Studies on tactical capacity planning with contingent capacities

Studies on Tactical Capacity Planning with Contingent Capacities

Gergely Mincsovics

Tactical capacity planning of production and services with contingent capacities / by Gergely Mincsovics. – Eindhoven : Technische Universiteit Eindhoven, 2008. – Proefschrift.

A catalogue record is available from the Eindhoven University of Technology Library

ISBN 978-90-386-1385-7

NUR 804

Keywords: Capacity Planning / Contingent Capacity / Temporary Capacity / Temporary Workforce / Capacity Acquisition Lead Time

Printed by Printpartners Ipskamp, Enschede, The Netherlands

Cover designed by Paul Verspaget

Inside: digital print, size B5 (170x240mm), 100g light coated paper-white, 1/1 black (double-sided black), milled-glued finishing.

Cover: 4/0 onesided full color, 240 grams white single-sided sulphate cardboard, gloss laminate.

Studies on Tactical Capacity Planning with Contingent Capacities

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op maandag 3 november 2008 om 16.00 uur

door

Gergely Zoltán Mincsovics

geboren te Boedapest, Hongarije

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. J.W.M. Bertrand
en
prof.dr.ir. J. van der Wal

Copromotor:
dr.ir. N.P. Dellaert

Acknowledgements

The four years of research resulting in this thesis has been supported by many in a direct or indirect way. It is hardly possible writing a totally exhaustive list of my acknowledgements. Nevertheless, there are some people who deserve special thanks.

Dr.ir. Nico Dellaert was not only my daily supervisor, but also a co-author of mine multiple times. I especially appreciate the large amount of time he dedicated to me, and invested in our joint scientific papers. I am particularly indebted for his patience, his guidance in research, and his confidence in me, as well as for the support I got from him to learn more about real-life engineering situations and to start using the Dutch language at work.

I am grateful for prof.dr.ir. Will Bertrand for being my leading, first supervisor. His explanations made me understand what engineering really is. I will try to follow his consistency in sustaining his high methodological standards in research and engineering. I particularly admire his constant availability, which was strongly helping in finishing this thesis on time.

I would like to thank prof.dr.ir. Jan van der Wal for accepting the second supervisor role, meaning involvement in the preparation of this thesis and reserving the time for our many meetings. His comments initiated revisions that created a stronger cohesion and essential enhancements in the mathematical correctness.

I am thankful to Tarkan Tan and Osman Alp, my co-authors, for allowing me to join their already fruitful cooperation. From them I learned a lot, primarily on publication style and communication. Their experience and professional insights was a guarantee for a high quality output.

Let me also express my gratitude to prof.dr. René de Koster, dr.ir. Ivo Adan, and prof.dr. Nico Vandaele who kindly accepted to participate in the core committee to judge the quality of this thesis. Their comments and suggestions were very much welcome, as these have brought valuable improvements towards the final version. All these acknowledgements are also intended for prof.dr.ir. Geert-Jan van Houtum, external thesis committee member, as well, whose building critics exhaustively addressed the entire thesis.

I would like to thank for the pleasant time for my former officemates (in chronological order) Erik Winands, Barıs Selçuk, Chihyun Jung, and Çagdas Büyükaramıklı, for the secretary, Florida Stritzko, Ineke Verbakel-Klok, Hennie van Gastel-Dinghs, for the faculty members, Gudrun Kiesmüller, Simme Douwe Flapper, Tom van Woensel, prof. Jan Fransoo, Marco Slikker, prof.

Ton de Kok, Rob Broekmeulen, Karel van Donselaar, Henny van Ooijen, Kai Huang, for the older PhD-candidate generation, who have already got their PhD degree, (in chronological order) Pieter van Nyen, Judith Spitter, Mustafa Kemal Doğru, Pim Ouwehand, Bram Kranenburg, Ulaş Özen, for the young generation, who are the present PhD-candidates like me, Ingrid Vliegen, Alina Curseu-Stefanut, Youssef Boulaksil, Kurtulus Öner, Ingrid Reijnen, Ola Gabali, Said Dabia, Michiel Jansen, Ben Vermeulen, Gönül Karaarslan, Kristel Hoen, and for the further members of the Operations, Planning, Accounting, and Control subdepartment, as well as for our visitors, including Özalp Özer, Charles Corbett, Steffen Klosterhalfen, Marko Jakšič, Yong Yue, and Gabriella Muratore.

There are plenty of people who gave me good advice, important counsel, made me develop my skills, or was just pleasure to talk with. I would like to thank them for the time I spent with them.

Onno Boxma and the whole Eurandom group, Jacques Resing, Bernardo D'Auria, Alexander Wolff. My Dutch teachers Nelleke de Vries, Elly Arkesteijn, Pieter uit den Bogaart, my Dutch course classmates; my LNMB classmates including Hossein Mansouri, Regina Egorova, as well as my teachers of other courses, Duncan Harkness, Astri Keizer, and my classmates there. The students I supervised in part: René Quirijns, Marthe Uitterhoeve, and Mark Stegeman. From the Information Systems subdepartment: Sven Till, Nataliya Mulyar, Ronny Mans, Alex Norta, and Monique Jansen-Vullers, from the Human Resource Management subdepartment: Eric van der Geer-Rutten-Rijswijk, Daphne Dekker, Floor Beeftink, and Marieke van den Tooren, from the LMS group: Twan Geenen, Gustavo Nellar, Maria Martinez Vilela, and Ildilkó Meertens, the former Quality and Reliability Engineering subdepartment, as well as the former and the present members of the Beta PhD Council. The IT support team (Bureau Automatiseering) including Koos Huibers, the Department Office team, Geertje Kramer and Jolande Matthijsse-van Geenen who organize the Beta Research School, Nel Evers and the present Personnel Department, the staff of the TM-library and that of the Tuimelaar. The OR in Health Care group, including Jan Vissers, Erwin Hans, Rafael Velásquez, Jeroen van Oostrum. Colin Ridley, who helped my research supplementing me with data and providing a lot of related information. Dragan Banjevic and Joseph Tan, who gave me useful advice. The Hungarian community of Eindhoven and Maastricht, my friends and family in Hungary; and my own family: my wife, Yu Da, and our daugther, Lina.

Gergely Mincsovics

August, 2008

Contents

1 Introduction **1**

1.1 The research topic under study 1

1.2 Characteristics of manufacturing systems and services 3

1.3 Research goals 4

1.4 Related literature 5

1.5 Research questions and methodology 7

1.6 Outline of the thesis 11

2 Integrated Capacity and Inventory Management **13**

2.1 Introduction and related literature 14

2.2 Model formulation 16

2.3 Analysis of the optimal policies 19

2.3.1 Optimal policy characterization 20

2.3.2 Note on the concave productivity function 22

2.3.3 Complementary slackness 23

2.4 Numerical results and discussion 25

2.4.1 Value of flexible capacity 26

2.4.2 Sensitivity analysis on $VFC\%$ for demand forecast error 28

2.4.3 Optimal level of permanent capacity 30

2.5 Conclusions and future research 31

3 Workload-Dependent Capacity Control in To-Order Systems 41

3.1 Introduction 42

3.2 Related literature 44

3.3 Model formulation 45

3.3.1 Definitions, assumptions and problem formulation . . . 46

3.3.2 Evaluation of cost functions 49

3.3.2.1 Derivation of the throughput time distribution 51

3.3.2.2 Moment approximation of the throughput time 52

3.4 Results 53

3.4.1 Value of workload-dependent capacity control 54

3.4.2 Sensitivity analysis on $VFC\%$ for estimated interarrival rates 58

3.4.3 Characterization of workload-dependent policies for uncapacitated, large scale settings 60

3.4.4 Characteristics of workload-dependent policies vs. fixed policies 61

3.4.5 An illustration to the use of workload-dependent policies in real-life 62

3.5 Conclusions and future research 63

4 Permanent-Contingent Budgeting in Services 67

4.1 Introduction 67

4.2 Models 70

4.2.1 Common modeling aspects 70

4.2.2 Overview of the models 70

4.3 Linear capacity shortage costs and restricting budget 71

4.3.1 Model development and analysis 72

4.3.2 A special case: gamma-distributed demand 73

4.3.3 Numerical illustration 74

4.4 Linear capacity shortage and budget deviation costs 76

4.4.1 Model development and analysis 76

4.4.2 Numerical illustration 78
4.5 Halfway quadratic capacity shortage costs 78
4.5.1 Halfway quadratic shortage costs and restricting budget 79
4.5.2 Halfway quadratic shortage and budget deviation costs . 80
4.6 Halfway relative quadratic capacity shortage costs 80
4.6.1 Halfway relative quadratic shortage costs and restricting budget 81
4.6.2 Halfway relative quadratic shortage and budget deviation costs 81
4.7 Results 82
4.7.1 Budget spending pattern 82
4.7.2 Sensitivity analysis on the budget spending pattern for demand forecast error 83
4.7.3 Shortage and shortage cost patterns 84
4.7.4 Budget deviations 86
4.8 Conclusions 87

5 Conclusions and Future Research 91
5.1 Answers and main conclusions 91
5.1.1 Production-to-stock aggregate planning 91
5.1.2 Production-to-order 94
5.1.3 Services 95
5.1.4 General conclusions 97
5.2 Few words on machines and other facilities 98
5.3 Comparison with subcontracting and dual sourcing models . . 99
5.4 Future research 101
5.4.1 Production-to-stock 101
5.4.2 Production-to-order 102
5.4.3 Services 102
5.4.4 An environment-unified model 103

References **105**

Summary **111**

Samenvatting **113**

About the author **115**

Chapter 1

Introduction

This chapter describes the studied research topic and summarizes the content of this thesis. First, the research topic is in our focus, where we elaborate on definitions, discuss the main representative articles, and position our work. Afterwards, we pose research questions, to be answered in the later chapters. Description of the research methodology used and the steps towards gaining the answers follow the research questions. Finally, we give an outline of the rest of the thesis.

1.1. The research topic under study

This thesis investigates tactical capacity planning of single stage, single item production-to-stock, production-to-order and service environments, which face stochastic demand and which, in response to demand uncertainty and fluctuations, can employ contingent capacity next to their in-house, permanent capacity.

By *tactical planning*, we refer to the middle level, medium term decision making in the framework of Anthony's hierarchical classification (see Anthony (1965)). The tactical level of decision-making is situated between the long term, strategic level (e.g. facility location planning), and the short term, operational level (e.g. job scheduling). On the one hand, the decisions at the strategic level constrain or direct the decisions at the tactical level. A very common example of constraining the tactical level decisions at the strategic level is to create budgets restricting the expenses for a given period of time. On the other hand, by decision-making at the tactical level one needs to anticipate the system's operational capabilities and flexibilities. The anticipation means

creating a simplified, aggregated view on the system's behavior. This simplification is often necessary because of the complexity at the tactical level of decision-making. These two practices (constraining and anticipating) are used overall in production-to-stock, production-to-order and service environments.

Capacity is a quite general term that in this thesis we limit to covering only the capacity of human production or service resources. Hence, we do not study storing capacity (warehouses, depots, reservoirs, buildings), nor capital equipment capacity, which need strategic considerations (power plants, networks' bandwidth, terminals, capital investments). Even for the limited definition of the word, capacity, we employ, categorization can be established in multiple ways. According to the *permanent-contingent capacity concept*, capacity is categorized into permanent and contingent capacity based on the time needed to commit a change in the size. Namely, changing the size of the permanent capacity has a longer lead-time than that of the contingent capacity. *Permanent capacity* we can relate to the regular in-house capacity, which is efficient and cheap, but is difficult to change in a short time. *Contingent capacity* is, as opposed to permanent capacity, used to follow quick changes in the demand, but these cost more. Examples for contingent capacity include contingent worker capacity, overtime, or external capacity reservations. Once we submit an order for an amount of contingent capacity, it may become available after a given lead-time, which we refer to as the *capacity acquisition lead-time* later.

If voluntary contingent capacity is used, then there is uncertainty about the amount of contingent capacity available; the contingent capacity order may be only partially fulfilled. This is called the voluntary compliance regime. In contrast, if contingent capacity is used on an obligatory basis, then a contract enforces the contingent capacity to become fully available in a given fixed lead-time, when it is asked for. This latter regime is called the forced compliance regime. This thesis discusses *tactical capacity planning* under the permanent-contingent capacity concept, addressing decisions of setting the size of the permanent capacity and/or selection of a contingent capacity control policy, such that we take into account the capability of facilitating contracted contingent capacity under forced compliance at the operational level.

One emerging form of contracted contingent capacity used is the workforce provided by external labor supply agencies (ELSAs), which served as a main motivation for this thesis. ELSAs establish agreements with companies guaranteeing contingent capacity availability in the form of temporary workforce. The agreement enforces the ELSA to deliver the workforce within a given fixed lead-time, which the ELSA uses for contacting the workforce pool they have registered, or for the searching process, in general. The main clients of ELSAs are companies having fluctuating demand, as fluctuating demand often asks

for frequent capacity changes. For such companies the main benefit in contracting an ELSA originates from the difference between the costs of changing the permanent and the contingent capacity: Since the temporary workers provided are employed by the ELSAs, decreasing the temporary workforce levels is different than decreasing permanent workers for the companies and does not bring any additional costs. On the other hand, ELSAs benefit from the higher costs of the contingent capacity, the pooling effect, as they are in touch with multiple companies, and from provision of related services. For statistical evidence on significant usage of ELSA workforce in different countries, we refer the reader to Tan and Alp (2005).

1.2. Characteristics of to-stock/to-order manufacturing systems, and services

This thesis is to contribute to the framework of operations management using mathematical knowledge. The goal of operations management research is to gain understanding the product creation and delivery processes and improve their performance. Operations research is a set of mathematical tools and techniques that support among others operations management to achieve this goal. We use operations research techniques throughout the thesis.

Delivering products often means delivering goods coupled with services. Hence, products are combinations of tangible physical items (goods) and intangible ones (services). The operations management literature often distinguishes the products, where the emphasis is put on the tangible part, from those where it is on the intangible part, calling the former manufacturing, and the latter services (see e.g. Waters (1996)). It is important to point out that this distinguishment does not mean that manufacturing and services are opposites or that they are separable. This distinguishment defines only an intended approach declaring if our focus should be on the tangible or on the intangible part of the product. In this thesis, we study both manufacturing and service systems.

Other obvious differences between goods and services are that goods can be kept in stock, while services cannot. Therefore, manufacturing systems are given the flexibility of employing inventories, while services are not. Those manufacturing systems that use finished-goods inventories are called production-to-stock systems (or in short, to-stock systems); those manufacturing systems that do not use finished-goods inventories are called production-to-order (or to-order) manufacturing systems. The reason for not using finished-goods inventories is often the lack of exact information on what to produce.

In this thesis, we study both to-order and to-stock manufacturing systems.

We can enhance our overview of the three types of systems discussed in this thesis, if we enumerate these systems in the order of their interaction level with the customers. To-stock systems' production is highly decoupled from the actual customer needs, as both the production lead-time and the inventory help to hedge against demand fluctuations and uncertainty. To-order systems are only moderately decoupled, as the production lead-time (or to that proportional, the work-in-process inventory) is their only mean for managing demand variations. Service systems have the most interaction with the customers and depend the most on the actual customer demand, because no decoupling is possible (or it is so limited that it can be considered being negligible). We can conclude that the stronger the dependence on the customers, the less freedom there is in hedging against demand uncertainty.

The build-up of this thesis reflects the discussed threefold structure of operations management: first we study to-stock, then to-order manufacturing systems, and we end with services. The research questions as well as the chapters are sequenced to present the three operations management environments in this order, in an increasing order of customer interaction.

1.3. Research goals

The management practice of using flexible resources to effectively meet fluctuating or uncertain demand has a long tradition and received sufficient scientific attention. However, the area of flexible capacity management, the (human) resource management practice of using contingent capacity is a recently established new tradition, and it is definitely underresearched in the operations management literature. The goal of our research is to provide an overview of the scientific literature related to permanent-contingent capacity management, and to make a contribution to the existing literature in all the three environments of operations management discussed in section 1.2.

We targeted conducting research on permanent-contingent capacity management knowing the scientific advancements and findings in the to-stock (Bradley and Arntzen (1999), Angelus and Porteus (2002)), to-order, and service environments (Pinker (1996), Tan and Alp (2005)). The following section summarizes the mentioned scientific papers and outlines the main research streams related to contingent capacity management.

1.4. Related literature

To position our research, we discuss the related operations management literature on capacity planning, and on planning with contingent capacity. We are not aware of books or book chapters that give an overview on tactical capacity planning under the permanent-contingent capacity concept. The two most related survey papers are fairly recent: van Mieghem (2003) is on capacity management and investment, Wu *et al.* (2005) is on capacity management in high-tech industries. However, even these surveys together cover only partially the papers, which served as a starting point for this thesis. By necessity, we summarize the key articles of our primary concern. These closely related articles form three main research streams. The first research stream contains models with capacity adjustment costs, the second is concerned with contingent capacity planning in services, and the last one is on aggregate planning.

The first research stream developed an answer for optimal capacity control with convex capacity adjustment costs (starting from Rocklin *et al.* (1984), Bentolila and Bertola (1990), and Dixit (1997)). The optimal solution of the model with non-stationary stochastic demand and finite or infinite planning horizon is called the ISD policy (Invest/Stay put/Disinvest), found in Eberly and van Mieghem (1997). For stationary finite horizon model with IID demands, a closed form expression characterizes the optimal investment strategy. The ISD policy is a target interval policy: when the actual capacity exceeds the lower bound, we invest, when it is within the boundaries, we take no action (stay put), when exceeding the upper bound, we disinvest. This ISD policy was shown to be optimal not just for capacity planning, but also in the more general framework of investment strategies. Although the models of this research stream give this general result on the optimality of the ISD policy, a major drawback of these models is that they do not represent production decisions, contingent capacity or the demand process, explicitly.

The second research stream comprises research that has been conducted on optimal capacity decision making under the *permanent-contingent concept* in services. The thesis of Pinker (1996) studies different types of capacity flexibilities (numerical, working time, and functional flexibilities) in a special type of service environment, where workload can be backordered (as e.g. a mail sorting facility). The contingent capacity has a *fixed acquisition lead time*, which is zero or one period, and what has been ordered becomes *always available* (or, alternatively, there is an infinite contingent capacity supply). The models presented in Pinker (1996) are discrete time and of finite horizon; the demand is represented as a sequence of independent random variables; both the contingent and the permanent capacity costs are linear. Some of the mod-

els incorporate absenteeism of permanent workers. For a synthesis of the work on the numerical and working time flexibilities, see also Pinker and Larson (2003).

The third research stream is concerned with production environments, where the main question is the coordination of production and capacity planning decisions. The models addressing such coordinated planning were given the name aggregate planning (see Holt *et al.* (1955)). In two case studies, Bradley and Arntzen (1999) demonstrate that companies can benefit tremendously by optimizing their capacity and production decisions simultaneously rather than making first the capacity decision and then the production decisions. Angelus and Porteus (2002) generalize some of the results of Eberly and van Mieghem (1997) to aggregate planning. For the case of no backorder, the ISD policy's target intervals are characterized for the case when demand stochastically increases at the beginning of the life cycle and decreases thereafter. Additionally, when backordering is allowed, it is shown that capacity and inventory are economic substitutes: The target intervals decrease in the initial stock level and the optimal unconstrained base stock level decreases in the capacity level.

Aggregate planning under the *permanent-contingent concept* with setup costs for production and contingent capacity ordering was studied in Tan and Alp (2005). Apart from the additional production decision and holding costs, their model follows the set of assumptions introduced in Pinker (1996). In particular, the contingent capacity acquisition lead time is fixed to zero, the ordered amount is always available without limit, and the contingent and permanent capacity costs are linear. They also adopt the discrete time, finite horizon dynamic programming approach with demand represented as a sequence of independent random variables. As for the production, they assume zero lead time. Having zero lead time both for production and for contingent capacity ordering leads to a simple order of events, repeated in each period. This order of events coincides with that in Angelus and Porteus (2002) except for the minor difference that there setting the total capacity level stands in place of the contingent capacity ordering and receipt. For this model, Tan and Alp (2005) conclude that the target interval policy is not optimal because of the irregular structure of the cost-to-go function, and they suggest that the setup costs account for the irregularities observed.

A part of the dual sourcing literature (see the review of Minner (2003)) also addresses the combined use of permanent and contingent resources. In particular, Rosenshine and Obee (1976) and Janssen and de Kok (1999) study the combination of standing and emergency shipments incoming from a fixed and a flexible supplier, respectively.

The subcontracting, outsourcing, and expediting literature contains further relevant papers to our topic. We discuss the most closely related work, Yang *et al.* (2005), in the introduction of Chapter 2. Later we elaborate on the similarities and dissimilarities between the permanent-contingent capacity management, the subcontracting, and the dual sourcing models in section 5.3 of Chapter 5.

In the next section, we pose research questions to establish a line of research for this thesis. Following Pinker (1996) and Tan and Alp (2005), our research addresses tactical capacity decisions under the *permanent-contingent concept*, where both the contingent and the permanent capacity costs are linear, and the contingent capacity is *always available* and has a *fixed* (non-negative) *acquisition lead time*.

1.5. Research questions and methodology

Our general interest is to give an overview on tactical capacity planning with contingent capacity, along various industries, and to take the next logical research steps. When taking these steps, we build on the literature we discussed in the previous section. We study the use of permanent and contingent capacity covering industries in production-to-stock, production-to-order, and service environments. In what follows, we pose research questions for these environments and explain their relevance.

1. Production-to-stock aggregate planning

a. What form do optimal policies have in aggregate planning with backordering for non-stationary demand under the permanent-contingent concept, if the contingent capacity acquisition lead time is zero?

We are inspired by Angelus and Porteus (2002), who study the single capacity source case with linear capacity adjustment costs. For the backordering case, their main result is that if we assume stochastic non-stationary demand with independent periods and immediate capacity acquisition, then a target interval policy is optimal, as well as capacity and inventory are economic substitutes.

Tan and Alp (2005) developed the permanent-contingent correspondent to the single capacity source, backordering model in Angelus and Porteus (2002). While the model in Angelus and Porteus (2002) is at the strategic level, the model in Tan and Alp (2005) is at the tactical level, incorporating more details on capacity, using the permanent-contingent concept. In the model in Angelus and Porteus (2002) there is only one type of capacity, which can be reduced

or extended at some linear costs, whereas the model in Tan and Alp (2005) incorporates a given level of permanent capacity, to be paid in all the periods, and an additional amount of contingent capacity, which is to be ordered, and paid proportionally to the ordered amount. Additionally, setup costs are incurred for ordering and production. These modeling differences, of course, can substantially affect the form of the optimal aggregate planning policies in question.

We intend to study the model in Tan and Alp (2005), without setup costs. The reason is that Tan and Alp (2005) show that the structural results in Angelus and Porteus (2002) do not hold under the permanent-contingent formulation if setups are present. What can be done is to step back a little in complexity by neglecting setups, and demonstrate analytical results for this case. Therefore, we aim at characterizing the optimal policy under the permanent-contingent concept with neither production nor capacity ordering setups.

b. Can we generalize the results answering (a) to fixed, positive contingent capacity acquisition lead times?

Our motivation originates from a particular sentence in Angelus and Porteus (2002). In their discussion, they address a call for extending their aggregate planning model: "This important generalization to the case of a positive capacity lead time with inventory carry-over merits further research". We address an extension to question (a) by considering fixed (deterministic), positive capacity acquisition lead times, and as such an extension to the model of Tan and Alp (2005) without setups.

The practical motivation for studying research questions (1a) and (1b) are issues by manufacturing of neon-light encasement boxes, which actively uses contingent workforce to respond for demand fluctuations. The acquisition of contingent workers takes two days. In this case, the appropriate level of permanent capacity is in question, and how the capacity and production decisions need to be coordinated.

2. Production-to-order

How can we use fast-response contingent capacity (approximately) optimally in production-to-order systems under setup costs and a fixed quoted customer lead time?

Fast-response capacity are those having insignificant acquisition lead-time. The ISD policy (see section 1.4, Eberly and van Mieghem (1997)) was developed for such capacity, but not for production-to-order systems. Although the ISD policy could be interpreted in a to-order environment, it would disregard both the work-in-process information and the due-date performance, which

are important factors.

While seeking for (approximately) optimal policies, we would like to specifically address the characteristics of a production-to-order setting. Therefore, optimality needs to be defined such that it includes due date performance for a quoted lead time. Furthermore, policies may utilize work-in-process information. The capacity acquisition lead-time we neglect in this question, because Eberly and van Mieghem (1997) assumes no lead-time, also, and we can find practical applications already without lead times.

In contrast to the previous question, we use setup costs now in order to conform to the capacity control with capacity adjustment costs research line. This way, we can establish the connection between the literature on capacity planning with adjustment costs and that of the contingent capacity planning. However, we need to overcome the complicating fact that the decisions in Eberly and van Mieghem (1997) are at the highly aggregated strategic level, whilst deciding on the contingent capacity ordering policy is at the tactical level. Naturally, at the tactical level more details about the system need to be studied.

One of our motivating examples from real life is in the pharmaceutical production, where expensive machines are used. The expensive machine may be used seven days a week, 24 hours a day. However, demand is sometimes less than what full production could satisfy. Therefore, capacity may be varied depending on the actual workload. The question is at which workload levels should the number of shifts changed. Later we give some further examples.

3. Services

What dynamics of contingent capacity usage does a budget constraint entail in service environments, where backordering is not possible?

For the manufacturing environments we have not studied the effect of the strategic level decisions on the tactical capacity decisions. Since the amount of contingent capacity planning literature on services exceeds far that on manufacturing, we can study a bit more complex service systems building on the accumulated knowledge. We selected the service environments for investigating the effect of constraining the tactical level by a decision made at a higher level.

Within services, an obvious example for a strategic constraint on the tactical capacity decisions is an annual budget. Hansen and van der Stede (2004) term budgeting an important control in almost all organizations. We study service systems, where the tactical capacity decisions are constrained by a given budget for a finite horizon. This budget is allocated to cover permanent

and contingent capacity cost over a finite horizon. In the beginning of the horizon the permanent capacity cost are incurred, and the rest of the budget is dynamically allocated to the contingent capacity.

For services, budget allocation along time has been already studied in the operation research literature (see Trivedi (1981)), but we are not aware of a production environment counterpart. A possible reason why budgeting was studied less in production environments than in services is that the additional possibilities the production systems have (inventory, work-in-process inventory, as discussed in Section 1.2) are generally budgeted together with the capacity, and the pooled budget creates a rather loose budget-constraint. Therefore, budgeting issues may be less relevant in production systems.

We focus on budgeted service environments, where backordering is not possible. If backordering is not possible, we talk about lost opportunities or quality deterioration, which can be, characterized by a loss (or capacity shortage penalty) function. Our reason for studying loss function formulations instead of backordering is that we can contribute the existing literature more in this way, since the backordering case has been already studied extensively in Pinker (1996). Moreover, Warner and Prawda (1972) have developed the concept of capacity planning related quality loss, giving a good starting point for the loss function case. Warner and Prawda (1972) couple the use of a loss function with the use of a capacity budget, inspiring the setting of our research question: how to minimize quality loss given a certain capacity budget. The emphasis in their research question is on the capacity usage dynamics. Although Trivedi (1981) studies capacity decisions under a budget constraint, the mixed-integer goal programming approach employed does not enable finding results about the dynamics. We are aware of neither analytical nor numerical studies that investigate longitudinal budget allocation dynamics.

A few army-hospitals serve as our motivation in this question. All these hospitals operate from a given annual budget, face uncertain demand, and have no significant budget uncertainty. The largest portion of the budget is dedicated to cover the permanent and contingent capacity costs. Once the permanent staffing level is given, we can be interested in how the remaining part of the capacity budget should be allocated along the year for using temporary staffing.

Research methodology

We use the axiomatic quantitative model based research methodology to approach all the research questions posed. This research methodology relies on the concept that real-life processes can be appropriately represented by quantitative models, meaning that the applicability of the observations resulting from the analysis of these models is likely. The quantitative models we study

are axiomatic, as they rely on idealized models of reality, and normative, since our model development and analysis are concerned with finding optimal decisions or policies. More details on the quantitative model based research methodology can be found in Bertrand and Fransoo (2002).

1.6. Outline of the thesis

The remaining part of the thesis consists of three main parts in line with the three groups of research questions. These three parts correspond to the production to stock (Chapter 2), production to order (Chapter 3) and the service environments (Chapter 4), respectively.

Chapter 2 addresses research questions 1a and 1b on to-stock systems. First, we discuss the related literature more extensively than when introducing the research questions. Our starting point is the model in Tan and Alp (2005) having setups omitted. We develop a natural extension of this model by including a fixed capacity acquisition lead-time. Since we allow this capacity acquisition lead-time to be zero as well, our model allows us to deal with both questions under a unified approach. We study the analytical properties of the model with respect to the optimal policy structure and economic substitution properties. Further insights are gained on the effect of the acquisition lead-time and the characteristics of the optimal decisions with the help of numerical experiments. We verify the validity of our conclusions with sensitivity analysis on demand forecast error. The content of this chapter has been presented in Mincsovics *et al.* (2008).

Chapter 3 is dedicated to deliver an answer for research question 2 on to-order systems. For this research question, we have not found real starting point in the literature on how to establish a tactical capacity planning model under the permanent-contingent concept. Therefore, we start with an extensive literature study in order to find the roots both in the theory and in real-life. We embrace the idea of workload-dependent capacity changes, already existing in the literature, and introduce stationary Markovian models to represent each of the workload-dependent policies. Standard evaluation of these models permits counting capacity, capacity level switching, and lost sales costs. Furthermore, we present two ways for due-date performance evaluation. The computational experiments performed let us show the situations for low and high benefits of using workload-dependent contingent capacity management policies as compared to using permanent capacity exclusively. We perform sensitivity analysis on the demand distribution to support our conclusions' validity. This chapter is based on Mincsovics and Dellaert (2008).

Chapter 4 comprises an analytical and a numerical part, both investigating research question 3 on budgeting in services. In both parts, we consider models with periods having independent, identically distributed demand with the dynamics, where the demand up to the forthcoming period is known with certainty. First, we formulate a dynamic programming model with a general capacity shortage function. In the analytical part, linear capacity shortage and budget deficit penalty costs are assumed. Under these assumptions, an approximate implicit analytical expression is developed for the optimal permanent capacity level; the optimal contingent capacity ordering is straightforward. The quality of the approximation is numerically presented. In the numerical part, two of more realistic shortage penalty functions are considered: these are the halfway quadratic and the halfway relative quadratic penalty functions. Employing these functions we evaluate our dynamic programming model via backward induction for various experimental settings. Finally, the optimal budget spending patterns are studied and compared with observed patterns reported in the empirical budget allocation literature. This chapter has been presented in Dellaert *et al.* (2008).

Chapter 5 is dedicated to drawing conclusions as well as to giving suggestions for future research. Because of the different assumptions and incomparable parameter space of the models in the different environments (to-stock, to-order, services), we found that discussing the similarities and dissimilarities in tactical capacity planning with contingent capacity would be too ambitious. Rather, we conclude the main insights for each environment separately.

Chapter 2

Integrated Capacity and Inventory Management with Contingent Capacity Acquisition Lead Times

We model a make-to-stock production system that utilizes permanent and contingent capacity to meet non-stationary stochastic demand, where a constant lead time is associated with the acquisition of contingent capacity. We determine the structure of the optimal solution concerning both the operational decisions of integrated inventory and flexible capacity management, and the tactical decision of determining the optimal permanent capacity level. Furthermore, we show that the inventory (either before or after production), the pipeline contingent capacity, the contingent capacity to be ordered, and the permanent capacity are economic substitutes. We also show that the stochastic demand variable and the optimal contingent capacity acquisition decisions are economic complements. Finally, we perform numerical experiments to evaluate the value of utilizing contingent capacity and to study the effects of capacity acquisition lead time, providing useful managerial insights. We verify the validity of our conclusions with sensitivity analysis on the demand forecast error.

2.1. Introduction and related literature

In a make-to-stock production system that faces volatile demand, system costs may be decreased by managing the capacity as well as the inventory in a joint fashion, in case there is some flexibility in the production capacity. In some production environments, it is possible to increase the production capacity temporarily while it may take some time to do so. We refer to this delay as capacity acquisition lead time. In this chapter we consider such a make-to-stock production system subject to periodic review in a finite-horizon under non-stationary stochastic demand, where our focus is on the effects of capacity acquisition lead time.

Throughout this chapter, we primarily consider the workforce capacity setting for ease of exposition. We use the temporary (contingent) labor jargon to refer to capacity flexibility. In that setting, we generally assume that the production quantity is proportional to the workforce size, permanent and contingent. Since the workforce productivity often diminishes, especially by the presence of machines in the production, we also address the case that the production quantity is a concave function of the workforce size.

Flexible capacity management refers to adjusting the total production capacity with the option of utilizing contingent resources in addition to the permanent ones. Since long-term changes in the state of the world can make permanent capacity changes unavoidable, we consider the determination of the permanent capacity level as a tactical decision that needs to be made only at the beginning of the planning horizon. This permanent capacity level will not be changed to the end of the horizon. The integrated inventory and flexible capacity management problem that we deal with in this chapter refers to determining the contingent capacity to be ordered which will be available in a future period as well as determining the optimal production quantity in a certain period given the available capacity which has been determined in an earlier period. We note that this problem is essentially a stochastic version of the aggregate production planning problem.

The dynamic capacity investment/disinvestment problem has been investigated extensively in literature. This problem aims at optimizing the total production capacity of firms at a strategic level to meet long-term demand fluctuations. Rocklin *et al.* (1984) show that a target interval policy is optimal for this problem. This policy suggests investing in (expanding) the capacity if its current level is below a critical value, disinvesting in (contracting) the capacity if its current level is above another critical value, and doing nothing otherwise. Eberly and van Mieghem (1997) later extend this result to environments with multiple resources. Further multidimensional optimality

results are shown by Gans and Zhou (2002) and Ahn *et al.* (2005). Angelus and Porteus (2002) show that target interval policy is still optimal for managing the capacity in the joint capacity and inventory management problem of a short-life-cycle product under certain assumptions. In general, the lead time for the realization of the capacity expansion and contraction decisions is neglected in this literature and Angelus and Porteus (2002) state that 'this important generalization to the case of positive capacity lead time with inventory carry-over merits further research'. The lead time issue is considered in the capacity expansion literature to a certain extent. Angelus and Porteus (2003) show optimality of the echelon capacity target policy for multiple resources, which can have different investment lead times and for which investments can be deferred. Ryan (2003) presents a summary of the literature on dynamic capacity expansions with lead times. There are two main differences between the dynamic capacity investment/disinvestment problem and the integrated inventory and flexible capacity management problem that we consider: (i) investment results in possession of capital goods, which still has some value at the time of divestment, whereas flexible capacity is not possessed, but acquired only for a temporary duration, (ii) investment decisions are strategic, while integrated inventory and flexible capacity management is tactical and operational.

The problem that is addressed in this chapter is closely related to the problems considered by Tan and Alp (2005), Alp and Tan (2008), and Yang *et al.* (2005). Tan and Alp (2005) deal with a similar problem environment where the lead time for capacity acquisitions is neglected and only the operational decisions are considered. Alp and Tan (2008) extend this analysis by including the tactical level decision of determination of the permanent capacity level. Both of these studies consider fixed costs that are associated with initiating production as well as acquiring contingent workers. We ignore such fixed costs in this chapter and focus on the effects of capacity acquisition lead time. When the fixed costs of the model in Alp and Tan (2008) and the capacity acquisition lead time of the model in this chapter are ignored, these two studies reduce to a common special case. We refer the reader to these two studies for analysis of this special case and also for a review of the literature on flexible capacity and inventory management for all aspects of the problem other than the capacity acquisition lead time.

Yang *et al.* (2005) deal with a production/inventory system under uncertain permanent capacity levels and the existence of subcontracting opportunities. Subcontracting takes a positive lead time, which is assumed to be one period longer than or equal to the production lead time and a fixed cost is associated with subcontracting. The optimal policy on subcontracting is shown

to be of capacity-dependent (s,S)-type. The authors also show that there is a complementarity condition between slack capacity and subcontracting: If subcontracting is more costly than production, no subcontracting will take place unless production capacity is fully utilized. There is a major operational difference between this form of subcontracting option and the use of contingent capacity as in our setting. Subcontracting affects the inventory level directly (any amount subcontracted increases the inventory position with full quantity), while contingent capacity gives extra flexibility as it allows underutilization of capacity at the time of production.

The rest of the chapter is organized as follows. We present our dynamic programming model in Section 2. The optimal policy and some of its properties are discussed in Section 3 and our computations that result in managerial insights are presented in Section 4. We summarize our conclusions and suggest some possible extensions in Section 5.

2.2. Model formulation

In this section, we present a finite-horizon dynamic programming model to formulate the problem under consideration. Unmet demand is assumed to be fully backordered. The relevant costs in our environment are inventory holding and backorder costs, and the unit cost of permanent and contingent capacity, all of which are non-negative. There is an infinite supply of contingent capacity, and any number of contingent workers ordered become available with a given time lag. The notation is introduced as need arises, but we summarize our major notation in Table 2.1 for ease of reference.

We consider a production cost component which is a linear function of permanent capacity in order to represent the costs that do not depend on the production quantity (even when there is no production), such as the salaries of permanent workers. That is, each unit of permanent capacity costs c_p per period, and the total cost of permanent capacity per period is Uc_p, for a permanent capacity of size U, independent of the production quantity. We do not consider material-related costs in our analysis, but it can easily be extended to accommodate this component. In order to synchronize the production quantity with the number of workers, we redefine the "unit production" as the number of actual units that an average permanent worker can produce; that is, the production capacity due to U permanent workers is U "unit"s per period. We also define unit production cost by contingent workers as c_c in the same unit basis. For ease of exposition we consider the productivity rates of contingent and permanent capacity to be the same, but our model can

T	:	Number of periods in the planning horizon
L	:	Lead time for contingent capacity acquisition
c_p	:	Unit cost of permanent capacity per period
c_c	:	Unit cost of contingent capacity per period
h	:	Inventory holding cost per unit per period
b	:	Penalty cost per unit of backorder per period
α	:	Discounting factor $(0 < \alpha \leq 1)$
D_t	:	Random variable denoting the demand in period t
$G_t(w)$	:	Distribution function of D_t
$g_t(w)$	:	Probability density function of D_t
U	:	Size of the permanent capacity
x_t	:	Inventory position at the beginning of period t before ordering
y_t	:	Inventory position in period t after ordering
θ_t	:	Contingent capacity available in period t (that is ordered in period $t-L$)
θ^t	:	$\begin{cases} (\theta_t, \theta_{t+1}, \ldots, \theta_{t+L-2}, \theta_{t+L-1}) & \text{if } 0 < t \leq T-L \\ (\theta_t, \theta_{t+1}, \ldots, \theta_{T-1}, \theta_T) & \text{if } T-L+1 \leq t \leq T \\ 0 & \text{if } t = T+1 \end{cases}$
$f_t(x_t, \theta^t, U)$	:	Minimum total expected cost of operating the system in periods $t, t+1, ..., T$, given the system state (x_t, θ^t, U)
$J_t(y_t, \theta^{t+1}, U)$	:	Cost-to-go function of period t excluding the period's capacity related costs, given the system state (y_t, θ^{t+1}, U)
s_t	:	Slack capacity in period t, after production
$\cdot^*$	:	Optimal solution
$\hat{\cdot}$	:	Unconstrained optimum
$\bar{y}_t$	:	Optimal inventory position after ordering in period t subject to $\theta_{t+L} = 0$
$\bar{\theta}^A_{t+L}$	:	Optimal contingent capacity ordered in period t subject to $y_t = x_t$
$\bar{\theta}^B_{t+L}$	:	Optimal contingent capacity ordered in period t subject to $y_t = x_t + \theta_t + U$

Table 2.1: Summary of Notation

accommodate different productivity rates as explained in Tan and Alp (2005).

We consider a finite decision horizon with periods indexed from 1 to T. In every period, a decision is made to determine the number of contingent workers to be

available in exactly L periods after the current period, as long as there are at least L periods before the end of planning horizon. If θ_t contingent workers are ordered in period $t-L$ then that many workers become available in period t at a total cost of $c_c\theta_t$ which is charged when they become available. In any period $t \leq T-L$, we keep a vector $\theta^t = (\theta_t, \theta_{t+1}, ..., \theta_{t+L-1})$ which consists of the number of contingent workers that are ordered in periods $t-L, t-L+1, ..., t-1$. In the next period, the vector θ^{t+1} consists of the information on the hired contingent workers for periods $t+1, t+2, ..., t+L-1$, carried from the vector θ^t, as well as the decision made for period $t+L$, θ_{t+L}, in period t. Since no contingent workers are ordered after period $T-L$, $\theta^t = (\theta_t, \theta_{t+1}, ..., \theta_{T-1}, \theta_T)$ for $T-L+1 \leq t \leq T$ and $\theta^{T+1} := 0$. The size of the permanent workforce, U, is determined only at the beginning of the first period, and it is considered to be fixed during the whole planning horizon.

The order of events in a period is as follows. At the beginning of period t, the initial inventory level, x_t, is observed, and the number of previously ordered contingent workers, θ_t, become available. The total amount of capacity at period t becomes $U+\theta_t$, which is the upper limit on the production quantity of this period. Then, the operational decisions, i.e. the production decision given the available capacity and the decision on the number of contingent workers to be available in period $t+L$, are made. According to the production decision, the inventory level is raised to $y_t \leq x_t + U + \theta_t$. We note that the optimal production quantity $(y_t - x_t)$ may result in partial utilization of the available capacity, which is already paid for. At the end of period t, the realized demand, $w_t \geq 0$ is met/backordered, resulting in a starting inventory for period $t+1$, $x_{t+1} = y_t - w_t$. The vector θ^{t+1} is constructed as explained above. We assume the demand to be independently but not necessarily identically distributed, and we denote the random variable corresponding to the demand in period t as $D_t \geq 0$ and its distribution function as G_t. Finally, we denote the minimum cost of operating the system from the beginning of period t until the end of the planning horizon as $f_t(x_t, \theta^t, U)$.

We define the holding-backorder cost function, $\mathcal{L}_t(z)$, for a given inventory level after production, z. We use the notation

$$\mathcal{L}_t(z) = h \int_{-\infty}^{z} (z-\omega)\, dG_t(\omega) + b \int_{z}^{\infty} (\omega - z)\, dG_t(\omega)$$

for this cost function, which corresponds with a news-vendor formula. With the help of this holding-backorder cost function, the problem of integrated Capacity and Inventory Management with Capacity Acquisition Lead Times (CILT) can be formulated as a dynamic programming model.

$$
\begin{aligned}
&(CILT)\\
f_t(x_t,\theta^t,U) &= Uc_p+\theta_t c_c+\\
&\quad\begin{cases}
\min\limits_{y_t\in[x_t;x_t+\theta_t+U]}\left\{\mathcal{L}_t(y_t)+\alpha E[f_{t+1}\left(y_t-D_t,\theta^{t+1},U\right)]\right\} & \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } T-L+1\le t\le T & \\
\min\limits_{\theta_{t+L}\ge 0,y_t\in[x_t;x_t+\theta_t+U]}\left\{\mathcal{L}_t\left(y_t\right)+\alpha E\left[f_{t+1}\left(y_t-D_t,\theta^{t+1},U\right)\right]\right\} & \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } 1\le t\le T-L &
\end{cases}\\
f_0(x_1) &= \min_{U\ge 0,\ \theta^1\ge 0} f_1\left(x_1,\theta^1,U\right)
\end{aligned}
$$

where $f_{T+1}(\cdot)\equiv 0$, and $0\le L\le T$.

We note that the number of contingent workers hired before the planning horizon begins, θ^1, is also optimized in the above formulation, assuming that those decisions are made in advance in an optimal manner. Nevertheless, all of our analytical results would hold for any given θ^1 as well.

When capacity acquisition lead time is zero ($L=0$), the minimization operator,

$\min\limits_{\theta_{t+L}\ge 0,y_t\in[x_t;x_t+\theta_t+U]}$ is to be read as $\min\limits_{\theta_t\ge 0}\ \min\limits_{y_t\in[x_t;x_t+\theta_t+U]}$, the cost $\theta_t c_c$ gets inside the minimization, and θ^t disappears from the state space. This two-dimensional minimization can be reduced to a single-dimensional one.

2.3. Analysis of the optimal policies

In this section, we first characterize the optimal solution to the problem that is modeled in Section 2.2. Then we introduce some properties of the optimal solution, including those that regard the utilization of the available capacity.

Let J_t denote the cost-to-go function of period t excluding the period's capacity related costs.

$$J_t\left(y_t,\theta^{t+1},U\right)=\mathcal{L}_t\left(y_t\right)+\alpha E\left[f_{t+1}\left(y_t-D_t,\theta^{t+1},U\right)\right]$$

Accordingly, $f_t(x_t,\theta^t,U)$ can be rewritten as

$$
\begin{aligned}
f_t(x_t,\theta^t,U) &= Uc_p+\theta_t c_c+\\
&\quad\begin{cases}
\min\limits_{y_t\in[x_t;x_t+\theta_t+U]} J_t(y_t,\theta^{t+1},U) & \text{if } T-L+1\le t\le T\\
\min\limits_{\theta_{t+L}\ge 0,\ y_t\in[x_t;x_t+\theta_t+U]} J_t(y_t,\theta^{t+1},U) & \text{if } 1\le t\le T-L
\end{cases}.
\end{aligned}
$$

Let $\left(\hat{y}_t, \hat{\theta}_{t+L}\right)$ be the unconstrained minimizer of the function $J_t(\cdot)$ for given state variables $\theta_{t+1}, ..., \theta_{t+L-1}$, and U. We use the following definitions in our further discussion for $t \in \{1, ..., T-L\}$:

$$\bar{y}_t := \underset{y_t \in [x_t; x_t+\theta_t+U], \theta_{t+L}=0}{\arg\min} J_t\left(y_t, \theta^{t+1}, U\right)$$

is the optimal production when no capacity is ordered;

$$\bar{\theta}^A_{t+L} := \underset{\theta_{t+L} \geq 0}{\arg\min} J_t\left(x_t, \theta^{t+1}, U\right)$$

is the optimal contingent capacity order when no production takes place; and

$$\bar{\theta}^B_{t+L} := \underset{\theta_{t+L} \geq 0}{\arg\min} J_t\left(x_t + \theta_t + U, \theta^{t+1}, U\right)$$

is the optimal contingent capacity order when full production takes place.

Let $\left(y^*_t, \theta^*_{t+L}\right)$ be the joint optimal production and contingent capacity hiring decision in period t given that the state variables are x_t, θ^t and U.

2.3.1 Optimal policy characterization

The optimal decisions in any period t (inventory level after production, y_t, and number of contingent workers hired, θ_{t+L}) are made by minimizing the function J_t over the feasible region. First, we characterize the solution of CILT in Theorem 2.1.

Theorem 2.1 *The following hold for any capacity acquisition lead time* $L = 0, 1, 2, \ldots T-1$.

1. *For any period* t $(1 \leq t \leq T)$, f_t *and* J_t *are (jointly) convex funtions.*
2. *For any period* t *such that* $1 \leq t \leq T-L$, *the optimal production and contingent capacity ordering policy is given by*

$$(y^*_t, \theta^*_{t+L}) = \begin{cases} (\hat{y}_t, \hat{\theta}_{t+L}) & \text{if } \hat{y}_t \in [x_t; x_t+\theta_t+U], \hat{\theta}_{t+L} \geq 0 \\ (x_t, \bar{\theta}^A_{t+L}) & \text{if } \hat{y}_t < x_t, \hat{\theta}_{t+L} \geq 0 \\ \left(x_t+\theta_t+U, \bar{\theta}^B_{t+L}\right) & \text{if } \hat{y}_t > x_t+\theta_t+U, \hat{\theta}_{t+L} \geq 0 \\ (\bar{y}_t, 0) & \text{if } \hat{y}_t \in [x_t; x_t+\theta_t+U], \hat{\theta}_{t+L} < 0 \\ (\bar{y}_t, \bar{\theta}^A_{t+L}) \text{ with } (\bar{y}_t - x_t)\bar{\theta}^A_{t+L} = 0 & \\ & \text{if } \hat{y}_t < x_t, \hat{\theta}_{t+L} < 0 \\ (\bar{y}_t, \bar{\theta}^B_{t+L}) \text{ with } (\bar{y}_t - x_t - \theta_t - U)\bar{\theta}^B_{t+L} = 0 & \\ & \text{if } \hat{y}_t > x_t+\theta_t+U, \hat{\theta}_{t+L} < 0 \end{cases}$$

Proof: See Appendix.

The convexity of J_t as stated in part 1 implies that the production quantity should bring the inventory level to the base-stock level $\hat{y}_t\left(\theta^t, \theta_{t+L}, U\right)$ for a given θ_{t+L}, where $\hat{y}_t\left(\theta^t, \theta_{t+L}, U\right)$ is the minimizer of J_t for a given $\left(\theta^t, \theta_{t+L}, U\right)$, as long as the base-stock level is in the interval $[x_t, x_t + \theta_t + U]$. Otherwise, $y_t^* = x_t$ if the base-stock level is less than x_t, meaning that no production should take place, and $y_t^* = x_t + \theta_t + U$ if the base-stock level is greater than $x_t + \theta_t + U$, meaning that all of the available capacity (permanent and contingent) should be utilized. With respect to the contingent capacity ordering decision, $\theta_{t+L}^* = \hat{\theta}_{t+L}\left(y_t, \theta^t, U\right)$ for any given y_t, where $\hat{\theta}_{t+L}\left(y_t, \theta^t, U\right)$ is the minimizer of J_t for a given $\left(y_t, \theta^t, U\right)$, as long as $\hat{\theta}_{t+L}\left(y_t, \theta^t, U\right) \geq 0$. Otherwise, no contingent capacity should be ordered. We also note that for periods $T - L + 1$ to T, the optimal level of inventory after production is given by a state-dependent base-stock policy, due to convexity of J_t. Part 2 of Theorem 2.1 characterizes the optimal integrated production and contingent capacity ordering decisions in terms of the unconstrained minimizer and $\bar{y}_t, \bar{\theta}_{t+L}^A$, and $\bar{\theta}_{t+L}^B$, which are the minimizers on the borders of the feasible domain ($\theta_{t+L} \geq 0,\ y_t \in [x_t; x_t + \theta_t + U]$). The first case corresponds to the situation where the unconstrained minimizer falls in the feasible region, and hence the unconstrained minimizer is the optimal solution. In the latter five cases the unconstrained minimizer is outside the feasible region, where the optimal solution is then on the boundary of the feasible region, due to convexity of J_t. The last two cases further characterize the optimal solution by imposing a condition (that we refer to as "complementary slackness property" in Section 2.3.3) when neither $\hat{y}_t$ nor $\hat{\theta}_{t+L}$ is within its feasible interval. Finally, part 1 also states that the recursive minimum expected cost function of the dynamic programming formulation, $f_t(x_t, \theta^t, U)$ is convex. Therefore, finding the optimal permanent capacity level, U^* is a convex optimization problem.

Remark 1 *If $c_c < c_p$, then $U^* = 0$.*

Remark 1 holds due to the fact that any solution, U' and θ_t' with $U' > 0$ would be dominated by the solution $U = 0$ and $\theta_t = U' + \theta_t'$ for all t.

In what follows we utilize the notion of supermodularity and submodularity to show properties on the pairwise relations of the variables and parameters in our model, which is a notion employed in economic theory often to explore economic complements and substitutes. A function which is supermodular (submodular) on two arguments implies that more of one of the arguments induces less (more) of the other (see Porteus (2002)). In particular, Theorem 2.2 identifies such relations in our problem environment between contingent capacity ordered, inventory position, permanent capacity, and demand.

Theorem 2.2 *For any period t ($1 \leq t \leq T$), and capacity acquisition lead time $L = 1, 2, \ldots, T-1$, the following hold.*

1. *$f_t\left(x_t, \theta^t, U\right)$ and $J_t\left(y_t, \theta^{t+1}, U\right)$ are supermodular functions.*
2. *J_t is submodular in $\left(D_t, \left(\theta^{t+1}, U\right)\right)$ and in (D_t, y_t), where $D_t \in D$, and D is the poset of discrete random variables with the first order stochastic dominance as partial order, $\left(\theta^{t+1}, U\right) \in \mathbb{R}^{L+1}$, on which the product order is the partial order.*

Proof: See Appendix.

Supermodularity of $J_t\left(y_t, \theta^{t+1}, U\right)$ implies, for example, that y_t and θ^{t+1} are economic substitutes: in any element we increase in θ^{t+1}, the optimal y_t is non-increasing. Naturally, it implies as well that substitution holds between y_t and U, any element of θ^{t+1} and U, or any two elements of θ^{t+1}. Supermodularity of $f_t\left(x_t, \theta^t, U\right)$ allows similar interpretation as that of $J_t\left(y_t, \theta^{t+1}, U\right)$: the inventory, the pipeline contingent capacity and the permanent capacity are economic substitutes. For example, a higher starting inventory eliminates the necessity for a higher permanent capacity.

Submodularity of the function J_t in (D_t, θ_{t+L}) given in part 2 indicates that D_t and θ_{t+L} are economic complements. That is to say, stochastically larger demand distributions lead to hiring more contingent capacity. A similar relation also exists between D_t and y_t. Note that the sub- and supermodularity results in Theorem 2.2 do not apply only to optimal decisions. For example, supermodularity of J_t in y_t and θ_{t+L} implies that the marginal cost of increasing y_t increases in θ_{t+L}. The reader is referred to Porteus (2002) and Topkis (1998) for further details on sub- and supermodular functions, and Puterman (1994) for partial ordering of random variables.

The following corollary to the second part of Theorem 2.2-*1.* helps to reduce the search space by providing bounds on the decision variables y_t and θ_{t+L}, using the fact that they are economic substitutes.

Corollary 1 *For any period t ($1 \leq t \leq T-L$), the (constrained) optimal solution of J_t is in the domain* $\left\{(y_t, \theta_{t+L}) \;:\; y_t \in [x_t; \bar{y}_t]\,,\ \theta_{t+L} \in \left[\bar{\theta}^B_{t+L}; \bar{\theta}^A_{t+L}\right]\right\}$.

2.3.2 Note on the concave productivity function

In most of the real-life cases the production quantity is not a linear function of the workforce size. Increasing the workforce size has diminishing returns on the production output. Passing some high workforce level the overall productivity

can even decrease. Our goal in this subsection is to study if the analytical results, formulated in Theorem 2.1 hold when the production quantity is a concave function of the workforce size.

For a given total workforce size (including both permanent and contingent workforce), u, we denote the maximal production quantity as $Q(u)$. We can reformulate the (CILT) model by replacing the minimization operators

$$\min_{y_t \in [x_t; x_t+\theta_t+U]} \quad \text{and} \quad \min_{\theta_{t+L} \geq 0, y_t \in [x_t; x_t+\theta_t+U]} \quad \text{with}$$
$$\min_{y_t \in [x_t; x_t+Q(\theta_t+U)]} \quad \text{and} \quad \min_{\theta_{t+L} \geq 0, y_t \in [x_t; x_t+Q(\theta_t+U)]} \quad \text{, respectively.}$$

We can characterize the modified (CILT) similar to the original (CILT).

Theorem 2.3 *For the modified (CILT) model, which includes concave productivity, for any capacity acquisition lead time $L = 0, 1, 2, \ldots T-1$ and for any period t ($1 \leq t \leq T$), f_t and J_t are (jointly) convex funtions.*

Proof: See Appendix.

The second part of Theorem 2.1 can be also inherited with minor modifications (all upper bound $x_t+\theta_t+U$ needs to be replaced with $x_t+Q(\theta_t+U)$, including the optimal production and contingent capacity ordering policy formula and the definition of $\bar{y}_t$ and $\bar{\theta}^B_{t+L}$).

In what follows, we continue studying the (CILT) model, where the production quantity is proportional to the workforce size.

2.3.3 Complementary slackness

In our model, we have two decision variables to be determined in every period: the inventory level after production and the contingent capacity ordered that will arrive L periods later. The production decision is bounded from above by the maximum amount of capacity available (the permanent capacity level plus the contingent capacity that was ordered L periods ago) whereas the contingent capacity ordering is only constrained to be non-negative. Let s_t denote the *slack capacity* in period t after the production decision is implemented, $s_t = x_t + U + \theta_t - y_t$. We define the complementary slackness property as follows:

Definition For any period t, there exists a Complementary Slackness Property (CSP) between slack capacity, s_t, and contingent capacity ordered, θ_{t+L}, only if $s_t\theta_{t+L} = 0$.

If a solution does not satisfy CSP, a positive contingent capacity is ordered for future use, while the current capacity which has already been paid for is not

fully utilized. If such a solution is optimal then ordering contingent capacity to be available L periods later is preferred to utilizing currently available capacity fully which might lead to carrying inventory. In case the optimal solution is known to satisfy CSP, this helps not only to further characterize the optimal solution, but also to simplify the solution of CILT. In particular, whenever the optimal solution satisfies CSP, the problem reduces to one-dimensional optimization problems. In what follows, we present some special cases where the optimal solution satisfies CSP.

For the special case where the demand is deterministic, it is straightforward to show that the optimal solution satisfies CSP if $\sum_{i=0}^{L-1} \alpha^i h < \alpha^L c_c$. This condition simply implies that it is less costly to carry inventory than to order contingent capacity, which assures that contingent capacity is never ordered unless available capacity is fully utilized.

Theorem 2.4 *When $L = 1$, the optimal solution satisfies CSP in the following cases:*

1. *In the infinite horizon problem with stationary and positive demand ($T \to \infty$, $D_t \equiv D > 0$), when $h < \alpha c_c$.*
2. *In the two-period problem, when $h\,(1 + \alpha) < \alpha c_c$.*

Proof: See Appendix.

Note that for the special case of $L = 1$, Theorem 2.4 is valid under reasonable cost parameter settings. The interpretation of the first part of the theorem is that if demand is the same (and non-zero) in each period being subject to the same uncertainty, then we can directly compare the cost coefficient of the two options. Namely, the optimality of complementary slackness follows if the unit holding cost is less than the discounted unit contingent capacity cost. The second part of the theorem gives a condition for a two-stage problem with changing demand. In contrast to the stationary case, the unit cost of the holding option is estimated with the unit holding cost of both periods. While an optimal solution which does not satisfy CSP might seem to be counter-intuitive, it turns out that in some cases this is true, as we illustrate in the following examples where CSP does not hold in the optimal solution.

Example 2.1 *For $T = 15$, $L = 2$, $h = 1$, $b = 5$, $c_p = 2.4$, $c_c = 3.2$, $\alpha = 1$, $U = 10$, $x_1 = 0$, $\theta_1 \geq 0$, and $\theta_2 \geq 0$, consider the following demand stream: $P(D_1 = 0) = 1$, $P(D_2 = 30) = 0.4$, $P(D_2 = 0) = 0.6$, $P(D_3 = \cdots = D_{11} = 0) = 1$ and $P(D_{12} = \cdots = D_{15} = 10) = 1$. The optimal decision is $(y_1^*, \theta_3^*) = (0, 10)$. That is, 10 units of contingent capacity is ordered while the available capacity is not fully utilized which violates CSP. The intuition behind*

this solution is as follows: Because the uncertainty will be resolved in period 2, any production before that may result in holding inventory for a number of periods. On the other hand, if the production capacity in period 3 is not increased -which requires requesting 2 periods in advance-, there may be high backordering costs in case positive demand in period 2 is materialized.

Example 2.2 *For* $T = 2$, $L = 1$, $h = 2.98$, $b = 5$, $c_p = 2.5$, $c_c = 3$, $\alpha = 0.99$, $U = 6$, $x_1 = 0$, *and* $\theta_1 = 0$, *let the demand follow normal distribution with* $E[D_1] = 3$, $Var[D_1] = 0.36$, *and* $E[D_2] = 21$, $Var[D_2] = 17.64$ *(both yielding coefficient of variation = 0.2). In this case,* $(y_1^*, \theta_2^*) = (4.8, 10.4)$, *violating CSP.*

2.4. Numerical results and discussion

The main goal of this section is to gain insights on how the value of flexible capacity and the optimal permanent capacity levels change as the following system parameters change: capacity acquisition lead time, unit cost of contingent capacity, backorder cost, and the variability of the demand. For this purpose, we conduct some numerical experiments by solving CILT. We use the following set of input parameters, unless otherwise noted: $T = 12$ (e.g. in days, weeks, or months), $b = 10$, $h = 1$, $c_c = 3$, $c_p = 2.5$ (all in units of €33, €250, or €1000, respectively), $\alpha = 0.99$, and $x_1 = 0$. We consider Normal demand with a coefficient of variation (CV) of 0.2 that follows a seasonal pattern with a cycle of 4 periods, where the expected demand is 10, 15, 10, and 5, respectively. Recall that the values of the pipeline of contingent capacity at the beginning of the first period are optimized in CILT, and accordingly the results containing different lead times are comparable.

Solution of CILT on a Pentium 4 with a 2.79GHz CPU and 1Gb RAM for the parameter set given above took less than 1 second for $L < 3$ and 14 seconds for $L = 3$. For longer lead times, the curse of dimensionality prevails and computational limitations become prohibitive.

In the results that we present, we use the term "increasing" ("decreasing") in the weak sense to mean "non-decreasing" ("non-increasing"). We provide intuitive explanations to all of our results below and our findings are supported in several numerical studies. However, like all experimental results, one should be careful in generalizing them, especially for extreme values of problem parameters.

2.4.1 Value of flexible capacity

The option of utilizing contingent capacity provides additional flexibility to the system and leads to reduction of the total costs, even though there is a certain lead time associated with it. We measure the magnitude of cost reduction in order to gain insight on the value of flexible capacity. We compare a flexible capacity (FC) system with an inflexible one (IC), where the contingent capacity can be utilized in the former but not in the latter. We define the absolute value of flexible capacity, VFC, as the difference between the optimal expected total cost of operating the IC system, ETC_{IC}, and that of the FC system, ETC_{FC}. That is, $VFC = ETC_{IC} - ETC_{FC}$. We also define the (relative) value of flexible capacity as the relative potential cost savings due to utilizing the flexible capacity. That is, $VFC\% = 100 \cdot VFC/ETC_{IC}$. We note that both VFC and $VFC\%$ are always non-negative because it is always an option not to use contingent capacity. We also note that the permanent capacity levels are optimized in both systems separately to ensure that the differences are not caused by the insufficiency of permanent capacity in the inflexible system.

We first test the value of flexibility with respect to the backorder and contingent capacity costs under different capacity acquisition lead times, by varying the value of one of the parameters while keeping the rest fixed. We present the results in Table 2.2, which support intuition in the sense that $VFC\%$ is higher when capacity acquisition lead time is shorter. These results also generalize the findings of Tan and Alp (2005) for $L = 0$ to the case of positive capacity acquisition lead times, such that $VFC\%$ is higher when contingent capacity cost is lower or backorders are more costly (equivalently, when a higher service level is targeted).

We note that, although $VFC\%$ decreases with an increasing lead time, the marginal decrease appears to be decreasing as L increases. Besides, we also observe that $VFC\%$ with higher lead times persists to be comparable with $VFC\%$ with lower lead times, meaning that flexibility is still valuable even when the capacity acquisition lead time is relatively long.

We also analyze the relation between the value of flexible capacity and the demand variability. The results presented by Alp and Tan (2008) indicate that the value of flexibility is not necessarily monotonic (i.e. it does not increase or decrease consistently) as the demand variability increases for the case where the lead time is zero. We find out that this continues to be true for the case where the lead time is strictly positive as well, because the system has the ability to adapt itself to changes in coefficient of variation, CV, by optimizing the permanent capacity level accordingly. Nevertheless, for increasing values

L	0	1	2	3
c_c		$VFC\%$		
1.0	63.35	58.94	57.63	57.36
2.0	36.35	31.50	28.34	27.18
2.5	22.87	17.90	14.57	12.71
3.0	14.91	10.30	8.55	7.50
3.5	11.10	7.26	6.27	5.61
4.0	8.92	5.58	4.91	4.21
5.0	6.02	3.18	2.98	2.74
8.0	1.75	0.42	0.37	0.34
b		$VFC\%$		
5	11.79	7.91	6.49	5.54
10	14.91	10.30	8.55	7.50
20	17.50	12.22	10.22	9.07
50	20.51	14.63	12.31	11.09
250	24.82	18.06	15.49	14.22

Table 2.2: $VFC\%$ as L, c_c, and b change (base values marked)

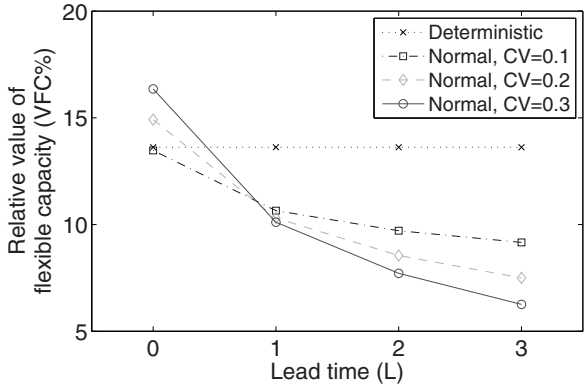

Figure 2.1: $VFC\%$ as a function of L for different demand streams

of the contingent capacity acquisition lead time we observe that the value of flexibility generally decreases when the demand variability increases as is the case in Figure 2.1. A longer capacity acquisition lead time deteriorates the effectiveness of capacity flexibility. This effect is amplified in case of higher

demand variability. In other words, since the capacity needs are more predictable for lower demand variability, use of contingent capacity -which has to be ordered one lead time ahead- becomes more effective as compared to the high variability case. This also explains why the decrease in the value of flexibility as lead time increases is steeper when the variability is higher.

2.4.2 Sensitivity analysis on $VFC\%$ for demand forecast error

Estimation of future demand is subject to error. Even if demand is predicted in stochastic terms, the forecasted distribution is often just an estimate in real-life situations. In order to calculate a more realistic value of the capacity flexibility, we represent some demand forecast error in our calculation.

We study the effect of the relative demand forecast error on the value of capacity flexibility via sensitivity analysis. In the numerical experiments we perform, the relative demand forecast error means misestimation of the demands' real expected value ($E[D_t]$) or coefficient of variation, for all t. Namely, the estimates are

$$\hat{E}[D_t] = E[D_t](1 + \Delta E[D_t]), \text{ and}$$

$$\hat{CV}[D_t] = CV[D_t](1 + \Delta CV[D_t]) \text{ for all } t = 1, 2, \ldots, T.$$

We perform sensitivity analysis on $VFC\%$ for demand forecast error. In our numerical experiments we calculate the $VFC\%$ values over the entire horizon while constantly over-/underestimated demand expectation or demand coefficient of variation. We preserve the same definition of $VFC\%$. The only change is that both the inflexible and the flexible expected total costs (ETC_{IC} and ETC_{FC}) are calculated under the misestimated demand stream.

Our numerical experiments use the base case (demand CV is 0.2, $c_c = 3$, and $b = 10$) extended by the values of demand coefficient of variation 0.1 and 0.3, $c_c = 2.6$, 3.5, and $b = 5$, 20, in all the possible $3^3 = 27$ combinations. Additionally, we vary ΔE to take the values -0.04, -0.02, 0, $+0.02$, and $+0.04$; and ΔCV to take the values -0.14, -0.07, 0, $+0.07$, and $+0.14$. The values ± 0.02 and ± 0.07 originate from simulating demand forecast error based on a history of 100 instances. Furthermore, the capacity acquisition lead time, L is 0, 1 or 2. The base case, which we found being a typical case in our sensitivity analysis, is presented in Table 2.3.

The $\Delta E = 0$ row as well as the $\Delta CV = 0$ row correspond to the estimated value of capacity flexibility already depicted in Figure 2.1. The $\Delta E > 0$ (< 0) rows or the $\Delta CV > 0$ (< 0) rows correspond to the real value of capacity flexibility for overestimated (underestimated) expected value or coefficient of

L	0	1	2
ΔE		$VFC\%$	
-0.04	14.57	10.00	7.18
-0.02	14.57	10.09	7.85
0.0	14.91	10.30	8.55
0.02	15.70	10.74	9.29
0.04	17.04	11.50	10.13
ΔCV		$VFC\%$	
-0.14	14.26	10.26	8.29
-0.07	14.57	10.26	8.43
0.0	14.91	10.30	8.55
0.07	15.31	10.38	8.67
0.14	15.76	10.52	8.80

Table 2.3: $VFC\%$ as L, ΔE, and ΔCV change, for the base case

variation. E.g. the $VFC\%$ value of 14.57% in the first row, first column (for $L = 0$) means that if the expected value was underestimated by 4% ($\Delta E = -0.04$), then our $VFC\%$ value estimate of 14.91% ($\Delta E = 0$) overshoots the (real) $VFC\%$ value, which is 14.57%.

Based on the 27 cases observed, we found that varied ΔE, or varied ΔCV results in similar $VFC\%$ behavior. We can typically observe convexly increasing shapes, particularly for short lead times, which straightens for longer lead times. However, we observed slightly concave shapes for $L = 2$ and $b = 20$: along changes in ΔE, $\Delta E = 0$ is peak value, and when ΔCV is varied, then ΔCV is decreasing. Except for the cases of long lead time and high backorder unit costs, overestimation of the expected demand induces underestimation of $VFC\%$, while underestimation of the expected demand gives less error in estimating $VFC\%$.

Not all the properties observed in Figure 2.1 hold under demand forecast error. Although we observed that the lower contingent capacity cost c_c and the shorter capacity acquisition lead times still consistently yield higher $VFC\%$, the value of flexibility as a function of the backorder cost coefficient becomes more irregular: especially for long lead times and large forecast errors, it becomes neither increasing nor decreasing (e.g. for $\Delta E = -0.1$, $\Delta CV = 0$, $L = 2$, and $c_c = 3.5$, for $b = 5$, 10, and 20, we have $VFC\% = 4.79$, 4.64, 7.16, respectively).

2.4.3 Optimal level of permanent capacity

In this section we investigate how the optimal level of permanent capacity changes as the problem parameters change. We present the data regarding some of our results in Table 2.4. We first note that the optimal permanent capacity decreases as the contingent capacity acquisition lead time decreases, in all of the cases that we consider. That is, since the decreased lead time makes the capacity flexibility a more powerful tool, it decreases the required level of permanent capacity. When c_c and L are small enough, the benefits of capacity flexibility becomes so prevalent that, even when $c_c > c_p$ the optimal permanent capacity level may turn out to be zero. We also note that the findings of Alp and Tan (2008) for the case of $L = 0$, which state that the optimal permanent capacity level decreases as contingent capacity cost decreases or backorder cost increases, also hold for positive capacity acquisition lead times.

L	0	1	2	3
c_c	U^* for $b = 10$			
2.5	0	0	0	0
2.51	0	0	2	3
2.6	3	3	4	6
3.0	7	7	8	9
3.5	8	9	10	10
4.0	9	10	10	10
5.0	10	11	11	11
8.0	11	12	12	12
demand	U^* for $b = 10$			
deterministic	7	7	7	7
normal, CV=0.1	7	8	8	8
normal, CV=0.2	7	7	8	9
normal, CV=0.3	6	7	9	9
demand	U^* for $b = 50$			
deterministic	7	7	7	7
normal, CV=0.1	7	7	7	8
normal, CV=0.2	6	7	8	9
normal, CV=0.3	5	6	8	9

Table 2.4: U^* as a function of the lead time, L, for varied c_c, b and demand distribution streams (base values are marked)

Similar to the value of flexibility, we observe that the optimal permanent capacity level is not necessarily monotonic in demand variability. Nevertheless, for longer capacity acquisition lead times or higher costs of contingent capacity, optimal permanent capacity level in general increases as demand variability increases. On the contrary, for shorter capacity acquisition lead times and lower costs of contingent capacity, optimal permanent capacity level in general decreases as demand variability increases.

2.5. Conclusions and future research

In this chapter the integrated problem of inventory and flexible capacity management under non-stationary stochastic demand is considered, when a fixed lead time is present for flexible capacity acquisition. Permanent productive resources may be increased temporarily by hiring contingent capacity in every period, where this capacity acquisition decision becomes effective with a given time lag. Other than the operational level decisions (related to the production and capacity acquisition levels), we also keep the permanent capacity level as a tactical decision variable which is to be determined at the beginning of a finite planning horizon. We provide insights into the effects of capacity acquisition lead time.

We first prove that all of the cost functions under consideration for decision making are convex. Moreover, we prove that the inventory (either before or after production), the pipeline contingent capacity, the contingent capacity to be ordered, and the permanent capacity are economic substitutes. We also show that the stochastic demand variable and the optimal contingent capacity acquisition decisions are economic complements; for stochastically larger demand streams, we observe higher contingent capacity levels in optimality. A similar interpretation is also true for stochastically larger demand streams and the optimal inventory levels obtained after production.

The convexity results help us to provide an optimal policy for the operational decisions and to find the optimal permanent capacity level. The optimal policy for managing contingent capacity ordering can be termed as a special target interval policy (see Eberly and van Mieghem (1997)), where the target interval reduces to a single point. Our convexity results also imply that, provided the capacity level, the optimal level of inventory after production is given by a state-dependent base-stock policy, where the dependency is on the capacity pipeline and the actual capacity. Furthermore, from the economic substitution results between inventory and capacity levels, it follows that the optimal target capacity level is a decreasing function of the inventory level, and that

the optimal inventory level is a decreasing function of any of the capacity dimensions including the permanent capacity.

A policy that might seem to be optimal is never to order contingent capacity unless the actual capacity is fully utilized, which we refer to as complimentary slackness property (CSP). We show through numerical examples that an optimal solution does not necessarily satisfy CSP. We also provide some cases where the optimal solution is assured to satisfy CSP.

By making use of our model, we develop some managerial insights. First of all, the value of flexibility naturally decreases with an increasing lead time. Consequently, there is a value in trying to decrease capacity acquisition lead time in the system through means such as negotiating with the external labor supply agency or forming a contingent labor pool perhaps within different organizations of the same company. This especially holds when the demand is highly variable. Nevertheless, the value of flexibility remains considerable even when the capacity acquisition lead time is relatively long. Therefore, the existence of a lead time in acquiring contingent capacity should not discourage the production company from making use of capacity flexibility, especially if the demand variability is not very high. Consequently, the managers should invest in higher levels of permanent capacity when capacity acquisition lead time and demand variability are high, and it is not wise to do so when the contingent capacity is a more "effective" tool in the sense that capacity acquisition lead time is short and the demand variability is high.

This research may be extended in several ways. Introducing an uncertainty on the permanent and contingent capacity levels would enrich the model. For example, the supply of contingent capacity may be certain for larger lead times whereas it may be subject to an uncertainty for shorter lead times. Some other extension possibilities include considering the fixed costs for production and/or acquisition of contingent capacity, including expansion and contraction decisions for the permanent capacity, considering multiple types of capacity flexibility (e.g. overtime, temporary workers, extra shifts), considering the possibility to carry over the contingent capacity and cancel previously ordered capacity, and developing efficient heuristic methods for the problem.

Appendix

Proof of Theorem 2.1:

We prove part 1 by induction. Note that $f_{T+1}(\cdot) = 0$ and is convex. Assume that $f_{t+1}(\cdot)$ is also convex. The function $J_t(y_t, \theta^{t+1}, U) = \mathcal{L}_t(y_t) +$

$\alpha E\left[f_{t+1}\left(y_t - D_t, \theta^{t+1}, U\right)\right]$ is convex because (i) $\mathcal{L}_t(y_t)$ is a convex function, (ii) $E\left[f_{t+1}\left(y_t - D_t, \theta^{t+1}, U\right)\right]$ is convex by the convexity preservation of the expected value operator (see Appendix A.5 in Bertsekas (1976)), and (iii) the convexity preservation of the linear combination with non-negative weights. Then note that the following minimization operators preserve the convexity of J.

$$\begin{aligned} g(x,\theta,U) &= \min_{y\in[x,x+\theta+U]} J(y,U), \\ h(x,\theta,U) &= \min_{\substack{y\in[x,x+\theta+U] \\ \delta\geq 0}} J(y,\delta,U) \end{aligned}$$

From Proposition B-4 of Heyman and Sobel (2004), coupled with the convexity preservation of affine mappings (see Hiriart-Urruty and Lemaréchal (1993)) it follows that the resulting g and h functions are convex when J is convex. Finally, $f_t(x_t,\theta^t,U) = Uc_p + \theta_t c_c + \begin{cases} g(x_t,\theta,U) & \text{for } T-L+1 \leq t \leq T \\ h(x_t,\theta,U) & \text{for } 1 \leq t \leq T-L \end{cases}$ is convex, which completes the proof of part 1. ■

Part 1 implies directly part 2. ■

Proof of Theorem 2.3:

Since the set $\{(x,U,\theta,y) : y \in [x, x+Q(\theta+u)]\}$ is convex for all concave $Q(.)$ function the following minimization operators also preserve the convexity of J.

$$\begin{aligned} g(x,\theta,U) &= \min_{y\in[x,x+Q(\theta+U)]} J(y,U), \\ h(x,\theta,U) &= \min_{\substack{y\in[x,x+Q(\theta+U)] \\ \delta\geq 0}} J(y,\delta,U) \end{aligned}$$

The rest of the proof is identical to the proof of Theorem 2.1. ■

Preliminaries to Proof of Theorem 2.2: We start with two lemmas that will help us with the proof of Theorem 2.2.

Lemma 2.1 *For any cost parameters, the newsboy function (holding-backorder cost function)*

$$\mathcal{L}(D,y) = h\int_{-\infty}^{y}(y-w)\,dF_D(w) + b\int_{y}^{\infty}(w-y)\,dF_D(w)$$

is submodular with $y \in \mathbb{R}$ (real) and $D \in \mathcal{D}$, where $\mathcal{D}$ is the poset of random variables with the first order stochastic dominance ($\preceq$) as partial order.

Proof : We need to show that $\mathcal{L}(D, y)$ is submodular; that is $\mathcal{L}(D^-, y^-) + \mathcal{L}(D^+, y^+) \leq \mathcal{L}(D^+, y^-) + \mathcal{L}(D^-, y^+)$ for all $D^-, D^+ \in \mathcal{D}$ and $y^-, y^+ \in \mathbb{R}$, for which $D^- \preceq D^+$ and $y^- \leq y^+$.We denote the cumulative distribution functions of D^- and D^+ by F^- and F^+, respectively. Then by the definition of stochastic dominance, if $D^- \preceq D^+$ then we have $F^-(w) \geq F^+(w)$ for all $w \in \mathbb{R}$. In the first step, we split integration intervals in $\mathcal{L}(D, y^+)$ and $\mathcal{L}(D, y^-)$ by y^- and y^+.

$$\begin{aligned}\mathcal{L}(D, y^+) &= h\int_{-\infty}^{y^-}(y^+ - w)\, dF_D(w) + h\int_{y^-}^{y^+}(y^+ - w)\, dF_D(w) \\ &\quad + b\int_{y^+}^{\infty}(w - y^+)\, dF_D(w)\end{aligned}$$

$$\begin{aligned}\mathcal{L}(D, y^-) &= h\int_{-\infty}^{y^-}(y^- - w)\, dF_D(w) + b\int_{y^-}^{y^+}(w - y^-)\, dF_D(w) \\ &\quad + b\int_{y^+}^{\infty}(w - y^-)\, dF_D(w)\end{aligned}$$

We denote the difference of the above standing two terms by $\Delta(D)$. One can show with the help of partial integration that $\Delta(D) := \mathcal{L}(D, y^+) - \mathcal{L}(D, y^-) = (h + b)\int_{y^-}^{y^+} F_D(w)\ dw$ holds.

In the final step, we subtract $\Delta(D^+)$ from $\Delta(D^-)$ and use the first order stochastic dominance of D^+ over D^-, meaning $F^-(w) \geq F^+(w)$ for all $w \in \mathbb{R}$.

$\Delta(D^+) - \Delta(D^-) = [\mathcal{L}(D^+, y^+) - \mathcal{L}(D^+, y^-)] - [\mathcal{L}(D^-, y^+) - \mathcal{L}(D^-, y^-)]$

$= (h + b)\int_{y^-}^{y^+}(F^+(w) - F^-(w))\ dw \leq 0$. This completes the proof. ■

Lemma 2.2 *Convex minimization operators resulting in supermodular functions*

Assume that y, x, θ, c are real numbers, z is a real vector, and g is a real valued function.

1. *If $g(y)$ is convex then $H(x, \theta) := \min_{y \in [x; x+\theta+c]} g(y)$ is supermodular ($\theta \geq 0$, $c \geq 0$).*

2. *If* $g(y,z)$ *is supermodular, then* $H(x,z) = \min_{y\in[x;x+c]} g(y,z)$ *is supermodular* $(c \geq 0)$.

3. *If* $g(y,z)$ *is supermodular, then* $H(\theta,z) = \min_{y\in[x;x+c+\theta]} g(y,z)$ *is supermodular* $(\theta \geq 0, c \geq 0)$.

Proof : Proof of part 1: We define the global optimum point as $\hat{y} := \min_{y\in\mathbb{R}} g(y) \in \mathbb{R} \cup \{-\infty, +\infty\}$. For supermodularity, we aim to show for all $x^- \leq x^+, \theta^- \leq \theta^+$ that

$$H\left(x^+,\theta^-\right) + H\left(x^-,\theta^+\right) \leq H\left(x^-,\theta^-\right) + H\left(x^+,\theta^+\right)$$

The domain of $H(x,\theta)$ can be divided into three parts.

$$H(x,\theta) = \begin{cases} g(x) \text{ increasing in } x & \text{, if } \hat{y} \leq x \\ g(\hat{y}) \text{ constant} & \text{, if } \hat{y} - c \leq x + \theta \text{ and } x \leq \hat{y} \\ g(x+\theta+c) \text{ decreasing in } x+\theta & \text{, if } x + \theta \leq \hat{y} - c \end{cases}$$

We can observe that $H\left(x^-,\theta^+\right) \leq H\left(x^-,\theta^-\right)$ holds for all x^-,θ^-,θ^+, when $\theta^- \leq \theta^+$. We distinguish two cases by where x^+ is situated:

When $x^+ \geq \hat{y} - c$, then $H\left(x^+,\theta^-\right) = H\left(x^+,\theta^+\right)$ holds, which implies that supermodularity inequality holds.

When $x^+ \leq \hat{y} - c$, then we can define a function with a single variable $h(x+\theta) := H(x,\theta)$ which is convex (as discussed in Theorem 2.1.1). Note that a function (H) is supermodular if it is defined as a single argument convex function (h) at its arguments' non-negative linear combination, due to Lemma 2.6.2.a in Topkis (1998). This completes the proof. ■

Proof of part 2: We introduce $y^A := \arg\min_{y\in[x^-;x^-+c]} g(y,z^-)$ and $y^B := \arg\min_{y\in[x^+;x^++c]} g(y,z^+)$ with $x^- \leq x^+$ and $z^- \leq z^+$. Now we can express $H(x^-,z^-)$ and $H(x^+,z^+)$ as $g\left(y^A,z^-\right)$ and $g\left(y^B,z^+\right)$, respectively. Furthermore, $H(x^-,z^+) \leq g\left(y^A,z^+\right)$ and $H(x^+,z^-) \leq +g\left(y^B,z^-\right)$ holds because $y^A \in [x^-;x^-+c]$ and $y^B \in [x^+;x^++c]$.

If $y^A \leq y^B$, then by supermodularity of g, we have

$H(x^-,z^+) + H(x^+,z^-) \leq g\left(y^A,z^+\right) + g\left(y^B,z^-\right) \leq g\left(y^A,z^-\right) + g\left(y^B,z^+\right) = H(x^-,z^-) + H(x^+,z^+)$ from which supermodularity of H follows.

If $y^B \leq y^A$, then both y^A and y^B are in the $[x^+,x^-+c]$ interval, which imply $H(x^+,z^-) \leq g\left(y^A,z^-\right)$ and $H(x^-,z^+) \leq g\left(y^B,z^+\right)$. Therefore,

$H(x^+,z^-)+H(x^-,z^+)\leq g\left(y^A,z^-\right)+g\left(y^B,z^+\right)=H(x^-,z^-)+H(x^+,z^+)$ from which supermodularity of H follows. ■

Proof of part 3: We introduce $y^-:=\underset{y\in[x;x+c+\theta^-]}{\arg\min}\ g(y,z)$ and $y^+:=\underset{y\in[x;x+c+\theta^+]}{\arg\min}\ g(y,z)$, for which $y^-\leq y^+$ obviously holds.

By supermodularity of $g(y,z)$, we have $H\left(\theta^-,z^-\right)+H\left(\theta^+,z^+\right)=$

$g(y^-,z^-)+g(y^+,z^+)\geq g(y^+,z^-)+g(y^-,z^+)=H\left(\theta^+,z^-\right)+H\left(\theta^-,z^+\right)$ implying the supermodularity of $H(\theta,z)$. ■

Proof of Theorem 2.2:

Proof of part 1 is by induction. The base step consists of the following substeps:

$f_{T+1}\equiv 0$ and $J_T(y_T,U)=\mathcal{L}_T(y_T)+f_{T+1}$ are obviously supermodular. $\underset{y_T\in[x_T;x_T+\theta_T+U]}{\min}\mathcal{L}_T(y_T)$ is supermodular in $\left(x_T,\theta^T,U\right)$ by Lemma 2.2.1. Finally, $f_T\left(x_T,\theta^T,U\right)=Uc_p+\theta_T c_c+\underset{y_T\in[x_T;x_T+\theta_T+U]}{\min}\mathcal{L}_T(y_T)$ and $J_{T-1}\left(y_{T-1},\theta^T,U\right)=\mathcal{L}_{T-1}(y_{T-1})+\alpha E\left[f_T\left(y_{T-1}-D_{T-1},\theta^T,U\right)\right]$ are supermodular because of the supermodularity preservation of the non-negative linear combination and limit operators (see Lemma 2.6.1 and Corollary 2.6.2 in Topkis (1998)).

The general inductive step includes substeps as in the base step, and one additional substep. That is to prove

$$\begin{cases}\underset{y_t\in[x_t;x_t+\theta_t+U]}{\min}J_t\left(y_t,\theta^{t+1},U\right) & \text{is supermodular in } \left(x_t,\theta^t,U\right)\\ & \text{with } \theta^t=(\theta_t,\theta_{t+1},\ldots,\theta_{T-1},\theta_T),\\ & \text{if } T<t+L \text{ and}\\ \underset{y_t\in[x_t;x_t+\theta_t+U],\theta_{t+L}\geq 0}{\min}J_t\left(y_t,\theta^{t+1},U\right) & \text{is supermodular in } \left(x_t,\theta^t,U\right)\\ & \text{with } \theta^t=(\theta_t,\theta_{t+1},\ldots,\theta_{t+L-1},\theta_{t+L}),\\ & \text{if } t+L\leq T\end{cases}$$

given that $J_t\left(y_t,\theta^{t+1},U\right)$ is supermodular and convex in $\left(y_t,\theta^{t+1},U\right)$. The first branch follows directly from Lemma 2.2, the second branch follows from the supermodularity preservation of the projection operator (see Topkis (1998)), additionally. ■

Proof of part 2, first statement: From part 1, we have $f_{t+1}\left(x_{t+1},\theta^{t+1},U\right)$ being supermodular. By the definition of supermodularity, for $x^-\leq x^+, z^-\leq z^+$ we have $f_{t+1}(x^+,z^-)+f_{t+1}(x^-,z^+)\leq f_{t+1}(x^-,z^-)+f_{t+1}(x^+,z^+)$, where $z^\pm=$

$\left(\theta_t^{\pm}, \ldots, \theta_{\min\{t+L,T\}}^{\pm}, U^{\pm}\right)$ are vectors such that $\theta_t^- \leq \theta_t^+$, $\ldots$, $\theta_{\min\{t+L,T\}}^- \leq \theta_{\min\{t+L,T\}}^+$, and $U^- \leq U^+$.

We introduce new variables $w^- := y - x^+$, $w^+ := y - x^-$ with an arbitrary y. For $w^- \leq w^+, z^- \leq z^+$ we have $f_{t+1}(y - w^-, z^-) + f_{t+1}(y - w^+, z^+) \leq f_{t+1}(y - w^+, z^-) + f_{t+1}(y - w^-, z^+)$ for all y. This means that $H_t(w, z) := f_{t+1}(y - w, z)$ is submodular for all t. By submodularity preservation of the expected value and the non-negative linear combination operators (see Topkis (1998)), $J_t = \mathcal{L}_t(y_t) + \alpha E\left[H_t\left(D_t, \left(\theta^{t+1}, U\right)\right)\right]$ is also submodular in $\left(D_t, \left(\theta^{t+1}, U\right)\right)$. ■

Proof of part 2, second statement: We denote the first order stochastic dominance by $\preceq$. Since $f_{t+1}\left(x_{t+1}, \theta^{t+1}\right)$ is convex for all t, we have $H_t(x) := f_{t+1}\left(x, \theta^{t+1}\right)$ convex for all t. $H_t^-(w, y) := H_t(y - w)$ is also submodular in (w, y) for all t (due to Lemma 2.6.2.b in Topkis (1998)).

We introduce $Q(w) := H_t^-(w, y^-) - H_t^-(w, y^+)$ with some $y^- \leq y^+$. Because $H_t^-(w, y)$ is submodular, it has non-increasing differences (see Theorem 2.6.1 in Topkis (1998)), so $Q(w)$ is non-increasing. Therefore, for any $D^- \preceq D^+$, we have $Q(D^+) \preceq Q(D^-)$ implying $E[Q(D^+)] \leq E[Q(D^-)]$, as well (see Proposition 4.1.1 and Lemma 4.7.2 in Puterman (1994)). The latter expression means that $E\left[H_t^-(D_t, y_t)\right]$ has non-increasing differences, which is equivalent with its submodularity in (D_t, y_t).

Finally, $J_t = \mathcal{L}(D_t, y_t) + \alpha E\left[H_t^-(D_t, y_t)\right]$ is submodular in (D_t, y_t) because of Lemma 2.1 and the submodularity preservation of the non-negative linear combination operator (see Corollary 2.7.2 in Topkis (1998)). ■

Proof of Theorem 2.4:

We prove the first statement indirectly. The limiting (y_1^*, θ_2^*) for $T \to \infty$ exist because of the discountedness (see Puterman (1994)), and $y_1^* = \hat{y}_1$ holds. Let $D_{\min} > 0$ be the smallest possible realization of D. Assume, that (y_1^*, θ_2^*) does not satisfy the complementary slackness property. We can define another feasible strategy (y_1, θ_2) such that $y_1 := y_1^* + \varepsilon$ and $\theta_2 := \theta_2^* - \varepsilon$ with $\varepsilon := \min\left\{\frac{D_{\min}}{2}, x_1 + \theta_1 + U - y_1^*, \theta_2^*\right\} > 0$. We study the cost difference ΔJ_1 between the two strategies.

$$\begin{aligned}
\Delta J_1 := \quad & J_1(y_1^*, \theta_2^*, U) - J_1(y_1, \theta_2, U) \\
= \quad & \Pr[\hat{y}_1 \in [x_2 + \varepsilon; \infty)]\left(\mathcal{L}(y_1^*) - \mathcal{L}(y_1) + \alpha c_c \varepsilon\right) \\
& + \Pr[\hat{y}_1 \in [x_2; x_2 + \varepsilon)] C_1 \\
& + \Pr[\hat{y}_1 \in (-\infty; x_2)] C_2
\end{aligned}$$

with $x_2 = y_1^* - D$, and some C_1 and C_2 expected costs for the remaining periods. By the first term of the summation, the two strategies follow the same sample paths from the second period on, while none of the latter two terms are possible, as $0 < \varepsilon < D_{\min}$. Thus, we have $\Pr[\hat{y}_1 \in [x_2; x_2 + \varepsilon)] = \Pr[\hat{y}_1 \in (-\infty; x_2)] = 0$ and $\Pr[\hat{y}_1 \in [x_2 + \varepsilon; \infty)] = 1$.

Therefore, $\Delta J_1 = \mathcal{L}(y_1^*) - \mathcal{L}(y_1^* + \varepsilon) + \alpha c_c \varepsilon$. Since $\mathcal{L}$ is convex and $\mathcal{L}' < h$, we have $\mathcal{L}(y_1^* + \varepsilon) - \mathcal{L}(y_1^*) < h\varepsilon$. Using the required $h < \alpha c_c$ sufficiency condition, we find $\Delta J_1 > -h\varepsilon + \alpha c_c \varepsilon > 0$. However, the positive ΔJ_1 contradicts with (y_1^*, θ_2^*) being the optimum. $\square$

The proof of the second part is as follows. For the two-period problem, J_1 can be expressed explicitly. For a given U, the curve of the intersection of J_1 with the plane $y_1 + \theta_2 = 0$ defines a new function, $\tilde{J}_1$, which we parameterize with variable y_1.

$$\begin{aligned}\tilde{J}_1(y_1) = \; & \mathcal{L}_1(y_1) + \alpha U c_p - \alpha y_1 c_c + \alpha \mathcal{L}_2(\hat{y}_2) \left[G_1(\omega)\right]_{y_1 - \hat{y}_2}^{U - \hat{y}_2} \\ & + \alpha \int_{-\infty}^{y_1 - \hat{y}_2} \mathcal{L}_2(y_1 - \omega) \; g_1(\omega) \; d\omega \\ & + \alpha \int_{U - \hat{y}_2}^{\infty} \mathcal{L}_2(U - \omega) \; g_1(\omega) \; d\omega\end{aligned}$$

We take the derivative of function $\tilde{J}_1(y_1)$ and look for negative values.

$$\begin{aligned}0 > \partial_{y_1} \tilde{J}_1(y_1) = \; & + \mathcal{L}_1'(y_1) - \alpha c_c - \alpha \mathcal{L}_2(\hat{y}_2) \, g_1(y_1 - \hat{y}_2) \\ & + \alpha \partial_{y_1} \int_{-\infty}^{y_1 - \hat{y}_2} \mathcal{L}_2(y_1 - \omega) \; g_1(\omega) \; d\omega\end{aligned}$$

As a result, we have the inequality,

$$\mathcal{L}_1'(y_1) + \alpha(h + b) \int_{-\infty}^{y_1 - \hat{y}_2} G_2(y_1 - \omega) \, g_1(\omega) \, d\omega < \alpha c_c + \alpha b G_1(y_1 - \hat{y}_2)$$

By increasing its LHS, we create a sufficient condition for this inequality to hold.

$$\begin{aligned}& \mathcal{L}_1'(y_1) + \alpha(h + b) \int_{-\infty}^{y_1 - \hat{y}_2} G_2(y_1 - \omega) \, g_1(\omega) \, d\omega \\ & \leq h + \alpha(h + b) \int_{-\infty}^{y_1 - \hat{y}_2} 1 \; g_1(\omega) \, d\omega \;\; = \;\; h + \alpha(h + b) \, G_1(y_1 - \hat{y}_2) \\ & \leq h(1 + \alpha) + \alpha b G_1(y_1 - \hat{y}_2)\end{aligned}$$

When we check if the increased LHS is still below its RHS, we find

$$h\left(1+\alpha\right)+\alpha bG_1\left(y_1-\hat{y}_2\right)<\alpha c_c+\alpha bG_1\left(y_1-\hat{y}_2\right)$$

which is equivalent to $h\left(1+\alpha\right)<\alpha c_c$.

Consequently, for a given U, $\{y_1+\theta_2=0,\ y_1\rightarrow+\infty\}$ is an always decreasing ray for $J_1\left(y_1,\theta_2,U\right)$ when $h\left(1+\alpha\right)<\alpha c_c$. Therefore, the constrained optimum of the first period satisfies the complementary slackness property. ■

Chapter 3

Workload-Dependent Capacity Control in Production-to-Order Systems

The development of job intermediation and the increasing use of the Internet allow companies to carry out ever quicker capacity changes. In many cases, capacity can be adapted rapidly to the actual workload. The use of fast response contingent capacity next to the permanent capacity is an especially important mean for coping with demand variations in production-to-order systems, where, as opposed to production-to-stock systems, inventory cannot be used as a buffer. We introduce a set of Markov chain models to represent workload-dependent capacity control policies. We present two analytical approaches to evaluate the policies' due-date performance based on stationary analysis. One provides an explicit expression of throughput time distribution, the other is a fixed-point iteration method that calculates the moments of the throughput time. We compare due-date performance, capacity, capacity switching, and lost sales costs to select optimal policies. We also give insight into which situations a workload-dependent policy is beneficial to introduce. Our results can be used by manufacturing and service industries when establishing a policy for dynamic capacity planning.

3.1. Introduction

Most production-to-order companies do not have a constant flow of orders. This often leads to a varying queue of customers. Despite the uncertainty in both arrival and service of orders, customers request tight due dates, and they are also resentful of late deliveries. To compensate for the variations in the orders' frequency, they cannot use inventory, but a good practical solution is to try to match the production rate to the actual workload or, when orders are statistically identical, to the length of the queue. Such a policy may bring benefits in labor cost reduction or in due-date performance. This chapter investigates the value of using an optimal workload-dependent policy, where "workload" means the number of orders in the system.

For studying workload-dependent capacity planning, we found a strong real-life motivation in the area of engineer-to-order (ETO). As a definition for ETO, we quote that of Gelders (1991): "In an engineer-to-order environment a company *designs* and *produces* products to customer order." In the same paper, Gelders concludes that ETO companies need "fast-response capacity," which is to be considered as a general characteristic of competitive ETO production.

Similarly to ETO companies, production-to-order (PTO) companies also need to adjust their capacity to meet customer demand (see e.g. Vollmann *et al.* (2005)). Apart from the motivating example for ETO in the paper of Gelders, we give three examples of PTO companies, where workload-dependent capacity management is applicable. In our first example, we describe a general manufacturing situation with an expensive bottleneck machine. In this case, the number of shifts the bottleneck machine works determines the production capacity. The capacity can be set between zero to three shifts, depending on the workload. Our second and third examples are companies, which typically face unforeseen variations in the number of orders. The second is a special translator service, where topic leaders share translation tasks and assign them to specialized free-lance translators. The third example is a data recording company, where audio tapes are transcribed to digital text format in order to make later searches possible. While workers listen to the tapes, they type the text into a computer.

In all the examples, using fast-response contingent capacity next to the permanent capacity is not just desirable, but also affordable. Production capacity in manufacturing does not usually require high educational labor; labor acquisition lead times are often very short. Although, in most service industry situations labor has specific knowledge, acquisition lead time substantially decreased in the last decade due to the increasing use of the Internet. In the case of the translator company, contact information of free-lance translators

is carefully maintained, so distributing new tasks takes just hours. In case of workers with high education, capacity flexibility can be gained by using overtime or by contracting workers for flexible working times (see e.g. Filho and Marçola (2001)).

In this chapter, we assume that we can afford instantaneous capacity changes as an abstraction of opportunity for fast-response contingent capacity. For our analysis, we consider a stationary, homogeneous Poisson arrival process of orders, exponentially distributed service time, and FCFS service discipline. We consider two tactical decisions. These are on the permanent capacity level and on the policy for the use of fast-response contingent workers. The contingent worker policy declares the necessary number of workers depending on the number of jobs in the system. The actual number of contingent workers (servers) is assumed to follow the contingent worker policy. Our objective is to minimize the total average cost per time unit, where the costs consist of labor costs, costs of contingent capacity adjustments, costs related to order acceptance, work-in-process holding, and early or tardy order completion.

For this problem, we present our model by applying two evaluation approaches. One approach aims at the exact calculation of the distribution of the throughput time (time spent in the system). Another approach, which can be used to deal with problems of a larger scale, is based on a moment-approximation of the throughput time. With respect to the effectiveness of the flexible capacity it is essential to increase and decrease capacity at the right moment. We determine the proper switching states for small scale problem instances by searching exhaustively. The optimal strategy is compared in terms of performance with the fixed capacity rules, which utilize permanent capacity only.

Previously, a number of researchers pointed out relations among service rate, work-in-process and due-date performance. One application of due-date performance measures is the evaluation of batch scheduling rules. In the scheduling rules, capacity level and work-in-process are both consistently present as parameters showing the strong relation to the due-date performance (see e.g. Philipoom *et al.* (1993)). Lead time setting is a topic where work-in-process is also recognized as an adequate parameter, which improves the performance of the rules (see e.g. Bertrand (1983)). These two examples reveal the connection between capacity level, work-in-process and due-date performance. However, there is no model in the literature that incorporates all these three closely related notions.

Our contribution to the existing literature is studying a queuing system with adaptive capacity to satisfy the objective of having reliable production lead times. Moreover, our examples and their analysis provide insight into when

a workload-dependent strategy is effective, and also what characteristics this strategy has.

The chapter is organized as follows. In the next section, we describe the related literature. Then, in Section 3.3, we give a mathematical formulation of our problem. Results on the effectiveness and characteristics of the strategy are shown in Section 3.4. Finally, conclusions and plans for future extensions are presented.

3.2. Related literature

Decisions on capacity changes were studied first in capacity expansion problems. In the case of deterministic demand with positive trend, Chenery (1952) found that gas pipelines constantly have extra capacity. This was the basis of his "excess capacity hypothesis" that says capacity is always larger than demand; optimal overcapacity is to be investigated by looking at economies of scale. Manne (1961) revised Chenery's hypothesis when extending his model. The extension of the model with a backlog option created an environment in which the "excess capacity hypothesis" no longer held. Manne also studied stochastic, stationary demand without the backlog option. This model resulted in a smaller deviation from what Chenery's model suggested. Luss (1982) gave a comprehensive review of the literature on capacity expansion.

Models with capacity expansion/reduction decisions and hiring/firing costs are usually studied by means of dynamic programming. One example is the continuous time DP model of Bentolila and Bertola (1990), where sensitivity analysis on firing costs was presented. Rocklin *et al.* (1984) studied a service system with non-stationary demand, including both capacity expansion/reduction decisions and hiring/firing costs. Rocklin et al. showed the optimality of the (S', S'')-policy known from inventory theory by the means of discrete time DP. In their model, demand must always be met; if demand exceeds available capacity, the capacity must be immediately increased to overcome the deficit. For the most recent developments on capacity expansion/reduction models see Eberly and van Mieghem (1997) (multi-factor investment strategies) and Angelus and Porteus (2002) (aggregate planning). These two papers provide similar optimality results as in Rocklin et al. for models, which do not require automatical capacity addition as a response for capacity shortage.

Pinder (1995) deduced an approximation of optimal workload-dependent capacity control policy for stationary demand. In the model of Pinder, capacity (resources) is treated as discrete; capacity adjustments are dependent on the actual number of jobs (work), which are particular points in common with the

content of this chapter. However, the workload-dependent policy class defined by Pinder seemed too broad to give an explicit formula of policy evaluation or to find an optimal solution. In addition, Pinder did not consider due-date performance in any form, which is an essential part of our model. Besides, we define a less broad policy class that entails less limitations in the analysis so that we can provide additional insights.

Queuing models have made use of servers with load-dependent service times for approximate performance analysis since the fundamental work of Avi-Itzhak and Heyman (1973). These servers are apt to represent a sub-network as they can be set to provide nearly the same characteristics. Marie (1979) introduced a similar approximation technique. A comparison of the two techniques is given in Baynat and Dallery (1993).

In some simple workload-dependent policy classes, the optimal (capacity) control policy can be analytically determined. Faddy (1974) introduced the class of P_λ^M policies for the control of water reservoirs. Namely, the policy P_λ^M sets the output rate to M if the water level in the reservoir reaches (or exceeds) λ, and the output rate is set to zero if the reservoir is empty. This policy is still a subject of research (see Kim *et al.* (2006)). Another simple policy class, the two-step service rule, was defined in Bekker and Boxma (2005). A policy of this class has a lower service rate, r_1, when the workload is not more than K, and a higher service rate, r_2, if the limit K is exceeded. Such threshold-type workload-dependent policies are also used for managing call-centers, border-crossing stations, and airport check-in desks (see Bhandari *et al.* (2008), the references therein, and Zhang (2005)). This policy class has the drawback that the undesired frequent service rate changes are not excluded. Tijms and van der Duyn Schouten (1978) proposed a different class of policies for inventory control, called two switch-over level rule. A rule is characterized by two inventory levels, $y_2 \leq y_1$. An inventory decrease to y_2/increase to y_1 triggers compensating raise/reduction. This design restrains frequent service rate changes. In the next section, we specify a class of control policies for managing capacity with a higher degree of freedom than the three simple classes discussed. As a result, finding the optimal solution in our policy class allows a better characterization of how to control capacity optimally.

3.3. Model formulation

In the context of this chapter, a workload-dependent capacity planning policy is defined by two "switching points", one down and one up switching point, for each pair of neighboring capacity levels. Switching points are specific workload

values, for which a job arrival or departure can trigger a switch in capacity. If the system is at an *up switching point*, and a job arrives, we switch from the lower capacity level to the upper. If the system is at a *down switching point*, and a job departs, we switch from the upper capacity level to the lower. Each up-switch incurs a hiring cost (c_h), and each down-switch incurs a firing cost (c_f). The firm has an admissible domain of permanent and contingent capacity levels. The feasible permanent capacity levels are the integers larger than or equal to $U_{\min} \geq 0$, and there is a joint constraint restricting the total capacity level being at most $C_{\max} \geq 0$. Naturally, the firm also needs to pay for the used capacity. The unit costs of permanent and contingent capacity per unit time are denoted by c_p and c_c, respectively.

Although this chapter does not investigate the impact of efficiency aspects, one may also take into account the psychological effect of lower or higher workload (see Bertrand and van Ooijen (2002)), and the efficiency of using different levels of capacity (see Schlichter (2005)). Practically, the service rate can be measured for each workload and capacity level ($\mu_{w,c}$) and used in our model.

The firm works with a fixed lead time (L). It has to pay charges for each time unit when a job is late (c_t). A smaller cost is due for holding if a job is ready before the due date (c_e). The same or somewhat less needs to be paid for the work-in-process holding (c_w). The number of jobs with which the firm can deal is constrained from above by a constant integer, $W_{\max}$. New jobs are refused if the firm already has $W_{\max}$ number of jobs pending. Consequently, lost sales can occur, which is penalized by a cost of c_l for each occasion. To sum up, we have an eight element vector of cost coefficients $(c_h, c_f, c_p, c_c, c_t, c_e, c_w, c_l)$. Although we use linear cost components for the ease of presentation, our analysis applies for general cost functions, as well.

In this section, we define a workload-dependent capacity planning policy class, and formalize the firm's capacity control problem.

3.3.1 Definitions, assumptions and problem formulation

We assume a stationary environment with a homogeneous Poisson arrival processes and exponential service times. Arrivals have a constant rate of λ; departures have, in general, capacity level and workload-dependent rates. We denote by $\mu_{w,U,\theta}$ the service rates for a given number of jobs, w (workload), and permanent/contingent capacity levels, U and θ. For the sake of better comprehension, we assume

$$\mu_{w,U,\theta} = \begin{cases} c\mu & \text{by joint processing, and} \\ \min\{w, U\}\,\mu & \\ \quad +\gamma\left(\min\left\{\max\{0, w-U\}, \theta\right\}\right)\mu & \text{by one-to-one processing} \end{cases}$$

for all w, with $c = U+\gamma\theta$, where γ is the average productivity rate of contingent workers, relative to the productivity of permanent workers. Joint processing refers to the situations, where all capacity can concentrate on a single job (as e.g. the three shifts case), while by one-to-one processing each job is processed by one unit of permanent or contingent capacity (as e.g. machine processing). In what follows, we use joint processing for the sake of easier presentation; the model and analysis of the one-to-one processing mode differs only slightly from the joint processing. Note that the $\gamma = 1$ case corresponds to assuming identical servers (machines, workers or shifts). Jobs are served according to FCFS discipline.

We define the set of workload-dependent capacity control policies for given values of $U_{\min}$, γ, $C_{\max}$ and $W_{\max}$ as $\Omega\left(U_{\min}, \gamma, C_{\max}, W_{\max}\right)$. A feasible policy can be characterized by a triple $\Psi = (U, \theta_{\max}, \Theta)$. The terms U, and $\theta_{\max}$ stand for the permanent, and maximum contingent capacity level used by the policy. Necessarily, these terms have to satisfy the inequalities, $U \geq U_{\min}$, $\theta_{\max} \geq 0$ and $U + \theta_{\max} \leq C_{\max}$. The term Θ denotes a $\theta_{\max}$-by-2 matrix, which describes all the switching points. For the cases when $\theta_{\max} = 0$, we have an empty matrix. We use indices $c \to c+\gamma$ or $c+\gamma \to c$ with $c \in \{U + \gamma\theta : 0 \leq \theta \leq \theta_{\max},\ \theta \textit{ is integer}\}$ to show that the switch occurs between the capacity levels c and $c+\gamma$ either up or down. Particularly, $\Theta_{c\to c+\gamma}$, and $\Theta_{c+\gamma\to c}$ are workloads, where the capacity is changed from c to $c+\gamma$, upwards, or from $c+\gamma$ to c, downwards, if a job arrival or departure occurs, respectively. The elements of Θ are constrained by $W_{\max}$, and by each other as follows,

- Constraints on the lowest down- and the highest up-switching points
  $1 \leq \Theta_{U+\gamma\to U}$ and $\Theta_{U+\gamma(\theta_{\max}-1)\to U+\gamma\theta_{\max}} \leq W_{\max} - 1$,
- Constraints between up-down switching points for each pair of capacity levels
  $\Theta_{c+\gamma\to c} \leq \Theta_{c\to c+\gamma} + 1$ for all $c \in \{U + \gamma\theta : 0 \leq \theta \leq \theta_{\max} - 1,\ \theta \text{ is integer}\}$,
- Constraints between down-switching points and between up-switching points
  $\Theta_{c+\gamma\to c} \leq \Theta_{c+2\gamma\to c+\gamma}$ and $\Theta_{c\to c+\gamma} \leq \Theta_{c+\gamma\to c+2\gamma}$
  for all $c \in \{U + \gamma\theta : 0 \leq \theta \leq \theta_{\max} - 2,\ \theta \text{ is integer}\}$.

In Figure 3.1, we show two workload-dependent capacity control policies, $\Psi = (1, 2, \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix})$ and $\Psi = (1, 2, \begin{pmatrix} 3 & 3 \\ 4 & 5 \end{pmatrix})$ in the set $\Omega\left(1, 1, 3, 6\right)$. E.g.

the former policy, $\Psi = (1, 2, \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix})$, has the permanent capacity and maximum contingent capacity level parameters set to $U = 1$ and $\theta_{\max} = 2$, up-switching points $\Theta_{1\to2} = 3$, $\Theta_{2\to3} = 4$, and down-switching points $\Theta_{2\to1} = 1$, $\Theta_{3\to2} = 2$. Other feasible policies are e.g. $\Psi = (1, 1, \begin{pmatrix} 3 & 1 \end{pmatrix})$, which uses only two capacity levels, or $\Psi = (1, 2, \begin{pmatrix} 3 & 4 \\ 4 & 5 \end{pmatrix})$, when capacity becomes a function of the workload. We assume that capacity adjustments can be done instantaneously, in parallel with the change in workload.

Within the defined set of workload-dependent capacity control policies, we define the subset of *fixed policies*, as the set of policies using only permanent capacity ($\theta_{\max} = 0$). Next, we define the set of *continuous fixed policies* broadening the set of fixed policies by allowing the permanent capacity level to be set to any real values between $U_{\min}$ and $C_{\max}$. The set of fixed policies is the intersection of the set of continuous fixed policies and the set of workload-dependent policies.

In order to simplify the exposition, we do not separate costs of hiring to firing, but use a single cost coefficient, c_s that penalizes capacity adjustments in general. Note that as stationarity implies balance of hiring and firing, we can aggregate the cost coefficients of hiring and firing to a switching cost coefficient: $c_s = (c_h + c_f)/2$. Based on this formula, cost counting with c_s provides the same result as separate cost counting with c_h and c_f.

For the cost coefficients mentioned in the model description, we have the related costs. These costs can be separated into two groups. The costs of capacity, switching, lost sales, and work-in-process holding are functions of the policy only. The costs of earliness and tardiness are functions of the lead time, additionally.

Now, we can formulate the capacity control problem as

$$\min_{\Psi\in\Omega(U_{\min},\gamma,C_{\max},W_{\max})} cost_{group1}(\Psi) + cost_{group2}(\Psi, L) \tag{1}$$

where L is the fixed lead time, $cost_{group1}(\Psi)$ is the sum of $cost_{capacity}(\Psi)$, $cost_{switching}(\Psi)$, $cost_{lostsales}(\Psi)$, $cost_{WIP}(\Psi)$ and $cost_{group2}(\Psi, L)$ is the sum of $cost_{earliness}(\Psi, L)$, and $cost_{tardiness}(\Psi, L)$ (see the description of the cost components in Table 3.1). E.g. the first policy (left) in Figure 3.1, we can expect to outperform the second policy (right) in all the costs but $cost_{capacity}$.

$cost_{group1}(\Psi)$	cost components dependent only on the policy
$cost_{capacity}(\Psi)$	permanent and contingent capacity costs
$cost_{switching}(\Psi)$	contingent capacity level adjustment costs
$cost_{lostsales}(\Psi)$	lost sales costs
$cost_{WIP}(\Psi)$	holding costs of the work-in-process
$cost_{group2}(\Psi, L)$	cost components dependent on the lead time, L, as well
$cost_{earliness}(\Psi, L)$	holding costs of the finished products
$cost_{tardiness}(\Psi, L)$	tardiness costs

Table 3.1: Description of the cost components (all are in terms of average daily costs)

3.3.2 Evaluation of cost functions

In this section, we aim to express the costs one by one. First, we derive the costs that are independent of the quoted lead time ($cost_{group1}$) except for $cost_{WIP}(\Psi)$. After that, we show two approaches to evaluate the lead time dependent costs ($cost_{group2}$) and $cost_{WIP}(\Psi)$. We note that in real-life situations, where the cost functions need to be evaluated for general arrival and service processes, our Poisson process based approaches can help finding the optimal policy by giving a starting guess, however simulation is necessary for tuning the policy parameters afterwards.

We create a Markov chain, $MC^{\Psi,\gamma,\lambda,\mu}$, according to the policy, the given arrival rate, and the given service rate unit. In this Markov chain, we have the states labeled by (w, c), the workload (number of jobs in the system), and the capacity usage. A policy Ψ can be given by the arrival matrix (A^{Ψ}), and the departure matrix (D^{Ψ}). These matrices are square; rows and columns are indexed by the states of the Markov chain. Elements are ones at the related arrival or departure arcs of the Markov chain and zeros elsewhere. The state space of $MC^{\Psi,\gamma,\lambda,\mu}$ depends on the policy. As an example, we show the transition rate diagram of $MC^{\Psi,\gamma,\lambda,\mu}$ in Figure 3.1 for the policies $\Psi = (1, 2, \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix})$ and $\Psi = (1, 2, \begin{pmatrix} 3 & 3 \\ 4 & 5 \end{pmatrix})$. Note that the previously defined policy class, $\Omega(U_{\min}, \gamma, C_{\max}, W_{\max})$, for $U_{\min} = 0$ contains policies with $U = 0$, which correspond to $M|M|\theta_{\max}|W_{\max}$ queuing systems for any $\theta_{\max} = 1, 2, \ldots, C_{\max}$.

Via steady state analysis, the first cost group can be easily evaluated. If we add λA^{Ψ} to μD^{Ψ} weighted in rows (w, c) by capacity c, we can obtain Q^{Ψ}, the transition rate matrix of the Markov chain.

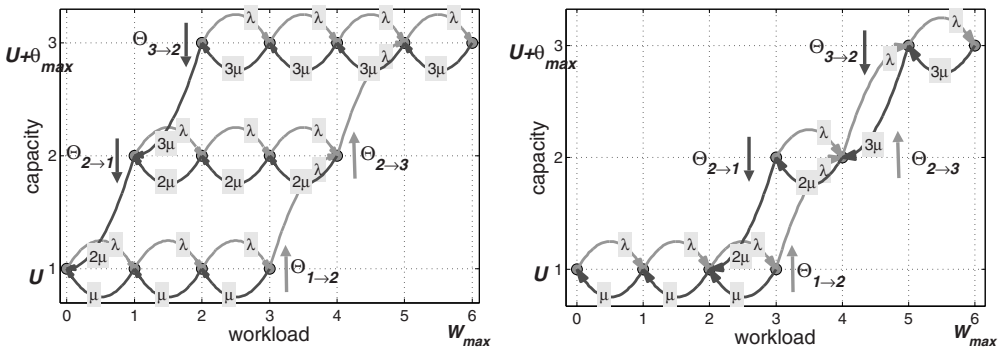

Figure 3.1: $MC^{\Psi,\gamma,\lambda,\mu}$ for $\Psi = (1, 2, [3, 1; 4, 2])$ and $\Psi = (1, 2, [3, 3; 4, 5])$

$$Q^{\Psi}_{(w_1,c_1),(w_2,c_2)} = \lambda A^{\Psi}_{(w_1,c_1),(w_2,c_2)} + c_1 \mu D^{\Psi}_{(w_1,c_1),(w_2,c_2)} \tag{2}$$

We define the diagonal elements of Q^{Ψ} as the negative state leaving rates: $v^{\Psi}_i := diag\left(Q^{\Psi}\right)_i = -\sum_{j \neq i} Q^{\Psi}_{i,j}$. Solving the linear equation system $\begin{cases} Q^{\Psi}\pi^{\Psi} &= \mathbf{0} \\ e\pi^{\Psi} &= 1 \end{cases}$, where e is the all one vector, we gain the limiting distribution π^{Ψ}, with elements $\pi^{\Psi}_{w,c}$ corresponding to the time fraction being in state (w, c). Eventually, we can specify all the costs of the first group.

$$\begin{aligned} cost_{capacity}(\Psi) &= c_p \sum_w \pi^{\Psi}_{w,U} + c_c \sum_{U<c} c \sum_w \pi^{\Psi}_{w,c} \\ cost_{adjustment}(\Psi) &= 2c_s\lambda \sum_{(w,c)\in\Upsilon} \pi^{\Psi}_{w,c} \\ cost_{lostsales}(\Psi) &= c_l \lambda P(lost\ sale) \\ cost_{WIP}(\Psi) &= c_w \lambda P(lost\ sale)\, E[X] \end{aligned} \tag{3}$$

where $P(lost\ sale) = \pi^{\Psi}_{W_{\max},U+\theta_{\max}}$ is the probability that a job is rejected, $E[X]$ is the throughput time's (or sojourn time's) expected value, and Υ is the set of states from which the up-switches in capacity are made. Recall that it is enough to count costs for the up-switches twice because of the stationarity assumption, and that the calculation of $cost_{WIP}$ needs the yet unknown expected throughput time.

In what follows, we complete the calculation of the work-in-process costs, and deduce the earliness and tardiness costs in two different ways. Both approaches aim at identifying the throughput time as a random variable so that the due-date performance of the policies can be determined.

3.3.2.1 Derivation of the throughput time distribution

In order to derive an explicit formula for the throughput time distribution, we observe an arbitrarily selected job while it is in the system. In addition to the system's workload and the used capacity, we observe the queue position of the pointed job. Therefore, we extend our Markov chain $MC^{\Psi,\gamma,\lambda,\mu}$ to an extended Markov chain $EMC^{\Psi,\gamma,\lambda,\mu}$, which has the queue position of the pointed job (q) as an extra dimension. $EMC^{\Psi,\gamma,\lambda,\mu}$ has states labeled by (w,c,q), the workload, capacity, and queue position of the pointed job, respectively, with $c \in \{U+\gamma\theta : 0 \leq \theta \leq \theta_{\max},\ \theta \ is\ integer\}$, and $0 \leq q \leq w \leq W_{\max}$. When the pointed job arrives, it enters one of the states, for which $0 < q = w$ hold; its lifetime ends when entering a state having $q = 0$. Because of the FCFS discipline, during the lifetime of the job, the queue position of the pointed job, q decreases when a job is ready, whereas it is unaffected by arrivals. Naturally, both arrivals and departures may affect its total time in the system.

We want to obtain the probability that starting from state r at time 0 we will be in state s at time t in $EMC^{\Psi,\gamma,\lambda,\mu}$. This probability we denote by $\bar{P}^{\Psi}_{r,s}(t)$. In the appendix, we show an example for $EMC^{\Psi,\gamma,\lambda,\mu}$ and derive an explicit expression for $\bar{P}^{\Psi}_{r,s}(t)$ by applying uniformization.

The states $(w,c,0)$, where the pointed job may exit the system, are modeled as absorbing states. This way, we can express the throughput time CDF as the sum of probabilities of getting to certain exit states as follows.

$$F^{r,\Psi}(t) = \sum_{s\in\{(w,c,0)\}} \bar{P}^{\Psi}_{r,s}(t) \tag{4}$$

is the throughput time CDF if the pointed job arrives when the system is in state $r = (w_r, c_r, q_r)$ with $q_r = w_r$.

The stationary distribution (π^{Ψ} for $MC^{\Psi,\gamma,\lambda,\mu}$) is different from the distribution right after the arrival of the pointed job, which we denote by π^{Ψ}_{A}. We can use the arrival matrix (A^{Ψ}) to determine the probability distribution of being in a state right after the arrival $\pi^{\Psi}_{A} = \pi^{\Psi}A^{\Psi}$. We can express the after arrival distribution of the extended Markov chain, $EMC^{\Psi,\gamma,\lambda,\mu}$, which we call $\bar{\pi}^{\Psi}_{A}$, from the after arrival distribution of $MC^{\Psi,\gamma,\lambda,\mu}$.

$$\bar{\pi}^{\Psi}_{A,(w,c,q)} = \begin{cases} \pi^{\Psi}_{A,(w,c)} & \text{, if } q = w \\ 0 & \text{, if } q \neq w \end{cases} \tag{5}$$

Once we have the extended starting distribution $\bar{\pi}^{\Psi}_{A}$, the throughput time CDF (F^{Ψ}) can be expressed.

$$F^{\Psi}(t) = \sum_{\substack{r=(w_r,c_r,q_r)\\ q_r=w_r}} \bar{\pi}^{\Psi}_{A,r} F^{r,\Psi}(t) \tag{6}$$

Finally we can give formulas for the lead time dependent cost functions.

$$cost_{earliness}(\Psi, L) = c_e \lambda (1 - P(lost\ sale)) \int_0^L (L - t)\ dF^{\Psi}(t)$$

$$cost_{tardiness}(\Psi, L) = c_t \lambda (1 - P(lost\ sale)) \int_L^{\infty} (t - L)\ dF^{\Psi}(t) \tag{7}$$

In the special case, when $c_e = c_w$, we can use a simplified formula for the sum of the throughput time related costs.

$$cost_{earliness}(\Psi, L) + cost_{tardiness}(\Psi, L) + cost_{WIP}(\Psi) = \lambda (1 - P(lost\ sale)) \left((c_t + c_e) \int_L^{\infty} (t - L)\, dF^{\Psi}(t) + c_e L \right) \tag{8}$$

3.3.2.2 Moment approximation of the throughput time

Apart from the distribution function, we also derive a moment approximation for the throughput time as an alternative for the steps from applying uniformization to the step of expressing $F(t)$ in (6). This approach allows evaluation of large scale problems, as it both speeds up the calculations and needs less memory. Besides, throughput time moments can be determined with higher accuracy than extracting them from the throughput time distribution.

We evaluate the moments based on the equation that describes the relation of the conditional expected throughput times. We denote the kth moment of the throughput time if starting from state r by $E\left[(X_r)^k\right]$ and the duration of the visit to state r by Z_r. Then we have

$$E\left[(X_r)^k\right] = \sum_s \bar{P}^{\Psi}_{r,s}\ E\left[(Z_r + X_s)^k\right] \tag{9}$$

where $\bar{P}^{\Psi}$ is the extended transition probability matrix, containing the probabilities of going from state r to some neighboring state s. Similar to the approach in the previous subsection, a state is described by the components (w, c, q) and therefore the definition of $\bar{P}^{\Psi}$ in (9) is identical to the one in the appendix, equation (17).

For small state spaces, we can solve this linear system by matrix inversion, but for larger state spaces, we need a vector iteration to determine the moments for the relevant states. This vector iteration can be established by indexing the equation by the iterator n, and declaring the starting state.

$$\begin{cases} E_{n+1}\left[(X_r)^k\right] = \sum_s \bar{P}^{\Psi}_{r,s}\ E_n\left[(Z_r + X_s)^k\right] \\ E_0\left[(X_r)^k\right] = 0 \end{cases} \tag{10}$$

Using the independence of conditional throughput time and visit duration, we can write

$$\begin{aligned} E\left[(Z_r + X_s)^k\right] &= \sum_{j=0}^{k} \binom{k}{j} E\left[(Z_r)^{k-j}(X_s)^j\right] \\ &= \sum_{j=0}^{k} \binom{k}{j} E\left[(Z_r)^{k-j}\right] E\left[(X_s)^j\right] \end{aligned} \tag{11}$$

We can notice in expression (11) that for the evaluation of higher moments all the previous ones are needed in the iteration, (10). In the general form, we evaluate the first K moments.

We can use the following algorithm. We increase the moment iterator k from 1 to K. For each k we take limit of n at ∞ with the vector iteration to evaluate the moments $E\left[(X_r)^k\right]$ one after the other, inductively. Similarly to (6), we weight the conditional moments by the after arrival distribution of the extended Markov chain, $EMC^{\Psi,\gamma,\lambda,\mu}$ (that is $\bar{\pi}_A^{\Psi}$), which gives the kth moment of the throughput time, $E\left[X^k\right]$.

$$E\left[X^k\right] = \sum_{\substack{r=(w_r,c_r,q_r) \\ q_r=w_r}} \bar{\pi}_{A,r}^{\Psi} E\left[(X_r)^k\right] \tag{12}$$

In practice, we can evaluate the equation for the first two moments, $E[X]$ and $E\left[X^2\right]$, and fit a suitable distribution, e.g. in the class of gamma distributions to approximate the throughput time CDF, $F^{\Psi}(t)$. Eventually, cost of earliness and tardiness can be found using equations labeled by (7).

3.4. Results

We try to achieve two goals with our numerical experiments. First, we would like to show in which situations workload-dependent capacity planning is worth using. We study the effect of setting high/low switching cost coefficient as well as the different settings of lead time, and arrival rates for a fixed service rate. Second, we would like to characterize the workload-dependent policies as compared to the fixed capacity policies, the policies that can use one capacity level only. We illustrate the practical use of workload-dependent policies in the end of the section.

Our model has its limitations. We find the optimal policy via enumeration. Therefore the number of policies is an important factor of the computation's duration. Using our exact, throughput time distribution evaluation approach is not possible in acceptable time for systems that can either use many capacity levels, or handle high number of orders economically. In these cases the number

of workload-dependent policies and their state space increases to large values, so we need to use the approximation approach (e.g. when $U_{\min} = 0$, $C_{\max} = 6$ and $W_{\max} = 10$, the number of policies is 16844, the largest state space by the due-date performance evaluation has 203 states). As an initial setting, we take $U_{\min} = 0$, $C_{\max} = 3$ and $W_{\max} = 6$ (number of policies is 288, maximum number of states is 57) for which we use the throughput time distribution evaluation approach. We study large scale settings in subsection 3.4.3, and 3.4.5, where we use the moment approximation approach.

We assume a service rate of $\mu = 0.04$/hour in all of experiments, and vary the interarrival rate, λ, from 0.01 to 0.12/hour with step size of 0.01/hour. We study lead time (L) values from the interval 0 to 180 hours with step size of 10 hours. In our cost coefficient test-bed the capacity cost coefficients have a pointed role. Namely, we fix the permanent capacity cost coefficient to 100€/hour, to normalize the test-bed. Our basis values for the contingent capacity productivity rate, γ, is 0.9 and for the contingent capacity cost coefficient is 110€/hour. We observe the model's behavior for different switching cost coefficients, interarrival rates, and lead time settings. For each of the remaining cost parameters we examined three values. To the lost sales coefficient, c_l the values, 3000, 4000, and 5000€/occasion are assigned. We do not distinguish the earliness and the work-in-process holding unit costs, and calculate according to formula (8). For $c_e = c_w$, we take the values 1, 2, and 5€/job·hour. Finally, c_t has values 10, 25, and 100€/job·hour, respectively.

3.4.1 Value of workload-dependent capacity control

We investigate the value of workload-dependency in capacity control in different environments. We compare workload-dependent capacity policies with fixed capacity policies. The (relative) value of workload-dependent capacity flexibility, $VFC\%$, we define as the difference between the expected total cost of the optimal fixed and the optimal workload-dependent capacity policy, which is divided by the expected total cost of the optimal fixed policy, and is expressed in percentages. For example, a $VFC\%$ value of 5 means that given a firm that uses the optimal fixed capacity policy, it can reduce its costs by 5% by introducing the optimal workload-dependent policy. We conduct numerical experiments to circumscribe the cases where high $VFC\%$ values are to be expected.

In Figure 3.2, we depict contour lines of $VFC\%$ value functions for c_s values 1000 and 3000€/occasion.[1] Contour lines are drawn at the $VFC\%$ values 2, 8, and 15. As an example for interpreting Figure 3.2, we take the $VFC\%$

[1]This type of figure is called contour map, first used in cartography. It shows level sets

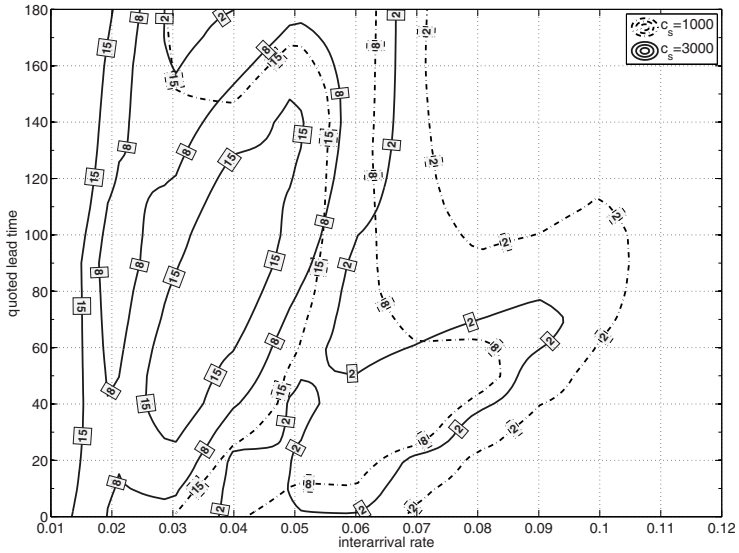

Figure 3.2: $VFC\%$ values for $U_{\min} = 0$, $C_{\max} = 3$, $W_{\max} = 6$, and $\mu = 0.04$

function for $c_s = 3000$€/job·hour, which correspond to the solid contour lines. E.g. we look at the $VFC\%$ function values on the line, $L = 40$: for $\lambda = 0.05$ or $\lambda = 0.09 \ldots 0.12$ the value is below 2, for $\lambda = 0.06$, 0.07 or 0.08 the value is between 2 and 8, for $\lambda = 0.02$, or 0.04 the value is between 8 and 15, for $\lambda = 0.01$ or 0.03, the value is above 15. The cost coefficients, c_l, c_e, and c_t take the values 5000€/occasion, 1€/job·hour, and 100€/job·hour, respectively. In what follows, we discuss the major properties of the $VFC\%$ value that can be observed in this figure. The $3^3 = 27$ combinations of the cost coefficients c_l, c_e, and c_t that we checked consistently support these properties and its reasonings, however they do not guarantee the properties to hold for different cost coefficient settings. An exception is the first property, which we present as a lemma with the sketch of the proof.

Lemma 1. When the switching cost coefficient decreases, the $VFC\%$ value increases

Proof. Since fixed capacity policies use a constant capacity level, they do not incur switching cost as opposed to the rest of the policies. While an increasing switching cost coefficient entail increasing total costs among the workload-dependent policies, the optimal fixed policy and its cost remain the same.

of a function, which has two arguments. Here, we plot two $VFC\%$ functions for c_s=1000 and 2000, respectively, with the arguments, interarrival rate and lead time

Consequently, the lower the switching cost coefficient, the higher the $VFC\%$ value.

Property 1. When the quoted lead time increases, the $VFC\%$ value decreases

Long quoted lead times, or equivalently, loose due dates penalize less the long waiting times. Long waiting times have a utilization smoothing effect similar to workload-dependent capacity flexibility. This means that when lead time (L) values are high enough, the $VFC\%$ values are low.

Property 2. When the arrival rate increases, the $VFC\%$ value decreases

As interarrival rate (λ) increases, we use correspondingly more capacity. U also increases for the optimal policy. This way, there is less and less room for workload-dependent policies to use contingent capacity. This results in a decreasing trend in the $VFC\%$ value.

Property 3. The effect of capacity discreteness: cuts in the $VFC\%$ value

Independently of the switching cost coefficient, one can observe some regions, where the $VFC\%$ value drops. The reason is the discreteness of the capacity. It would make a difference if we could set an optimal fixed capacity level on a continuous basis. We plot the integer contours of the optimal continuous capacity level in Figure 3.3 keeping the contour lines of Figure 3.2. There are regions where the optimal fixed policies get close to the continuous fixed optimum, and regions where they are distant. Where the continuous fixed optimum is close to an integer, the (discrete) fixed capacity policies perform reasonably well, while where the continuous fixed optimum is far from an integer level, fixed capacity policies perform poor. As the workload-dependent policies are less affected by the discreteness of the capacity, the $VFC\%$ value decreases around the integer contours of the optimal continuous fixed capacity.

Property 4. Steep increase in the $VFC\%$ value for small interarrival rates for $U_{\min} = 0$

To the left from the continuous fixed capacity contour at the value of one in Figure 3.3, the $VFC\%$ value steeply increases when the interarrival rate gets smaller. Fixed policies cannot adapt to low interarrival rates, as the fixed policy with a fixed zero capacity level is highly uneconomical, whereas the workload-dependent policy can make use of the zero capacity level for low workload values. As a result, the $VFC\%$ value is high for low interarrival rates, when $U_{\min} = 0$.

Our further experiments showed that Property 4 does not generally hold. In particular, we need to differentiate between closable systems, which we define

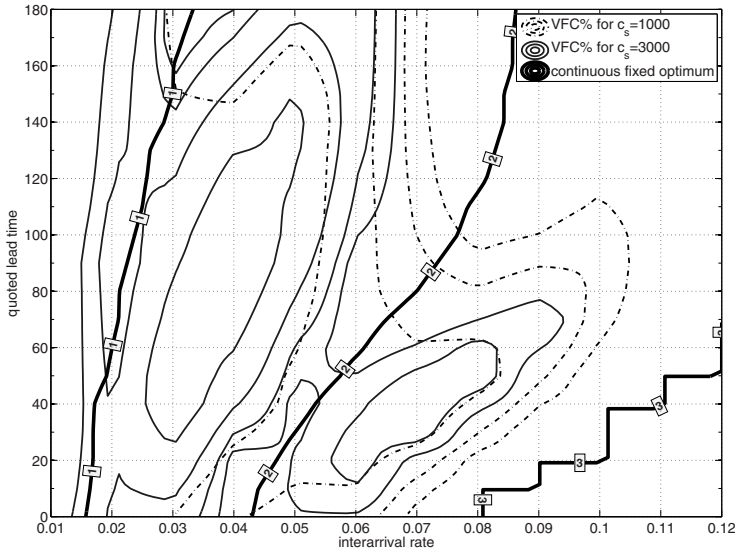

Figure 3.3: Continuous fixed optimum for $U_{\min} = 0$, $C_{\max} = 3$, $W_{\max} = 6$, and $\mu = 0.04$

as $U_{\min} = 0$, and non-closable systems having $U_{\min} > 0$. We adjust property 4 for the case of non-closable systems. Figure 3.4 depicts what happens if we change $U_{\min}$ in the parameter setting of Figure 3.2 from zero to one.

Property 4.* Low $VFC\%$ values for small interarrival rates and $U_{\min} > 0$

If we consider a non-closable system with $U_{\min} = 1$, then the constant one fixed capacity is optimal among both the fixed and the workload-dependent policies when the interarrival rate tend to zero, resulting in $VFC\% = 0$.

We also performed sensitivity analysis in $W_{\max}$ and $C_{\max}$. We observed an increase of the $VFC\%$ value in both cases. The reason may be that if we increase the $W_{\max}$ or $C_{\max}$ value by one, the number of fixed capacity policies remains the same or increases only by one, respectively, whereas the increase in the number of workload-dependent capacity policies is very large in general. The increase in $W_{\max}$ induces an overall increase in $VFC\%$ values, while the increase of $C_{\max}$ from $C_{\max}^{old} = 3$ to $C_{\max}^{new} = 4$ affects the graph of $VFC\%$ values only at λ values above around $\mu C_{\max}^{old}$.

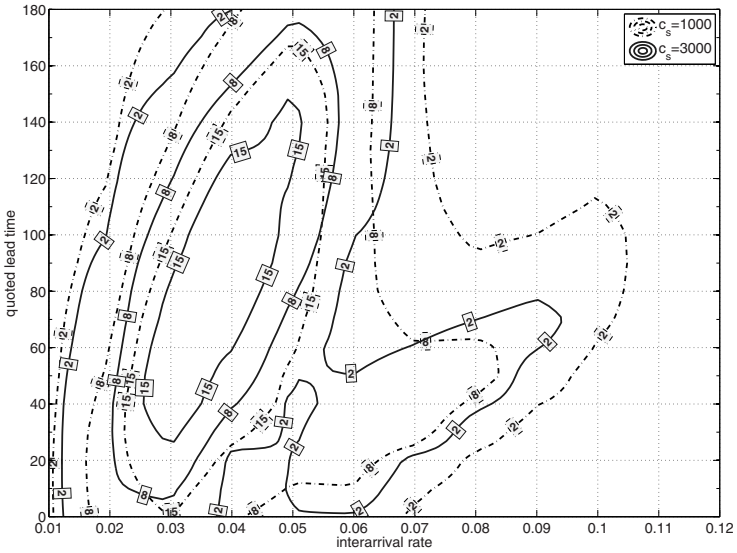

Figure 3.4: $VFC\%$ values for $U_{\min} = 1$, $C_{\max} = 3$, $W_{\max} = 6$, and $\mu = 0.04$

3.4.2 Sensitivity analysis on $VFC\%$ for estimated interarrival rates

In real-life situations the interarrival rate is not known with certainty, so its estimation is subject to error. Naturally, the value of capacity flexibility, $VFC\%$, is also not certain, but an estimate. It is therefore important to study how significant $VFC\%$ estimation errors the various levels of interarrival rate estimation error entail. In other words, it is important to perform a sensitivity analysis on the value of workload-dependent capacity control for estimated interarrival rates.

We study the effect of the relative interarrival rate estimation error, which we denote by $\Delta\lambda$, on the estimation of the value of capacity flexibility, $VFC\%$. The interarrival rate estimation can be expressed as

$$\hat{\lambda} = \lambda\,(1 + \Delta\lambda). \tag{13}$$

We perform sensitivity analysis on $VFC\%$ for interarrival rate estimation error. In our numerical experiments we calculate the $VFC\%$ values for over-/underestimated interarrival rates. The definition of $VFC\%$ remains the same as before. The only difference is that both the fixed and the workload-dependent costs are calculated under the misestimated interarrival rates.

Our numerical experiments use the already introduced $3^3 = 27$ combinations

of the cost coefficients c_l, c_e, and c_t, and we additionally take four values of the switching cost coefficient, c_s: 0, 1000, 2000, and 3000, and three quoted lead-time values, L: 30, 60, and 90, with $U_{\min} = 0$, $C_{\max} = 3$, and $W_{\max} = 6$. Table 3.2 shows the real value of introducing the workload-dependent capacity control, the average of the 3^3 cases, for the cost coefficient setting of Figure 3.2 ($c_l = 5000$ $c_e = 1$, and $c_t = 100$), for a quoted lead-time of $L = 60$, and $c_s = 3000$. The $\Delta\lambda = 0$ line corresponds to the estimated value of capacity flexibility already depicted in Figure 3.2, and the $\Delta\lambda \neq 0$ lines correspond to the real value of capacity flexibility for various nonzero relative error levels of the interarrival rate estimation. The $\Delta\lambda$ values of -0.3, -0.15, 0, 0.15, 0.3 were selected based on a simulation that measures demand forecast error for an interarrival time history of 50 jobs.

Although in the 27 cases we typically observed that an overestimated (underestimated) interarrival rate ($\Delta\lambda < 0$) implies an underestimated (overestimated) $VFC\%$ value (in line with Property 2), few exceptions occur for higher switching cost coefficients, which we demonstrate in Table 3.2 (see $\lambda = 0.02$, 0.03, 0.06, 0.07, and 0.08). These few exceptions correspond to λ values where the $\Delta\lambda = 0$ case has locally increasing $VFC\%$ values, because then Property 2 does not hold.

	interarrival rate (λ)									
$\Delta\lambda$	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
-0.30	44.40	0.87	5.96	30.90	21.04	-0.45	-0.65	-0.09	19.43	0.00
-0.15	34.53	2.21	12.06	23.17	11.58	-0.48	0.32	1.87	10.39	0.00
0.00	25.60	5.01	19.61	15.44	3.56	0.42	2.18	4.74	2.17	0.00
0.15	17.68	8.51	26.42	7.68	-1.70	1.70	4.88	7.85	-3.98	0.00
0.30	10.73	12.76	32.00	1.01	-4.54	3.56	7.61	10.39	-7.75	0.00

Table 3.2: Average $VFC\%$ values for $L = 60$, and $c_s = 3000$

The positive and negative absolute $VFC\%$ estimation errors are similar in absolute value. For $L = 60$ the positive and negative worst-case scenarios yield -15.01, and 20.09 absolute error, respectively. The absolute error does not decrease substantially with higher interarrival rates (for $\lambda = 0.9, \ldots 0.12$ the worst-case absolute errors are -10.72, and 16.77), which results in high worst-case relative errors -40.40%, and 111.19%. For lower interarrival rates (for $\lambda = 0.1, \ldots 0.8$) worst-case relative errors are much lower: -6.32%, and 16.29%. Worst-case absolute error of the $VFC\%$ estimation on the whole parameter-scenario setting was 20.30 for $L = 90$. Thus, the dangerous settings are the

high λ values, where underestimation of λ can lead to drastic overestimation of $VFC\%$, while only moderate estimation errors can be expected for lower λ values.

3.4.3 Characterization of workload-dependent policies for uncapacitated, large scale settings

An uncapacitated situation means that one can always hire the necessary capacity. That capacity is not limited, can be expressed either by $C_{\max} = W_{\max}$ or $C_{\max} = \infty$. We study the uncapacitated setting with $U_{\min} = 0$, $\lambda = 0.10$/hour, $\mu = 0.04$/hour, $\gamma = 0.9$, $L = 30$ hours, and cost coefficients, $(c_p, c_c, c_s, c_l, c_w, c_e, c_t) = (100, 110, 1000, 3000, 5, 5, 100)$, and observe changes of the optimal workload-dependent policy for an increasing $W_{\max}$ value. Our findings hold for all the other cost coefficient combinations from our test-bed.

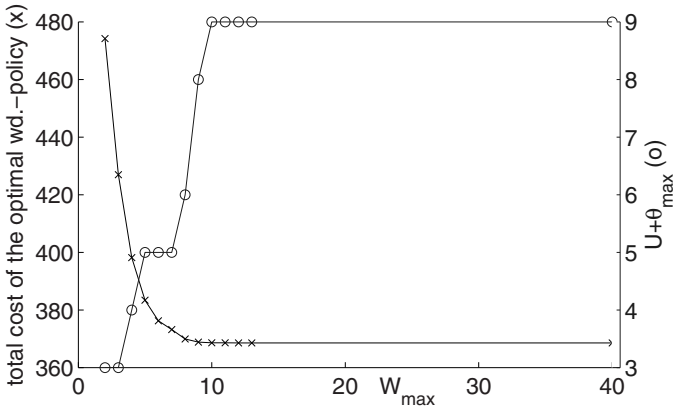

Figure 3.5: Total cost and maximum capacity of optimal workload-dependent policies for changing $W_{\max}$

We found that the capacity/lost sales cost component is monotone increasing/decreasing in $W_{\max}$ for the optimal policy. Moreover, the simultaneous stagnation of these two costs entails the stagnation of the total costs. Figure 3.5 shows that

- the total cost is decreasing in $W_{\max}$ up to a certain value (about $W_{\max} = 6$); thereafter it stagnates as the visiting probabilities of the high workload levels become very small, and it converges in a generally decreasing manner (not necessarily monotonically decreasing). As a consequence, it is a safe

alternative to allow arbitrarily high workload values in the uncapacitated case.

- the maximum capacity level ($U+\theta_{\max}$) and the number of used capacity levels ($\theta_{\max}$) show a monotone increasing, converging shape. U is unaffected; here it is always 2.

We also observed that optimal switching points of the lower levels tend to change less and less, when $W_{\max}$ increases. In Figure 3.6, the first six pairs of switching points can be seen, in which optimal policies did not differ for large enough $W_{\max}$ values ($13 \leq W_{\max} \leq 40$). We can see that up- (to the right) and down-switching points (to the left) limit the attainable states in a closely linear manner.

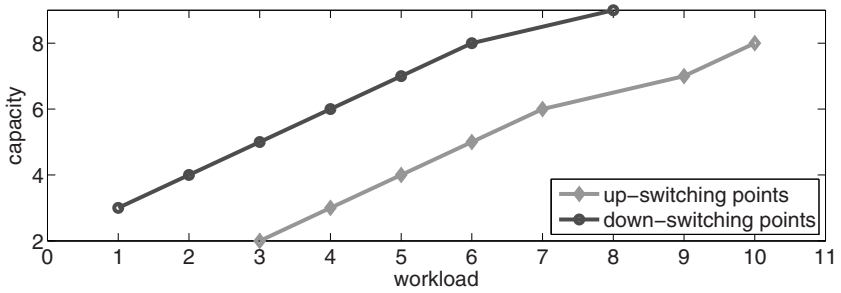

Figure 3.6: First six pair switching points of optimal policies for $13 \leq W_{\max} \leq 40$

3.4.4 Characteristics of workload-dependent policies vs. fixed policies

We try to answer the question of why workload-dependent policies outperform the fixed policies. Workload-dependent policies often compensate for switching costs by using less capacity, avoiding lost sales, and providing a throughput that gives less earliness-tardiness costs. The primary cost savings these policies achieve are normally in tardiness and lost sales cost. In the example shown in Table 3.3, the optimal workload-dependent policy beats the fixed optimum in all the cost components except $cost_{switching}$.

We emphasize that $cost_{capacity}$ and $cost_{earliness} + cost_{tardiness} + cost_{WIP}$ both decreased. Thus, the throughput time distribution adapts better to the lead time (L), which in turn results in a better due date performance. The reason

parameter	λ	μ	L	c_p	c_c	c_s	c_l	$c_e = c_w$	c_t
value	0.07	0.04	50	100	110	1000	3000	5	100

policy class	optimal Ψ	$VFC\%$	capacity	switch.	lost s.	WIP+early+tardy
workl.-dep.	(1,2,[3,1;4,2])	0	193.0	17.9	11.4	45.6
fixed	(2,0,[ ])	7.3	200	0	19.4	69.6
cont. fixed	2.18	6.1	218.2	0	14.7	52.4

Table 3.3: An example, where all cost components are decreased by workload-dependency, except for the switching cost.

for the better adaption is twofold. With workload-dependency the expected value of the throughput time gets closer to the actual lead time. In addition, the variance of throughput time decreases. Table 3.4 shows the expected value and standard deviation of throughput times ($E[X]$ and $Std[X]$) for the previous example, for which the corresponding Markov-chain, $MC^{(1,2,[3,1;4,2]),\lambda,\mu}$ was illustrated in Figure 3.1.

Policy	$E[X]$	$Std[X]$
$\Psi = (1, 2, [3, 1; 4, 2])$	37.5	21.5
$\Psi = (2, 0, [\,])$	39.0	30.3
continuous fixed 2.18	33.0	26.9

Table 3.4: Expected value and standard deviation of throughput time in the same setting.

3.4.5 An illustration to the use of workload-dependent policies in real-life

This subsection we dedicate to illustrating how our procedure can be applied for evaluation of real-life labor arrangements options. We compare two fast-response labor arrangements alternatives to using permanent capacity exclusively: the use of overtime, and the use of temporary labor.

A firm produces colored neon light figures using a quoted production lead time of 4 weeks (L). On average, the firm receives 3.8 orders weekly (λ). One worker can manufacture one neon light figure in a week (μ). In the present situation, all 4 regular workers work 8 hours each day. The manager would like to introduce a fast-response flexible labor arrangement in order to cope

better with the varying workload. He compares two alternatives, well-known in practice: the overtime and the on-call temporary labor arrangements.

Including the present situation, we define three workload-dependent policy classes corresponding to the alternatives: the fixed, the overtime, and the temporary labor policy classes. The classes we parameterize with the number of regular workers, an integer, which we denote by U, and the switching points, Θ. The capacity levels of the classes are U, $\left(U, \frac{9U}{8}\right)$,$(U, U+1, \ldots, U+5)$, respectively. Overtime costs 1.5 times the regular capacity cost, but it does not incur switching cost, whereas temporary labor costs 1.25 times the regular capacity and it does incur switching cost. The service rate at workload w, and capacity level c is $\mu_{w,c} = c\mu$, as before. New orders are rejected if the firm already has 40 jobs ($W_{\max}$).

Table 3.5 summarizes the results for the cost coefficients $(c_p, c_s, c_l, c_w, c_e, c_t) = (50€/\text{hour}, 100€/\text{occasion}, 3000€/\text{occasion}, 5e/\text{week}, 5e/\text{week}, 100€/\text{job·hour})$. We conclude that the temporary labor arrangement is the manager's best alternative in this particular case.

fast-response flexible labor arrangement	opt. #reg. workers (U)	optimal switching points (Θ)	total costs
fixed	4	( )	8586.7
overtime	4	(21 22)	8506.8
temporary labor	3	$\begin{pmatrix} 14 & 16 & 18 & 21 & 23 \\ 10 & 11 & 13 & 14 & 15 \end{pmatrix}^T$	8275.6

Table 3.5: Expected costs of the different labor arrangements.

3.5. Conclusions and future research

Under the assumption that fast-response capacity changes are possible, we defined a set of workload-dependent capacity planning policies. We introduced a general capacity control model to find the optimal workload-dependent policy with respect to permanent/contingent capacity, capacity switching, lost sales, work-in-process holding, and earliness/tardiness costs for a fixed quoted lead time. Providing formulas for the cost components one by one via stationary analysis, we evaluated this model, and gave insight into the value of the workload-dependent capacity management policies.

We measured the cost savings by using the workload-dependent policies as

compared to the fixed policies, which can use only one capacity level. Large switching cost coefficients, high demand rates, and long quoted lead times are detrimental, while high workload limits ($W_{\max}$) are beneficial for the savings. Capacity discreteness can strongly affect the cost savings, as workload-dependent policies can counteract non-integer capacity needs, while fixed policies cannot. We showed that when the necessary capacity is between level 0 and level 1, we need to differentiate two cases: if the zero permanent capacity level is feasible (closeable system) then the savings are particularly high, whereas if the zero level is infeasible the savings are low.

In the uncapacitated case, we observed that using a sufficiently high order-acceptance rate, or equivalently a high workload limit ($W_{\max}$), is a safe choice when selecting the workload-dependent strategy. We found that for high workload limits, the optimal capacity up- and down-switching points tend to change less and less, and appear to form two lines. This observation may facilitate future research on the policy class comprising this linear type of policies.

Finally, we revealed that compared to the optimal fixed capacity policies, the optimal workload-dependent capacity planning policies can achieve a better due-date performance. In particular cases they can also spare capacity, and decrease lost sales probability at the same time (as shown in Table 3.3).

Appendix

In this appendix, we apply uniformization for $EMC^{\Psi,\gamma,\lambda,\mu}$ in order to obtain the probability that starting from state r at time 0 we will be state s in t time.

We can define both $\bar{A}^{\Psi}$ extended arrival matrix, and $\bar{D}^{\Psi}$ extended departure matrix for $EMC^{\Psi,\gamma,\lambda,\mu}$ based on A^{Ψ}, and D^{Ψ}, as follows.

$$\bar{A}^{\Psi}_{(w_1,c_1,q_1),(w_2,c_2,q_2)} = \begin{cases} 1 & \text{, if } A^{\Psi}_{(w_1,c_1),(w_2,c_2)} = 1, \text{ and } q_1 = q_2 \\ 0 & \text{, otherwise} \end{cases}$$

$$\bar{D}^{\Psi}_{(w_1,c_1,q_1),(w_2,c_2,q_2)} = \begin{cases} 1 & \text{, if } D^{\Psi}_{(w_1,c_1),(w_2,c_2)} = 1, \text{ and } q_1 - 1 = q_2 \\ 0 & \text{, otherwise} \end{cases} \tag{14}$$

for all (w_1, c_1, q_1), and (w_2, c_2, q_2). The transition rate diagram of $EMC^{\Psi,1,\lambda,\mu}$ for the policy $\Psi = (1, 2, [3, 1; 4, 2])$ can be seen in Figure 3.7.

The extended transition rate matrix, $\bar{Q}^{\Psi}$, can be given in two steps like before.

$$\bar{Q}^{\Psi}_{(w_1,c_1,q_1),(w_2,c_2,q_2)} = \lambda \bar{A}^{\Psi}_{(w_1,c_1,q_1),(w_2,c_2,q_2)} + \mu_{c_1} \bar{D}^{\Psi}_{(w_1,c_1,q_1),(w_2,c_2,q_2)} \tag{15}$$

if $(w_1, c_1, q_1) \neq (w_2, c_2, q_2)$.

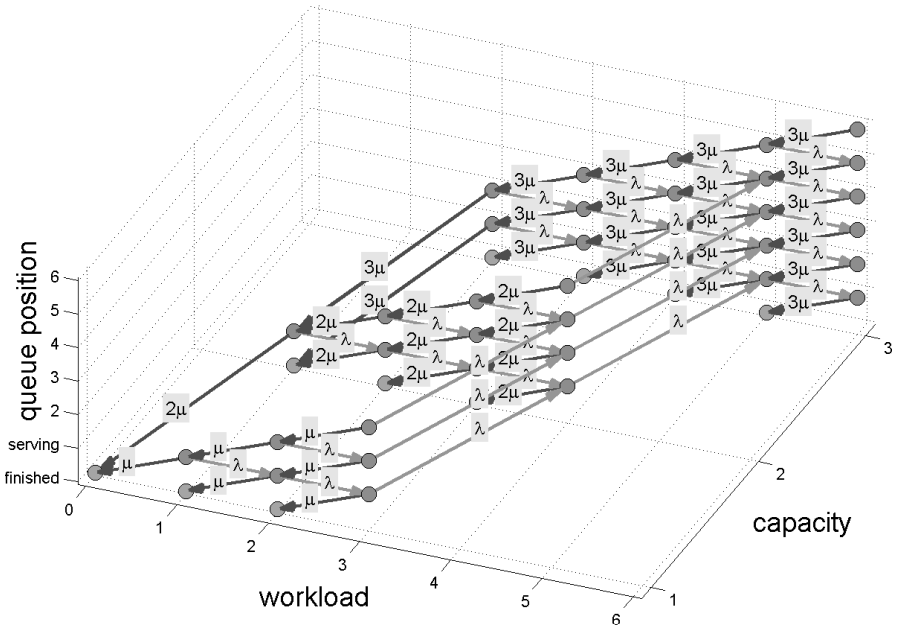

Figure 3.7: $EMC^{\Psi,\gamma,\lambda,\mu}$, for $\gamma = 1$ and $\Psi = (1, 2, [3, 1; 4, 2])$

Next, we define the vector of state leaving rates, $\bar{v}^{\Psi}$.

$$\bar{v}_s^{\Psi} = \sum_{s \neq r} \bar{Q}_{s,r}^{\Psi} \tag{16}$$

Now, the remaining diagonal elements of $\bar{Q}^{\Psi}$ can be defined as $\bar{Q}_{s,s}^{\Psi} = -\bar{v}_s^{\Psi}$.

From $\bar{Q}^{\Psi}$ we can derive the extended transition probability matrix, $\bar{P}^{\Psi}$.

$$\bar{P}_{r,s}^{\Psi} = \begin{cases} \frac{\bar{Q}_{r,s}^{\Psi}}{\bar{v}_r^{\Psi}} & \text{, if } r \neq s \\ 0 & \text{, if } r = s \end{cases} \tag{17}$$

We use the uniformization method; we add loop arcs to most of the states in the extended Markov chain ($EMC^{\Psi,\gamma,\lambda,\mu}$) so that the distribution of the time between two subsequent events (arrival or departure or loop) becomes identical. The uniformized event occurrence rate is $\bar{v}^{U\Psi} = \max_s \bar{v}_s^{\Psi}$. The uniformized extended transition probability matrix, elementwise, is

$$\bar{P}_{r,s}^{U\Psi} = \begin{cases} \frac{\bar{v}_r^{\Psi}}{\bar{v}^{U\Psi}} \bar{P}_{r,s}^{\Psi} & \text{, if } r \neq s \\ 1 - \frac{\bar{v}_r^{\Psi}}{\bar{v}^{U\Psi}} & \text{, if } r = s \end{cases} \tag{18}$$

Using the matrix, $\bar{P}^{U\Psi}$, we can determine the probability of getting from state

r to s in time t.

$$\bar{P}^{\Psi}_{r,s}(t) = \sum_{n=0}^{\infty} \left(\left(\bar{P}^{U\Psi} \right)^n \right)_{r,s} e^{-\bar{v}^{U\Psi} t} \frac{\left(\bar{v}^{U\Psi} t \right)^n}{n!} \tag{19}$$

Chapter 4

Permanent-Contingent Budgeting in Services

We model budgeted services with fixed and variable expenditures and illustrate this general setting with the situation, where the fixed and variable expenditures correspond to permanent and contingent capacity costs, respectively. No backordering of tasks is allowed, instead, capacity shortages are penalized in each period. Dedicated to optimizing the permanent and contingent capacity levels, six models are developed, distinguishing restricting and guiding budget environments, as well as linear, quadratic and relative quadratic capacity shortage penalty cost functions. We provide some analytical results on the linear shortage cost case, and perform numerical experiments for all the six models, showing among others the optimal budget spending patterns.

4.1. Introduction

Lately, budgeting and budget allocation received particular research attention. This attention may be due to the fact that there are still many fundamental questions remaining in this field and that budgets are still the most often used means of financial control, primarily in large organizations. Hansen and van der Stede (2004) term budgeting as being an important control in almost all organizations. Their survey, which study 37 manufacturing and 20 service firms in 2001, shows that an annual budget was used in 77% of the (surveyed) US companies. Although the budgeting practice has been already established for many decades, there is still no clear understanding, how one should operate

a given budget, throughout the budgeted period. In this chapter, we study the budget spending problem with fixed and variable costs. In the beginning of the budgeted period, a part of the budget is allocated to cover permanently occurring, fixed expenses, which can decrease the variable costs originating from contingent expenditures.

Managers may face budget, demand and cost uncertainty while they allocate the budget over time. These uncertainties can make the allocation task rather difficult. In the following paragraphs, we discuss two studies containing data on monthly spending patterns.

Balakrishnan *et al.* (2007) is a remarkable empirical study, which indicates a strong support of the so-called saving-dissaving behavior under budget lapsing (when unspent funds are lost). According to the saving-dissaving model, managers build reserves in the beginning of the year against budget uncertainty and/or demand uncertainty. These reserves are consumed later on, especially in the end of the year, where it results in a peak spending.

Zimmerman (1976) suggests that the high end of year spending is a rational response to the budget uncertainty, experienced throughout the year. Assuming existence of cardinal utility, he summarizes analytical results on the spending behavior. Additionally, he performs additional regression calculations to capture the budgeting uncertainty in an empirical setting. Based on the analytical results and the regression, his conclusion is that there is a rational basis for the practice of deferring expenditures until uncertainty is resolved.

We remark that the data collection in these two studies took place in quite different environments. The more recent study, Balakrishnan *et al.* (2007), is in a hospital setting, where the demand uncertainty drives the spending behavior. By the older study, Zimmerman (1976), presents data on the budgeting pattern of a single research institute, where the budget uncertainty leads the longitudinal budget allocation decisions. We also note that cost-uncertainty plays in neither of the studies a significant role. Naturally, additional empirical studies could strengthen the results of Zimmerman (1976). Nevertheless the empirical observations coincide with the analytical claims, so we may conclude that we gained understanding of what spending pattern the budget uncertainty can entail. However, it is not the case as far as demand uncertainty is concerned. Although Balakrishnan *et al.* (2007) provides an empirical evidence of the saving-dissaving spending behavior in response to demand uncertainty, it is still an open question if this spending behavior was rational. The goal of this chapter is to answer this question, providing either justification or unjustification in analytical and numerical ways.

To help understanding, we illustrate the budget spending problem with dis-

tinct fixed and variable costs via a longitudinal capacity planning situation. We interpret the fixed and variable costs as permanent and contingent capacity costs, respectively. In this setting, we use disutility instead of utility, which takes the form of a cardinal capacity shortage penalty. We note that another setting could be the use of an advertising budget, which is a lot more complicated because of the necessary representation of the competing companies (see Zufryden (1975)).

Our main contribution with this chapter is that we give a plausible explanation for the saving-dissaving spending behavior under the permanent-contingent capacity concept. We support this explanation by a broad set of numerical experiments. In the end of the chapter we suggest another, alternative explanation, additionally.

In the capacity planning literature, there is a good basis for studying permanent-contingent budgeting under demand uncertainty. On the one hand, capacity planning for services with permanent and contingent capacities has been already studied under stochastic demand (see Pinker (1996)), on the other hand budgeting has been also addressed in connection with capacity planning (see Trivedi (1981)). The application of the cardinal utility concept can be also found in the capacity planning literature. In particular, Warner and Prawda (1972) develops the concept of capacity shortage penalty cost. The shortage penalty cost is a convexly decreasing function of the capacity; it hits zero at some point called the required capacity, and it remains zero afterwards. We mention that an alternative to the penalty cost concept is to target a prespecified service level (see e.g. Abernathy *et al.* (1973)) throughout the year.

We can draw parallels between this chapter and some specific two-stage supply chain models. We illustrate it with the following perishable (or alternatively a short life-cycle) production example, which is similar to the two-stage case in van Houtum *et al.* (2007) except for the perishability. The raw material arrive in long, monthly periods to stock. The production stage consumes the raw material on stock day by day and delivers perishable goods to satisfy demand. There is also an option to contract a supplier for delivering fixed quantities of the perishable product, daily. Here, the permanent capacity corresponds with the external supplier of perishable goods, the contingent capacity corresponds with the production quantity, the budget corresponds with the monthly stock of raw material, the capacity shortage penalty cost corresponds with a linear lost sales penalty.

4.2. Models

We study various problem settings of determining the permanent capacity level and contingent capacity levels under stochastic demand. We develop several models in the following sections. The next subsection introduces the common points of these models, and we discuss their differences afterwards.

4.2.1 Common modeling aspects

We define permanent-contingent capacity planning problems under stochastic demand, in which we minimize the annual capacity shortage costs over a finite horizon of periods taking into account the presence of a budget. The horizon length is H, and the periods are indexed by integers, 1, 2, ..., $H-1$, H. We assume that the demand in each period is identically distributed with some known, arbitrary demand distribution, and that each period's demand is independent from demand of all the other periods.

Our capacity planning model is a sequence of budget-allocation decisions. The budget is allocated to cover expenses of the permanent and contingent capacity. In the beginning of the horizon, we take the the tactical level decision on the permanent capacity level (U), which will be present throughout the whole year. At the operational level, calling some contingent capacity (θ_t) may take place in each period of the year. In all these periods, we assume the following order of events,

- Demand of the coming period, D_t, is revealed, and the remaining budget, b_t, is observed.
- Contingent capacity may be called, which amounts to θ_t.
- Capacity shortage costs are registered.

Thus, demand of the actual period is known in advance, before the contingent capacity decision, while only the distribution of demand is known for later periods. In line with the capacity shortage penalty cost concept in Warner and Prawda (1972), surplus capacity in any periods is assumed to be lost.

4.2.2 Overview of the models

We present a categorization of the models we address. Our models are developed systematically, so that we can study the role of the budget and the capacity shortage cost function in depth.

In our models, we represent the role of the budget in two ways.

- Either capacity shortages are minimized for a given budget,
- or capacity shortages and budget deviations are jointly minimized.

The first budget model type corresponds to the situations alike the army hospital in Balakrishnan *et al.* (2007), and the second type can be regarded as the generalization of the first to include situations, where budget overspendings are common. We point out that in the second model type such coefficients are to be given to capacity shortages and budget deviations that make them comparable, and establish their proper balance. The goal programming approach in Trivedi (1981) addressed the tuning of these coefficients. Our second type models follow the budget deviation penalty assumption in Trivedi (1981): the end-of-year budget surplus is rewarded linearly, having the coefficient, c_b^+, while budget deficit is penalized linearly with another, larger coefficient, c_b^-.

The capacity shortage cost function can take the following forms in our models,

- halfway linear
- or halfway quadratic
- or halfway relative quadratic.

The prefix 'halfway' we added to all the function types, referring to that all the function types are zero when the capacity exceeds the demand, since surplus capacity is assumed to be lost. We consider the halfway linear shortage cost functions because of their simplicity, the halfway quadratic one because it is a traditional assumption (see Warner and Prawda (1972)), and the halfway relative quadratic one because we suspect it being more realistic, than the halfway quadratic function.

Table 4.1 enlists the sections showing their correspondence with the models.

4.3. Linear capacity shortage costs and restricting budget

This section is devoted to the analytical results we derive under the assumption that the capacity shortage cost is a linear function, and that the budget may not be overspent. We first develop a simple model to provide analytical formulas, and then turn to a numerical illustration emphasizing the efficiency of the cost function approximation we propose.

Models	Budget model type	Capacity shortage cost function type
Section 3	restricting budget	halfway linear
Section 4	penalized budget deviations	halfway linear
Section 5	restricting budget, penalized budget deviations	halfway quadratic
Section 6	restricting budget, penalized budget deviations	halfway relative quadratic

Table 4.1: The models in the following sections

4.3.1 Model development and analysis

Let c_s be the penalty cost associated with one unit of capacity shortage (or equivalently, lost sale), and let $(a)^+$ be defined as $(a)^+ := \max\{0, a\}$. The total annual costs are then given by

$$c_s \sum_{t=1}^{H} (D_t - U - \theta_t)^+ \tag{1}$$

with an expected value of

$$c_s \sum_{t=1}^{H} \int_{U+\theta_t}^{\infty} (x - U - \theta_t)^+ \, dF(x) \tag{2}$$

where $F(x)$ denotes the demand probability distribution function per period, and θ_t is determined dynamically, based on the budget available in the beginning of period t, b_t. Because of the linear shortage costs, it is never advantageous to accept some shortage in a period t to avoid shortages in later periods, as the future savings can never exceed the current extra costs. This feature makes the decision about contingent capacity simple.

As long as the remaining budget is large enough we use contingent capacity to exactly meet the actual demand, otherwise we use all remaining budget for the contingent capacity. Denoting the unit contingent capacity cost by c_c, the contingent capacity used in period t can be expressed as

$$\theta_t = \begin{cases} (D_t - U)^+ & \text{if } b_t \geq c_c(D_t - U)^+ \\ b_t / c_c & \text{otherwise} \end{cases} \tag{3}$$

where D_t can be interpreted both as a revealed demand value and as a yet unrevealed random demand.

This means that the only decision left is the determination of the permanent capacity level, U. Suppose that the cost of permanent capacity per unit per

period is c_p. When we have a budget of B, the total contingent capacity is a function of U. This function, $T(U)$, has the form

$$T(U) = \frac{B - c_p H U}{c_c} \geq \sum_{t=1}^{H} \theta_t. \tag{4}$$

The annual shortage cost, which is a random variable, can be expressed for a given random demand stream D_1, D_2, ... D_H, with the difference between the total excess demand that the permanent capacity could not satisfy and the total contingent capacity used, as follows.

$$C_s(U) = c_s \left(\left(\sum_{t=1}^{H} (D_t - U)^+ \right) - T(U) \right)^+ \tag{5}$$

Our goal is to find the optimal permanent capacity level, U, which minimizes the expected annual shortage costs. Since it is difficult to obtain an analytical expression of the optimal U, we provide an approximation of expected annual shortage costs. In particular, we bring the expected value operator inside of the outer positive part, and use that the demand of the periods are identically distributed. Denoting the expected excess demand that the permanent capacity could not satisfy by $R(U) = E\left[(D_t - U)^+\right]$ for all t, we can develop the approximation

$$E[C_s(U)] \approx c_s \left(H \cdot R(U) - T(U)\right)^+ \tag{6}$$

Minimizing the approximation shown in the right-hand side of (6) is considerably easier. Its minimization is equivalent with the minimization of the term inside the positive part operator. Substituting back with $T(U)$, we obtain a newsvendor problem, having the following straightforward approximation of the optimal permanent capacity level,

$$U^* \approx \underset{0 \leq U}{\arg\min} \left\{ H \cdot R(U) - \frac{B - HPc_p}{c_c} \right\} = F^{-1}\left(\frac{c_c - c_p}{c_c}\right) \tag{7}$$

In subsection 4.3.3, we illustrate the accuracy of this approximation via numerical experiments.

4.3.2 A special case: gamma-distributed demand

We can proceed from equation (5) in an alternative, possibly more precise way, if we assume that the demand per period is given by a gamma distribution. In particular, we can determine the first and second moments of the total excess demand that the permanent capacity could not satisfy, $\sum_{t=1}^{H} (D_t - U)^+$. As the summed terms are independent, and identically distributed, we can gain the moments of the sum as

$$E\left[\sum_{t=1}^{H}(D_t-U)^+\right]=HE\left[(D_t-U)^+\right]$$

and (8)

$$Var\left[\left(\sum_{t=1}^{H}(D_t-U)^+\right)\right]=H\ Var\left[\left((D_t-U)^+\right)\right]$$

Now assume the special case of the gamma-distributed demand, so that we have

$$D_t \text{ is of } \Gamma_{\alpha,\lambda} \text{ for all } t=1,2,\ldots,H \tag{9}$$

with the distribution function

$$F_D(x)=\Pr\left[D_t\le x\right]=\int_0^x \frac{\lambda e^{-\lambda y}(\lambda y)^{\alpha-1}}{\Gamma(\alpha)}dy \tag{10}$$

Then the first two moments of the expected excess demand that the permanent capacity could not satisfy can be obtained by

$$E\left[(D_t-U)^+\right]=\int_U^\infty (x-U)\,dF_D(x)=\int_U^\infty \frac{(x-U)\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}dx$$
$$=\frac{\alpha}{\lambda}\left(1-\Gamma_{\alpha+1,\lambda}(U)\right)-U\left(1-\Gamma_{\alpha,\lambda}(U)\right)$$

and (11)

$$E\left[\left((D_t-U)^+\right)^2\right]=\int_U^\infty (x-U)^2\,dF_D(x)=\int_U^\infty \frac{(x-U)^2\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}dx$$
$$=\frac{\alpha(\alpha+1)}{\lambda^2}\left(1-\Gamma_{\alpha+2,\lambda}(U)\right)-2\frac{\alpha U}{\lambda}\left(1-\Gamma_{\alpha+1,\lambda}(U)\right)+U^2\left(1-\Gamma_{\alpha,\lambda}(U)\right)$$

Once the moments are obtained, a suitable distribution function can be fit to approximate $\sum_{t=1}^{H}(D_t-U)^+$, which we call G. Finally, an approximation of the expected shortage cost can be calculated as

$$E\left[C_s(U)\right]\approx c_s\int_{T(U)}^\infty (z-T(U))^+\,dG(z), \tag{12}$$

corresponding to equation (5).

4.3.3 Numerical illustration

We consider the following example. The budget B is set to 3250 over 50 periods. The capacity shortage cost and the cost for the permanent capacity are both equal to 1 and the cost for one unit of contingent capacity is 2.5. Figure 4.1 displays the average results of 50000 simulations, for the exact and approximate shortage costs, with a value of U ranging from 30 to 70, assuming a normal and a gamma distribution for the demand, with average of 50 units

and a standard deviation of 20 units. With a normally distributed demand, the best permanent capacity level is 55 units for both methods (exact and approximation), although there is some difference between the two total costs for medium and high permanent capacity level. This illustrates the usefulness of the approximation.

When demand is gamma-distributed, the shortage quantity can be determined by using (12) for the appropriate permanent capacity levels. As the 'exact' results of the simulation are always within 1 percent of the approximation, there is only one line representing both costs in the figure. The optimal permanent capacity is 53, slightly different from the 55 for the normal distribution. Minimizing (12) leads to a permanent capacity of 53.12, whereas (7) leads to a permanent capacity of 52.44.

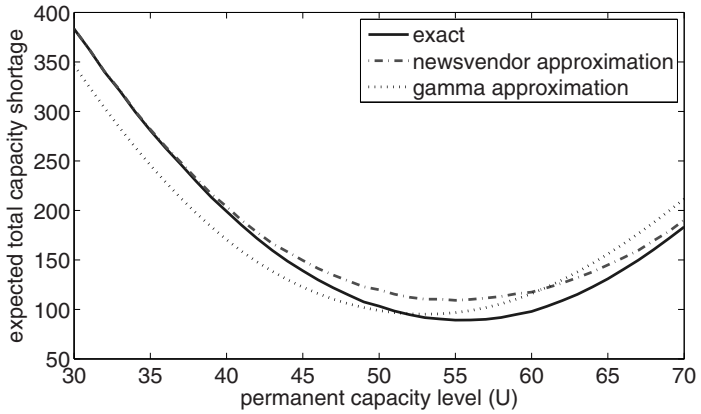

Figure 4.1: Expected capacity shortage cost as a function of the permanent capacity level

This illustration supports that it is sufficient to optimize the capacity use per period with formula (7). Since this approximation formula is budget-independent and it works well, we can conclude that the budget has only little influence on the optimal permanent capacity level.

4.4. Linear capacity shortage and budget deviation costs

The situation in which we include costs for budget deviations is slightly more complicated, but also more realistic. The objective is to allocate the budget dynamically so as to find a good balance between capacity shortages and budget deficit for a given annual budget.

4.4.1 Model development and analysis

We consider linear costs for budget deficits, and linear reward for budget surpluses. The budget deficit is assumed to be at least as much penalized as the budget surplus is rewarded. We express the budget deficit cost ($C_b^-(U)$) and the budget surplus reward ($-C_b^+(U)$) in two ways: first with the difference of the budget and the total capacity costs, second with the difference of the demand and the total contingent capacity used.

$$C_b^-(U) = c_b^- \left(c_p HU + c_c \sum_{t=1}^{H} \theta_t - B \right)^+$$

and (13)

$$C_b^+(U) = -c_b^+ \left(B - c_p HU - c_c \sum_{t=1}^{H} \theta_t \right)^+$$

where c_b^- is the unit cost of the budget shortage and c_b^+ is the unit cost of the budget excess.

We minimize the expected value of the total cost, which is the sum of the capacity shortage, the budget deficit and the budget surplus costs.

$$\min_{\substack{0 \le U \\ TC(U)=C_s(U)+C_b^+(U)+C_b^-(U)}} E\left[TC(U)\right], \tag{14}$$

for which we distinguish four cases, as follows.

Case 1. When $c_b^+ c_p < c_b^+ c_c \le c_b^- c_c < c_s$, all demand is satisfied (at the expense of incurring budget penalties), which implies $C_s(U) = 0$. Consequently, the total costs have only two components, which we can express as

$$\begin{aligned} TC(U) &= C_b^-(U) + C_b^+(U) \\ &= c_b^- c_c \left(\sum_{t=1}^{H} (D_t - U)^+ - T(U) \right)^+ - c_b^+ c_c \left(T(U) - \sum_{t=1}^{H} (D_t - U)^+ \right)^+ . \end{aligned} \tag{15}$$

We can use the same recipe as in (6) to obtain the following simple approxi-

mation,

$$\begin{array}{l} E\left[TC\left(U\right)\right] \approx c_b^- c_c \left(HR\left(U\right) - T\left(U\right)\right)^+ - c_b^+ c_c \left(T\left(U\right) - HR\left(U\right)\right)^+ \\ = c_b^- \left(c_p U + c_c R\left(U\right) - \frac{B}{H}\right)^+ - c_b^+ \left(\frac{B}{H} - c_p U - c_c R\left(U\right)\right)^+ \end{array} . \quad (16)$$

The approximation suggests that the optimal permanent capacity level, $U^* \geq 0$, is close to the maximizer of

$$\frac{B}{H} - c_p U - c_c R\left(U\right). \quad (17)$$

This maximization reduces to the same newsvendor problem, as in (7), yielding the same optimal permanent capacity level,

$$U^* \approx \underset{0 \leq U}{\arg\min} \left\{c_c R\left(U\right) + c_p U\right\} = F^{-1}\left(\frac{c_c - c_p}{c_c}\right). \quad (18)$$

Case 2. When $c_b^+ c_p < c_b^+ c_c < c_s < c_b^- c_c$, demand is satisfied only if there is budget left (to avoid budget deficit), which yields $C_b^-\left(U\right) = 0$. In this case, the total costs have again two components, namely,

$$\begin{array}{l} TC\left(U\right) = C_s\left(U\right) + C_b^+\left(U\right) \\ = c_s \left(\sum\limits_{t=1}^{H} \left(D_t - U\right)^+ - T\left(U\right)\right)^+ - c_b^+ c_c \left(T\left(U\right) - \sum\limits_{t=1}^{H} \left(D_t - U\right)^+\right)^+ \end{array} . \quad (19)$$

We can proceed similarly to Case 1 to obtain the approximation,

$$\begin{array}{l} E\left[TC\left(U\right)\right] \approx c_s \left(HR\left(U\right) - T\left(U\right)\right)^+ - c_b^+ c_c \left(T\left(U\right) - HR\left(U\right)\right)^+ \\ = \frac{c_s}{c_c} \left(c_p U + c_c R\left(U\right) - \frac{B}{H}\right)^+ - c_b^+ \left(\frac{B}{H} - c_p U - c_c R\left(U\right)\right)^+ \end{array} , \quad (20)$$

which leads to the same observation, and approximation, as in Case 1, equation (18).

Case 3. When $c_b^+ c_p < c_s < c_b^+ c_c \leq c_b^- c_c$, budget is used for permanent capacity exclusively and exhaustively, which results in $T\left(U\right) = C_b^-\left(U\right) = 0$ and $C_b^+\left(U\right) = -c_b^+\left(B - c_p HU\right)^+$. The total costs can be written as

$$TC\left(U\right) = C_s\left(U\right) + C_b^+\left(U\right) = c_s \sum_{t=1}^{H} \left(D_t - U\right)^+ - c_b^+\left(B - c_p HU\right)^+. \quad (21)$$

In this case, the expected total cost formula can be formulated as

$$E\left[TC\left(U\right)\right] = c_s HR\left(U\right) - c_b^+\left(B - c_p HU\right)^+, \quad (22)$$

which is another newsvendor equation, yielding

$$U^* = F^{-1}\left(\frac{c_s - c_b^+ c_p}{c_s}\right) \leq \frac{B}{Hc_p}. \quad (23)$$

If the budget is sufficiently high, it will totally be spent, meaning $U^* = B/Hc_p$.

Case 4. In the extreme case, when $c_s < c_b^+ c_p$, we would not spend any budget at all.

4.4.2 Numerical illustration

We consider the experimental setting of the previous section, with gamma-distributed demand. Additionally, we take the budget deviation unit costs, $c_b^- = 2$ and $c_b^+ = 1$.

In Case 1, for a unit cost of capacity shortage, $c_s = 6$, the optimal permanent capacity level is 52.76, the suggested newsvendor approximation, equation (18), provides the estimation of 52.44, and the two-moment approximation, equation (12), gives an estimate of 53.12. Actually, the optimal permanent capacity does not deviate much from the one in the situation described in the previous section, as the budget surplus term, $C_b^+(U)$, is small.

In Case 2, taking a shortage cost coefficient, $c_s = 4$, the optimal permanent capacity level is 52.68, and the whole budget is used with probability $80-90\%$. Therefore, the outcome is hardly different from that of the previous section, showing an almost identical pattern for capacity shortages.

To illustrate Case 3, we considered a capacity shortage unit cost of $c_s = 1$. The optimal permanent capacity for the gamma-distributed demand is 47.36 with a yearly expected shortage quantity of 456.12 units.

4.5. Halfway quadratic capacity shortage costs

In this section, we address the model with the traditional halfway quadratic capacity shortages cost function (see Warner and Prawda (1972)) under strictly limiting budget as well as under penalized budget deviations. The shortages cost function has the form,

$$C_s(U) = c_s \sum_{t=1}^{H} \left((D_t - U - \theta_t)^+\right)^2. \tag{24}$$

While in the linear capacity shortages costs case, the determination of the permanent capacity level was the only difficult question, by the halfway quadratic costs the contingent capacity decisions are not straightforward. It is also different that by the halfway quadratic costs, it becomes particularly important to avoid large shortages in the last periods. First, we consider the restricting budget case and afterwards the penalized budget deviation case.

This situation we model as a dynamic program and evaluate it recursively via a backward induction algorithm. A limited state space is necessary to calculate the decisions. Hence we assume that we always use an integer amount of contingent capacity per period and that the demand per period follows some discrete distribution. We expect that this discretization has only a minor

influence on our analysis and results.

In what follows, we build the dynamic programming models for the restricting and guiding budget situations. First, we define expected capacity shortage (and budget deviation) cost-to-go functions recursively, starting from the last period. Then, we discuss the budget situation (restricting/guiding) specific parts.

4.5.1 Halfway quadratic shortage costs and restricting budget

In this subsection, we build the dynamic programming model for the restricting budget case, and give suggestion for a simplification in the calculations.

We develop the recursive formula of the expected capacity shortage cost-to-go from period t on, $f_t(b_t)$, where b_t denotes the remaining budget at the beginning of period t. All identical formulas correspond to the periods $t = 1, 2, \ldots, H$. At the start of period t, we observe demand realization, D_t, and only afterwards make our contingent capacity (θ_t) decision to minimize the overall sum of the actual capacity shortage costs, $c_s\left((D_t - U - \theta_t)^+\right)^2$, and the future (capacity shortage and budget deviation penalty) costs, $f_{t+1}(b_t - c_c\theta_t)$.

$$f_t(b_t) = E_{D_t}\left[\min_{0\le\theta_t} c_s\left((D_t - U - \theta_t)^+\right)^2 + f_{t+1}(b_t - c_c\theta_t)\right] \tag{25}$$

for all $t = 1, 2, \ldots, H$.

The minimal total expected capacity shortage cost, $f_0(B)$, can be found if we optimize the permanent capacity level at the beginning of the horizon, as

$$f_0(B) = \min_{0\le U} f_1(B - c_pU). \tag{26}$$

The part specific for the restricting budget is the closing stage of the dynamic program. That the budget is restricting, we can express as $b_{H+1} \ge 0$, where b_{H+1} is the budget in the end of period H. The closing cost-to-go we define as

$$f_{H+1}(b_{H+1}) \equiv 0. \tag{27}$$

Notice that because last period's demand is known before the last period's decision making takes place, it is always optimal to choose $\theta_H^* = \frac{b_H}{c_c}$, which results in $b_{H+1} = 0$.

We note that by applying the total expectation theorem one can achieve shorter calculation times. In particular, we can distinguish two cases: when the permanent capacity suffices to meet demand ($D_t \le U$), and when it does not ($D_t > U$). Consequently, we can rewrite $f_t(b_t)$ as

$$f_t(b_t) = E_{D_t}\left[\min_{0\le\theta_t} c_s\left((D_t - U - \theta_t)^+\right)^2 + f_{t+1}(b_t - c_c\theta_t)\,|D_t \le U\right] P(D_t \le U)$$
$$+E_{D_t}\left[\min_{0\le\theta_t} c_s\left((D_t - U - \theta_t)^+\right)^2 + f_{t+1}(b_t - c_c\theta_t)\,|D_t > U\right] P(D_t > U). \quad (28)$$

It is easy to recognize that for $D_t \le U$, the optimal contingent capacity decision is $\theta_t^* = 0$. Using this substitution, we can simplify the formula of $f_t(b_t)$ to

$$f_t(b_t) = E_{D_t}\left[f_{t+1}(b_t)\,|D_t \le U\right] P(D_t \le U)$$
$$+E_{D_t}\left[\min_{0\le\theta_t} c_s\left((D_t - U - \theta_t)^+\right)^2 + f_{t+1}(b_t - c_c\theta_t)\,|D_t > U\right] P(D_t > U), \quad (29)$$

which can be used for speeding up the calculations.

4.5.2 Halfway quadratic shortage and budget deviation costs

In this subsection, we build the dynamic programming model for the case when budget deviations are allowed. As in Section 4.4, the unit penalty cost of budget deficit is c_b^+, while each unit of budget surplus is rewarded with c_b^-.

We need to make only little changes in the previous model (in subsection 4.5.1. The cost-to-go formula, equation (25), and the total expected cost formula, equation (26), remain the same. The calculation fastening simplification formula, equation (29), is also applicable.

The only change in the model is in the final stage, stage $H+1$. To represent that budget deviations are allowed, and penalized, we rewrite equation (27) to

$$f_{H+1}(b_{H+1}) = c_b^-\,(-b_{H+1})^+ - c_b^+\,(b_{H+1})^+. \quad (30)$$

4.6. Halfway relative quadratic capacity shortage costs

Next to studying the halfway quadratic capacity shortage cost function suggested for calculational purposes by Warner and Prawda (1972), we introduce the relative quadratic cost function, which we consider as being more realistic. Our starting point is the observation that using the quadratic penalty means that we regard the situation with one unit of demand and no capacity as severe as having ten units of demand and nine units of capacity. We consider this phenomenon as being unrealistic, and as being a representational shortcoming of the halfway quadratic shortage cost function. By introducing the relative

quadratic shortage cost function, we eliminate this unwanted phenomenon. This cost function is defined as

$$C_s(U) = c_{cs} \sum_{t=1}^{H} \frac{\left((D_t - U - \theta_t)^+\right)^2}{D_t} \tag{31}$$

Again, we study the restricting budget case first, and the budget deviation penalty case afterwards.

4.6.1 Halfway relative quadratic shortage costs and restricting budget

In this subsection, we obtain the dynamic programming model under halfway relative quadratic shortage costs and restricting budget by changing the shortage cost formula in the model of subsection 4.5.1. We summarize the dynamic program formulation in the following lines.

$$f_t(b_t) = E_{D_t}\left[\min_{0 \le \theta_t} c_s \left(\frac{(D_t - U - \theta_t)^+}{D_t}\right)^2 + f_{t+1}(b_t - c_c\theta_t)\right] \tag{32}$$

for all $t = 1, 2, \ldots, H$,

$$f_0(B) = \min_{0 \le U} f_1(B - c_p U), \tag{33}$$

$$f_{H+1}(b_{H+1}) \equiv 0, \tag{34}$$

with $b_{H+1} = 0$. The simplification in accordance with equation (29) applies here as well.

4.6.2 Halfway relative quadratic shortage and budget deviation costs

We can arrive at the halfway relative quadratic shortage cost model with budget deviations combining the models in 4.5.2 and 4.6.1, as

$$f_t(b_t) = E_{D_t}\left[\min_{0 \le \theta_t} c_s \left(\frac{(D_t - U - \theta_t)^+}{D_t}\right)^2 + f_{t+1}(b_t - c_c\theta_t)\right] \tag{35}$$

for all $t = 1, 2, \ldots, H$,

$$f_0(B) = \min_{0 \le U} f_1(B - c_p U), \tag{36}$$

$$f_{H+1}(b_{H+1}) = c_b^-(-b_{H+1})^+ - c_b^+(b_{H+1})^+. \tag{37}$$

Simplification (29) again applies.

4.7. Results

We perform numerical experiments to study the optimal budget spending, shortage, and shortage cost patterns. Furthermore, we investigate the final budget deviations of different models. We consider the same set of parameters as default that we used in subsection 4.3.3 (horizon $H = 50$, demand is discretized gamma distributed with a mean of 50 and standard deviation of 20, budget $B = 3250$, capacity shortage cost coefficient $c_s = 1$, permanent capacity cost coefficient $c_p = 1$, and contingent capacity cost coefficient $c_c = 2.5$). Additionally, we used budget deficit coefficient $c_b^- = 10$, and budget surplus coefficient $c_b^+ = 5$ as default values.

4.7.1 Budget spending pattern

While in almost all the guiding budget cases we observed rather flat spending patterns, in the restricting budget cases these were mostly *concave and decreasing*. The most typical spending patterns we present in Figure 4.2, showing the spending pattern outcomes in percentages of the mean spending, for the halfway relative quadratic shortage costs, under the default parameter setting.

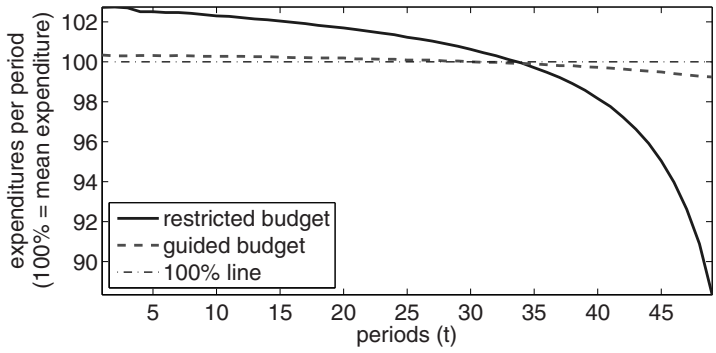

Figure 4.2: Budget spending pattern under halfway relative quadratic shortage cost function

The restricting budget cases utilize a larger range of spending values, generally (in Figure 4.2 from 102.7% to 88.34%) than the guiding budget cases (in Figure 4.2 from 100.3% to 99.24%), which is a result of the steeper drop of the restricting budget case in the last periods. Our intuitive explanation is

that for a constantly high demand in the last periods, heavy cutbacks may be required when the budget is limited, whereas in the guiding budget case it is not generally necessary.

One exception from the *concave and decreasing* shape is the linear shortage cost, restricted budget case, where the spending pattern is decreasing, but it is concave only for the first 40 periods, and becomes convexly decreasing afterwards. In the first 12 periods it decreases from 142.4% only to 141.7%, while in the rest of the periods it drops steeply to 17.8%.

Among the flat spending patterns of guiding budget cases, there are patterns having a slight convex increase on the whole horizon length. For instance the halfway quadratic shortage cost case with $c_b^- = 5$ and $c_b^+ = 1$ gives an optimal spending pattern, which increases from 99.5% to 101.3%. This spending pattern is rather flat in the first 30 periods, reaching there 99.9%, and bends more and more upwards in the last 20 periods.

4.7.2 Sensitivity analysis on the budget spending pattern for demand forecast error

Parameter estimation of future demand is subject to error. Even if demand is predicted in stochastic terms, the forecasted distribution parameters are only estimates in real-life situations. These estimations may reflect risk-averse, neutral, or risk-seeking spending behavior. We call a spending behavior risk-averse if it consistently overestimates the demand or its variance to be met later, risk-seeking if it underestimates, and neutral if it does not over- or underestimate. In this subsection, we investigate the effect of the risk-averse, neutral, or risk-seeking behaviors on the spending pattern.

We perform numerical experiments including into the model the consistent forecast error of the demands' real expected value ($E[D_t]$) and standard deviation ($Std[D_t]$), for all t. We take all the decisions based on forecast using the demand expected value and standard deviation estimates

$$\hat{E}[D_t] = E[D_t](1 + \Delta E[D_t]), \text{ and}$$
$$\hat{Std}[D_t] = Std[D_t](1 + \Delta Std[D_t]) \text{ for all } t = 1, 2, \ldots, H. \quad (38)$$

Considering the coefficients $c_s = c_p = 1$, $c_b^- = 5$ and $c_b^+ = 1$, we vary $\Delta E[D_t]$ as -0.1, -0.05, 0, $+0.05$, and $+0.1$, in combination with $\Delta Std[D_t]$ as -0.2, -0.1, 0, $+0.1$, and $+0.2$. The experimental values originate from simulating demand forecast error based on a history of 50 instances. Table 4.2 presents the trends in the spending patterns of the different cases.

We term a spending pattern increasing (decreasing) if it is increasing (decreas-

	$\Delta Std\,[D_t]$				
$\Delta E\,[D_t]$	-0.2	-0.1	0	0.1	0.2
-0.1	incr	incr	incr	incr	flat
-0.05	incr	incr	incr	flat	flat
0	incr	incr	flat	flat	decr
0.05	incr	flat	decr	decr	decr
0.1	decr	decr	decr	decr	decr

Table 4.2: Trends in the budget spending pattern for demand parameter estimates $\hat{E}\,[D_t] = 50$ $\hat{Std}\,[D_t] = 20$ (incr/decr = increasing/decreasing trend, flat = no trend).

ing) in more than 90% of the horizon, and we term the pattern flat, in the rest of the cases. Alternatively, we may call a spending pattern increasing (decreasing) if it is increasing (decreasing) in more than 3% from the beginning to the end, in percentages of the average (as in 4.2). Both of the two interpretations yield the results of Table 4.2.

We can observe the risk-averse, neutral, and risk-seeking spending behaviors at different parts of Table 4.2. Note that negative $\Delta E\,[D_t]$ or $\Delta Std\,[D_t]$ values corresponds to overestimation of the expected value or the standard deviation of the demand, respectively, while the positive values correspond to their underestimation. Thus, we can associate negative $\Delta E\,[D_t]$ and $\Delta Std\,[D_t]$ values with the risk-averse behavior (left, upper part), the zero values with the neutral behavior (middle), and positive values with the risk-seeking behavior (right, bottom part). We can conclude that risk-averse demand forecast results in an increasing budget spending trend, a neutral behavior gives flat pattern, and risk-seeking behavior leads to an increasing trend in expenditures.

4.7.3 Shortage and shortage cost patterns

In Figure 4.3 and 4.4, we show expected shortage and shortage cost patterns of the halfway relative quadratic shortage cost, restricted budget case, as well as for a low budget penalty case with $c_b^- = 0.04$ and $c_b^+ = 0.02$, and a high budget penalty case with $c_b^- = 0.16$ and $c_b^+ = 0.08$. By analyzing the two figures together, an interesting point to start with is that although the guided, high penalty case has a larger expected shortage than the restricted case in the last periods, it incurs less shortage cost on average. The relatively long tail of the shortage probability density of the restricted case is the explanation to this phenomenon (we discuss it in more detail by Figure 4.5). In line

with our expectations, the lower the budget penalty, the less attention budget deficits/surplusses get, the more freedom in the capacity decisions, and, consequently, the lower the shortage costs.

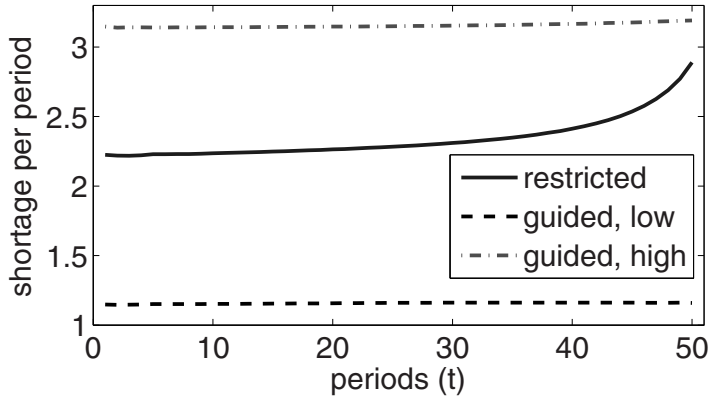

Figure 4.3: Expected shortage pattern under halfway relative quadratic shortage cost function

The quadratic shortage cost case, shows similar behavior as the one depicted and described above. Naturally differently, in all the linear shortage cost cases, expected shortage costs equal to the expected shortages times c_s. The shape of these graphs coincide with that of the spending pattern times -1, apart from linear scaling. The shortage (cost) graph increases from 0 to 0.07 by period 15, then increases steeply to 5.7 by the end. Regarding that the mean demand is 50, this corresponds to a service level of close to 100% in the first periods, and 88% in the end. This we can compare with the much less affected relative quadratic (or quadratic) cases, where a more stable service level is provided, which is always above 94%.

An interesting halfway quadratic shortage cost case, with $c_b^- = 5$ and $c_b^+ = 1$, is shown in Figure 4.5. We normalized both the shortage and shortage costs by their mean value. We can observe that even if the expected shortage has a decreasing trend, the expected shortage cost is increasing, particularly in the last periods. This phenomenon results from the quadratic shortage cost structure and the long tail of the shortage probabilities: although large shortages occur only with small probabilities, these still get emphasis because of the quadratically increasing penalty structure.

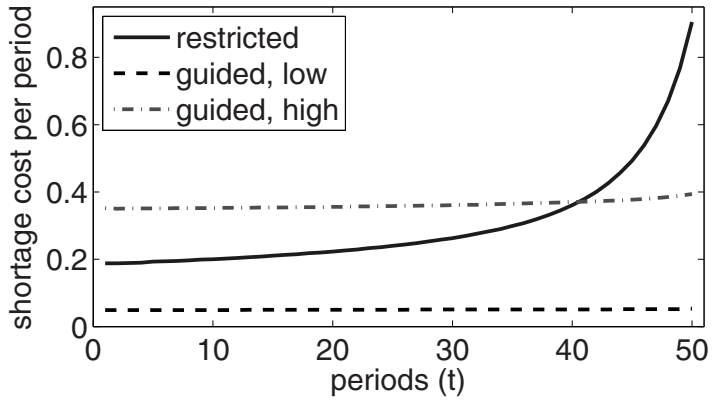

Figure 4.4: Expected shortage cost pattern under halfway relative quadratic shortage cost function

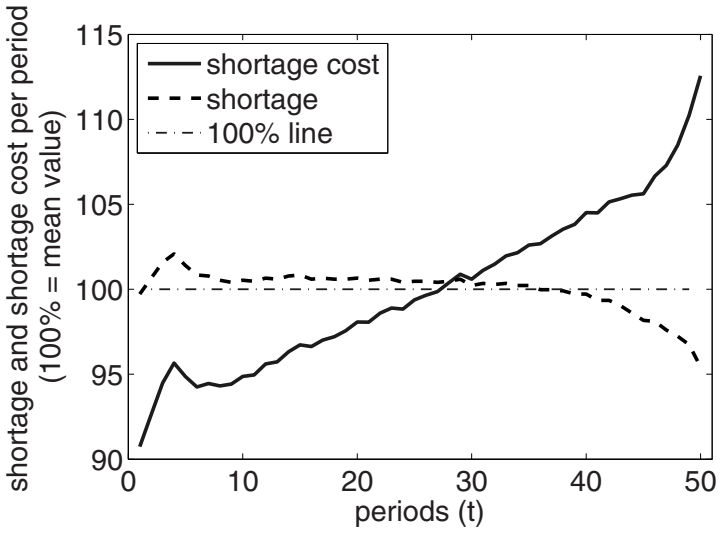

Figure 4.5: Relative shortages and shortage costs under halfway quadratic shortage cost function and guiding budget

4.7.4 Budget deviations

In Figure 4.6 and 4.7, we show budget deviation probability density functions in the halfway relative quadratic shortage cost, restricted budget case, as well

as for a low budget penalty case with $c_b^- = 0.04$ and $c_b^+ = 0.02$, a high budget penalty case with $c_b^- = 0.16$ and $c_b^+ = 0.08$, and an equal penalty case with $c_b^- = c_b^+ = 0.01$.

The distinguished role of zero can be especially noticed in the restricted case. At zero its probability density reaches 0.31, which exceeds the boundary of the figure. Figure 4.7 shows the graphs in Figure 4.6 in the environment of 0, zoomed. There we can see that the lower the budget penalty coefficients, the less budget enforcement, the less probability is accumulated at zero, and the more budget deviations and deficit are allowed. In the case, when $c_b^- = c_b^+$, the peak at zero disappears, since the value zero loses its particular role.

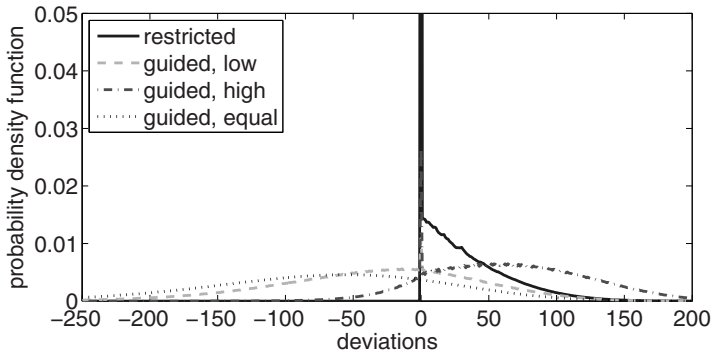

Figure 4.6: Budget deviations of models with relative quadratic shortage cost function

4.8. Conclusions

In many companies, fixed yearly budgets are allocated to the heads of departments to cover their fixed and variable expenses during the year.

In this chapter, we addressed the problem of longitudinal budget allocation to fixed and variable expenditures. We illustrated this general problem with a capacity budgeting task in services, where the budget is dedicated to cover permanent (fixed) and contingent (variable) capacity costs. We developed six different models determining the optimal permanent and contingent capacity levels so as to minimize the capacity shortage and budget deviation penalty costs. These models differ in the budget and in the capacity shortage cost modeling.

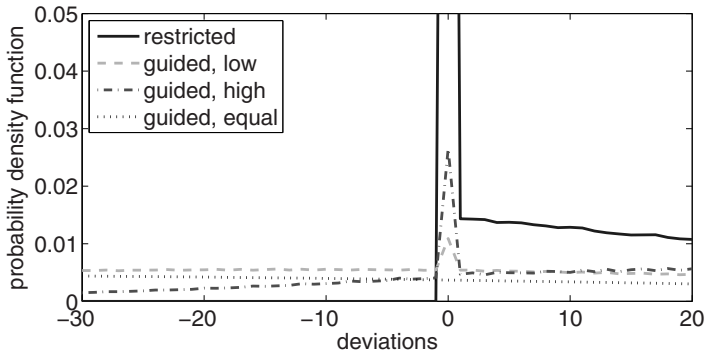

Figure 4.7: Budget deviations of models with relative quadratic shortage cost function, zoomed to the range around zero

We considered models with restricting budgets, where the budget strictly limits the total expenses, and models with guiding budgets, where budget deficits are allowed at some costs. Besides, we considered halfway linear, halfway quadratic, and halfway relative quadratic capacity shortage penalty costs, where the halfway refers to the penalty costs being zero when capacity exceeds demand.

In both the restricting and guiding budget cases, we developed analytic formulas in the linear shortage cost case, and found that near-optimal solutions can be found by using a newsvendor equation. For quadratic and relative quadratic cost functions, we proposed a solution with dynamic programming.

Through numerical experiments, we found that the service level during the year is more stable if the cost function is a halfway quadratic or relative quadratic instead of a linear one. This result may direct preference to not using linear cost functions in most of the real-life situations.

Furthermore, we found that the spending, the expected shortage and the expected shortage cost patterns are all rather flat in the guiding budget cases, as compared to the restricting budget cases. Interestingly, we found that the typical optimal expenditure pattern is concave and decreasing instead of the increasing patterns shown in Balakrishnan *et al.* (2007), which correspond with the saving-dissaving behavior observed in multiple restricting budget real-life cases. Most interestingly, decreasing optimal patterns belonged to all of the restricting budget cases we observed, suggesting that the saving-dissaving behavior is not rational. We think that further investigation is necessary to fully understand the rational behind the saving-dissaving behavior. E.g. future re-

search may show the effect of demand autocorrelation on the optimal spending pattern. We expect that positive autocorrelation can give an alternative explanation for the saving-dissaving spending behavior.

Although, in the guiding budget, halfway quadratic shortage penalty cost case we found situations, where the optimal spending pattern is slightly increasing, e.g. from 99.5% to 101.3%, even there we found no indication of the optimality of starting saving behavior. This result is in line with what Zimmerman (1976) suggests and empirically supports, even if we did not consider budget uncertainty in our model. Furthermore, performing sensitivity analysis on the spending pattern with consistently under- and overestimated demand, we found that the saving-dissaving pattern can be an indicator of a risk-averse attitude.

Finally, we found situations in the quadratic penalty case, where the expected shortages decrease, whereas the expected shortage costs increase. We explained this phenomenon by the long tail of the shortage probability density function, which tail gains a higher importance because of the quadratic shortage cost structure. We also studied budget deviations, and showed that the budget deviations have peak probability mass at zero especially in the restricted budget case, and that the less the budget is enforced, the more deviations, larger deficits we can expect. When the budget surplus is rewarded with the same weight as the deficit is penalized, the peak probability mass at zero disappears.

Chapter 5

Conclusions and Future Research

This last chapter summarizes our research studies on permanent-contingent capacity management and the conclusions drawn in each of them. Based on the conclusions, we give answers to the research questions posed in the first chapter, in Section 1.5. Furthermore, the study's applicability for machine and facility capacities is discussed. Finally, promising lines of future research are drafted.

5.1. Answers and main conclusions

We dedicate the following subsections to answering the research questions posed in the beginning of the thesis. All these subsections start with recalling the research questions.

5.1.1 Production-to-stock aggregate planning

1a. What form do optimal policies have in aggregate planning with backordering for non-stationary demand under the permanent-contingent concept, if the contingent capacity acquisition lead time is zero?

1b. Can we generalize the results answering (a) to general contingent capacity acquisition lead times?

Our first research questions addressed the integrated problem of inventory

and flexible capacity management. We investigated both research questions (1a) and (1b) via establishing and studying a single, but generic model. This generic model includes a fixed lead time for flexible capacity acquisition, which may be zero.

Our analytical results gave a characterization of the optimal policy via mathematical statements for any fixed non-negative integer lead time. The statements we proved were concerned with the convexity and sub-/supermodularity of cost-to-go functions.

The convexity results bring three-fold contribution. First, on the basis of convexity we can characterize the optimal policy to some extent: provided the capacity level, the optimal level of inventory after production is given by a state-dependent base-stock policy, where the dependency is on the capacity pipeline and the actual capacity. Secondly, convexity leads us to surmise little sensitivity for small mistakes in making decisions. Namely, it suggests that a small deviation from the optimal decision does not lead to drastic cost increase. Third, knowing convexity of the cost-to-go functions we could largely increase the efficiency of the optimal decision's calculation. These calculations include that of the optimal permanent capacity level, that of the optimal contingent capacity level, and that of the optimal production decision.

We proved sub-/supermodularity relations between multiple variables, which are the arguments of the cost-to-go functions. These relations help in further characterizing the optimal policy. From the economic substitution results between inventory and capacity levels, it follows that the optimal target capacity level is a decreasing function of the inventory level, and that the optimal inventory level is a decreasing function of any element of the contingent capacity pipeline vector and the permanent capacity. Moreover, we proved that the inventory (either before or after production), the pipeline contingent capacity, the contingent capacity to be ordered, and the permanent capacity are economic substitutes. We also showed that the stochastic demand variable and the optimal contingent capacity acquisition decisions are economic complements meaning that stochastically larger demand streams imply higher contingent capacity levels in optimality. A similar interpretation is also true for stochastically larger demand streams and the optimal inventory levels obtained after production.

Numerical experiments supplement our insights into the optimal permanent capacity level selection. We observed that the optimal level is not necessarily monotonic in demand variability. Nevertheless, for longer capacity acquisition lead times or higher costs of contingent capacity, optimal permanent capacity level in general increases as demand variability increases. On the contrary, for

shorter capacity acquisition lead times and lower costs of contingent capacity, optimal permanent capacity level in general decreases as demand variability increases.

By our experiments, we also paid attention to the value of flexibility, $VFC\%$, that expresses the relative difference between not using and using contingent capacity. Intuitively appealing, we observed that $VFC\%$ is higher when capacity acquisition lead time is shorter. Our results also generalize the findings of Tan and Alp (2005) for $L = 0$ to the case of positive capacity acquisition lead times, such that $VFC\%$ is higher when contingent capacity cost is lower or backorders are more costly (equivalently, when a higher service level is targeted). We note that, although $VFC\%$ decreases with an increasing lead time, the marginal decrease appears to be less and less as L increases. Besides, we also observed that $VFC\%$ with longer lead times persists to be comparable with $VFC\%$ with shorter lead times, meaning that flexibility is still valuable even when the capacity acquisition lead time is relatively long. We also analyze the relation between the value of flexible capacity and the demand variability. The results presented by Alp and Tan (2008) indicate that the value of flexibility is not necessarily monotonic (i.e. it does not increase or decrease consistently) as the demand variability increases for the case where the lead time is zero. We find out that this continues to be true for the case where the lead time is strictly positive as well, because the system has the ability to adapt itself to changes in coefficient of variation, CV, by optimizing the permanent capacity level accordingly. Nevertheless, for increasing values of the contingent capacity acquisition lead time we observe that the value of flexibility generally decreases when the demand variability increases. A longer capacity acquisition lead time deteriorates the effectiveness of capacity flexibility. This effect is amplified in case of higher demand variability. In other words, since the capacity needs are more predictable for lower demand variability, use of contingent capacity, which has to be ordered one lead time ahead, becomes more effective as compared to the high variability case. This also explains why the decrease in the value of flexibility as lead time increases is steeper when the variability is higher.

Not all the properties reported in the previous paragraph hold when demand forecast error is included in the model. Although we observed that the lower contingent capacity cost c_c and the shorter capacity acquisition lead times still consistently yield higher $VFC\%$, the value of flexibility as a function of the backorder cost coefficient becomes more irregular: especially for long lead times and large forecast errors, it becomes neither increasing nor decreasing.

Additionally, we studied a policy (class) that we observed to be optimal in many cases in our numerical experiments. This policy orders no contingent

capacity unless the actual capacity is fully utilized, which we refer to as complimentary slackness property (CSP). We showed through numerical examples that an optimal solution did not necessarily satisfy CSP. A typical example for CSP-policy being sub-optimal is having a highly uncertain period followed by several periods with no demand. We also provided some cases, for which we proved sufficient conditions assuring the optimal decisions to satisfy CSP. Through these cases we can recognize the importance of the contingent capacity - holding cost coefficient ratio: the optimal decision is likely to satisfy CSP if the ratio is high, and it is more likely to meet situations with an optimal decision not satisfying CSP, when the ratio is low.

5.1.2 Production-to-order

2. How can we use fast-response contingent capacity (approximately) optimally in production-to-order systems under setup costs and a fixed quoted customer lead time?

Research question 2 is concerned with production-to-order systems, where fast-response capacity changes are possible. We studied a set of Markov chain models, each instance representing a capacity planning policy, which may use workload information. Evaluation of the Markov models allowed comparisons between the workload-dependent policies with respect to permanent/contingent capacity, capacity switching, lost sales, work-in-process holding, and earliness/tardiness costs, as well as finding the optimal policy in different situations.

The calculation procedure we developed results in the optimal workload-dependent policy under the assumption of homogenous Poisson arrival process and exponential service times. Its main inputs are the interarrival rate, the service rates for each capacity level, and the maximum workload allowed. It is assumed that the jobs arriving by the time the maximum workload level has been achieved are lost. The permanent/contingent capacity, capacity switching, and lost sales costs are determined via steady-state analysis, whereas the work-in-process holding, and earliness/tardiness costs are determined via transient analysis.

We also characterized situations, where it is beneficial to use fast-response contingent capacity. We defined the value of flexibility, $VFC\%$, measured as the relative cost difference between the optimal fixed (which can use only one capacity level) and the optimal workload-dependent policy. We drew the conclusions that generally large switching cost coefficients, high demand rates, and long quoted lead times are detrimental, while high workload limits ($W_{\max}$)

beneficial for the value of flexibility. We also raised awareness of capacity discreteness, which can strongly affect $VFC\%$, as workload-dependent policies can counteract non-integer capacity needs, while fixed policies cannot. Additionally, we revealed that compared to the optimal fixed capacity policies, the optimal workload-dependent capacity planning policies can achieve a better due-date performance. In particular cases they can spare capacity, and decrease lost sales probability at the same time.

Following our experiments, we can surmise the form of an approximately optimal policy (class). In the uncapacitated case, we observed that using a sufficiently high order-acceptance rate, or equivalently a high workload limit ($W_{\max}$), is a safe choice when selecting the workload-dependent strategy. We found that for high workload limits, the optimal capacity up- and down-switching points tend to change less and less, and appear to form two lines. This observation may facilitate future research on the policy class comprising this linear type of policies.

5.1.3 Services

What dynamics of contingent capacity usage does a budget constraint entail in service environments, where backordering is not possible?

Our last research question addressed the effect of a budget constraint on the planning of service capacities. We realized that the longitudinal budget allocation to permanent and contingent capacities can be generally handled as budget allocation to fixed and variable expenditures. This observation broadened the scope and applicability of the answer we might find to this question.

We developed six different models determining the optimal permanent and contingent capacity levels so as to minimize the capacity shortage and budget deviation penalty costs. These models differ in the budget and in the capacity shortage cost modeling.

We considered models with restricting budgets, where the budget strictly limits the total expenses, and models with guiding budgets, where budget deficits are allowed at some costs. Besides, we considered halfway linear, halfway quadratic, and halfway relative quadratic capacity shortage penalty costs, where the halfway refers to the penalty costs being zero when capacity exceeds demand.

In both the restricting and guiding budget cases, we developed analytic formulas in the linear shortage cost case, and found that near-optimal solutions can be found by using a newsvendor equation. For quadratic and relative

quadratic cost functions, we proposed a solution with dynamic programming.

Through numerical experiments, we found that the service level during the year is more stable for halfway quadratic or relative quadratic cost functions than for linear cost functions. Since the unstable level of service is not acceptable in most of the real-life situations, our results show that, generally, linear capacity shortage cost functions do not represent well the usual allocation preferences. An option that may remedy the mentioned drawbacks of using linear shortage cost is to include cost of service instability in the model (e.g. penalizing the service level differences).

Furthermore, we found that the spending, the expected shortage and the expected shortage cost patterns are all flat in the guiding budget cases, as compared to the restricting budget cases. Interestingly, we found that the typical optimal expenditure pattern is concave and decreasing in contrast to the increasing patterns shown in Balakrishnan *et al.* (2007), which correspond with the saving-dissaving behavior observed in multiple restricting budget real-life cases. Most interestingly, decreasing optimal patterns belonged to all of the restricting budget cases we observed, suggesting that the saving-dissaving behavior is not rational.

In the guiding budget, halfway quadratic shortage penalty cost case we did discover situations, where the optimal spending pattern is slightly increasing with an even stronger growth towards the end (we gave an example for an increase from 99.5% to 101.3%, in percentages of the mean expenditure), but even there we have found no indication of the optimality of the starting saving behavior reported in Balakrishnan *et al.* (2007).

The few slightly increasing spending patterns we found can be termed similar to what Zimmerman (1976) suggests and empirically supports. The underlying dissaving behavior suggested was there, however, attributed to being a result of budget uncertainty in our model, while our examples show increasing spending patterns under budget certainty. Consequently, the spending patterns, which were found typical in Zimmerman (1976) are not necessarily the result of budget uncertainty.

In order to gain insight into risk-averse and risk-seeking attitudes, we performed sensitivity analysis on the spending pattern with consistently under- and overestimating demand throughout the horizon. We found that the attitudes towards risk ranging from risk-seeking to risk-averse behavior correspond typically with patterns having from decreasing to increasing spending trends. Consequently, risk-averse attitude can be the rational for saving-dissaving spending patterns.

We also suggest studying the spending pattern under autocorrelated demand

structure. Namely, we find it likely that a positive autocorrelation can be an alternative explanation for the saving-dissaving spending behavior.

The expected shortages and the expected shortage costs do not generally follow the same pattern. We found situations in the quadratic penalty case, where the expected shortage decreases, whereas the expected shortage costs increase. We explained this phenomenon by the long tail of the shortage probability density function, which tail gains a higher importance because of the quadratic shortage cost structure.

We also investigated the probability distributions of the final budget deviations. We showed that the special role of the targeted zero final budget manifests in a peak probability mass at zero of the budget deviation probability distribution. This peak is especially high in the restricted budget case. When the budget surplus is rewarded with the same weight as the deficit is penalized, the special role that the zero final budget value plays vanishes, and the peak probability mass at zero disappears. Our other important conclusion was that the less the budget is enforced, the more deviations, and the larger deficits we can expect.

5.1.4 General conclusions

Up to now we discussed our conclusions on to-stock, to-order, and service systems. After having the main results summarized, we can make a qualitative comparison between our models. If the outcomes of our different models lead to the same conclusion, it suggests its general applicability.

Our first observation is on the influence of service level on the relative value of capacity flexibility. We associate an increasing service level with an increasing backordering cost in Chapter 2, with an increasing lateness cost in Chapter 3, and with an increasing shortage penalty coefficient in Chapter 4. Some results we already presented in Chapter 2, Table 2.2. We show some further results for the base-case of Chapter 3 in Table 5.1. Based on these results, our general conclusion is that a higher service level target implies an increasing value of capacity flexibility.

tardiness penalty (€/job·hour)	60	70	80	90	100
$VFC\%$	1.11	1.57	2.41	3.21	4.07

Table 5.1: $VFC\%$ increases for an increasing service level in the to-order model in Chapter 3

Our second observation is on how the demand forecast error affects the relative value of capacity flexibility. We observed in Chapter 2 and in Chapter 3 that a structural overestimation of demand, which can be associated with a risk-averse behavior, leads generally to underestimating the relative value of capacity flexibility. Exceptions could be found in both cases: in Chapter 2 for long lead times and high backorder unit costs, in Chapter 3 for high capacity switching unit costs.

The last of our observation is that even when the contingent capacity costs are particularly high (e.g. twice as much as the permanent capacity costs), its value is still significant (e.g. in Table 2.2 in Chapter 2 it is $3-6\%$).

5.2. Few words on machines and other facilities

One could ask how far permanent-contingent capacity concept and our results on human workforce management apply for the case of machines or other facilities. Permanent machine capacity can be interpreted as the capacity of the in-house resources. Contingent capacity of a machine or facility can be interpreted as lease(, rent) or capacity reservation of the resource.

Leasing takes place usually in case of mobile machines (e.g. trucks) and smaller tools (e.g. mowers), whereas capacity reservation is more associated with large immobile machines and facilities (e.g. furnaces). Table 5.2 presents a few examples for leasing and capacity reservation in production and service provision. We remark that matching leasing with production is not common (our example is the leasing of a video camera so that a film is produced).

	leasing	capacity reservation
production	film \ video camera	medicine \ production facility
services	transportation \ truck	calculation \ computer-cluster

Table 5.2: Examples for leasing and capacity reservation in production and service systems.

Capacity reservation of a medicine production facility can be a good example for a production system, where a fixed capacity reservation lead-time plays an important role. We could claim that the models in Chapter 2 and 3 can be a good representation of such systems, and that the obtained results apply. However, it is better to pay attention to two aspects, which can ruin the model's applicability. On the one hand, the facility may not provide ample

amount of contingent capacity, and the provided amount may not be fixed as the contingent capacity providers often introduce production rationing. On the other hand, when logistics costs of raw materials and finished goods are comparable to the contingent capacity costs, the logistics and capacity ordering decisions need coordination. We also remark that renting production facilities resembles more to outsourcing if the contingent capacity is never under-utilized.

Temporary lease of helicopters to be ready for delivering emergency shipments is an example for using contingent capacity in services. The costs of the standard and the emergency delivery modes incur fixed and variable expenditures, respectively. It is also usual to have a common budget for transportation or logistics cost. Therefore, the content of Chapter 4 is applicable here. Renting or capacity reservation of some additional computer-clusters is similarly a good example. Note that the limited capacity availability in both of these two examples can undermine using our results.

5.3. Comparison of the permanent-contingent capacity, the subcontracting, and the dual sourcing models

We compare the permanent-contingent capacity management, the subcontracting, and the dual sourcing strategies from decision structure and flexibility point of view based on production-to-stock cases. In the service and to-order environments both the dual sourcing and subcontracting is less relevant, because in these environments the need for a rapid response on demand makes these options less attractive. We note that the to-stock dual sourcing models are also called dual supply models, and that we consider here the dual supply models with deterministic production lead-times[1]. We summarize our comparison in Table 5.3.

Janssen and de Kok (1999) study the use of a fixed and a flexible supplier. The fixed supplier delivers a fixed quantity each period, while the flexible supplier satisfies additional orders generated by a (R, S) replenishment policy. The (R, S) replenishment policy initiates replenishment in the beginning of each review period (that has length of R), so that the inventory position is raised to the order-up-to level, S. In case the inventory position is not less than S, no replenishment order takes places. We can make correspondence

[1]Dual supply models with stochastic lead times primarily address order-splitting, which has no sensible counterpart in the permanent-contingent capacity planning.

	Dual sourcing	Subcontracting	Perm.-cont.
reference	Janssen & de Kok (1999)	Yang *et al.* (2005)	Chapter 2
production decision	No (Yes)	Yes	Yes
resources	fixed/flexible sources	in-house/subcontractor	perm./cont.
underutilize	none	in-house	both

Table 5.3: The comparison of dual sourcing, subcontracting and permanent-contingent management strategies

between the fixed supplier and the permanent capacity, as well as between the flexible supplier and the contingent capacity. Similarly, we can associate the in-house capacity and the subcontractor in the subcontracting models with the permanent and contingent capacity, respectively (see Yang *et al.* (2005), discussed already in Chapter 2).

The concepts behind using the combination of a fixed supplier/in-house/permanent capacity and a flexible supplier/subcontractor/contingent capacity are quite similar: the level demand is satisfied by former, whereas the demand fluctuations are absorbed by the latter. However, there is a fundamental difference between the deliveries of the supplier or the subcontractor and the own capacity usage: once a replenishment order is placed for the supplier, it cannot be (partly) withdrawn if low demand is observed in the meanwhile; in contrast, ordering capacity has the advantage that when arrived it may be still underutilized.

We can observe that the lead times play an important role when pointing out the difference between the dual sourcing, subcontracting and permanent-contingent management. When production/delivery lead times equal zero in the model, this major difference between the dual sourcing, the subcontracting and the permanent-contingent capacity planning models disappear. The models, where no lead time is present are could be generally interpreted and applicable for all the three strategies (see Tan and Alp (2005) and Bradley (2004)).

We remark that there is no production stage considered in the model of Janssen and de Kok (1999) (the sources deliver finished goods). We note that there are other dual sourcing models, where the sources deliver the raw material or to a production stage, which may produce to stock. However, these models also do not consider production decision, just model the production lead time (see e.g. Yan *et al.* (2003)).

5.4. Future research

This section adds some ideas for future research. We stated in the very beginning of this thesis that we investigate tactical capacity planning of single stage, single item cases. Natural, but complicated extensions of the models in Chapters 2-3-4 are considering systems having multiple production stages or multiple products/services. In what follows, we go through the models in the thesis, and elaborate on their less complicated and more model-specific extensions. We finalize this last section by focusing on an environment-unified model, which could help in the assessment of firm's environment transition plans.

5.4.1 Production-to-stock

One interesting extension of the to-stock model is the relaxation of the assumption of independent demand periods. There are various ways of including demand dependency in the model. The analytically more attractive way is the use of a (first of higher order) Markov-modulated demand model with time-independent transitions. In this case the model's state space is extended with dimensions, in number equal to the demand model's order. We expect all the convexity and super/submodularity results to hold in this setting. Evaluation of the models is likely possible for a first-order demand model with 5-10 demand states and a capacity acquisition lead time of two periods. Typically, the numerically more attractive ways are the ones using first-order demand models. Among the first-order demand models we underline the benefits of the AR(1) demand model (first-order autoregressive), with which numerical calculations with varied correlation coefficient can be easily interpreted and presented.

The model's extension to general fixed production lead times is more complicated. Imagine that the production lead time is 3 periods, and we start to produce 12 items in the beginning of the first period. In the second and the third periods we stop using our capacity. Should then 4 items be ready in the end of the third period or none? The answer depends on the very operational structure of the system. If the 4 items were ready in the end of the first period, then the model boils down to the one in Chapter 2. If the 4 items are not ready in the first period, but ready in the end of the third, then additional state space dimensions need to archive the production started each period one lead-time length backwards. Additionally, a FIFO-processing paradigm may be assumed. For the case when less than 4 or none of the items are ready by the end of the third period, we do not see any reasonable model

extension. An exception might be when all the work-in-process not ready in production lead-time is lost. Losing outdated work-in-process creates a limit on the state space dimensions, and the resulting model may be relevant in the food industry.

5.4.2 Production-to-order

Extensions to our to-order model are numerous. We summarize the most relevant ones only. First, more general interarrival and service time distributions can be handled by using phase-type distributions instead of exponential. Unfortunately, employing phase-type distributions would heavily increase the number of states in the model. Therefore, approximations may be considered, where the less visited states behave memoryless. Second, service time parameters ($\mu_{w,c}$) can be set so that they correspond to situations where only one server can be assigned to a job or represent other dependencies on actual workload or capacity level. Third, switching costs for starting or suspending production are usually higher than increasing or decreasing production rate when production is already running. A simple extension can assign different costs for the different types of switchings. Fourth, a generalization of our model to positive capacity alteration lead times could reveal to what extent the assumption of instantaneous capacity changes is restrictive in workload-dependent capacity planning. However, this extension seems to be more complicated.

5.4.3 Services

We have analytical findings only for the service budgeting case with linear capacity shortage penalty. Employing the more interesting halfway quadratic and halfway relative quadratic capacity shortage penalty functions seem to be complicated for exact analysis. Our unpublished further analytical effort show that using a quadratic penalty function is promising. The analysis of the model under quadratic capacity shortage penalty may lead to compact formulas of the optimal permanent and contingent capacity decisions.

To further study the saving-dissaving spending behavior, we suggest including demand autocorrelation in the model. Positive autocorrelation may be an alternative explanation for the increasing spending patterns.

Empirical research has a potential for contribution in describing the capacity shortage and budget penalty functions. A good starting point for such research can be the economics literature on cardinal utility functions and their

measurement (see Schoemaker (1982)).

5.4.4 An environment-unified model

While the previous subsections on future research were engaged in flexible capacity management within a given environment, we now turn our attention to more generic observations and suggestions on the basis of this thesis and the existing literature. Our goal in this subsection is to present the way towards establishing an environment-unified model.

Some pieces of literature have already been concentrated on combining to-stock, and to-order models, providing an optimal choice for the type of production (e.g. Adan and van der Wal (1998) study a Markov-chain with dimensions of work-in-process and finished goods inventory levels; Rajagopalan (2002) minimizes of inventory costs under a service constraint via analysis of an $M|G|1$ queue). Yet, we are aware of no unified model of the three type of systems that includes services. Such unified model would make possible e.g. the quantitative comparison of the value of contingent capacity in the different environments. A unified model can also help to assess e.g. the capacity cost implications of firms' profile change to a new environment. Furthermore, it has particular educational value, since it gives an overview of operations management.

Modeling the combination of to-stock and to-order systems including contingent capacity management is straightforward based on Chapter 3 and Adan and van der Wal (1998). Similarly to the first model in Adan and van der Wal (1998), using that inventory and work-in-process are complementary, one can establish a single-dimensional state-space, where negative numbers refer to a number of items on stock and the positive numbers to orders waiting to be processed. This state-space can be extended by a capacity dimension as in Chapter 3.

For some service systems, to-order models can be still applicable. These are the services, which we can perceive as a guarantee for satisfying a set of orders, which are not completely known in advance. E.g. hospitals, airlines, and insurance companies provide services that provide coverage over a set of small orders (e.g. ordering a drink can be part of a flight service). The majority of the possible orders that can be demanded is often previously given in contracts and/or protocols (e.g. coverage of health insurance packages). As compared to the to-order systems, services have the extra complexity of anticipating the costs of the various possible order patterns that may come. While there is only a single order paid and a single possible outcome in the case of to-order

systems, by services one pays for covering the anticipated costs of yet unknown composite and pattern of future orders.

We propose an approach that may be employed to create a unified model of to-stock, to-order, and service systems. In particular, for studying the value of capacity flexibility in the different environments, we can take the two steps of the previous two paragraphs: (1) extend the first model in Adan and van der Wal (1998) by a capacity dimension as in Chapter 3, and (2) counting for multiple orders in multiple order pattern scenarios. It is important to take into account the cost of not immediately delivering (setup between services and to-order), and the costs of being not customer-specific (setup between to-order and to-stock).

We have some expectations regarding the results that a unified model of to-stock, to-order, and service environments can yield with respect to the value of contingent capacity. The basis of our expectations is a qualitative comparison that continues the contemplation on the characteristics of the three environments in Section 1.2. There we have already remarked that the level of customer involvement and the (production/customer order) lead time length are the main factors that affect whether to give more emphasis to the services or to the goods. Increasing the lead time from zero on, we can span a range from services, via to-order, to production-to-stock systems. Naturally, a short lead time necessary by services induces more pressure on towards employing capacity flexibility than a longer lead time, which is attributed to production systems. Consequently, we can expect that the value of capacity flexibility is higher for services and lower for production environments.

References

Abernathy, W.J., Baloff, N., Hershey, J.C., and Wandel, S. 1973. A three-stage manpower planning and scheduling model – a service-sector example. *Operations Research*, **21**(3), 693–711.

Adan, I.J.B.F., and van der Wal, J. 1998. Combining make to order and make to stock. *OR Spektrum*, **20**, 73–81.

Ahn, H., Righter, R., and Shanthikumar, J.G. 2005. Staffing decisions for heterogenous workers with turnover. *Mathematical Methods of Operations Research*, **62**(3), 499–514.

Alp, O., and Tan, T. 2008. Tactical Capacity Management under Capacity Flexibility in Make-to-Stock Systems. *IIE Transactions, to appear.*

Angelus, A., and Porteus, E.L. 2002. Simultaneous Capacity and Production Management of Short-Life-Cycle, Produce-to-Stock Goods Under Stochastic Demand. *Management Science*, **48**(3), 399–413.

Angelus, A., and Porteus, E.L. 2003. *On Capacity Expansions and Deferrals.* Tech. rept. Graduate School of Business, Stanford University, CA.

Anthony, R.N. 1965. *Planning and control systems: A framework for analysis.* Boston: Harvard Business School Press.

Avi-Itzhak, B., and Heyman, D.P. 1973. Approximate queuing models for multiprogramming computer systems. *Operations Research*, **21**, 1212–1230.

Balakrishnan, R., Soderstrom, N.S., and West, T.D. 2007. Spending Patterns with Lapsing Budgets: Evidence from US Army Hospitals. *Journal of Management Accounting Research*, **19**, 1–23.

Baynat, B., and Dallery, Y. 1993. A unified view of product-form approximation techniques for general closed queueing networks. *Performance Evaluation*, **18**, 205–224.

Bekker, R., and Boxma, O.J. 2005. *An M/G/1 queue with adaptable service speed*. SPOR-Report 2005-09, Eindhoven University of Technology, The Netherlands, submitted for publication.

Bentolila, S., and Bertola, G. 1990. Firing costs and labour demand: How bad is eurosclerosis? *The Review of Economic Studies*, **57**, 381–402.

Bertrand, J.W.M. 1983. The effect of workload dependent due-dates on job shop performance. *Management Science*, **29**, 799–816.

Bertrand, J.W.M., and Fransoo, J.C. 2002. Operations management research methodologies using quantitative modeling. *International Journal of Operations and Production Management*, **22**, 241–264.

Bertrand, J.W.M., and van Ooijen, H.P.G. 2002. Workload based order release and productivity: a missing link. *Production Planning and Control*, **13**, 665–678.

Bertsekas, D. 1976. *Dynamic Programming and Stochastic Control*. New York: Academic Press.

Bhandari, A., Scheller-Wolf, A., and Harchol-Balter, M. 2008. An exact and efficient algorithm for the constrained dynamic operator staffing problem. *Management Science*, **54**, 339–353.

Bradley, J.R. 2004. A Brownian approximation of a production-inventory system with a manufacturer that subcontracts. *Operations Research*, **52**(5), 765–784.

Bradley, J.R., and Arntzen, B.C. 1999. The simultaneous planning of production, capacity, and inventory in easonal demand environments. *Operations Research*, **47**(6), 795–806.

Chenery, H.B. 1952. Overcapacity and the acceleration principle. *Econometrica*, **20**, 1–28.

Dellaert, N.P., Jeunet, J., and Mincsovics, G.Z. 2008. *Optimizing permanent and temporary workforce under a budget constraint*. Working paper, Technische Universiteit Eindhoven, The Netherlands.

Dixit, A. 1997. Investment and employment dynamics in the short run and the long run. *Oxford Economic Papers*, **49**(1), 1–20.

Eberly, J.C., and van Mieghem, J.A. 1997. Multi-factor Dynamic Investment under Uncertainty. *Journal of Economic Theory*, **75**, 345–387.

Faddy, M.J. 1974. Optimal control of finite dams: discrete (2-stage) output procedure. *Journal of Applied Probability*, **11**, 111–121.

Filho, E.V.G.A., and Marçola, J.A. 2001. Annualized hours as a capacity planning tool in make-to-order or assemble-to-order environment: an agricultural implements company case. *Production Planning and Control*, **12**, 388–398.

Gans, N., and Zhou, Y.P. 2002. Managing learning and turnover in employee staffing. *Operations Research*, **50**(6), 991–1006.

Gelders, L.F. 1991. Production control in an 'engineer-to-order' environment. *Production Planning and Control*, **2**, 280–285.

Hansen, S.C., and van der Stede, W.A. 2004. Multiple facets of budgeting, an exploratory analysis. *Management Accounting Research*, **15**, 415439.

Heyman, D.P., and Sobel, M.J. 2004. *Stochastic Models in Operations Research, Vol. II.* Mineola, New York: Dover Publications.

Hiriart-Urruty, J.-B., and Lemaréchal, C. 1993. *Convex Analysis and Minimization Algorithms, Vol. I.* Berlin, Germany: Springer-Verlag.

Holt, C.C., Modigliani, F., and Simon, H.A. 1955. A Linear Decision Rule for Production and Employment Scheduling. *Management Science*, **2**(1), 1–30.

Janssen, F., and de Kok, T. 1999. A two-supplier inventory model. *International Journal of Production Economics*, **59**, 395–403.

Kim, J., Bae, J., and Lee, E.Y. 2006. An optimal P_{λ}^{M}-service policy for an M/G/1 queueing system. *Applied Mathematical Modelling*, **30**, 38–48.

Luss, H. 1982. Operations research and capacity expansion problems: a survey. *Operations Research*, **30**, 907–947.

Manne, A.S. 1961. Capacity expansion and probabilistic growth. *Econometrica*, **29**, 632–649.

Marie, R.A. 1979. An approximate analytical method for general queueing networks. *IEEE Transactions on Software Engineering*, **5**, 530–538.

Mincsovics, G.Z., and Dellaert, N.P. 2008. Workload-dependent capacity control in production-to-order systems. *IIE Transactions, to appear.*

Mincsovics, G.Z., Tan, T., and Alp, O. 2008. Integrated capacity and inventory management with capacity acquisition lead times. *European Journal of Operations Research, to appear.*

Minner, S. 2003. Multiple-supplier inventory models in supply chain management – a review. *International Journal of Production Economics*, **81-82**, 265–279.

Philipoom, P.R., Malhotra, M.K., and Jensen, J.B. 1993. An evaluation of capacity sensitive order review and release procedures in job shops. *Decision Sciences*, **24**, 1109–1133.

Pinder, J.P. 1995. An approximation of a Markov decision process for resource planning. *Journal of the Operational Research Society*, **46**, 819–830.

Pinker, E.J. 1996. *Models of flexible workforce management in uncertain environments.* thesis, Massachusetts Institute of Technology.

Pinker, E.J., and Larson, R.C. 2003. Optimizing the use of contingent labor when demand is uncertain. *European Journal of Operational Research*, **144**, 39–55.

Porteus, E.L. 2002. *Foundations of Stochastic Inventory Theory.* Stanford, California, USA: Stanford University Press.

Puterman, M.L. 1994. *Markov Decision Processes.* Wiley, New York, USA: John Wiley & Sons, Inc.

Rajagopalan, S. 2002. Make to order or make to stock: model and application. *Management Science*, **48**(2), 241–256.

Rocklin, S.M., Kashper, A., and Varvaloucas, G.C. 1984. Capacity expansion/contraction of a facility with demand augmentation dynamics. *Operations Research*, **32**, 133–147.

Rosenshine, M., and Obee, D. 1976. Analysis of a standing order inventory system with emergency orders. *Operations Research*, **24**(6), 1143–1155.

Ryan, S.M. 2003. Capacity expansion with lead times and autocorrelated random demand. *Naval Research Logistics*, **50**(2), 167–83.

Schlichter, M. 2005. *Stork Fokker - The production planning and control system.* Report, Logistics Management Systems, Technische Universiteit Eindhoven, The Netherlands, 2005, p55.

Schoemaker, P.J.H. 1982. The expected utility model, its variants, purposes, evidence and limitations. *Journal of Economic Literature*, **20**(2), 529–563.

Tan, T., and Alp, O. 2005. *An Integrated Approach to Inventory and Flexible Capacity Management under Non-stationary Stochastic Demand and Setup Costs*. Tech. rept. WP 132. Technische Universiteit Eindhoven.

Tijms, H.C., and van der Duyn Schouten, F.A. 1978. Inventory control with two switch-over levels for a class of M/G/1 queueing systems with variable arrival and service rate. *Stochastic Processes and Their Applications*, **6**, 213–222.

Topkis, D.M. 1998. *Supermodularity and Complementarity*. Princeton, New Jersey, USA: Princeton University Press.

Trivedi, V.M. 1981. A Mixed-Integer Goal Programming Model for Nursing Service Budgeting. *Operations Research*, **29**(5), 1019–1034.

van Houtum, G-J., Scheller-Wolf, A., and Yi, J. 2007. Optimal control of serial inventory systems with fixed replenishment intervals. *Operations Research*, **55**(4), 674–687.

van Mieghem, J.A. 2003. Capacity management investment and hedging: review and recent developments. *Manufacturing and Service Operations Management*, **5**(4), 269–302.

Vollmann, T.E., Berry, W.L., Whybark, D.C., and Jacobs, F.R. 2005. *Manufacturing planning and control for supply chain management*. New York: McGraw-Hill.

Warner, DM, and Prawda, J. 1972. A mathematical programming model for scheduling nursing personnel in a hospital. *Management Science*, **19**(4), 411–422.

Waters, D. 1996. *Operations management: producing goods and services*. Amsterdam: Addison-Wesley.

Wu, S.D., Erkoc, M., and Karabuk, S. 2005. Managing capacity in the high-tech industry, a review of literature. *The Engineering Economist*, **50**, 125158.

Yan, H., Liu, K., and Hsu, A. 2003. Optimal ordering in a dual-supplies system with demand forecast updates. *Production and Operations Management*, **12**(1), 30–45.

Yang, J., Qi, X., and Xia, Y. 2005. A Production-Inventory System with Markovian Capacity and Outsourcing Option. *Operations Research*, **53**(2), 328–349.

Zhang, Z.G. 2005. On the three threshold policy in the multi-server queueing system with vacations. *Queueing Systems*, **51**, 173–186.

Zimmerman, J.L. 1976. Budget Uncertainty and the Allocation Decision in a Nonprofit Organization. *Journal of Accounting Research*, **14**(2), 301–319.

Zufryden, F.S. 1975. Optimal Multi-Period Advertising Budget Allocation within a Competitive Environment. *Operational Research Quarterly*, **26**(4), 743–754.

Summary

Studies on Tactical Capacity Planning with Contingent Capacities

The main motivation for this thesis is the emerging use of the workforce provided by external labor supply agencies (ELSAs). This temporary workforce we refer to as contingent capacity. Temporarily extending the in-house, permanent capacity by contingent capacity in peak demand periods can create a competitive advantage for companies, which need to meet uncertain and fluctuating demand. This thesis investigates tactical capacity planning of single stage, single item production-to-stock, production-to-order and service environments.

In production-to-stock environments, our focus is on aggregate production and capacity planning, when there is an option of using contingent capacity next to the permanent capacity. We analyze the optimal coordinated production and capacity decisions when the lead times for contingent capacity acquisition is zero or positive. Our main contribution is that we analytically show the structure of the optimal policy and characterize the economic relations between the inventory level, the permanent capacity, and the contingent including those in the pipeline. We also point out important differences between the zero and the positive contingent capacity acquisition lead time cases. Namely, while for the zero lead time case increasing demand variability generally increases the value of using contingent capacity, its effect is reversed when a positive lead time is present.

In production-to-order environments, we study workload-dependent capacity management policies. The permanent production capacity can be increased and decreased at specific workload levels by employing contingent resources. A workload-dependent capacity management policy specifies these workload values, which trigger capacity level switches, and the permanent capacity level. We perform numerical experiments to learn about the structure of the optimal policy with respect to capacity, capacity switching, lost sales, work-in-process cost, and due-date performance. We show the effect of discreteness of the

capacity levels, and what potential the use of contingent capacity for given quoted lead time - order interarrival time combination could have. Our results also indicate that a special policy class is possibly close to optimal in many cases.

In service environments, we concentrate on the use of a capacity budget, which can be allocated to permanent and contingent capacity to meet demand over a finite horizon. In the beginning of the horizon the permanent capacity cost are incurred, and the rest of the budget is dynamically allocated to cover the contingent capacity expenditures. Our main result is that we give a possible explanation for the saving-dissaving behavior, which some scientists observed with in their empirical study.

Finally, we answer the research questions posed in the beginning of the thesis, draw conclusions, and point at interesting future research questions.

Samenvatting

Studies Aangaande Tactische Capaciteitsplanning met Contingente Capacitiet

De belangrijkste aanleiding voor deze dissertatie is het toenemend gebruik van personeel dat wordt geleverd door externe uitzend- en detacheringsbureaus. Een dergelijk bureau duiden we aan met het Engels acroniem ELSA (voor external labor supply agency) en het tijdelijk personeel dat ze verschaffen noemen we 'contingente capaciteit'. Het tijdelijk uitbreiden van interne, permanente capaciteit met dergelijke contingente capaciteit tijdens perioden van piekvraag kan een concurrentievoordeel opleveren voor bedrijven; zij kunnen daarmee tegemoet komen aan de onzekere en fluctuerende vraag. Deze dissertatie behandelt tactische capaciteitsplanning van enkelvoudige enkelstuksbewerking produktie-op-voorraad, produktie-op-bestelling en dienstverleningsomgevingen.

In produktie-op-voorraad omgevingen ligt de nadruk op de geaggregeerde productie- en capaciteitsplanning waarbij er de mogelijkheid is om naast de permanente ook contingente capaciteit in te zetten. We analyseren de optimaal gecoördineerde productie- en capaciteitsbeslissingen waarbij de levertijd voor het verkrijgen van contingente capaciteit nul of positief is. Onze belangrijkste bijdrage is de analytische afleiding van de structuur van het optimaal beleid en de economische relaties tussen voorraadniveau, permanente capaciteit en contingente capaciteit (inclusief die nog in de pijplijn). We stipuleren ook belangrijke verschillen wanneer de levertijd voor contingente capaciteit niet nul maar positief is. Terwijl in geval van een levertijd van nul bij een toenemende variabiliteit het nut van contingente capaciteit in het algemeen toeneemt, is het effect namelijk juist omgekeerd wanneer er een positieve levertijd is.

In produktie-op-bestelling omgevingen bestuderen we beleid voor werklastafhankelijk capaciteitsbeheer. De permanente productiecapaciteit kan worden uitgebreid en ingekrompen bij specifieke niveaus van werklast door gebruik te maken van contingente capaciteit. Een werklastafhankelijk capaciteitsbeheerbeleid specificeert zowel de werklastniveaus die dergelijke ca-

paciteitsverandering voorschrijven, alsook het niveau van de permanente capaciteit. We voeren numerieke experimenten uit om de structuur van het optimale beleid te doorgronden in relatie tot capaciteit, capaciteitsveranderingen, gemiste verkopen, onderhanden werk kosten en levertijdprestatie. We isoleren de rol van de geheeltalligheid van de capaciteitsniveaus en de potentie van het gebruik van contingente capaciteit voor combinaties van gegeven gequoteerde levertijd en bestelling-tussenaankomsttijd. Onze bevindingen maken ook duidelijk dat er een speciale klasse van beleid in vele gevallen bijna-optimaal is.

In een dienstverleningsomgeving richten we ons op het gebruik van een capaciteitsbudget dat gealloceerd kan worden over permanente of contingente capaciteit om aan vraag te voldoen gedurende een eindige horizon. In het begin van de horizon worden de kosten voor permanente capaciteit gemaakt terwijl de rest van het budget dynamisch gealloceerd wordt om kosten voor contingente capaciteit te dekken. De voornaamste bevinding is een mogelijke verklaring voor het uitstel- en inhaalgedrag hetgeen door sommige wetenschappers in empirische studies geobserveerd is.

Uiteindelijk beantwoorden we de onderzoeksvraag die aan het begin van de dissertatie gesteld is, trekken we conclusies en verschaffen interessante vragen voor vervolgonderzoek.

About the author

Gergely Mincsovics was born in Budapest, Hungary, in 1980, and lived his first twelve years in a small, nice town, Szentendre. After moving to Budapest, he attended there the Árpád Grammar School taking specialization in mathematics. The programming experience he acquired during these years as a hobby directed him to the Eötvös Loránd Science University's Software Design and Mathematics joint Bachelor-Master of Science program to continue his studies. Among the few, he finished this BSc-MSc program in the prespecified 5 years. While doing his study, he become an instructor / teaching assistant of more subjects from already his second year on to the end of the five years. He was actively sporting and he took part in organizing introductory camps for the new-coming first year students. In the last two years he took the specializations, mathematical modeling, optimization, statistics and database management. He graduated in 2004, completing his master software addressing the implementation of the N-body problem's Barnes-Huts parallel calculation model, as well as completing his thesis, entitled "The Dirichlet and the inverted Dirichlet distribution," under the supervision of Tamás Szántai.

After graduating, Gergely got married with Yu Da, who he got know at the university. In the same year, they both got employment from the Technical University of Eindhoven, The Netherlands. Since then, Gergely has been working as a Research Assistant at the Department of Technology Management, taking part of the PhD-program of the Beta Research School. He has conducted research on the role and use of contingent capacity in to-stock, to-order and service capacity planning environments under the supervision of dr.ir. N.P. Dellaert, prof.dr.ir. J.W.M. Bertrand, and prof.dr.ir. J. van der Wal. This PhD dissertation is a result of Gergely's research activities between 2004 and 2008. Additionally, he studied capacity planning issues in the health care sector, which do not form part of this thesis, but are available on request.

On 3rd November, 2008, Gergely will defend his PhD dissertation at the Technical University of Eindhoven, in Hall 5, Auditorium, at 16:00pm.