le \bar{a}_i - t \\ \frac{\rho_i}{\bar{b}_i - \bar{a}_i}(s - (\bar{a}_i - t)) & \text{for } \bar{a}_i - t \le s \le \bar{a}_i \\ \frac{t\rho_i}{\bar{b}_i - \bar{a}_i} & \text{for } \bar{a}_i \le s \le \bar{L}_i(0) - t \\ \frac{\rho_i}{\bar{b}_i - \bar{a}_i}(\bar{b}_i - s) & \text{for } \bar{L}_i(0) - t \le s \le \bar{b}_i \\ 0 & \text{for } s > \bar{b}_i \end{cases} . \tag{C.4}$$

If $\bar{L}_i(0) - t \le \bar{a}_i$ then

$$\bar{G}_i(s,t) = \begin{cases} 0 & \text{for } s \le \bar{a}_i - t \\ \frac{\rho_i}{\bar{b}_i - \bar{a}_i}(s - (\bar{a}_i - t)) & \text{for } \bar{a}_i - t \le s \le \bar{L}_i(0) - t \\ \rho_i & \text{for } \bar{L}_i(0) - t \le s \le \bar{a}_i \\ \frac{\rho_i}{\bar{b}_i - \bar{a}_i}(\bar{b}_i - s) & \text{for } \bar{a}_i \le s \le \bar{b}_i \\ 0 & \text{for } s > \bar{b}_i \end{cases} . \tag{C.5}$$

The non-trivial values of $\bar{G}_i(s,t)$ can be broken-down into three distinct parts. They are separated by the two partitions $\bar{a}_i$ and $\bar{L}_i(0) - t$. The earliest due-date function $\bar{L}_i(t)$ has three forms depending on which part of $\bar{G}_i(s,t)$ the server has competed work up to. We can find when these epochs occur and then characterize $\bar{L}_i(t)$. Let t_{i1} be the time when $\bar{L}_i(t)$ reaches the first partition and t_{i2} be the time of the second partition. A singe prime (i.e. τ'_{i1}) will mark the first case where $\bar{a}_i$ is the first partition, a double prime for the second where $\bar{L}_i(0) - t$ is the first partition. We shall use similar conventions for $\bar{L}_i(0)$.

Let's first examine the $\bar{a}_i < \bar{L}_i(0) - t$ case. The time until $\bar{a}_i$ is reached must satisfy

$$t'_{i1} = \int_{\bar{a}_i - t'_{i1}}^{\bar{a}_i} \frac{\rho_i}{\bar{b}_i - \bar{a}_i}(s - (\bar{a}_i - t'_{i1}))ds. \tag{C.6}$$

Solving for t'_{i1} we get

$$t'_{i1} = \frac{\bar{b}_i - \bar{a}_i}{\rho_i}\left(1 - \sqrt{1 - \frac{2\tau\rho_i(1 - \rho_i)}{\bar{b}_i - \bar{a}_i}}\right). \tag{C.7}$$

With this, we have for $t \in (\tau(1-\rho_i), t'_{i1})$

$$\bar{L}'_{i1}(t) = \bar{a}_i - t + \sqrt{\frac{2(\bar{b}_i - \bar{a}_i)(t - \tau(1-\rho_i))}{\rho_i}} \tag{C.8}$$

Similarly t'_{i2} must satisfy

$$t'_{i2} = \int_{\bar{a}_i - t'_{i2}}^{\bar{a}_i} \frac{\rho_i}{\bar{b}_i - \bar{a}_i}(s - (\bar{a}_i - t'_{i2}))ds + \int_{\bar{a}_i}^{\bar{L}_i(0) - t'_{i2}} \frac{s\rho_i}{\bar{b}_i - \bar{a}_i} ds. \tag{C.9}$$

Thus we have

$$t'_{i2} = \bar{L}_i(0) - \bar{a}_i - \frac{\bar{b}_i - \bar{a}_i}{\rho_i} + \sqrt{(\bar{L}_i(0) - \bar{a}_i - \frac{\bar{b}_i - \bar{a}_i}{\rho_i})^2 + 2\frac{\bar{b}_i - \bar{a}_i}{\rho_i}\tau(1-\rho_i)} \tag{C.10}$$

and so for $t \in (t'_{i1}, t'_{i2})$

$$\bar{L}'_{i2}(t) = \bar{a}_i + \frac{\bar{b}_i - \bar{a}_i}{\rho_i}\frac{t - \tau(1-\rho_i)}{t} - \frac{1}{2}t. \tag{C.11}$$

Lastly for $t \in (t'_{i2}, \tau)$

$$\bar{L}'_{i3}(t) = \bar{b}_i - \sqrt{t^2 - (\bar{b}_i - \bar{L}_i(0))^2 + 2(\bar{b}_i - \bar{a}_i)t - \frac{2(t - \tau(1-\rho_i))}{\rho_i}(\bar{b}_i - \bar{a}_i)}. \tag{C.12}$$

For the second case $\bar{L}_i(0) - t$ is less than $\bar{a}_i$. By similar manipulations as performed above we have

$$t''_{i1} = \frac{\rho_i}{2(\bar{b}_i - \bar{a}_i)}(\bar{L}_i(0) - \bar{a}_i)^2 \tag{C.13}$$

and for $t \in (\tau(1-\rho_i), t''_{i1})$

$$\bar{L}''_{i1}(t) = \bar{a}_i - t + \sqrt{\frac{2(\bar{b}_i - \bar{a}_i)(t - \tau(1-\rho_i))}{\rho_i}}. \tag{C.14}$$

Again we can find t''_{i2} as before and get

$$t''_{i2} = \frac{\rho_i}{1-\rho_i}\left[\frac{(\bar{L}_i(0) - \bar{a}_i)^2}{2(\bar{b}_i - \bar{a}_i)} + (\bar{a}_i - \bar{L}_i(0))\right] + \tau \tag{C.15}$$

and have for $t \in (t''_{i1}, t''_{i2})$

$$\bar{L}''_{i2}(t) = \frac{t - \tau(1 - \rho_i)}{\rho_i} + \bar{L}_i(0) - t - \frac{(\bar{L}_i(0) - \bar{a}_i)^2}{2(\bar{b}_i - \bar{a}_i)}. \qquad (C.16)$$

Lastly for $t \in (t''_{i2}, \tau)$

$$\bar{L}''_{i3}(t) = \bar{b}_i - \sqrt{(\bar{b}_i - \bar{L}_i(0))^2 + 2(\bar{b}_i - \bar{a}_i)t - \frac{2(t - \tau(1 - \rho_i))}{\rho_i}(\bar{b}_i - \bar{a}_i)}. \qquad (C.17)$$

With this machinery, we can derive the average cost per cycle. First, we find which case holds based on x_i^c and τ. In total we have three cases: 1) If $\bar{L}_i(0) - \bar{a}_i < t'_{i1}$ (i.e. the steady-state distribution "front" hits the earliest due-date "line" before t'_{i1}) then $\bar{L}_i(t)''$ is always used; 2) if $\bar{L}_i(0) - \bar{a}_i > t'_{i1}$ and $\bar{L}_i(0) - \bar{a}_i < \tau$ (i.e. the earliest due-date function hits the steady-state "front" above the earliest due-date line and the steady-state front at some time would have hit the earliest due-date line) then $\bar{L}_i(t)'$ is used for $t \in (\tau(1 - \rho_i), \bar{L}_i(0) - \bar{a}_i)$ and $\bar{L}_i(t)''$ for $t \in (\bar{L}_i(0) - \bar{a}_i, \tau)$; and 3) if $\bar{L}_i(0) - \bar{a}_i > \tau$ (i.e. the steady-state "front" never hits the earliest due-date line), then $\bar{L}_i(t)'$ is always used.

If $x_i^s > (\bar{b}_i - \bar{a}_i)/2$ then $\bar{L}_i(0) - \bar{a}_i < 0$ which is less than t'_{i1}. Therefore in this case $\bar{L}_i(t)''$ is always used. Moreover the cost per cycle is the same as that in the deterministic due-date case. The cost per cycle for class i is then the same as equations 4.12, 4.13 and 4.14 with the $\bar{l}_i - x_i^c/\rho_i$ term replaced by $(\bar{b}_i + \bar{a}_i)/2 - x_i^c/\rho_i$.

If $x_i^s < (\bar{b}_i - \bar{a}_i)/2$ then the cost per cycle is more complicated. Let's calculate the backorder (the first term in equation 3.8) and holding cost (the last two terms of equation 3.8) parts of the cost per cycle separately.

Backordering only occurs if $\bar{a}_i$ is less than $\tau(1 - \rho_i)$. If $\bar{L}_i(0)$ is less than $\min(t'_{i1}, t''_{i1})$ then only the triangular shaped region of $\bar{G}_i(s, t)$ will fall below zero over the course of the cycle. In this case, the earliest due-date function $\bar{L}_i(t)$ leaves the backorder region (i.e. becomes positive) at the point t_{i0}^1 which satisfies

$$t_{i0}^1 = \int_{\bar{a}_i - t_{i0}^1}^{0} \frac{\rho_i}{\bar{b}_i - \bar{a}_i}(s - (\bar{a}_i - t_{i0}^1))ds \qquad (C.18)$$

and so

$$t_{i0}^1 = \frac{\bar{b}_i - \bar{a}_i}{\rho_i} + \bar{a}_i - \sqrt{(\frac{\bar{b}_i - \bar{a}_i}{\rho_i} + \bar{a}_i)^2 - \bar{a}_i^2 - 2\frac{\bar{b}_i - \bar{a}_i}{\rho_i}\tau(1-\rho_i)}. \qquad \text{(C.19)}$$

The backorder cost for the cycle is then

$$-\bar{b}_i \int_{\tau}^{t_{i0}^1} \bar{a}_i - t + \sqrt{\frac{2(\bar{b}_i - \bar{a}_i)(t - \tau(1-\rho_i))}{\rho_i}} dt. \qquad \text{(C.20)}$$

If $\bar{L}_i(0)$ is greater than t''_{i1} then both the triangular shaped and flat part of $\bar{G}_i(s,t)$ is in the backorder region. Thus, backorders are completed at time t_{i0}^2 when

$$t_{i0}^2 = \int_{\bar{a}_i - t_{i0}^2}^{L_i(0) - t_{i0}^2} \frac{\rho_i}{\bar{b}_i - \bar{a}_i}(s - (\bar{a}_i - t_{i0}^2))ds + \int_{L_i(0) - t_{i0}^2}^{0} \rho_i ds. \qquad \text{(C.21)}$$

and so

$$t_{i0}^2 = \frac{\rho_i}{1-\rho_i}\left[\frac{(\bar{L}_i(0) - \bar{a}_i)^2}{2(\bar{b}_i - \bar{a}_i)} - \bar{L}_i(0)\right] + \tau \qquad \text{(C.22)}$$

The backorder cost is

$$-\bar{b}_i\left[\int_{\tau}^{t''_{i1}} \bar{a}_i - t + \sqrt{\frac{2(\bar{b}_i - \bar{a}_i)(t - \tau(1-\rho_i))}{\rho_i}} dt + \int_{t''_{i1}}^{t_{i0}^2} \rho_i dt\right]. \qquad \text{(C.23)}$$

There are four cases for the holding cost per cycle: 1) the steady-state distribution front has passed into the backorder region; 2) the steady-state front makes contact with $\bar{L}_i(t)$ below the earliest due-date line but above zero; 3) the steady-state "front" make contact above the earliest due-date line and would have had time to hit the earliest due-date line; and 4) the steady-state "front" does not have enough time to hit the earliest due-date line.

Thus the case 1) holding cost occurs if t_{i0}^2 is more than t''_{i1} and the cost is

$$h_i\left[\int_{t_{i0}^2}^{t''_{i2}} \left(\bar{L}''_{i2}(t)\right) dt + \int_{t''_{i2}}^{\tau} \left(\frac{\rho_i}{\bar{b}_i - \bar{a}_i}(\bar{L}''_{i3}(t) - \bar{a}_i) + (1 - \rho_i\frac{\bar{L}''_{i3}(t) - \bar{a}_i}{\bar{b}_i - \bar{a}_i})\bar{L}''_{i3}(t)\right) dt\right]. \qquad \text{(C.24)}$$

Case 2) is for t_{i0}^2 and $\bar{L}_i(0) - \bar{a}_i$ less than t''_{i1} and the cost is

$$\bar{h}_i \Big[\int_{t_{i0}^1}^{t''_{i1}} \left(\bar{L}''_{i1}(t) \right) dt + \int_{t''_{i1}}^{t''_{i2}} \left(\bar{L}''_{i2}(t) \right) dt + \int_{t''_{i2}}^{T} \left(\tfrac{\rho_i}{\bar{b}_i - \bar{a}_i} (\bar{L}''_{i3}(t) - \bar{a}_i) + (1 - \rho_i \tfrac{L''_{i3}(t) - \bar{a}_i}{\bar{b}_i - \bar{a}_i}) \bar{L}''_{i3}(t) \right) dt \Big]. \tag{C.25}$$

For case 3), $\bar{L}_i(0) - \bar{a}_i$ is greater than t'_{i1} and less than τ. The holding cost for the cycle is then

$$\begin{aligned} \bar{h}_i \Big[& \int_{t_{i0}^1}^{t'_{i1}} \left(\bar{L}'_{i1}(t) \right) dt + \int_{t'_{i1}}^{t'_{i2}} \left(\tfrac{\rho_i}{\bar{b}_i - \bar{a}_i} (\bar{L}'_{i2}(t) - \bar{a}_i) + (1 - \rho_i \tfrac{L'_{i2}(t) - \bar{a}_i}{\bar{b}_i - \bar{a}_i}) \bar{L}'_{i2}(t) \right) dt \\ & + \int_{t'_{i2}}^{L_i(0) - \bar{a}_i} \left(\tfrac{\rho_i}{\bar{b}_i - \bar{a}_i} (\bar{L}'_{i3}(t) - \bar{a}_i) + (1 - \rho_i \tfrac{L'_{i3}(t) - \bar{a}_i}{\bar{b}_i - \bar{a}_i}) \bar{L}'_{i3}(t) \right) dt \\ & + \int_{\bar{L}_i(0) - \bar{a}_i}^{T} \left(\tfrac{\rho_i}{\bar{b}_i - \bar{a}_i} (\bar{L}''_{i3}(t) - \bar{a}_i) + (1 - \rho_i \tfrac{L''_{i3}(t) - \bar{a}_i}{\bar{b}_i - \bar{a}_i}) \bar{L}''_{i3}(t) \right) dt \Big]. \end{aligned} \tag{C.26}$$

For the last case, $\bar{L}_i(0) - \bar{a}_i$ is greater than t'_{i1} and τ. The holding cost is

$$\begin{aligned} \bar{h}_i \Big[& \int_{t_{i0}^1}^{t'_{i1}} \left(\bar{L}'_{i1}(t) \right) dt + \int_{t'_{i1}}^{t'_{i2}} \left(\tfrac{\rho_i}{\bar{b}_i - \bar{a}_i} (\bar{L}'_{i2}(t) - \bar{a}_i) + (1 - \rho_i \tfrac{L'_{i2}(t) - \bar{a}_i}{\bar{b}_i - \bar{a}_i}) \bar{L}'_{i2}(t) \right) dt \\ & + \int_{t'_{i2}}^{T} \left(\tfrac{\rho_i}{\bar{b}_i - \bar{a}_i} (\bar{L}'_{i3}(t) - \bar{a}_i) + (1 - \rho_i \tfrac{L'_{i3}(t) - \bar{a}_i}{\bar{b}_i - \bar{a}_i}) \bar{L}'_{i3}(t) \right) dt \Big]. \end{aligned} \tag{C.27}$$

As can be seen, the average cost per cycle for the uniform due-date case is more complicated than the deterministic due-date example. Much of the structure, however, carries over from the deterministic due-date case. If the cycle start x_i^s is above $(\bar{b}_i - \bar{a}_i)/2$ then the cost structures are the same and one would expect a form of switching curves identical to the SELSP ones shifted by $\rho_i(\bar{b}_i + \bar{a}_i)/2$. Thus, we expect the same qualitative value for due-dates as in the deterministic case. Differences begin to occur when the cycle start moves close to the positive workload constraints. Cost per cycle increases rapidly as the machine is forced to work on orders with due-dates farther in the future than would have been the case in a deterministic due-date setting with a fixed due-date of $(\bar{b}_i + \bar{a}_i)/2$. If the cycle center was moved into this region, the higher inventory cost would favor a shorter cycle length than would be found in the deterministic due-date case.

Finding the optimal policy, however, by optimizing the average cost per cycle with respect to cycle center x_i^c and cycle length τ appears to be numerically intensive. Nonetheless, the calculation performed here could be used for the potential function optimization outlined in Chapter 3.

Bibliography

[1] R. Anupindi and S. Tayur. Managing stochastic multi-product systems: Model, measures, and analysis. *Operations Research*, 1994. To Appear.

[2] F. Baccelli, Z. Liu, and D. Towsley. Extremal scheduling of parallel processing with and without real-time constraints. *Journal of the Association for Computing Machinery*, 40:1209–1237, 1993.

[3] D. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1995.

[4] K. Bourland and C. Yano. The strategic use of capacity slack in the economic lot scheduling problem with random demand. *Management Science*, 40:1690–1704, 1995.

[5] O. Boxma, H. Levy, and J. Westrate. Efficient visit frequencies for polling tables: Minimization of waiting cost. *Queueing Systems*, 9:133–162, 1994.

[6] S. Browne and U. Yechiali. Dynamic priority rules for cyclic-type queues. *Advances in Applied Probability*, 21:432–450, 1989.

[7] A. Carr, Z. A Gullu, P. Jackson, and J. Muckstadt. An exact analysis of a production-inventory strategy for industrial planners. Technical report, School of Operations Research and Industrial Engineering, Cornell University, Ithica, NY, 1993.

[8] E. Coffman, A. Puhalskii, and M. Reiman. Polling systems with switchover times: A heavy-traffic averaging principle. Technical report. Bell Laboratories, Lucent Technologies, Murray Hill, NJ, 1995.

[9] E. Coffman, A. Puhalskii, and M. Reiman. Polling systems with zero switchover times: A heavy-traffic averaging principle. *Annals of Applied Probability*, 5, 1995. 681-719.

[10] D. Cox and W. Smith. *Queues*. Chapman and Hall, London, 1961.

[11] G. Dobson. The economic lot scheduling problem - achieving feasibility using time-varying lot sizes. *Operations Research*, 35:764–771, 1987.

[12] I. Duenyas and M. Van Oyen. Stochastic scheduling of parallel quwuws with setup costs. Technical report, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, 1993.

[13] S. Elmaghraby. The economic lot scheduling problem (elsp): Review and extensions. *Management Science*, 24:587–598, 1978.

[14] A. Federgruen and Z. Katalan. The stochastic economic lot scheduling problem: Cyclic base-stock policies with idle times. Technical report, Graduate School of Business, Columbia University, New Yrk, NY, 1993.

[15] A. Federgruen and Z. Katalan. Make-to-stock or make-to-order: that is the question; novel answers to an ancient debate. Technical report, Graduate School of Business, Columbia University, New York, NY, 1994.

[16] G. Gallego. Scheduling the production of several items with random demands in a single facility. *Management Science*, 16:1579–1592, 1990.

[17] S. Graves. The multi-product production cycling problem. *AIIE Transactions*, 12:233–240, 1980.

[18] A. Ha. Optimal dynamic scheduling policy for a make-to-stock production system. *Operations Research*, 1993. To Appear.

[19] R. Hariharan and P. Zipkin. Customer-order information, leadtimes, and inventories. *Management Science*, 41:1599–1607, 1995.

[20] M. Hofri and K. Ross. On the optimal control of two queues with server set-up times and its analysis. *SIAM Journal of Computing*, 16:399–420, 1987.

[21] J. Hsu. On the general feasibility test of scheduling lot sizes for several products on one machine. *Management Science*, 29:93–105, 1985.

[22] D. Iglehart and W. Whitt. Multiple channel queues in heavy traffic. *i* and *ii*. *Advanced Applied Probability*, 2:150–177, 1970.

[23] L. Karlin and H. Taylor. *A Second Course in Stochastic Processes.* Academic Press, New York, 1981.

[24] G. Klimov. Time sharing service systems i. *Theory of Probability and Its Applications*, 19:532–551, 1974.

[25] H. Kushner. *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations.* Academic Press, New York, NY, 1977.

[26] H. Kushner and P. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time.* Springer-Verlag, New York, NY, 1992.

[27] R. Leachman and A. Gascon. A heuristic scheduling policy for multi-item, single machine production systems with time-varying stochastic demands. *Management Science*, 34:377–390, 1988.

[28] T. Lennon Olsen. Approximations for the waiting time distributions in polling models with and without state-dependent setups. Technical report, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, 1996.

[29] P. Mandl. *Analytical Treatment of One-Dimensional Markov Processes.* Springer-Verlag, New York, NY, 1968.

[30] D. Markowitz, M. Reiman, and L. Wein. The stochastic economic lot scheduling problem: Heavy traffic analysis of dynamic cyclic policies. Technical report,

Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, 1996. Submitted for publication.

[31] P. Morse. *Queues, Inventories and Maintenance.* John Wiley, New York, NY, 1958.

[32] V. Nguyen. Fluid and diffusion approximations of a two-station mixed queueing network. *Mathematics of Operations Research*, 20:321–354, 1995.

[33] V. Nguyen. On base-stock policies for make-to-order/make-to-stock production. Technical report, Sloan School of Management, MIT, Cambridge, MA, 1995.

[34] M. Van Oyen and I. Duenyas. Control of a queueing system with sequence-dependent set-ups. Technical report, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, 1992.

[35] D. Pandelis and D. Teneketzis. Optimal stochastic dynamic scheduling in multi-class queues with tardiness and/or earliness penalties. *Probability in the Engineering and Information Sciences*, 8:491–509, 1994.

[36] A. Pena and P. Zipkin. Dynamic scheduling rules for a multiproduct make-to-stock queue. Technical report, Graduate School of Business, Columbia University, New York, NY, 1993.

[37] J. Qiu and R. Loulou. Multiproduct production/inventory control under random demands. *IEEE Transactions on Automatic Control*, 40:350–356, 1995.

[38] M. Reiman. Open queueing networks in heavy traffic. *Mathematics of Operations Research*, 9:441–458, 1984.

[39] M. Reiman and L. Wein. Dynamic scheduling of a two-class queue with setups. *Operations Research*, 1996. To Appear.

[40] R. Righter. Optimal scheduling of multiclass stochstic systems. *Probability in the Engineering and Information Sciences*, 10:229–241, 1996.

[41] A. Sharifnia, M. Caramanis, and S. Gershwin. Dynamic setup scheduling and flow control in manufacturing systems. *Discrete Event Dynamic Systems: Theory and Applications*, 1:149–175, 1991.

[42] C. Sox and J. Muckstadt. Optimization-based planning for the stochastic lot scheduling problem. Technical report, Operations Research and Industrial Engineering Dept, Ithica, NY, 1995.

[43] H. Takagi. *Analysis of Polling Systems.* MIT Press, Cambridge, MA, 1986.

[44] J. Van Mieghem. Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. *The Annals of Applied Probability*, 1994. To Appear.

[45] M. Veatch and L. Wein. Scheduling a make-to-stock queue: Index policies and hedging points. *Operations Research*, 1993. To Appear.

[46] L. Wein. Dynamic scheduling of a multiclass make-to-stock queue. *Operations Research*, 40:724–735, 1992.

[47] Y. Zheng and P. Zipkin. A queueing model to analyze the value of centralized inventory information. *Operations Research*, 38:296–307, 1990.

[48] P. Zipkin. Computing optimal lot sizes in the economic lot scheduling problem. *Operations Research*, 39:56–63, 1991.