Models of Flexible Workforce Management in Uncertain Environments

by

Edieal Jacob Pinker

S.M., Massachusetts Institute of Technology (1993)

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

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1996

© Massachusetts Institute of Technology 1996. All rights reserved.
Models of Flexible Workforce Management in Uncertain Environments

by

Edieal Jacob Pinker

Submitted to the Sloan School of Management on February 16, 1996, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

Today many firms are seeking ways to respond to variability in their demand for labor. Traditionally the responses have involved the hiring/layoff of workers. The ability to adapt the size of a workforce to the amount of work is a measure of numerical flexibility. This approach has been well studied in the labor economics and operations management literature, and while it is suitable for dealing with workload variability, on a monthly or longer time scale, it is not for shorter ones. Specifically, many firms are introducing working time and functional flexibility as ways in which to respond to this variability without involving hiring/layoff.

In this thesis we develop models of workforce management in stochastic environments. These models provide managers with a tool for making decisions about the utilization of flexible workforces and a framework within which to assess the benefits of different forms of flexibility. All the models are structured around two layers of optimization. The inner, decides the optimal use of the labor on hand given daily uncertainty in demand and absenteeism, while the outer-layer decides what types and how many workers to have on hand to minimize the expected costs of the dynamic labor utilization over some planning period.

We first focus on working time flexibility representing everything from overtime to short-timing. We then introduce functional flexibility. For all the models we derive theoretical results about their mathematical structure. We do extensive numerical experiments of the different models that demonstrate the strong linkages between the benefits of different forms of flexibility, stochasticity of the environment, and the timing of decisions relative to the arrival of information.
In the final section, we draw upon field experience with the USPS and a mutual fund company, and the labor economics and human resource literature to discuss the important management, quality of work life, and equity issues involved in implementing different flexible workforce arrangements. We discuss how the models developed can help address some of them and where they fall short.

Thesis Supervisor: Richard C. Larson
Title: Director, Center for Advanced Educational Services
Acknowledgments

In most of its facets, a doctoral degree program is an individualistic activity. While taking courses, ones grades are determined individually. During oral exams one stands alone in front of a questioning committee. Student’s tuition is funded individually and in the end a Ph.D. dissertation has one official author. who is also the sole recipient of the degree. Despite all this. I have never felt completely isolated while on this journey, because only through the help, support, and guidance of many people have I been able to achieve so much.

I’d like to thank Dick Larson, with whom I’ve worked from Day 1. He has always helped me keep a clear view of what my research should be and been an endless source of ideas and energy. I also thank my two other thesis committee members, Steve Graves and Jim Rebitzer for providing alternative perspectives on the research that have greatly enriched my thesis and stimulated my thinking. I also thank Oded Berman for his readiness to offer advice and guidance on things professional and personal.

It would have been impossible to produce the work I have, if not for the resources and comfort provided by the Operations Research Center. For these I thank Laura. Paulette, their senses of humor and the two dozen receptionists I’ve seen go by. With all due respect to the faculty, I have learned the most from my fellow students. the Center’s greatest asset. Many of them, past and present, have become my close friends.

I owe the most to my wife and parents. Monica has truly shared the experience with me and her presence has made every defeat less demoralizing and every victory more exciting and satisfying. My parents have long ago provided me with the fundamental ingredients of success, by setting great personal and professional examples. I dedicate this thesis to them, it is as much the fruit of their labors as mine.
Contents

1 Introduction .................................................. 12
   1.1 Types of labor flexibility .................................. 15
   1.2 Conceptual models ......................................... 22
   1.3 Definitions and Problem Classification .................. 23
       1.3.1 Definitions ........................................... 24
       1.3.2 Problem classification ................................ 27
   1.4 Outline of thesis ......................................... 28

2 One Job Problem Formulation .................................. 32
   2.1 Formulations ............................................. 33
       2.1.1 Assumptions ......................................... 33
       2.1.2 Set-up ............................................... 35
       2.1.3 Same-period notification, no absenteeism .......... 38
       2.1.4 Next-period notification, no absenteeism .......... 41
       2.1.5 Same-period notification with absenteeism ....... 45
       2.1.6 Next-period notification with absenteeism ....... 54
       2.1.7 Extensions .......................................... 57
   2.2 Special Cases ........................................... 59
   2.3 Summary ................................................ 60

3 Multiple-Job Problems Formulation and Analysis ............ 62
   3.1 General formulation and assumptions .................... 64
5.2.2 Functional flexibility ........................................ 137
5.2.3 Interaction between types of flexibility .................. 141
5.3 System behavior ................................................. 144
  5.3.1 Call-in Usage Trajectories .............................. 144
  5.3.2 Backlog trajectories ..................................... 148
5.4 Summary .......................................................... 151

6 Implementation Issues and Conclusions ....................... 153
  6.1 Field sites ..................................................... 155
    6.1.1 Mutual fund company .................................. 155
    6.1.2 Mail sorting facility ................................... 156
  6.2 What jobs are suitable for FWAs? .......................... 157
  6.3 The supply side of the flexible labor market .............. 161
    6.3.1 Job security ............................................. 161
    6.3.2 Working hours .......................................... 164
  6.4 Hidden costs to employers ................................... 168
  6.5 Managerial and organizational obstacles to FWA ......... 170
    6.5.1 False expectations ...................................... 171
    6.5.2 " The time to panic is now " .......................... 172
    6.5.3 Summary .................................................. 174
  6.6 Summary and implementation guidelines .................... 174
    6.6.1 Tensions .................................................. 175
    6.6.2 Employment packages and employee parameters ....... 177
    6.6.3 Job security ............................................. 179

7 Conclusions and future research .............................. 180
  7.1 Contributions ............................................... 180
  7.2 Future Research .............................................. 181
Contents

7.2.1 Direct extensions .......................... 182
7.2.2 Related new models ....................... 182

Bibliography .................................. 184

A ................................................. 190
  A.1 Work arrival processes .................... 190
  A.2 Parameters ................................ 192
  A.3 Technical issues .......................... 193

B ................................................. 195
  B.1 Causes of Unplanned Overtime in Mail Sorting Facility .......................... 195
  B.2 Observations .............................. 196
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Schematic of operational staffing decisions</td>
<td>23</td>
</tr>
<tr>
<td>1-2</td>
<td>Tree of Workforce Problems</td>
<td>27</td>
</tr>
<tr>
<td>2-1</td>
<td>State Space Schematic</td>
<td>40</td>
</tr>
<tr>
<td>2-2</td>
<td>Order of events with next-period notification</td>
<td>43</td>
</tr>
<tr>
<td>2-3</td>
<td>Order of events for next-period notification with redefined periods</td>
<td>44</td>
</tr>
<tr>
<td>2-4</td>
<td>Order of events for same-day notification with absenteeism case (i)</td>
<td>47</td>
</tr>
<tr>
<td>2-5</td>
<td>Order of events for same-day notification with absenteeism case (ii)</td>
<td>50</td>
</tr>
<tr>
<td>2-6</td>
<td>Order of events for same-day notification with absenteeism case (iii)</td>
<td>52</td>
</tr>
<tr>
<td>2-7</td>
<td>Order of events for next-period notification with absenteeism case (i)</td>
<td>55</td>
</tr>
<tr>
<td>3-1</td>
<td>Order of events for same-period notification with absenteeism case (i)</td>
<td>71</td>
</tr>
<tr>
<td>3-2</td>
<td>Order of events for next-period notification with absenteeism case (i)</td>
<td>77</td>
</tr>
<tr>
<td>4-1</td>
<td>Bounding the 2-job outer optimization problem</td>
<td>116</td>
</tr>
<tr>
<td>5-1</td>
<td>Expected cost vs. staffing level in 1-job problem</td>
<td>127</td>
</tr>
<tr>
<td>5-2</td>
<td>Expected cost vs. staffing in 2-job problem</td>
<td>129</td>
</tr>
<tr>
<td>5-3</td>
<td>Benefits of working time flexibility for different work arrival distributions</td>
<td>131</td>
</tr>
<tr>
<td>5-4</td>
<td>Expected costs with call-in flexibiliy relative to costs without flexibility</td>
<td>132</td>
</tr>
<tr>
<td>5-5</td>
<td>Expected Cost vs. G when N=5</td>
<td>133</td>
</tr>
<tr>
<td>5-6</td>
<td>Expected Cost vs. G when N=9</td>
<td>134</td>
</tr>
</tbody>
</table>
5-7 Expected cost vs. G for different information regimes  
5-8 Expected cost versus correlation in 2-job problems for W1 arrivals, comparing work arrivals W1 and W3  
5-9 Expected cost versus correlation in 2-job problems for W1 arrivals, comparing different levels of crosstraining  
5-10 Utilizing call-in and functional flexibility jointly  
5-11 Trajectory of expected call-in utilization, for next-period notification for W1 and different values of G  
5-12 Trajectory of variation in call-in utilization, for next-period notification for W1 and different values of G  
5-13 Trajectory of expected call-in utilization, comparing next and same-period notification for W1 and G=.6  
5-14 Trajectory of variation in call-in utilization, comparing next and same-period notification for W1 and G=.6  
5-15 Trajectory of expected backlog, for next-period notification with W1 and different values of G  
5-16 Trajectory of variation in backlog, for next-period notification with W1 and different values of G  
5-17 Trajectory of expected backlog, comparing next and same-period notification with W1 and G=.6  
5-18 Trajectory of variation in backlog, comparing next and same-period notification with W1 and G=.6  
6-1 The non-equivalence of the sets of secondary and flexibility suitable jobs  
A-1  
A-2
List of Tables

1.1 Mapping of flexibility to work environments .......................... 20

4.1 Cost for different staffing levels of example problem ................. 108
4.2 Search path in example problem ........................................ 109

5.1 2-job problem results with no working time flexibility. ............. 143
5.2 1-job problem results with call-in flexibility. .......................... 143

B.1 Planned and actual mail volume and hours worked on days with ‘unusual’ behavior. Note: Volume figures are in thousands of pieces of mail. ................................................................. 197
Chapter 1

Introduction

There are many competitive pressures on American companies today that force them to find ways to cut their operating costs and to improve the quality of their products and services. Firms are trying to minimize inventories, and eliminate defects. They are also reorganizing themselves to eliminate redundant activities and to streamline decision processes. It is inevitable that they consider methods to control labor costs. When a firm reorganizes itself and eliminates activities there is usually a reduction in labor costs, but there are also methods by which firms are attempting to reduce labor costs directly. There are two fundamental approaches taken to directly reducing labor costs\(^1\). The first is to reduce the labor input required to produce the firm's product by automating tasks previously performed by humans or by providing workers with training and/or tools that increase their productivity. The second approach is to increase labor flexibility to allow firms more options in how they utilize and compensate workers.

The first approach has been practiced throughout history and is strongly driven by technological advancement. An example of automation is to replace manual letter

\(^1\)Reduction of unit labor costs might be more appropriate, since a firm might seek to increase output without significant increase in labor costs.
sorters who can sort mail at a top rate of approximately 3600 pieces per hour with an optical character reader that sorts the mail at 36,000 pieces per hour. An example of increasing productivity is to provide mail handlers with motorized carts to move large quantities of mail around a sorting facility instead of dragging a smaller amount by hand. In the first example a job or task was eliminated, in the second example a task was made more productive.

The second approach, increasing labor flexibility, is a more recent phenomenon. There is substantial anecdotal evidence that firms are using a variety of flexible work arrangements to reduce costs, for example: Nurses in the Minneapolis-St. Paul hospital system have an ‘on-call’ status in which they may be called into work when there is a need for extra staff. While on-call and not at the hospital these nurses are paid the minimum wage [Hos92]. At GM there has been a significant increase in the use of overtime in the past year to expand production without expanding the workforce [Kil94]. At Volkswagen a recent labor agreement established the full-time work week to be 38.5 hours and gave managers the flexibility to spread these hours in different ways across the week [Gum95]. The United States Postal Service (USPS) has a class of workers called part-time flexible (PTF) that do not have fixed work schedules during the week [Ser94]. Throughout the economy there is evidence of firms trying to match their workforce to their workload as closely as possible using flexible techniques in a ‘just-in-time’ way [Bro87]. There is also historical statistical evidence documenting increases in particular forms of flexible work arrangements, namely temporary work\(^2\).

The other side of this phenomenon is that some of the arrangements cited above demonstrate that employers have difficulty in utilizing flexible work arrangements. The large amounts of overtime at GM led to labor strife. We have observed that although PTF’s are an option in USPS mail sorting facilities they are often not utilized. The payment of minimum wages to nurses on-call seems quite arbitrary and

\(^2\)The average daily employment of temporary workers has increased from .4 million in 1980 to nearly 2.0 million in 1994, based upon Bureau of Labor Statistics data [Fed95].
suggests a failure to assess the true impact of the policy. These difficulties are not unexpected, because even though, from a theoretical perspective, increased flexibility is a good thing\textsuperscript{3}, from a practical perspective there are many important questions that firms must answer before making decisions about flexible staffing. With respect to a particular work environment these questions are:

- What is flexibility?
- What is it used for?
- How do you decide what kinds and how much flexibility to utilize?

These questions are not specific to staffing decisions and it is useful to structure our thinking about them with the general model of flexibility presented in [dG94]. In this paper the model of flexibility is defined around technologies and environments. Technologies may be more or less flexible and environments may be more or less diverse. In the framework the flexibility of a technology serves as a "hedge" against the diversity in the environment. It requires that there is a way to rank order the environments in terms of diversity. In cases where the environments and technologies have more than one component it is difficult to rank diversity and flexibility in order to make good decisions about what technologies to use and which environments to match them to. We do not actually determine such an ordering. However, following this model and [Tre92] we develop a mapping of different types of labor flexibility to the environments that they are useful in. There are several types of labor flexibility prevalent in today's workplace and in [Tre92] they are classified as numerical, working time, functional, and pay flexibility.

\textsuperscript{3}From an economic or mathematical viewpoint increased flexibility implies fewer constraints or more options for a decision maker which cannot worsen the situation.
1.1 Types of labor flexibility

Numerical flexibility governs the employers options in altering the size of their workforce and allows a firm to better match their workforce to workload, when there is workload variability over a time scale of months. Working time flexibility governs the scheduling of work hours and the number of hours worked, and allows a firm to better match its workforce to workload when there is workload variability from day to day or within days. Functional flexibility governs many organizational issues such as job definition, supervisory hierarchies, and internal mobility. This flexibility can increase the range of activities a workforce is capable of performing and therefore make it more adaptable to work demand variability. It can also be a necessary condition for a firm to be able to apply different operational techniques such as JIT and continuous improvement. Pay flexibility governs a firm’s ability to tie wages more closely to the firm’s economic performance and/or the worker’s individual performance.

We can think of these different types of flexibility as different labor technologies that are each most applicable to different environments. We are interested in environments in which the primary problem is matching workforce size to workload when there is variability in workloads. Therefore we will focus our attention on those forms of labor flexibility that directly relate to this problem, namely numerical, working time, and to some degree functional flexibility. The variability in a workload can be characterized by the degree to which it is stochastic and the time scale of the variability. We consider the following time scales: months, weeks, days, and hours. For each of these time scales there are examples of work environments in which there is deterministic or seasonal workload variability, examples in which there is stochastic variability, and examples of combinations of both. We use the term stochastic broadly, to describe work arrivals that have a random component. We do not preclude correlation between work arrivals at different times. Each form of variability may benefit from a different form of labor flexibility.
1.1 Types of labor flexibility

Monthly variability in workload can be a result of unexpected changes in demand for a firm's services, a change in the type of services required or a seasonal effect. In all cases, a firm may need to expand or contract its workforce for a period of one or more months but usually not more than 6 months\(^4\). The problem of deciding how to adapt the workforce size in such situations was considered for manufacturers in [HM60]. The same approach applies for services without the ability to build inventory. In [HM60] the only option for adjusting workforce sizes is to hire or fire workers. Today there are other options commonly used to achieve the same effect such as, subcontracting excess work and contracting contingent workers\(^5\). The ability to contract contingent workers is a form of numerical flexibility, a simple model of using contingent workers for numerical flexibility appears in [Abr86]. When a firm faces a shift in demand from one mix of services to another it may be aided by functional as well as numerical flexibility. Cross-trained (or multiskilled) workers can be shifted, from a low demand job to a higher demand job, thereby reducing the firm's need to lay-off workers. from the low demand job, and its need to hire new workers, for the high demand job. Both actions involving significant costs\(^6\). In these situations firms also use working time flexibility in the form of overtime and short-time to smooth their labor requirements [Mic87].

When workload varies on a weekly, daily or hourly basis working time and functional are the most applicable forms of flexibility. It is not feasible to hire and fire on a regular basis from week to week therefore numerical flexibility is not relevant in any significant scale\(^7\). When workloads vary deterministically, working time flexibility is

---

\(^4\)We assume that the firm has already taken measures such as scheduling vacations and low priority tasks for low workload times of year.

\(^5\)The term 'contingent worker' has been used in many different ways. We will use the definition in [Pol89]: "Any job in which an individual does not have an explicit or implicit contract for long-term employment or one in which the minimum hours worked can vary in a non-systematic manner."

\(^6\)Cross-training also involves significant costs and many other complications such as, assigning supervisory responsibility to workers who perform a variety of tasks, setting pay incentives, paying for training, and maintaining skill levels [Kle94].

\(^7\)It is feasible, and common, to use temporary workers to cover for absences and unfilled positions.
manifested in days-off scheduling and the use of part-time work. The benefits of such
flexibility are demonstrated in [BLP94, Bak73, MR73]. The benefits of crosstraining
(functional flexibility) for within-day variability is demonstrated in [BLP94, War72].

When workload varies stochastically within a day, there is a practical limit to
working time adjustments. Once a scheduled worker has arrived to work it is difficult
to reassign them to different hours or to send them home. The most common working
hour adjustment is to offer overtime hours. In mail processing plants managers can
move the start time of a worker’s shift up to two hours, but must pay an overtime
premium for those shifted hours and must be able to make the adjustment before
the worker has arrived. The most common phenomenon is for mail, the workload, to
arrive in the predicted quantities, but at later than planned times. In these situations,
a large amount of overtime is usually utilized. Short-timing is a less common option
and is used when workload is light. Functional flexibility can be beneficial for within
day variability if the variability shifts work from one job to another. This occurs
in stochastic flow shops and the benefits of crosstraining for these environments has
been shown in [Tre89].

When workload varies stochastically from day to day or week to week there is an
opportunity to benefit from functional flexibility, as in the previous case, but there
is also a significant opportunity to benefit from working time flexibility. The ability
to adjust worker’s hours up or down each day or week is manifested in many ways.
Overtime is an example of working time flexibility that allows and employer to expand
hours to meet higher demand. Short-timing is a system by which employee’s hours are
reduced in times of decreased demand without layoffs. Comp-time is a combination
of the above in which workers might work longer hours one week but compensate for
it by working shorter hours the following week. I.e. it is a restricted ability to bank
work hours.

Another representation of this kind of flexibility we use in this thesis is call-in
workers. Call-in workers are workers who are called in to work on short notice when there is a need for them. A firm employing call-in workers would draw upon two sources of labor. The primary source, is the pool of regular staff who have fixed schedules and are permanent employees of the firm. The second source would be call-in workers who do not have preset schedules, i.e. workers with flexible working time. These workers could be permanent employees or contracted from an external temporary employment agency. In either case the firm would be required to guarantee some minimum pay over the employment period to each call-in worker. If the workers are called in more than the guaranteed amount they would be compensated accordingly. An aggregate model of this approach, with no backlogging of work from day-to-day, is presented in [BL93a] and [BL94].

An interesting hierarchical model for planning the use of long-, medium-, and short-term flexibilities is presented in [WR93]. In this paper an attempt is made to model the use of short-timing, temporary workers, overtime, and crossstrained floaters in a service environment with stochastic workloads and absenteeism. The classification of the model as service based is used to justify the lack of inventories and/or backlog. Furthermore it is assumed that there is an endless supply of temporary workers. These assumptions together lead to a formulation in which the daily staffing decisions are independent of one another. This approach is successful in modeling the impact of long-term planning decisions on the availability of resources for short-term staffing decisions but fails to accurately represent the short-term system behavior.

There are many work environments that would fall under the category of services with day-to-day workload variability. For example, in the transaction processing center of a mutual fund company the amount of work that arrives in the mail each day is highly variable. This workload is subject to variability in mail service, economic trends, market events, and response to marketing promotions. Customer’s requests for transactions must be processed very quickly, preferably on the day of receipt, and
1.1 Types of labor flexibility

obviously cannot be processed before they arrive. Staffing to the average workload leads to the accumulation of backlogged work and poor customer service. Staffing at a level that is sufficient, with high probability, is very costly and involves a low utilization of the workforce. Another example is a hospital. Patients arrive to an emergency room randomly each day and, as many medical services are distributed to clinics and centers, a higher proportion of hospital workloads are emergency cases and therefore less predictable and manageable. The hospital cannot afford to be understaffed since patients must receive their treatments in a timely manner. On the other hand, medical personnel are highly trained and expensive, therefore overstaffing can be a great financial burden.

The use of call-in workers has many potential benefits. For the employer, call-in workers can provide an expanded labor capacity with a lower cost than an equivalent labor capacity composed of entirely full-time regular workers. Call-in workers also provide a consistent source of expanded capacity making it easier for a firm to maintain consistent workforce performance standards. For employees, the use of call-in workers will reduce the amount of overtime worked and should reduce the overall amount of wage received per worker employed by the firm. On the other hand the call-in worker arrangement can suggest a long term commitment of employment. Furthermore, the reduction in wages paid come with an increase in leisure time for regular workers who work less overtime and for call-in workers who do not work full-time hours. Although the uncertainty in schedule is a negative factor, overall a call-in arrangement offers a stable source of income for workers who do not want to work full-time.

Some employers try to reap these benefits today but in very adhoc ways. The USPS has a class of employees called casual employees that are not required by union contract to have regular schedules. These workers are utilized to fill staffing gaps in operations caused by absenteeism or vacations etc. [Ser94]. In some financial services
companies part-time data entry clerks are informally promised 16 hours of work each week and are sent home early without pay if workload is light. this is an example of short-timing. In neither case has a systematic analysis been done to determine staffing needs in light of the existing scheduling flexibility.

We summarize the different types of labor flexibility, their environments and the associated literature in table 1.1. We consider 6 different work environments each specified by the time frame of the variability in demand and whether the variation is deterministic or stochastic. For each environment an ‘x’ marks the types of labor flexibility that can be effective for it. Beneath each ‘x’ we cite literature in which the particular combination of flexibility and environment was modeled.

<table>
<thead>
<tr>
<th>Environments - by type of demand variation</th>
<th>Within day</th>
<th>Daily/Weekly</th>
<th>Monthly +</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part-time</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>[BLP94]</td>
<td>[Bak73]</td>
<td>[MS83]</td>
</tr>
<tr>
<td>Short-time</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Overtime</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>[WR93]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>On-call</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Temporary</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>[BL94]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functional</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>[War72]</td>
<td>[BLP94]</td>
<td>[Tre89]</td>
</tr>
<tr>
<td>Numerical</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1: Mapping of flexibility to work environments

From the table we can see that overtime and functional flexibility are useful in

The limitation of this approach is that short-timed workers have already come in to work and must be compensated for the inconvenience.
1.1 Types of labor flexibility

every situation. However, in practice, there are limits on the amount of overtime
that can be performed by an individual worker over an extended period of time since
excessive overtime can reduce productivity. The effectiveness of functional flexibility
is limited by the cost of training and as we show in chapter 5 it is also limited by
the nature of the variability in work mix within a work environment. As a result it
is important to consider other forms of labor flexibility.

We focus our interest on environments with stochastic variability on a daily or
weekly level. There are some existing models that could apply to this environment
but they do not address the problems relevant to the financial services and hospital
examples described previously. These have following characteristics in common:

- Workload is stochastic.

- Workload can be backlogged.

- There is no 'finished goods' inventory.

- The time scale does not allow for hiring/firing of workers.

The short time scale and lack of inventory are, in general, features of services
that distinguish them from manufacturing environments [BL93b]. Therefore, the ex-
isting stochastic models for flexibly managing resources in manufacturing are not
transferable to the type of problems described above. Unfortunately the volumi-
nous literature on flexible and agile manufacturing has not been matched within the
workforce management literature in general, or the service workforce management
literature in particular\(^9\). This is despite the fact that it is now well established that

\(^9\)I think the reason for this is the traditional dominance of manufacturing and production of
the Operations Research/Management Science literature. Most of the problems in these areas are
formulated as machine driven problems with labor being included as an additional cost but not
a primary concern. The main reason for this may be the capital intensive nature of production
processes, but I think there is also a historical dimension related to management traditions of trying
to manage around labor rather than with labor. We won't get into this here. As the U.S economy
1.2 Conceptual models

Our effort is built around the following conceptual models, C1 and C2.

**C1:** A firm must decide how many workers of different types to hire/contract for a planning horizon in order to minimize its costs from backlogged work and labor usage. The worker types are distinguished from one another by their flexibility and costs. The cost of making the decision in C1 is partially determined by what happens during the course of the planning horizon. We model this activity with conceptual model C2.

**C2:** In each time period work arrives to the workplace and the regular staff of the firm, works on it. The firm must decide how to draw upon its flexible workers to process work, in a way that minimizes costs from backlog and labor usage over the entire planning horizon. This decision process is depicted schematically in Figure 1-1.

Model C1 is fundamentally a one-period labor demand problem. The cost function is a production function with a variety of labor inputs. The firm is deciding how much of each input to buy in order to maximize profit or minimize cost. Most economic analysis of labor demand involve relatively simple and aggregated production functions\(^{10}\). As a result they do not give much insight into the dynamics of a

---

\(^{10}\)In [Ham93] dynamic models of labor demand are formulated and analyzed, but are structured for longer time scales that are most relevant for numerical flexibility, not working time flexibility.
Figure 1-1: Schematic of operational staffing decisions

firm's utilization of flexible workforces on a shorter time scale. Model C2 provides the setting for learning about these dynamics.

Using these two conceptual models we define a framework for modeling labor flexibility on an operational level. Using the terms and taxonomy of this framework we formulate a family of workforce management models that represent utilization of working time and functional flexibility on the operational decision making level of a firm. We find that the behavior of a firm, with regard to flexibility on the operational level of model C2, has important ramifications for labor demand on the higher planning level of model C1.

1.3 Definitions and Problem Classification

To make the proper staffing decisions structured by the conceptual models we need to further specify the work environments. We think of staffing decisions as the selection of a production technology. In this section we define the terms that characterize different work environments on a greater level of detail than has been done thus far.
With this detail we construct a modeling framework in which we can identify the right matches of staffing technologies to work environments.

We develop a general set of models for workplaces similar to the financial services and hospital examples described previously. We assume that all of these workplaces follow operating policies with the following general structure: The firm process all work that arrives each day with only regular staff. If there is excess work a decision is made to draw upon non-regular labor capacity or not. Excess capacity is drawn from the call-in workforce and overtime hours. If excess work remains it is backlogged. Different workplaces will be distinguished from one another by the following characteristics: notification, backlog tolerances, workload distributions, absenteeism, call-in worker contracts, crosstraining, and their cost structures.

In this section we define what we mean by the above distinguishing characteristics of workplaces.

### 1.3.1 Definitions

- **Period** - We will model the problem using a general time unit called a period. This can be considered to be a day, a week, etc. On the level of a period we make decisions about drawing upon call-in workers or overtime and how much backlog to allow. The day is the time unit for which the problem is most naturally defined but we will not restrict ourselves to this scale.

- **Planning Period** - A planning period is some number of periods for which we make the decision of how to size the workforce.

- **Notification** - The timing of decisions regarding the utilization of non-regular or flexible labor resources within a time period, is strongly dependent upon when managers know\(^{11}\) what their workload is. For example, if a manager has

\(^{11}\text{In reality there will always be uncertainty in the workload so we consider knowledge of workload}\)
a good idea in period 1. of what the workload will be in period 2, she can call in workers in period 1 for period 2. In this situation, it is reasonable to expect the call-in workers to be able to come in. If a manager only knows the period 2 workload after the start of period 2 it is less likely that she can get call-in workers for the same period. If overtime is possible the decision to utilize it may be at a different point in time then call-in decisions and the arrival of workload information.

- **Backlog policies** - There are many policies that a firm may have for dealing with backlog and ways to represent these policies. For example, a firm might assess a cost to the backlog in the system each period. This cost might be linear, quadratic or have some other functional form. A firm might also have no tolerance for backlog, which implies that any backlog is assigned an infinite cost. A firm may also view backlog as unsatisfied demand that is lost with some penalty. The differences between these policies strongly affect the dynamics of the system operation and the staffing policies used.

- **Workload Statistics** - We characterize new work arrival distributions by considering the following two questions: 1) Are the workloads in each period identically distributed, and 2) are the period workloads independent?

- **Absenteeism** - Another source of variability in a workplace is absenteeism. We consider the cases when there is no absenteeism and when there is absenteeism in one or both of the regular and call-in staffs.

- **Call-in worker contracts** - We view call-in workers as being contracted for a planning period. These workers could be regular staff that are temporarily placed into this scheduling category, permanent call-in workers, or externally contracted temporary workers. In any event we assume that call-in workers to be a low variance estimate of workload.
are guaranteed a minimum number of payed work periods for the contracted planning period. If call-in workers are needed more than the guaranteed number of periods then they must receive additional compensation.

- Crosstraining - In firms where there are multiple jobs requiring different worker qualifications, crosstraining can be used to increase worker management flexibility. If the workload of one job is not strongly correlated with the workload at another job, it is possible that having some workers who are crosstrained to perform each job will reduce labor costs.

- Cost Structures - In formulating the cost functions for the models we take into account the following sources of labor costs: Benefits costs for regular and call-in workers, hourly wages for all workers and overtime wage premiums. We also assess a penalty for carrying backlog.

One might ask why there is a need for another taxonomy of service operations, when there are so many in the literature already. For example, queueing theory provides a classification scheme that characterizes systems by the stochastic processes governing arrival and service of customers to a service facility, the number of servers and service priority rules, see [Kle75]. However, this classification does not really give any structure to the allocation of labor and staffing decisions, which are our interest here. [DL94] presents a taxonomy for service capacity problems built around five dimensions: type of service, resources, cost of services, demand characteristics, and the cost of not satisfying demand. In this scheme cost of services include the fixed and variable cost of labor, aggregating all work arrangements into this one dimension. Such a taxonomy does not provide enough resolution to structure the modeling of flexible workforces, and for the same reasons neither do the classification schemes of [WB92] and [Lov83].
1.3.2 Problem classification

While all the above characteristics may distinguish work environments we focus on a subset that most strongly defines the structures of the models we develop. These characteristics are: Notification, Crosstraining, and Backlog policy. We use the first two characteristics to define a family of workforce management problems. This family is outlined in the form of a tree in Figure 1-2.

**Flexible Staffing Model Tree:**

```
Determine Staffing Levels
  /
One Job (No Crosstraining)   Multiple Jobs (Crosstraining)
  /
  /
Same Period Notification   Next Period Notification
```

Figure 1-2: Tree of Workforce Problems

On the most general level, the problem we want to solve is: *What are the optimal staffing levels that would minimize labor and backlog penalty costs?* On the next level, we model one job and multiple jobs with no crosstraining as a single problem type. We develop a distinct model for environments with multiple jobs and crosstraining. On the lowest branching level, on the problem tree, we distinguish between the two extreme notification regimes, same-period and next-period. Same-period represents the situation in which flexible staffing decisions are made with complete workload information. I.e. the manager knows the exogenous work arrival for a period early enough that she can make all staffing decisions for the same period that she has
information. Next period represents the situation in which staffing decisions are made without any information about new exogenous work arrivals. I.e. the workload information arrives too late for a manager to make staffing decision for the same-period of the information. In reality, the situation is somewhere in between these two extremes. We use the extreme cases to define the spectrum of problems. There are very many different backlog policies that could be in place, within a workplace, so we don’t explicitly represent them in the problem tree. Within the context of each of the problems, on the tree, we have to consider the effects of the backlog policy, absenteeism and non-homogeneous and/or dependent work arrival distributions.

1.4 Outline of thesis

To summarize, we have identified an important workforce management problem area that has yet to be thoroughly considered in either the Operations Research, Labor Economics or Human Resource Management literatures. The problem is how to incorporate and utilize workforce flexibility in matching workforces to workloads in service environments when workloads vary stochastically on relatively short time scales. There are many labor intensive workplaces with such variability and therefore with potentially much to gain from research in this area.

The fields of Operations Research and Management Science perform three intellectual functions that we denote as: 1) satisfying economic rationality assumptions, 2) revealing insights about human activities, and 3) technological innovation. The first function is based upon the role OR/MS models play in the optimal or rational allocation of resources. Economic models usually assume that actors are behaving in a utility maximizing way but this assumption implies that the actors have a way of determining what this is. This kind of optimization is a primary function of OR/MS. The second function is a result of analyzing new OR/MS models. By striving to represent human decision processes, in as precise a way as possible, OR/MS models can
often provide new insights into these processes through experimentation with, and mathematical analysis of the models. The third function results in cases where an application of OR/MS techniques allows decision makers to resolve problems that were intractable before. This can open up new areas of activity and/or give a competitive advantage to an economic agent. In this thesis we perform, with varying degrees of success, all three functions with respect to the workforce management problem stated above.

Using the definitions and problem classification of the previous section we formulate a set of models of flexible workforce management in stochastic environments. These models allow us to perform the three functions in the following way:

- The models provide us with tools with which to make cost minimizing staffing decisions that otherwise could not be made, thereby eliminating an inefficiency of the labor market.

- By analyzing the models mathematically and through numerical experiments we are able to identify the determinants of the benefits to a firm of various forms of workforce flexibility.

- The models are also a technological innovation because by providing a way to assess the value of different flexible work arrangements they make a wider range of staffing options available to employees and employers.

In Chapter 2 we formulate the models that form the 1-job branch of the problem tree in Figure 1-2 representing the use of call-in and overtime flexibility. We make distinctions between the two notification schemes and three different ways of representing absenteeism. We also discuss straightforward extensions of these formulations and give interpretations of a number of special cases. In Chapter 3 we formulate the models that form the 2-job branch of the problem tree in Figure 1-2. These models
include the working time flexibility of the 1-job models but also incorporate functional flexibility. We formulate the models in two ways representing different training geometries.

In Chapter 4 we analyze the mathematical problems formulated in Chapters 2 and 3 and relate them to other mathematical problems that appear in the OR/MS literature. We prove that the objective function of the problem to determine the optimal staffing level is convex in the staffing. We also characterize the optimal staffing decisions on a period level for special instances of the problems. Finally, we suggest some techniques for speeding up the solution of the various optimization problems involved in the models.

In Chapter 5 we conduct a series of numerical experiments that identify the factors that determine the effectiveness of different flexible work arrangements. We find that there are strong links between the relative benefits of flexibility, the information available to a decision maker, and the degree and type of stochasticity in the system. Having seen the potential benefits of flexibility demonstrated by the analysis of the models, it is natural to ask if these staffing arrangements can be successfully implemented in practice. This is the subject of Chapter 6. Since we are modeling the labor aspect of work systems there are many more side effects to operational policies than if we were modeling machines. As is determined in [Lee85] the characteristics of a work environment as defined by the working time arrangement and content of the job, effect turnover, absenteeism, morale and productivity of employees. All these factors can be important determinants of a work system's performance over time. We do not try to include all these labor supply issues in our models, rather we synthesize our field observations, experience with our models and evidence form the Human Resource Management and Labor Economics literature to develop an understanding of how they affect the implementation of these arrangements. To accomplish this we develop a complete characterization of work environments that are suitable for
flexible workforce arrangements. We draw upon dual labor market theory to create guidelines for adapting the implementation of the models to the concerns of workers in different segments of the market.

Drawing upon the issues raised in discussing implementation we complete Chapter 6 and the thesis with a description of interesting areas for future research in workforce management problems and a summary of our conclusions from conducting this research.
Chapter 2

One Job Problem Formulation

In this chapter we formulate models of working time flexibility for the 1-job branch of the problem tree. We represent the flexibility with call-in workers and overtime. In this situation the manager must decide:

**Problem 2.1** How many regular workers $N$ and call-in workers $M$ to staff for a planning period of length $V$ time periods, so that the expected labor and backlog costs are minimized over the entire planning period.

We initially assume that there is no absenteeism and that the new work that arrives each period is independent and identically distributed to all other periods and formulate the same-day and next-day notification problems. We then formulate both problems with absenteeism, in doing this we show how absenteeism can be manifested in several different ways requiring distinct formulations. We complete the formulation section with a discussion of extensions to situations in which exogenous work arrival distributions are independent but not identical and situations in which partial work arrival information is available before staffing decisions are made.

The problem formulations all address problem 2.1 and assume that in each period managers have two decisions to make, how many call-in shifts to utilize and how many overtime shifts to utilize. These decisions are assumed to be made optimally
and dynamically. The major difference between the different problems formulated in this chapter is the information available about workloads and absenteeism when these two decisions must be made. This means that for each problem in this branch of the problem tree we formulate a different dynamic program for the period by period decision making.

After formulating the various problems we discuss how these formulations can be used to characterize a variety of real-world work arrangements by setting the problem parameters to specific values.

2.1 Formulations

In all the models in this chapter we make some general assumptions. Some of these are relaxed in the discussion of extensions. In addition, within each particular formulation we have model specific assumptions which are discussed in the appropriate sections. Before formulating the problems we outline the general assumptions.

2.1.1 Assumptions

Work arrivals: Exogenous work arrivals each period are i.i.d random variables. We discuss a relaxation of this constraint in the extensions section.

Overtime: In all the models we place bounds upon the amount of overtime that can be performed each period. In all cases this bound is expressed as a function of the total staffing pool trained for each job. Since the dependency on $M$, the call-in worker pool, does not take into account the actual number of call-in workers on-site when overtime is assigned, odd situations may arise. For example, let us assume that $N = 5$ and $M = 5$, and the bound on overtime is $OT_{\text{max}}(N, M) = .2(N + M) = 2$, signifying that no worker should work more than .2 of a shift on overtime. In this case it is possible that on a particular day (assuming no absenteeism) we call-in only
one worker. This means that 6 workers are on-site and the intent of the constraint is to restrict their overtime to \((.2)(6) = 1.2\) shifts worth. The constraint, however, allows up to 2 shifts worth of overtime. In theory we can make this situation much more extreme by considering staffing levels that are predominantly call-in.

We now make some arguments as to why these anomalies should not have an important impact upon the behavior of the models. If we assume that the overtime wage per shift is greater than that of regular workers rate per shift, then we know that if the expected overtime usage each period exceeds one, we would improve the staffing situation by adding another regular worker, thereby also increasing the amount of allowable overtime.

This argument does not address the role of peak points of overtime usage on the system behavior. However, we can argue that these instances in which large amounts of overtime are used are those in which large amounts of backlog are possible. These situations are anticipated by call-in usage. Finally, we can argue that in practice, in extreme situations, firms will bend or break the overtime limit rules to "get the job done", often calling upon managers and others to augment the work pool. In the models with absenteeism we base the overtime constraint on expected available workers. To summarize, the overtime bound is modeled as a "soft constraint" with respect to how it is implemented in reality.

**Productivity rates:** We do not model productivity in perfect detail. We do not consider any interactions between productivities of different workers, or of productivities changing over time. We also do not rely upon any hierarchy of productivity. There is wide debate on relative productivities of regular workers, contingent workers and overtime work. Some argue that contingent workers are more productive because they have fewer distractions at work. Some say that overtime work is less productive because of fatigue. While others argue that overtime is more productive because it is supervised more closely and workers stretch out the work during regular hours in
order to create opportunities for overtime.

**Absenteeism:** We assume that each worker is present with a probability $p$, independently of all other workers. $p$ may vary among types of workers. In the formulations of the models with absenteeism there is an implicit assumption that regular workers are paid for periods they are absent. This is similar to practice, in that workers are given an allowance of sick and personal days. For call-in workers there is an implicit assumption that a period in which a call-in worker is solicited and is unable to work does not count toward their guaranteed minimum number of periods of work. This means that call-in workers have in effect an allowance of sick and personal days as do regular workers.

In the formulation we also assume that there is no penalty to workers, call-in or regular, for excessive absences. The rationale behind this is that an unreliable employee, in terms of excessive absences, will be fired relatively soon after the behavior was detected, and that over all reliable employees absences average out over many periods.

**Call-in protocol:** We assume that call-in workers are solicited in a sequential way that balances the calls among all of them so that the differences in work hours between any two in the pool can be at most one shift’s worth.

### 2.1.2 Set-up

In each period the amount of work that can be processed is determined by the number of regular workers present, the number of call-in workers utilized and the number of overtime shifts utilized, any excess is backlogged to the next period. In the different problems, the timing of the call-in and overtime decisions may be different but the tradeoffs involved in these decisions are the same. Let’s consider these tradeoffs here.
2.1 Formulations

If we do not utilize call-in or overtime work and the workload in a period $t$ does not exceed the available regular staff's capacity there are no backlog costs and no new staffing costs, since the cost of the $N$ regular workers is a sunk cost. If the workload exceeds the regular staff's capacity we have backlog costs unless a sufficient number of call-in and/or overtime shifts are utilized. I.e. we must tradeoff the cost of backlog with the cost of overtime and/or call-in worker utilization. The cost of overtime is a linear function of the number of hours utilized in period $t$. The cost of call-in workers is dependent upon the cumulative use of call-in workers up until period $t$; because, we contract $M$ call-in workers for the planning horizon with a guarantee of a fraction $G$. of $V$, paid periods of work in the planning horizon. The payment for $GV$ per call-in worker periods of work is a sunk cost and any periods worked in excess of $GV$ periods per call-in worker incurs extra costs. This means that the overtime/call-in decision, in each period $t$, is a dynamic decision based upon cumulative call-in utilization, expected future utilization, and future backlogs.

To make these various costs more tangible we now define the cost parameters that will be used in all the problem formulations. First, we recall our definition of the shift as the number of hours a worker works within a time period at their ordinary wage\(^1\). Second, we only define labor costs over the course of a single planning horizon (composed of an arbitrary number of time periods). Third, we assume that each worker has three components to their compensation: A salary received for each shift worked, a benefits component for each shift worked, and a fixed component for being part of the workforce during the planning period in question. The cost parameters we use in the formulations are as follows:

\[
C_f = \text{fixed cost for each worker, call-in or regular that is a member of the workforce for the planning period. It includes fixed component of compensation and fixed costs per worker for the firm. E.g. human resource departments costs. services}
\]

\(^1\) Later we introduce the constant $\pi$ as the number of units of work processed per shift of work.
2.1 Formulations

available to all employees regardless of status, etc.

\( C_{rw} = \) per-shift cost of a regular worker that combines the benefits and salary.

\( C_{cw} = \) per-shift cost of a call-in worker that combines the benefits and salary.

\( C_{ot} = \) the premium paid per-worker per shifts worth of overtime worked.

\( C_b = \) the per-time period penalty incurred by the firm for every unit of work backlogged.

\( C_B = \) the penalty incurred by the firm for every unit of work backlogged in the final period of the planning period.

We have assumed linear costs of \( C_{ot} \) for overtime shifts worked and \( C_{cw} \) for call-in worker shifts paid for. We have specified linear backlog penalties \( C_b \) and \( C_B \).

Note: We could formulate the models with an arbitrary functional form for backlog penalties. We use linear penalties here to be consistent with the numerical examples and analysis sections. Having understood these trade-offs and cost parameters we can now state the problem more completely as:

**Problem 2.2** What staffing level \( S = (N, M) \) should a firm contract, over a planning period of length \( V \), to minimize the expected labor and backlog cost incurred when optimal call-in/overtime decisions are made dynamically each period, if the call-in workers are guaranteed at least a fraction \( G \), of \( V \), paid periods of work per planning period.

The implication of this problem statement is that for each problem in the one-job branch of the problem tree the expected cost of a staffing level \( S \) is the expected cost of the optimal solution to a finite horizon, labor allocation, dynamic problem. We formulate problem 2.2 as:

\[ \min_S C_{rw}NV + C_{cw}MGV + C_f(N + M) + E[f^S_1(x_1, \kappa_1)] \]
subject to:

\[ S \geq (0, 0) \]

\[ g(S) \leq 0 \]

\[ S \text{ integer} \]

where \( C_f \) is a fixed cost for each worker, call-in or regular, and \( C_{rw} \) and \( C_{cw} \) are respectively per-period costs for employing regular and call-in workers that include hourly wages and pro-rated benefits. \( g(S) \) is a constraint based upon the relative sizes of the call-in and regular workforces. The expression \( f^S_1(x_1, \kappa_1) \) is the expected cost of making optimal dynamic staffing decisions over the planning horizon given a staffing level \( S \), with \( x_1 \) units of work in the system, and \( \kappa_1 \) guaranteed shifts of call-in workers unused in period 1.

### 2.1.3 Same-period notification, no absenteeism

**Definitions**

The problem has \( V \) stages that are equivalent to periods.

The state of the system at each stage \( t \) is given by the vector \((\kappa_t, x_t)\).

\( \kappa_t \) is unused portion of the total call-in worker guarantee \( MGV \) at the end of stage \( t-1 \).

\( x_t \) is defined to be the workload in the system at the start of stage \( t \). This work is composed of new work that can be thought of arriving between stages \( t-1 \) and \( t \) which we will call \( d_t \), and work left in the system at the end of stage \( t-1 \).

**Decisions**

\( u_t \) is the number of call-in workers utilized in stage \( t \).
2.1 Formulations

\( \omega_t \)  the number of overtime shifts utilized in stage \( t \).

Constraints

\[ u_t \leq M \]

\[ \omega_t \leq OT_{\text{max}}(N, u_t) \]

Where \( OT_{\text{max}}(N, u_t) \) is a function representing the maximum amount of overtime that can be performed by a complement of workers \( (N, u_t) \). I.e. \( N \) regular workers are always available and \( u_t \) call-in workers are on site in stage \( t \).

Transitions

\[ \kappa_t = [\kappa_{t-1} - u_{t-1}]^+ \]

\[ x_t = [x_{t-1} - \pi_{rw} N - \pi_{cw} u_{t-1} - \pi_{ot}\omega_{t-1}]^+ + d_t \]

where the \( \pi \)'s are constants that convert units of people-shifts to units of work for different types of labor. The geometry of the state space is depicted in figure 2-1.

Costs  The costs incurred in stage \( t \) are the costs of backlog, the cost of overtime and a cost for call-in worker usage in excess of the guarantee. The cost at stage \( t \) is given by:

\[ C_{\text{cw}}[u_t - \kappa_t]^+ + C_{\text{ot}}\omega_t + C_b[x_t - \pi_{rw} N - \pi_{cw} u_t - \pi_{ot}\omega_t]^+ \]

To save space we define:

\[ b_t = \text{Max}(x_t - \pi_{rw} N - \pi_{cw} u_{t-1} - \pi_{ot}\omega_{t-1}, 0) \]

The end of planning horizon cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined
as:

$$C_V(b_V) = C_B(b_V)$$

The cost to go function in stage $V$ is:

$$f_V(\kappa_V, x_V) = \text{Min}_{u_t, \omega_t} \left\{ C_{cw}[u_t - \kappa_t]^+ + C_{ot} \omega_V + C_V(b_V) \right\}$$

and in stage $t$:

$$f_t(\kappa_t, x_t) = \text{Min}_{u_t, \omega_t} \left\{ C_{cw}[u_t - \kappa_t]^+ + C_{ot} \omega_t + C_b(b_t) + E_{d_{t+1}}[f_{t+1}(\kappa_{t+1}, x_{t+1})] \right\}$$

The interpretation of this equation\(^2\) is: If you are in state $(\kappa_t, x_t)$ in stage $t$, the expected cost you incur from stage $t$ until the end of the planning horizon is the minimum expected cost over all decisions $(u_t, \omega_t)$ of the overtime and backlog cost in stage $t$, plus the cost that is incurred from stage $t + 1$ until the end of the planning horizon.

\(^2\)The notation used in this formulation is based upon the conventions of [Ber87].
period. When we solve this dynamic program we will start with $f_V()$ and work back to $f_1()$.

2.1.4 Next-period notification, no absenteeism

In the previous section we have assumed that all staffing decisions (i.e. how many call-in workers and overtime shifts to utilize), for a period, have been made with perfect information about the workload in that period. In this section we consider the scenario where we do not have perfect information about the workload when we make the decision to call-in workers. This situation can arise because, for example, we must notify call-in workers by 5pm on Tuesday if we want them to work on Wednesday. It can also occur, for example, when we can notify call-in workers by 7am Wednesday morning that we want them to work that day, but do not have perfect knowledge of the workload for Wed. by that time. However, next-period notification describes all situations in which call-in worker decisions are made for a time period with incomplete knowledge of workload for that period. In this thesis we only consider next-period notification scenarios in which exogenous work arrivals in each period are independent and identically distributed. Therefore, information about workload is unchanged until a new work arrival occurs.

Within this scenario we can consider two subcases, relating to when we make overtime decisions. Subcase (i): Overtime must also be allocated before we have perfect workload information. Subcase (ii): Overtime may be allocated after information has become available.

I.e. In subcase (i) overtime and call-in decisions are made at the same time and in subcase (ii) the overtime decision may be made later. These two cases arise in practice in the sense that in many workplaces managers may not require overtime and therefore must seek volunteers. These volunteers are more likely to be found if sought out earlier as represented in subcase (i). On the other hand there are workplaces in
which employers may require overtime (within some limits) and therefore do not have provide notice to the employees as is represented in subcase (ii).

**Subcase (i) Formulation** Since the staffing decisions are being made earlier relative to the workload information arrival we redefine the states and decisions appropriately:

**State Definitions** We will define the state of the system with two dimensions \((x_t, \kappa_t)\) subject to a one dimensional disturbance \(d_t\).

We define \(x_t\) to be the amount of work left in the system at the end of stage \(t - 1\), i.e., the backlog. We define \(\kappa_t\) to be the remaining hours in the call-in worker guarantee by the end of stage \(t\).

We define \(d_t\) to be the random quantity of new work arriving to the system in period \(t\) to be worked on in period \(t\).

**Decisions** \(u_t, \omega_t\) are respectively the amount of call-in workers and overtime shifts solicited in period \(t\) for utilization in period \(t + 1\).

These decisions are constrained as follows:

\[
    u_t \leq M
\]

\[
    \omega_t \leq OT_{\text{max}}(N, u_t)
\]

State transitions occur as follows:

\[
    x_{t+1} = [x_t + d_t - \pi_{rw}N - \pi_{cw}u_t - \pi_{ot}\omega_t]^+
\]

\[
    \kappa_{t+1} = [\kappa_t - u_t]^+
\]
2.1 Formulations

The final period cost function (or terminal value function) is defined as:

\[ C_V(\kappa_{V+1}, \bar{x}_{V+1}) = C_B(x_{V+1}) \]

The cost to go function in stage V is:

\[ f_V(\kappa_V, x_V) = \min_{u_V, \omega_V} \left\{ C_{oV} \omega_V + C_b(x_{V-1}) + C_{cw} [u_V - \kappa_V]^+ + E_{d_V} [C_V(\kappa_{V+1}, x_V)] \right\} \]

and in stage t:

\[ f_t(\kappa_t, x_t) = \min_{u_t, \omega_t} \left\{ C_{ot} \omega_t + C_b(x_t) + C_{cw} [u_t - \kappa_t]^+ + E_{d_t} [f_{t+1}(\kappa_{t+1}, x_{t+1})] \right\} \]

**Subcase (ii) Formulation** In this subcase we make the assumption that the timing of the call-in decision for a period and the overtime decision for that period are separated by the arrival of information about the workload (see Figure 2-2). To accommodate this we redefine the start and end of a period. A period begins with the exogenous arrival of new work followed by the overtime decision for that period and call-in decision for the next-period (see Figure 2-3).

![Figure 2-2: Order of events with next-period notification](#)

**Definitions** We define a two dimensional state space: \((x_t, \kappa_t)\) where:
2.1 Formulations

Figure 2-3: Order of events for next-period notification with redefined periods

\( \kappa_t \) is unused portion of the total call-in worker guarantee \( MGV \) at the end of stage \( t - 1 \).

\( x_t \) is defined to be the workload in the system, in excess of the call-in workers on hand, at the start of stage \( t \). \( x_t \) may be negative.

We define a single random variable \( d_t \) for each stage \( t \) as, the amount of exogenous work arriving to the system at the start of stage \( t \).

**Decisions** We have two decision variables: \( u_t \), the number of call-in workers utilized in stage \( t \), and \( \omega_t \) the number of overtime shifts utilized in stage \( t \).

**Constraints**

\[ u_t \leq M \]

\[ \omega_t \leq OT_{\text{max}}(N, M) \]

Where \( OT_{\text{max}}(N, M) \) is a function representing the maximum amount of overtime that can be performed by a complement of workers \( (N, M) \). I.e. a standard limit on overtime has been set based upon the total staffing level\(^3\).

\(^3\)In theory basing the overtime limit on the total staffing level allows for situations in which overtime is assigned in amounts that require more call-in workers to be present than are actually
State Transitions

\[ x_t = [x_{t-1} - \pi_{rw}N - \pi_{ol\omega_t}]^+ + d_t - \pi_{cw}u_{t-1} \]

\[ \kappa_t = [\kappa_{t-1} - u_{t-1}]^+ \]

**Cost Functions**  The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

\[ C_V(x_V - \pi_{ol\omega_V}) = C_B(x_V - \pi_{ol\omega_V}) \]

The cost-to-go function in stage V is:

\[ f_V(\kappa_V, x_V) = \min_{\omega_{V}} \{C_{ol\omega_V} + C_V(x_V - \pi_{ol\omega_V})\} \]

in stages \( t \):

\[ f_t(\kappa_t, x_t) = \min_{\omega_t, u_t} \{C_{ol\omega_t} + C_b(x_t - \pi_{ol\omega_t}) + C_{cw}[u_t - \kappa_t]^+ + E_{dt+1}[f_{t+1}(\kappa_{t+1}, x_{t+1})]\} \]

### 2.1.5 Same-period notification with absenteeism

We now consider the staffing problem when there is absenteeism among the regular and call-in worker pools. We assume that absences among regular workers are independent of absences among call-in workers and that absences are independent of workload. As in the no absenteeism case, we always have two staffing decisions to make for each period, namely how many call-in workers to use and how many overtime workers to use. We model the multiple decision points by splitting each exogenous work arrival cycle, or time period, into two stages for the DP staffing engine.

utilized that day. In practice this anomaly is only significant for extreme problem parameter choices. See discussion in the assumptions section.
2.1 Formulations

Therefore, while we considered a $V$ stage problem in the no absenteeism case we now consider the problem to have $2V$ stages with stage 1 being the first stage. There are several different scenarios for how absenteeism can affect the decision making process. We characterize these scenarios by the relative timing of absence information and staffing decisions. We first discuss the interpretation of our representation of absenteeism in practice. We then formulate the three different scenarios for absenteeism.

**Interpretation:** Since absences are not counted toward call-in worker's guarantees, to some degree it represents the worker's autonomy. Traditionally one might think of an absence as a period in which a workers is incapable of working. We can also think of it as a period in which a worker chose not to work when solicited, because it was not convenient for him. In this sense the greater the probability of absenteeism for a worker in the model, the more autonomy they have over their work hours.

**Case (i)** In this case we assume that we make the call-in decision when we know the regular worker absenteeism and workload. We then assume that we make the overtime decision when we know the call-in worker absenteeism. In this case we are assuming that regular workers are giving some notice about their absence and that we know this information when we start to solicit, call-in workers. We then solicit call-in workers until we find the amount we want or have exhausted the available ones. At this point in time we know how many regular workers and call-in workers are actually at work, how much work there is, and make the overtime decision. The order of events is depicted in figure 2-4.

**Definitions** We define a two dimensional state space: $(x_t, \kappa_t)$ where:

$\kappa_t$ is unused portion of the total call-in worker guarantee $MGV$ at the end of stage $t - 1$. 
2.1 Formulations

Figure 2-4: Order of events for same-day notification with absenteeism case (i)

\( x_t \) is defined to be the workload in the system at the start of stage \( t \).

We define the following random variables in each odd numbered stage:

\( n_t \) is the number of regular workers who are present in stage \( t \).

\( d_t \) is the amount of exogenous work arriving to the system in stage \( t \).

We define \( s_t = d_t - \pi_{rw} n_t \) to be the newly arrived staffing 'shortage' for stage \( t \). I.e. if \( s_t \) is positive it means that the new work arriving to the system exceeds the regular staff. If \( s_t \) is negative it means that there are more regular workers present than needed for the new work and therefore the excess staff can work on the backlog, if any, from the previous day.

For the even-numbered stages we define the random disturbance: \( m_t \) as the number of call-in workers who are available in stage \( t \).

**Decisions** As with the random variables we have different decisions variables for odd and even stages.

If \( t \) is odd we decide: \( u_t \), the number of call-in workers utilized in stage \( t \).

If \( t \) is even we decide: \( \omega_t \) the number of overtime shifts utilized in stage \( t \).
Constraints

\[ u_t \leq M \]

\[ \omega_t \leq OT_{\text{max}}(\bar{n}_t, \bar{m}_t) \]

Where \( OT_{\text{max}}(\bar{n}_t, \bar{m}_t) \) is a function representing the maximum amount of overtime that can be performed by a complement of workers \((n_t, m_t)\). I.e a standard limit on overtime has been set based upon the expected number of regular workers present and call-in workers available.

State Transitions  For odd stages \( t + 1 \):

\[ x_{t+1} = [(x_t - \pi_{ot}\omega_t)^+ + s_{t+1}]^+ \]

\[ \kappa_{t+1} = \kappa_t \]

Note: We are defining the workload at the start of an odd stage as the work beyond the processing capability of the regular staff.

For even stages \( t + 1 \):

\[ x_{t+1} = [x_t - \pi_{cw} \min[u_t, m_t]]^+ \]

\[ \kappa_{t+1} = [\kappa_t - \min[u_t, m_t]]^+ \]

Cost Functions  The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

\[ C_2(x_{2V'} - \pi_{ot}\omega_{2V}) = C_B(x_{2V} - \pi_{ot}\omega_{2V}) \]
2.1 Formulations

The cost-to-go function in stage $2V$ is:

$$f_V(\kappa_{2V}, x_{2V}) = \min_{\omega_{2V}} \{C_{ot}\omega_{2V} + C_{2V}(x_{2V} - \pi_{ot}\omega_{2V})\}$$

and in even stages $t$:

$$f_t(\kappa_t, x_t) = \min_{\omega_t} \left\{ C_{ot}\omega_t + C_b(x_t - \tau_{ot}\omega_t) + E_{s_{t+1}}[f_{t+1}(\kappa_{t+1}, x_{t+1})]\right\}$$

and in odd stages $t$:

$$f_t(\kappa_t, x_t) = \min_{u_t} \left\{ E_{m_{t+1}}[C_{cw}[\min[u_t, \kappa_t]] + f_{t+1}(\kappa_{t+1}, x_{t+1})]\right\}$$

Case (ii) In this case we assume that we make the call-in decision without knowing the regular worker absenteeism, only the workload. We then assume that we make the overtime decision with complete information about call-ins, regulars, and the workload. In this case we are assuming that the regular workers do not give notice of absences and just don't show up. I.e. when we find out how many regular workers are present it is too late to solicit more call-in workers.

We can see from the timeline in figure 2-5 that as in case (i) there are two decision points, one for call-in workers and one for overtime workers. We again model this by splitting each exogenous workload arrival cycle into two stages for the DP staffing engine. We define the same two dimensional state space as in case(i): $(x_t, \kappa_t)$.

We define the random variable $d_t$ in each odd numbered stage as the amount of exogenous work arriving to the system in stage $t$.

For the even-numbered stages we define the following random variables:

$n_t$ is the number of regular workers who are present in stage $t$.

$m_t$ is the number of call-in workers who are available in stage $t$. 
2.1 Formulations

Figure 2-5: Order of events for same-day notification with absenteeism case (ii)

**Decisions** If \( t \) is an even stage we make the decision: \( u_t \), the number of call-in workers utilized in stage \( t \).

If \( t \) is an even stage we make the decision: \( \omega_t \) the number of overtime shifts utilized in stage \( t \).

**Constraints**

\[
 u_t \leq M \\
 \omega_t \leq OT_{\text{max}}(\bar{n}_t, \bar{m}_t)
\]

Where \( OT_{\text{max}}(\bar{n}_t, \bar{m}_t) \) is defined as in case(i).

**State Transitions** For odd stages \( t + 1 \):

\[
x_{t+1} = [[x_t - \pi_0(\omega_t) + d_{t+1}]^+]^+
\]

\[
\kappa_{t+1} = \kappa_t
\]
2.1 Formulations

For even stages $t + 1$:

$$x_{t+1} = [x_t - \pi_{cw}\min[u_t, m_t] - \pi_{tw}n_t]^+$$

$$\kappa_{t+1} = [\kappa_t - \min[u_t, m_t]]^+$$

Cost Functions The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

$$C_{2V}(x_{2V} - \pi_{ot}\omega_{2V}) = C_B(x_{2V} - \pi_{ot}\omega_{2V})$$

The cost to go function in stage $2V$ is:

$$f_V(\kappa_{2V}, x_{2V}) = \min_{\omega_{2V}} \{C_{ot}\omega_{2V} + C_{2V}(x_{2V} - \pi_{ot}\omega_{2V})\}$$

and in even stages $t$:

$$f_t(\kappa_t, x_t) = \min_{\omega_t} \{C_{ot}\omega_t + C_b(x_t - \pi_{ot}\omega_t) + E_{d_{t+1}}[f_{t+1}(\kappa_{t+1}, x_{t+1})]\}$$

and in odd stages $t$:

$$f_t(\kappa_t, x_t) = \min_{u_t} \{E_{m_{t+1}, n_{t+1}}[C_{cw}\min[u_t, m_t] - \kappa_t]^+ + f_{t+1}(\kappa_{t+1}, x_{t+1})]\}$$

Case (iii) In this case we assume that we make the call-in decision without knowing the regular worker absenteeism and without knowing the call-in worker absenteeism but make the overtime decision with complete information. The implication here is that the regular workers do not give notice of their absences and that the call-in workers who are absent are ones that have said they would come in to work but do not show up.

We can see from the timeline in figure 2-6 that as in case (i) there are two decision
points, one for call-in workers and one for overtime workers. We again model this by splitting each exogenous workload arrival cycle into two stages for the DP staffing engine.

Figure 2-6: Order of events for same-day notification with absenteeism case (iii)

We define the same two dimensional state space \((x_t, \kappa_t)\) where:

\(\kappa_t\) is unused portion of the total call-in worker guarantee \(MGV\) at the end of stage \(t - 1\).

\(x_t\) is defined to be the workload in the system at the start of stage \(t\).

We define the random variable \(d_t\) in each odd numbered stage, as the amount of exogenous work arriving to the system in stage \(t\).

For the even-numbered stages we define the random disturbances \(n_t\) as the number of regular workers who are present in stage \(t\) and \(m_t\) as the number of call-in workers who we want to come in, that actually do in stage \(t\).

The definition of the disturbance \(m_t\) is different here than in cases (i) and (ii). We now model the call-in absenteeism as occurring among the subgroup of workers who are requested.
Decisions  If $t$ is an even stage we make the decision: $u_t$, the number of call-in workers asked for in stage $t$.

If $t$ is an even stage we make the decision: $\omega_t$ the number of overtime shifts utilized in stage $t$.

Constraints

$$u_t \leq M$$

$$\omega_t \leq OT_{\text{max}}(\bar{n}_t, \bar{m}_t)$$

Where $OT_{\text{max}}(\bar{n}_t, \bar{m}_t)$ is defined as in cases (i) and (ii).

State Transitions  For odd stages $t + 1$:

$$x_{t+1} = [(x_t - \pi_{ot}\omega_t)^+ + d_{t+1}]^+$$

$$\kappa_{t+1} = \kappa_t$$

For even stages $t + 1$:

$$x_{t+1} = [x_t - \pi_{cw}m_t - \pi_{rw}n_t]^+$$

$$\kappa_{t+1} = [\kappa_t - m_t]^+$$

Cost Functions  The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

$$C_{2V}(x_{2V} - \pi_{ot}\omega_{2V}) = C_B(x_{2V} - \pi_{ot}\omega_{2V})$$
2.1 Formulations

The cost to go function in stage V is:

\[ f_V(\kappa_{2V}, x_{2V}) = \min_{\omega_{2V}} \{ C_{ot}\omega_{2V} + C_{2V}(x_{2V} - \pi_{ot}\omega_{2V}) \} \]

and in even stages \( t \):

\[ f_t(\kappa_t, x_t) = \min_{\omega_t} \{ C_{ot}\omega_t + C_b(x_t - \pi_{ot}\omega_t) + E_{dt+1}[f_{t+1}(\kappa_{t+1}, x_{t+1})] \} \]

and in odd stages \( t \):

\[ f_t(\kappa_t, x_t) = \min_{\mu_t} \{ E_{mt+1,n_{t+1}}[C_{cw}[m_t - \kappa_t]^{+} + f_{t+1}(\kappa_{t+1}, x_{t+1})] \} \]

2.1.6 Next-period notification with absenteeism

In this case we assume that we make the call-in decision when we know the regular worker absenteeism but before we know the new work arrival. We then assume that we make the overtime decision when we know the call-in worker absenteeism and the workload situation. In this case we are assuming that regular workers are giving some notice about their absence and that we know this information when we start to solicit, call-in workers. We solicit call-in workers until we find the amount we want or have exhausted the available ones. At this point in time we know how many regular workers and call-in workers will actually be at work and how much work there is and make the overtime decision. The order of events is depicted in figure 3-2.

We define a two dimensional state space: \( (x_t, \kappa_t) \) where:

\( \kappa_t \) is unused portion of the total call-in worker guarantee \( MGV \) at the end of stage \( t - 1 \).

\( x_t \) is defined to be the workload in the system at the start of stage \( t \).

We define the following random variables in each odd numbered stage:
2.1 Formulations

Figure 2-7: Order of events for next-period notification with absenteeism case (i)

- $n_t$ is the number of regular workers who are present in stage $t$.

- $d_t$ is the amount of exogenous work arriving to the system in stage $t$.

For the even-numbered stages we define the random variable: $m_t$ as the number of call-in workers who are available in stage $t$.

**Decisions**  As with the random variables we have different decisions for odd and even stages.

- If $t$ is odd we decide: $u_t$, the number of call-in workers utilized in stage $t$.
- If $t$ is even we decide: $\omega_t$ the number of overtime shifts utilized in stage $t$.

We place the following constraints on the decisions:

$$u_t \leq M$$

$$\omega_t \leq OT_{\text{max}}(\bar{n}_t, \bar{m}_t).$$

Where $OT_{\text{max}}(\bar{n}_t, \bar{m}_t)$ is a function representing the maximum amount of overtime that can be performed by a complement of workers $(n_t, m_t)$. I.e a standard limit on overtime has been set based upon the expected number of regular workers present and call-in workers available. To simplify notation we define $a_t = \min[u_t, m_t]$ as the actual number of call-in workers utilized in stage $t$. 
State Transitions  For odd stages $t + 1$:

$$x_{t+1} = [x_t - \pi_{ot} \omega_t]^+ - n_t$$

$$\kappa_{t+1} = \kappa_t$$

Note: $x_t$ may take on negative values for odd $t$. If $x_t$ is positive it means that the backlogged work in the system exceeds the regular staff processing capacity. If $x_t$ is negative it means that there are more regular workers present than needed for the backlog currently in the system and therefore the excess staff can work on the new work yet to arrive that period.

For even stages $t + 1$:

$$x_{t+1} = [x_t - \pi_{cw} a_t + d_t]^+$$

$$\kappa_{t+1} = [\kappa_t - a_t]^+$$

Cost Functions  The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

$$C_{2V}(x_{2V} - \pi_{ot} \omega_{2V}) = C_B(x_{2V} - \pi_{ot} \omega_{2V})$$

The cost-to-go function in stage 2V is:

$$f_V(\kappa_{2V}, x_{2V}) = \min_{\omega_{2V}} \{C_{ot} \omega_{2V} + C_B(x_{2V} - \pi_{ot} \omega_{2V})\}$$

and in even stages $t$:

$$f_t(\kappa_t, x_t) = \min_{\omega_t} \{C_{ot} \omega_t + C_b(x_t - \pi_{ot} \omega_t) + E_{n+1}[f_{t+1}(\kappa_{t+1}, x_{t+1})]\}$$
and in odd stages $t$:

$$f_t(\kappa_t, x_t) = \min_{u_t} \left\{ E_{m_{t+1}, d_{t+1}} [C_{uw} (a_t - \kappa_t)^+ + f_{t+1}(\kappa_{t+1}, x_{t+1})] \right\}$$

The assumption that the systems starts with no backlogged work requires some discussion. The formulation clearly does not require this to be the case and we could also draw the starting work in the system from some probability distribution representing the backlog in the system at the end of $V$ periods. We could generate such a distribution with numerical experiments. Alternatively we could set the terminal stage backlog cost very high to drive the system to empty itself. In the numerical results chapter we discuss this issue more.

2.1.7 Extensions

In most practical applications of the staffing models we have developed here, there will be different exogenous work arrival processes for different time periods. Typically there will also be exhibited a day of the week seasonality in the work arrival profile. It is also common that some partial work arrival information is available before staffing decisions are made. We can accommodate both of these phenomena in a straightforward fashion within the framework of the models formulated in this chapter.

**Seasonality** Implicit in the models is the assumption that all the regular workers are scheduled to work in each time period. This can be interpreted as all regular workers being full-time workers. When workloads vary from time period to time period because of a day of the week type of seasonality it is common to use part-time workers, i.e. workers who are not scheduled to work each time period, to match workforce to workload. The models we have developed here are geared toward the stochastic variability as opposed to the deterministic/seasonal variability. Therefore, we do not explicitly model the scheduling of part-time workers. However, it is possible
to consider call-in workers to be flexible part-time workers.

In this more general workload arrival case we model the new work arriving in period $t$ as $d_t = s_t + r_t$ where $s_t$ is a deterministic component of workload and $r_t$ is the stochastic component that is distributed according to some probability mass function $F_t(r)$. Note that this PMF is dependent upon the time period.

If the deterministic components $s_t$ are significant we can assume that a separate schedule for regular workers including part-time regulars has been determined to satisfy the deterministic needs. We can then proceed with our ordinary models concerned with only the stochastic component $r_t$. If the deterministic component of the work arrival is not significant with respect to the stochastic component we can just view $d_t$ as being a stochastic variable as we did in all our models.

To summarize, we do not really have to change our models to adapt to non-identical workload arrivals. All we do is use the time period dependent distributions in all calculations of expected values involving workload in the dynamic programs.

**Partial information**  The notification regimes we have formulated in this chapter are extremes of the possible amounts of information available to a manager when she makes staffing decisions. In practice, between one period's work arrival and the next period's call-in decisions a manager might receive partial information about work arrivals.

To accommodate such a partial or mixed information regime, we would extend out models by simply adding another work arrival point. I.e. at the beginning of odd periods there would be an exogenous work arrival drawn from one stochastic process, while for even periods there would be an arrival drawn from another process. These processes would be independent. The odd arrival would represent the partial workload information indicating the minimum arrival for the time period. The total arrival would be completed by the even period arrival, We do not present the formulation of this extension here, but in Chapter 5 we do present results of numerical experiments
with a partial information notification regime.

2.2 Special Cases

The models we have formulated are very general in many senses. We have not assumed any special probability distributions for the random variables in the model. We also have created a structure that can represent many of the realistic work force management arrangements that exist today. These arrangements are modeled by choosing special values for some of the different parameters in the models. We now demonstrate how this can be done to represent different working time flexibility arrangements.

**Traditional workplace**  In the traditional, inflexible workplace there are only regular workers and overtime available to the managers. The problem of determining how many regular workers to staff and how to utilize overtime can be represented in our models by setting $M = 0$. The outer optimization becomes a problem of choosing the cost minimizing value for $N$ and the dynamic decisions made within the inner optimizations involve managing the use of overtime to reduce backlog.

**Temporary workers**  Today many workplaces rely on temporary agencies to supply them with labor when the regular workforce is insufficient. These temporary workers are only paid for the days they work. This arrangement can be represented in our models by setting $G = 0$. This means that the call-in workers are not guaranteed any days of work. This implies that there is no need to select a value of $M$. It is assumed that there is always some employment agency that can provide temporary help on short notice.
Short-timing/Flextime  In some workplaces (more commonly in Europe than the United States) management works out arrangements with workers in which the hours of regular fulltime employees are set over longer time spans than usual. For example, a fulltime worker will be defined to be any worker that works 40 hours per five-day work week. This definition does not require that each worker work eight hours per day. Although, there might be a restriction that within any day a worker may only be required to work upto 10 hours. This form of flexibility can be represented by our models as well. If we consider call-in workers to be the flexible workers we can set $G = 1$ and only charge an overtime premium for work done in excess of the guarantee $MGV$.

Everything in between  While we have shown the way the most common workforce management practices can be represented by our models it is important to note that these special cases are extremes. This suggests that what is done in practice is only a very limited representation of the arrangements available to a firm. Our models provide a framework for making decisions about a wide spectrum of work arrangements that form compromises between the special cases described above.

2.3  Summary

In this chapter we have formulated the general one-job problem 2.2 as an optimization over two dimensions; regular workforce size $N$ and call-in workforce size $M$. The value of the objective function at a particular solution $S = (N, M)$ is the expected cost of the optimal solution to a finite horizon dynamic program. We have formulated all the one-job problems depicted in the problem tree in figure 1-2. That is, we have formulated the various problems as distinguished by notification and absenteeism. All the dynamic programming formulations are based upon a Markovian structure that governs the work in the system and the cumulative utilization of call-in workers.
In the formulations involving absenteeism we characterized several different cases distinguished by the relative timing of the two staffing decisions made each period and information about workload and absenteeism. Each such case required a distinct formulation. Finally we showed how the formulations could be extended to situations in which the exogenous new work arrival each time period were not identically distributed. In all cases we have assumed that they are independent.
Chapter 3

Multiple-Job Problems

Formulation and Analysis

In this chapter we formulate and analyze the various staffing problems for the multiple-job branch of the problem tree. Our interest in modeling multiple-jobs and crosstraining is to develop insight into the benefits of functional flexibility caused by the pooling of randomness in the system. Our intuition suggests that when all workers are crosstrained for two jobs we can view the system as a one-job system that has less variability than two jobs taken independently. In this case, all other things being equal, it is the covariance of the two work arrival distributions that determine the benefits of crosstraining.

In practice we can expect that all other things are not equal. Different jobs have different backlog penalties, workers have varying productivity rates at different jobs, and of course different levels of training draw different salaries. The manager must decide what mix of workers to employ and how to assign them to tasks. In this chapter, we model these decisions based upon functional flexibility. At the same time we continue to consider working time flexibility and therefore create a framework for studying the interplay of these varied forms of flexibility.
We have been using the terms functional flexibility and crosstraining interchangeably, implying an enlarged set of skills on the part of the crosstrained or functionally flexible worker. In this thesis we define crosstraining in a broader sense that involves what might be called geographical or organizational flexibility. In these categories we represent workers whose flexibility does not involve different skills than other workers, but rather an ability or willingness to work in different locations or within different organizations at a job with essentially the same content.

For example a city hospital system might draw upon a geographically disparate floating pool of nurses that agree to work in any of the hospitals that require extra staffing because of absences and/or high patient loads. In terms of professional skills, these "floaters" may be no different than any of the absent nurses they cover for. I.e. they may be extremely inflexible with regard to the content of the jobs they do. However, they may be flexible with respect to which hospital and/or nursing unit they work in.

We start with the special case of two jobs in which there are workers trained to perform one or both jobs. The fundamental problem is basically the same as problem 2.1. However, in the case of two-jobs we must consider there to be multiple types of call-in and regular workers characterized by their training while keeping track of workloads in two jobs which adds considerably to the complexity of the problem. To model situations with more than two jobs we assume that worker's qualifications for the different jobs can be represented by a pyramid structure. In such a system, there is a skill hierarchy such that we can order J jobs from 1 to J with all workers who are qualified to work at job \( j \leq J \) are qualified to work at all jobs \( i \leq j \). It is our belief that in practice, in a firm with multiple jobs and types of workers, staffing decisions can be split up into multiple decisions each adequately supported by the use of either a two-job model or a pyramid model.

In chapter 2 we saw that there are distinct formulations for the various sub-
branches of the problem tree representing same-period vs. next-period notification and similarly for the various absenteeism scenarios. In looking at the multiple-job case we consider a more restrictive subset of possible problems. Namely, we model same-period and next-period notification without absenteeism, and with absenteeism type (i) \(^1\). When modeling the multiple-job problem we have several choices to make that were not present in the single-job branch of the problem tree. We must decide upon the mechanism for assigning crossstrained workers to jobs within a period, how to model the splitting of individual worker shifts, and how to differentiate productivity rates for different types of workers.

In the next section we discuss the general features of all the multiple-job models formulated in this chapter as well as the general assumptions and out justifications for them. Assumptions and modeling choices specific to individual formulations are discussed within the appropriate sections of this chapter.

### 3.1 General formulation and assumptions

In all models in this chapter we make some general assumptions. In addition, within each particular formulation we have specific assumptions which are discussed in the appropriate sections. Before formulating the problems we outline the general assumptions.

#### 3.1.1 Assumptions

**Work arrivals** Exogenous work arrivals each period are i.i.d random variables. This constraint can be relaxed as is described in the extensions to the 1-job problems.

\(^1\)Absenteism type (i) is when regular worker absenteeism is known before the call-in decision is made, and call-in absenteeism is discovered when the solicitations are being made. Therefore, once a call-in worker has agreed to come in they do with probability 1.
3.1 General formulation and assumptions

**Fungibility of hours:** We assume that work hours are fungible. By this we mean that we do not have to assign workers to jobs and/or overtime work in discrete blocks of time. This means that a crosstrained worker can work regular hours at one job and overtime at another. This also means that the models are formulated in such a way as to allow for situations in which crosstrained worker hours are divided between jobs in partial shifts. If there are \( J \) jobs in the system we can assume that assignments are made so that at most \( J - 1 \) workers have split shifts. Furthermore, the more workers inconvenienced in this way, the larger blocks of time that can be spent at each job. These observations together suggest that this assumption is not unreasonable.

**Overtime:** We bound overtime in the same way as we did in the 1-job models. The discussion of the overtime assumptions in Chapter 2 applies here as well.

**Productivity rates:** As in the 1-job models we do not consider any interactions between productivities of different workers, or of productivities changing over time. We also do not rely upon any hierarchy of productivity. There is wide debate on relative productivities of regular workers, contingent workers and overtime work. Some argue that contingent workers are more productive because they have fewer distractions at work.

We also make assumptions about the degree to which productivity is differentiated.

Each regular worker type \( i \) has a productivity \( \pi_{iw}^r \) at job \( i \).

Each call-in worker type \( i \) has a productivity \( \pi_{iw}^{ci} \) at job \( i \).

All workers performing overtime at job \( j \) have productivity \( \pi_j \).

All crosstrained workers of type \( i > j \) have productivity \( \pi_{ij}^{cf} \) at job \( j \).

These assumptions are made for simplification, as were the assignment assumptions. However they are not very limiting in the sense that it is hard to imagine a
workplace with more than four gradations of productivity for any well defined job. For each job \( j \) we have established a productivity rate for regular workers for whom job \( j \) is their primary job, for call-in workers for whom job \( j \) is their primary job, for workers who do not regularly work at job \( j \), and for workers who work at job \( j \) during overtime hours.

**Assignment rules:** The most general form of the two-job problem allows for 6 different types of workers and assumes that when all work has arrived to the system, crosstressed workers are assigned to tasks optimally. We might assume that crosstressed workers are assigned to a job in whole shift units of time and may only perform overtime work at the same job. Alternatively, we might assume that crosstressed overtime may be performed at a job other than the one at which the worker put in their regular hours. Another alternative has crosstressed hours split continuously between different jobs. There are many combinations of these three alternatives, each with its own level of complexity. Instead of an optimal assignment of crosstressed we can also follow a priority system in which one job has first access to the pool of available crosstressed workers and the lower priority job has access to only those crosstressed workers that have not been utilized by the higher priority job.

**Call-in protocol:** We make the same assumptions here about the protocol for call-in solicitation and absenteeism as in Chapter 2.

### 3.1.2 Set-up

The overall structure and terminology of the multiple-job problems are the same as the one-job problems. However, in order to represent the several distinct jobs we introduce the following notation:

\[ C_{\text{rw}}^i = \text{per-shift cost of a regular worker of type } i \text{ that combines the benefits and salary.} \]
3.1 General formulation and assumptions

$C^i_{cw} =$ per-shift cost of a call-in worker of type $i$ that combines the benefits and salary.

$C^j_{ot} =$ the premium paid per-worker per shifts worth of overtime worked at job $j$.

$C^j_f(\cdot) =$ a function for the per-time period penalty incurred by the firm for job $j$ work backlogged.

$C^j_B(\cdot) =$ a function for the penalty incurred by the firm for job $j$ work backlogged in the final period of the planning period.

$G_i =$ the fraction of the $V$ periods of work guaranteed to a call-in worker of type $i$.

$M_i =$ the number of call-in workers of type $i$ contracted for the planning period.

$N_i =$ the number of regular workers of type $i$ contracted for the planning period.

$\pi^{cw}_{ij} =$ the productivity rate of call-in worker type $i$ at job $j$.

$\pi^{rw}_{ij} =$ the productivity rate of regular worker type $i$ at job $j$.

$\pi_j =$ the productivity rate of all workers at job $j$ when performing overtime.

In the multiple-job problem there are potentially many different types of call-in and regular workers. Namely types 1 to $K$, where each type of worker is only qualified to work at some subset of the $J$ jobs in the system. The multiple-job problem can be stated as follows:

**Problem 3.1** What staffing level $S = (N, M)$ should a firm contract over a planning period of length $V$, to minimize the expected labor and backlog cost incurred when optimal call-in/overtime decisions are made dynamically each period, if the call-in workers are guaranteed at least a fraction $G$ respectively, of $V$, paid periods of work per planning period, where $N, M, \text{ and } G$ are each $K$-vectors.
3.2 Two-job problem

We formulate problem 3.1 as:

\[ \min_S \sum_{i \in [1,K]} \left[ C_{ri}^i N_i V + C_{ci}^i M_i G_i V + C_f(N_i + M_i) \right] + f_1^S(x_1, \kappa_1) \]

subject to:

\[ S \geq 0 \]

\[ g(N, M) \geq 0 \]

\[ S \text{ integer} \]

As in the one-job problems, the expression \( f_1^S(x_1, \kappa_1) \) is the optimal expected cost of the dynamic staffing decisions made over the planning horizon given a staffing level \( S \), where \( x_1 \) is a \( J \)-vector and \( \kappa_1 \) is a \( K \)-vector. Also, as in the one-job problems, \( g(N, M) \) is a constraint base upon the relative sizes of the call-in and regular workforces. We note that in this general statement of the multiple-job problem we have a \( 2K \) dimensional outer optimization problem. Furthermore the inner dynamic optimization problems are much more complex than in the one-job case\(^2\).

3.2 Two-job problem

In this section we formulate the dynamic programming problem embedded in problem 3.1 for the special case of a work environment with two jobs. As in chapter 2 we formulate same-period and next-period notification models with and without absenteeism. We only consider type (i) absenteeism in this chapter to narrow the discussion.

\(^2\)The degree of added complexity will vary depending upon the modeling approach taken.
3.2 Two-job problem

3.2.1 Same-period notification formulations

No absenteeism

The two job model in this case is almost exactly the same as in the one job case. The state space is five dimensional to keep track of the workload in two jobs and the remaining guarantee for the three types of call-in workers. In each stage we have four call-in decisions to make, i.e. how many of each type of worker to utilize and how to assign crosstrained workers. We must also make an assignment decision for crosstrained regular workers. As a result of our assumptions about productivity we aggregate overtime and therefore have two overtime decisions.

Definitions We define \( S_i = (N_i, M_i) \) to be the staffing of workers singly trained for job \( i = 1, 2 \), and \( S_3 = (N_3, M_3) \) to be the staffing of crosstrained workers.

We define a five dimensional state space: \( (x_t^1, x_t^2, \kappa_t^1, \kappa_t^2, \kappa_t^3) \) where:

\( \kappa_t^i \) is unused portion of the total type \( i \) call-in worker guarantee \( M_iG_iV \) at the end of stage \( t - 1 \).

\( x_t^j \) is defined to be the job \( j \) workload in the system at the start of stage \( t \).

We define the following random variable in each stage:

\( d_t^j \) is the amount of exogenous job \( j \) work arriving to the system at the beginning of stage \( t \).

Decisions In each stage \( t \) we make the decisions:

\( u_t^{ij} \), the number of type \( i \) call-in workers utilized in stage \( t \) for job \( j \).

\( r_t^{ij} \), the number of regular workers of type \( i \) that are utilized in stage \( t \) for job \( j \).

\( \omega_t^j \), the number of shifts of overtime at job \( j \) in stage \( t \).
3.2 Two-job problem

Constraints

\[ \sum_j u_{t}^{ij} \leq M_i \]

\[ \sum_j r_{t}^{ij} \leq N_i \]

\[ \omega_{t}^1 \leq OT_{\text{max}}(N_j + N_3, M_j + M_3) \]

\[ \omega_{t}^1 + \omega_{t}^2 \leq OT_{\text{max}}(N_1 + N_2 + N_3, M_1 + M_2 + M_3) \]

Where \( OT_{\text{max}}(N, M) \) is a function representing the maximum amount of overtime that can be performed by a complement of workers \((N, M)\). Note that lack of qualifications for a job, represented by a productivity of zero, implies that \( u_{t}^{12} = u_{t}^{21} = r_{t}^{12} = r_{t}^{21} = 0 \), \( \forall t \)

State Transitions

\[ x_{t+1}^{i} = [x_{t}^{i} - \sum_i (\pi_j \omega_{t}^i + \pi_i^{cu} u_{t}^{ij} + \pi_i^{rw} r_{t}^{ij})]^{+} + d_{t+1}^{i} \]

\[ \kappa_{t+1}^{i} = [\kappa_{t}^{i} - \sum_j u_{t}^{ij}]^{+} \]

Cost Functions To simplify notation we define the backlog in job \( j \) at the end of stage \( t \) as:

\[ b_{t}^{j} = [x_{t}^{j} - \sum_i (\pi_j \omega_{t}^{j} + \pi_i^{cu} u_{t}^{ij} + \pi_i^{rw} r_{t}^{ij})]^{+} \]

The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

\[ C_{V}(b_{V}) = \sum_{j=1}^{2} C_{B}^{j}(b_{V}^{j}) \]
The cost-to-go function in stage \( V \) is:

\[
f_V(\kappa_V, x_V) = \min_{r_V, u_V, \omega_V} \left\{ \sum_{j=1}^{2} \left[ C_{O}^{j} \omega_{t}^{j} + C_{B}^{j}(b_{t}^{j}) \right] + C_{cw}^{i} \left[ \sum_{j} (u_{t}^{ij}) - \kappa_{t}^{i} \right]^{+} \right\}
\]

and in stages \( t \):

\[
f_t(\kappa_t, x_t) = \min_{r_t, u_t, \omega_t} \left\{ \sum_{j=1}^{2} \left[ C_{O}^{j} \omega_{t}^{j} + C_{B}^{j}(b_{t}^{j}) \right] + C_{cw}^{i} \left[ \sum_{j} (u_{t}^{ij}) - \kappa_{t}^{i} \right]^{+} + E_{dt+1}[f_{t+1}(\kappa_{t+1}, x_{t+1})] \right\}
\]

**Absenteeism**

When we have absenteeism in the 2-job case we have a more complicated time-line of decisions than in the 1-job case, see figure 3-1. While we make the call-in solicitation decision after we know the workload but before we know the availability of the call-in workers it is possible that we could make the call-in worker assignments after having complete information about the system.

![One Period Diagram](image)

**CW = Call-in Workers**  **RW = Regular Workers**  **O.T. = Overtime**  **Abs = Absenteeism**  **Dec. = Decision**

Figure 3-1: Order of events for same-period notification with absenteeism case (i)

If we make call-in job assignments after accumulating all information we must keep track of the solicitation decision made before we knew the call-in availability and therefore expand the state space. With respect to regular workers, we may also
require keeping track of the crosstrained absenteeism since our assignment of workers may depend on their availability. If we follow some crosstrained worker priority assignment rule we do not have to keep track of this information. This is like making call-in job assignments when we solicit them so the call-in workers and managers know which worker is responsible to which manager. We call this an adaptive priority rule because it can set a different priority in each stage depending upon the state of the system. In this formulation, we assume the use of an adaptive priority rule that is adjusted when all system information is available, to reduce paid lost time. For example, we might set the priority for period $t$ such that 50 percent of the available crosstrained workers are assigned to each job. If the amount of work in job one turns out to be less than could be processed by the workers assigned to job 1 the excess crosstrained workers are reassigned to job two, resulting in more that 50 percent working at job two.

**Decisions** In odd stages $t$ we decide:

$u_i^t$, how many call-in workers of type $i$ to solicit for the following stage.

$\alpha_i^0$, the fraction of crosstrained call-in workers that actually arrive to assign to job 1.

In even stages $t$ we decide:

$\omega_j^t$, the amount of overtime to perform in job $j$ in stage $t$.

$\alpha_i^1$, the fraction of crosstrained regular workers that actually arrive to assign to job 1.

**Constraints**

$\omega_j^t \leq OT_{max}(\bar{n}_j, \bar{m}_j, \bar{n}_3, \bar{m}_3)$. 
3.2 Two-job problem

\[ \sum_j \omega^j_t \leq OT_{\text{max}}(\sum_i \bar{n}_i, \bar{m}_i) \]

\[ 0 \leq \alpha^0_t, \alpha^1_t \leq 1 \]

\[ u^i_t \leq M_i \]

I.e a standard limit on overtime has been set based upon the expected number of regular workers present and call-in workers available.

**State Transitions** For \( t + 1 \) odd:

\[ x^1_{t+1} = [d^1_{t+1} - [x^1_t - \pi^1_1 \omega^1_t]^+ - \pi^{rw}_{11} n^1_t - \pi^{rw}_{31} (\alpha^1_t n^3_t + Q^1_t)]^+ \]

\[ x^2_{t+1} = [d^2_{t+1} - [x^2_t - \pi^2_2 \omega^2_t]^+ - \pi^{rw}_{22} n^2_t - \pi^{rw}_{32} ((1 - \alpha^1_t)n^3_t + Q^2_t)]^+ \]

\[ \kappa^i_{t+1} = \kappa^i_t \]

\( Q^i_t \) is the contribution of crossstrained regular workers who are not needed at the other job and is computed as follows:

\[ Q^1_t = \frac{1}{\pi^{rw}_{32}} [\pi^{rw}_{32} (1 - \alpha^1_t)n^3_t - [d^2_{t+1} - [x^2_t - \pi^2_2 \omega^2_t]^+ - \pi^{rw}_{22} n^2_t]^+]^+ \]

\[ Q^2_t = \frac{1}{\pi^{rw}_{31}} [\pi^{rw}_{31} \alpha^1_t n^3_t - [d^1_{t+1} - [x^1_t - \pi^1_1 \omega^1_t]^+ - \pi^{rw}_{11} n^1_t]^+]^+ \]

For \( t + 1 \) even, we define \( a^i_t = \min[u^i_t, m^i_t] \) to simplify notation and have the following state transitions:

\[ x^1_{t+1} = [x^1_t - \pi^{cw}_{11} a^1_t - \pi^{cw}_{31}(\alpha^0_t a^3_t - Z^1_t)]^+ \]
\[ x_{t+1}^2 = [x_t^2 - \pi_{22}^{cw} a_t^2 - \pi_{32}^{cw} ((1 - \alpha_t^0) a_t^3 - Z_t^2)]^+ \]

\( Z_t^i \) is defined as the contribution of crosstrained call-in workers not needed at other job. Where,

\[ Z_t^1 = \frac{1}{\pi_{32}^{cw}}[(1 - \alpha_t^0) \pi_{32}^{cw} a_t^3 - [x_t^2 - \pi_{22}^{cw} a_t^2]^+]^+ \]

\[ Z_t^2 = \frac{1}{\pi_{31}^{cw}}[\alpha_t^0 \pi_{31}^{cw} a_t^3 - [x_t^1 - \pi_{11}^{cw} a_t^1]^+]^+ \]

\[ \kappa_t^i = [\kappa_t^i - a_{t+1}^i]^+ \]

**Cost Functions** The end of planning horizon cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

\[ C_{2V}(x_{2V} - \pi_j \omega_{2V}) = \sum_{j=1}^2 C_{2V}^j (x_{2V}^j - \pi_j \omega_{2V}^j) \]

The cost-to-go function in stage 2V is:

\[ f_{2V}(\kappa_{2V}, x_{2V}) = \min_{\omega_{2V}} \left\{ \sum_{j=1}^2 \left[ C_{2V}^j (x_{2V}^j + C_{2V}^j (x_{2V}^j - \pi_j \omega_{2V}^j)) \right] \right\} \]

and in even stages \( t \):

\[ f_t(\kappa_t, x_t) = \min_{\omega_t, \alpha_t^0} \left\{ \sum_{j=1}^2 \left[ C_{2V}^j (x_t^j + C_{2V}^j (x_t^j - \pi_j \omega_t^j)) \right] + E_{mt+1, dt+1} [f_{t+1}(\kappa_{t+1}, x_{t+1})] \right\} \]

and in odd stages \( t \):

\[ f_t(\kappa_t, x_t) = \min_{\omega_t, \alpha_t^0} \left\{ E_{mt+1} \left[ \sum_{i=1}^3 C_{cw}^i (x_t^i - \kappa_i^i)^+ + f_{t+1}(\kappa_{t+1}, x_{t+1}) \right] \right\} \]
3.2 Two-job problem

3.2.2 Next-period notification

As in the same-period notification formulations we only formulate the no-absenteeism and absenteeism type (i) problems for next-period notification.

No absenteeism

We define a five dimensional state space: \((x^1_t, x^2_t, \kappa^1_t, \kappa^2_t, \kappa^3_t)\) where:

- \(\kappa^i_t\) is unused portion of the total type \(i\) call-in worker guarantee \(MG_iV\) at the end of stage \(t - 1\).

- \(x^j_t\) is defined to be the job \(j\) workload in the system at the start of stage \(t\).

In each stage \(d^j_t\) is a random variable defining the amount of exogenous job \(j\) work arriving to the system in stage \(t\).

Decisions In stage \(t\) we decide:

- \(u^i_t\), how many call-in workers of type \(i\) to solicit for the following stage.

- \(\alpha_t\), the fraction of all crosstrained workers assigned to job 1.

- \(\omega^j_t\) the amount of overtime to perform in job \(j\) in stage \(t\).

Constraints

\[
\omega^j_t \leq OT_{max}(N_j, M_j, N_3, M_3).
\]

\[
\sum_j \omega^j_t \leq OT_{max}(\sum_i N_i, \sum_i M_i)
\]

\[
0 \leq \alpha_t \leq 1
\]

\[
u^i_t \leq M_i
\]

I.e a standard limit on overtime has been set based upon the expected number of regular workers present and call-in workers available.
3.2 Two-job problem

State Transitions  For stage $t + 1$:

$$x_{t+1}^1 = \left[ (x_t^1 - \pi_1 \omega_t^1)^+ + d_{t+1}^1 - \pi_{11}^w N_t^1 - \pi_{11}^{cu} u_t^1 - \pi_{31}^{ct} \alpha_t (N_3 + u_t^3) \right]^+ - Q_t^1$$

$$x_{t+1}^2 = \left[ (x_t^2 - \pi_2 \omega_t^2)^+ + d_{t+1}^2 - \pi_{22}^w N_t^2 - \pi_{22}^{cu} u_t^2 - \pi_{32}^{ct} (1 - \alpha_t) (N_3 + u_t^3) \right]^+ - Q_t^2$$

where,

$$Q_t^1 = \left[ \pi_{31}^{ct} (1 - \alpha_t) (N_3 + u_t^3) - (x_t^1 - \pi_1 \omega_t^1)^+ + d_{t+1}^1 - \pi_{22}^w N_t^2 - \pi_{22}^{cu} u_t^2 \right]^+$$

$$Q_t^2 = \left[ \pi_{32}^{ct} \alpha_t (N_3 + u_t^3) - (x_t^1 - \pi_1 \omega_t^1)^+ + d_{t+1}^1 - \pi_{11}^w N_t^1 - \pi_{11}^{cu} u_t^1 \right]^+$$

$$\kappa_{t+1}^i = [\kappa_t^i - u_{t+1}^i]^+$$

Cost Functions  The end of planning horizon cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

$$C_V(x_V - \pi \omega_V) = \sum_{j=1}^{2} C_B^j (x_V^j - \pi j \omega_V^j)$$

The cost-to-go function in stage $V$ is:

$$f_V(\kappa_V, x_V) = \min_{\omega_V} \left\{ \sum_{j=1}^{2} \left[ C_{at}^j \omega_{Vj}^j + C_{B}^j (x_V^j - \pi j \omega_V^j) \right] \right\}$$

and in stage $t$:

$$f_t(\kappa_t, x_t) = \min_{\omega_t, \alpha_t, u_t} \left\{ \sum_{j=1}^{2} \left[ C_{at}^j \omega_t^j + C_B^j (x_t^j - \pi j \omega_t^j) \right] + \sum_{i \in \{1, 2, 12\}} C_{cu}^i [u_t^i - \kappa_t^i]^+ + E_{dt+1} [f_{t+1}(\kappa_{t+1}, x_{t+1})] \right\}$$
3.2 Two-job problem

Absenteeism

As stated before, we are focusing on the next-day notification with absenteeism case (i). In this case, we assume that we make the call-in decision when we know the regular worker absenteeism but before we know the new work arrival. We then assume that we make the overtime decision when we know the call-in worker absenteeism and the workload situation. In this case, we are assuming that regular workers are giving some notice about their absence and that we know this information when we start to solicit call-in workers. We then solicit call-in workers until we find the amount we want or have exhausted the available ones. At this point in time, we know how many regular workers and call-in workers will actually be at work and how much work there is and make the overtime decision. The order of events is depicted in figure 3-2.

![Diagram of time periods with events]

**Figure 3-2: Order of events for next-period notification with absenteeism case (i)**

As in the same-period notification problem with absenteeism we use an adaptive priority assignment rule. Unlike the same-period case we do not reassign excess crosstrained regular workers in odd stages, because we do not know if they are excess or not, until the work arrival information becomes available.

We define a five dimensional state space: \((x_1^1, x_2^1, \kappa_1^1, \kappa_2^1, \kappa_3^1)\) where:
3.2 Two-job problem

\( \kappa_i^t \) is unused portion of the total type \( i \) call-in worker guarantee \( MG_iV \) at the end of stage \( t - 1 \).

\( x_i^t \) is defined to be the job \( j \) workload in the system at the start of stage \( t \).

We define the following random variables in each odd numbered stage:

\( n_i^t \) is the number of type \( i \) regular workers who are present in stage \( t \).

\( d_i^t \) is the amount of exogenous job \( j \) work arriving to the system in stage \( t \).

For the even-numbered stages we define the random variable: \( m_i^t \) as the number of type \( i \) call-in workers who are available in stage \( t \).

**Decisions** In odd stages \( t \) we decide:

\( u_i^t \), how many call-in workers of type \( i \) to solicit for the following stage.

\( \alpha_i^0 \), the fraction of crosstrained call-in workers that actually arrive to assign to job 1.

In even stages \( t \) we decide:

\( \omega_i^t \) the amount of overtime to perform in job \( j \) in stage \( t \).

\( \alpha_i^1 \), the fraction of crosstrained regular workers that actually arrive to assign to job 1.

**Constraints**

\[ \omega_i^t \leq OT_{max}(\bar{n}_j, \bar{m}_j, \bar{n}_3, \bar{m}_3). \]

\[ \sum_j \omega_i^t \leq OT_{max}(\sum_i \bar{n}_i, \bar{m}_i) \]

\[ 0 \leq \alpha_i^0, \alpha_i^1 \leq 1 \]
3.2 Two-job problem

\[ u_i^t \leq M_i \]

I.e a standard limit on overtime has been set based upon the expected number of regular workers present and call-in workers available.

**State Transitions**  For \( t + 1 \) odd:

\[
x_{t+1}^1 = [x_t^1 - \pi_1 \omega_t^1]^+ - \pi_{11}^r n_t^1 - \pi_{31}^r \alpha_t^1 n_t^3
\]

\[
x_{t+1}^2 = [x_t^2 - \pi_2 \omega_t^2]^+ - \pi_{22}^r n_t^2 - \pi_{32}^r (1 - \alpha_t^1) n_t^3
\]

\[
\kappa_{t+1}^i = \kappa_t^i
\]

Note: \( x_t^j \) may take on negative values for odd \( t \). If \( x_t^j \) is positive it means the backlogged work in the system exceeds the regular staff processing capacity. If \( x_t^j \) is negative it means there are more regular workers present than are needed for the backlog currently in the system and therefore the excess staff can work on the new work yet to arrive that period.

For \( t + 1 \) even, we define \( a_t^j = min[u_t^j, m_t^j] \) to simplify notation and have the following state transitions:

\[
x_{t+1}^1 = [x_t^1 + d_{t+1}^1 - \pi_{11}^{cw} a_t^1 - \pi_{31}^{cw} (\alpha_t^0 a_t^3 - Z_t^1)]^+
\]

\[
x_{t+1}^2 = [x_t^2 + d_{t+1}^2 - \pi_{22}^{cw} a_t^2 - \pi_{32}^{cw} ((1 - \alpha_t^0) a_t^3 - Z_t^2)]^+
\]

\( Z_t^j \) is defined as the contribution of crosstrained call-in workers not needed at
3.2 Two-job problem

other job. Where,

\[ Z_t^1 = \frac{1}{\pi_{32}^{c_w}} [(1 - \alpha_t^0) \pi_{32}^{c_w} a_{t}^3 - [x_t^2 + d_{t+1}^2 - \pi_{22}^{c_w} a_{t}^2]^+]^+ \]

\[ Z_t^2 = \frac{1}{\pi_{31}^{c_w}} [\alpha_t^0 \pi_{31}^{c_w} a_{t}^3 - [x_t^1 + d_{t+1}^1 - \pi_{11}^{c_w} a_{t}^1]^+]^+ \]

\[ \kappa_{t+1} = [\kappa_t - a_{t+1}]^+ \]

**Cost Functions**  The end of planning horizon cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

\[ C_{2V}(x_{2V} - \pi_j \omega_{2V}) = \sum_{j=1}^{2} C_{Bj}^j(x_{2V}^j - \pi_j \omega_{2V}^j) \]

The end of planning period cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

\[ C_{2V}(x_{2V} - \pi_j \omega_{2V}) = \sum_{j=1}^{2} C_{Bj}^j(x_{2V}^j - \pi_j \omega_{2V}^j) \]

The cost-to-go function in stage 2V is:

\[ f_V(\kappa_{2V}, x_{2V}) = \min_{\omega_{2V}} \left\{ \sum_{j=1}^{2} \left[ C_{Bj}^j(x_{2V}^j - \pi_j \omega_{2V}^j) + C_{2Vj}(x_{2V}^j - \pi_j \omega_{2V}^j) \right]^+ \right\} \]

and in even stages \(t\):

\[ f_t(\kappa_t, x_t) = \min_{\omega_t, \alpha_t} \left\{ \sum_{j=1}^{2} \left[ C_{Bj}^j(x_t^j - \pi_j \omega_t^j)^+ \right] + En_{t+1}[f_{t+1}(\kappa_{t+1}, x_{t+1})] \right\} \]


and in odd stages $t$:

$$f_t(\kappa_t, x_t) = \min_{u_t, \alpha_t} \left\{ E_{m_{t+1}, d_{t+1}} \left[ \sum_{i \in [1, 2, 12]} C_{cu}^i [a_i^t - \kappa_i^t]^+ + f_{t+1}(\kappa_{t+1}, x_{t+1}) \right] \right\}$$

### 3.3 Pyramid problem

In the pyramid training structure model, we assume a worker of type $i$ is trained to perform all jobs $j \leq i$, where $i, j \in [1, J]$, $J$ being the number of jobs in the system. We also assume that assignments to tasks within a period follow the same hierarchy, meaning that if job $i$ is understaffed, we first assign workers of type $j + 1$ to assist at job $j$, if they are available, before any other workers. If we elect to assign more workers we assign type $j + 2$ workers etc. In this scheme each worker has a primary responsibility to a particular task and will support efforts in a task further "down" the training hierarchy only when staffing exceeds workload at the primary job.

These assumptions considerably simplify the models of multiple-jobs and are realistic as well. In a pyramid training structure the workers with the most skills are the most highly paid and with the highest prestige, and therefore would be reluctant to perform tasks that many worker types are qualified for, or lower prestige jobs. For example, in a hospital, a doctor might be qualified to perform almost all nursing tasks besides her own specialized skills. This does not mean that she will be asked to take a patients temperature if there are nurses available to do this.

#### 3.3.1 Same-period notification

**No Absenteeism**

We define a $2J$ dimensional state space, where $J$ is the number of distinct jobs in the system: $(x_t, \kappa_t)$ where:
3.3 Pyramid problem

κ<sub>t</sub> is a vector of unused portions of the total call-in worker guarantees MG<sub>i</sub>V at the end of stage t − 1 for each of the J different worker types.

x<sub>t</sub> is defined to be the vector of workloads in the different jobs in the system at the start of stage t.

In each stage d<sub>t</sub> is a random variable defining the amount of exogenous job j work arriving to the system in stage t. d<sub>t</sub> is the corresponding vector.

**Decisions** In stage t we decide:

u<sub>t</sub>, how many call-in workers of type i to solicit for the current stage.

ω<sub>t</sub> the amount of overtime to perform in job j in stage t.

**Constraints**

ω<sub>t</sub> ≤ OT<sub>max</sub>(Σ<sub>i≥j</sub> N<sub>i</sub>, Σ<sub>i≥j</sub> M<sub>i</sub>).

Σ<sub>i≥j</sub> ω<sub>t</sub> ≤ OT<sub>max</sub>(Σ<sub>i≥j</sub> N<sub>i</sub>, Σ<sub>i≥j</sub> M<sub>i</sub>) ∀ j.

u<sub>t</sub> ≤ M<sub>i</sub>

**State Transitions** To simplify the expressions in this formulation we define the following notation: e<sub>t</sub><sup>j−1</sup> is the number of excess higher skilled workers available in stage t for job j − 1. That is:

\[ e_t^{j-1} = [e_t^j + N_j + u_t^j - "demand for j"]^+ \]

More formally this is:

\[ e_t^{j-1} = \frac{1}{\pi_j^{rw}}[\pi_j^{rw} N_j - x_t^j]^+ + \frac{1}{\pi_j^{cw}}[\pi_j^{cw} u_t^j - [x_t^j - \pi_j^{rw} N_j]^+]^+ + \frac{1}{\pi_j^{ct}}[\pi_j^{ct} e_t^j - [x_t^j - \pi_j^{rw} N_j - \pi_j^{cw} u_t^j]^+]^+ \]
The backlog in job \( j \) for at the end of stage \( t \) is defined as:

\[
b_t^j = [x_t^j - \pi^w_j N_j - \pi^u_j u_t^j - \pi^e_t e_t^j - \pi^\omega_j \omega_t^j]^+ \]

We can now write the state transitions as for stage \( t + 1 \):

\[
x_{t+1}^i = b_t^i + d_{t+1}^i
\]

\[
\kappa_{t+1}^i = [\kappa_t^i - u_{t+1}^i]^+
\]

**Cost Functions** The end of planning horizon cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

\[
C_V(b_V) = \sum_{j=1}^{J} C_B^j(b_V^j)
\]

The cost-to-go function in stage \( V \) is:

\[
f_V(\kappa_V, x_V) = \min_{\omega_V, u_V} \left\{ \sum_{j=1}^{J} \left[ C_B^j(u_V^j) + \sum_{i \leq j} C_c^i [u_V^i - \kappa_V^i]^+ \right] \right\}
\]

and in stage \( t \):

\[
f_t(\kappa_t, x_t) = \min_{\omega_t, u_t} \left\{ \sum_{j=1}^{J} [C_B^j(u_t^j) + C_c^i [u_t^i - \kappa_t^i]^+ + E_{t+1}^i f_{t+1}(\kappa_{t+1}, x_{t+1})] \right\}
\]

Where, \( \omega_t, u_t, \kappa_t, \) and \( x_t \) are all vectors.

**Absences**

With absenteeism we again have to split each stage into two parts. We define a \( 2J \) dimensional state space, where \( J \) is the number of distinct jobs in the system: \((x_t, \kappa_t)\)
where:

\( r_t \) is a vector of unused portions of the total call-in worker guarantees \( MG,V \) at the end of stage \( t - 1 \) for each of the \( J \) different worker types.

\( x_t \) is defined to be the vector of workloads in the different jobs in the system at the start of stage \( t \).

In each stage \( d^j_t \) is a random variable defining the amount of exogenous job \( j \) work arriving to the system in stage \( t \). \( d_t \) is the corresponding vector.

**Decisions** In stage \( t \) we decide:

\( u^i_t \), how many call-in workers of type \( i \) to solicit for the next stage if \( t \) is odd.

\( \omega^j_t \) the amount of overtime to perform in job \( j \) in stage \( t \) if \( t \) is even.

**Constraints**

\[ \omega^j_t \leq OT_{\max}(\sum_{i \geq j} N_i, \sum_{i \geq j} M_i). \]

\[ \sum_{i \geq j} \omega^i_t \leq OT_{\max}(\sum_{i \geq j} N_i, \sum_{i \geq j} M_i) \forall j. \]

\[ u^i_t \leq M_i \]

**State Transitions** To simplify the expressions in this formulation we define the following notation for odd stages \( t \):

\[ e^{i-1}_t = \frac{1}{\pi_j^{\text{rw}}} [0 - x^j_t]^+ + \frac{1}{\pi_j^{\text{cw}}} [\pi_j^{\text{cw}} a^j_t - [x^j_t]^+]^+ + \frac{1}{\pi_j^{\text{ct}}} [\pi_j^{\text{ct}} e^j_t - [x^j_t]^+ - \pi_j^{\text{cw}} a^j_t]^+]^+ \]

\( e^{i-1}_t \) represents the number of excess higher skilled workers available in stage \( t \) for job \( j - 1 \).
We can now write the state transitions as for odd stages $t + 1$:

$$x_{t+1}^j = [x_t^j - \omega_t^j \pi_j]^+ + d_{t+1}^j - \pi_j^{rw} n_t^j$$

$$\kappa_{t+1}^i = \kappa_t^i$$

For even stages $t + 1$:

$$x_{t+1}^j = [x_t^j - \pi_j^{cw} a_t^j - \pi_t^e b_t^j]^+$$

$$\kappa_{t+1}^i = [\kappa_t^i - a_{t+1}^i]^+$$

Cost Functions  The end of planning horizon cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

$$C_{2V}(x_{2V} - \pi \omega_{2V}) = \sum_{j=1}^2 C_B^j(x_{2V}^j - \pi_j \omega_{2V}^j)$$

The cost-to-go function in stage $2V$ is:

$$f_V(\kappa_{2V}, x_{2V}) = \min_{\omega_{2V}} \left\{ \sum_{j=1}^2 \left[ C_{oV}^j \omega_{2V}^j + C_B^j(x_{2V}^j - \pi_j \omega_{2V}^j) \right] \right\}$$

In even stages $t$:

$$f_t(\kappa_t, x_t) = \min_{\omega_t^j} \left\{ \sum_{j=1}^J \left[ C_{oV}^j \omega_t^j + C_B^j(x_t^j - \pi_j \omega_t^j) \right] + E_{d_{t+1}n_{t+1}}[f_{t+1}(\kappa_{t+1}, x_{t+1})] \right\}$$

In odd stages $t$:

$$f_t(\kappa_t, x_t) = \min_{u_t} \left\{ E_{m_{t+1}} \left[ \sum_{i \leq j} C_{cu}^i [a_t^i - \kappa_t^i]^+] + f_{t+1}(\kappa_{t+1}, x_{t+1}) \right\}$$
3.3 Pyramid problem

3.3.2 Next-period notification

No Absenteeism

We define a $2J$ dimensional state space, where $J$ is the number of distinct jobs in the system: $(x_t, \kappa_t, \kappa_t, \kappa_t)$ where:

- $\kappa_t$ is a vector of unused portions of the total call-in worker guarantees $MG_i V$ at the end of stage $t - 1$ for each of the $J$ different worker types.

- $x_t$ is defined to be the vector of workloads in the different jobs in the system at the start of stage $t$.

In each stage $d^j_t$ is a random variable defining the amount of exogenous job $j$ work arriving to the system in stage $t$. $d_t$ is the corresponding vector.

Decisions In stage $t$ we decide:

- $u^i_t$, how many call-in workers of type $i$ to solicit for the current stage.

- $\omega^j_t$ the amount of overtime to perform in job $j$ in stage $t$.

Constraints

- $\omega^j_t \leq OT_{\text{max}}(\sum_{i \geq j} N_i, \sum_{i \geq j} M_i)$.

- $\sum_{i \geq j} \omega^i_t \leq OT_{\text{max}}(\sum_{i \geq j} N_i, \sum_{i \geq j} M_i) \forall j$.

- $u^i_t \leq M_i$

State Transitions To simplify the expressions in this formulation we define the following notation:

$$e_t^{-1} = \frac{1}{\pi_t^{w}} [0 - x_t^j]^+ + \frac{1}{\pi_t^{w}} [\pi_t^{rw} N_j - [x_t^j]^+]^+ + \frac{1}{\pi_t^{ct}} [\pi_t^{ct} e_t^j - [e_t^j]^+ - \pi_t^{cw} N_j]^+]^+$$
3.3 Pyramid problem

e_t^{i-1} represents the number of excess higher skilled workers available in stage $t$ for job $j - 1$. The backlog in job $j$ for at the end of stage $t$ is defined as:

$$b_t^i = [x_t^i - \pi_{j}^{ct} N_j - \pi_{j}^{ct} c_t^i - \pi_{j}^{cw} \omega_t^i]^+$$

We can now write the state transitions as for stage $t + 1$:

$$x_{t+1}^i = b_t^i + d_{t+1}^i - \pi_{j}^{cw} u_t^i$$

$$\kappa_{t+1}^i = [\kappa_t^i - u_{t+1}^i]^+$$

**Cost Functions** The end of planning horizon cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

$$C_V(b_V) = \sum_{j=1}^{J} C_B^j(b_V^j)$$

The cost-to-go function in stage $V$ is:

$$f_V(\kappa_V, x_V) = \min_{\omega_V, u_V} \left\{ \sum_{j=1}^{J} \left[ C_B^j (b_V^j) + \sum_{i \leq j} C_{cw}^i [u_V^i - \kappa_V^i]^+ \right] \right\}$$

and in stage $t$:

$$f_t(\kappa_t, x_t) = \min_{\omega_t, u_t} \left\{ \sum_{j=1}^{J} \left[ C_B^j (b_t^j) + \sum_{i \leq j} C_{cw}^i [u_t^i - \kappa_t^i]^+ + E_{d_{t+1}} [f_{t+1}(\kappa_{t+1}, x_{t+1})] \right] \right\}$$

**Absenteeism**

With absenteeism we again have to split each stage into two parts. We define a $2J$ dimensional state space, where $J$ is the number of distinct jobs in the system: $(x_t, \kappa_t)$ where:
3.3 Pyramid problem

$\kappa_t$ is a vector of unused portions of the total call-in worker guarantees $MG_iV$ at the end of stage $t - 1$ for each of the $J$ different worker types.

$x_t$ is defined to be the vector of workloads in the different jobs in the system at the start of stage $t$.

In each stage $d_t^j$ is a random variable defining the amount of exogenous job $j$ work arriving to the system in stage $t$. $d_t$ is the corresponding vector.

**Decisions** In stage $t$ we decide:

$u_t^i$, how many call-in workers of type $i$ to solicit for the next stage if $t$ is odd.

$\omega_t^j$ the amount of overtime to perform in job $j$ in stage $t$ if $t$ is even.

**Constraints**

$\omega_t^j \leq OT_{\text{max}}(\Sigma_{i \geq j} N_j, \Sigma_{i \geq j} M_j)$.

$\Sigma_{i \geq j} \omega_t^i \leq OT_{\text{max}}(\Sigma_{i \geq j} N_i, \Sigma_{i \geq j} M_i) \forall j$.

$u_t^i \leq M_i$

**State Transitions** To simplify the expressions in this formulation we define the following notation for odd stages $t$:

$e_t^{j-1} = \frac{1}{\pi_{j}^{rw}}[0 - x_t^j]^+ + \frac{1}{\pi_{j}^{cw}}[\pi_{j}^{cw}a_t^j - [x_t^j]^+]^+ + \frac{1}{\pi_{j}^{ct}}[\pi_{j}^{ct}c_t^j - [[x_t^j]^+ - \pi_{j}^{cw}a_t^j]^+]^+$

$e_t^{j-1}$ represents the number of excess higher skilled workers available in stage $t$ for job $j - 1$.

We can now write the state transitions as for odd stages $t + 1$:

$x_{t+1}^i = [x_t^i - \omega_t^i\pi_j]^+ - \pi_{j}^{rw}n_t^i$
\[ \kappa_{t+1}^i = \kappa_t^i \]

For even stages \( t + 1 \):

\[ x_{t+1}^i = \left[ x_t^i + d_{t+1}^i - \pi_j^{cw} a_t^i - \pi_t^{ct} e_j \right]^+ \]

\[ \kappa_{t+1}^i = [\kappa_t^i - a_{t+1}^i]^+ \]

**Cost Functions** The end of planning horizon cost (or terminal value function) is used to allow a different cost for backlog remaining at the end of the planning period and is defined as:

\[ C_{2V}(x_{2V} - \pi_{2V}) = \sum_{j=1}^{2} C_B^j (x_{2V}^{j} - \pi_j \omega_{2V}^{j}) \]

The cost-to-go function in stage \( 2V \) is:

\[ f_V(\kappa_{2V}, x_{2V}) = \min_{\omega_{2V}} \left\{ \sum_{j=1}^{2} \left[ C_{oV}^j \omega_{2V}^j + C_B^j (x_{2V}^{j} - \pi_j \omega_{2V}^{j}) \right] \right\} \]

In even stages \( t \):

\[ f_t(\kappa_t, x_t) = \min_{\omega_t} \left\{ \sum_{j=1}^{J} \left[ C_{oV}^j \omega_t^j + C_B^j (x_t^j - \pi_j \omega_t^j) \right] + E_{nt+1} [f_{t+1}(\kappa_{t+1}, x_{t+1})] \right\} \]

In odd stages \( t \):

\[ f_t(\kappa_t, x_t) = \min_{\omega_t} \left\{ E_{dt+1, m_{t+1}} \left[ \sum_{i \leq J} C_{cw}^i [a_t^i - \kappa_t^i]^+ + f_{t+1}(\kappa_{t+1}, x_{t+1}) \right] \right\} \]
3.4 Special cases

Focusing on the 2-job problem formulation, there are several special cases of interest that can change how we approach the problems.

No crosstraining: If there are no crosstrained workers we can clearly separate the two jobs into two separate problems.

Full crosstraining: If all workers are crosstrained we can combine the two job streams into one and consider the problem a 1-job problem under certain conditions. The conditions are that the ratio of the backlog penalties for each job stream should be the same as the ratios of the productivity rate of each type of worker at each job. I.e., for some constant $\gamma$:

$$\frac{C^1_k}{C^2_0} = \frac{\pi^{cw}_{31}}{\pi^{cw}_{32}} = \frac{\pi^{tw}_{31}}{\pi^{tw}_{32}} = \frac{\pi_{31}}{\pi_{32}} = \gamma$$

If these conditions prevail then we can define a combined arrival stream with work arrivals in each period $t$ defined as:

$$d^3_t = \gamma d^1_t + d^2_t$$

No call-in workers: If we can consider a work environment with only functional flexibility then the staffing decision each period is how to assign workers to job and how much overtime to use. If we make the further assumption that either overtime is never used or it is always used in periods where workload exceeds regular staff, we can view the work environment as a multi-class queueing system. We discuss this relation to queueing problems in Chapter 4.
3.5 Summary

In this chapter we have formulated models of functional and working time flexibility in work environments with multiple job types. We have presented a very general model of the case when there are 2 jobs and a specially structured model for 2 or more jobs. Following the structure of the problem tree in Figure 1-2 we have distinct formulations for same vs. next-period notification regimes, with and without absenteeism. The models all involve the solution of the problem 3.1, an optimization over six dimensions. The value of the objective function at a particular solution \( S(N, M) \), where \( N \) and \( M \) are vectors, is the expected cost of the optimal solution to a finite horizon dynamic program. All the dynamic programming formulations are based upon a Markovian structure that governs the work in the system and the cumulative utilization of call-in workers as in the 1-job problem.
Chapter 4

Computational analysis of models

In chapters 2 and 3 we formulated a family of mathematical problems, based upon models of work force flexibility in stochastic work environments. These models have represented workforce management decision making at two levels with a great degree of realism. As a result some of the related mathematical problems are quite complex and their solution computationally burdensome. Our primary goal in developing these models was to gain an understanding of the dynamics of utilizing flexible workers of different kinds and what factors influence staffing decisions in work environments with flexible workers. This goal is achieved in Chapter 5 through series of numerical experiments. A secondary goal of this modeling effort is to provide managers with a set of decision support tools to aid in managing flexible workforces. To create practical tools the mathematical problems underlying the decision making must be computationally tractable. In this chapter, we present results that contribute to making the solution of the staffing problems computationally tractable.

We have formulated the problems as two nested optimization problems. The inner problem is the optimal, in an expected value sense, dynamic allocation of preset staffing resources over a finite planning horizon. The outer problem is the determination of the staffing resources to have available to minimize the cost of the inner
problem plus the cost of maintaining such a workforce.

In this chapter, we develop results about both these problems. For the inner optimization we find ways to characterize the optimal staff utilization decisions that reduce work involved in making optimal staffing decisions within each stage of the dynamic programming problem. Based on the analysis of the inner optimization problem we are able to prove that in the continuous version of the problem the objective function of the outer optimization is convex in the staffing levels $S$. We are able to do this for both problems $P1$ and $P2$. This allows us to develop search procedures that can greatly reduce the computational work of the outer optimization.

Clearly, there is more analysis that can be done of these problems. We also suspect that the best prospects for practical decision support tools lie in different formulations of the models. We believe this to be the case because while we have been able to develop techniques for speeding up the optimization processes at both levels, we are still faced with the problem of large state spaces in the dynamic programming engine. This is not really an important factor in the 1-job problem but in the 2-jobs problems it is. Nonetheless, the models formulated here are very detailed and realistic, therefore they serve as important points of comparison for other representations of the staffing problems defined in this thesis. Therefore, we want alternative formulations of the problems to have similar structural properties to the models presented here. This means that even if the results derived in this chapter do not lead to more computationally tractable solution approaches they provide us with a way to validate

---

1For example, let’s consider a 2-job problem where all workers have a productivity of 1 unit of work per shift, and the work load arrival distribution ranges from 0 units to 15 units of work in each job. Allowing for backlog we allow work in the system in each job to range from 0 to 30 units. Let’s assume we wanted to determine the cost of running such a system with 5 singly trained call-in workers in each job and 15 cross-trained call-in workers all with a guarantee $G = .5$ over 20 time periods. In this case the state space would have: $(32)(32)(.5)(20)(5)[(5)(20)(5)][(5)(20)(5)][(5)(20)(15)] = 3.84e + 8$ states. If there is a difference between productivity rates for different types of workers we must increase the resolution of the dimensions representing work in the system and thus further increase the size of the state space. We also note that modeling a variety of types of call-in workers also greatly increases the state space.
alternative formulations.

In the first and second sections we analyze the 1 and 2-job problems respectively. In the analyses we prove the convexity of the inner and outer-optimization problems. For each problem class we also characterize optimal staffing decisions for special cost structures, and propose some search strategies for the outer-optimization problems. In the third section, we relate the problems to other mathematical problems in the literature. Finally, in the fourth section we discuss the areas for future research on the computational side of the problems formulated in this thesis.

4.1 One-job problems

In this section we focus on the inner and outer optimization problems for the 1-job problems. We show that convexity properties of the inner-optimization problem can be carried over to the outer optimization. We then make some observations about the outer optimization and use them to develop a search algorithm for the optimal staffing level.

4.1.1 Inner optimization

For the same-period notification problems we show that for special cases of the parameter settings, determined by the relative cost-productivity ratios of the different types of labor, we can characterize the optimal decisions made within each stage. We then show that in general, when we consider the continuous variable versions of the models the cost-to-go functions in each stage are convex in the state variables, for both same-period and next-period notification.

Special cases for same-period notification If we consider different relative cost structures we can characterize the control decisions made each period. We assume that the backlog costs in each period, including the final period are linear and, equal.
4.1 One-job problems

We also assume that exogenous work arrivals are iid. Within the dynamic programming problem there are then three cost parameters, $C_{cw}, C_{ot},$ and $C_b$. Their relative values will determine the control decisions made each period. In this analysis we will assume all the state and control variables to be continuous. Those results that are independent of the continuous approximation are noted.

\[
\frac{C_{cw}}{\pi_{cw}} \leq \frac{C_{ot}}{\pi_{ot}} \leq C_b
\]

In this case, in each period in which $x_t$ exceeds $N$ it is optimal to use as many call-in workers and overtime workers as are available to eliminate backlog. Furthermore, we always use all the call-in workers available before utilizing any overtime. This means in each such period $t$, $u_t = \text{Min}[M, x_t - N\pi_{rw}]$ and $\omega_t = \text{Min}[\text{OT}_{\text{max}}(N, M), x_t - N\pi_{rw} - u_t\pi_{cw}]$. We notice that this cost structure results in the control decisions being only dependent upon the state variable $x_t$.

\[
\frac{C_{ot}}{\pi_{ot}} \leq \frac{C_{cw}}{\pi_{cw}} \leq C_b
\]

In this case, in each period in which $x_t$ exceeds $N$, it is optimal to use as many call-in workers and overtime workers as are available to eliminate backlog. However, the decision to use overtime versus call-in workers is dependent upon the state variable $\kappa$. When $\kappa > 0$ we can utilize up to $\text{Min}[M, \kappa]$ call-in workers at no cost. On the other hand when $\kappa = 0$ the call-in workers are less cost effective than overtime so we utilize all the allowable overtime shifts before call-in shifts beyond the guarantee.

The above results are dependent, in part, upon the observation that in the same-period notification problem we do not 'hoard' the guaranteed call-in worker shifts. In other words, anytime there is an opportunity to utilize call-in workers at no cost we do so. The reasons for this phenomenon are: 1) in a finite horizon problem early backlog has potentially greater costs than later backlog, and 2) we never utilize call-in workers unless work is available for them but there is a risk of not utilizing all the shifts paid for by the guarantee.
4.1 One-job problems

\[
\frac{C_{ot}}{\pi_{ot}} \leq C_b \leq \frac{C_{cw}}{\pi_{ot}} \quad \text{In this case, we operate in a similar fashion as the previous as long as we are utilizing guaranteed call-in shifts. As soon as } \kappa = 0 \text{ we shift into an operating mode in which we always use as much overtime as allowed to prevent backlog but only use call-in workers when the backlog threatens to exceed a certain threshold.}
\]

\[
\frac{C_{cw}}{\pi_{cw}} \leq C_b \leq \frac{C_{ot}}{\pi_{ot}} \quad \text{In this case, it is optimal to use as many call-in workers as are available to eliminate backlog. This means in each such period } t. \ u_t = \text{Min}[M, x_t - N\pi_{rw}]. \text{ We only utilize overtime if the backlog threatens to exceed a certain threshold.}
\]

\[
C_b \leq \frac{C_{ot}}{\pi_{ot}} \leq \frac{C_{cw}}{\pi_{cw}} \quad \text{In this case, the only simple statement we can make about staffing decisions is that we do not 'hoard' the guaranteed shifts and we always use overtime shifts before non-guarantee call-in hours. When the cost hierarchy is: } C_b \leq \frac{C_{cw}}{\pi_{cw}} \leq \frac{C_{ot}}{\pi_{ot}} \text{ we always use call-in shifts before overtime, even non-guaranteed shifts.}
\]

**General results** We saw above, in the cases where the unit cost of backlog is less than the costs of processing a unit of work with call-in workers or overtime, it is not possible to characterize optimal staffing decisions in a simple way. Our approach to analyzing the general same-period problems is to note that the same-period notification without absenteeism is a special case of the same-period problem with absenteeism and probability of absence of all workers equal to 0. Therefore, showing the convexity of the problem of the absenteeism problem is sufficient to demonstrate it for the no-absenteeism case.

We also note that the next-period with absenteeism formulation only differs, from the same-period one, by the stage in which an expected value over exogenous work arrivals takes place. In the analysis of the next-period problems, it becomes apparent that this distinction has no bearing upon the demonstration of convexity. This implies that demonstrating convexity for the next-period with absenteeism problems is
sufficient for demonstrating it for all the variations of the 1-job problem. This result is not surprising since we expect that the should be diminishing returns to increases in staff size. However, in the numerical results chapter we show how the behavior of the system in terms of backlog and call-in usage is significantly different in the same-period versus next-period regimes. Interestingly, these differences to not effect the convexity of the inner optimization cost functions in terms of the state variables \((x, \kappa)\).

We use the problem with absenteeism type (i) to generate our results about the mathematical properties of these problems. It should be clear that similar types of analysis can be performed for the models of the other types of absenteeism.

**Preliminaries** We first review some observations and facts that we use in the proofs of the convexity results.

**Proposition 4.1** If \(H(X, Y)\) is convex then \(L(X) = \min_{Y \in U} \{H(X, Y)\}\) is convex for a vectors \(X\) and \(Y\) and convex set \(U\).

**Proof** Define \((X_3, Y_3) = \lambda_1(X_1, Y_1) + \lambda_2(X_2, Y_2)\) with \(\lambda_2 = 1 - \lambda_1\), to be a convex combination of \((X_1, Y_1)\) and \((X_2, Y_2)\). Also define \(Y_i^*\) to be the \(Y\) that minimizes \(H(X_i, Y)\). By the convexity of of \(H\) we have that:

\[
H(X_3, Y_3) \leq \lambda_1 H(X_1, Y_1) + \lambda_2 H(X_2, Y_2)
\]

Therefore,

\[
H(X_3, \lambda_1 Y_1^* + \lambda_2 Y_2^*) \leq \lambda_1 H(X_1, Y_1^*) + \lambda_2 H(X_2, Y_2^*)
\]

Because \(U\) is a convex set \(\lambda_1 Y_1^* + \lambda_2 Y_2^*\) is a feasible value for \(Y\). We also know that \(H(X_3, Y_3^*) \leq H(X_3, \lambda_1 Y_1^* + \lambda_2 Y_2^*)\) by the minimization properties. Therefore,

\[
H(X_3, Y_3^*) \leq \lambda_1 H(X_1, Y_1^*) + \lambda_2 H(X_2, Y_2^*)
\]
4.1 One-job problems

Proving the convexity of $L(Y)$.

We also observe that the state transitions modelled throughout the dynamic programming formulations in the thesis are homogeneous mappings $g(x)$ such that $\lambda g(x) = g(\lambda x)$. Furthermore these transitions all involve taking the positive part of an expression, i.e. $g(x) = \max(h(x), 0)$ as a result they are subadditive. By subadditive we mean $g(x) + g(y) \leq g(x + y)$. Homogeneous mappings, otherwise known as gauge functions, are convex if and only if they are subadditive, see [Baz93].

Another important observation is that the cost functions within the dynamic programming formulations have the following properties: (1) $f_t(x_t, \kappa_t)$ is increasing in $x_t$, and (2) $f_t(x_t, \kappa_t)$ is decreasing in $\kappa_t$. Observation (1) says that the more work in the system the more expensive it is to operate. Observation (2) says that the more guaranteed shifts of call-in workers unused the lower the operating costs within the dynamic optimization.

Results

**Lemma 4.1** $f_{2V}(x_{2V}, \kappa_{2V})$ is a convex function of $(x_{2V}, \kappa_{2V})$.

**Proof** The final stage, $2V$ is an even stage in which we make an overtime decision.

When, $C_B \leq \frac{C_{cw}}{\pi_{cw}} \leq \frac{C_{ot}}{\pi_{ot}}$ or $C_B \leq \frac{C_{ct}}{\pi_{ot}} \leq \frac{C_{cw}}{\pi_{cw}}$ or $\frac{C_{ct}}{\pi_{cw}} \leq C_B \leq \frac{C_{ot}}{\pi_{ot}}$ we never utilize overtime:

$$\Rightarrow f_{2V}(x_{2V}, \kappa_{2V}) = C_B x_{2V} \text{ a convex function}$$

When, $\frac{C_{cw}}{\pi_{cw}} \leq \frac{C_{ot}}{\pi_{ot}} \leq C_B$ or $\frac{C_{ct}}{\pi_{ot}} \leq \frac{C_{cw}}{\pi_{cw}} \leq C_B$ or $\frac{C_{ct}}{\pi_{ot}} \leq C_B \leq \frac{C_{cw}}{\pi_{cw}}$ we use as much overtime as we can to prevent any backlog.

$$\Rightarrow f_{2V}(x_{2V}, \kappa_{2V}) = C_{ot} \min\{\frac{OT_{max}}{\pi_{ot}}, \frac{x_{2V}}{\pi_{ot}}\} + C_B [x_{2V} - \frac{OT_{max}}{\pi_{ot}}]^+ \text{ which is a convex function}$$

\[ \blacksquare \]. Note that $f_{2V}(x_{2V}, \kappa_{2V})$ is independent of $\kappa_{2V}$ since we only make overtime decisions and this is the final stage.
4.1 One-job problems

**Lemma 4.2** $f_{2V-1}(x_{2V-1}, \kappa_{2V-1})$ is a convex function of $(x_{2V-1}, \kappa_{2V-1})$.

**Proof** In stage $t = 2V - 1$, an odd stage, we make call-in utilization decisions. When, $C_B \leq \frac{C_{cw}}{\pi_{cw}} \leq \frac{C_{ot}}{\pi_{ot}}$ or $C_B \leq \frac{C_{ot}}{\pi_{ot}} \leq \frac{C_{cw}}{\pi_{cw}}$ or $\frac{C_{ot}}{\pi_{ot}} \leq C_B \leq \frac{C_{cw}}{\pi_{cw}}$ we do not use any more call-in workers than we have guaranteed remaining since $\frac{C_{cw}}{\pi_{cw}} > C_B$. This means we try to utilize $u^* = \min(\kappa_t, M)$ call-in workers, and actually use $a = \min(u^*, m)$, where $m$ is the number of call-in workers who are available.

$$f_t(x_t, \kappa_t) = C_b E_d[a][f_{2V}([x_t + d - a\pi_{cw}]^+, [\kappa_t] - a]^+]$$

a convex function in $(x_t, \kappa_t)$.

When, $\frac{C_{cw}}{\pi_{cw}} \leq \frac{C_{ot}}{\pi_{ot}} \leq C_B$ or $\frac{C_{ot}}{\pi_{ot}} \leq \frac{C_{cw}}{\pi_{cw}} \leq C_B$ or $\frac{C_{cw}}{\pi_{cw}} \leq C_B \leq \frac{C_{ot}}{\pi_{ot}}$ we use as many call-in workers as we can to prevent backlog:

$$f_t(x_t, \kappa_t) = \min\{C_{cw}[u_t - \kappa_t]^+ + E_d[a][f_{2V}([x_t + d - a\pi_{cw}]^+, \kappa_{t+1})]\}$$

We can rewrite $f_t(x_t, \kappa_t)$ as

$$\min_{u_t} H(x_t, \kappa_t, u_t)$$

In the continuous approximation, the convexity of $H()$ implies the convexity of $f_t()$. To complete the proof of the lemma we need to prove that $H()$ is convex in $x_t, \kappa_t, u_t$.

**Convexity of $H(x_t, \kappa_t, u_t)$** We can write:

$$H(x_t, \kappa_t, u_t) = C_{cw} E_d[[a - \kappa_t]^+] + E_d[a][f_{2V}(h(x_t, d, a), g(\kappa_t, a))]$$

Both $h()$ and $g()$ are homogeneous mappings such that $\lambda h(x_t, d, a) = h(\lambda x_t, \lambda d, \lambda a)$ (similarly for $g()$) and both are convex. We have seen that $f_{2V}$ does not depend on $\kappa$ so we can ignore this dimension. Since $f_{2V}$ is clearly an non-decreasing function in $x$ we have that $f_{2V}(h(x_t, d, a))$ is a non-decreasing convex function of a convex function.
4.1 One-job problems

The result is a convex function in \((x_t, d, a)\). Taking expected values with respect to \(d\) and \(a\) also preserves the convexity of \(f_{2V}()\). Since \(C_{cw}[u_t - \kappa_t]^+\) is convex, we have that \(H()\) is convex in \((x_t, \kappa_t, u_t)\). ■

Lemma 4.3 \(f_{2V-2}(x_{2V-2}, \kappa_{2V-2})\) is a convex function of \((x_{2V-2}, \kappa_{2V-2})\).

**Proof**  When \(t = 2V - 2\),

\[
f_t(x_t, \kappa_t) = \min_{\omega_t \pi_{ot} \leq x_t} \{C_{ot}\omega_t + C_b(x_t - \omega_t \pi_{ot}) + E_n[f_{t+1}(x_t - \omega_t \pi_{ot} - n\pi_{rw}, \kappa_t)]\}
\]

Dropping the stage subscript we can write this cost-to-go function as:

\[
f(x, \kappa) = L(x, \kappa) + \frac{C_{ot}}{\pi_{ot}} x
\]

Where, \(L(x, \kappa) = \min_{[x - OT_{max_{\pi_{ot}}}^+] \leq b \leq x} \{G(b, \kappa)\} ; \)

\[
G(b, \kappa) = (C_b - \frac{C_{ot}}{\pi_{ot}})b + E_n[f_{t+1}(b_t - n\pi_{rw}, \kappa_t)];
\]

and \(b = x - \omega_t \pi_{ot}\)

By Lemma 4.2 we know that \(f_{t+1}(x_{t+1}, \kappa_{t+1})\) is convex for \(t = 2V - 2\).

\[
\Rightarrow E_n[f_{t+1}(b_t - n\pi_{rw}, \kappa_{t+1})] \text{ is convex in } (b_t, \kappa_{t+1})
\]

Note: \(\kappa_{t+1} = \kappa_t\) for even stages \(t\). Since \(G(b, \kappa)\) is a convex function we can show that \(L(x, \kappa)\) is convex as follows: If \(x'_o = \lambda_1 x'_1 + \lambda_2 x'_2\) for \(\lambda_1, \lambda_2 \geq 0\) and \(\lambda_1 + \lambda_2 = 1\) and similarly \(\kappa_o = \lambda_1 \kappa_1 + \lambda_2 \kappa_2\), we define the following sets:

\[
S_o \equiv \{[x'_o - OT_{max\pi_{ot}}]^+, x'_o]\}
\]

\[
S_1 \equiv \{[x'_1 - OT_{max\pi_{ot}}]^+, x'_1]\}
\]
4.1 One-job problems

\[ S_2 \equiv [(x'_2 - OT_{\text{max}} \pi_{ot})^+ \cdot x'_2] \]

By the definition of \( L() \) we have that:

\[ L(x'_o, \kappa_o) \leq G(b, \kappa_o) \text{ for all } b \in S_o \]

or,

\[ L(x'_o, \kappa_o) \leq G(\lambda_1 b_1 + \lambda_2 b_2, \lambda_1 \kappa_1 + \lambda_2 \kappa_2) \text{ for all pairs } (b_1, b_2) : b_i \in S_i \]

by the convexity of \( G() \) we have:

\[ L(x'_o, \kappa_o) \leq \lambda_1 G(b_1, \kappa_1) + \lambda_2 G(b_2, \kappa_2) \text{ for all pairs } (b_1, b_2) : b_i \in S_i \]

\[ \Rightarrow L(x'_o, \kappa_o) \leq \lambda_1 \min_{b_1 \in S_1} G(b_1, \kappa_1) + \lambda_2 \min_{b_2 \in S_2} G(b_2, \kappa_2) \]

Therefore, \( L(x, \kappa) \) is convex.

We see then that \( f_t(x_t, \kappa_t) \) is a sum of convex functions and is therefore convex as well. \[ \blacksquare \]

Lemma 4.4 \( f_{2V-3}(x_{2V-3}, \kappa_{2V-3}) \) is a convex function of \( (x_{2V-3}, \kappa_{2V-3}) \).

Proof When \( t = 2V - 3 \),

\[ f_t(x_t, \kappa_t) = \min_{u_t} \{ C_{cw} E_a[u_t - \kappa_t]^+] + E_{d_{t+1}, a}[f_{t+1}([[x_t - a\pi_{cw} + d_{t+1}]^+, [\kappa_t - a]^+) ] \} \]

As before we define \( a = \min(u_t, m_{t+1}) \) where \( m_{t+1} \) is the number of call-in workers who are actually available. We can write \( f_t(x_t, \kappa_t) \) as follows (dropping the subscripts):

\[ f_t(x, \kappa) = \min_{u_t} \{ E_a[G(a, \kappa, x)] \} \]
where:

\[ G(a, \kappa, x) = C_{cw}(a - \kappa) + E_d[f_{t+1}([x - a\pi_{cw} + d]^+, 0)] \text{ if } \kappa < a \]

and,

\[ G(a, \kappa, x) = E_d[f_{t+1}([x - a\pi_{cw} + d]^+, \kappa - a)] \text{ if } a < \kappa \]

We define \( R1 \) to be the set of points in the \((a, \kappa, x)\) space such that \( \kappa < a \), and \( R2 \) to be the set of points in the same space such that \( a < \kappa \). It is easy to show that within each region \( G(a, \kappa, x) \) is convex. We prove here that the convexity holds across regions.

Assume we have two points \( p_1 \in R1 \) and \( p_2 \in R2 \). For non-negative \( \lambda_1 \) and \( \lambda_2 \) such that \( \lambda_1 + \lambda_2 = 1 \) we define \( p_3 = \lambda_1 p_1 + \lambda_2 p_2 \) such that \( p_3 \in R1 \). We first prove that \( G(p_3) \leq \lambda_1 G(p_1) + \lambda_2 G(p_2) \).

If we define \( Q_3 = G(p_3) - \lambda_1 G(p_1) + \lambda_2 G(p_2) \) we have that:

\[ Q_3 = C_{cw}(\lambda_2 a_2 - \lambda_2 \kappa_2) + E_d \left[ f([\lambda_1 x_1 + \lambda_2 x_2 - (\lambda_1 a_1 + \lambda_2 a_2)\pi_{cw} + d]^+, 0) \right. \\
\left. - f([\lambda_1 [x_1 - a_1\pi_{cw} + d]^+, 0) - f(\lambda_2 [x_2 - a_2\pi_{cw} + d]^+, \kappa_2 - a_2) \right] \]

Since \( \lambda_2 a_2 - \lambda_2 \kappa_2 \) is non-positive and the most that a unit increase in \( \kappa \) can reduce cost is \( C_{cw} \) we can see that \( Q_3 \leq 0 \).

We now define \( p_4 = \lambda_1 p_1 + \lambda_2 p_2 \) such that \( p_4 \in R1 \). We first prove that \( G(p_4) \leq \lambda_1 G(p_1) + \lambda_2 G(p_2) \).

If we define \( Q_4 = G(p_4) - \lambda_1 G(p_1) + \lambda_2 G(p_2) \) we have that:

\[ Q_4 = -C_{cw}(\lambda_1 (a_1 - \kappa_1)) + E_d \left[ f([\lambda_1 x_1 + \lambda_2 x_2 - (\lambda_1 a_1 + \lambda_2 a_2)\pi_{cw} + d]^+, (\lambda_1 \kappa_1 + \lambda_2 \kappa_2 - \lambda_1 a_1 - \lambda_2 a_2) \right. \\
\left. - f([\lambda_1 [x_1 - a_1\pi_{cw} + d]^+, 0) - f(\lambda_2 [x_2 - a_2\pi_{cw} + d]^+, \kappa_2 - a_2) \right] \]

For similar reasons as \( Q_3 \), we have that \( Q_4 \leq 0 \). This proves that \( G(a, \kappa, x) \) is
convex. This implies that \( E_{a}[G(a, \kappa, x)] \) is convex in \((\kappa, x, u)\) and therefore we can write:

\[
 f_t(x_t, \kappa_t) = \min_{u_t} H(x_t, \kappa_t, u_t)
\]

In the continuous approximation, the convexity of \( H() \) implies the convexity of \( f_t() \).

We can apply these arguments inductively to show that:

**Proposition 4.2** \( f_t(x_t, \kappa_t) \) is convex, for all \( t \) with no assumptions about the cost hierarchy.

These results allow us to make the following statement about the outer optimization problem \( P1 \):

**Proposition 4.3** The objective function of problem \( P1 \):

\[
 C(S) = C_{rw}NV + C_{cw}MGV + C_f(N + M) + f_1^S(x'_1, \kappa_1)
\]

is convex.

**Proof** From proposition 4.2 we know that \( f_1(x_1, \kappa_1) \) is convex in \((x_1, \kappa_1)\). If the system starts with some amount of work \( x_0 \geq 0 \), then, taking into account the number of regular workers available in stage 1, we have:

\[
 x_1 = x_0 - E_n[n\pi_{rw}] = x_0 - \pi_{rw}E[n]
\]

Based upon our assumption of the independence of absences of individual workers, \( E[n] = pN \), where \( p \) is the probability a regular worker is present. Therefore \( x_1 \) is a
4.1 One-job problems

linear function of $N$ and $\kappa_1$ is a linear function of $M$. $(MGV)$.

\[ \Rightarrow f_1(x_1, \kappa_1) \text{ is convex in } S = (N, M) \]

We have noted before that the deterministic part of $C(S)$ is a combination of linear functions.

\[ \Rightarrow C(S) \text{ is the sum of linear terms and a convex term.} \]

\[ \Rightarrow C(S) \text{ is convex in } S. \]

Looking at lemma 4.3 we see that in even stages selecting the optimal amount of overtime to utilize is equivalent to selecting an optimal amount of backlog to accept or tolerate in the system. Let us denote the backlog that minimizes $G_t(b_t, \kappa_t)$ for a particular $\kappa_t$ in period $t$ to be $\beta_t(\kappa_t)$. The backlog tolerance dictates the use of overtime as follows: When $x_t > 0$ the optimal overtime usage is:

\[ \omega_t^* = \frac{x_t - b_{t-1}^*}{\pi_{ot}} \]

Where, $b_t^*$ is define by the following:

\[ b_t^* = \begin{cases} 
  x_t & \text{if } \beta_t(\kappa_t) \geq x_t \\
  \beta_t(\kappa_t) & \text{if } [x_t - \pi_{ot} OT_{max}(N, M)]^+ \leq \beta_t(\kappa_t) < x_t \\
  [x_t - \pi_{ot} OT_{max}(N, M)]^+ & \text{if } \beta_t(\kappa_t) < [x_t - \pi_{ot} OT_{max}(N, M)]^+ 
\end{cases} \]

Note the similarity to $(s, S)$ order-up-to polices in inventory control theory. We elaborate on the relation to inventory control later.
4.1 One-job problems

4.1.2 Outer optimization

In this section, we discuss how we can use the convexity results derived from the inner optimization analysis together with other observations to make the search for an optimal staffing level more efficient. First, we make a number of observations and then discuss how we would utilize them to conduct an efficient search for an optimal staffing level.

Observations

- (i) We only consider $S$ such that $N\pi_{rw} + M\pi_{cw} \leq d_{max}$. I.e. we do not staff beyond the maximum exogenous work arrival.

- (ii) Only consider $S$ such that $N\pi_{rw} + M\pi_{cw} + OT_{max}\pi_{ot} \geq \bar{d}$. This is a form of system stability constraint that is not in general an explicit constraint of the model but that in practice would always be included.

- (iii) Computing the cost of staffing arrangements with $M = 0$ is very quick since we only consider a one-dimensional state space and only make overtime decisions each stage.

- (iv) The larger $M$ is, the more computation required to solve the problem. This follows from the dependence of the state space size upon $MGV$, the larger search space for optimal call-in decisions within each stage.

- (v) We can quickly calculate upper bounds on the optimal $M$ for any value of $N$. We elaborate on this below.

Because of observation (iv) it is important to find good upper bounds on $M$. The following proposition provides a method for determining upper bounds on the optimal value of $M$ for a specific value of $N$ by only using values of $C(S)$ for which $M = 0$. As noted in observation (iii) these are quick calculations.
Proposition 4.4 For any value of $N$ the optimal $M^*$ satisfies:

$$
M^* \leq \min \left\{ M : C(N + \left\lfloor \frac{M \pi_{cw}}{\pi_{rw}} \right\rfloor, 0) - C(N + \left\lfloor \frac{(M-1) \pi_{cw}}{\pi_{rw}} \right\rfloor, 0)
+ \left[ \left\lfloor \frac{(M-1) \pi_{cw}}{\pi_{rw}} \right\rfloor - \left\lfloor \frac{M \pi_{cw}}{\pi_{rw}} \right\rfloor \right](C_{rw}V + C_f) + (GVC_{cw} + C_f) \geq 0 \right\}
$$

Proof For a staffing level $S = (N, M)$ define $D(S) = C_{cw}.MGV + C_{rw}.NV + C_f(N + M)$, i.e. the deterministic part of the cost function $C(S)$. We know that when we increase the number of call-in workers by one there is an increase in $D(S)$. In addition we know that there is a decrease in the stochastic component of the cost $f_1^{N}(x_1, \kappa_1)$ since there is more labor available and therefore backlog costs are reduced. Since $C(S)$ is convex we also know that if $C(N, M) - C(N, M + 1) \leq 0$ then $C(N, M + 1) - C(N, M + 2) \leq 0$, i.e. once $C(S)$ starts to increase in the $M$ direction it continues to increase. This means that for a specific $N$, if we find the smallest $M$ such that $C(N, M - 1) - C(N, M) \leq 0$ it is an upper bound on the $M^*$ that minimizes $C(N, M^*)$.

In the dynamic programming term of the cost function we receive a larger decrease in cost from adding a regular worker than from adding an equivalent, in terms of processing capability, number of call-in workers. This is simply because there is never a cost associated with utilizing regular workers within the context of the inner optimization. This implies that if we make the following definitions: $S_1 = (N + \left\lfloor \frac{M \pi_{cw}}{\pi_{rw}} \right\rfloor, 0), S_2 = (N + \left\lfloor \frac{(M-1) \pi_{cw}}{\pi_{rw}} \right\rfloor, 0)^2, S_3 = (N, M),$ and $S_4 = (N, M - 1)$

$$\Rightarrow f_1^{S_2}(x_1, \kappa_1) - f_1^{S_1}(x_1, \kappa_1) \geq f_1^{S_4}(x_1, \kappa_1) - f_1^{S_3}(x_1, \kappa_1)$$

\footnote{Because we must have integer values for $N$ we have to round the equivalent regular staffing levels. This makes the bounds less accurate.}
4.1 One-job problems

By definition, \( f^S_1(x_1, \kappa_1) = C(S) - D(S) \).

\[
\Rightarrow C(S_2) - C(S_1) + D(S_1) - D(S_2) \geq C(S_3) - C(S_3) + D(S_3) - D(S_4)
\]

If \( L.H.S \leq (GVC_{cw} + C_f) \) it implies that the fixed cost from adding a call-in worker exceeds any potential benefit. Therefore, the smallest \( M \) that makes \( L.H.S. \leq (GVC_{cw} + C_f) \) is an upper bound on optimal \( M \). ■

**Search strategy** We now discuss how we make use of the observations of the previous section to make our search for the optimal staffing level \( S \) as efficient as possible.

Step 1: We compute \( C(S) \) for all \( S \) such that \( M = 0 \) and \( N \in [\frac{d}{\pi_r w}, \frac{d_{max}}{\pi_r w}] \).

Step 2: Determine the minimum feasible value of \( N \): \( N_{min} \), given the constraints of problem \( P1 \).

Step 3: For each value of \( N \in [N_{min}, \frac{d_{max}}{\pi_r w}] \) determine the minimum upper bound given the constraints of \( P1 \), observation (i), and the bounds given by proposition 4.4.

Step 4: Starting at largest \( N \) such that upper bound on \( M \) is not zero, do a binary search on \( M \) to find \( M^* \).

Step 5: For \( N - 1 \) do a binary search on \( M \) that is bounded from below by the optimal \( M^* \) from the previous search. Repeat this step until \( N_{min} \) or optimal objective value exceeds value from previous search.

Step 6: Optimal staffing \( S \) from last search performed is the global optimal.

Using this search strategy allows us to reduce the number of runs of the dynamic programming engine we need to do with larger values of \( M \). In practice it performs very effectively as we demonstrate in the following example:
4.1 One-job problems

<table>
<thead>
<tr>
<th>M by N</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>216.9</td>
<td>170.1</td>
<td>171</td>
<td>184.6</td>
<td>202</td>
<td>220.8</td>
<td>240.3</td>
<td>260.1</td>
<td>280</td>
</tr>
<tr>
<td>1</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>233.1</td>
<td>171.2</td>
<td>166.6</td>
<td>178.8</td>
<td>195.6</td>
<td>214.2</td>
<td>233.5</td>
<td>253.3</td>
<td>273.2</td>
</tr>
<tr>
<td>2</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>256.3</td>
<td>176.2</td>
<td>162.9</td>
<td>173.1</td>
<td>189.3</td>
<td>207.6</td>
<td>226.8</td>
<td>246.5</td>
<td>266.4</td>
</tr>
<tr>
<td>3</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>280.5</td>
<td>184.8</td>
<td>160.3</td>
<td>167.6</td>
<td>183</td>
<td>201.1</td>
<td>220.2</td>
<td>239.8</td>
<td>259.6</td>
</tr>
<tr>
<td>4</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>200.5</td>
<td>160.5</td>
<td>162.5</td>
<td>176.8</td>
<td>194.5</td>
<td>213.5</td>
<td>233.1</td>
<td>252.8</td>
<td>272.8</td>
</tr>
<tr>
<td>5</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>216</td>
<td>163</td>
<td>158.1</td>
<td>170.7</td>
<td>188</td>
<td>206.8</td>
<td>226.3</td>
<td>246.1</td>
<td>266</td>
</tr>
<tr>
<td>6</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>243</td>
<td>172</td>
<td>156.2</td>
<td>165.1</td>
<td>181.6</td>
<td>200.2</td>
<td>219.6</td>
<td>239.3</td>
<td>259.2</td>
</tr>
<tr>
<td>7</td>
<td>268.8</td>
<td>184.8</td>
<td>157.5</td>
<td>160.3</td>
<td>175.4</td>
<td>193.6</td>
<td>212.9</td>
<td>232.6</td>
<td>252.4</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
</tr>
<tr>
<td>8</td>
<td>200.4</td>
<td>162.6</td>
<td>157.4</td>
<td>169.3</td>
<td>187.1</td>
<td>206.2</td>
<td>225.8</td>
<td>245.7</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
</tr>
<tr>
<td>9</td>
<td>171.5</td>
<td>157.6</td>
<td>164.1</td>
<td>180.6</td>
<td>199.5</td>
<td>219.1</td>
<td>238.9</td>
<td>258.8</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
</tr>
<tr>
<td>10</td>
<td>161.3</td>
<td>160.5</td>
<td>174.2</td>
<td>192.8</td>
<td>212.2</td>
<td>232</td>
<td>252</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
</tr>
<tr>
<td>11</td>
<td>160.5</td>
<td>168.8</td>
<td>186.2</td>
<td>205.5</td>
<td>225.3</td>
<td>245.2</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
</tr>
<tr>
<td>12</td>
<td>165.5</td>
<td>179.9</td>
<td>198.8</td>
<td>218.5</td>
<td>238.4</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
</tr>
<tr>
<td>13</td>
<td>174.4</td>
<td>192.2</td>
<td>211.8</td>
<td>231.6</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
</tr>
<tr>
<td>14</td>
<td>185.8</td>
<td>205</td>
<td>224.8</td>
<td>244.8</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
<td>Inf</td>
</tr>
</tbody>
</table>

Table 4.1: Cost for different staffing levels of example problem

Example of implementation of search strategy In this example we solve a particular instance of a same-period notification problem without absenteeism. We assume that work arrivals are according to a distribution $W1$ that is depicted in the numerical results chapter 5. For our purposes, here it is important to note that for this work arrival pattern $d_{min} = 0$, $d_{max} = 14$ and $\bar{d}$ is between 5 and 6 units of work. The productivity of regular workers is one unit of work per shift and for call-in workers it is .8 units per shift. We assume that there are no fixed costs for workers, that $C_{rw} = 1$, $C_{rw} = 1.1$ and that the penalty for each unit of backlog per stage is one cost unit. In this problem, we set the call-in worker guarantee, $G = .6$. In table 4.1 we present the costs of the different feasible staffing levels. Finally, we assume that problem $P1$ has a side constraint $g(N, M) \geq 0$ that restricts $S$ to the set of $(N, M)$ such that $N - M \geq 0$, i.e. there must be at least as many regular workers as call-in workers. In table 4.2 we display the sequence of steps involved in finding the optimal staffing level $S$. In block I we have marked with large X's the region of staffing levels that involve over-staffing. We have marked , with small x's, the region of unstable solutions. Note that the feasible range for $N$ with $M = 0$ is $N \in [6, 14]$. In block II we have marked off the solutions that are made infeasible by the side constraint with a g. In block III we have marked the upper bounds on $M^*$ for each value of $N$, as
calculated from proposition 4.4, and have marked them with a $U^3$.

In block IV we show the sequence of solutions we look at in the search for the optimum. Recall that the evaluation of each solution in the sequence requires computing the expected cost of the inner-dynamic optimization for that particular $S$. We mark with a * the optimal $S$ for each fixed $N$. Note that when $N = 5$ we start our search with $M = 2$. This is because when $N = 6$ the optimal $M$ was 2, which becomes a lower bound on the optimal $M$ for $N < 6$.

---

3The squares marked with a $U$ are considered potential optimal solutions.
4.2 Two-job problems

In general the 2-job problem is computationally extremely difficult. As noted in the introduction to this chapter, it can involve a 5-dimensional state space that very easily can lead to problems of enormous size. Furthermore the search space for the outer optimization is six dimensional, where most solutions can only be evaluated by solving large dynamic programming problems. The above difficulties apply to the general problem but in practice we may take advantage of specific features of the practical problem being solved to greatly reduce these difficulties. In this section, we describe a collection of general observations and reasonable approximation schemes to solving the 2-job problem that are not all useful in all cases, but that all give insight into the mathematical structure of the problems and the sources of computational difficulties.

Among other things, in this section, we argue that in general the 2-job problems have the same convexity properties of the 1-job problems. For the inner-optimization problem this means that the cost-to-go functions are convex in the state vectors. For the outer-optimization problem this means that the objective function of problem 3.1 is a convex function of the staffing level $S$. In this section we only informally prove that these statements are true by pointing out how we can follow the same arguments of the convexity proofs for the one-job problems. These arguments are also valid for the two-job problem and the pyramid n-job problems.

4.2.1 Inner optimization

In the same-period notification problems we can simplify the staffing decision-making within each period, for some special cases of the cost hierarchy. These simplifications are straightforward extensions of those made for the 1-job problems and are useful for decisions about when to use call-in workers before accepting backlog or utilizing overtime. The question of how to assign crosstrained workers to jobs within a period
4.2 Two-job problems

is much more difficult.

For the same-period notification case we can observe the following about the usage of crosstrained call-in workers: (1) We always use guaranteed hours of singly trained workers before using those of crosstrained workers. (2) If we have the choice of crosstrained or singly trained workers shifts beyond the guarantee we always utilize the worker with the lowest cost to productivity ratio. Statement (1) derives from the fact that in the same-period case we only utilize call-in workers we need and that we do not hoard guaranteed shifts. Therefore if in a period there is a need for call-in workers for job i and $\kappa_i > 0$ we utilize type i workers because we do not know if in the future there will be a need for them. On the contrary, we know that whatever type of work is in the system, if it exceeds the regular worker’s capability, we can utilize crosstrained call-in workers. Statement (2) is just another way of saying we always use the most cost effective resource.

In the next-period notification problems, the above statements do not necessarily hold. Since we do not know what the demand for labor will be when we call in workers, we tend to prefer calling in crosstrained workers. since they have a higher probability of being utilized.

4.2.2 General convexity results

In this section we outline demonstrations of convexity properties of 2-job problems. We do this for the next-period notification models with absenteeism. A similar approach can also be applied to the same-period notification models. We first describe the line of reasoning we use:

All stages have cost functions with the following general form:

$$ f_t(X, K) = \min_{Y \in \mathcal{U}} \left\{ E_D [G(X, K, Y, D) + f_{t+1}(h(X, Y, D), g(K, Y, D))] \right\} $$

where, $X, K$ are state variable vectors, $Y$ is a decision variable vector, $D$ is a
random vector, and \( U \) is a convex set.

To show that \( f_t(X, K) \) is convex we need to follow the following steps:

1 Show that \( f_{t+1}(X, K) \) is convex.

2 Show that \( G(X, K, Y, D) + f_{t+1}(h(X, Y, D), g(K, Y, D)) \) is convex in \( (X, K, Y, D) \).

If we do (1) and (2) we can argue that \( E_D [G(X, K, Y, D) + f_{t+1}(h(X, Y, D), g(K, Y, D))] \) is a convex function \( H(X, K, Y) \). We can then argue that:

\[
f_t(X, K) = \min_{Y \in U} \{ H(X, K, Y) \} \Rightarrow f_t(X, K) \text{ is convex.}
\]

In the even stages \( t \):

\[
f_t(\kappa_t, x_t) = \min_{\omega_t, \alpha_t} \left\{ \sum_{j=1}^{2} \left[ C^j_t \omega_t^j + C^j_0 [x_t^j - \pi_j \omega_t^j]^+] + E_{n_{t+1}} [f_{t+1}(\kappa_{t+1}, x_{t+1})] \right] \right\}
\]

We can write \( x_{t+1} = h(X, Y, D) \) where: \( Y = (\omega_t, \alpha_t) \) and \( D = n_{t+1} \). The function \( h \) is defined explicitly in chapter 3 but can be shown to be a convex function. \( K_{t+1} = K_t \).

We assume that \( f_{t+1}(\kappa_{t+1}, x_{t+1}) \) is convex and drop the time subscripts.

\[
\Rightarrow f(K_3, \lambda_1 h(X_1, Y_1, D_1) + \lambda_2 h(X_2, Y_2, D_2))
\]

where, \( X_3 = \lambda_1 X_1 + \lambda_2 X_2 \) etc. Since \( h \) is convex and \( F \) is increasing in \( X \):

\[
f(K_3, h(X_3, Y_3, D_3)) \leq f(K_3, \lambda_1 h(X_1, Y_1, D_1) + \lambda_2 h(X_2, Y_2, D_2))
\]

\[
\Rightarrow f_{t+1} \text{ is convex in } (K, X, Y, D)
\]

The backlog and overtime cost terms are convex also so we have shown that \( f_t(\kappa_t, x_t) \) is convex if \( f_{t+1} \) is.
4.2 Two-job problems

In odd stages $t$:

$$f_{t}(\kappa_{t}, x_{t}) = \min_{u_{t}, a_{t}^{i}} \left\{ E_{m_{t+1}, d_{t+1}} \left[ \sum_{i \in [1,2,12]} C_{cu}^{i} [a_{t}^{i} - \kappa_{t}^{i}]^{+} + f_{t+1}(\kappa_{t+1}, x_{t+1}) \right] \right\}$$

$$= \min_{Y \in U} \left\{ E_{D} \left[ G(X, K, Y, D) + f_{t+1}(g(K, Y, D), h(X, Y, D)) \right] \right\}$$

In this case $Y = (u_{t}, a_{t}^{0})$ and $D = (m_{t+1}, d_{t+1})$. We again assume that $f_{t+1}$ is convex and make use of the facts that it is increasing in $x$ and decreasing in $\kappa$. We consider two vectors $(K_1, X_1, Y_1, D_1)$ and $(K_2, X_2, Y_2, D_2)$ and define their convex combination:

$$(K_3, X_3, Y_3, D_3) = \lambda_1 (K_1, X_1, Y_1, D_1) + \lambda_2 (K_2, X_2, Y_2, D_2)$$

We want to show that:

$$G(K_3, Y_3, D_3) + f(g(K_3, Y_3, D_3), h(X_3, Y_3, D_3)) - \lambda_1 [G(K_1, Y_1, D_1) + f(g(K_1, Y_1, D_1), h(X_1, Y_1, D_1))]$$

$$- \lambda_2 [G(K_2, Y_2, D_2) + f(g(K_2, Y_2, D_2), h(X_2, Y_2, D_2))] \leq 0$$

$G(K_3, Y_3, D_3)$ is convex which implies that :

$$\Delta G = G_3 - \lambda_1 G_1 - \lambda_2 G_2 \leq 0$$

We have assumed that $f(\kappa, x)$ is convex:

$$\Rightarrow f(\lambda_1 g_1 + \lambda_2 g_2, \lambda_1 h_1 + \lambda_2 h_2) \leq \lambda_1 f(g_1, h_1) + \lambda_2 f(g_2, h_2)$$

We actually need to show that:

$$\Delta G + f(g_3, h_3) - f(\lambda_1 g_1 + \lambda_2 g_2, \lambda_1 h_1 + \lambda_2 h_2) \leq 0$$
4.2 Two-job problems

By the convexity of \( g \) and \( h \) we know that \( g_3 \leq \lambda_1 g_1 + \lambda_2 g_2 \) and that \( h_3 \leq \lambda_1 h_1 + \lambda_2 h_2 \). These inequalities are componentwise. In any of the \( \kappa^i \) directions the most that \( f \) can increase is \( C_{cw}^i \). This increase is compensated for, by a decrease in \( G \). Therefore, although it is in a higher dimensional space, we have that \( G(K,Y,D) + f(g(K,Y,D), h(X,Y,D)) \) is convex as in Lemma 4.4 for the 1-job problems. The convexity of \( f_t(\kappa_t, x_t) \) follows.

For \( t \) even and odd we have shown that the convexity of \( f_{t+1}(\kappa_{t+1}, x_{t+1}) \) is enough to prove the convexity of \( f_t(\kappa_t, x_t) \). To complete the analysis we only need to observe that \( f_{2V}(\kappa_{2V}, x_{2V}) \) is convex.

We note that:

- There is nothing in the structure of the pyramid models that should invalidate the arguments made in this section. Thus we conjecture that the convexity properties hold for the pyramid models as well.

- We also conjecture that, if we were to generalize the formulation of the two-job problems to n-jobs the convexity results would generalize as well.

4.2.3 Outer optimization

As noted before the 2-job outer optimization problem can involve up to a 6-dimensional search. If we do not allow crosstraining workers, the problem clearly reduces to two 1-job problems. If we include crosstraining then it is a 1,2 or 3-dimensional problem if we use only regular workers. We can add 1,2 or 3 more dimensions by considering call-in workers. In analyzing the outer optimization problem we want to achieve two goals, one is to reduce the number of solutions we must consider as possible optima, the other is to reduce the number of computationally expensive solutions we must consider. Computationally expensive solutions are those that involve larger values of \( M \) and more non-zero values of \( M_1, M_2, \) and \( M_3 \).
4.2 Two-job problems

In this section, we first present some general observations about the 2-job problem that are similar to those made for the 1-job problem. We then propose some strategies for taking advantage of these observations together with some approximation schemes to solve the problem to near optimality with as little work as possible.

Observations

- (i) We only consider $S$ such that: $N_i \pi_{ii}^{rw} + M_i \pi_{ii}^{cw} \leq d_{i_{max}}^i$ for $i = 1$ or $2$, and such that: $\sum_{i=1}^{3} N_i \pi_{ii}^{rw} + M_i \pi_{ii}^{cw} \leq \sum_{j=1}^{2} d_{j_{max}}^j$. I.e. we do not staff beyond the maximum exogenous work arrival.

- (ii) Only consider $S$ such that both $N_3 \pi_{11}^{ct} + M_3 \pi_{11}^{ct} - [\tilde{d}_1 - N_1 \pi_{11}^{rw} - M_1 \pi_{11}^{cw}]^+ \geq 0$, and $\frac{n_1}{\pi_{11}^{ct}} \{N_3 \pi_{22}^{ct} + M_3 \pi_{22}^{ct} - [\tilde{d}_1 - N_1 \pi_{11}^{rw} - M_1 \pi_{11}^{cw}]^+ \} - [\tilde{d}_2 - N_2 \pi_{22}^{rw} - M_2 \pi_{22}^{cw}]^+ \geq 0$. This is a form of system stability constraint that is not in general an explicit constraint of the model but in practice would always be included.

- (iii) Computing the cost of staffing arrangements with $M = 0$ is very quick since we only consider a two-dimensional state space and only make overtime and crossstrained assignment decisions each stage.

- (iv) The larger $M$, the more computation required to solve the problem. This follows from the dependence of the state space size upon $MGV$, and the larger search space for optimal call-in decisions within each stage.

- (v) We can calculate upper bounds on the optimal $M$ for any value of $N$ relatively quickly.

- (vi) We can use the results of solving the one-job and partial 2-job problems to place bounds on solutions to the full 2-job problems. We elaborate on this below.
4.2 Two-job problems

We now elaborate on observation (vi). If we solve the 2-job problem constrained to \( N_1 = N_2 = M_1 = M_2 = 0 \); i.e. only crosstraining workers, regular and call-in, we get upper bounds on \( N_3 \) and \( M_3 \) for the unconstrained problem by just solving inner-optimizations that are 3-dimensional instead of 5. This follows from the convexity of \( C(S) \). Similarly if we solve both 1-job problems independently, i.e. no crosstraining, we get upper bounds on \( N_1, N_2, M_1, M_2 \), also by the convexity of \( C(S) \). We can solve a series of problems easier than the general 6-dimensional problem to generate a set of upper bounds on the outer decision variables. We can then take the minimum of the bounds for each variable and use that to restrict our search space. This process is depicted in figure 4-1.

![Figure 4-1: Bounding the 2-job outer optimization problem](image)

In this figure each vertical line represents the range of possible values for a decision variable, 0 being at the bottom. Each curve in the figure represents the optimal solution to a restricted version of the 2-job problem. For example the thin dotted line represents the optimal staffing level for the 1-job problem for job type 1. Note, it only takes positive values for the variables \( N_1 \) and \( M_1 \). The solid thick line represents the solution to the 2-job problem when only crosstrained workers are allowed, therefore only \( N_3 \) and \( M_3 \) take on positive values. Each time we solve such a restricted version of the problem we get another curve. Each crossing of a vertical line, at a positive point, is another upper bound for the variable represented by that line. In figure 4-1
there are two upper bounds for $N_3$. One is derived from the solution to the problem with only crosstrained workers. The other bound is derived from the solution to the problem with only regular workers. The latter provides a tighter bound.

With relatively the same amount of difficulty we can also solve the problem with $S$ restricted to the set of solutions with $M_2 = M_3 = 0$ and similarly for $M_1 = M_2 = 0$ and $M_1 = M_3 = 0$. In all these cases, we have restricted ourselves to solving problems with only one type of call-in worker at a time. Solving these restricted problems not only provides bounds on the optimal values of the decision variables, it also gives us an idea if there is much benefit from allowing more types of workers than in the restricted problems.

Although it is impossible to make general and formal statements about this, it seems that in practice it is unlikely that a solution with four or five nonzero dimensions of $S$ will be greatly inferior to a solution with six. What we imply by this statement is the diminishing returns to flexibility. I.e., if there are benefits to having crosstrained workers the added benefits of having call-in workers are less than the benefits achieved by just adding call-in workers, we demonstrate this in chapter 5. To take advantage of this phenomenon we should solve the restricted versions of the 2-job problem in a systematic way:

First, we should solve the two related 1-job problems with only singly trained regular workers. We can then solve the 2-job problem with only regular workers. The difference between the two solutions should give us an idea about the benefits of crosstraining flexibility. If we then solve the 2-job problem, with call-in crosstrained workers, instead of the $N_3$ workers we can get an idea about the benefits of call-in flexibility beyond crosstraining flexibility. Finally, we can solve the two 1-job problems with regular and call-in workers to see what kind of benefits to expect from call-in flexibility alone. As a result of these comparison we may determine that we are unlikely to improve much on a solution with only $N_1$, $N_2$, and $M_3$. Therefore, we
4.2 Two-job problems

do not even attempt to solve the problem with more than three decision variables.

Unfortunately, even though we may have situations in which it is clear that a
solution, with only $N_3$ and $M_3$ non-zero, is near optimal we may be diverted from
them by side constraints in problem $P2$. These side constraints may limit the ratio
of crosstrained to singly trained workers. In this case, we may have to consider
solutions with all three types of call-in workers to achieve a satisfactory result. Often
these side constraints limit the search space considerably, but we cannot ignore the
possibility of having to calculate a long series of $C(S)$ values for $S$ that require
solving an inner optimization problem with a 5 dimensional state space. For these
situations we propose the following heuristic for calculating the expected cost of the
inner optimization problem.

$k$-scaling heuristic The main source of growth in size of the state space is in the
$k$ dimension. Every time we consider the cost of a solution with $M_i$ call-in workers of
type $i$ we increase the state space by a factor $M_iG_iV$ over when we did not include
call-in workers type $i$. The factor $V$ is very impractical since it represents the number
of periods in the planning horizon. The $k$-scaling heuristic is a reformulation of the
inner optimization problem. We redefine $k_i^t$ to be the average number of shifts of
call-in workers type $i$ used until time period $t$. Each period we round this number to
the nearest integer, or fraction of an integer.

The above scaling results in the $k$ dimension of the state space being reduced to
taking on values from 0 to $M_i$. Based upon how much we want to reduce the size
of the problem, we can decide what the resolution should be. If we only use the
integers from 0 to $M_i$, we reduce the state space by a factor of $G_iV$. If for example
$G_i = .5$ for all $i$ and $V = 20$ periods we can reduce the state space size by a factor
of $(GV)^3 = 1000$. This, redefinition of $k$, can be applied selectively to a particular
worker type $i$, and it can be applied with different resolution to each $i$.

The drawback of this reformulation is that it results in rounding error in the
accounting of call-in worker utilization. During the dynamic optimization, decisions may be made during a time period to take advantage of beneficial rounding. It is impossible to determine theoretically what kind of errors this will generate. We hypothesize that many of these errors cancel each other out, and that in problem instances with high guarantees, i.e. \( G \geq .6 \) the relative errors are low.

Because we assume that call-in workers are utilized in whole shift increments it is natural to discretize the \( \kappa \) dimension of the state space for the purpose of implementing the dynamic programming algorithm. Discretizing the \( x \) dimension of the state space is not obvious and it can play an important role in the computational feasibility of the problem.

**Discretization of \( x \) and the productivity scaling heuristic** The discretization of \( x \) is determined by the productivity rates of the different types of workers. If all workers have the same productivity rate we can rescale the work arrival distribution by this rate and discretize the whole problem in units of worker shifts. However, let's say we have a work arrival distribution that takes on values of 0 to 10 and that regular workers have productivity 1.0 units per shift and call-in workers have productivity .8 units per shift. In this case we must keep track of units of work in increments of .2. The discretization leads to an \( x \) state variable that takes on 51 values instead of 11\(^4\). I.e. if the smallest difference between two productivity rates is \( \Delta \) the discretization of the work dimension increases by a factor of \( \frac{1}{\Delta} \).

Since we are very sensitive to the size of the state space, in the 2-job problems, we can initially solve problems where we multiply all productivity rates by an integer factor. This increases \( \Delta \) by the same factor and thus reduces the size of the state space. We can use this scaled problem to zero-in, on a few possible optimal solutions, before increasing the resolution. A possible compromise in this area is to scale all the

\(^4\)If we keep track of backlog it is of course larger.
cost parameters for different workers by their productivity rates and thereby reflect
differences in the cost parameters only. We can then rescale solutions to get actual
staffing levels. In order to represent the different productivity rates of crosstrained
workers at different jobs we would need to scale the work arrival distributions in some
way as well. This approach might lose some of the detail in productivity differences
but eliminates the need to represent the workload state variable with a fine grain
measure.

In the many numerical tests we have conducted, we have found that the objective
function $C(S)$ is usually relatively flat around optimal solutions with several solutions
nearly as good as the optimal. This suggests the two approximation approaches
discussed in this section, the $\kappa$–scaling heuristic and the productivity rate scaling
heuristic, should perform well.

4.3 Relation to other Operations Research problems

The problems formulated in this thesis have some similarities to problems in inventory
control and queueing. We discuss the extent of those similarities here.

4.3.1 Relationship to inventory control theory

Looking first at the one-job branch of the problem tree the problems modeled in
this thesis can be considered to be inventory control problems in which the work in
the system can be considered to be negative inventory. This inventory has an upper
bound 0, since we can not create a finished goods inventory. In each period we make
an ordering/production decision. This decision is how much labor effort to apply to

\footnote{Another way to view the inability to inventory is to consider labor to be a perishable good with
a one period lifespan.}
increase the inventory, i.e. reduce workload. We do not try to give a survey of the inventory management literature but rather refer the reader to the survey in [Por90]. This characterization of the problem is more complicated for multiple jobs.

1-job problems  In each period we have new demand, which represents the shortage in regular workers, i.e. $s_t = \pi_{rw} n_t - d_t$. This can be negative or positive. In each period we can choose between two different technologies with which to produce. One is using call-in workers, the other is using overtime. Each technology has an upper bound on production in each period. To compare our models to inventory models it is useful to make the assumption that call-in workers cost is some linear factor that is the same each period (like overtime), we have a piecewise linear cost function and standard inventory problem.

The same-period notification model, with no-absenteeism, is then a multi-period, stochastic inventory control problem with a piecewise convex linear cost function for ordering. This has been well studied in the literature. Absenteeism of regular workers can be made part of the demand process and absenteeism of call-in workers can be thought of as uncertainty in the yield of an order.

Representing next-period notification is the same as introducing a lead time to the call-in technology of one period. Overtime can be considered to be an emergency ordering option with no lead time but higher costs. Again, these models or ones similar to them have already been studied in the literature.

In our problem, the cost of the call-in technology cannot be simply represented within a period as a linear function. Instead we think of three production technologies, one is free and depleted over time (the guaranteed call-in work hours), the second is the call-in workers payed a linear wage, and the third is the overtime hours paid a higher linear wage. The bound on per period usage of the free technology would be linked to the bound on the non-free call-in technology, i.e. $M$. We also would have to represent a joint bound on the free and call-in technology production each period.
which adds another wrinkle. The issue of the call-in guarantee is a significant point of departure from regular inventory problems in terms in that it requires keeping track of control decisions over time in addition to inventory positions. However, we have seen in this chapter that much of the mathematical structure of operating policies is the same for our problems as for inventory problems. Namely, the optimal staffing decisions made within a period in the 1-job problems can be characterized as generalized base stock policies, see [Por90]. We also note that whenever the call-in guarantee has been completely utilized, i.e. when \( \kappa \) reaches 0, we can eliminate this dimension of the state space and consider the problem a standard inventory control problem.

2-job problems To view the 2-job problems as inventory problems we need think of workers representing 3 products, one for each demand stream and a third product that is substitutable for each of the others but doesn't have it's own demand. The substitutability represents crosstraining. Deciding how many crosstrained workers to utilize for each job can be viewed as a multi-product inventory control with a resource constraint. This has been studied in [Eva67] but not for as complicated a cost structure as we have.

4.3.2 Relation to queueing

If we view the utilization of regular and call-in workers separately in the 2-job problem we see that the problem has a blend of inventory and queueing problems. The utilization of crosstrained call-in workers is akin to a multi-product inventory control with a resource constraint. The assignment of crosstrained regular workers to jobs is more like the assignment of a flexible server in a multi-class queueing system.

We can think of each period's exogenous work arrival as the arrival of a customer to a queueing system. The amount of work that arrives in a period determines the
service time of the customer representing that periods arrival. In the 2-job problem there are two streams of customers arriving to a facility with three different servers. Server 1 represents all the singly trained workers for job 1, server 2 represents all the singly trained workers for job 2, and server 3 represents all the crosstrained workers. In this simplified representation of the 2-job problem, each period we must decide how to allocate server 3. In this model we are assuming deterministic arrivals and general distribution service times.

This kind of problem has been addressed recently in the queueing/inventory control literature see [Zip95], [Har95], [Wei92], [GvR95]. The latter three of these results are reliant upon heavy-traffic Brownian approximations to develop heuristic control policies. The worker assignment problem in the 2-job problems is much more complex because the utilization of a crosstrained worker by assigning them to a job has costs. These costs determine whether or not a call-in worker is utilized in a period or not, similarly for overtime work. This aspect of the models developed in this thesis put them beyond the analysis techniques of the cited papers. Other complications to these approaches are caused by the finite planning horizon we consider and doubts about the heavy traffic assumptions.

4.4 Future research directions

The focus of this thesis is on the modeling of complex flexible work arrangements and on using these models to increase our understanding of when different forms of flexibility are more effective. To achieve these goals we do not necessarily have to solve instances of the problems that are the most computationally challenging, as can be seen in the next chapter. Because of this we have not performed all the analysis that could be done to turn the models into tools that can be efficiently

---

6As is seen in chapter 5, it is very common for these system to never be in any kind of steady-state.
4.4 Future research directions

implemented. In this section we discuss some potential directions for future research into the computational side of these models.

The most promising direction is simulation, especially if we want to consider multiple-job problems with more than 2-jobs. To run effective simulations, we need to derive some satisfactory staffing policies for the period by period decision-making. To do this, more work must be done to understand the relationships between backlog penalties, call-in and overtime utilization, over time. Fortunately, since it is relatively easy to solve the problems optimally for problem instances with little or no flexibility we can have good baselines for comparing the effectiveness of the heuristic policies that are proposed.

A side benefit of using heuristic policies for period-by-period decision making is that it may be easier to implement simpler heuristic policies than optimal ones that are more sensitive to the state of the system and therefore more complicated. This suggests another direction for interesting research. We have not done much work toward understanding the sensitivity of staffing decisions to the system state and time period. In chapter 5, we have some empirical results that suggest that within the interior of a planning horizon the system achieves some form of steady state. We would like to know what determines the existence of steady-state, whether or not we can characterize the system behavior during it, and determine when it begins and ends.
Chapter 5

Numerical Results

In chapters 2 and 3 we have formulated a family of models that represent a wide range of types of labor flexibility. We also discussed reasonable extensions of these models that could be of interest in different practical settings. In chapter 4 we derived a set of theoretical results about the mathematical problem defined by the models formulated in chapters 2 and 3. Some of these results gave us insight into how the optimization of the models should be performed. Some of these results also gave us insight into what staffing decisions are optimal under certain conditions. In this chapter, we present results of numerical experiments that provide insights into the benefits of the different types of labor flexibility modeled in this thesis and the ways in which we can expect the systems modeled to behave under different conditions.

In the first section, we make some of the ideas of flexible work arrangements more concrete with some numerical examples. In these examples, we show how the use of call-in workers and crosstrained workers can change the staffing patterns of a workplace. In the second section, we conduct a thorough investigation into the determinants of the benefits of flexibility. For working time flexibility we demonstrate the linkages between its benefits, stochasticity of the work environment, and information. For functional flexibility or crosstraining we demonstrate how it can impact
the benefits of call-in flexibility and how the correlation of work arrival streams in the 2-job problems effects the benefits of cross-training.

In the third section, we present examples of the trajectories of call-in usage and backlog over time. These examples show that the qualitative behavior of the system differs strongly when decisions are made with different amounts of information. In appendix A we discuss a variety of technical issues that were involved in implementing the various models on the computer and present a record of the parameters used in the runs.

In performing the various numerical experiments we have attempted to isolate the quantities that are most relevant for the demonstration being made. This means that we set many parameters equal to one another varying only one and seeing its effect on the cost of operating the system. As a result most of the runs do not use data that would resemble any realistic problem. The exceptions to this are the work arrival distributions which to some degree are motivated by observations of actual workplaces.

In general our goal in performing these numerical experiments is not to solve any specific staffing problem within the setting of a specific work environment. Rather, it is to illuminate the complex relationships between the different factors that interact within the models. A result of these experiments is that we observe that the behavior of the models is not in contradiction to our expectations of reality. Therefore, the experiments also serve as a validation of the models and support their use in practice.

5.1 General observations

In Figure 5-1 we plot the cost of operating the system with different staffing levels. These results are based upon parameter set 1 and work arrivals $W1$ (see appendix). In all cases the call-in guarantee is $G = .6$ and no overtime is allowed. The curve marked 'Same-period N=0' shows the expected cost of operating the system with
the above parameters when only call-in workers are used and we have same-period notification. For this case the horizontal axis marks the number of call-in workers staffed. In the graph marked 'Next-period N=0' we have the same results plotted except with next-period notification. The graph marked 'M=0' shows the expected cost of operating the system when only regular workers are utilized. This is the same for same-period and next-period notification. One way to think about the effect of

![Graph](image)

Figure 5-1: Expected cost vs. staffing level in 1-job problem

flexibility in our model is as a way to increase the utilization of labor. Since flexibility allows us to better match workers to workload, the number of productive hours of each worker increases. In our models, worker production reduces cost by preventing backlog. When there are few workers they are heavily utilized and therefore flexibility does not make much difference to costs. As we increase the staff we reduce utilization which in turn reduces backlog cost but eventually increases total cost. Adding a flexible worker does not decrease utilization as quickly as adding regular workers but does reduce backlog costs equally. That is, we can achieve the same production level with fewer workers or fewer paid hours. We can get a general sense of how call-in
flexibility works from Figure 5-1. We see that:

- For low staffing levels call-in workers and regular workers are all heavily utilized and therefore result in similar costs.

- For high staffing levels costs start to increase linearly for all cases, with the cost of additional call-in workers being less than that of regular workers.

- For staffing levels around the minima each additional call-in worker is better than an additional regular worker because they work fewer hours to process the same amount of work.

- Same-period notification is better than next-period notification for call-in workers.

In Figure 5-2 we plot the expected cost of operating the system versus the number of workers for different values of $\rho$ for the two job problem with parameter set 1\(^1\). Where $\rho$ is the correlation between the two work arrival streams. We again use $W_1$ arrivals and do not allow overtime.

In Figure 5-2 we can see how functional flexibility can increase the utilization of workers in a similar way as working time flexibility does. We also see that the properties of the work arrival processes determines the extent of this utilization increase.

In the following sections of the chapter we investigate the causes of the above phenomena in detail.

5.2 Determinants of the benefits of flexibility

In this section we investigate the factors that determine the benefits of the different forms of flexibility and how they interact with one another. First, we consider working time flexibility focusing on the 1-job problems. Second, we consider functional

\[^1\text{All parameters distinguished by job type are the same.}\]
5.2 Determinants of the benefits of flexibility

Operating costs plotted versus staff size for 2-job problem with different correlations between arrival streams

Number of Workers

- Corr 0
- Singly Trained
- Corr 5
- Corr .5
- Corr -1

Figure 5-2: Expected cost vs. staffing in 2-job problem

flexibility in the 2-job problems. Finally, we discuss how working time and functional flexibility interact with one another when utilized within the same work environment.

5.2.1 Working time flexibility

We have explicitly modeled two forms of working time flexibility, call-in workers and overtime. We define the degree of call-in flexibility in terms of the call-in guarantee, G. The lower the value of G the more flexibility we have in utilizing call-in workers. The degree of overtime flexibility is defined by the limit on overtime hours worked by each worker represented by the parameter OT_{max}. Within this chapter OT_{max}, is the maximum fraction of a shift that a worker can work overtime in a period. The larger the value of OT_{max} the more flexibility in the system.

Other indirect determiners, of the degree of flexibility, are the costs of utilizing call-in workers or overtime. For example, if the per-shift cost of a call-in worker, C_{cw}, is very high, it greatly limits our flexibility in using them. Similarly, if the overtime premium, in a work place, is very high we won’t utilize overtime very often. To
5.2 Determinants of the benefits of flexibility

simplify matters, in all the examples in this section, we assume that the per-shift cost of all workers is the same. That means that we set $C_{rw} = C_{rw} = C_{ot}$. Clearly, in practice, it would be hard to find a workplace in which the overtime pay rate is equal to that of regular hours. However, for our purposes, keeping all costs equal allows us to isolate the impact of the flexibility of interest, be it overtime or call-in.

System stochasticity

In chapter 1 we discussed that the motivation for finding flexible approaches to utilizing and managing work forces is to find ways to handle stochasticity within a work environment. The implication is: the more uncertainty there is in a work environment, as caused by demand variability and absenteeism, the greater the need, and therefore the benefits of flexibility.

We demonstrate this phenomenon by looking at the same-period notification case of the 1-job problem. We first test the effect of different degrees of variability in the work arrival stream. We then see how working time flexibility is beneficial in dealing with absenteeism.

Arrival variability Using parameter set 1 (see appendix), we optimize the staffing for three different work arrival distributions, $W_1, W_2, W_3$ that each have different degrees of variability as defined by their coefficients of variation. $W_1$ is a scaled representation of the work arrival to a new accounts processing area of a mutual fund company. $W_2$ and $W_3$ are artificially constructed to provide contrast to $W_1$. Since we are assuming that all the per-shift costs of workers are the same, it is always optimal to staff with call-in workers and not utilize regular workers at all. When $G = 1$ call-in workers are equivalent to regular workers. As $G$ decreases to 0 the optimal number of call-in workers, $M^*$ increases. To see the impact of overtime flexibility we fix $M = 0$ and find the optimal regular worker pool size $N^*$, when $OT_{max} = .25$. We compare all these cases to the base, no-flexibility case where $M = 0$ and $OT_{max} = 0$. 

5.2 Determinants of the benefits of flexibility

In figure 1 we plot the optimal expected cost of the objective function of problem $P1$ versus different values of $G$ with no overtime. I.e. each point marks the cost of operating the system when the optimal number of call-in workers are contracted for the planning horizon, given a work arrival process and a guarantee, $G$. Each curve plots the relationship of cost to the guarantee for the different work arrival process. In all cases, we see that the costs increase as the guarantee increases, i.e. as the flexibility decreases. The steepness of a curve shows the sensitivity of the cost to call-in flexibility for the work arrivals in question. We can see that for almost all $G$, the $W1$ curve is steeper than the others and that the $W2$ curve is steeper than the $W3$ curve. This follows the same ordering as the coefficients of variation of the three processes. On the same graph, we plot the costs of operating the system with only

![Comparison of benefits of call-in flexibility for different work arrival distributions](image)

**Figure 5-3:** Benefits of working time flexibility for different work arrival distributions

regular workers and overtime with $OT_{max} = .25$. In all cases, we see that overtime flexibility is approximately equivalent to operating with call-in workers and $G = .9$. This demonstrates that in same-period notification regimes a small amount of call-in flexibility can be superior to overtime flexibility.
5.2 Determinants of the benefits of flexibility

In figure 5-4 we show the benefits of call-in flexibility, with $G = .6$, relative to the no-flexibility case\textsuperscript{2}. In each graph, we have also marked the cost of operating the system if each period we could somehow have the exact number of workers necessary to process all the work. This forms a lower bound on the benefits of any form of flexibility. We can see that, in the cases with low work arrival variability, there is significantly less room for improvement, over the no-flexibility case. However, we can also observe that in the $W2, W3$ cases we are able to achieve the lower bound with $G = .6$. In fact, if we look at the graphs in figure 1, we see that most of the improvement in cost is achieved by setting $G = .8$ for $W2$ and $W3$. To summarize, while the potential benefits from working time flexibility are greater when there is greater workload variability, a small amount of flexibility can have a large benefit even when variability is low.

![Chart showing savings from G=0.6, gap with lower bound, and lower bound over work arrivals for W1, W2, and W3.]

Figure 5-4: Expected costs with call-in flexibility relative to costs without flexibility

\textsuperscript{2}By no-flexibility we mean only regular workers are used without overtime.
5.2 Determinants of the benefits of flexibility

Absenteism  In chapter 2 we discussed three ways we could model absenteeism, within the general modeling framework formulated there. The three differ in the relative times at which absenteeism information becomes available to the decision-maker. We focus on the type (i) absenteeism in which the decision-maker knows the regular worker absenteeism before making the call-in decision, and knows the call-in worker availability when making the overtime decision.

In these experiments, we compare the benefits from increased flexibility for the same-period notification problem with and without absenteeism. We assume that the probability of a call-in or regular worker being unavailable or absent in a period is .1 and independent of other workers. We use work arrivals $W3$. In practice, we would try to respond to regular worker's absenteeism with working time flexibility. Therefore, in these examples, we restrict ourselves to staffing levels with significant amounts of regular workers. In figures 5-5 and 5-6 we plot the cost of operating the system with different amounts of call-in flexibility. Each pair of curves represents results with different fixed amounts, $N$, of regular workers.

![Graph: Expected Cost vs. G when N=5](image)

Figure 5-5: Expected Cost vs. $G$ when $N=5$
5.2 Determinants of the benefits of flexibility

![Graph showing expected cost vs. call-in guarantee, G when N=9](image)

Figure 5-6: Expected Cost vs. G when N=9

We can see, in each case, the cost is more sensitive to flexibility when there is absenteeism. Note: this is true even though call-in workers are absent at the same rate as regular workers. The reason the benefits have increased is that the amount of stochasticity in the system in the system has increased when we allow absenteeism.

**Information**

In the preceding examples we have seen that call-in worker flexibility can have a powerful positive effect on the expected cost of operating the system in a same-period notification regime. In the next-period notification regime we do not expect call-in flexibility to be as beneficial. The primary advantage of call-in workers is the ability to avoid "paid lost time" or situations in which there are more workers present than are needed to process the work in the system. When same-period notification is possible, call-in workers are only utilized when there is work for them. As a result, while it is possible for there to be periods in which there are excess regular workers in the facility, any call-in worker present in the facility during some period is fully
5.2 Determinants of the benefits of flexibility

utilized. This situation is markedly different from the next-period notification regime. When only next-period notification is possible, we decide how many call-in workers to have present in a period before knowing the work arrival quantity for that period. As a result, it is possible to have paid-lost-time for call-in workers. Therefore, the benefits of call-in flexibility are less.

The difference between same-period and next-period notification can be characterized in terms of the information available to the decision-maker when he/she makes decisions about the utilization of call-in workers or overtime. In the same-period case, the total amount of work in the system in a period is known when both the call-in and overtime decisions are made. In the next-period case only the backlog component of work is known. When absenteeism type (i) is present the decision maker also knows the regular worker absenteeism in both cases.

When the backlog penalty is relatively high the level of backlog in the system is low and therefore the workload information available in the next-period is low. When the backlog penalty is lower, the backlog level rises and becomes a more significant portion of the total workload information, and therefore, more information is available in the next-period case. However, since backlog penalties are lower, the impact of backlog on the system performance is less significant, and therefore the overall benefits of call-in flexibility, in the next-period notification regime, remain small.

The two cases, same- and next-period notification, represent extremes of information availability. In practice, we would expect that some information about work arrivals would always be available to the decision-maker when staffing decisions must be made. In chapter 2 we discussed an extension of the models to a mixed-information case, in which within each period work arrival information becomes available at two points in time. The first, is before any staffing decisions are made, and the second, is after call-in decisions are made, but before overtime decisions are made. In figure 5-7 we plot expected operating cost versus call-in guarantee for same-period, next-
5.2 Determinants of the benefits of flexibility

period, and mixed-info regimes. We use work arrival process W2 as the inputs. For
the mixed-info regime W2 is a two part process representing the two stages in which
work information arrives to the decision maker (see the appendix). We convolve these
two parts into a single W2 arrival process that is used as inputs into the same-period
and next-period regimes.

![Comparison of benefits of working-time flexibility for different information regimes](image)

**Figure 5-7: Expected cost vs. G for different information regimes**

In figure 5-7 we can see that the benefits of call-in flexibility are greater in the
cases with more information. On the same graph we have also plotted the cost of
operating the system with only regular workers and overtime with $OT_{max} = .25$. We
see that, when there is less information available when call-in decisions are made, the
benefits of overtime relative to call-in flexibility are higher. This is because, in all
situations, overtime decisions are made with complete information.

If the work arrival processes were not identically distributed each period, but
rather were periodic, e.g. arrival processes that were day-of-week dependent, then
the information situation would be different. In this more general case, the particular
period or day of the week would add information in all the information regimes dis-
cussed above. This added information would increase the benefits of call-in flexibility
in all cases.
5.2 Determinants of the benefits of flexibility

Summary

To summarize, we have seen that the potential benefits of working time flexibility must be assessed with an understanding of the work environment that goes beyond just the wages, productivity, and flexibility of the workers themselves. With the numerical experiments described above we have demonstrated that the benefits of flexibility are greater in systems with greater uncertainty, in their requirements for labor. We have also demonstrated that the information available when staffing decisions are made plays, a no less important, role in this assessment.

5.2.2 Functional flexibility

In formulating the 2-job problems we have created a framework for investigating the potential benefits of functional flexibility. For functional flexibility we want to consider three issues:

1. What impact does correlation between the two work arrival processes have on the benefits of crosstraining?

2. How much functional flexibility is needed?

3. What impact does the stochasticity of the work arrival processes have on the benefits of crosstraining?

In addressing all these issues we isolate functional flexibility by ignoring working time flexibility. This means that, in all the experiments, the only choice of workers are regular workers, either singly or crosstrained. The specific parameter used for the experiments are listed in the appendix, but in general we set all costs and productivities equal for all types of workers. We also use symmetric work arrival processes with equal backlog penalties for jobs 1 and 2.
5.2 Determinants of the benefits of flexibility

Correlation

We expect that if the two work arrival streams are positively correlated the benefits of crosstraining would be reduced relative to the case in which the arrival streams are uncorrelated or negatively correlated. The extreme case is when the correlation is 1. In this case, it is equivalent to consider the system as two independent jobs with only singly trained workers. I.e., there is no benefit from functional flexibility. This is to be contrasted with the case where the correlation is -1. In this case, we should need considerably fewer workers than the correlation 1 case since, when one job has a heavy workload the other has a light workload. We expect functional flexibility to have its greatest benefits here.

To demonstrate these ideas in a concrete way, we conduct experiments with the 2-job problem and \( W1 \) arrivals to each job and see how the system performs when we change the correlation of the two arrival streams. I.e. we think of a work arrival process of pairs of work quantities \((d_1, d_2)\), where \( d_i \) is the work arriving for job \( i \). The pairs \((d_1, d_2)\) come from some joint probability mass function \( \Delta_j(d_1, d_2) \) that has a correlation \( \rho_j \). To conduct the experiments we generate the functions \( \Delta_j \) so that they have marginal probability mass functions \( W1 \) and correlation \( \rho_j \). The functions satisfying these conditions are not unique. The ones we use in the experiments are based on the bivariate normal distribution and are generated with an algorithm described in the appendix.

In figure 5-8 we have plotted the expected cost of operating the system with only crosstrained workers for different values of the correlation \( \rho \). Each point represents the minimum expected cost if the work arrival process is governed by a joint probability mass function \( \Delta_j(d_1, d_2) \) with a particular correlation \( \rho_j \) where for all points the marginal pmf's for each job are the same as the \( W1 \) arrival process. On the same graph we have plotted the same information for the system with \( W3 \) based arrivals.

It is clear from this figure that the greater the correlation between the work arrivals
5.2 Determinants of the benefits of flexibility

![Graph showing cost versus correlation for W1 and W3 arrivals.]

Figure 5-8: Expected cost versus correlation in 2-job problems for W1 arrivals, comparing work arrivals W1 and W3

In the two jobs the less benefit to be gained from functional flexibility.

**How much flexibility?**

In practice, it may be unlikely that a workplace would operate with all workers crosstrained. The reasons for this are discussed in chapter 6. Therefore, it useful to understand what degree of flexibility is really necessary to benefit from crosstraining, in a 2-job environment. As with working time flexibility, it is necessary to define a measure of the degree of functional flexibility. With working time flexibility we had explicit parameters $G$ and $OT_{max}$ in the models that served this role. For functional flexibility we define the degree of flexibility $CT_{ratio}$ as the maximum fraction of all workers that can be crosstrained. This kind of constraint would appear as a side constraint of the outer optimization problem $P2$.

To investigate this question, we conduct numerical experiments in which we compare the cost of operating the system with different degrees of flexibility $CT_{ratio}$ for work arrivals based on the $W1$ arrival process with different correlations. The results
of these experiments are displayed in figure 5-9. From the figure, we can see that

Figure 5-9: Expected cost versus correlation in 2-job problems for W1 arrivals, comparing different levels of crosstraining

when we have more functional flexibility, i.e. when $CT_{ratio}$ is bigger, we can reduce the costs of operating the system more than when $CT_{ratio}$ is smaller. However, we also see that when the work arrival streams are positively correlated, and therefore the potential benefit of crosstraining is smaller, a lesser degree of functional flexibility is sufficient to achieve all or most of the potential benefits.

**Stochasticity**

In figure 5-8 we have plotted the cost of optimal staffing with regular crosstrained workers for $W1$ and $W3$ based arrival processes with different correlations. $W3$ has significantly less variability than $W1$ and comparing the benefits of crosstraining for these two arrival patterns should give us an idea of the affect of stochasticity on functional flexibility. We see that the $W1$ curve is steeper signifying greater benefits from crosstraining when there is greater stochasticity in the system.
5.2 Determinants of the benefits of flexibility

This result is what we would expect considering the source of crosstraining benefits. Crosstraining allows us to pool the variability of the different work arrival streams. If the individual streams are less variable to begin with, we do not expect that pooling them will reduce the variability of the total work arrivals by as much, as when the original arrival streams have greater variability.

5.2.3 Interaction between types of flexibility

The numerical experiments we have discussed thus far, have all involved one form of workforce flexibility taken in isolation. We have also identified the important determinants of the potential benefits of different forms of flexibility. Based upon these results we should be able to say, given a specific work environment, which form of flexibility would be most beneficial. For example in a 2-job workplace with high variability in work arrivals to each job, high correlation between jobs and early work arrival information we would focus on call-in flexibility. If however, the information about work arrivals was late coming and there was little correlation between jobs, we would focus on crosstraining.

In practice, we can expect there to be constraints limiting the degree of each type of flexibility. In such situations, we would draw upon a combination of forms of flexibility to reduce costs. We find that, when multiple forms of flexibility are utilized in one work environment, the benefits are not completely additive; i.e., the effects of one can diminish the effects of the other. We demonstrate this by conducting the experiment depicted in figure 5-10.

We take the case of 2-job work arrivals with W1 marginals and 0 correlation. We set the backlog penalties for both jobs to be the same, and set all productivity rates to be the same for all types of labor. In this setting, if we only use crosstrained workers, we can represent the work arrivals as a single stream that is the sum of the two original work arrival streams. We determine the expected costs of operating the
5.2 Determinants of the benefits of flexibility

![Graph showing expected cost vs. call-in guarantee, G for singly-trained and cross-trained call-in workers.]

Figure 5-10: Utilizing call-in and functional flexibility jointly

system with only cross-trained call-in workers for different guarantees $G$ and compare them with the result when we only use cross-trained regular workers (or set $G = 1$). We compare these results with the benefits from using singly-trained call-in workers.

In figure 5-10, we see that as we decrease $G$ there is a steeper decline in cost for the singly trained case versus the cross-trained case. We explain this phenomenon as a direct consequence of what we have observed about the effect of workload stochasticity on the benefits of call-in flexibility. When we have cross-trained workers, and can combine the work arrival streams into one, there is a reduction in variance. The reduction in variance reduces the benefits of call-in flexibility. We can conclude that, in addition to there being diminishing returns to increased flexibility of a certain type, there are also diminishing returns to adding types of flexibility. However, there is still considerable benefit to having both forms of flexibility as can be seen by the gap between the curves. This suggests that the two forms of flexibility are addressing different types of variability. This makes sense. Call-in or, in general, working time flexibility allows adjustment to variability in work volume while functional flexibility
Table 5.1: 2-job problem results with no working time flexibility.

<table>
<thead>
<tr>
<th>Arrivals</th>
<th>$\rho$</th>
<th>Ct-ratio</th>
<th>$C(S^*)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>1</td>
<td>0.25</td>
<td>342</td>
</tr>
<tr>
<td>W1</td>
<td>1</td>
<td>0.5</td>
<td>342</td>
</tr>
<tr>
<td>W1</td>
<td>1</td>
<td>1</td>
<td>342</td>
</tr>
<tr>
<td>W1</td>
<td>0</td>
<td>0.25</td>
<td>308</td>
</tr>
<tr>
<td>W1</td>
<td>0</td>
<td>0.5</td>
<td>300</td>
</tr>
<tr>
<td>W1</td>
<td>0</td>
<td>1</td>
<td>298</td>
</tr>
<tr>
<td>W1</td>
<td>-1</td>
<td>0.25</td>
<td>285</td>
</tr>
<tr>
<td>W1</td>
<td>-1</td>
<td>0.5</td>
<td>253</td>
</tr>
<tr>
<td>W1</td>
<td>-1</td>
<td>1</td>
<td>226</td>
</tr>
</tbody>
</table>

Table 5.2: 1-job problem results with call-in flexibility.

<table>
<thead>
<tr>
<th>Arrivals</th>
<th>$G$</th>
<th>$C(S^*)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>0.2</td>
<td>214.96</td>
</tr>
<tr>
<td>W1</td>
<td>0.4</td>
<td>217.3</td>
</tr>
<tr>
<td>W1</td>
<td>0.6</td>
<td>235.12</td>
</tr>
<tr>
<td>W1</td>
<td>0.8</td>
<td>278.48</td>
</tr>
<tr>
<td>W1</td>
<td>1</td>
<td>342</td>
</tr>
</tbody>
</table>

is an adjustment mechanism for variability in work mix.

In Table 5.1 we have a summary of the results plotted in Figure 5-9. We can compare these with the results of operating the same system with singly trained call-in workers, i.e. the upper curve in Figure 5-10, which appear in Table 5.2.

Comparing the two tables we can see that if the costs and productivities of workers are the same, we can expect call-in flexibility to be more effective than crosstraining. To summarize, in this section, we have seen that there are several factors that determine whether a form of labor adjustment is effective or not at reducing costs. In other words the labor technology selected should match the work environment. We have also seen that the working time and functional flexibility are very different instruments of labor adjustment that also interact with one another.
5.3 System behavior

We have seen that the benefits of call-in flexibility are very different depending upon the information available when staffing decisions are made. In this section we demonstrate how the overall system behavior is very different as well. By system behavior we mean how the state of the system changes over time. In all the models we have represented the work environment (or system) with a two-dimensional state space. One dimension kept track of the work in the system and the other kept track of the call-in worker usage. Analogously, here we want to consider how the amount of backlog in the system varies with time and how the call-in worker utilization varies with time as well.

For any particular staffing level $S$, we can compute the expected amount of backlog per-period and its variance, and the expected value and variance of the call-in worker usage per-period, as a by-product of solving the inner dynamic optimization problem. In this section, we solve a same-period notification problem for a set of realistic parameters and use the optimal staffing levels $S^*$ to generate the period by period trajectory of backlog and call-in usage. We do the same for the next-period version of the same problem with different values of $G$. We conduct the experiments with problem parameter set 2 (see appendix). In this experiment we solve outer optimization problems with the side constraint $M < N$.

5.3.1 Call-in Usage Trajectories

We now consider what the pattern of call-in utilization looks like. In figures 5-11 and 5-12 we look at the period by period expected call-in worker usage and its coefficient of variation for next-period notification with absenteeism, work arrivals $W_1$, and staffing level $S = (5, 2)$ for different values of $G$. (This staffing level is optimal for the $G = .6$ case.) We make the following observations about figure 5-11:
5.3 System behavior

Figure 5-11: Trajectory of expected call-in utilization, for next-period notification for W1 and different values of G

Obs1: In each period the expected call-in usage increases with G.

Obs2: For each G there is a peak late in the planning horizon.

Obs3: For each G the call-in usage decreases sharply toward 0 after the peak.

Obs4: Call-in usage always increases from period 1 to period 2.

Obs1 is explained by the fact that for higher values of G more call-in worker shifts have been guaranteed and therefore, within the context of period by period operations, there are more free call-in shifts to use. Obs2 is explained by the call-in guarantee as well. We know that if toward the end of the planning horizon we have shifts remaining in the call-in guarantee it is worthwhile to use them up in a less conservative manner than in earlier periods, this causes the observed peak. Note, the peak is preceded by a decreasing trend, and both phenomenon are more marked as G increases.

Obs3 is expected in this case since call-in worker costs per shift are greater than backlog costs. At the end of the planning period it is most likely that the guaranteed shifts will have been used up; furthermore, the finite horizon makes the system more tolerant of backlog. These three factors make call-in utilization less attractive. Obs4 is explained by the initial conditions of the system. We assumed that the system starts out empty which means that it will probably have more work in the second period, than the first, since there might be some backlog after the first period. Therefore, in
expected value, there will be a greater need for workers in period 2 than in period 1. We make the following observations about figure 5-12:

Figure 5-12: Trajectory of variation in call-in utilization, for next-period notification for W1 and different values of G

Obs1a: Variation decreases with G.

Obs2a: Variation increases dramatically in the last periods.

Obs3a: Initial coefficient of variation for $G = .2,.4$ is very high.

Obs4a: For $G = .4$ variation starts as $G = .2$ but converges toward the $G = .6$ case; while, $G = .6$ starts as $G = .8$ and converges toward $G = .4$.

Obs1a is closely related to Obs1. When we change $G$ we don’t change the randomness in work arrivals and absenteeism and since expected call-in usage is higher when $G$ is higher it makes sense that the coefficient of variation decreases with $G$. However, this relationship is even stronger because the reduced usage of call-in workers brought about by lowering $G$ results in an increase in backlog levels and the variability of backlog. Obs2a can be explained by the sharp drop in call-in worker usage. Obs3a can be explained by the low call-in usage in the first periods caused by the initial conditions. Obs4a seems to suggest that when the call-in guarantee is substantial, $G = .8$, or small, $G = .2$, there is little difference between system behavior in the body of the
5.3 System behavior

planning horizon, i.e. some form of steady-state. When $G = .4$ the behavior of the system seems to start like a $G = .2$ system and then converge to a $G = .6$ system. This behavior shows that for intermediate values of $G$ the system is very sensitive to the remaining call-in guarantee and therefore never achieves any sort of steady state.

To get a sense of the behavior of the system in the same-period notification case we compare it with the behavior of a next-period notification system with the same parameters. In figures 5-13 and 5-14, we have the expected value and coefficient of variation of call-in utilization by period for W1 work arrivals, no absenteeism, $G = .6$ and $S = (4, 3)^3$ for the same-period case. On the same graphs we have the trajectories for $S = (5, 2)^4$ for the next-period case.

![Figure 5-13: Trajectory of expected call-in utilization, comparing next and same-period notification for W1 and G=.6](image)

![Figure 5-14: Trajectory of variation in call-in utilization, comparing next and same-period notification for W1 and G=.6](image)

---

3Optimal staffing for same-period notification with side constraint $M < N$.  
4Optimal for $G = .6$ in next-period notification.
5.3 System behavior

The main observations are:

Obs1b: Expected call-in usage is higher in same-period case than next-period.

Obs2b: Same-period variation in call-in usage is higher than in next-period case.

Obs3b: In same-period case there is little change in system behavior from period to period.

We explain these observations by noting that because call-in decisions are made in the same-period case, when there is complete information about the workload in the system, the realizations of the work arrival process directly determine call-in usage. Uncertain impact of the call-in workers in the next-period case keeps their usage lower, and the anticipatory nature of the decision keeps its variability lower as well. Since the work arrivals are iid. each period we can also explain Obs3b by the direct relation between same-period call-in usage and work arrivals.

To summarize, the behavior of the system with respect to call-in worker utilization over the course of the planning horizon can be very different for the same-period notification case than for the next-period and can be strongly effected by the call-in guarantee. We also see that the effects of the initial conditions and the finite horizon are not important over the majority of the planning horizon, but the finite call-in guarantee can effect the system throughout the planning horizon (see Obs1, Obs2, and Obs4a).

5.3.2 Backlog trajectories

Using the same test problems and solutions as for the call-in trajectories we also plot the trajectory of backlog over time. In figure 5-15 we plot the expected backlog for each period for different values of $G$. It is difficult to distinguish between the different trajectories but we can clearly see that toward the end of the planning horizon the tolerance for backlog increases.
5.3 System behavior

![Graph showing expected backlog over time periods for different values of G.]

Figure 5-15: Trajectory of expected backlog, for next-period notification with W1 and different values of G.

In figure 5-16 we can see that the variability of the backlog is very different for different values of G. It is difficult to interpret these trajectories but we make the following observations:

Obs1c: For all G the coefficient of variation drops sharply at the end of the planning horizon.

Obs2c: For G = .2 the coefficient of variation stays about the until the last few periods.

Obs3c: The behavior of the system changes more for the G = .8 and G = .4 cases than for the G = .6 case.

![Graph showing coefficient of variation over time periods for different values of G.]

Figure 5-16: Trajectory of variation in backlog, for next-period notification with W1 and different values of G.

Obs1c can be explained by looking at the behavior of figure 5-15 toward the end of the planning horizon. Since in all cases the expected backlog goes up sharply we
expect the coefficient of variation to be decreased. When \( G = .2 \) the call-in guarantee is used up very quickly and therefore the system achieves a steady-state until it is close enough to the end of the planning horizon for the tolerance for backlog to increase, which explains Obs2c. In the interior portion of the planning horizon the behavior of \( G = .6 \) case is steadier than the \( G = .4, G = .8 \) cases. We might expect that \( G = .6 \) would have behavior that is somewhere between the \( G = .4, G = .8 \) cases. The reason it doesn’t is that, in all these cases, we are operating the system with a staff \( S = (5, 2) \) that is optimal for \( G = .6 \). In general, we expect that an optimal solution to the outer optimization without side constraints would achieve a steady-state type behavior within the interior of the planning horizon. If we operate the system with a staffing level \( S \) that is only optimal in a constrained version of problem \( P1 \) we expect its behavior to be transient throughout most of the planning horizon.

In figures 5-17 and 5-18 we compare the behavior of the backlog in the system for the same-period notification optimal solution to the \( G = .6 \) problem and the next-period notification optimal solution, as we did for call-in worker usage.

![Graph showing expected backlog over time](image)

Figure 5-17: Trajectory of expected backlog, comparing next and same-period notification with \( W1 \) and \( G=.6 \)

We see that, as in the case of call-in worker usage, the same-period notification system behaves as if it is in a steady-state. However, while we saw above that there was greater variation in call-in usage for the same-period relative to the next-period case, the roles are reversed for backlog. The variation in system backlog is greater for the next-period notification system. This is further demonstration of how, under the same-period notification regime, we are better able to transfer the variability in
5.4 Summary

Figure 5-18: Trajectory of variation in backlog, comparing next and same-period notification with W1 and G=.6

work loads onto the call-in workers. This ability is the essence of the information advantage of the same-period case.

5.4 Summary

In this chapter we have conducted a battery of numerical experiments to demonstrate how the effects of flexibility can only be well understood within the context of a particular work environment. That is, we must take a holistic view of the system that includes the degree of stochasticity, the correlation of work arrival streams, and the timing of information about work arrivals and absenteeism, relative to staffing decisions. We summarize our findings as follows:

- The greater the variability of the system, the greater the returns to flexibility.

- The benefits of call-in flexibility relative to overtime and functional flexibility is determined by the information available when call-in decisions are made.

- There are diminishing returns to flexibility and even modest levels of flexibility can cause significant cost reductions.
• Positive correlation of work arrival streams in the 2-job problems reduces the benefits of crosstraining, while negative correlation increases the benefits.

• Working time flexibility and functional flexibility are not completely substitutable for one another and working time is ceteris paribus more effective.

We also found that, with respect to call-in flexibility, the system behaved very differently under different information regimes. The major observations were:

• Call-in utilization each period was more variable when more information was available.

• Within the interior of the planning horizon the system tend to achieve a steady state for optimal staffing levels of the unconstrained problems\(^5\). Otherwise the behavior is transient throughout the planning horizon.

These results, though not obvious, make intuitive sense. This reasonableness, of the results, suggests that our models are valid representations of flexible work arrangements in stochastic environments. The sensitivity of the behavior of the system and cost benefits of flexibility to environmental factors also demonstrate it is necessary to draw the distinctions between different models that we have done in figure 1-2.

In this chapter we have focused on employer’s incentives to use flexible work arrangements, in other words the demand side of the flexible labor market. In the next chapter we investigate the disincentives to employers that are not modeled and the employees interests, or supply side of the market.

\(^5\)By unconstrained we mean problems with the same parameters but no side constraints in the outer optimization problem that restrict the choice of \(S\).
Chapter 6

Implementation Issues and Conclusions

Our analysis of the models in Chapter 5 focused on the benefits to employers of flexibility. In this chapter, we investigate the prospects for the successful implementation of the flexible work arrangements (FWAs). To truly be able to assess these prospects it is necessary to characterize the work environments that are suitable for flexibility and investigate the factors that can influence the implementation that have not been included in our models. We identify three sets of such factors. One set consists of the potential hidden costs to employers of implementing FWAs, that we have not modeled explicitly. Another set derives from organizational and managerial obstacles to implementation of FWAs. The third set derives from considering the concerns of employees. In chapter 1 we described how the models formulated in this thesis can be viewed as labor demand models, that is they represent the process by which firms decide how many and what kinds of workers to employ. In particular, the models minimize cost for employers, showing a bias toward their needs, not employee's. In practice an employer decides how many workers of each type to employ, but also what employment package to offer. I.e. combinations of wages, hour, guarantees etc.
The set of employment packages that an employer can offer is restricted by the labor supply of workers who are interested in such packages. As a result, successful implementation of FWAs requires some understanding of the supply side of the flexible labor market.

We find that it is useful to view these issues from the perspective of dual labor market theory\(^1\). Our approach is to first establish that the characteristics of jobs suitable for FWAs define a class of jobs that is not equivalent to secondary jobs. Then using a dualist approach we analyze the major supply side issues and determine separate guidelines for implementing FWAs in primary and secondary jobs. We draw upon our field observations of a USPS mail processing plant and the transaction processing center of a large mutual fund company, experience with our models, and the human resource management and labor economics literature.

In the first section of the chapter, we give short descriptions of the two work environments we observed. In the second section, we characterize jobs that are suitable for FWAs. In the third section, we address the supply side factors not explicitly modeled that affect the implementation of FWAs, and in the fourth we review the potential hidden costs to employers of implementing FWAs. In the fifth section we again draw upon our field observations to demonstrate how even when the opportunity to benefit from FWAs exists they are avoided because of organizational obstacles and managerial preconceptions. In the sixth and final section we summarize the discussion and

\(^1\)According the economic theory of dual labor markets described in [BP80], labor markets are segmented into more than one sector. This segmentation implies that there are discontinuities in the market such that human behavior cannot be captured by a single model applied in continuous fashion across all segments. For the purposes of this thesis it is sufficient to talk of two sectors the primary and secondary sectors. The secondary sector is defined as jobs that: “are characterized by relatively low wages, poor working conditions ... little security of career advancement. They are essentially unskilled, either requiring no skill at all or utilizing basic human skills and capacities shared by virtually all adult workers.” Primary sector jobs are in contrast, “better paying, with generally more pleasant working conditions and greater social status. They typically provide considerable security, a certain opportunity for advancement toward higher wages, and more attractive opportunities over the course of a work life.” From this perspective the behavior of workers, in terms of their work preferences and the incentives that motivate them, must be considered different for each sector of the market.
establish implementation guidelines.

6.1 Field sites

As part of this research we spent the first six months of 1994 studying the operations of a USPS mail processing plant, in Massachusetts, from the perspective of staffing problems. The USPS employees approximately 700,000 employees, many of whom work in facilities such as this one, and has an operating budget of approximately $50 billion. The USPS projected a deficit of $1.7b, a major part of which was caused by overtime [AP94]. These facts suggested to us that the staffing problems of the USPS should be of interest. In addition, there has been a proliferation of companies offering similar services and facing similar operational problems, such as UPS, Federal Express, DHL etc.

We also spent one year, June 1994 to June 1995, studying a variety of staffing problems in a large mutual fund company. The problems of this company are shared by the entire mutual fund industry as well as a variety of financial services companies. All these companies have uncertainty in their demand for services and have large transaction processing centers.

These two field sites were rich sources of insights, into the problems FWAs can help solve and the practical problems involved in instituting FWAs. We now summarize some of our observations of each of these environments.

6.1.1 Mutual fund company

We studied the transaction processing center of a large mutual fund company based in Boston, Massachusetts. At this center, customer requests arrive daily through the mail. These requests range from opening new accounts to changes of address to account redemptions. The operation is split into two functional groups. The first
handles primarily new account setups, while the second handles everything else which come under the heading redemptions and account maintainences.

Workers who actually process the work are titled Associates. Their jobs mainly involve interacting with a computer database to carry out the customer's transaction requests. They also identify errors and incomplete forms, and are responsible for contacting customers to resolve these problems. All Associates are part time workers scheduled to work 4 hours per day, five days per week. Most Associates are married women with children who are supplementing their husband's income. Their average job tenure is 3 years and average wage is $11/hr.

The company recently built a large facility in a suburban area and employs several hundred people from the surrounding communities. The management is very pleased with the local labor pool, finding them highly literate (important for job success), and dependable. The Associates are not unionized.

There is considerable seasonality in the work over the course of the calendar year, with tax season having the heaviest work load. There is also high variability in the amount of new work arriving each day with the associated time pressures of financial transactions. When work is light, managers can send excess workers home but pay them to the nearest two hours worth of work.

6.1.2 Mail sorting facility

We also studied a mail sorting facility or processing and distribution center of the United States Postal service. The facility is one of the most modern in the country with a high level of automation and excellent record of performance. It is medium sized, with approximately 750 employees. The postal workers are unionized and, in the plant, are represented by two unions, the Mail Handlers Union and the Mail Clerk's Union which represent the two major job categories.

Under the union contracts there are several well defined flexible job classes, namely
6.2 What jobs are suitable for FWAs?

full time flexible, part time flexible, casual and transitional. These classes are defined as follows:

- Full-time flexible: Career employees, scheduled 5 days per week, 40 hours per week, guaranteed 8 hours pay upon reporting for work. Flexible days off and reporting times.

- Part-time flexible: Career employees, minimum 4 hours, maximum 8 hours per day. No more than 40 hours per week and guaranteed 4 hours upon reporting. Flexible days off and reporting time.

- Casual: Non-career, schedule can be changed daily not guaranteed full-time hours each week.

- Transitional: Employees in jobs scheduled to be abolished. Same scheduling as casuals.

In this facility there were no part time or full-time flexible employees only transitional and casual. From interviews with supervisors we found that these employees had very regular schedules. We also found that there were periods of high overtime in the past that were reduced through managerial pressure not operational changes\(^2\). Overtime was well sought after, by employees, with 400 people on the overtime desired list and the reduction in overtime was not well received.

---

\(^2\)Overtime went down from 15% of all hours worked to 8% when the plant manager increased the accountability of tour managers.
not done without reason, many of these jobs have the characteristics of secondary jobs, especially with regard to lack of job security. However, this practice has attached a negative stigma to FWAs that is not entirely deserved. Polivka [Pol89] has pointed out that contingent work, though sharing much in common with secondary work, is distinct and there is value to analyzing it distinctly. We extend this notion to the wider range of flexibility developed in this thesis. That is, we argue that the jobs suitable for FWAs can be both primary and secondary jobs from the perspective of dual market theory.

We have developed the models in this thesis for work environments in which there is significant short term stochasticity in the demand for labor caused by workload stochasticity and absenteeism. The types of flexibility we have modeled are forms of labor adjustment to this kind of variability. As a result, the kinds of jobs that can be performed by flexible workers of this kind are those that involve relatively short tasks\(^3\). In this sense, the forms of work force flexibility modeled here (in particular working time flexibility) are implementable in a restricted set of work environments.

Based upon our models and observations the major characteristics of jobs that are suitable for flexible work arrangements are: short-task, easy to monitor, in services, and with stochastic environments. The jobs must involve short-task so that the irregularity of flexible staffing does not disrupt productivity. They must be easy to monitor because otherwise guaranteeing productivity might involve paying higher wages to workers to make the threat of dismissal a strong disincentive to shirking\(^4\). They must be in services with stochastic environments because otherwise it is unlikely that there would be benefits from flexibility, as we have defined the staffing problems

\(^{3}\)For example processing new account applications in a mutual fund company is a job with short tasks. A worker can end her shift without leaving any incomplete individual tasks. In contrast a software design project is a job with long tasks. A software designer can be hired on a contingent basis to work on the project but this represents numerical flexibility which is not modeled in this thesis.

\(^{4}\)We derive this idea from the effort regulation models presented in [RT95, LR95].
here.

The major characteristics of secondary jobs are: low-skill, low-wage, easy to monitor, and lacking security\(^5\). It seems that there is considerable overlap between secondary jobs and jobs suitable for flexible work arrangements. In addition, the jobs described in our field observation also fit the description of secondary jobs. However, can we describe substitute teachers, on-call doctors and nurses as members of the secondary market?

In looking at the BLS contingent and alternative work survey of Feb. 1995 [Sta95a] we can explain some of this overlap. While 34.5% of non-contingent work is in services, about 54% of contingent workers are in services\(^6\). These are jobs with less security making them more likely to be secondary jobs. Easy monitoring is also common to both categories of jobs. However, based upon our definitions it is not necessary for flexible jobs to be low-skilled and low-wage. One could also argue that service jobs with short tasks are in general easy to monitor\(^7\). Therefore, we can see that flexibility suitable jobs can include many secondary jobs, but the two classes of jobs are not equivalents. We depict this non-equivalence with a Venn diagram of job characteristics in Figure 6-1.

In this figure the flexibility suitable jobs, that are also secondary jobs, are shaded black. The hatched region to the right of this represents the set of secondary jobs.

---

\(^5\)Unions can make secondary jobs higher paid and more secure, as in the case of the USPS. We discuss this in the next section. For the purpose of this discussion we ignore the role of unions.

\(^6\)The definition of contingent is the same as in [Pol89], see chapter 1.

\(^7\)Our reasoning here is that in most service related jobs there is at least the customer present to monitor performance. For example, a hospital patient complains if they are not satisfied with the care provided by a nurse. In the case of short tasks it is also easier to set a work pace standard. As a result we have observed that despite having at times found postal workers sleeping behind sacks of mail, the tight time commitments of mail sorting and automated equipment controls the pace of work. In the mutual fund company there are also strict deadlines by which new accounts must be processed and the use of computers weds each item of work to an identified Associate. In the telephone call center of the same company workers time spent on calls is logged automatically and supervisors periodically listen in on conversations. We do not claim that these are all examples of primary jobs but rather use them to illustrate the connection between quick paced work in services and the ease of monitoring work performance.
6.2 What jobs are suitable for FWAs?

Figure 6-1: The non-equivalence of the sets of secondary and flexibility suitable jobs

The other hatched region is the set of flexibility suitable jobs that are primary sector as well. Note, that we have not marked the set of jobs with security, even though security is a key determinant of the sector a job falls into. We have excluded it because job security is not a relevant concept in the time scales we are considering. As a result our models are neutral with respect to this issue. This suggests that many of the negative effects for workers caused by FWAs can be mitigated by job security.

Since we cannot classify all flexibility suitable work settings as secondary market jobs we cannot analyze the supply side of the staffing problems in only one way. Rather, to determine how to implement a FWA a firm must first identify the segment of the labor market it is operating in. If it is in the primary segment the preferences of its labor pool and the incentives appropriate to them differ from work places in the secondary segment.
6.3 The supply side of the flexible labor market

Even though we have developed the models with a bias toward minimizing the operational costs of the employer we have included many parameters that represent directly and indirectly the concerns of employees. These parameters are the call-in guarantee, the overtime limit, the various productivity rates, the cost parameters, and the absenteeism rate. The realistic combinations of these parameters represent the majority of labor supply issues. While the models accept any combination of values for these parameters an employer would be hard pressed, for example, to find a worker that is never absent, works every day, is willing to work overtime equal to a full shift and has a productivity rate higher than all other workers. We have not modeled the relationships between these parameters but clearly, understanding these relationships is crucial for a successful implementation of a FWA.

Following a dualist approach we note that these relationships are different for primary versus secondary jobs. We make one important deviation from this. We have noted that our models are based upon a short planning horizon during which it is not feasible to adjust to variability in the work environment with numerical flexibility. As a result job security is not a part of our models. Job security is, however, an important component of the definition of secondary and primary jobs and an important determinant of the incentive structure of a workplace. Therefore, we first discuss the role of job security as a background feature important to any implementation of a FWA. We then discuss the relation between work hours, compensation and productivity.

6.3.1 Job security

The flexible workforce arrangements we have discussed have an interesting connection to job security. Traditionally, firms in the United States have operated under a tradition of employment at will. In this system, firms would respond to down turns in demand for their products or services by laying-off workers. When demand picked
up, layed-off workers were often rehired. However, the layoff/hiring options were primarily the only instruments of labor force adjustments utilized by employers. This is in contrast to Western Europe where regulations promoting job security created incentives for work sharing programs to adjust to the demand shocks. Under these programs, workers are not laid off but their hours are reduced, when demand increases the hours of the workers are increased.

Today, employers in the U.S. have modified their labor adjustment strategies somewhat. Their new strategy has been called a "core-periphery" strategy by some, [App87]. In this strategy a firm has core employees with usually full-time hours, benefits, and good long-term employment prospects. The firm supplements this core with a periphery of mainly contingent workers, who are used to absorb fluctuations in the demand for the firm's outputs. Hiring and layoffs are first carried out in the periphery. The peripheral workers, by their contingent relationship with the firm, are often not protected by the same job security regulation that full-time permanent employees have. In this way, employers can continue the employment at will tradition albeit with only the peripheral workers.

**Working time flexibility and job security** There is considerable debate among economists as to whether or not worksharing policies actual increase or decrease employment and whether they are preferred or not, from a societal welfare perspective. This debate is not directly addressed by our models. For a neoclassical view of this see [Laz86]. A more positive view of work-sharing arrangements, based on experience with short-time compensation programs instituted in California, Germany and Arizona, see [MM84].

From the perspective of the models developed in this thesis, a flexible worker may or may not be a long term employee of a firm. The flexibility we represent is geared

---

8 Depending upon the duration of the down turn and the firms strategy for expanding production when market conditions improve, many workers are never rehired.
toward short term variability of the work environment, not long term persistent down
turns in demand for a firm’s outputs. However, the models do suggest ways in which
to implement work-sharing arrangements in a way that improves their value to a
firm, i.e. by using call-in flexibility. In Chapter 5, we have seen that call-in flexibility
usually results in an absolute increase in the numbers of employees\(^9\), and decrease
in the number of shifts or hours worked per employee. This is a way to spread the
work among more employees, in a manner that is more efficient, from an employer’s
perspective, than simply reducing the hours of full-time regular workers across the
board.

**Functional flexibility and job security**  Functional flexibility can increase job
security for employees in two ways. The first way is based upon a human capital
argument while the second is based directly on flexibility. If the firm had to invest
in training crosstrained workers it has an incentive to retain them in order to protect
this investment, see [RM86] and [Bec93]. In addition the new skill also provides the
worker with a hedge against changes in the activities of the firm. If work is decreased
in one area of activity the crosstrained worker has greater ability to move to an area
of the firm that is still very active.

However, functional flexibility need not have any of these benefits for employ-
ment security. An extreme form of firm specific functional flexibility is geographical
flexibility which, as described in Chapter 3, does not actually require any new mar-
ketable skills from the employee. Therefore, there is no investment in the worker
nor functional mobility. Unfortunately, firm specific functional flexibility can also in-
volve nothing more than a breaking down of job definitions. This results in a worker
having little control over their job content and perhaps a lowering of morale as jobs
that workers identified with became jobs that “anyone can do”. This is most likely

\(^9\)Assuming equal productivity rates for all employees.
6.3 The supply side of the flexible labor market

a problem in low skill jobs where job definition and classifications are an important tool for protecting worker's rights\textsuperscript{10}, i.e. secondary jobs. In fact this breakdown of protective job delineations can actually hurt job security.

Of all the parameters related to worker's concerns listed above, job security has the most impact on worker's pay. Most worker's being risk averse prefer a lower wage and secure job rather than higher wage and uncertain job with the same expected earnings. Therefore, throughout the discussion that follows all projected increases worker's pay should be lessened if we are considering an environment with secure jobs.

To summarize, our models do not have a bias toward any form of job security arrangements, as this issue is, to a great degree, beyond the temporal scope of the decision making we are modeling. Yet, the models do provide a method for implementing worksharing plans that for workers, tradeoff increased job-security with flexibility demands.

6.3.2 Working hours

In the models we have developed the major benefit to employers of working time flexibility comes from savings in paid lost time. The result of these savings is that, in general, call-in\textsuperscript{11} flexible workers work fewer hours than regular full-time workers. The questions we want to answer here are: are these shortened hours preferred by some workers? If so, how do we relate these preferences to employer's needs and to compensation?

\textsuperscript{10}The distinction between Mail Handlers and Clerks in the USPS mail sorting facilities is a prominent example of this.

\textsuperscript{11}We do not discuss overtime here because it is a relatively well understood phenomenon and the incentives for overtime are set by law.
6.3 The supply side of the flexible labor market

**Hours preferences** The evidence on this subject appears, at first, to be conflicting. On one hand, there is evidence there are many workers who are over-employed. By over-employment we mean workers who work more hours than they would choose, given other options. For example workers who perform consistent and large amounts of overtime each week. There have been studies that show that many workers would be willing to trade future increases in earnings for more leisure time, see [Sch91]. There is increasing corporate awareness of the need to address work/family issues in a comprehensive way [HP93], [Chr95] because of worker's difficulties in finding the time to balance work and household responsibilities.

On the other hand, there is evidence many workers are under-employed. By under-employment we mean workers who by choice would work longer hours than they do currently. For example, workers categorized in the Bureau of Labor Statistics CPS as being contingent workers that would prefer non-contingent employment [Sta95a]. In [GA92], the sharp rise in contingent employment between 1982-88 is explained by employer's needs not employee's. They determine that demand side needs combined with lessened labor bargaining power caused the increase as opposed to supply side needs. The implication is that overwhelmingly contingent employment was created by employer's need for flexibility in adjusting to marker volatility for their products and services combined with a greater ability to dictate employment terms vis-a-vis unions. Increased worker desire for these types of working arrangements was not shown to be a significant factor.

We do not believe these different analyses are truly contradictory. They suggest there are many workers who would want shorter and more flexible work hours but not in low-wage jobs with no benefits. Or phrased differently, workers in secondary jobs prefer wages to hours whereas in the primary segment workers are more likely to be over-employed and seeking shorter and more flexible hours. This distinction, between hours preferences, in the different segments of the labor market, influences
the relationship between hours and compensation.

**Relation to compensation and productivity** In [RT95] and [LR95], an employer has an incentive to offer long hours in jobs where it is difficult to monitor employees productivity because the threat of dismissal is a great disincentive to shirking. Such a situation is most characteristic of a primary job. In this scheme, the employer has a disincentive to having shorter, more flexible hours because overall an employee will be paid less and fear the consequences of shirking less. From an employee’s perspective, shorter hours at the same rate may be more desirable because of the value they put on leisure. Since there is a disincentive to the employer to provide these shorter hours the employee is stuck with personally suboptimal hours of work. The benefits to employers, of flexibility, can change the incentives in this situation somewhat. If, for example, the shorter hours for employees involve some flexibility that the employer benefits from it can partially negate the disincentive of shorter hours created by the risk of shirking. Conversely, we can use this analysis to show that the benefit of flexibility are reduced in work environments in which employers must be concerned with effort regulation.

Within the secondary market none of the above give and take is likely to occur. Jobs are typically easy to monitor and therefore there is less of a reason to use rigid hours to regulate effort. In fact there is evidence that within this market workers find it is much easier to find the hours of work they desire. [Dey90] demonstrates the general flexibility of part-time work hours in Britain. Statistics also show that many people cobble together almost full-time work with a few part-time jobs, which demonstrates their ability to find a variety of work hours among short hour jobs.

**The competing needs of employers and employees** There is some inherent tension between what flexible work hours are preferred by employees and what type of hours reduce employer’s costs. In general, we assume that the closer work hours
match employee's preferences the greater satisfaction they have with their work. We can expect this to improve their productivity, a gain for the employer. With respect to absenteeism rates we can expect there to be an inverse relation between average hours worked per unit time and absence rates. This relationship has been born out in studies of job sharing and part time work arrangements [Ron84]. This relationship can make increases in call-in flexibility more beneficial to employers. Yet, a call-in work arrangement can make work hours less convenient for employees since true flexibility for workers would involve some autonomy in their choice of working hours [Chr90] and so, the benefits of shorter-hours can be canceled out for workers with irregular and unpredictable schedules.

To summarize, throughout this thesis we have presented workforce flexibility as a way to reduce the costs of stochasticity of the work environment, for the employers. From our consideration of the worker's needs we have seen that flexibility does not eliminate stochasticity, it rather shifts its burden from the employer to the employee. The labor management literature has shown us that workers are dissatisfied with the extreme flexibility of temporary employment. On the other hand analysis of our models has shown that autonomy and stability of hours, factors that could increase worker satisfaction, are exactly the factors that reduce the benefits of flexibility for the employer.

**Working time flexibility** Clearly, all things being equai. high productivities and low absenteeism are cost reducing, however, productivity rates maybe related to the number of hours worked and therefore the call-in guarantee and pay rate. This follows from effort regulation models such as [RT95]. In these models, a workers effort or productivity is, in part, determined by the hours and wage package. The reasoning here is that, the longer the hours, the more a worker has at stake, and therefore puts forth more effort. Here, the employer's imposition of standard long hours discourages
shirking and screens out less committed workers\textsuperscript{12}. As a result the benefits of call-in workers would not necessarily increase monotonically with decreases in $G$ because an employer may need to increase $C_{cw}$ to maintain high enough productivity levels.

For all crosstrained workers, any improvements they feel in job content can be easily canceled out by haphazard implementation of the policy. If crosstrained workers are frequently being moved from one job to another in a way that reduces their productivity and confuses any incentives for performance they may become increasingly frustrated with the arrangement.

\section{Hidden costs to employers}

In chapter 5 we have seen that the benefits of flexibility are strongly effected by the type of variability that exists in the work arrivals and in the correlation between arrival streams. We also saw that working time flexibility was strongly influenced by the information available to decision makers. There are a number of potential hidden costs that can limit the benefits to employers that we have not modeled. We characterize three types of potential "hidden costs" of flexible arrangements, those related to working time flexibility, those related to functional flexibility, and those related to the strategic planning of firms.

\textbf{Working time flexibility} However, the number of hours worked per unit time, pay rates and job content do impact turnover rate [Ron84] which incur costs not explicitly modeled here. Another hidden cost not captured in our objective function, is a potential reduction in the productivity of managers caused by an increase in flexibility. Flexible working time and functional flexibility can require managers to make many more and complicated staffing decisions than they have in the past or are

\textsuperscript{12}It also leads to workers not being open about their hours preferences.
trained to make. A call-in and or overtime policy requires line managers to spend
time deciding and then soliciting call-in workers and overtime shifts.

A final hidden cost is the impact on morale and productivity of a two tiered hu-
man resource policy. It is found that many companies that are progressive from a
work/family perspective, only do so with regard to full-time permanent employees.
They often have a two-tiered benefits and work/family policy for regular and con-
tingent workers, with attendant tensions between workers [Chr95]. In our models
work/family policies can also be viewed as fixed costs that apply to all workers and
therefore include flexible ones equally. In fact, many of the results of chapter 5 are
based upon parameter settings in which call-in workers receive the same pro-rated
compensation as regular workers. Many forms of benefit arrangements can be repre-
sented, but the hidden cost of those that cause the above tensions are not modeled.

Functional flexibility To achieve the benefits of functional flexibility a firm might
be forced to create its own supply of crosstrained workers. In cases where job content
is very firm specific, the burden of training workers, to have the correct mix of skills,
falls upon the firm and involves costs not modeled here. In these cases, there are also
few incentives for workers to want to participate in crosstraining. The firm may be
forced to pay higher wages to workers that join a crosstraining program. When the
skills needed for crosstraining are not firm specific, a firm that invests in training its
single skilled workers must worry about retaining them in the future.

Crosstraining places similar demands on managers while adding confusion about
supervision and responsibility for work. For example, does a worker crosstrained for
jobs A and B report to a supervisor of job A, or a supervisor of job B? If the worker’s
performance is poor in job B, which supervisor is responsible for it? The models
do not take into account the complexity of a work environment as a factor in the
objective function.

In the 2-job models we allow for differing productivity rates among workers as-
signed to different jobs, but we assume that throughout the planning horizon these are fixed. In reality we might expect that the productivity of cross-trained workers is effected by the amount of time they spend at each job and how often they switch from task to task. There may be learning and forgetting effects that, depending upon the length of a planning horizon, in absolute temporal terms, strongly influence productivity. We do not model these effects.

Short sighted policies A final concern about labor flexibility, from the employer’s perspective, that is often mentioned in the literature, is that it is a short-term strategy that in the long run fails the firm. The rational, as presented in [HB90], is that a trend toward flexibility is leading firms to “compete on the principal of cheap labour, rather than on the basis of technological improvement and the upgrading of the skills of their employees.”

First, in our models, there is no predisposition to represent flexible workers as “cheap labour”. Second, the concern that a focus on flexibility can be used as a short term substitute for technological and business process improvement, and thus cause the entrenchment of primitive business practices that are not viable in the long term, shows a very limited view of flexibility. This position may be valid within a manufacturing setting, but as discussed in Chapter 1, in many services the variability of the work environment is a persistent and short term phenomenon that demands a response. In chapter 5, we demonstrate that flexible work arrangements form a very competitive response. This point is made in [Wal90] when analyzing case studies of flexible work arrangements in hotels and retail stores in Britain.

6.5 Managerial and organizational obstacles to FWA

Most often, we can expect that when a firm is considering implementing FWAs it is for an existing operation. Such an operation has an existing staff of workers and
managers and work patterns. As is the case with any attempt at change within an organization there tends to be changing a system in which most participants are comfortable. Based upon our field observations we discuss some obstacles that we believe are relevant to the implementation of FWAs. In the mutual fund company the managers had false expectations of being able to eliminate the workload uncertainty. In the USPS facility, the managers made no attempt to reduce the uncertainty. In each of the two work environments, in different ways, the managers had difficulty integrating the uncertainty in the environment into their staffing decision making and planning.

6.5.1 False expectations

In the mutual fund company the customer transaction requests arrived each morning in the mail. Only once the envelopes had been opened and their contents sorted was it certain how much work there was for each department. The quantity of mail and its contents was highly variable since it was influenced by the economic situation, the behavior of the financial markets, the marketing activities of the company and the vagaries of the U.S. mail. The management of the facility we observed was almost obsessed with forecasting the work volume. They held the belief that with the right methods we should be able to predict with great accuracy the number of new account applications that would arrive each day. There was some limited success in using phone inquiry data to predict total weekly new accounts volumes, but translating those predictions into good daily forecasts for staffing purposes was not successful.

Implementing a FWA in this setting required the management to accept that there was a significant degree of uncertainty in the daily work volumes and to develop ways to reduce its cost impacts through new staffing policies. There was great resistance to this. It appeared that the cause of the resistance was fear of uncertainty and an incentive system strongly biased toward service. Adopting some FWA would most
likely mean a reduction in staff without a decrease in the work volume uncertainty. The managers found it difficult to accept that they would be able to perform better with fewer people. In this case, performance was mainly judged in terms of service level, or, how quickly customer’s transactions were processed. The struggle over introducing FWAs was actually a struggle between elements within the company that wanted to cut costs and those who felt that cutting cost would be at the expense of their ability to perform the tasks on which their performance is judged.

In this operation an inability to understand how staffing decisions could be made better without reducing uncertainty led to a view of the situation as the struggle described above.

6.5.2 "The time to panic is now"

During one visit to the mail sorting facility a supervisor described his modus operandi as: "The time to panic is now." By this he meant that he always assumed the worst and at the first sign of trouble would do anything he could to increase his staff, usually through overtime. This characterized the behavior of most of the managers. The individual tour\textsuperscript{13} managers are very experienced and capable at getting the mail processed quickly. I.e. if a pallet of mail arrives to the loading dock they know how to push it through the system as quickly as possible. When there is a disruption to the regular schedule such as receiving more mail than planned for that day\textsuperscript{14} or if the mail arrives to the facility late they are also effective at coming as close as possible to the scheduled distribution. Unfortunately, these situations are handled as crises and result in excessive use of overtime.

---

\textsuperscript{13}In mail processing operations the day is split into three tours with the most work being done in tours 1 and 3. Tour 3 starts at 3pm and ends at 11 pm. It is responsible for processing all outgoing mail. Tour 1 starts at 11pm and ends at 6:30am the following day. It is responsible for processing all in coming mail. Worker's hours are termed shifts and they can overlap the boundaries of tours.

\textsuperscript{14}The USPS prepares a plan for every day of the year for each facility. The plan specifies a forecast of mail volume for that day and number of work hours to be used in processing it.
Looking at a history of planned and actual mail volumes and work hours for 1994 we observed that uncertainty in mail volume did not explain deviations from planned work hours, (see appendix B). We conjectured that there were several possible causes of these deviations that could be predicted and adjusted too. To demonstrate that this was, in fact, the case we conducted a survey of supervisors\textsuperscript{15} at the facility in all tours over the course of 3 weeks in January 1995. The survey asked to note unusual events that required adjustments in staffing and what adjustments were made. We compared the responses with the work hour and mail volume data for the same period. With this data we could piece together a picture of what caused the use of overtime on different days. These causes tended to be predictable, a summary of this study appears in the appendix. For example, on one day mail arrived late to Tour 3 because bad weather slowed down the collection of mail that day. It stands to reason that it was possible for the branch offices responsible for collection to determine in the morning that there would be collection delays. If the manager of Tour 3 had this information he could adjust the start times of some shifts to avoid the subsequent overtime\textsuperscript{16}. This is rarely done.

As in the mutual fund company there is a failure to integrate the uncertainty of the environment into the staffing decisions. In the USPS, there is a similar incentive problem as well. A Tour manager’s primary responsibility is to process the mail in a timely fashion, the overall cost to the facility is not given much weight. In fact, we found that the Tour managers were not very involved in the plant’s budget process. Furthermore, aggressively seeking information to use in staffing decisions could in the end be counterproductive. Ultimately, such a policy would reduce overtime and

\textsuperscript{15}Supervisors are members of the management who have responsibility over individual or clusters of, processing operations.

\textsuperscript{16}According to the contract between the USPS and the American Postal Workers Union, workers are paid a overtime premium for hours worked outside of their regularly scheduled hours [Ser94]. As a result the savings are in fewer hours worked at regular pay for full-time regular workers. For the flexible workers there is no such out-of-schedule premium.
anger workers. In this particular case the reliance of worker’s on overtime and the power of the unions made the actual implementation\textsuperscript{17} of FWAs more difficult, if not impossible.

6.5.3 Summary

From these two examples we see how difficult it can be, in practice, to actually implement novel work arrangements. If we were able to represent all the myriad of labor supply issues, and to ‘reveal’ the hidden costs to employers, so that the models were perfectly accurate, there would still remain obstacles to successful implementation.

6.6 Summary and implementation guidelines

We now summarize what we have learned about implementing FWAs from considering the supply side of the flexible labor market and the hidden costs of, and obstacles to FWAs. Based upon these learnings we suggest some guidelines for how to approach implementation. The following three points are clear from out discussion:

- (1) There is an unavoidable tension between employer and employee flexibility needs.

- (2) The relationship between employment packages and employee parameters must be understood within the context of the appropriate sector of the labor market.

- (3) Job security has a strong effect on many of the labor supply issues, even though it is beyond the scope of the models themselves.

\textsuperscript{17}We note that officially flexible schedules were written into the labor contracts, but in practice it did not appear that they were implemented.
6.6 Summary and implementation guidelines

6.6.1 Tensions

While we cannot eliminate this tension, the models provide a framework for achieving a compromise between these competing objectives. We can represent some of the worker preferences mentioned here and see their effect on the benefits of flexibility for the employer. As was discussed in the analysis section of chapter 2, the parameters of the model allow for the designation of very complex employment packages for call-in workers. Based upon our field experience, we also suggest some common sense approaches to reduce the burden of flexibility on the workers. We then outline the main concerns for employers in satisfying employees needs within primary and secondary labor markets.

Reducing tension via the model: The guarantee, \( G \), not only determines the minimum wage they can expect to receive, it also determines the regularity of work hours. In analyzing the trajectories of call-in usage in chapter 5 we saw that as \( G \) increases the variation in call-in shifts used each period decreases. We have, thus far, mostly described the notification scheme as a representation of the information available to the manager at decision making times. We recall from chapter 1 that our definition of the different schemes is really from the perspective of the notice required by the call-in workers to be expected to come into work. The more notice a worker has, the more flexibility they have in planning their personal lives around work hours. Clearly, this situation highlights the potential tensions between the needs of employers and employees since same-period notification, which gives workers very little flexibility, provides employers with the greatest benefits from call-in arrangements.

The final problem parameter that determines the flexibility of a work arrangement for workers is the absenteeism rate. The way the problems are structured call-in workers receive the guaranteed portion of their compensation regardless of their absences.
6.6 Summary and implementation guidelines

We can expect, if a worker is absent many more times than the average rate, they would probably be replaced. Nonetheless, we can think of the absenteeism rate of call-in workers as a measure of their autonomy to choose the periods in which they work. The absence rate enables the model to take into account the worker's degree of autonomy in determining the performance of a particular work arrangement.

Our models do not seek to optimize working conditions for employees by determining the arrangements that maximize their combined utility from working hours flexibility and compensation. However, the models do offer quite sophisticated ways to represent worker preferences. The result is that an employer that is developing a workforce management policy, that is sensitive to the welfare and interests of workers, has a way of representing this sensitivity with the model's parameters\(^{18}\). What the models cannot do, is represent any relationships between an employer's sensitivity to worker's flexibility needs and their performance\(^ {19}\).

Reducing tension with common sense: In the mutual fund company we observed, when work loads were light workers could be sent home, receiving the nearest two hours worth of pay. In this company, the work was performed in two shifts, morning and afternoon. In the morning, it was not clear how much work had arrived in the mail, and therefore it was not possible to cancel shifts, before the morning shift had arrived to work. Usually the amount of work was known before the afternoon shift arrived and therefore there was time to cancel those shifts. From the firm's perspective, it was optimal to only cancel the afternoon shifts based upon complete information. However, afternoon workers did not want to lose work hours and therefore some of the burden of cancellation was shifted to morning shift people, who were sent home early based upon partial workload information. Since these workers were

\(^{18}\)Of course, an employer that is not sensitive to worker's welfare and interest may use the models as well.

\(^{19}\)The evidence in this area is mostly anecdotal. Modeling these relations is an interesting area for future research.
already on site they were payed for at least two hours and the staffing decision was made with much less information than in the optimal situation.

One possible solution to this problem is to hire morning and afternoon workers under different agreements. Morning shifts could be regular workers and afternoon workers could be call-in workers who receive a guaranteed minimum pay, as in the models in this thesis. Another possible solution is to only cancel afternoon shifts but to rotate workers between morning and afternoon shifts from day to day or week to week. The first solution places the entire burden of hours flexibility on one segment of the workers but compensates them for it. The second solution spreads the flexibility among all workers. In both cases the employer makes shift cancellation decisions with the most information, which we have seen to be a decisive factor in its effectiveness.

6.6.2 Employment packages and employee parameters

Our discussion of employee needs and preferences has been based upon dual labor market theory, with our distinguishing between FWAs in primary and secondary sector jobs. Here we suggest some guidelines for how employers implementing FWAs in these different sectors should think about the hours, wage packages they offer.

Secondary jobs: Workers in secondary jobs are more concerned with being compensated enough rather than finding jobs with the right number of hours. We conclude this from our discussion of the effects of flexibility on workers and our observation of the USPS and mutual fund company.

In the mail sorting plant we found that although there exists a part-time flexible work option, in the union contracts, it was not utilized. Rather, flexible positions were entirely non-career positions. We also saw that there was a great demand for overtime. This suggests that there was greater interest among workers for access to overtime than flexible hours. Since the workers are members of a powerful union,
the flexible jobs which could potentially involve less overtime, were delegated along union membership and seniority lines, as opposed to hours preferences. In the mutual fund company, we found that the jobs were part time by design, to satisfy worker's needs. Nonetheless, the workers were very protective of their weekly hours because they wanted a guaranteed steady stream of pay.

Based upon these observations, an employer that operates in the secondary labor market, and wants to attract workers that are productive and dependable, must be most concerned with protecting worker's income when introducing flexible work arrangements. The implication of this guideline is that an employer should implement flexibility in a way that reduces the absolute staffing level, while maintaining the salary levels of remaining employees.

**Primary jobs:** Workers in primary jobs are more likely to be working more hours than they would like, than workers in secondary jobs. As a result introducing flexibility into the work place has a potential to benefit both employers and employees. Flexible working time arrangements can reduce the hours worked for employees bringing them more in line with their preferences. For employers, the shorter hours may limit their ability to regulate effort but the potential loss in productivity can be compensated for by the benefits of flexibility. We have also noted that: even the primary jobs that are suitable for flexible arrangements tend to be easy to monitor thus reducing the shirking problem.

Based upon these observations, an employer that operates in the primary labor market, and wants to successfully implement flexible work arrangements should be concerned that the arrangements either are both not burdensome on employees and provide them with a measure of flexibility. i.e. pay premiums to compensate for uncomfortable schedules must be relatively high for primary workers.
6.6 Summary and implementation guidelines

6.6.3 Job security

Throughout this chapter we have discussed issues that can all be influenced by job security. As can be seen from environments like those in our field studies, job security can make a secondary sector job much more like a primary sector job\textsuperscript{20}. Job security also reduces the wages and therefore flexibility premiums an employer might pay. Maybe most importantly, job security for flexible workers means that they are not marginalized for providing flexibility. This implies that they be treated to the same employment benefits and support programs as regular employees. This can lead to eliminate or lessen tensions between haves and have nots in a workplace [Chr95].

In this chapter we have seen that there are many aspects of implementing FWAs, that are not modeled in this thesis. This is not a tragedy, not everything should be modeled mathematically. It is, however, important to have an understanding of the influence of those factors, not modeled, on the use of the models. We have tried to develop that understanding here.

\textsuperscript{20}Postal workers are for the most part career employees and the mutual fund associates have an average job tenure of 3 years, for a job they have high productivity at in just a few months.
Chapter 7

Conclusions and future research

Our goal in this thesis was to answer the following questions about labor flexibility:

(1) What is it?

(2) What is it used for?

(3) How do you decide what kinds and how much flexibility to utilize?

We defined labor flexibility as the degree to which a firm can adjust its workforce to its actual demand for labor. We created a high level mapping of types of flexibility to work environments in terms of the time scale of the variability in the environment. We chose to focus on environments with variability on the scale of days and weeks, since we found the literature sparse in this area. In answering questions (2) and (3) we made the following contributions:

7.1 Contributions

- We created a taxonomy for flexible workforce management problems in these environments.
• Based upon the taxonomy we formulated a family of models that can serve as tools to answering question (3).

• We analyzed the models to develop our ability to use the models as tools in a computationally practical way.

• We conducted numerical experiments that showed how closely linked the benefits of different type of flexibility are to the specific work environment, as defined by our taxonomy. Namely, the notification regime and the nature of the variability of the system determine the usefulness of flexibility, thereby answering question(2).

• We showed how the forms of flexibility modeled in this thesis fit within the broader workforce issues discussed in the human resource management and labor economics literature. In doing so, we demonstrated, drawing upon dual labor market theory, how successful approaches to implementation of flexible work arrangements must take into account the sector of the labor market a firm is operating in.

This thesis represents a step into a problem area that is underdeveloped. An effort has been made to present a framework for future research in this area. As a result the research presented here leads naturally into several interesting directions for more research, which we outline in the next section.

7.2 Future Research

Future research areas motivated by this thesis can be divided into those that represent direct extensions of the work done here and those that involve the formulation of completely different models.
7.2 Future Research

7.2.1 Direct extensions

Backlog As noted in the introductory chapter there are many ways to represent backlog and we have really only modeled one of these in this thesis. We can represent different forms in this thesis without changing the aspects of the models governing the dynamics of the system behavior.

Computational performance We have noted in Chapter 4 that the multiple job models, as formulated in this thesis are difficult to work with from a computational perspective. At the end of that chapter we have suggested some possible research to address this limitation.

Networks of jobs We have only modelled jobs performed individually or in parallel. There are many work environments in which distinct jobs are organized in networks where the workload in one is determined by the output of another.

7.2.2 Related new models

Job security In discussing the role of FWA in job security we said that flexibility can make work sharing arrangements more attractive to employers. Currently the models assume a stationary work arrival distribution. This foregoes any need to hire or layoff workers once a staffing level has been set. If the work arrival process were to change, e.g. if its mean were reduced, the need for labor would decrease as well. We would like to consider how the option of changing workers from regular status to flexible status can compare with layoffs. To achieve this we would need to simplify the models and then model longer time horizons that have non-stationary arrival processes.

Employment packages In this chapter, we have discussed how hours, pay, and productivity of workers are linked. A richer model of flexible staffing would involve
employers deciding on employment packages of hours and pay in addition to how many worker to staff. The employment package would determine the productivity of workers which would partially determine the number required and the way in which they are utilized. A model like this would link the supply and demand sides of the labor market.

Learning and forgetting  Productivity of workers is also determined by their training and experience, not just their employment package. When workers are crosstrained they do not necessarily perform each job they are trained at. frequently enough to maintain proficiency. I.e. some forgetting of skills may occur. Conversely, consistent assignment to a job may increase a worker's productivity as a result of learning. It would be interesting to model the productivity of a workplace utilizing crosstrained workers. One decision variable for employers would be the amount of initial training to provide a worker in anticipation of learning and forgetting on the job. An extension of such a model would be one that scheduled refresher training sessions for workers.

Temporary help providers  It is also possible to think of the staffing problems presented in this thesis from the perspective of a firm that provides temporary help to other firms adjusting their workforces. The temporary agency must decide how many and what types of contracts to enter with firms who's demand for labor is uncertain.
Bibliography


Appendix A

A.1 Work arrival processes

In the numerical experiments conducted in section 1 we have drawn upon the following set of work arrival processes: $W_1, W_2$ and $W_3$. Each of these processes is i.i.d. and has a probability mass function graphed below.

**W1 and W3** Work arrival process $W_1$ is a scaled representation of the arrival of new accounts applications to a large mutual fund company in Boston, and $W_3$ is a higher mean lower variance process generated for contrast. In figure A-1 we depict the probability mass functions for $W_1$ and $W_3$ which have means 5.4 and 9.1 units of work per period respectively and coefficients of variation .56 and .2 respectively.

**W2** Arrival process $W_2$ is generated by combining a two part process used for the mixed information experiments. It was chosen arbitrarily as a high mean and low variance inputs to contrast with $W_1$. In figure A-2 we depict the probability mass functions for the two part and combined $W_2$ arrival process. $W_2$ has mean 8.6 and coefficient of variation .23. Part 1 represents the arrival of information before call-in decisions are made, and part 2 represents the arrival of information after the call-in decision is made but before the overtime decision. Note that part 1 represents a substantially smaller portion of the total work arrivals in a period than part 2.
A.1 Work arrival processes

![Graph showing W1 and W3 work arrival probability mass functions.]

Figure A-1:

**W4** arrival process W4 is used in the experiments to show the benefits of call-in flexibility in the presence of crosstraining. It is generated by adding the arrivals from two W1 processes to get an i.i.d. process with mean 10.8 units of work per period and coefficient of variation .4. Note that this variability is less than that of the W1 process take individually.

**2-job work arrival processes** For the 2-job problems we generate joint pmf's \( \Delta_j(d^1, d^2) \) that have a correlation \( \rho_j \) and are based on one of the 1-job arrival processes in the sense that the marginal pmf's of \( d^1 \) and \( d^2 \) follow the same pmf as the 1-job arrival process. We generate these \( \Delta \) as follows:

If for example we want to generate a joint pmf \( \Delta_j(d^1, d^2) \) that has correlation near .5 and marginal PMFs like W1 we follow the following steps:

1. For every possible value of \( d^1, d^1_i \) find the pair \((u_i^0, u_i^1)\) such that the cumulative probability of \( d^1_i \) is equal to \( u_i^1 \) and the cumulative probability of \( d^1_{i-1} \) is equal to \( u_i^0 \). We do the same for \( d^2 \) but in our case the two arrival streams are symmetric so we present things in terms of \( d^1 \).

2. Using the inverse of the cumulative standard normal density function map the pairs \((u_i^0, u_i^1)\) to pairs \((y_i^0, y_i^1)\). We actually use an approximation to the inverse of the
standard normal CDF:

\[ \Phi^{-1}(u) = \frac{u^{.135} - (1 - u)^{.135}}{.1975} \]

At this point for every pair \((d^1, d^2)\) we have two pairs \((y^1_1, y^1_1)\) and \((y^2_2, y^2_1)\) that mark off a rectangular region in \(R^2\).

(3) For every pair \((d^1, d^2)\), using a standard bivariate normal CDF with correlation \(\rho = .5\) we find the probability of a standard bivariate normal random variable falling within the square in \(R^2\) determined in step (2). This probability is the probability mass we assign to \(\Delta(d^1, d^2)\).

In general the resulting PMF, \(\Delta(d^1, d^2)\), will not have a correlation that is exactly .5. In our case since the marginals of \(d^1\) and \(d^2\) have PMFs as in \(W1\) which are close to binomial, the correlation of \(\Delta(d^1, d^2)\) turns out to be remarkably close to .5.

### A.2 Parameters

Parameter set 1 is used in the experiments demonstrating the determinants of the benefits of working time flexibility. They are:
A.3 Technical issues

\[ C_{rw} = C_{cw} = C_{ot} = 1.0 \text{ and } C_f = 0 \]

\[ C_b = C_B = 1, \text{ and } V = 20 \]

\[ pi_{rw} = \pi_{cw} = \pi_{ot} = 1 \]

Parameter set 2 is used in the experiments about call-in and backlog trajectories, they are:

\[ C_{rw} = 1.2, C_{cw} = 1.0, C_{ot} = 1.5 \text{ and } C_f = 0 \]

\[ C_b = 1, C_B = 1.5, \text{ and } V = 20 \]

\[ pi_{rw} = 1.0, \pi_{cw} = .8, \pi_{ot} = 1.2 \]

A.3 Technical issues

Finite horizon issues We will now consider the question of what is the appropriate way to take into account the effect of final period backlog on the future performance of the system. Since we have constructed finite horizon models our primary concern is that late in the planning horizon we will start to accept higher levels of backlog than would be realistic. Setting the final period backlog cost at a very high level, to force the system to empty by the end of the planning horizon artificially creates a significant difference between our tolerance for backlog in the final period versus others.

Our approach to this problem is to optimize the staffing level over multiple planning horizons, rather than just one. That is, if the planning horizon has length \( V \), we would solve the staffing problem over \( kV \) periods for some integer \( k \). Every \( V \) periods we reset the call-in guarantee remaining to \( MGV \). In this scheme we do not assign any special penalty to backlog remaining at the end of the \( kV \)th period. The larger the value of \( k \) the less the edge effects.
A.3 Technical issues

In this section we have presented numerical results for specific problem instances that are relatively insensitive to the finite horizon. We found that after at most 3 planning horizons the end of planning horizon backlog stabilized and that the optimal staffing levels did not change from those determined for the single planning horizon. This need not always be the case and can be explained by the backlog penalty. The one period backlog penalty is high relative to the labor costs, therefore backlog is kept to a very low level in all periods. This means that successive planning horizons are almost independent of one another.

**Truncated state space** None of the model formulations implemented in this chapter have explicit constraints limiting the amount of backlog that is carried in the system, and therefore place no limits on the total amount of work in the system. In implementing the models we put an artificial constraint on the work in the system to limit the memory and computational requirements of solving the dynamic programs. In all cases we limit the amount of backlog that may be carried from one period to the next to be the maximum possible work arrival. Any staffing decision that would result in a larger backlog is given a very large penalty. We then truncate the work dimension of the state space to a value twice the maximum possible work arrival in a period. This approach is reasonable since no service providing facility would tolerate such large backlogs. Furthermore, a facility that experienced such backlogs regularly would quickly become unstable.
Appendix B

B.1 Causes of Unplanned Overtime in Mail Sorting Facility

There are two possible reasons for unplanned overtime on any given day: 1) actual mail volume exceeds planned mail volume, and 2) actual productivity is lower than planned productivity. Based upon the data it appears that both reasons are at play here. In figure 2 we graph the percent error in hours and volume from 10/1/93 to 2/25/94. In figure 3 we do the same for the second data set. It is easy to see that there are many days on which the actual volume was less than the planned volume yet there are more actual work hours than planned work hours. These days demonstrate that there are productivity effects that are affecting the use of overtime. There are several factors that impact productivity, namely: Mail Volume, Time of mail arrival, Mail content, and Machine failure.

- Mail volume - It seems that since much of the automation equipment is under utilized on light volume days the marginal increase in labor needs, caused by an increase in volume, would initially be small thereby increasing productivity. That is, for any staffing level the planned productivity level can be improved with moderate increases in volume. However, a very large increase in volume might stress the system and degrade productivity. On the other hand, if mail
volume is less than planned, productivity will be lower than planned because the facility is overstaffed (another way of thinking about paid lost time).

- Time of mail arrival - Most workers in the facility work regular schedules. As a result, if mail arrives early it will have to wait. If mail arrives late it means that workers will be idle while waiting for it, resulting in a productivity reduction.

- Mail Content - For example, if a high proportion of the mail is machine readable than productivity will be higher than if much of the mail must go through the LSM's. Another example of mail content impacting productivity is if the mail is wet and has trouble going through automation.

- Machine Breakdown - The highest productivity is achieved with the automation equipment. Any machine failures in these areas can greatly reduce productivity for the entire plant. Similarly, problems with the remote ISS site can have the same effect.

There are obviously many more possible "glitches" in the system but these seem to be the most serious ones.

### B.2 Observations

In looking at the FHP volume and work hours data I computed the percent difference (or error) of the actual value vs. the planned value. That is I computed $100\times(\text{Actual-Plan})/\text{Plan}$. I then compared the volume errors with the hours errors. In an ideal scenario we would want the two values to be highly positively correlated. This would mean that when volume was higher than expected so would work hours and when volume was lower than expected we would want work hours to be lower than plan, as well. On the other hand we know that since there are many factors effecting productivity we should not expect the two errors to be linearly related. By linearly
related we mean that every one percent increase of volume over plan will not result in the same percent increase in work hours. However, if volume is close to plan or below plan we would want hours to be close to plan as well. If we look at Figure 4 we can see that there are many days in which the percent error in volume and hours are very different. We will use the results of a survey of supervisors conducted between 1/12/95 and 1/27/95 to figure out what events caused these deviations\(^1\). The following are notable days on the graph (for which we have some more information):

<table>
<thead>
<tr>
<th>Date</th>
<th>Planned Vol</th>
<th>Actual Vol</th>
<th>% Diff</th>
<th>Planned Hrs.</th>
<th>Actual Hrs.</th>
<th>% Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/8</td>
<td>1349</td>
<td>1913</td>
<td>42</td>
<td>2251</td>
<td>2509</td>
<td>11</td>
</tr>
<tr>
<td>1/9</td>
<td>2533</td>
<td>2180</td>
<td>-14</td>
<td>3958</td>
<td>4182</td>
<td>6</td>
</tr>
<tr>
<td>1/12</td>
<td>2794</td>
<td>2680</td>
<td>-4</td>
<td>4087</td>
<td>4273</td>
<td>5</td>
</tr>
<tr>
<td>1/13</td>
<td>2864</td>
<td>2754</td>
<td>-4</td>
<td>4072</td>
<td>4307</td>
<td>6</td>
</tr>
<tr>
<td>1/16</td>
<td>1795</td>
<td>2139</td>
<td>19</td>
<td>2785</td>
<td>3005</td>
<td>8</td>
</tr>
<tr>
<td>1/18</td>
<td>2796</td>
<td>2844</td>
<td>2</td>
<td>4144</td>
<td>4376</td>
<td>6</td>
</tr>
<tr>
<td>1/19</td>
<td>2831</td>
<td>2876</td>
<td>2</td>
<td>4173</td>
<td>4398</td>
<td>5</td>
</tr>
<tr>
<td>1/22</td>
<td>1378</td>
<td>1588</td>
<td>15</td>
<td>2295</td>
<td>2414</td>
<td>5</td>
</tr>
<tr>
<td>1/23</td>
<td>2643</td>
<td>2492</td>
<td>-6</td>
<td>4087</td>
<td>4235</td>
<td>4</td>
</tr>
<tr>
<td>1/27</td>
<td>2953</td>
<td>2910</td>
<td>-4</td>
<td>4225</td>
<td>4340</td>
<td>3</td>
</tr>
<tr>
<td>1/28</td>
<td>1352</td>
<td>1783</td>
<td>32</td>
<td>2128</td>
<td>2037</td>
<td>-4</td>
</tr>
</tbody>
</table>

Table B.1: Planned and actual mail volume and hours worked on days with 'unusual' behavior. Note: Volume figures are in thousands of pieces of mail.

Interpretations: 1/8: On this day the actual volume was 40% higher than planned but the actual hours used were only 11% more than planned. This means that productivity was significantly higher than planned. The most obvious explanation for this is that the plan would have under utilized the processing capacity of the workers. I.e. 1/8 was a Sunday and planned volume was low enough that the automated sorting machines were not working at their full speed (note the actual volume is significantly lower than the average for the graph of 2,300,000 pieces per day). The extra volume that arrived caused the regularly scheduled automation workers to become more productive since they now could run their machines at full speed.

\(^1\)An example of a survey form appears in Exhibit 1.
B.2 Observations

1/9: On this day actual volume was 14% lower than planned but actual hours were 6% more than planned. We do not have any more detailed information for this day to determine why productivity was lower, but we can notice that this low volume day immediately followed a high volume day and begs the question: Could we have predicted the low volume on 1/9 and adjusted staffing in some way?

1/12 and 1/13: For these days there was not much written in the surveys except for their being snow on the 12th and some problems in Flats on the 13th. On both days volume was lighter than expected but more hours were needed. We could guess that the snow delayed the arrival of mail to the facility. In any event, there usually is warning of snow so there could have been staffing adjustments made. For example, if branch offices responsible for collection were having difficulty with operations because of snow they would have time to inform the M/E GMF.

1/16: This was Martin Luther King's Birthday and volume was higher than expected (19%) but hours were only 8% higher, i.e. higher than planned productivity. The explanation for this is probably the same as on Sunday the 8th.

1/18 and 1/19: On both these days volume was very close to plan yet a total of 240 unplanned overtime hours were used. According to the survey on both days mail arrived late.

1/22 and 1/23: These are a Sunday, Monday pair like the 8th and 9th, with higher than expected volume on Sunday followed by lower volume on Monday. Also, the surveys report that on 1/23 there were a lot of pallets on the platform and a heavy load of flats. These reports suggest that the mail content affected the productivity.

1/27: Volume was a little bit less than planned yet 115 hours of unplanned overtime were used. The surveys report that Tour 3 received a lot of late meter mail that day.

1/28: This is another day of generally low volume in which the higher than planned volume caused productivity to improve. We have no information about what happened to some of the planned hours that were not used.
To summarize the above analysis: We saw that productivity variation occurs quite often and that a little bit of added information can suggest reasonable explanations for its causes. These explanations also suggest that much of this variation is within our control.
Exhibit 1: Sample survey sheet from mail sorting facility

NAME: C. Casey / A. Scaggs
OPERATION: 831/881 871/872 874/894
DATE: 1/2095

1. What is your staffing for today? 45
   How many absences? 1

2. Any events alter your operation today? (late arriving mail, staffing problems, machine breakdown, etc.)
   JUST WET MAIL

3. Did you use unscheduled overtime to finish your operation today? YES

4. Comments related to workhour/mail volume relationship for today.
   (A lot of the mail was wet, wet, wet. Heavy rain all day.
   Mail running slow threw machines due to rain.)