Technische Universität München

TUM School of Management

Stochastic Models for Performance Analysis and Optimization of Design and Control Policies in Manufacturing Systems

Miray Öner Közen

Vollständiger Abdruck der von der Fakultät für Wirtscaftswissenschaften der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften (Dr. rer. pol.) genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Rainer Kolisch
Prüfender der Dissertation:	1.	Univ.-Prof. Dr. Stefan Minner
	2.	Univ.-Prof. Dr. Heinrich Kuhn

Die Dissertation wurde am 30.01.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Wirtscaftswissenschaften am 15.05.2017 angenommen.

Abstract

This thesis focuses on the decision-making problems pertaining to management of make-to-order manufacturing systems. It covers the analysis of long-term decisions regarding how to design as well as the short-term decisions regarding how to control a manufacturing system. It proposes stochastic models for the performance analysis and optimization of these decisions.

The priority dispatching problem of a make-to-order manufacturing firm, which serves customers with predetermined expectations for the amount of time to wait is analyzed. Tardiness penalties are incurred whenever these expectations are not met. The problem is modeled as a Markov decision process. It is found that obtaining a near optimal performance is possible by employing simple priority sequencing rules that are in line with the tardiness penalty structure. The model is also extended to consider the problem in which the firm makes a leadtime quote, in combination with an appropriate price quote, to arriving customers individually. The customers respond the quote by accepting it or by taking their businesses elsewhere. The results show that, it is possible for the firm to attract more customers and, at the same time, increase the service level if marketing and operations groups collaborate for developing a joint quotation and dispatching policy.

The traditional guidelines for the improvement and the optimization of performance of serial production lines are revisited by introducing models that appreciate the fundamental differences between human and machine operators in their behavior when processing jobs. A simulation model is developed for comparing paced and unpaced production lines based on their efficiency. The results show that an unpaced design is superior in many real-world settings, however, its superiority is overestimated by models ignoring the characteristics of human work behavior. The exact model developed for revisiting the guidelines for the optimal design of unpaced lines considers that minimizing the output variability or maximizing the service level might be more important than maximizing throughput for today's manufacturing firms that perform just-in-time production. It is found for a line with human operators that these objectives can be achieved by allocating the total available buffer spaces as well as the total workload in a decreasing pattern, rather than following the traditional guidelines.

Acknowledgements

First, I would like to express my grateful thanks to my supervisor Prof. Dr. Stefan Minner who offered invaluable guidance, support and encouragement throughout my Ph.D. studies. The opportunity to pursue highly interesting research projects and to be a part of a great team at Technische Universität München, have been of great value for me. I would like to offer my special thanks to Prof. Dr. Heinrich Kuhn for providing insightful comments and for being part of the examination committee, as well as, to Prof. Dr. Rainer Kolisch for being my mentor during my Ph.D. studies and for being the chairman of the examination committee.

I am very thankful to the current and former members of the chair of Logistics and Supply Chain Management at TUM School of Management: Yuka Akasaka, Szymon Albinski, Christian Bohner, Dr. Maximilian Budde, Dr. Pirmin Fontaine, Evelyn Gemkow, Sebastian Malicki, Christian Mandl, Thitinan Pholsook, Jun.-Prof. Dr. Anna-Lena Sachs, Dr. Martin Stößlein, Florian Taube, Dariush Tavaghof Gigloo for offering me valuable feedback, support and friendship.

I extend my thanks to my beloved family for encouraging me to follow my ambitions in life. I thank to my husband Soner Közen for continuously motivating me and giving me the strength to carry on whenever I needed. Finally, special thanks to my expected daughter Mine Közen for further strengthening me.

Contents

List of Tables

List of Figures

List of Acronyms

ANOVA	Analysis Of Variance.
BAS	Blocking After Service.
CT	Cycle Time.
CV	Coefficient of Variation.
CRL	Customer-Required Leadtime.
CTMC	Continuous Time Markov Chain.
DC	Decide upon Completion.
EDD	Earliest-Due-Date.
FCFS	First-Come-First-Served.
JIT	Just-In-Time.
KPI	Key Performance Indicator.
LDD	Latest-Due-Date.
LS	Leadtime Sensitive.
MC	Markov Chain.
MDP	Markov Decision Process.
MET	Maximum Endurance Time.
MILP	Mixed Integer Linear Programming.
MTO	Make-To-Order.
OM	Operations Management.
PA	Position Arriving order.
PS	Price Sensitive.
SL	Service Level
S/OPN	Slack per remaining Operation.
TL	Tightness Level.
TT	Takt Time.
WIP	Work-In-Process.
WS	Workstation.

1. Introduction

1.1. Motivation

Today's manufacturing firms compete in highly challenging environments, since the efforts for finding and adopting solutions to provide better products cheaper and faster have substantially increased already during the last decades. As a result, the pressure on these firms to make more informed decisions and to act nearer to the optimal has increased, as well as the complexity involved in the decision problems they face.

In order to support decision makers in better understanding the behavior of a complex system and provide them with accurate guidelines, models should incorporate those aspects that are known to have a decisive role on the dynamics of a manufacturing system. Developing such models has been and is still being of great interest to operations management (OM) researchers. Uncertainty, which can have external or internal sources, is one of these aspects and it has been well appreciated in the field. The focus on customized production, high system responsiveness and delivery reliability, as well as, the concept of just-in-time production are among others. Realizing that human operators, who are still commonly used as resources in production systems, behave fundamentally differently than machines, is also an important aspect for better understanding the dynamics of a production system.

This thesis is motivated by the pressing need for models of highly stochastic and dynamic manufacturing environments that incorporate the above mentioned complicating aspects. The thesis mainly focuses on numerical models that satisfy this requirement while allowing an exact performance analysis and optimization. The optimal decision-making policy for a complex system provides a number of valuable information to decision-makers. It leads to useful guidelines for optimal decision-making, to an understanding of the best possible system performance and how close to it the performances of simpler policies are. The decision problems that are taken under investigation in this thesis include both short- and long-term decisions pertaining to management of manufacturing systems such as, the optimal control of manufacturing systems and the design of serial production lines, respectively.

1.2. Problem definition

In order to gain a competitive advantage by means of improving the extent to which their products meet the customers' unique needs, companies choose to offer a high variety of products. Keeping safety stocks for each product variant may either not be feasible or prohibitively costly for some of these firms. As a result, many companies move to manufacturing based on a make-to-order (MTO) strategy. Adapting this strategy to satisfy customer demand requires successful leadtime management in the system. Attaining a good performance in matching the production leadtimes with the customer-required leadtimes is crucial since on-time delivery strongly influences the perceived quality of service. Therefore, the key external customer-related performance indicators of MTO systems are the percentage of orders delivered on-time and expected tardiness of orders (see e.g. Hopp and Spearman (2001)). These are linked to the internal performance measures such as system utilization, work-in-process and production leadtime.

The management of MTO systems involves several decision-making problems, some of which are: which orders to accept, how to quote due-dates to customers, how to schedule the orders and what capacity to install. There is a large body of research investigating these managerial issues separately or in combination. The reviews (e.g. see Keskinocak and Tayur (2004) and Slotnick (2011b)) categorize the studies based on various features such as: arrival of orders being static or dynamic, arrival times and processing times being deterministic or stochastic, the objective function under consideration (e.g. maximizing profit, minimizing tardiness-related measures) and applied methodology (e.g. analytical models, simulation models, heuristics). Due to the complexity of analyzing such systems, most of the studies that deviate from a static setting and simple first-come-first-served (FCFS) discipline in processing either apply simulation or disregard some source of uncertainty, e.g., in arrival and/or service processes. In this thesis, the priority sequencing problem in a make-to-order production system, to which orders arrive dynamically following a stochastic process with stochastic processing and customer-required leadtimes, is modeled as a Markov decision process (MDP).

Businesses with the ability to charge prices that are in good relation to the value a product creates for a specific customer at the time of purchase can obtain higher revenues (Cross and Dixit, 2005). An important criterion for the customers when appraising the value of a product is the delivery leadtime (Keskinocak and Tayur, 2004). Making dynamic price/leadtime quotes is a way for firms to increase revenue, since it allows for charging prices that match the created value more precisely.

Timbuk2 is a custom messenger bag producer that applies such a policy. It typically quotes a leadtime of two weeks to retailers, who are price-sensitive, while offering only a leadtime of a few days to direct consumers who feel less tolerant about waiting, and combines the additional value created for direct consumers with higher price quotes (Plambeck, 2004). A semiconductor manufacturer that assembles final products to order distinguishes between the quotes of customers in peak demand phases during which resources become scarce. This is because timely delivery of the product is crucial for some. These customers are willing to pay a premium for receiving short leadtime quotes (Guhlich et al., 2015).

Despite their positive short-term impact on revenues, short leadtime quotes may hurt the on-time delivery reputation if possible lateness is not carefully considered. Typically, marketing specifies what value to provide at which price, focusing on revenue, while operations aim at delivering the value as promised (e.g. on-time), focusing on cost. However, research at the marketing-operations interface acknowledges that this might lead to sub-optimal solutions. In general, marketing and operations groups should act in coordination. Ideally, they should collaborate in order to develop a joint policy (Tang, 2010). In this thesis, the joint pricing, leadtime quotation and due-date-based order dispatching problem is modeled as an MDP.

The assembly line design problems faced by a first tier automotive equipment supplier, which largely uses human operators in the lines and adopts a just-in-time production policy, motivates the second part of the work presented in this thesis. In the literature, the performance evaluation and optimization of serial production lines are well-investigated. It attracted the attention of researchers since 60's (see e.g. Dallery and Gershwin (1992) and Papadopoulos and Heavey (1996) for extensive reviews). After Hillier and Boling (1966) conjectured the existence of the "bowl phenomenon", which suggests unbalancing the workload allocation in an unpaced production line with stochastic processing times for achieving higher throughput, unbalancing of unpaced lines have received a considerable interest. Hudson et al. (2015) provides a recent review on research in this direction. However, almost all of the existing studies utilize processing time assumptions that are not well-suited for analyzing lines with human operators.

As identified by Boudreau et al. (2003) the operations management (OM) literature typically models human work behavior assuming that humans operate independent of each other and that they are stationary resources (e.g. no learning or tiredness). Bendoly et al. (2006) stress the importance and the benefits of incorporating human

characteristics into existing operations models to overcome their lack in precision of representing real-world phenomena.

In the area of production line design, Powell and Schultz (2004) were able to refute the classical guideline (e.g. Conway et al. (1988)) saying that throughput degrades with line length. They showed that built-in balancing mechanisms attributable to human behavior at least partially compensate for the higher degree of interference in longer lines. Similarly, Schultz et al. (1998) found that, in contrast to the prevailing opinion, low buffer lines are not necessarily worse than high buffer lines because higher interdependence between workers leads to higher motivation and productivity. The direct and immediate feedback about the relative work speed especially encourages slower workers to increase efforts to not cause idle times for co-workers. Such a state-dependent behavior of human operators requires revisiting production line performance assessment, as well as, the guidelines for their optimal design.

In this thesis, first, paced and unpaced lines are compared in the light of behavioral results. Such a comparison provides scientific guidance for the overall design of a line without the prior commitment to any of the line types. Second, the optimal design of an unpaced line is analyzed. In particular, the workload and buffer space allocation problems are investigated separately and simultaneously under various objectives.

This thesis contributes to the huge body of literature on manufacturing systems by answering the following research questions:

1. Where is the efficient frontier of the on-time probability and the expected tardiness of orders and how close to this frontier the system performs when simple priority sequencing rules are employed?
2. How much can profitability be increased and how are the related key performance indicators (KPI) affected when the joint optimization of (i) price/leadtime quotation and (ii) dispatching is considered, rather than a sequential approach for making decisions (i) and (ii)?
3. Which type of an assembly line is better for a firm that utilizes human workforce: paced or unpaced? Does the conclusion differ when models that ignore the human aspect are used?
4. What are the optimal workload and buffer space allocation guidelines for an unpaced line with human operators and how different these are from the

guidelines for a line with machine operators?

1.3. Outline

The models introduced by this thesis are mainly divided into two groups based on the type of decisions under consideration: (1) short-term decisions related to the control of make-to-order manufacturing systems (2) long-term decisions related to the design of serial production lines. In particular, the first group considers the order selection and scheduling decisions. Each group consists of two models.

In Chapter 2, we review the relevant literature in two sections each of which covers studies related one group. Chapter 3 considers the priority sequencing problem in a make-to-order production system where the time allowance for the orders to be fulfilled is externally determined by the customers. The model utilizes the theory of Markov processes and takes the minimization of the long-run average tardiness cost as the objective. The chapter is based on the working paper Öner Közen and Minner (2016c) that is currently under review in *European Journal of Operational Research*.

In Chapter 4, we present a model that extends Öner Közen and Minner (2016c). The joint optimization problem of a profit maximizing make-to-order manufacturing firm that (i) dynamically quotes a price/leadtime pair to arriving prospective customers, who then decide whether or not to place an order by trading off the price and leadtime, and (ii) dispatches placed orders, is modeled as a Markov decision process (MDP). This model allows comparison of firm's profitability under two scenarios: sequential and joint decision-making by marketing and manufacturing groups. The chapter is based on Öner Közen and Minner (2016b).

The two models that consider long-term decisions regarding the design of serial production lines are presented in Chapters 5 and 6. Both chapters focus on those lines in which operations are carried out by humans and, in the light of behavioral findings, revisit the traditional guidelines for line design.

Chapter 5 presents a simulation model that takes into account operators' reaction to their immediate environment with the desire to avoid causing idleness, their tiredness due to showing this reaction, as well as, their level of flexibility and experience. This model allows a comparison of paced and unpaced lines based on the steady-state efficiency and the characteristics of the output process. The chapter is based on Öner Közen et al. (2016) that is currently under review for publication

in the *Special Issue on "Human factors in industrial and logistic system design" in Computers and Industrial Engineering.*

The second model, which considers the state-dependent adjustments in processing rates of workers as well as their tiredness due to making these adjustments, uses an exact Markovian approach. Chapter 6 presents the model and the analysis of the workload and buffer allocation problems as separate and simultaneously solved optimization problems under various objectives. The chapter is based on Öner Közen and Minner (2016a).

In Chapter 7, we summarize the main findings of the thesis and propose directions for future research.

2. Related Literature

The chapter is structured as follows. In Section 2.1, the literature related to models presented in Chapters 3 and 4, i.e., the literature on the order selection and scheduling problem, is reviewed. In Section 2.2, the review of the literature on the performance analysis and optimization of production lines with human operators, which is extended by the models presented in Chapters 5 and 6, is given.

2.1. Order selection and scheduling in make-to-order manufacturing systems

Slotnick (2011b) provides a review of studies on order acceptance and scheduling decisions. Keskinocak and Tayur (2004) review the literature on due-date management with (as well as without) order selection decisions. We mainly divide the existing research related to our work into two groups based on the type of analysis: performance evaluation or optimization.

2.1.1. Performance evaluation

In the literature, a significant number of studies use simulation for modeling a dynamic MTO manufacturing environment and investigating the performance of different scheduling policies. Rajendran and Holthaus (1999) and Jayamohan and Rajendran (2000) state that, due to the simplicity of their implementation, the dispatching rules are often used in real-life manufacturing systems and investigate the performance of dispatching rules based on several performance measures, including the proportion of tardy jobs, mean tardiness and maximum tardiness. Barman (1998) considers the combination of popular simple priority rules in a three-stage flow shop. He finds that, for mean and maximum tardiness, the earliest-due-date (EDD) rule is the most effective when used at any of the three stages in combination with the other rules. However, he also finds that it is the least effective when it comes to reducing the proportion of tardy jobs.

Some researchers present queuing models under due-date-based priority sequencing. Jackson (1961) considers a queuing system with dynamic job priorities. The queuing system operates under the EDD rule. The author provides conjectures about the upper tails of the waiting time distribution of jobs based on the results from simulation studies. Kleinrock and Finkelstein (1967), Goldberg (1977), Goldberg (1980) and Bramson (2001) investigate EDD dispatched queuing systems.

None of these studies provides the exact distribution of the time spent in the system. On the other hand, heavy traffic approximations are provided by Doytchinov et al. (2001), who consider a single-server queuing system with the EDD queuing discipline and the distribution of customer lateness. Kruk et al. (2011) consider a single-server EDD dispatched queuing system where the customer service stops when the due-date is reached. Altendorfer and Jodlbauer (2011) use queuing theory to derive analytical expressions for the percentage of orders completed before their due-date and for the expected tardiness of orders in an M/M/1 queue that receives orders with exponentially distributed customer-required leadtimes, assuming that the queuing discipline is FCFS.

Others analyze the performance of policies for the selection and scheduling of orders. Ebben et al. (2005) consider a job shop environment working under MTO and compare different order acceptance and scheduling policies with the objective of maximizing the utilization rate while keeping the number of accepted orders completed before their due-dates above a certain threshold. They consider the EDD dispatching rule and a mixed integer linear programming (MILP) model for generating order processing schedules and state that the MILP model for scheduling jobs did not outperform the EDD rule. An explanation provided by authors for observing this result is the MILP model being given a limited computation time. Rogers and Nandi (2007) consider a multi-stage manufacturing system for investigating order acceptance, order release and scheduling decisions. The objective is to maximize the profit in a system where a penalty has to be paid whenever an order is completed after its due-date. The tardiness cost is proportional to the product of job tardiness and job revenue. They consider the EDD rule and the minimum slack per remaining operation first rule (S/OPN) to schedule jobs and state that S/OPN performs better than EDD. Moreira and Alves (2009) investigate order acceptance, due-date assignment, order release and scheduling decisions simultaneously. They model a job shop environment and compare different policies based on nine performance measures, including the due-date related measures of mean tardiness and percent tardy. Their results show that the EDD rule improves performance based on both of these measures. Van Foreest et al. (2010) analyze a single-server MTO system where arriving orders for different product families require deterministic production times and a fixed customer-required leadtime. They develop heuristic policies that accept an arriving order if it can be scheduled, considering that no order is allowed to be completed later than its due-date and that there are inter-family setup times. They compare several heuristic policies based on the resulting machine utilization rate.

2.1.2. Optimization

The research considering optimization of order selection decisions is divided into two categories: (1) order selection assuming FCFS discipline in order processing or (2) order selection in combination with scheduling decisions. Each category is further divided into two, according to the type of decision-making: static or dynamic. In each sub-category, we present studies applying the following sequence: order selection via accept/reject decisions, price quotation, leadtime quotation and joint price and leadtime quotation decisions.

2.1.2.1. Static order selection decisions assuming FCFS

Altendorfer and Minner (2015) investigate the optimal capacity investment under different order selection policies that are in the form of direct accept/reject decisions. They assume exogenously determined stochastic customer-required leadtimes. In the first policy, the decision is made by the customer who expects the production leadtime that guarantees an on-time probability of η to be less than the time he is willing to wait. A second policy considers the manufacturer as a decision maker. He accepts or rejects an order taking the current system state and the targeted overall on-time probability into account. The authors show that in the latter case, optimizing capacity investment and order acceptance policy simultaneously provides a high cost saving potential. Chatterjee et al. (2002) investigate order selection by means of leadtime quotation. They consider a firm that offers delivery guarantees to its customers, i.e., a certain amount of reimbursement when the order is delivered later than the promised due-date. They analyze and give real-life examples for, the following two types of reimbursements: (1) an amount that is proportional to tardiness (2) a fixed amount no matter how late the delivery is. The marketing department of the firm makes leadtime quotes facing the trade-off between promising short leadtimes to attract customers and paying high tardiness penalties. They show that when unit tardiness cost is independent of the processing time, there is a critical processing time value above which it is optimal to quote a leadtime of zero. This also holds for the case of a fixed tardiness cost.

Pekgün et al. (2008) solve the problem of price and leadtime quotation under service level constraints in a setting where these decisions are not made by a single decision maker. The decentralized case, in which marketing chooses a price and manufacturing chooses a leadtime optimizing their own objectives, is compared to the centralized case, in which price and leadtime quotes are optimized simultaneously. They demonstrate the inefficiencies that result from decentralized decision-making and find that in a decentralized setting, firm's profits are lower although the total

generated demand is higher. Zhao et al. (2012) also investigate the problem of price and leadtime quotation under service level constraints. In their study, the customers are divided into two categories: price- and leadtime-sensitive customers. They compare the uniform quotation approach, in which the firm uses a single price/leadtime pair, with the differentiated quotation approach in which the firm offers a menu of price/leadtime pairs. A uniform or a differentiated quotation approach being more beneficial may depend on several parameters such as the proportion of leadtime sensitive customers and the desired service level.

Palaka et al. (1998) and Ray and Jewkes (2004) investigate price and leadtime quotation in combination with capacity decisions. The former study shows that the capacity utilization should be lower when customer leadtime sensitivity and/or congestion related costs at the firm and/or the lateness penalty is higher. In the latter study, price is not modeled as a decision variable but as a function of the quoted leadtime. The study finds that under some conditions the decisions are substantially sub-optimal if the relationship between price and delivery time is ignored.

2.1.2.2. Dynamic order selection decisions assuming FCFS

Defregger and Kuhn (2007) use a discrete time Markov decision model for the optimization of dynamic order acceptance decisions in an MTO production system consisting of a machine and a finished goods inventory with limited capacity. The orders arrive stochastically with externally specified attributes such as the capacity usage, the maximum leadtime and the profit margin. The authors show that optimizing order accept/reject decisions, which allows reserving capacity for high margin orders, is a better strategy than accepting all orders as long as the maximum leadtime constraints can be met. Savaşaneril et al. (2010) investigate the leadtime quotation problem by modeling the system as an M/M/1 base-stock queue and formulating an MDP. They find that less sensitivity to leadtimes, increases the benefit of quoting longer but more reliable leadtimes and more sensitivity to leadtimes increases the benefit of holding inventories. Slotnick (2014) also uses an MDP model for investigating the leadtime quotation problem. She takes into account the long-term effects of delivery performance on the customers' decision on whether to accept or reject the quoted leadtime and suggests the quotation decisions to be made considering firm's past on-time delivery performance in addition to market characteristics.

2.1.2.3. Static order selection decisions in combination with scheduling

Easton and Moodie (1999) investigate the problem of price and leadtime quotation for the case where the time between firm's quote and customer's decision is not negligible. If a new customer arrives during this time, the firm makes the quote bearing a risk of incurring a tardiness cost due to available capacity being uncertain. They assume that obtained orders are processed based on a given rule , e.g., EDD. The probability that a customer accepts a price/leadtime pair is modeled using an S-shaped logistical response function. The tardiness penalty model regards terms for the probability of a tardy job and the expected amount of tardiness. However, their model ignores future customer arrivals. They show that their approach outperforms simple rules that estimate leadtimes based on minimum, maximum or expected shop load. Watanapa and Techanitisawad (2005) extend the work of Easton and Moodie (1999) to consider multiple customer classes and resequencing of orders. Their results show that employing an EDD sequencing rule for processing orders leads to an increased number of winning bids and higher bid prices, on the other hand, to a higher tardiness penalty per order because it results in some orders to be repeatedly postponed.

2.1.2.4. Dynamic order selection decisions in combination with scheduling

Germs and Van Foreest (2011) extend Van Foreest et al. (2010) and model the order acceptance and scheduling problem as an MDP. They use the optimal order acceptance and scheduling policy to benchmark the performance of simple heuristic policies. As opposed to this study, Chapters 3 and 4 present models that incorporate stochastic processing and customer-required leadtimes.

A group of studies approach the order selection and sequencing problem using approximation methods and investigate policies ensuring service within order leadtimes in a multi-product setting. In these studies, orders for each product join dedicated queues, the sequencing decision of interest is the allocation of the server effort among these queues and formulation of the leadtime constraints involve translation of maximum allowed leadtimes into maximum allowed number of orders in respective queues. The fundamental difference between models presented in Chapters 3 and 4 and the ones in this group is that the chapters consider the time in system in comparison to the quoted leadtime explicitly by keeping track of the remaining time of orders until the due-date. Maglaras and Van Mieghem (2005) and Ata (2006) consider the problem of order acceptance/rejection and sequencing assuming that order leadtimes are exogenously determined.

The study of Çelik and Maglaras (2008), which also belongs to the above mentioned group, investigates leadtime quotation and dynamic pricing decisions in combination with order sequencing and expediting. The authors consider a make-to-order firm that offers a menu of price/leadtime pairs in which prices are dynamically set and arriving customers decide which product to buy if any. They find that pricing decisions do not depend on the product-level queue lengths but on the aggregate system load and that when sequencing orders, priority should be given to the order closest to violating its leadtime. Charnsirisakskul et al. (2006) optimize order acceptance, pricing and scheduling (production quantity in each period) decisions simultaneously using a deterministic mixed integer model where quoted prices influence demanded quantities. In their model, the tardiness cost is proportional to the number of periods and the quantity. They find that when there is no inventory flexibility, leadtime flexibility becomes more useful and that price flexibility is often more useful than leadtime flexibility.

Plambeck (2004) considers two customer classes which differ in their willingness to pay and tolerance for delay. Prices for each customer class and the capacity (service rate) are up-front decided, while leadtimes are dynamically quoted. The objective is to maximize profit such that each order is processed within the quoted leadtime. They show that it is asymptotically optimal to prioritize the impatient customer class when allocating server effort. Ata and Olsen (2009) provide an approximating diffusion control problem for investigating the dynamic leadtime and price quotation decisions. They consider a monopoly and eliminate the dynamic pricing problem. The firm makes an up-front capacity decision and considers maximization of revenue minus capacity costs as the objective. They prove that a threshold policy is asymptotically optimal. Neither Plambeck (2004) nor Ata and Olsen (2009) address the possibility of tardiness in order completion. Slotnick (2011a) uses a finite horizon discrete time Markov decision model for investigating leadtime quotation decisions under minimum batch size requirements. Increased frequency of balking due to long leadtime quotes affects the waiting time of the already accepted orders by slowing down the rate at which the buckets are filled. She finds that shorter lead times should be quoted for an arriving order as the amount of that product increases in the system.

Duenyas and Hopp (1995) model the due-date quotation problem as a semi-Markov decision process, where the demand is sensitive to the quoted due-dates. They investigate the optimal due-date quotation and order scheduling problem with the objective of maximizing the long-run average profit. They show that, once the due-

dates are quoted, the optimal policy processes orders using the EDD rule when the tardiness penalty is proportional to tardiness. However, when the tardiness penalty is a fixed cost, this result does not hold. Chapter 3 presents an MDP model for optimization of the sequence in which the orders are processed with the objective of minimizing the long-run average cost resulting from the tardiness of orders. The results confirm the optimality of the EDD rule for the case where the penalty is proportional to tardiness. Furthermore, they shed light on the case where a fixed cost is involved in the tardiness penalty.

Duenyas (1995) investigates the joint problem of dynamic leadtime quotation and order sequencing considering multiple customer classes. Upon acceptance of the quote, the system manager places the order at any position in the queue. In case of tardiness, the firm incurs a proportional penalty. He states that the optimal policy processes customer orders according to the earliest-due-date-first principle when tardiness costs and processing time distributions are identical. For achieving higher profitability, one should employ leadtime quotation and order sequencing policies that take customer price and leadtime preferences into account. The model presented in Chapter 4 is similar to Duenyas's in several aspects such as the use of the theory for Markov decision processes, inclusion of the remaining time of orders until the due-date in the state description, the use of this information for order sequencing and the assumption of heterogeneous customers. To name the principal differences, Chapter 4 considers also a fixed cost in the tardiness penalty, dispatching decisions upon order completion and a dynamic price quotation. As noted by Öner Közen and Minner (2016c), whenever a fixed tardiness cost plays a role, order sequencing decisions made upon process completion dominate the ones made upon arrival.

Ata and Olsen (2013) consider two classes of customers and the problem of a firm offering an incentive compatible menu of price/leadtime pairs in which both components are dynamically set. Their model assumes that the production decisions are made at discrete points in time. They prove that a discrete time version of the $Gc\mu$ rule, where the customer class with the largest value of marginal cost times service rate is prioritized, is asymptotically optimal. As opposed to the models presented in Chapters 3 and 4, their model assumes that quoted leadtimes must be met.

2.2. Production lines with human operators

2.2.1. Performance evaluation

The complexities involved in production lines with stochastic operation times and finite buffer spaces make exact analyses difficult. They are generally limited to small system sizes, e.g., two- to three-stage systems with small buffer sizes. Nevertheless, various authors used exact methods for investigating the expected throughput of such lines. Lau (1986b), Lau (1986a), Rao (1975a) consider a two-stage production system with no buffers. Under these assumptions, the analysis of the expected throughput measure reduces to an analysis of the expected value of the maximum of two random variables. When the processing times in at least one of the workstations is assumed to be according to an exponential distribution, the case of larger buffer sizes can be analyzed by the analysis of equivalent queuing systems such as M/M/1/K, M/G/1/K, e.g., Rao (1975b). The extension of the analysis to processing times with phase-type distributions is also possible by using continuous time Markov models, e.g., Buzacott and Kostelski (1987), Berman (1982).

The expected throughput of a three-stage system with no buffers and exponentially distributed processing times is investigated analytically by Hunt (1956) also via a continuous time Markov chain model. Muth (1973) introduces a different approach, which is called the holding time model, for analyzing a production line with no buffers. Hillier and Boling (1967) provide an exact numerical model assuming exponential or Erlang distributed processing times. They model a multi-stage production line as a continuous time Markov chain. Their model can solve for the performance measures of systems up to six stages. Altıok (1985) also suggests a numerical solution for the problem. In contrast to Hillier and Boling (1967), he assumes that the machines are subject to breakdowns and the processing and repair times have phase-type distributions. Furthermore, Papadopoulos (1996), Knott (1970) and Blumenfeld (1990) conduct approximate analyses to obtain analytical expressions for the throughput. Knott (1970) and Blumenfeld (1990) assume a balanced line while Papadopoulos (1996) allows different mean processing times in workstations of a K-station production line with exponentially distributed service times and no buffer spaces.

Another stream of research focuses on developing models for investigating higher moments of the inter-completion time distribution as well as the distribution itself. Hendricks (1992) models a serial production line with exponential processing times and investigates the variability of line's output process as well as the effect of buffer

allocation decisions on this measure. Lagershausen and Tan (2015), who introduce a method for determining the inter-event time distributions of jobs in queuing networks that can be modeled as a continuous time Markov chain, present a summary of the research in this direction.

2.2.2. Comparison of paced and unpaced configurations

For single-stage work tasks, the unpaced work rate represents the upper bound on output, which means that no operator can perform better in paced than in unpaced conditions (Murrell (1972); Sury (1965)). If two or more operators are connected to form an assembly line, the output of that line is no longer just dependent on the performance of any single operator, but also on the interaction of operators in the group. For a 3-station line, Davis (1965) analytically finds that the time between two finished items in paced conditions is never as low as in unpaced conditions, and that modifications of the line that tend to decouple the worker from the pacing mechanism enhance line performance. Sury (1971) confirms that considerable output potential is lost when operators are linked in a paced assembly line. The reason for this phenomenon is that, in unpaced conditions, workers can already start their work on a new item before their co-workers have finished theirs (Muth and Alkaff, 1987).

2.2.3. Optimization

Hillier and Boling (1966) analyze short production lines (up to four stations) with exponentially distributed processing times and conjecture that it is optimal to allocate the mean processing times in such a way that workstations near the two ends of the line are assigned with higher values compared to the ones toward the center of the line, when buffer spaces are uniformly distributed. Hillier and Boling (1977) approach the workload allocation problem theoretically and state that a bowl phenomenon applies in a system for which the three properties (reversibility, symmetricity and monotonicity) can be shown to hold. Hillier and Boling (1979) investigate the optimal allocation of workload in a production system up to six stages and Erlang distributions. Their study confirms the bowl phenomenon. It provides results on the effect of line length, available buffer spaces and the variance of processing times.

Rao (1976) investigates the effect of variability imbalance between different production stages on the optimal workload allocation. His results show that in some cases, variability imbalance overweights the bowl phenomenon observed in mean

processing times. The optimal workload allocation is balanced when the middle stage has a coefficient of variation of 0.5 and reverse bowl-shaped when the middle stage is deterministic. Pike and Martinj (1994) study longer lines (longer than 30 workstations) with small buffer spaces (up to one unit) and state that a bowl-shaped allocation leads to higher throughput compared to a balanced allocation. They further find that an optimal two-level allocation of workload performs as good as an optimal multi-level workload allocation. Moreover, they observe that the bowl allocation is robust, i.e., it is superior to balanced even when the degree of imbalance differs from the optimal. They find that allocating the workload according to the bowl phenomenon is not beneficial when the coefficient of variation is lower and the buffer sizes are larger. Hillier and So (1996) also focus on the robustness of the bowl phenomenon. They show that it is possible to obtain 95% of the potential gain provided by the optimal bowl-shaped allocation even when the amount of unbalance differs by 25%. When the optimal amount of unbalance is unclear, it is suggested to err on the small side.

Hillier et al. (1993) study the optimal allocation of buffer spaces given an equal allocation of the workload and observe the storage bowl phenomenon in the optimal solution where larger spaces are allocated to the interior buffers. Hillier and So (1995) investigate three different design problems (workload allocation, buffer allocation and server allocation) separately and in combination. They suggest determining first the server, then the buffer and finally the workload allocation as a heuristic scheme. Recently, Hillier (2013) studied the unpaced line design problem considering maximization of profit (revenue minus inventory holding cost) as the objective while optimizing workload and buffer space allocation simultaneously. He finds that when the holding costs due to work-in-process inventory are taken into account, the optimal workload and buffer allocation guidelines differ significantly from the guidelines for maximizing throughput. Moreover, the optimization of the workload allocation is found to have a significantly larger impact than the optimization of the buffer allocations.

2.2.4. Human aspect and its incorporation into production line design

Because human beings are commonly used as resources in production systems, understanding the nature of human work is important when investigating decisions pertaining to design of assembly lines. However, it is not an easy task since it is relevant to research in several disciplines such as ergonomics and psychology. Existing research appreciates human aspects such as learning, performance heterogeneity, ergonomics as well as human behavior.

Folgado et al. (2015) conduct two empirical studies for assessing the influence of worker heterogeneity on line output by using industrial data collected from an assembly line that produces automotive components. They study two different pacing mechanisms. One system imposes a fixed takt time ("rigid pacing", Murrell (1972)) while the other paces workers by an hourly production target ("system with margins", Murrell (1972)). In the latter case, they find that slower workers show higher variability than faster workers. The difference in workers' performance disappears in the rigid system, which leads to a 13% higher output in their study.

Carnahan et al. (2001), Otto and Scholl (2011) and Battini et al. (2016) investigate the assembly line balancing problem (ALBP) after incorporating the ergonomic aspect into it. Particularly, they criticize optimizing task allocation decisions solely from the economical view point, because this might lead to severe consequences for operators' physical well-being. They propose Pareto optimization, in other words, using an objective function that combines economic and ergonomic aspects. Carnahan et al. (2001) use an objective function which equally weights fatigue and cycle time, Otto and Scholl (2011) investigate the trade-off between the number of workstations and ergonomic risks while Battini et al. (2016) analyze the trade-off between time smoothness and energy smoothness of task allocations. Battini et al. (2011) develop an integrated framework for assembly systems design that simultaneously takes technological variables such as assembly times and ergonomics variables such as human diversity into account. Applying this framework to two industrial case studies, they show an increase in productivity by up to 15% while lowering fatigue levels and injuries.

Dode et al. (2016) integrate worker fatigue and learning effects into their simulation models and propose a line design for consumer electronics production by taking human factors into account. The proposed line allows up to 33% lower fatigue dosage compared to the existing line. Furthermore, a model that accounts for human learning, estimates 10.5% higher output compared to a model that ignores this effect. Neumann and Medbo (2016) also take human learning into account when comparing two types of assembly lines based on throughput during ramp-up: a serial and a parallel flow line. They find that serial lines facilitate faster learning and a shorter ramp-up time, however, the latter flow type overtakes the former at a point in time, due to providing higher throughput potential.

Furthermore, there is a considerable amount of literature discussing the human reaction to the flow of work in their immediate environment when performing tasks

from the behavioral perspective. Edie (1954) states that the processing times decrease with the amount of congestion. The studies by Doerr et al. (1996) and Schultz et al. (1998) suggest that the average processing time of workers are shorter in low inventory systems. Schultz et al. (1999) focus on psychological aspects such as goal setting, feedback and peer pressure for predicting how the workers adjust themselves and for explaining the previous findings about low inventory systems. Hertel et al. (2000) show that increasing difference between the abilities of the two peers increases the effort of the weaker member. Schultz et al. (2003) assert that visible feedback on the performance increases the pace at which the workers operate. Kc and Terwiesch (2009) use data from two health-care delivery services and show that the rate at which workers provide service increases with the system load. However, if the system remains highly loaded for a long duration, the service rate decreases. By using real-world manufacturing data, Schultz et al. (2010) show that there is a reaction by workers to the speed of their co-workers, which varies from one to the other. Other studies that suggest a dependency between a worker's behavior and his coworkers include; Falk and Ichino (2006), Mas and Moretti (2009), Siemsen et al. (2007), Gould and Winter (2009).

However, only a small number of studies use these findings to model human behavior more realistically when analyzing production lines. An important finding by the study of Schultz et al. (1998) is that the lines with low inventory can be as efficient as high inventory lines due to the adaptive behavior of workers. Powell and Schultz (2004) investigate the relationship between the line length and the throughput of a production system under the assumption that workers adjust their speeds dependent on the system state. Their findings show that the lines are more efficient in the existence of this behavior. Furthermore, the efficiency deterioration due to increasing the number of stages in a production line is not as large as estimated by studies ignoring this behavioral effect. Heimbach et al. (2012) consider the workload allocation problem and study the effect of state-dependent behavior on the optimal allocation of workload. The shape of the optimal allocation changes with the speed adjustment factor. A bowl-shaped allocation is observed for the case where the speed adjustment is zero, whereas a balanced or a reverse bowl-shaped allocation is observed for moderate and large values of the adjustment parameter. The last two studies can be criticized for modeling workers as stationary resources which are always capable of increasing their speed up-to the required amount as long as the condition for speed-up is satisfied. This turns out to be a strong assumption especially when workload allocation problem is considered. In Heimbach et al. (2012), the results for high values of the speed-up parameter ($f = 0.9$) suggest nearly all the workload to be assigned to the worker in the middle station in a three-stage

system when the inter-stage buffers are small (single buffer space). Schultz et al. (2010) consider the problem of optimizing the order of workers in a production line with the objective of maximizing the worker output. They suggest that the workers should be ordered from fastest to slowest in such a way that each worker can only see the faster one in front.

Chapter 5 compares paced and unpaced assembly lines using a simulation model while Chapter 6 investigates the optimal line design problem by introducing an exact model. The state-dependent worker behavior is incorporated into both models in combination with fatigue.

3. Impact of Priority Sequencing Decisions on On-Time Probability and Expected Tardiness of Orders in MTO Production Systems with External Due-Dates

We model the priority sequencing problem in a make-to-order (MTO) production system where the customers specify the amount of time they are willing to wait for their orders to be fulfilled as a Markov decision process (MDP). The objective function is the sum of a fixed and a variable cost of tardiness that combines two external customer-related criteria: "on-time probability of orders" and "expected tardiness of orders". We benchmark several simple rules against the optimal policy and analyze the efficient frontier of on-time probability and expected tardiness. The numerical results show that it is possible to obtain near optimal performance by employing simple rules. Whenever a fixed cost of tardiness is involved, the optimal priority sequencing policy deviates from the earliest-due-date (EDD) principle, however, an adjusted EDD rule performs well. Furthermore, postponement of priority sequencing decisions until the next completion improves performance.

3.1. Introduction

In this chapter, we consider an MTO environment where orders arrive dynamically following a stochastic process with stochastic customer-required leadtimes. The customer-required leadtime is the amount of time a customer is willing to wait for his order. We assume that whenever an order is not completed within this time window, the firm bears a penalty cost due to tardiness occurrence and duration. By means of making priority sequencing decisions, the firm aims at minimizing the tardiness penalties.

This problem finds applications, for example, in the capital goods industry where the production is inevitably customer-specific and it is crucial to meet customers' delivery time expectations. Based on the results of a customer survey, American National Tooling and Machining Association (NTMA) state that "schedule and delivery problems" is the most frequently cited problem type by the customers and that the firms "...wanting to stay competitive should devote much attention to scheduling and timely delivery among all the other fundamentals...". In addition, late deliveries might mean additional costs for a firm because in the delivery agreements, customers commonly include clauses such as "Time is of the essence for this

order." (www.ntma.org). In the capital goods industry, the production leadtime, as well as, the lifespan of finished products are very long and typically measured in years. It might be more convenient in such a production environment to define the order arrival and completion processes based on probabilities of occurrences in a period (e.g. a month) because for firms, estimating these probabilities might be easier compared to estimating the rates required by continuous models. This chapter offers a discrete time model which allows directly using the probability estimations. Discrete time analysis is commonly used in computer and telecommunications systems' analysis as well as production systems' analysis (see e.g. Schleyer and Furmans (2007), Jolai et al. (2008), Artalejo et al. (2008)). Another important aspect is that the raw material and/or subassemblies needed for producing the final good, e.g. a building, machinery or an aircraft, might be very expensive and/or space consuming. Therefore, it might not be possible for a firm to carry a large number of pending orders. In this model, we incorporate an upper-bound on the number of orders that can be accommodated in the system.

In general, the tardiness penalties are not limited to costs due to contractual obligations. They may include penalties such as loss of goodwill, potential loss of future business and discounts if the firm commits to a delivery guarantee, i.e., offers a price discount when the product is not delivered before or on the due-date. Chatterjee et al. (2002) provide several real-life examples of delivery guarantees involving two types of price discounts: a fixed amount no matter how tardy the delivery is and an amount that increases with tardiness. Most of the literature models the tardiness penalty as a variable cost that is proportional to the amount of tardiness. Exceptions include Easton and Moodie (1999) who consider a tardiness penalty that includes terms for the probability of a tardy job and the expected amount of tardiness.

The customer-related performance of the system, along with the cost optimal sequencing of orders, strongly depends on how tardiness of orders is penalized. If the penalty is linearly proportional to tardiness, the cost minimizing decisions should result in the minimum "expected tardiness of orders". On the other hand, if the tardiness penalty is a fixed cost, the cost minimizing decisions should result in the maximum "on-time probability of orders". Also note that a sequence of decisions that improves one of the measures does not necessarily improve the performance based on the other measure. For instance, the orders for which the due-date has already been reached might be allocated with the least priority for improving the on-time probability measure. However, this increases the amount of tardiness for such orders and might worsen the expected tardiness measure. We consider the

tardiness penalty to be the sum of a fixed and a variable cost. Thus, we combine the two widely investigated performance criteria for generating managerial insights into the decisions that improve one of the two measures while compromising on the other. A decision maker can input the value of two costs, which are determined in accordance with the customers' expectations/preferences, to the model. In the extreme case, if only one of the performance criteria (e.g. expected tardiness) is important to customers, the decision maker can set the weight of the other criterion to zero (e.g. set the fixed cost to zero) and reduce the analysis to a single criterion optimization.

Whenever a fixed tardiness cost plays a role, we also need to ask the following question: How much benefit is possible by delaying priority sequencing decisions to the next order completion instead of deciding upon arrival? For this purpose, we consider two different decision-making strategies. Under the first strategy, the decision regarding the initial position of an arriving order in the queue of orders is made upon each arrival. Under the second strategy, the decision about which order to process next is made upon the completion of an order. Moreover, due to the existence of a fixed penalty, simple rules other than the earliest-due-date-first (EDD) rule need to be investigated. We consider several simple rules and propose a new rule for determining the processing sequence of orders.

In order to isolate the effect of priority sequencing decisions from other possible effects in a complex system, we consider the firm as a single aggregate unit serving orders. To this end, we utilize a single server model as commonly done in the related literature (see e.g. Baker and Bertrand (1982), Wein (1991), Duenyas and Hopp (1995), Germs and Van Foreest (2011)).

Section 3.2 presents the Markov decision model. Section 3.3 gives expressions for calculating the internal and customer-related performance measures of the optimal policy obtained from the MDP. It also covers the special case where the priority sequencing decisions are made in accordance with simple rules. In Section 3.4, results of the numerical study are reported. Section 3.5 summarizes findings and discusses limitations of this work.

3.2. Markov decision model for priority sequencing

We assume that there is a single server to process arriving orders. There is a limited waiting space. The number of orders the system can accept is K. All arriving orders are accepted as long as K is not reached. The stochastic arrival and service process

evolves as follows. Time is divided into sufficiently small discrete periods so that in each period only one of the three possible events can happen with associated probabilities: arrival of an order with probability γ, completion of an order (i.e. departure) with probability β, no order arrival or completion with probability $\theta = 1 - \gamma - \beta$. Under these assumptions, the inter-arrival and processing times of orders can be described as geometric random variables with parameters γ and β, respectively.

The amount of time, i.e. the number of periods that the customers are willing to wait for their orders, is called the customer-required leadtime (CRL). It is unknown before and realized upon arrival. It can take integer values with a minimum value of one. Since the arrival and completion of an order cannot both occur in the same period, a customer-required leadtime of zero is not considered. We assume finite support and denote the highest possible customer-required leadtime CRL_{max}. The due-date of an order is calculated as the sum of its arrival time and the customer-required leadtime. As soon as the due-date of an order is determined, it is *fixed*. The remaining time until the due-date decreases as time moves forward and is called the remaining leadtime of orders. A tardiness penalty is incurred whenever an order is completed after its due-date. The fixed cost is incurred if the processing of an order is completed later than its due-date, regardless of the amount of tardiness, whereas the variable cost increases linearly with the amount of tardiness.

The priority sequencing decisions are made based on the information about the due-dates, equivalently, based on the remaining leadtimes of orders. Preemption of the order that is being processed is not allowed. The following two decision-making strategies are under consideration:

- PA (Position Arriving Order): The decision about the initial position of the arriving order in the queue of orders is made upon each arrival. This strategy models the case where the firm needs to make a decision upon order arrival and it is considered e.g. by Duenyas and Hopp (1995).
- DC (Decision upon Completion): The decision which order to process next is made upon the completion of an order. This strategy models the case where the firm has the flexibility to release itself from making a decision upon arrival and maintaining an ordered queue.

We assume an infinite planning horizon and model this system as a discrete time Markov decision process to find the priority sequencing policy which minimizes the long-run average tardiness cost per time unit.

3.2.1. State space

A discrete time Markov decision model takes each period as a decision epoch. However, when the decision-making strategy is PA (DC), only the periods with arrival (completion) are decision epochs. As a consequence, the state description needs to include an indicator variable to distinguish between the periods with and without an arrival (a completion). For modeling the PA strategy, the information on the CRL of the arriving order is also needed.

The state of the MDP is denoted as $(e, n, r_1, r_2, ..., r_n)$ where e provides information about the event occurring in that period, and in case of an arrival, CRL of the arriving order. $e = -1$ indicates "completion", $e = 0$ indicates "no arrival or completion" and $e = 1, 2, ..., CRL_{max}$ indicates "arrival" with $CRL = e$. n represents the number of orders that the arriving order *finds* in the system when $e > 0$ and the number of orders in the system after the event (e.g. completion) otherwise. r_i, $i = 1, 2, ..., n$ denote the remaining leadtime of the i^{th} order. We consider r_1 as the remaining leadtime of the order that is currently being served. Under the first decision-making strategy, these orders are kept in sequence so that upon service completion, the order in the second position automatically moves to the first position (to server).

The set of all possible states, i.e. the state space, is denoted by S.

3.2.2. Action space

Let us denote the set of all possible decisions in state s by $K(s)$. Under the PA strategy, with n orders already in the system, an arriving order can be placed at any position between 2 and $n + 1$. The position "1" is not an option because it represents the order in process and the processing of orders is assumed to be non-preemptive. We denote the set of all states with an order arrival which finds n orders in the system by $S^{e>0,n}$. Then, $K(s) = \{2, ..., n + 1\}$, $\forall s \in S^{e>0,n}$ and $K(s) = \{0\}$ otherwise.

Under the DC strategy, when the completed order leaves n orders behind, the order to be processed next, i.e. the order to be placed at position 1, is selected among these n orders. Let the set of all states with an order completion that leaves n orders behind be denoted by $S^{e=-1,n}$. Then, $K(s) = \{1, ..., n\}$, $\forall s \in S^{e=-1,n}$ and $K(s) = \{0\}$ otherwise.

3.2.3. Transition probabilities

The probability that a certain state is visited at period t depends on the decision made in the state at the previous period. We define $p_{s',s}(k)$ as the probability that at a decision epoch the system will be in state s if decision k was made at the previous period in state s', $k \in K(s')$.

Under the PA strategy, the probabilities of transition to an arbitrary state $s = (e, n, r_1, ..., r_n)$ from all possible previous states can be written as follows.

- For "arrival" periods where $e > 0$:
  The transition probability to state s is equal to the probability that an order with a CRL of e arrives ($\gamma P\{CRL = e\}$) or equal to 0, dependent on the state in the previous period (s').
  - If no arrival occurred in the previous period:

$$\begin{aligned} &p_{s',s}(k) = \gamma P\{CRL = e\} \\ &s' = (e', n, r_1 + 1, ..., r_n + 1), e' \in \{-1, 0\} \\ &k = 0, n < K \end{aligned} \tag{3.1}$$

    In words, the transition probability to state s is equal to $\gamma P\{CRL = e\}$ if s' is a state where there are n orders each with one more period until the due-date.

  - If the previous period was an arrival period:

$$\begin{aligned} &p_{s',s}(k) = \begin{cases} \gamma P\{CRL = e\} & r_k \geq 0 \\ 0 & r_k < 0 \end{cases} \\ &s' = (r_k + 1, n - 1, r_1 + 1, ..., r_{k-1} + 1, r_{k+1} + 1, ..., r_n + 1) \\ &k \in \{2, ..., n\}, 1 < n < K \end{aligned} \tag{3.2}$$

    In words, the transition probability to state s is equal to $\gamma P\{CRL = e\}$ if s' is a state where there are $n - 1$ existing orders, the CRL of the arriving order is equal to $r_k + 1$ and the decision k was made. Transition from s' to state $s = (e, n, r_1, ..., r_n)$ by making decision k is not possible for $r_k < 0$ because the CRL of an arriving order is greater than or equal to 1 by definition. Note that this can be the case when an order remains in the system after its due-date is reached.

- For "no arrival" periods where $e \leq 0$:

The transition probability to state s is equal to β, θ or 0, dependent on the state in the previous period (s').

- If no arrival occurred in the previous period:

$$\begin{aligned} & p_{s',s}(k) = \beta \\ & s' = (e', n+1, r, r_1+1, ..., r_n+1), e' \in \{-1, 0\} \\ & k = 0, n < K, r \leq CRL_{max} - 1 \end{aligned} \tag{3.3}$$

The transition probability to state s is equal to β if s' is a state where there are n+1 orders with remaining leadtimes r, r_1+1, ..., r_n+1 for any $r \leq CRL_{max}$−1.

$$\begin{aligned} & p_{s',s}(k) = \theta \\ & s' = (e', n, r_1+1, ..., r_n+1), e' \in \{-1, 0\} \\ & k = 0 \end{aligned} \tag{3.4}$$

The transition probability to state s is equal to θ if s' is a state where there are n orders each with one more period until the due-date.

- If the previous period was an arrival period:

$$\begin{aligned} & p_{s',s}(k+1) = \begin{cases} \beta & r_k \geq 0 \\ 0 & r_k < 0 \end{cases} \\ & s' = (r_k+1, n, r, r_1+1, ..., r_{k-1}+1, r_{k+1}+1, ..., r_n+1) \\ & k \in \{1, ..., n\}, 0 < n < K, r \leq CRL_{max} - 1 \end{aligned} \tag{3.5}$$

In words, the transition probability to state s is equal to β if s' is a state in which there are n existing orders, the CRL of the arriving order is equal to $r_k + 1$ and the decision $k + 1$ is made.

Finally,

$$\begin{aligned} & p_{s',s}(k) = \begin{cases} \theta & r_k \geq 0 \\ 0 & r_k < 0 \end{cases} \\ & s' = (r_k+1, n-1, r_1+1, ..., r_{k-1}+1, r_{k+1}+1, ..., r_n+1) \\ & k \in \{2, ..., n\}, n > 1 \end{aligned} \tag{3.6}$$

The transition probability to state s is equal to θ if s' is a state in which there are $n - 1$ existing orders, the CRL of the arriving order is equal to $r_k + 1$ and the decision k was made.

The transition probabilities under the DC strategy can be obtained using the same logic and considering "completion" and "no completion" periods.

3.2.4. Tardiness penalties

Each state has a corresponding cost summarized by the following expression.

$$\begin{aligned} cost_s = &\text{ (Number of orders with } r_i = 0)\cdot\text{(fixed cost)} \\ &+ \text{(Number of orders with } r_i \leq 0)\cdot\text{(unit variable cost)} \qquad \forall s \in S \end{aligned} \tag{3.7}$$

For an order, a fixed cost is incurred at the period in which its remaining leadtime becomes zero ($r = 0$). In addition, a unit variable cost applies at the period in which $r = 0$ and at each of the following periods during which the order remains in the system. This is because s describes the state of the system after the event and at the period in which an order is completed, r is not represented in s. Thus, for an order that is completed, for example, 3 periods later than its due-date, the unit variable cost is incurred 3 times, from the period with $r = 0$ to the period with $r = -2$.

3.2.5. Computing the optimal policy

In order to obtain the optimal policy, we solve the following linear program. (Tijms (2003, Chapter 6)).

$$\min \sum_{s \in S} \sum_{k \in K(s)} cost_s \cdot x_{s,k} \tag{3.8}$$

subject to

$$\sum_{k \in K(s)} x_{s,k} = \sum_{s' \in S} \sum_{k \in K(s')} x_{s',k} \cdot p_{s',s}(k) \qquad \forall s \in S \tag{3.9}$$

$$\sum_{s \in S} \sum_{k \in K(s)} x_{s,k} = 1 \tag{3.10}$$

$$x_{s,k} \geq 0 \qquad \forall s \in S, \forall k \in K\,(s) \tag{3.11}$$

The decision variable $x_{s,k}$ is the long-run fraction of decision epochs at which the system is in state s and decision k is made. $cost_s$ indicates the cost of visiting state s given by (3.7). Note that there is no direct cost associated with making the decision $k \in K\,(s)$ in state s at the moment of decision. However, under a stationary policy, which makes the decision $k \in K\,(s)$ whenever the system is in state s, the decisions determine the steady-state frequency of visiting each state by

acting on the transition probabilities. Thus, they are determinants of the long-run average cost.

We implement this linear program in Xpress-MP 7.9. In the implementation, it is not necessary to distinguish between the cases where $e = 0$ or $e = -1$ when PA strategy is considered because a decision is made in neither case. Therefore, we use $e = 0$ to indicate a dummy decision period. Under the DC strategy, a decision is made only in states where $e = -1$. Similarly, we use $e = 0$ to indicate a dummy decision period and do not distinguish between $e = 0, 1, 2, ..., CRL_{max}$. Furthermore, there is no lower bound for the remaining leadtime of an existing order. It can take any integer value less than or equal to CRL_{max}. When modeling remaining leadtimes, we use $r_i = -1$ to represent a negative remaining leadtime for the i^{th} order in the system. The cost structure allows this since a unit variable cost applies for an order at each period after the due-date is reached.

The advantage of using the linear programming approach is that it provides not only the value of the objective function, but also the steady-state probabilities of visiting each state under the optimal policy. Once the optimal values of the decision variables are obtained, the calculation of the performance measures under the optimal policy, for which formulas are given in the next section, takes place as a post processing. Due to the use of $r = -1$ in representing $r < 0$, we obtain the probability of tardiness of an order to be equal to l by calculating the probability of the order visiting the state $r = -1$ for the l^{th} time.

3.3. System performance based on internal and customer-related measures

3.3.1. Performance under the optimal policy

The optimal long-run average cost obtained by solving the MDP provides information about the tardiness-related system performance. However, this might not be the only measure of interest. After all, information regarding e.g. the expected response time (production leadtime) or amount of tardiness a company provides to its customers has managerial value. Thus, in this section, we focus on performance evaluation of the optimal policy based on the key internal and customer-related performance indicators, namely, the expected number of orders in the system, the expected production leadtime, the on-time probability and the expected tardiness of orders. In the following, we give expressions for calculating these measures using the optimal values of $x_{s,k}$'s.

3.3.1.1. Expected number and production leadtime of orders

The steady-state probability of the number of orders in the system being n (π_n) is equal to the sum of the long-run fraction of decision epochs at which the system is in a state where there is a total of n orders.

$$\pi_n = \begin{cases} \sum_{s \in S^{e>0,n-1}} \sum_{k \in K(s)} x_{s,k} + \sum_{s \in S^{e \leq 0,n}} \sum_{k \in K(s)} x_{s,k} & n \in \{1, ..., K\} \\ \sum_{s \in S^{e \leq 0,n}} \sum_{k \in K(s)} x_{s,k} & n = 0 \end{cases} \tag{3.12}$$

$S^{e \leq 0,n}$ is the set of all states with an order completion that leaves n orders behind or with no arrival and completion having n orders in the system. Using the well-known relationship provided by the queuing theory (e.g. Gross et al. (2008, Chapter 2)), we can obtain the expected production leadtime of orders (W).

$$E[W] = \frac{\sum_{n=0}^{K} \pi_n}{\gamma \cdot (1 - \pi_K)} \tag{3.13}$$

Note that no information about processing times of orders is utilized when priority sequencing decisions are made. Thus, the expected number of orders in the system and the expected production leadtime are the same under all priority sequencing policies.

3.3.1.2. On-time probability and expected tardiness

Given that a completion occurs in an arbitrary period t, the probability that the completed order is on-time is given by the probability that in period $t-1$ the remaining leadtime of the order in process is one or larger. Considering that a completion can only occur when the system is not idle (with probability $1 - \pi_0$) and that a decision made under the PA strategy does not affect the order in position 1, the on-time probability (η) can be obtained by using the following relationship,

$$\eta_{PA} = \frac{\sum_{s \in S} \sum_{k \in K(s)} x_{s,k} \cdot 1_{\{r_1 \geq 1\}}}{1 - \pi_0} \tag{3.14}$$

where

$$1_{\{r_1 \geq 1\}} = \begin{cases} 1 & r_1 \geq 1 \text{ in state } s \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

Under the DC strategy, if period $t-1$ is a decision period in state $(-1, n, r_1, r_2, ..., r_n)$ and decision k is made, the remaining leadtime of the completed order at time t is $r_k - 1$. Thus, in this case real and dummy decision periods should be distinguished.

$$\eta_{DC} = \frac{\sum_{s \in S^{e \geq 0}} \sum_{k \in K(s)} x_{s,k} \cdot 1_{\{r_1 \geq 1\}} + \sum_{s \in S^{e=-1}} \sum_{k \in K(s)} x_{s,k} \cdot 1_{\{r_k \geq 1\}}}{1 - \pi_0} \tag{3.16}$$

$S^{e=-1}$ is the set of all states with an order completion. $S^{e \geq 0}$ is the set of all states with an order arrival or with no arrival/completion.

Let T denote the tardiness of an order. Then, $P\{T = 0\}$ is equal to η, because the on-time probability gives the probability that an order is completed with no tardiness. The probability that the completed order at time t is late by l periods is equal to the probability of the order in process at time $t-1$ having a remaining leadtime of $1-l$. Thus, the probability distribution of tardiness is given as follows.

$$P\{T_{PA} = l\} = \begin{cases} \eta_{PA} & l = 0 \\ \frac{\sum_{s \in S} \sum_{k \in K(s)} x_{s,k} \cdot 1_{\{r_1 = 1-l\}}}{1-\pi_0} & l = 1, 2, ..., \infty \end{cases} \tag{3.17}$$

Similar to the on-time probability, r_k needs to be taken into account for the decision periods where decision $k \in K(s)$ is made under the DC strategy.

$$P\{T_{DC} = l\} = \begin{cases} \eta_{DC} & l = 0 \\ \frac{\sum_{s \in S^{e \geq 0}} \sum_{k \in K(s)} x_{s,k} \cdot 1_{\{r_1 = 1-l\}} + \sum_{s \in S^{e=-1}} \sum_{k \in K(s)} x_{s,k} \cdot 1_{\{r_k = 1-l\}}}{1-\pi_0} & l = 1, 2, ..., \infty \end{cases} \tag{3.18}$$

3.3.2. Performance under simple rules

The MDP model can also serve as a performance evaluation tool for simple rules when run with additional constraints that fix the priority sequencing decision k in each state according to the rule. In this case the linear program given in Section 3.2.5 reduces to a system of equations from which $x_{s,k}$'s can be obtained. When this is combined with the formulas given in Section 3.3.1, the expected number of orders in the system, the expected production leadtime, the on-time probability and the expected tardiness of orders under a given rule can be evaluated.

Since the literature suggests that the most commonly used priority sequencing rules in practice are due-date-based and the EDD rule is often employed (Keskinocak and Tayur (2004), Hopp and Spearman (2001), Rajendran and Holthaus (1999)), we consider EDD and other due-date-based rules that are essentially its variants, in addition to the first-come-first-served (FCFS) discipline.

3.3.2.1. Queuing model under FCFS discipline

The first-come-first-served (FCFS) discipline is frequently used for benchmarking the performance of priority sequencing policies. The performance measures of the queuing system working under the FCFS rule can be obtained by restricting $K(s) = \{n+1\}\ \forall s \in S^{e>0,n}$ in the MDP. Also, analytical expressions can be obtained for this basic rule. In the following, we give the probability distribution of waiting time in the queue (W^f_{queue}) and production leadtime (W^f).

The waiting time of an arriving order in the queue is the sum of processing times of orders in the queue at the time of arrival and the remaining processing time of the order in process. We know that the processing time of a single order is geometrically distributed and, because of the memoryless property of a geometric random variable (Nelson (1995, Chapter 4)), this also holds for the remaining processing time of the order in process. Thus, when an arriving order finds n orders in the system (including the order in process), its waiting time in the queue is the sum of n geometric random variables. This is a negative binomial random variable with parameters (n, β). When an arriving order finds the system idle ($N_{system} = 0$), it moves directly to processing. Then, $P\{W^f_{queue} = 0\} = \pi_0$ and $P\{W^f_{queue} = t\}$ for $t > 0$ can be written as follows.

$$\begin{aligned} P\{W^f_{queue} = t'\} &= \sum_{n=1}^{\min\{t', K-1\}} P\{N_{system} = n\} P\{W^f_{queue} = t' \mid N_{system} = n\} \\ &= \sum_{n=1}^{\min\{t', K-1\}} \pi_n \binom{t'-1}{n-1} (1-\beta)^{t'-n} \beta^n \end{aligned} \tag{3.19}$$

where

$$\pi_n = \frac{(\gamma/\beta)^n (1-\gamma/\beta)}{1-(\gamma/\beta)^{K+1}} \qquad \gamma \neq \beta.$$

In (3.19), the summation over the number of orders (n) goes up to $\min\{t', K-1\}$. This is because the processing of an order takes at least one period, thus, $P\{W^f_{queue} = t' \mid N_{system} = n\} = 0$, when n is larger than t'.

The expected waiting time in the queue is:

$$E\left[W^f_{queue}\right] = \frac{1}{\beta}\left(\frac{\gamma/\beta}{1-\gamma/\beta} - \frac{K\left(\gamma/\beta\right)^K}{1-\left(\gamma/\beta\right)^K}\right). \tag{3.20}$$

The production leadtime is the sum of the waiting time in the queue and the processing time of the order itself. Therefore, we obtain the number of periods (trials) required to have the $(n+1)^{st}$ completion (a negative binomial random variable with parameters $(n+1,\beta)$), given that the arriving order finds n orders in the system, $N_{system} = n$. Due to the fact that an arrival and completion cannot occur in the same period, the minimum value for production leadtime is one.

$$\begin{aligned} P\{W^f = t'\} &= \sum_{n=0}^{\min\{t'-1,K-1\}} P\{N_{system} = n\} P\{W^f = t' \mid N_{system} = n\} \\ &= \sum_{n=0}^{\min\{t'-1,K-1\}} \pi_n \binom{t'-1}{n} (1-\beta)^{t'-(n+1)} \beta^{n+1} \quad t' = 1,2,3...,\infty \end{aligned} \tag{3.21}$$

In (3.21), the summation over the number of orders (n) goes up to $\min\{t'-1, K-1\}$ since $P\{W^f = t' \mid N_{system} = n\} = 0$, when $n+1$ is larger than t'.

The expected production leadtime is:

$$E\left[W^f\right] = \frac{1}{\beta}\left(\frac{1}{1-\gamma/\beta} - \frac{K\left(\gamma/\beta\right)^K}{1-\left(\gamma/\beta\right)^K}\right). \tag{3.22}$$

Details about the derivation of expressions (3.20) and (3.22) are presented in Appendix A. It should be noted that when the probabilities in these expressions are replaced by rates, the results coincide with those known for the continuous time M/M/1/K system (see Gross et al. (2008, Chapter 2)).

Finally, the on-time probability of orders and the tardiness probabilities can be obtained by using the following relationships.

$$\eta^f = P\left\{W^f \le CRL\right\} = \sum_{c=1}^{CRL_{max}} P\left\{CRL = c\right\} P\left\{W^f \le c\right\} \tag{3.23}$$

$$P\left\{T^f = l\right\} = \begin{cases} P\left\{W^f \le CRL\right\} & l = 0 \\ \sum\limits_{c=1}^{CRL_{max}} P\left\{CRL = c\right\} P\left\{W^f = l + c\right\} & l = 1,2,...,\infty \end{cases} \tag{3.24}$$

3.3.2.2. Queuing model under simple due-date-based rules

The due-date-based simple rules that we consider and the additional constraints needed in the MDP for reducing the action space in each state in accordance with the rule are given below. These rules can be divided into two classes: static and dynamic. Under static rules, the priority assigned to an order remains the same while under dynamic rules, it might change over time.

1. EDD (earliest-due-date):

$$K(s) = \begin{cases} \{2\} & e < r_2 \\ \{k\} & r_{k-1} \leq e < r_k \quad k \in \{3, ..., n\}, \\ \{n+1\} & e \geq r_n \end{cases} \qquad \forall s \in S^{e>0,n} \tag{3.25}$$

   where e is the CRL of the arriving order for states with $e > 0$ and FCFS applies for two orders with the same due-date as a tie-breaker. This is a static rule since due-dates of orders do not change over time.

2. EDD^{otp}(adjusted earliest-due-date):

   This is a new due-date-based rule we propose. Under the EDD rule, higher priority is assigned to an order with a closer due-date (smaller remaining leadtime) even when the due-date is in the past. However, considering orders that are already "late" as higher priority orders than the orders with a positive chance of being completed "on-time" might not be a good policy for improving the on-time probability. Thus, we propose adjusting the EDD rule in such a way that whenever the due-date of a queuing order is reached, it is considered lower priority than orders with larger but positive remaining leadtimes.

   As opposed to the EDD rule, EDD^{otp} is a dynamic rule since the priorities might change over time. Consequently, we consider this rule in two different versions: EDD^{otp}_{PA} and EDD^{otp}_{DC}. The former is applied upon each arrival and considers placing the arriving order at an appropriate position in the queue while the latter considers the selection of the next order to process upon an order completion.

   - Under EDD^{otp}_{PA}, the arriving order is assigned with a higher priority than queuing orders that are already late. Note that, due to the discrete time assumption, the queuing orders with a remaining leadtime of 0 or 1 also do not have a chance of being completed on-time. The CRL of the arriving order is compared with the remaining leadtimes of existing orders, starting from the end of the queue. The arriving order continues to overtake orders

as long as the overtaking condition is satisfied ($e < r_k$ or $r_k \leq 1$). It gets the $k' + 1^{st}$ position when $1 < r_{k'} \leq e$ is satisfied for the first time. Thus,

$$K(s) = 1 + \max\{k \in \{1, ..., n\} | \, 1 < r_k \leq e\} \quad \forall s \in S^{e>0,n} \tag{3.26}$$

We assume that the order in process cannot be overtaken.

- Under EDD_{DC}^{otp}, the highest priority is given to the order with the smallest remaining leadtime that is strictly larger than 0. Because, when an order moves to processing with a remaining leadtime of 0, it has no chance of being completed on-time. If none of the orders satisfy this condition ($r_i > 0$), $FCFS$ applies.

3. LDD (latest-due-date):

   If most of the queuing orders are already late, even doing the opposite of what the EDD rule suggests and giving the highest priority to the order with the latest-due-date might be promising in terms of on-time probability performance. In this case, employing LDD would mean keeping the order with the largest probability of on-time completion for a given production leadtime (W) in front of the queue since $P\{W \leq r\}$ is larger for larger values of r.

$$K(s) = \begin{cases} \{2\} & e > r_2 \\ \{k\} & r_{k-1} \geq e > r_k \quad k \in \{3, ..., n\}, \\ \{n+1\} & e \leq r_n \end{cases} \quad \forall s \in S^{e>0,n} \tag{3.27}$$

   FCFS applies for two orders with the same due-date. Similar to the EDD rule, this is a static rule.

3.4. Numerical study

We first analyze the performance of optimal policies obtained under varying cost functions and system conditions, considering different decision-making strategies. Secondly, we benchmark simple rules against the optimal policy and compare them with each other.

The maximum system size is $K = 5$ and the upper bound of the CRL distribution is $CRL_{max} = 7$ unless stated otherwise. We vary the following system parameters where the underlined values form a base case. The processing time parameter (β) is the probability that a completion occurs and $\beta \in \{0.3, 0.4, \underline{0.5}\}$. The capital goods industry can also be referred here to example what these numbers in real-life might correspond to. For example, the Boeing Company announced that the

production for the 747-8 program will be limited to 6 airplanes per year starting from September 2016 (www.boeing.com). $\beta = 0.5$ corresponds to such a setting if one takes the time period in the model as a month.

VDMA (Verband Deutscher Maschinen- und Anlagenbau, Mechanical Engineering Industry Association), which represents a large number of mostly medium-sized companies in the capital goods industry in Germany, reports that the average capacity utilization in this industry was 84.7% and 84.5% in 2014 and 2015, respectively (www.vdma.org). We define the utilization level (ρ) as the probability that the system is not idle ($\rho = 1 - \pi_0$) and consider $\rho \in \{0.5, \underline{0.7}, 0.85\}$ in order to cover a wide range of scenarios.

We determine the value for γ, i.e. the order arrival probability, that results in a system utilization approximately equal to the selected level. The γ values selected for all (β, K) combinations considered in the numerical study are given in Table 3.1 together with the resulting utilizations (ρ).

β	K	γ	ρ
0.30	5	0.228	0.70
0.40	5	0.304	0.70
0.50	5	0.380	0.70
0.50	5	0.254	0.50
0.49	5	0.510	0.85
0.50	10	0.354	0.70
0.50	20	0.351	0.70

Table 3.1 γ values used in the numerical study

When γ is set equal to β, the resulting utilization level is $\rho = 0.833$. For obtaining a utilization level of 0.85, we set $\gamma = 0.51$ and $\beta = 0.49$. Due-date tightness levels (TL) represent different distributions of the CRL and we consider $\{\text{TL } 1, \text{TL } 2, \underline{\text{TL } 3}\}$. TL 1 corresponds to (nearly) equal probabilities for observing customer-required leadtimes between 1 and CRL_{max}. The probabilities for observing larger CRL's are higher under TL 2, which generates less tight due-dates. They are even higher under TL 3. In order to obtain TL's, we use a truncated opposite geometric distribution where the range of the geometric random variable has a finite maximum (CRL_{max}) and the highest probability of occurrence is associated with this maximum value.

$$P\{CRL = c\} = \frac{\delta(1-\delta)^{CRL_{max}-c+1}}{1-(1-\delta)^{CRL_{max}+1}-\delta} \tag{3.28}$$

TL 1, TL 2 and TL 3 are obtained by setting the parameter δ to 0.001, 0.1 and 0.2, respectively. Therefore, less due-date tightness means a higher mean and less variance in CRL. Since both the variance and mean of the CRL change with the tightness levels, we also conduct experiments with a varying CRL variance (σ) under a fixed mean and with a varying mean CRL (μ) under a fixed variance. For this purpose, we consider five other CRL distributions. These distributions are defined according to the probabilities given in Table 3.2. Note that $P\{CRL = i\} = 0$, $\forall i \in \{2, ..., CRL_{max} - 1\} \setminus \{\mu\}$.

Name	(μ, σ)	$P\{CRL = 1\}$	$P\{CRL = \mu\}$	$P\{CRL = CRL_{max}\}$
Dist 1	$(6.00, 0.00)$	0	1	0
Dist 2	$(5.00, 0.00)$	0	1	0
Dist 3	$(4.00, 0.00)$	0	1	0
Dist 4	$(4.00, 1.99)$	0.22	0.56	0.22
Dist 5	$(4.00, 3.00)$	0.5	0	0.5

Table 3.2 Distribution of CRL's for varying μ and σ

Thus, there is a total of eight different CRL distributions under consideration (TL 1-3 and Dist 1-5).

The experiments were performed on an Intel(R) Core(TM) i7-4790 M CPU, 3.6 GHz, 32 GB RAM. An optimization run under the PA strategy takes on average 814 seconds and varies between 590 and 1070 seconds for instances where the CRL distribution is determined according to TL 1-3. The average time reduces to 23 seconds for instances where the CRL distribution is determined according to Dist 1-5. This is because, in the former case each value in the range of possible CRL realizations $(1, 2, ..., CRL_{max})$ has a non-zero probability of occurrence while in the latter case this is only true for at most three CRL values. On the other hand, an optimization run under the DC strategy takes on average 19 seconds and varies between 14 and 28 seconds in the former (TL 1-3) while it takes on average 1 second in the latter case (Dist 1-5). The reason is that $e \in \{0, 1, 2, ..., CRL_{max}\}$ for the PA strategy, while $e \in \{-1, 0\}$ for the DC strategy in the implementation. Keeping other influencing factors, i.e. range of possible n's and r_i's, the same, this leads to a larger state space when the PA strategy is considered.

3.4.1. Analysis of the optimal policy

In this section, we focus on the customer-related performance of the optimal policies obtained under varying cost functions and system conditions, using the PA and DC strategies. Tables 3.3-3.6 present the on-time probability (η), the expected

tardiness ($E[T]$), and the value of the objective function ($Cost$) for an optimal policy. The tables show the effect of TL, ρ, β, μ and σ on these performance measures, respectively. The cost function, i.e. the value of the fixed and unit variable cost, and the decision-making strategy are varied in each table. Figures 3.1-3.4 visualize the results presented in Tables 3.3-3.6, respectively.

3.4.1.1. Effect of cost function on performance measures

Figure 3.1 shows the results under an increasing fixed cost in detail for the base case ($\rho = 0.7$, $\beta = 0.5$ and TL 3). Each dot represents the system performance when the priority sequencing decisions are optimized under the cost structure given as pairs (fixed, unit variable cost) below them. Therefore, the two extreme cases are: (0,1) where only the variable cost is relevant and (1,0) where only the fixed cost plays a role.

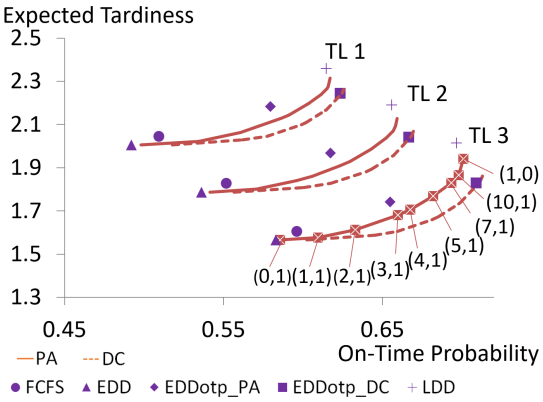

Figure 3.1 Effect of TL's ($\rho = 0.7, \beta = 0.5$)

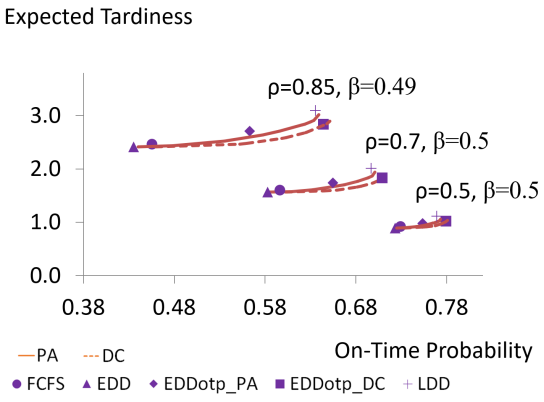

Figure 3.2 Effect of utilization (TL 3)

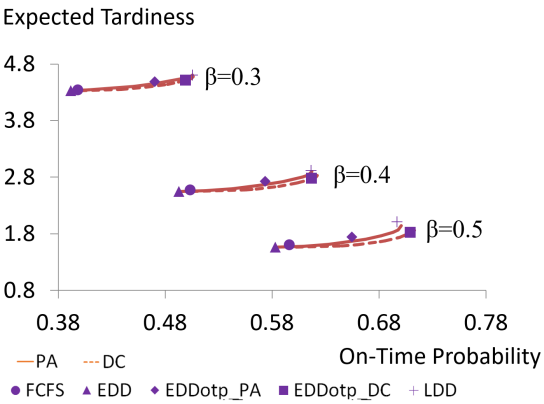

Figure 3.3 Effect of processing time parameter ($\rho = 0.7$, TL 3)

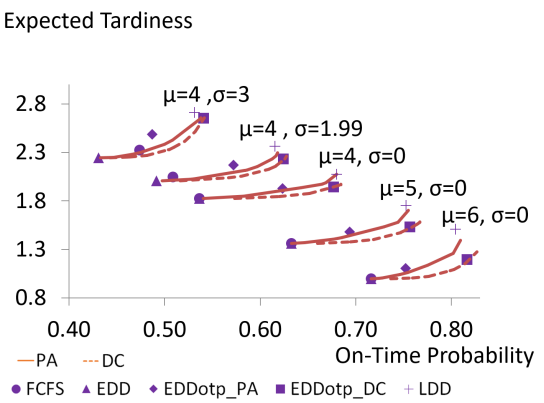

Figure 3.4 Effect of CRL variance and mean ($\rho = 0.7, \beta = 0.5$)

The results show that the (0,1) pair provides the minimum expected tardiness. Furthermore, the optimal priority sequencing policy leads to a higher on-time probability whenever a larger fixed cost applies. However, the achievement of a better

performance based on the on-time probability measure comes together with a higher expected tardiness. The optimal priority sequencing policy leads to the maximum on-time probability when (1,0) applies. The system performances under the optimal policies for different cost pairs form a Pareto curve on which it is not possible to improve one criterion without worsening the other. The system performances under simple rules are not Pareto-efficient because by employing better priority sequencing policies it is possible to improve one of the two criteria without worsening the other. This can be seen from the dots that represent the system performance under the simple rules being in the inner area of the Pareto frontier in all figures.

3.4.1.2. Effect of system conditions on Pareto curves

Pareto curves move to the upper-left side of the graph, which indicates a poorer performance based on both criteria as the CRL becomes tighter (Figure 3.1), utilization increases (Figure 3.2) and the order processing time gets longer and more volatile (i.e. β decreases, Figure 3.3). We observe the same effect for increasing variance and decreasing mean of CRL (Figure 3.4).

						PA				DC			
Instance	ρ	β	μ	σ		(0,1)	(1,1)	(4,1)	(1,0)	(0,1)	(1,1)	(4,1)	(1,0)
Base	0.7	0.5	4.86	1.88	η	0.585	0.609	0.667	0.701	0.602	0.645	0.685	0.713
			(TL 3)		$E[T]$	1.565	1.577	1.706	1.941	1.565	1.588	1.674	1.864
					$Cost$	0.550	0.691	1.067	0.105	0.550	0.683	1.030	0.101
1	0.7	0.5	4.42	1.97	η	0.541	0.571	0.629	0.659	0.557	0.601	0.643	0.670
			(TL 2)		$E[T]$	1.787	1.802	1.929	2.128	1.787	1.808	1.894	2.068
					$Cost$	0.628	0.784	1.199	0.120	0.628	0.776	1.167	0.116
2	0.7	0.5	4.00	2.00	η	0.498	0.534	0.591	0.617	0.518	0.562	0.602	0.626
			(TL 1)		$E[T]$	2.005	2.023	2.144	2.314	2.005	2.029	2.108	2.263
					$Cost$	0.705	0.874	1.328	0.135	0.705	0.867	1.301	0.132

Table 3.3 Effect of TL's on the performance of optimal policies

The parameters ρ, β, μ and σ affect not only the position of the Pareto curves but also their spread over the x and y axes. A wider spread indicates a larger error in case of a misspecification of the performance measure, i.e. in case the decision maker considers the expected tardiness measure when optimizing decisions while the on-time probability is more important for the customers and vice-versa. Thus, we look at the relative difference between the performance of the optimal policies for (0,1) and (1,0) cost pairs based on the two measures ($\frac{\eta^{(1,0)}-\eta^{(0,1)}}{\eta^{(0,1)}}$ and $\frac{E\left[T^{(1,0)}\right]-E\left[T^{(0,1)}\right]}{E\left[T^{(0,1)}\right]}$). We report these values in column Δ of Tables 3.4-3.6.

The misspecification error (Δ) increases in the utilization level for both performance measures (Table 3.4). It decreases for the expected tardiness as parameter β (Table 3.5) and the mean CRL (Table 3.6) decrease. On the other hand, with decreasing β and μ, it increases for the on-time probability measure. This effect is due to the

Instance	ρ	β	μ	σ		PA (0,1)	PA (1,1)	PA (4,1)	PA (1,0)	PA Δ	DC (0,1)	DC (1,1)	DC (4,1)	DC (1,0)	DC Δ
3	**0.5**	0.5	4.86	1.88	η	0.725	0.734	0.759	0.773	6.6%	0.733	0.749	0.769	0.781	6.5%
			(TL 3)		$E[T]$	0.895	0.900	0.956	1.061	18.5%	0.895	0.904	0.945	1.033	15.4%
					$Cost$	0.224	0.291	0.480	0.057		0.224	0.288	0.467	0.055	
Base	**0.7**	0.5	4.86	1.88	η	0.585	0.609	0.667	0.701	19.7%	0.602	0.645	0.685	0.713	18.4%
			(TL 3)		$E[T]$	1.565	1.577	1.706	1.941	24.0%	1.565	1.588	1.674	1.864	19.1%
					$Cost$	0.550	0.691	1.067	0.105		0.550	0.683	1.030	0.101	
4	**0.85**	0.49	4.86	1.88	η	0.440	0.478	0.596	0.639	45.1%	0.460	0.552	0.603	0.651	41.7%
			(TL 3)		$E[T]$	2.412	2.432	2.705	3.013	24.9%	2.412	2.464	2.589	2.894	20.0%
					$Cost$	1.004	1.230	1.800	0.150		1.004	1.212	1.739	0.145	

Table 3.4 Effect of utilization (ρ) on the performance of optimal policies

Instance	ρ	β	μ	σ		PA (0,1)	PA (1,1)	PA (4,1)	PA (1,0)	PA Δ	DC (0,1)	DC (1,1)	DC (4,1)	DC (1,0)	DC Δ
Base	0.7	**0.5**	4.86	1.88	η	0.585	0.609	0.667	0.701	19.7%	0.602	0.645	0.685	0.713	18.4%
			(TL 3)		$E[T]$	1.565	1.577	1.706	1.941	24.0%	1.565	1.588	1.674	1.864	19.1%
					$Cost$	0.550	0.691	1.067	0.105		0.550	0.683	1.030	0.101	
5	0.7	**0.4**	4.86	1.88	η	0.495	0.518	0.590	0.617	24.6%	0.504	0.555	0.602	0.622	23.4%
			(TL 3)		$E[T]$	2.548	2.562	2.723	2.901	13.9%	2.548	2.574	2.692	2.833	11.2%
					$Cost$	0.716	0.856	1.226	0.108		0.716	0.849	1.204	0.106	
6	0.7	**0.3**	4.86	1.88	η	0.393	0.409	0.496	0.505	28.8%	0.401	0.438	0.492	0.506	26.4%
			(TL 3)		$E[T]$	4.333	4.341	4.540	4.607	6.3%	4.333	4.351	4.481	4.580	5.7%
					$Cost$	0.913	1.040	1.382	0.104		0.913	1.036	1.373	0.104	

Table 3.5 Effect of β on the performance of optimal policies

due-dates becoming harder to meet as the expected production leadtime increases or the expected amount of time customers are willing to wait decreases. When this is the case, attaining an improvement in the expected tardiness becomes harder while it is still possible to improve on-time probability by employing stiffer policies such as giving late orders the least priority.

Finally the error one makes in the expected tardiness performance ($E[T]$) due to optimizing decisions using (1,0) cost pair instead of (0,1), increases in the customer-required leadtime variability (Table 3.6).

3.4.1.3. Effect of decision-making strategy

The results show that the DC strategy strictly dominates the PA strategy. The curve obtained by considering the former is always slightly to the right side of the one obtained by considering the latter strategy (Figures 3.1-3.4), which indicates a better performance in on-time probability without worsening the expected tardiness and vice-versa. Comparing $Cost$ values for PA and DC strategies in all instances of Tables 3.3-3.6, one can see that optimal policies under both strategies lead to the same objective value when the penalty is proportional to tardiness ((0,1) cost pair). The DC strategy provides additional cost savings when the tardiness penalty involves a fixed cost.

Instance	ρ	β	μ	σ		PA (0,1)	PA (1,1)	PA (4,1)	PA (1,0)	PA Δ	DC (0,1)	DC (1,1)	DC (4,1)	DC (1,0)	DC Δ
7	0.7	0.5	**6.00**	0.00	η	0.716	0.724	0.775	0.810	13.0%	0.736	0.759	0.803	0.827	12.2%
			(Dist 1)		$E[T]$	0.995	1.002	1.134	1.392	39.9%	0.995	1.006	1.094	1.276	28.2%
					$Cost$	0.350	0.449	0.715	0.067		0.350	0.438	0.662	0.061	
8	0.7	0.5	**5.00**	0.00	η	0.633	0.646	0.744	0.755	19.3%	0.660	0.694	0.747	0.767	16.2%
			(Dist 2)		$E[T]$	1.363	1.369	1.581	1.702	24.9%	1.363	1.377	1.475	1.582	16.1%
					$Cost$	0.479	0.605	0.915	0.086		0.479	0.591	0.873	0.082	
9	0.7	0.5	**4.00**	**0.00**	η	0.540	0.583	0.669	0.680	25.8%	0.573	0.620	0.684	0.684	19.4%
			(Dist 3)		$E[T]$	1.826	1.847	1.986	2.074	13.6%	1.826	1.845	1.967	1.973	8.0%
					$Cost$	0.642	0.796	1.163	0.112		0.642	0.782	1.135	0.111	
10	0.7	0.5	4.00	**1.99**	η	0.497	0.527	0.597	0.618	24.4%	0.516	0.567	0.610	0.627	21.5%
			(Dist 4)		$E[T]$	2.008	2.023	2.152	2.297	14.4%	2.008	2.037	2.130	2.259	12.5%
					$Cost$	0.705	0.877	1.323	0.134		0.705	0.868	1.297	0.131	
11	0.7	0.5	4.00	**3.00**	η	0.433	0.447	0.500	0.538	24.2%	0.443	0.464	0.504	0.541	22.2%
			(Dist 5)		$E[T]$	2.243	2.250	2.384	2.655	18.4%	2.243	2.251	2.329	2.656	18.4%
					$Cost$	0.788	0.985	1.540	0.162		0.788	0.979	1.515	0.161	

Table 3.6 Effects of CRL mean (μ) and variance (σ) on the performance of optimal policies

The reason for this observation is the following. Under the PA strategy, the decisions regarding priorities of orders are made upon arrival. An order with a small but still positive remaining leadtime at this point in time might be given a higher priority than the arriving one. However, there is a positive probability that the remaining processing time of the currently processed order is larger than the remaining leadtime of this order and a higher priority is actually given to an order that is already late before its processing starts. On the other hand, under the DC strategy, the decision is delayed until the next completion, which is the point where the uncertainty about the remaining processing time of the currently processed order is resolved. Thus, the latter strategy benefits from "wait and see".

Decision upon completion provides the biggest benefit when the variability in customer-required leadtimes is low and their mean value is high. The objective value for the (1,0) cost pair decreases from 0.067 to 0.061, which corresponds to nearly 9% improvement, when $\mu = 6$, $\sigma = 0$ (Table 3.6). The advantage one gains by postponing decisions until the next order completion is typically degraded by an increasing variability in customer-required leadtimes and decreasing value of their mean. The improvement in the objective value for the (1,0) cost pair is less than 1% when $\mu = 4$ and $\sigma = 3$ (Table 3.6).

3.4.2. Benchmarking simple priority rules against the optimal policy

In this section, we benchmark simple rules against the optimal policy. Tables 3.7-3.10 present the performance of simple rules based on the percentage cost gap. For a simple rule, the percentage cost gap is defined as the percentage increase in the

value of the objective function when the rule is employed instead of the optimal policy ($\frac{Cost_{Rule}-Cost_{Opt}}{Cost_{Opt}} \cdot 100\%$). The simple rule that provides the smallest percentage cost gap is marked in bold in each case for each cost pair.

Instance	ρ	β	μ	σ		Percentage cost gap using PA				Percentage cost gap using DC			
						$FCFS$	EDD	EDD^{otp}_{PA}	LDD	$FCFS$	EDD	EDD^{otp}_{DC}	LDD
Base	0.7	0.5	4.86	1.88	(0,1)	2.6	**0.0**	11.3	28.7	2.6	**0.0**	16.9	28.7
			(TL 3)		(1,1)	2.1	**0.7**	6.1	17.8	3.4	**2.0**	9.1	19.3
					(4,1)	6.1	6.5	**2.9**	6.3	9.9	10.3	**2.1**	10.1
					(1,0)	35.1	39.4	15.4	**1.4**	40.8	45.4	**1.4**	5.7
1	0.7	0.5	4.42	1.97	(0,1)	2.3	**0.0**	10.1	22.5	2.3	**0.0**	14.3	22.5
			(TL 2)		(1,1)	2.0	**0.9**	5.4	13.6	3.1	**2.0**	7.6	14.8
					(4,1)	6.1	6.8	**2.6**	4.5	9.0	9.7	**1.7**	7.4
					(1,0)	31.5	36.1	12.3	**1.0**	35.6	40.4	**0.9**	4.2
2	0.7	0.5	4.00	2.00	(0,1)	2.0	**0.0**	8.8	17.7	2.0	**0.0**	11.9	17.7
			(TL 1)		(1,1)	1.9	**1.0**	4.6	10.3	2.8	**1.9**	6.2	11.3
					(4,1)	6.0	6.8	**2.3**	3.2	8.3	9.1	**1.3**	5.4
					(1,0)	28.2	32.7	9.8	**0.7**	31.1	35.7	**0.6**	3.0

Table 3.7 Effect of TL's on the percentage cost gap of simple priority rules

The percentage cost gap resulting from the use of the EDD rule is observed to be 0% in Tables 3.7-3.10 whenever the tardiness penalty is proportional to tardiness (fixed cost=0), as shown by Duenyas and Hopp (1995). Thus, in Figures 3.1-3.4, the system performance under EDD is represented by a point that is exactly at the same level on the y-axis as the optimal policy for the (0,1) cost pair. It falls slightly to the left since the on-time probability obtained by employing the EDD rule is slightly worse. Although the resulting long-run average costs are exactly the same, the priority sequencing decisions given by the optimal policy under PA and DC for the (0,1) cost pair are not identical to each other and to the ones given by the EDD rule. In other words, making priority sequencing decisions differently for some states does not result in a different objective value when fixed cost=0. Suppose that the system is in a state where $(r_1, r_2) = (3, 1)$ and the position of a new arriving order with $CRL = 2$ is decided. The EDD rule suggests $k = 3$, while the optimal policy for (0,1) cost pair under the PA strategy suggests $k = 2$. The total variable tardiness cost incurred for these three orders is the same under both decisions, no matter the processing times of orders. If each order takes one period to process, incurred total variable tardiness cost is: $0 + 1 + 1 = 2$ under $k = 3$, and $0+0+2 = 2$ under $k = 2$. On the other hand, $k = 2$ leads to an on-time completion of the arriving order (with $CRL = 2$), while the decision $k = 3$ leads to its late completion. The two decisions are different in terms of the number of late orders they result in, but this is not reflected in the objective value when fixed cost=0.

Whenever a fixed cost of tardiness is involved, the optimal policy deviates from the EDD principle. Adhering to EDD results in a large percentage cost gap

Instance	ρ	β	μ	σ		Percentage cost gap using PA $FCFS$	EDD	EDD^{otp}_{PA}	LDD	Percentage cost gap using DC $FCFS$	EDD	EDD^{otp}_{DC}	LDD
3	**0.5**	0.5	4.86	1.88	(0,1)	2.7	**0.0**	9.1	24.6	2.7	**0.0**	13.7	24.6
			(TL 3)		(1,1)	2.1	**0.5**	4.9	15.4	3.1	**1.5**	7.3	16.6
					(4,1)	4.3	4.3	**2.2**	6.2	7.2	7.1	**1.7**	9.1
					(1,0)	19.3	22.0	8.6	**1.7**	23.6	26.3	**0.7**	5.4
Base	**0.7**	0.5	4.86	1.88	(0,1)	2.6	**0.0**	11.3	28.7	2.6	**0.0**	16.9	28.7
			(TL 3)		(1,1)	2.1	**0.7**	6.1	17.8	3.4	**2.0**	9.1	19.3
					(4,1)	6.1	6.5	**2.9**	6.3	9.9	10.3	**2.1**	10.1
					(1,0)	35.1	39.4	15.4	**1.4**	40.8	45.4	**1.4**	5.7
4	**0.85**	0.49	4.86	1.88	(0,1)	2.0	**0.0**	12.3	28.3	2.0	**0.0**	17.6	28.3
			(TL 3)		(1,1)	1.8	**0.8**	6.5	17.1	3.2	**2.2**	9.6	18.8
					(4,1)	7.3	8.1	**3.1**	5.3	11.1	11.9	**2.0**	9.0
					(1,0)	50.9	56.5	21.0	**1.0**	56.2	61.9	**2.0**	4.5

Table 3.8 Effect of utilization (ρ) on the percentage cost gap of simple priority rules

when the relative value of the fixed cost is high. Furthermore, an increase in the utilization level (ρ) and in the processing time parameter (β), and a decrease in the CRL variance (σ), increase this gap (Tables 3.8-3.10, respectively). Under a high utilization level ($\rho = 0.85$), the percentage cost gap resulting from the use of EDD for the pair (1,0) exceeds 56%. This is due to the effects observed in Figures 3.2-3.4. In other words, the potential benefit provided by policies appropriate for use when there is a fixed cost (e.g. EDD^{otp}_{PA}) is higher when the utilization is higher (Table 3.8), the expected processing time is shorter and the variability in processing times is lower (i.e. β is higher Table 3.9), and CRL variability is lower (Table 3.10).

Instance	ρ	β	μ	σ		Percentage cost gap using PA $FCFS$	EDD	EDD^{otp}_{PA}	LDD	Percentage cost gap using DC $FCFS$	EDD	EDD^{otp}_{DC}	LDD
Base	0.7	**0.5**	4.86	1.88	(0,1)	2.6	**0.0**	11.3	28.7	2.6	**0.0**	16.9	28.7
			(TL 3)		(1,1)	2.1	**0.7**	6.1	17.8	3.4	**2.0**	9.1	19.3
					(4,1)	6.1	6.5	**2.9**	6.3	9.9	10.3	**2.1**	10.1
					(1,0)	35.1	39.4	15.4	**1.4**	40.8	45.4	**1.4**	5.7
5	0.7	**0.4**	4.86	1.88	(0,1)	1.0	**0.0**	6.9	14.7	1.0	**0.0**	9.3	14.7
			(TL 3)		(1,1)	0.9	**0.4**	3.5	8.6	1.7	**1.2**	4.9	9.5
					(4,1)	4.5	4.9	**1.6**	2.1	6.4	6.9	**0.8**	4.0
					(1,0)	29.7	32.5	11.4	**0.1**	31.5	34.3	**1.5**	1.5
6	0.7	**0.3**	4.86	1.88	(0,1)	0.3	**0.0**	3.5	6.3	0.3	**0.0**	4.3	6.3
			(TL 3)		(1,1)	0.3	**0.2**	1.7	3.4	0.7	**0.6**	2.2	3.8
					(4,1)	3.0	3.2	0.7	**0.4**	3.7	3.9	**0.2**	1.1
					(1,0)	21.7	23.0	7.1	**0.0**	21.9	23.2	1.5	**0.2**

Table 3.9 Effect of β on the percentage cost gap of simple priority rules

The potential benefit gained by using policies that mitigate the amount of tardiness, such as EDD, increases with a decreasing utilization level (ρ), increasing processing time parameter β (lower mean and lower variance in processing times) and increasing CRL variance (σ). This can be seen from the increase of the percentage cost gap resulting from the use of the $FCFS$ discipline under the (0,1) cost pair in Tables 3.8, 3.9 and Table 3.10, respectively.

Instance	ρ	β	μ	σ		Percentage cost gap using PA				Percentage cost gap using DC			
						$FCFS$	EDD	EDD^{otp}_{PA}	LDD	$FCFS$	EDD	EDD^{otp}_{DC}	LDD
7	0.7	0.5	**6.00**	0.00	(0,1)	**0.0**	**0.0**	11.0	51.4	**0.0**	**0.0**	20.3	51.4
			(Dist 1)		(1,1)	**0.0**	**0.0**	5.8	33.2	**2.6**	**2.6**	10.7	36.6
					(4,1)	4.7	4.7	**3.0**	12.6	13.1	13.1	**2.6**	21.6
					(1,0)	49.0	49.0	30.1	**2.8**	63.7	63.7	**6.0**	12.9
8	0.7	0.5	**5.00**	0.00	(0,1)	**0.0**	**0.0**	8.7	28.8	**0.0**	**0.0**	12.5	28.8
			(Dist 2)		(1,1)	**0.4**	**0.4**	3.8	16.2	**2.8**	**2.8**	5.5	18.9
					(4,1)	8.7	8.7	**3.9**	5.3	13.9	13.9	**0.8**	10.4
					(1,0)	49.8	49.8	24.9	**0.9**	57.7	57.7	**4.4**	6.2
9	0.7	0.5	**4.00**	**0.00**	(0,1)	**0.0**	**0.0**	5.5	13.6	**0.0**	**0.0**	6.4	13.6
			(Dist 3)		(1,1)	**1.1**	**1.1**	1.7	5.7	2.9	2.9	**1.9**	7.6
					(4,1)	11.2	11.2	3.7	**1.3**	14.0	14.0	**0.2**	3.9
					(1,0)	44.9	44.9	17.7	**0.0**	47.0	47.0	2.5	**1.4**
10	0.7	0.5	4.00	**1.99**	(0,1)	1.9	**0.0**	8.1	17.6	1.9	**0.0**	11.3	17.6
			(Dist 4)		(1,1)	1.6	**0.8**	4.1	10.0	2.7	**1.9**	5.6	11.2
					(4,1)	6.5	7.4	**3.1**	3.6	8.7	9.5	**1.3**	5.7
					(1,0)	28.7	33.3	12.2	**0.8**	31.9	36.5	**0.9**	3.3
11	0.7	0.5	4.00	**3.00**	(0,1)	3.6	**0.0**	10.9	21.0	3.6	**0.0**	18.2	21.0
			(Dist 5)		(1,1)	1.7	**0.3**	7.0	13.5	2.3	**0.9**	11.6	14.2
					(4,1)	**1.0**	3.1	3.6	4.7	**2.8**	4.8	4.1	6.5
					(1,0)	13.9	23.1	11.0	**1.5**	14.8	24.1	**0.1**	2.3

Table 3.10 Effects of CRL mean (μ) and variance (σ) on the percentage cost gap of simple priority rules

When the PA strategy is considered, the simple rule with the smallest percentage cost gap moves from EDD to EDD^{otp}_{PA} and then to LDD as the relative value of the fixed cost increases. This effect is observed in all instances except the one where the variability in customer-required leadtimes is high ($\sigma = 3$). When $\sigma = 3$, $FCFS$ performs closer to optimal than EDD^{otp}_{PA}, as can be seen from Figure 3.4 as well as from Table 3.10 where the best performing rule for the cost pair (4,1) is $FCFS$ under both strategies.

When the DC strategy is considered, the percentage cost gap resulting from the use of LDD is rarely the smallest. Only for the cost pair (1,0) in instances 6 and 9 in Tables 3.9 and 3.10 this is the case. Thus, as the relative value of the fixed cost increases, the simple rule with the smallest percentage cost gap moves from EDD to EDD^{otp}_{DC} and it stays there for further increases, in all other instances. In the worst case among the ones considered here, the percentage cost gap resulting from the use of the best performing simple rule is 3.9 under the PA strategy and 6.0 under the DC strategy (for instances 8 and 7 in Table 3.10). Thus, results show that the simple rules perform well in comparison to the optimal.

This is an important result from the practical point of view because, in terms of ease of applicability, simple rules are superior to the optimal policy in the following sense. First, simple rules are given policies, i.e. no computational effort is required to obtain them. Second, as discussed by Hopp and Spearman (2001), they only require

polynomial time sorting algorithms for their implementation. The computational complexity of sorting n orders is $O(nlogn)$ when the queue is maintained unsorted. The complexity is less in the sorted case. On the other hand, as discussed by Littman (1997), using linear programming, the MDP can be solved in polynomial time in the number of states. However, the number of states increases exponentially in problem size. Thus, the computational time required for obtaining the optimal policy grows exponentially in problem size.

3.4.3. Comparing simple priority rules with each other

In this section, we compare simple rules with each other. Table 3.11 presents the performance of simple rules based on the on-time probability (η) and expected tardiness ($E[T]$) measures.

When all arriving customers are willing to wait for the same amount of time, i.e. Dist 1, 2 or 3 applies for the CRL distribution, the earliest-due-date always belongs to the order that had arrived first. Thus, the simple rules $FCFS$ and EDD are equivalent and the results presented in Table 3.11 show the same η and $E[T]$ values for EDD and $FCFS$ in instances 7-9. Furthermore, both Figure 3.4 and instances 9-11 in Table 3.11 show that increasing customer-required leadtime variability (σ) increases their performance difference with regard to both measures.

In addition, the following observations can be made from all instances in Table 3.11. The on-time probability is worse under EDD than it is under the basic $FCFS$ whenever $\sigma > 0$. This is an interesting result since the EDD rule takes into account the information regarding due-dates while $FCFS$ does not. The reason for this effect is the following: in all instances investigated up-to now (base case and instances 1-11), the maximum customer-required leadtime is 7, on the other hand, the expected production leadtime is 4.937 in the base case ($E[W]$ is obtained using (3.13), equivalently one can consider (3.22) under $\gamma = 0.38$, $\beta = 0.5$, $K = 5$). Thus, a situation in which orders with already reached due-dates are still in the system, may frequently occur. This means that under the EDD rule, the arriving orders with close due-dates are kept waiting for orders that are already late to be completed, which also prevents them from being completed on-time and hurts the on-time probability measure.

On the other hand, EDD_{PA}^{otp} and EDD_{DC}^{otp} treat the queuing orders with no chance of on-time completion differently and LDD gives the highest priority to the order with the highest probability of being completed on-time. Therefore, as also visualized

Instance	ρ	β	μ	σ		$FCFS$	EDD	EDD^{otp}_{PA}	EDD^{otp}_{DC}	LDD
Base	0.7	0.5	4.86	1.88	η	0.596	0.583	0.655	0.709	0.697
			(TL 3)		$E[T]$	1.606	1.565	1.742	1.830	2.015
1	0.7	0.5	4.42	1.97	η	0.552	0.536	0.617	0.666	0.656
			(TL 2)		$E[T]$	1.828	1.787	1.968	2.042	2.190
2	0.7	0.5	4.00	2.00	η	0.509	0.492	0.579	0.623	0.614
			(TL 1)		$E[T]$	2.045	2.005	2.183	2.244	2.360
3	**0.5**	0.5	4.86	1.88	η	0.729	0.723	0.754	0.779	0.769
			(TL 3)		$E[T]$	0.920	0.895	0.977	1.019	1.115
4	**0.85**	0.49	4.86	1.88	η	0.455	0.435	0.563	0.644	0.635
			(TL 3)		$E[T]$	2.461	2.412	2.708	2.837	3.094
5	0.7	**0.4**	4.86	1.88	η	0.503	0.493	0.573	0.617	0.617
			(TL 3)		$E[T]$	2.573	2.548	2.725	2.784	2.923
6	0.7	**0.3**	4.86	1.88	η	0.398	0.392	0.470	0.499	0.505
			(TL 3)		$E[T]$	4.346	4.333	4.485	4.519	4.607
7	0.7	0.5	**6.00**	0.00	η	0.716	0.716	0.752	0.816	0.804
			(Dist 1)		$E[T]$	0.995	0.995	1.105	1.197	1.507
8	0.7	0.5	**5.00**	0.00	η	0.633	0.633	0.694	0.757	0.753
			(Dist 2)		$E[T]$	1.363	1.363	1.482	1.532	1.754
9	0.7	0.5	**4.00**	**0.00**	η	0.536	0.536	0.623	0.677	0.680
			(Dist 3)		$E[T]$	1.826	1.826	1.926	1.944	2.074
10	0.7	0.5	4.00	**1.99**	η	0.509	0.491	0.572	0.624	0.615
			(Dist 4)		$E[T]$	2.045	2.008	2.171	2.234	2.361
11	0.7	0.5	4.00	**3.00**	η	0.474	0.431	0.487	0.541	0.531
			(Dist 5)		$E[T]$	2.324	2.243	2.488	2.651	2.713

Table 3.11 The customer-related performance of simple priority rules

in Figures 3.1-3.4, they provide better on-time probability performance with larger expected tardiness than $FCFS$. Of these three rules, the EDD^{otp}_{DC} rule provides the best on-time probability in all instances except 6 and 9, in which the best on-time probability is obtained by employing LDD. For instance 5, η under EDD^{otp}_{DC} is slightly better than η under LDD (in fourth digit after the comma).

While LDD gives the second best on-time probability whenever it does not give the best, it always results in the largest expected tardiness. This means that a trade-off between two measures is apparent between EDD^{otp}_{PA} and LDD but EDD^{otp}_{DC} outperforms LDD by providing a better performance based on both, in all instances except 6 and 9.

3.4.4. Comparing simple priority rules with each other in larger problems

In this section, we vary the parameters maximum system size (K) and the maximum possible customer-required leadtime CRL_{max}, both of which determine the size of the problem. We consider $K \in \{5, 10, 20\}$ and $CRL_{max} \in \{15, 20\}$. Although in these cases the problem size is too large to allow an exact analysis and optimization,

i.e. benchmarking against the optimal policy is not possible, we compare simple rules with each other. This allows us to identify the effects that remain the same, as well as those that show differences to the ones observed in smaller sized problems in Section 3.4.3.

We use simulation to investigate the performance of the simple rules based on η and $E[T]$. We model the discrete time MTO system described in Section 3.2 and implement it in AnyLogic 7. The simulation is stopped after 150,000 orders are completed, the warm up period ends after 20,000 orders are completed and 150 replications are conducted. We consider varying customer-required leadtime tightness levels and take $\rho = 0.7$ and $\beta = 0.5$. Tables 3.12 and 3.13 present results under varying CRL_{max} and K respectively. The long-run average cost under the cost pair (4,1) is given in addition to η and $E[T]$ as an indicator of the combined performance.

CRL_{max}	Tightness		$FCFS$	EDD	EDD^{otp}_{PA}	EDD^{otp}_{DC}	LDD
15	TL 1	η	0.740	0.744	0.769	0.812	0.770
			(±0.000)	(±0.000)	(±0.000)	(±0.000)	(±0.000)
		$E[T]$	1.048	0.823	0.992	1.286	1.676
			(±0.002)	(±0.002)	(±0.002)	(±0.002)	(±0.003)
		Cost	0.734	0.649	0.674	0.716	0.912
			(±0.001)	(±0.001)	(±0.001)	(±0.001)	(±0.001)
15	TL 2	η	0.838	0.853	0.861	0.890	0.839
			(±0.000)	(±0.000)	(±0.000)	(±0.000)	(±0.000)
		$E[T]$	0.623	0.437	0.522	0.790	1.300
			(±0.001)	(±0.001)	(±0.001)	(±0.002)	(±0.003)
		Cost	0.447	0.360	0.379	0.432	0.683
			(±0.001)	(±0.001)	(±0.001)	(±0.001)	(±0.001)
15	TL 3	η	0.910	0.926	0.928	0.943	0.889*
			(±0.000)	(±0.000)	(±0.000)	(±0.000)	(±0.000)
		$E[T]$	0.320	0.207	0.237	0.413	0.987
			(±0.001)	(±0.001)	(±0.001)	(±0.002)	(±0.002)
		Cost	0.238	0.176	0.184	0.225	0.503
			(±0.001)	(±0.001)	(±0.001)	(±0.001)	(±0.001)
20	TL 1	η	0.804	0.829	0.839	0.865	0.812
			(±0.000)	(±0.000)	(±0.000)	(±0.000)	(±0.000)
		$E[T]$	0.786	0.508	0.629	0.969	1.453
			(±0.001)	(±0.001)	(±0.002)	(±0.002)	(±0.003)
		Cost	0.551	0.418	0.447	0.530	0.775
			(±0.001)	(±0.001)	(±0.001)	(±0.001)	(±0.001)
20	TL 2	η	0.912	0.938	0.939	0.947	0.890*
			(±0.000)	(±0.000)	(±0.000)	(±0.000)	(±0.000)
		$E[T]$	0.336	0.169	0.206	0.400	0.974
			(±0.001)	(±0.001)	(±0.001)	(±0.001)	(±0.002)
		Cost	0.242	0.147	0.158	0.216	0.497
			(±0.001)	(±0.000)	(±0.000)	(±0.001)	(±0.001)
20	TL 3	η	0.970	0.983	0.983	0.984	0.935*
			(±0.000)	(±0.000)	(±0.000)	(±0.000)	(±0.000)
		$E[T]$	0.107	0.044	0.052	0.113	0.635
			(±0.000)	(±0.000)	(±0.000)	(±0.001)	(±0.002)
		Cost	0.080	0.040	0.042	0.062	0.315
			(±0.000)	(±0.000)	(±0.000)	(±0.000)	(±0.001)

Values after ± give the half-width of a 95% confidence interval.

Table 3.12 Effect of CRL_{max} on the performance of simple rules in larger problem sizes ($K = 5$)

The results show that some of the effects observed for small problems change when we have large problems while some remain the same. In particular, the following observations can be made from all instances given in Tables 3.12 and 3.13. Both the negative impact of EDD on the on-time probability and the negative impact of EDD^{otp}_{PA} on the expected tardiness disappear when due-dates are easier to meet, i.e. $CRL_{max} \in \{15, 20\}$. Thus, EDD and EDD^{otp}_{PA} dominate the $FCFS$ discipline. On the other hand, EDD^{otp}_{DC} and LDD typically improve the on-time probability compared to $FCFS$ while expected tardiness degrades. However, under LDD, a small improvement in η corresponds to a large worsening in $E[T]$. Although in Tables 3.7-3.10 LDD has always shown a lower percentage cost gap than FCFS under the cost pair (1,0), in some instances of Tables 3.12-3.13, the on-time probability under LDD is also worse than it is under $FCFS$. The on-time probability under LDD is marked in tables to show such instances. EDD^{otp}_{DC} outperforms LDD by providing a better performance based on both measures, similarly to results presented in Table 3.11. Furthermore, a trade-off of two performance measures is still apparent between EDD^{otp}_{DC}, EDD^{otp}_{PA} and EDD.

Since the EDD^{otp}_{PA} and EDD^{otp}_{DC} rules, which were near optimal for small sized systems, also show good performance in these experiments, they appear to be robust and therefore superior to EDD and LDD.

3.5. Conclusions

The impact of priority sequencing decisions on the customer-related performance of a make-to-order (MTO) production system was investigated. The priority sequencing problem was modeled as a Markov decision process (MDP). The objective function was defined as the sum of a fixed and a variable cost of tardiness, which allowed the investigation of the two commonly used performance criteria "on-time probability" and "expected tardiness". With the MDP model, it is possible to compare simple priority rules with the optimal policy while it is only possible to compare different rules with each other when the optimal policy is unknown.

The numerical results show that it is possible to achieve near optimal system performance by employing simple due-date-based rules. Nevertheless, the optimal policy, as well as the simple rule with the closest performance, heavily depend on the relative value of the fixed cost in tardiness penalty. As the value of the fixed cost increases, the performance of EDD deteriorates. When the fixed cost plays a role and the due-dates are difficult to meet, adhering to processing based on the EDD discipline results in a percentage cost gap of more than 56%. On the other hand,

K	Tightness		$FCFS$	EDD	EDD^{otp}_{PA}	EDD^{otp}_{DC}	LDD
10	TL 1	η	0.744	0.748	0.769	0.846	0.792
			(± 0.000)	(±0.001)	(±0.000)	(±0.000)	(±0.000)
		$E[T]$	1.441	1.106	1.258	1.860	2.567
			(±0.005)	(±0.004)	(±0.005)	(±0.007)	(±0.008)
		Cost	0.865	0.742	0.765	0.869	1.193
			(±0.003)	(±0.002)	(±0.002)	(±0.003)	(±0.003)
10	TL 2	η	0.859	0.874	0.878	0.927	0.867
			(±0.000)	(±0.000)	(±0.000)	(±0.000)	(±0.000)
		$E[T]$	0.747	0.505	0.551	1.046	2.006
			(±0.003)	(±0.003)	(±0.003)	(±0.005)	(±0.007)
		Cost	0.460	0.353	0.364	0.469	0.889
			±0.002)	(±0.002)	(±0.002)	(±0.002)	(±0.003)
10	TL 3	η	0.927	0.940	0.940	0.967	0.910*
			(±0.000)	(±0.000)	(±0.000)	(±0.000)	(±0.000)
		$E[T]$	0.351	0.235	0.243	0.551	1.595
			(±0.002)	(±0.002)	(±0.002)	(±0.004)	(±0.007)
		Cost	0.226	0.167	0.169	0.240	0.685
			(±0.001)	(±0.001)	(±0.001)	(±0.002)	(±0.003)
20	TL 1	η	0.728	0.731	0.754	0.842	0.788
			(±0.001)	(±0.001)	(±0.001)	(±0.000)	(±0.000)
		$E[T]$	1.801	1.496	1.641	2.284	3.046
			(±0.010)	(±0.009)	(±0.011)	(±0.012)	(±0.013)
		Cost	1.014	0.903	0.921	1.023	1.366
			(±0.005)	(±0.004)	(±0.005)	(±0.005)	(±0.005)
20	TL 2	η	0.838	0.853	0.858	0.923	0.864
			(±0.001)	(±0.001)	(±0.001)	(±0.000)	(±0.000)
		$E[T]$	1.065	0.825	0.864	1.438	2.457
			(±0.009)	(±0.009)	(±0.009)	(±0.012)	(±0.013)
		Cost	0.601	0.496	0.502	0.613	1.053
			(±0.004)	(±0.004)	(±0.004)	(±0.005)	(±0.005)
20	TL 3	η	0.906	0.916	0.917	0.962	0.907
			(±0.001)	(±0.001)	(±0.001)	(±0.000)	(±0.000)
		$E[T]$	0.610	0.504	0.517	0.905	2.021
			(±0.006)	(±0.006)	(±0.006)	(±0.009)	(±0.011)
		Cost	0.346	0.294	0.298	0.371	0.840
			(±0.003)	(±0.003)	(±0.003)	(±0.003)	(±0.004)

Values after ± give the half-width of a 95% confidence interval.

Table 3.13 Effect of K on the performance of simple rules in larger problem sizes ($CRL_{max} = 20$)

the proposed EDD^{otp} rule performs close to optimal, which also works well when the due-dates are easier to meet. Moreover, when there is a fixed cost, delaying the priority sequencing decisions to the next order completion instead of deciding upon arrival provides further improvement potential.

Due to the nature of priority sequencing decisions, the size of the state space increases exponentially in the problem size. The major limitation of an exact analysis is therefore the fact that required computational times easily become prohibitive for larger sized problems.

4. Dynamic Pricing, Leadtime Quotation and Due-Date-Based Priority Dispatching

We model the marketing-operations collaboration problem as a Markov decision process (MDP) to obtain the optimal quotation and dispatching policy numerically. We further investigate the sub-optimality of several sequential approaches. Our numerical results show that sub-optimality is negligible when the tardiness penalty is proportional to tardiness and the customer sensitivities to price and leadtime quotes are similar. However, it is considerable when tardiness of orders is penalized with a fixed cost and the customers differ significantly in their sensitivity to price and leadtime. By joint optimization, it is possible to make more appealing price/leadtime quotes to customers and at the same time reach a better service level. On the other hand, the joint optimization can also suggest the lowering of a firm's service level in order to achieve higher profits by serving more customers.

4.1. Introduction

This chapter considers a firm that provides service to a market of price- and leadtime-sensitive customers and processes orders following a make-to-order fashion (MTO). The customer arrivals and order completions evolve according to a stochastic process. The firm makes individual price/leadtime quotes to dynamically arriving prospective customers, who then make the decision whether or not to place an order by trading off these two aspects of service. The customers are not homogeneous in their sensitivity to price and leadtime. If a customer accepts the quote, his order along with its promised leadtime, becomes an input to manufacturing. If completing an order takes longer than the promised leadtime, the firm may suffer a loss of goodwill or future business, may incur extra shipping costs or lose revenue if it has been offering price discounts in case of tardiness. The firm's objective is to maximize the expected profit, which is the margin earned from placed orders minus tardiness penalties. By quoting prices and leadtimes, the firm controls its demand, i.e., makes an order selection decision by influencing which of the prospective customers finally places an order. By dispatching orders accordingly, the firm aims at decreasing tardiness penalties. In such a business environment, it is necessary for the firm to consider interdependencies between marketing- and operations- (manufacturing) related decisions.

We compare two scenarios regarding the firm's approach to taking such interdependencies into account. In the first scenario (sequential approach), marketing

quotes price/leadtime pairs in coordination with manufacturing, which then dispatches obtained orders with given leadtimes. In the second scenario, marketing and manufacturing fully collaborate and make these decisions jointly (simultaneous optimization approach). The decisions in the first scenario are made as follows. Marketing considers the firm's profit maximization as the objective, i.e., it takes the consequences of its decisions for manufacturing into account. Thus, it manages the trade-off between quoting shorter leadtimes to attract more customers and increase their willingness to pay and incurring higher tardiness penalties. Upon arrival of a prospective customer, marketing knows which type of customer he is. However, it has incomplete information regarding the current state of manufacturing. Although the current number of orders in the manufacturing system and the maximum number of pending orders it allows are known to marketing, it has no information about the sequence in which they are going to be processed. Marketing therefore assumes a first-come-first-served (FCFS) discipline in processing. However, manufacturing may follow alternative policies for dispatching.

We answer the following research questions: (1) How much can the profitability be increased? (2) How are the system utilization, the service level of the firm and the selection of different customer types affected when the simultaneous optimization approach, rather than one of the sequential approaches, is considered? (3) How do the tardiness penalty structure and the market-related characteristics affect the answers to the first two questions?

The chapter is structured as follows: the model is presented in Section 4.2. Section 4.3 introduces the investigated KPI's and describes their computation. Section 4.4 presents numerical results and Section 4.5 summarizes findings.

4.2. Markov decision model

We model an MTO firm as a discrete time queuing system with a single server and a limited waiting space. We assume that time is divided into sufficiently small discrete periods so that only one of the following three possible events can happen with associated probabilities in each period: (i) Arrival of a prospective customer with probability γ, (ii) completion of an order (i.e. departure) with probability β or (iii) no customer arrival or order completion with probability $\theta = 1 - \gamma - \beta$.

In terms of their sensitivity, customers are divided into two types. Leadtime sensitive customers (LS) are willing to pay more for receiving shorter leadtimes. Price sensitive customers (PS) are willing to wait more for paying less. Given that an ar-

rival occurs, the probability that we are dealing with a leadtime sensitive customer is ζ. Upon arrival of a prospective customer, the customer type is realized and the decision about what price/leadtime pair (p, L) to quote is made. There are lower and upper limits to quotable price and leadtime values. Upper limits represent the maximum amount a customer is willing to pay (p_{max}) and the maximum amount of time he is prepared to wait (L_{max}) for the product. Lower limits, on the other hand, represent the minimum possible price and leadtime that can be quoted for the product, which are denoted p_{min} and L_{min} respectively. For example, p_{min} can be the expected cost and L_{min} can be the expected time for processing an order.

Customers demand improved service at a lower price. Therefore, we model the probability that a customer accepts a given quote using a function that decreases both in price and leadtime. In particular, we use the S-shaped logistical response function, following Easton and Moodie (1999) and Watanapa and Techanitisawad (2005), and model the probability of acceptance of a (p, L) quote as follows.

$$P^i(p, L) = \begin{cases} \left[1 + \xi_0 e^{\xi_L^i (L - L_{min}) + \xi_p^i (p - p_{min})}\right]^{-1} & p_{min} \le p \le p_{max}, \\ & L_{min} \le L \le L_{max} \quad i = LS, PS \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Customers know ranges of possible price and leadtime offer (p_{min} and L_{min}) that they can get on the market. The probability that the firm's quote is accepted by customer type i decreases with factor ξ_p^i as the difference between offered price and p_{min} increases, and with factor ξ_L^i as the difference between quoted leadtime and L_{min} increases. The values of these sensitivity parameters can be determined by using historical data. We assume $\xi_p^{PS} \ge \xi_p^{LS}$ and $\xi_L^{PS} \le \xi_L^{LS}$. Note that for $\xi_0 > 0$, even when the firm offers p_{min} in combination with L_{min}, the probability that the customer accepts the quote is not equal to 1. This accounts for possibly other factors influencing the customers' choice.

If the customer accepts the quote, the order joins the production system and a revenue $R = p$ is earned; otherwise, the customer is lost. Whenever a prospective customer accepts the quote, it is referred to as "order".

We incorporate an upper bound K on the number of pending orders that the system can accommodate. This upper bound represents the financial or space-related limitations that may apply for a firm. For example, consider a firm that orders raw

material for processing customer orders. The raw material can consume available physical space in the production facility and/or budget. If there is no such limit applicable for the system of interest, K can be set to a value in the analysis that is large enough to allow (practically) all arriving customers to receive a quote. Thus, it reduces to an algorithmic parameter. On the other hand, since it is the parameter that limits the state space, the curse of dimensionality makes an MDP approach unfavorable for use in such cases.

A prospective customer is automatically rejected when K is reached. The due-date of an order is equal to its arrival time plus the quoted leadtime. When the remaining time of an order until the due-date becomes negative (i.e. the order is late), a tardiness penalty is incurred. The tardiness penalty is assumed to be the sum of a fixed and a variable cost. We assume that the customers want to get the product as soon as it is completed, therefore there is no penalty for completing orders before their due-dates. A variable processing cost is incurred in each period during which an order is being processed.

The orders that join the production system are put into a pool and the decision which of the pending orders to process next is made upon completion of the currently processed order based on the information about the remaining time of orders until their due-date. The pending orders are numbered according to their arrival sequence, so that number one is always taken for processing, when the special case of FCFS applies.

The above described joint pricing, leadtime quotation and priority dispatching problem is modeled as a discrete time Markov decision process where every period is a decision epoch. An infinite planning horizon is considered. The objective is maximizing of the long-run average profit per time unit.

4.2.1. State and action space

We define states of the MDP such that they describe the system at a period after the event occurrence (arrival, completion or no arrival nor completion) and before the decision. The state description needs to include an indicator variable to distinguish between three possible events since the decisions made at periods with arrival differ from those made at periods of completion and at all other periods there is no decision to make. Furthermore, the decisions at arrival periods are made by distinguishing between the two customer types. Thus, the state of the MDP is denoted as $(ind, n, r_1, r_2, ..., r_n)$ where $ind = 0$ represents a period with no order arrival or

completion, $ind = 1$ with an order completion, $ind = LS$ with a prospective LS customer arrival, and $ind = PS$ with a prospective PS customer arrival. n denotes the number of orders in the system. It is the number of orders a completed order leaves behind when $ind = 1$. When $ind = LS, PS$, it is the number of orders that an arriving prospective customer *finds* in the system. So, after receiving the firm's quote if the customer decides to place an order, the total number of orders in that period becomes $n+1$. r_i, $i = 1, 2, ..., n$ denote the remaining time of order i until the due-date. r_1 is the remaining time until the due-date for the order that is currently being served, if there is any. We denote the state space by S and the set of states in which the indicator variable is equal to $ind = 0, 1, LS, PS$ by S^{ind}.

Define $K(s)$ as the set of all possible actions in state s. For $ind = 1$, $K(s) = \{1, ..., n\}$, i.e. one of the n orders left behind the completed order is selected to be processed next. For $ind = LS, PS$, $K(s) = \{(p, L)|p \in \{p_{min}, ..., p_{max}\}, L \in \{L_{min}, ..., L_{max} + 1\}\}$. Due to (4.1), making quote $(p, L_{max} + 1)$ for any $p \in \{p_{min}, ..., p_{max}\}$ means that the customer is rejected. The values in the set of possible leadtimes to quote are integer numbers, since time is discrete. The values in the set of possible prices to quote can be real numbers. Finally, there is no decision, $K(s) = \{\}$, in dummy decision periods ($ind = 0$).

4.2.2. Transition probabilities

We present the transition probabilities for three different cases.

Case 0: *ind=0*

In this case, we consider a period t, in which no order arrival or completion occurred. Therefore, there is no decision to make. Thus, the next system state depends only on the current state and the event that will occur at the next decision epoch $(t+1)$. We define $p^0_{s,s'}$ as the probability that the system will be in state s' at the next decision epoch if the current state is s.

- For $s = (0, n, r_1 + 1, ..., r_n + 1)$, $\forall n \in \{1, ..., K - 1\}$,

$$p^0_{s,s'} = \begin{cases} \theta & s' = (0, n, r_1, r_2, ..., r_n) \\ \beta & s' = (1, n-1, r_2, ..., r_n) \\ \gamma \cdot \zeta & s' = (LS, n, r_1, r_2, ..., r_n) \\ \gamma \cdot (1 - \zeta) & s' = (PS, n, r_1, r_2, ..., r_n) \\ 0 & otherwise \end{cases} . \tag{4.2}$$

If the current state is $s = (0, n, r_1 + 1, ..., r_n + 1)$ and, at the next decision epoch $(t + 1)$, no order arrival or completion occurs (θ), the system moves to state $s' = (0, n, r_1, r_2, ..., r_n)$. The number of orders remains the same, while the remaining times of the orders until their due-date decrease by one. In s', the indicator also takes the value zero, since no order arrival or completion occurred in $t + 1$. If, at the next decision epoch, an order completion occurs (β), the system moves to state $s' = (1, n-1, r_2, ..., r_n)$. Since the completion of the order in processing has happened, s' includes the remaining times until the due-date, which have also decreased by one, only for $n-1$ orders left behind, i.e. $r_2, ..., r_n$.

If at the next decision epoch, a leadtime sensitive customer arrives $(\gamma \cdot \zeta)$, the system moves to state $s' = (LS, n, r_1, r_2, ..., r_n)$. The number of orders remains n although an arrival occurred, since for $ind = LS$, n indicates the number of orders that the customer *finds* in the system. At this point, neither the quote nor the decision by the customer whether to place an order or not are made. The logic for the case of a PS customer arrival is the same.

- For $s = (0, 0, null)$, an order completion in the next decision epoch is not possible.

$$p^0_{s,s'} = \begin{cases} 1-\gamma & s' = (0, 0, null) \\ \gamma\zeta & s' = (LS, 0, null) \\ \gamma\,(1-\zeta) & s' = (PS, 0, null) \\ 0 & otherwise \end{cases}. \tag{4.3}$$

- For $s = (0, K, r_1 + 1, ..., r_K + 1)$ an arriving customer is automatically rejected.

$$p^0_{s,s'} = \begin{cases} 1-\beta & s' = (0, K, r_1, r_2, ..., r_K) \\ \beta & s' = (1, K-1, r_2, ..., r_K) \\ 0 & otherwise \end{cases}. \tag{4.4}$$

Case 1: *ind=1*

In this case, we consider a period t in which an order completion occurred. Therefore, a decision about which order to process next is made. We define $p^1_{s,s'}(k)$ as the probability that the system will be in state s' at the next decision epoch if the decision k is made when the current state is s.

- For $s = (1, n, r_1 + 1, ..., r_n + 1)$, $\forall n \in \{1, ..., K-1\}$ and $\forall k \in K(s) = \{1, ..., n\}$,

$$p^1_{s,s'}(k) = \begin{cases} \theta & s' = (0, n, r_k, r_1, ..., r_{k-1}, r_{k+1}, ..., r_n) \\ \beta & s' = (1, n-1, r_1, ..., r_{k-1}, r_{k+1}, ..., r_n) \\ \gamma\zeta & s' = (LS, n, r_k, r_1, ..., r_{k-1}, r_{k+1}, ..., r_n) \\ \gamma(1-\zeta) & s' = (PS, n, r_k, r_1, ..., r_{k-1}, r_{k+1}, ..., r_n) \\ 0 & otherwise \end{cases}. \tag{4.5}$$

  If, at the next decision epoch $(t+1)$, no order arrival or completion occurs (θ), the system moves to state $s' = (0, n, r_k, r_1, ..., r_{k-1}, r_{k+1}, ..., r_n)$. In s', the remaining time of the orders until their due-dates are one period less than they were in state s. Furthermore, r_k is represented by the third state variable since order k is selected for processing.

  If an order completion occurs (β) at the next decision epoch, the system moves to state $s' = (1, n-1, r_1, ..., r_{k-1}, r_{k+1}, ..., r_n)$. Note that, since order k has been moved to processing at period t, the completed order is the one with a remaining time of r_k until the due-date at period $t+1$.

- For $s = (1, 0, null)$, which is the state where the system is left idle after an order completion, (4.3) holds.

$$p^1_{s,s'}(k) = p^0_{s,s'} \tag{4.6}$$

- $s = (1, K, r_1 + 1, ..., r_K + 1)$, is not possible, since a completed order cannot leave a full system behind.

Case 2: *ind=LS,PS*

In this case, we consider a period t in which a prospective customer with a certain type arrived. Therefore, a decision about which (p, L) pair to quote is made. The probability that the system visits a certain state at the next decision epoch is dependent on what decision is made in the current state, as well as whether the customer accepts the quote or not. We define $p^2_{s,s'}(k|a = 1)$ as the probability that the system will be in state s' at the next decision epoch if the decision k is made when the current state is s and the customer has accepted the quote. This probability is denoted by $p^2_{s,s'}(k|a = 0)$ if the customer has rejected quote k.

- For $s = (i, n, r_1 + 1, ..., r_n + 1)$, $i = LS, PS$, $\forall n \in \{1, ..., K-1\}$ and $\forall k \in K(s) =$

$\{(p, L)|p \in \{p_{min}, ..., p_{max}\}, L \in \{L_{min}, ..., L_{max} + 1\}\}$,

$$p^2_{s,s'}(k|a = 1) = \begin{cases} \theta & s' = (0, n + 1, r_1, r_2, ..., r_n, L - 1) \\ \beta & s' = (1, n, r_2, ..., r_n, L - 1) \\ \gamma\zeta & s' = (LS, n + 1, r_1, r_2, ..., r_n, L - 1) \\ \gamma\,(1 - \zeta) & s' = (PS, n + 1, r_1, r_2, ..., r_n, L - 1) \\ 0 & otherwise \end{cases} \tag{4.7}$$

If the quote (p, L) is made in state s, the customer accepted the quote, and no arrival or completion occurs (θ) in the next decision epoch $(t + 1)$, the system moves to state $s' = (0, n + 1, r_1, r_2, ..., r_n, L - 1)$. Since a new order with a leadtime of L is obtained and added to the pool of orders at period t, s' has $n + 1$ orders with the newly added order having $L - 1$ remaining periods until its due-date.

If the quoted (p, L) pair is not accepted by the customer, no new order joins the system. Thus, transition probabilities are the same as the ones given in (4.2).

$$p^2_{s,s'}(k|a = 0) = p^0_{s,s'} \tag{4.8}$$

The transition probabilities are conditioned on the decision of the customer whether to accept the quote or not. The probability that one or the other happens, given by (4.1), will come into consideration when we form the expression for the long-run average profit in equation (4.13).

- For $s = (i, 0, null)$, $i = LS, PS$,

$$p_{s,s'}(k|a = 1) = \begin{cases} \theta & s' = (0, 1, L - 1) \\ \beta & s' = (1, 0, null) \\ \gamma\zeta & s' = (LS, 1, L - 1) \\ \gamma\,(1 - \zeta) & s' = (PS, 1, L - 1) \\ 0 & otherwise. \end{cases} \tag{4.9}$$

If the quoted (p, L) pair is not accepted by the customer, the transition probabilities are the same as in (4.3).

$$p^2_{s,s'}(k|a = 0) = p^0_{s,s'} \tag{4.10}$$

- $s = (i, K, r_1 + 1, ..., r_K + 1)$, $i = LS, PS$, is not possible since a customer who arrives when the system is full is automatically rejected.

4.2.3. Cost structure

We consider the following cost structure to calculate the total tardiness cost incurred whenever a certain state is visited.

$$\begin{aligned} tcost_s = &\text{ (Number of orders with } r_i = 0)\cdot\text{(fixed cost)} \\ &+ \text{(Number of orders with } r_i \leq 0)\cdot\text{(unit variable cost)} \qquad \forall s \in S \end{aligned} \tag{4.11}$$

Note that the fixed cost is due as soon as the remaining time of an order until the due-date becomes 0. Furthermore, at this period and at each of the following periods during which the order stays in the system, a unit variable cost is incurred. The reason is the following. The state description does not include the remaining time until the due-date for the completed order in a completion period. Suppose that the remaining time until the due-date is "0" for the order in processing at a certain period and it is completed in the following period. This order experiences one period of lateness but a remaining time of -1 for this order is never represented in the system state.

Furthermore, a processing cost (c) is incurred in each period that an order spends in processing. Thus, this cost applies in each period unless the system is idle.

$$pcost_s = \begin{cases} c & \text{an order is in process} \\ 0 & \text{otherwise} \end{cases} \qquad \forall s \in S \tag{4.12}$$

4.2.4. Computing the optimal policy

Define g^* as the maximum long-run average profit per time unit. Let $g(R)$ be the long-run average profit per time unit that results from the actions of a stationary policy R. For obtaining the policy R with a $g(R)$ that is sufficiently close to g^*, we use the value-iteration algorithm. We compute the value function $V_t(s)$ from the following relationship.

$$V_t(s) = \begin{cases} -tcost_s - pcost_s + \sum_{s' \in S} p^0_{s,s'} V_{t-1}(s'), & \forall s \in S^0 \\ \max_{k \in K(s)} \left\{ -tcost_s - pcost_s + \sum_{s' \in S} p^1_{s,s'}(k) V_{t-1}(s') \right\}, & \forall s \in S^1 \\ \max_{k \in K(s)} \left\{ P^i(k) \left(p - tcost_s - pcost_s + \sum_{s' \in S} p^2_{s,s'}(k|a=1) V_{t-1}(s') \right) \right. & \\ \left. + (1 - P^i(k)) \left(-tcost_s - pcost_s + \sum_{s' \in S} p^2_{s,s'}(k|a=0) V_{t-1}(s') \right) \right\}, & \forall s \in S^i,\ i = LS, PS \end{cases} \tag{4.13}$$

The first line applies for the states where no order arrival or completion occurred and therefore no decision is made. The second line applies when a departure occurred

and the decision regarding which order to process next is made. The third equation applies for the states with an order arrival in which the decision regarding the price/leadtime quote is made. The probability that the customer accepts the quote depends on the decision. If the quote is accepted, a revenue that is equal to p is earned (line 3). The direct cost of visiting a state ($tcost_s + pcost_s$) appears in all lines of (4.13) and is incurred regardless of the decision.

Our algorithm stops at iteration t if the following criterion is satisfied:

$$\frac{\max_{s\in S}\{V_t(s) - V_{t-1}(s)\} - \min_{s\in S}\{V_t(s) - V_{t-1}(s)\}}{\min_{s\in S}\{V_t(s) - V_{t-1}(s)\}} \leq \epsilon \tag{4.14}$$

which guarantees $\frac{g^* - g(R(t))}{g^*} \leq \epsilon$, where $R(t)$ is the policy at iteration t. For further details about the approach, see Tijms (2003, Chapter 6).

4.3. Performance of the optimal policy

The long-run average profit per time unit is not the only measure of interest related to the performance of the optimal policy. There are a number of system-, customer- and tardiness-related measures that provide useful managerial insights. We focus on the utilization level (probability that the system is not idle) as a system-related performance measure, the percentage of LS and PS customers that accept firm's quote as customer-related performance measures and the on-time probability of an order (also referred as service level) as a tardiness-related measure.

Once the optimal policy is obtained, one can model the system as a Markov chain (MC) that evolves under this policy and compute further measures. Solving the equilibrium distribution of the MC is an approach towards this end, since it provides all the information on the system behavior. However, usually it is not necessary to have such detailed information for computing a performance measure. It is also computationally demanding to obtain for large Markov chains. Another approach is using value-iteration. Tijms (2003, Chapter 6) discusses, and gives examples for showing, the usefulness of this approach for computing the performance measures of a Markov chain. We follow the latter approach.

4.3.1. Utilization level

The steady-state probability of having m orders in the system is denoted π_m. For a given m, we use the following relationship to compute the value function $V_t^m(s)$

recursively.

$$V_t^m(s) = \begin{cases} \mathbf{1}_{\{n=m\}} + \sum_{s' \in S} p^0_{s,s'} V^m_{t-1}(s'), & \forall s \in S^0 \\ \mathbf{1}_{\{n=m\}} + \sum_{s' \in S} p^1_{s,s'}(k^*) V^m_{t-1}(s'), & \forall s \in S^1 \\ P^i(k^*)\left(\mathbf{1}_{\{n+1=m\}} + \sum_{s' \in S} p^2_{s,s'}(k^*|a=1) V^m_{t-1}(s')\right) & \\ + (1 - P^i(k^*))\left(\mathbf{1}_{\{n=m\}} + \sum_{s' \in S} p^2_{s,s'}(k^*|a=0) V^m_{t-1}(s')\right), & \forall s \in S^i,\ i = LS, PS \end{cases} \tag{4.15}$$

where

$$\mathbf{1}_{\{n=m\}} = \begin{cases} 1 & n \text{ in state } s \text{ is equal to } m \\ 0 & \text{otherwise.} \end{cases}$$

The relationship between π_m and equation (4.15) is as follows.

$$\min_{s \in S}\{V_t^m(s) - V_{t-1}^m(s)\} \leq \pi_m \leq \max_{s \in S}\{V_t^m(s) - V_{t-1}^m(s)\} \tag{4.16}$$

The utilization level is $\rho = 1 - \pi_0$.

4.3.2. Percentage of LS and PS customers who accept the quote

The percentage of LS (PS) customers that accept the quote is denoted by ϕ_{LS} (ϕ_{PS}). We compute the value function $V_t^{type}(s)$ with $type = LS$ ($type = PS$) from:

$$V_t^{type}(s) = \begin{cases} \sum_{s' \in S} p^0_{s,s'} V^{type}_{t-1}(s'), & \forall s \in S^0 \\ \sum_{s' \in S} p^1_{s,s'}(k^*) V^{type}_{t-1}(s'), & \forall s \in S^1 \\ P^i(k^*)\left(\mathbf{1}_{\{i=type\}} + \sum_{s' \in S} p^2_{s,s'}(k^*|a=1) V^{type}_{t-1}(s')\right) & \\ + (1 - P^i(k^*))\left(\sum_{s' \in S} p^2_{s,s'}(k^*|a=0) V^{type}_{t-1}(s')\right), & \forall s \in S^i,\ i = LS, PS \end{cases} \tag{4.17}$$

here, $\mathbf{1}_{\{i=type\}}$ takes value 1 if $i = type$ and 0 otherwise.

The solution of the value-iteration algorithm gives the steady-state probability that an LS (PS) customer places an order. In case of a leadtime sensitive customer, this probability is equal to $\gamma \cdot (1 - \pi_K) \cdot \zeta \cdot \phi_{LS}$, namely, to the probability that a prospective customer arrives at a period in which the system capacity is not yet reached ($\gamma \cdot (1 - \pi_K)$), multiplied by the probability that it is an LS customer (ζ) and by the percentage of LS customers that accept the quote (ϕ_{LS}). A similar logic applies for the steady-state probability that a price sensitive customer places an

order. Thus, the relationship between ϕ_{LS} and ϕ_{PS} with equation (4.17) is:

$$\min_{s\in S}\{V_t^{LS}(s) - V_{t-1}^{LS}(s)\} \leq \gamma \cdot (1-\pi_K) \cdot \zeta \cdot \phi_{LS} \leq \max_{s\in S}\{V_t^{LS}(s) - V_{t-1}^{LS}(s)\} \quad (4.18)$$

$$\min_{s\in S}\{V_t^{PS}(s) - V_{t-1}^{PS}(s)\} \leq \gamma \cdot (1-\pi_K) \cdot (1-\zeta) \cdot \phi_{PS} \leq \max_{s\in S}\{V_t^{PS}(s) - V_{t-1}^{PS}(s)\} \quad (4.19)$$

4.3.3. On-time probability

For obtaining the on-time probability (η), we compute the value function $V_t'(s)$ recursively from the following relationship.

$$V_t'(s) = \begin{cases} \mathbf{1}_{\{r_1\geq 1\}} + \sum_{s'\in S} p_{s,s'}^0 V_{t-1}'(s'), & \forall s \in S^0 \\ \mathbf{1}_{\{r_{k^*}\geq 1\}} + \sum_{s'\in S} p_{s,s'}^1(k^*) V_{t-1}'(s'), & \forall s \in S^1 \\ \mathbf{1}_{\{r_1\geq 1\}} + P^i(k^*)\left(\sum_{s'\in S} p_{s,s'}^2(k^*|a=1) V_{t-1}'(s')\right) & \\ + (1-P^i(k^*))\left(\sum_{s'\in S} p_{s,s'}^2(k^*|a=0) V_{t-1}'(s')\right), & \forall s \in S^i,\ i = LS, PS \end{cases} \quad (4.20)$$

where

$$\mathbf{1}_{\{r_1\geq 1\}} = \begin{cases} 1 & r_1 \geq 1 \text{ in state } s \\ 0 & \text{otherwise} \end{cases}$$

because, given that an order is completed at an arbitrary period t, the completion is on-time if the remaining time of the order until the due-date is one or larger at period $t-1$.

The solution of the value-iteration algorithm gives the steady-state probability that, if an order is completed in an arbitrary period, it is on-time. Since a completion can only occur when the system is not idle (with probability $1-\pi_0$), the relationship between η and equation (4.20) is:

$$\min_{s\in S}\{V_t'(s) - V_{t-1}'(s)\} \leq \eta \cdot (1-\pi_0) \leq \max_{s\in S}\{V_t'(s) - V_{t-1}'(s)\} \quad (4.21)$$

(4.14) also applies as the stopping criterion for iterations described in this section.

4.4. Numerical study

In the numerical study, we first analyze the decisions made by marketing under a sequential approach. Secondly, we answer the following research questions: (1) How much can the profitability be increased? (2) How are the system utilization, the service level of the firm and the selection of different customer types affected when the simultaneous optimization approach, rather than one of the sequential approaches, is considered? (3) How do the tardiness penalty structure and the market-related characteristics affect the answers to the first two questions?

We take $K = 5$ and set parameters that define the arrival and service processes respectively as $\gamma = 0.5$ and $\beta = 0.4$. We select an arrival probability that is larger than the order completion probability to generate a high frequency of prospective customer arrivals so that the system has the chance to obtain more customers if better price/leadtime quotes are made. The set of quotable prices and leadtimes are $p \in \{p_{min}, p_{min}+1, ..., p_{max}\}$ (i.e. integer values from p_{min} to p_{max}), $L \in \{L_{min}, L_{min}+1, ..., L_{max}+1\}$ with $p_{min} = L_{min} = [E[Y]] = [1/\beta] = 3$ and $p_{max} = L_{max} = 10$. $E[Y]$ denotes the expected processing time of an order.

We use tardiness penalty structures, in which the tardiness penalty involves: only a fixed cost, only a variable cost and both a fixed and a variable cost. Since $K = 5$ and $E[Y] = 2.5$, the expected production leadtime of the K^{th} arrived pending order is 12.5 under a FCFS processing. If this order is quoted a leadtime of $L_{max} = 10$, the expected tardiness is 2.5. We take (fixed cost, unit variable cost) pairs $\in \{(6,0),(3,1),(0,2)\}$ all of which result in the same tardiness cost when an order is late by 3 periods. The processing cost per time unit is $c = 0$. The stopping criterion of the value-iteration algorithm is set to $\epsilon = 10^{-5}$.

We assume that there is a common price sensitivity parameter (ξ_p) as well as a common leadtime sensitivity parameter (ξ_L) defined for the system. We then consider different sensitivities of two customer types to be specified based on a parameter ($\kappa \in [0,1]$) that determine the heterogeneity between the two customer types:

$$\begin{aligned} \xi_L^{LS} &= \xi_L + \kappa \cdot \xi_L \qquad & \xi_L^{PS} &= \xi_L - \kappa \cdot \xi_L \\ \xi_p^{LS} &= \xi_p - \kappa \cdot \xi_p \qquad & \xi_p^{PS} &= \xi_p + \kappa \cdot \xi_p \end{aligned} \tag{4.22}$$

where $\kappa = 0$ means that customer types do not differ in their sensitivity to price and leadtime, i.e. that there is a single customer type, while $\kappa = 1$ makes LS customers only sensitive to leadtime and PS customers only sensitive to price. We consider

$\xi_0 = 0.1$, $\xi_L = 0.75$ and $\xi_p = 0.75$ as a base case for customer sensitivity (Case 1). These values are selected to be similar to the ones considered by Easton and Moodie (1999) and Watanapa and Techanitisawad (2005). Both of these studies take $\xi_0 = 0.1$. Easton and Moodie (1999) take $\xi_L = 0.5$ and $\xi_p = 0.75$ while Watanapa and Techanitisawad (2005) take ξ_L values between 0.3 and 0.9 and ξ_p between 0.3 and 0.75. Then we double the leadtime sensitivity (Case 2), the price sensitivity (Case 3) and finally both of them (Case 4), for numerically exercising the impact of customer sensitivity parameters. In each case, we also vary the proportion of LS customer arrivals $\zeta \in \{0.05, 0.5, 0.95\}$ and the contrast parameter $\kappa \in \{0.2, 0.6, 1\}$ respectively. These variations result in nine sub-cases.

The sequential approaches and their computation are explained in the following:

- Marketing (Step 1): Marketing makes (p, L) quotes to arriving prospective customers with the objective of maximizing the long-run average profit per time unit. It uses information about the type of a prospective customer, the current number of orders in the manufacturing system and the upper bound on the number of pending orders. On the other hand, it has no information regarding the sequence in which the orders are processed, i.e. about the remaining times until the due-date for orders that are currently in the system. Marketing assumes FCFS in order processing and keeps track of the remaining time of orders until the due-date under this assumption for evaluating the tardiness penalties. This problem is a special case of the optimization problem described in Section 4.2. Hence, in order to compute the optimal solution to this problem, we run our MDP together with the constraint that $K(s) = \{1\}$ in all states with $ind = 1$.
- Manufacturing (Step 2):

  Approach 1. Manufacturing processes orders based on FCFS. The optimal solution to this approach is the one obtained in the first step because the order processing policy is in accordance with the assumption of the marketing department.

  Approach 2. Manufacturing processes orders based on EDD. The EDD rule is a simple and effective priority rule that is often used in practice (Keskinocak and Tayur, 2004). The optimal solution to this approach is computed by running the MDP in which the (p, L) pairs taken from the first step and the priority dispatching decisions are fixed in accordance with the EDD rule. This means that for any state $s = (1, n, r_1, ..., r_n)$ $K(s) = \{k\}$, where k is chosen such that $r_k = \min\{r_1, r_2, ..., r_n\}$. For orders with equal remaining

time until the due-date and for orders with negative remaining time until the due-date, FCFS applies.

Approach 3. Manufacturing processes orders according to the optimal priority dispatching policy. The optimal solution to this approach is also obtained by running the MDP with the (p, L) pairs taken from Step 1. In this case there is no restriction on the dispatching decisions.

Note that all three approaches involve simultaneous optimization of dynamic price and leadtime quotation and coordination of marketing with manufacturing. Their sub-optimality, if any, comes from not harmonizing these decisions with dispatching decisions, i.e. from not collaborating for a joint policy.

4.4.1. Analysis of marketing decisions under sequential approaches

We analyze the price and leadtime quotes made by the marketing department of the firm under a sequential approach. For the base case, Table 4.1 gives the optimal (p, L) decisions that distinguish between different customer types (PS and LS) and different system loads (n). It presents the effects of the contrast parameter (κ), the proportion of arriving LS customers (ζ) and the tardiness penalty structure. Whenever it is not varied, $\kappa = 0.6$, $\zeta = 0.5$ and the penalty structure is $(3, 1)$. Figures 4.1-4.4 visualize the effects observed in Table 4.1. The level at which a parameter is fixed for obtaining a visualization is indicated in the figure's caption.

	$\kappa = 0.2$		$\kappa = 0.6$		$\kappa = 1$		$\zeta = 0.05$		$\zeta = 0.5$		$\zeta = 0.95$		(6,0)		(3,1)		(0,2)	
n	p	L	p	L	p	L	p	L	p	L	p	L	p	L	p	L	p	L
0 PS	5	3	4	5	4	10	4	5	4	5	5	4	4	5	4	5	4	5
1 PS	5	6	5	7	5	10	4	8	5	7	5	9	4	9	5	7	5	7
2 PS	5	8	5	10	6	10	4	10	5	10	6	10	5	10	5	10	5	10
3 PS	6	10	7	10	8	10	6	10	7	10	8	10	6	10	7	10	7	10
4 PS	8	10	9	10	10	10	8	10	9	10	10	10	8	10	9	10	10	10
0 LS	6	3	10	3	10	3	10	3	10	3	10	3	10	3	10	3	10	3
1 LS	8	3	10	3	10	4	10	3	10	3	10	3	10	3	10	3	10	3
2 LS	10	4	10	4	10	5	10	4	10	4	10	5	10	3	10	4	10	5
3 LS	10	6	10	7	10	7	10	6	10	7	10	7	10	3	10	7	10	8
4 LS	10	9	10	9	10	10	10	8	10	9	10	10	10	3	10	9	10	10

Table 4.1 Effect of κ, ζ and the tardiness penalty structure on marketing decisions (Case 1: $\xi_L = 0.75$ and $\xi_p = 0.75$)

In most cases, the optimal price and leadtime values to quote are non-decreasing in the number of orders in the system (Figure 4.1). However, in some cases where tardiness is penalized according to (6,0), an increase in n decreases the optimal leadtime quote. The lower leadtime quote is combined with a higher price quote. Recall that the probability of a customer placing an order is positively affected

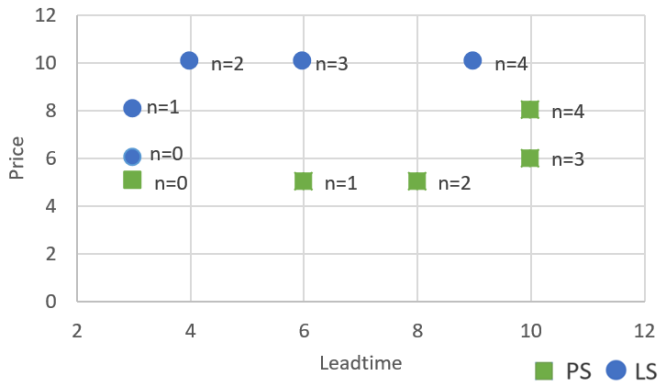

Figure 4.1 Effect of system load ($\kappa = 0.2$)

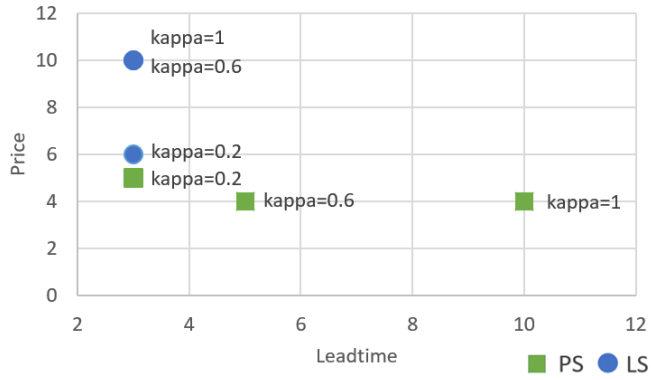

Figure 4.2 Effect of κ ($n = 0$)

by a decrease in price or leadtime. Thus, the optimal solution suggests making a lower leadtime quote under a higher system load actually for being able to charge a high price while keeping the probability of acceptance at a certain level. A similar logic applies for the optimal quotes in some cases under the (0,2) penalty structure. The optimal price quote decreases with an increase in n and the lower price quote is combined with a higher leadtime quote. These cases are mostly observed when $\kappa = 0.2$, in which setting the firm cannot benefit from the heterogeneity of customer preferences. Apparently, when the system is subject to the (0,2) penalty structure, increasing customers' willingness to wait becomes more valuable in these cases, than obtaining higher revenues. On the other hand, when the (6,0) structure applies, an increased willingness to pay is preferable although it comes together with a higher probability of tardiness.

The heterogeneity of two customer types (κ) significantly affects the optimal quotes (Figure 4.2). As expected, the quotes for the two customer types become more and more contrasting as κ increases. For example, when $\kappa = 0.2$, a prospective customer that arrives at an idle system ($n = 0$) receives the lowest leadtime quote possible, regardless of the customer type. If he is a price sensitive customer, the quoted price

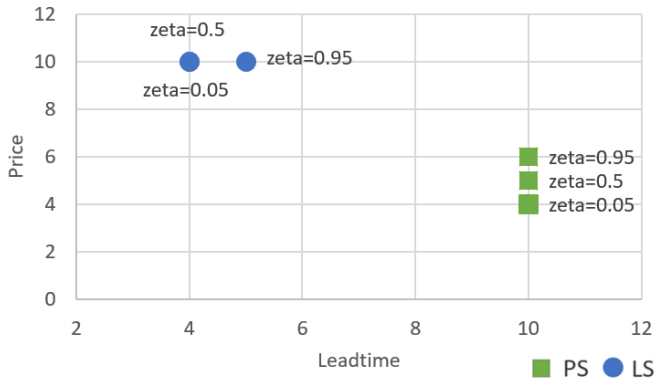

Figure 4.3 Effect of ζ ($n = 2$)

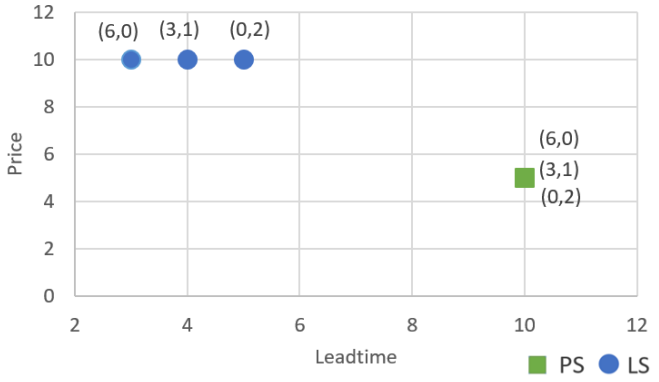

Figure 4.4 Effect of tardiness penalty structure ($n = 2$)

is 5, otherwise it is 6. If $\kappa = 0.6$, a PS customer receives the quote (4,5), since such a customer is more sensitive to price and less sensitive to leadtime in this case. For an LS customer, it becomes more important to be quoted the minimum possible leadtime and less important what the price is. Therefore, the optimal quote for an LS customer changes to (10,3).

The effect of an increasing proportion of LS customer arrivals (ζ) under (3,1) cost structure and $\kappa = 0.6$ is in the direction of increasing the price quote to price sensitive customers and the leadtime quote to leadtime sensitive customers (Figure 4.3). When moving from the (6,0) penalty structure to (3,1) and from (3,1) to the (0,2) penalty structure, the leadtime quote that LS customers receive increases, starting from $n = 2$ (Figure 4.4). This means that it becomes harder to offer low leadtimes when the system load, as well as the relative value of variable cost in the tardiness penalty, is higher.

4.4.2. Analysis of the potential improvement via simultaneous optimization

In order to answer the first research question, we compare the maximum long-run average profit obtained under the sequential approaches to the one obtained using the simultaneous approach. Define $\Delta_{Seq,i}$, the percentage sub-optimality of i^{th} sequential approach, as:

$$\Delta_{Seq,i} = \frac{g^* - g_{Seq,i}}{g^*} \cdot 100$$

where $g_{Seq,i}$ denotes the maximum long-run average profit per time unit under the i^{th} sequential approach.

Table 4.2 presents the sub-optimality of sequential approaches together with the maximum long-run average profit obtained by using the simultaneous approach (g^*) for all sensitivity cases. In each case, first the proportion of LS customer arrivals (ζ) and then the customer heterogeneity (κ) is varied. The highest sub-optimality value observed for a sequential approach within each case is marked in bold.

When tardiness of orders is penalized according to the variable only cost structure (0,2), all sequential approaches perform very close to the optimal. Under this penalty structure, the highest sub-optimality value is 1.08%, which is observed for the first sequential approach in case 4.7. Furthermore, the sub-optimality of all approaches is close to 0% when customers are similar in their sensitivity to price and leadtime ($\kappa = 0.2$). This means that it is nearly optimal to simply process orders based on FCFS. An explanation for this observation is the following. Under this penalty structure, the optimal rule for order processing is EDD (see Duenyas (1995)). This can also be seen in Table 4.2 from the sub-optimality of the second sequential approach (EDD) always being the same as the one of the third approach (optimal sequencing) under (0,2). Furthermore, it is known that FCFS and EDD are equivalent when leadtimes are constant while an increasing leadtime variability increases their difference (see e.g. Öner Közen and Minner (2016c)). When the customers are alike in their sensitivity to leadtime, the variability in quoted leadtimes is low. As a result, the sub-optimality due to employing FCFS in order processing and thus the sub-optimality due to assuming FCFS in the first step of a sequential optimization approach, is negligible.

Increasing κ and the relative value of the fixed cost in the tardiness penalty mostly increases the sub-optimality. The inefficiencies resulting from not harmonizing the

Case	Sub-case	κ	ζ	(6,0) $\Delta_{Seq,1}$	(6,0) $\Delta_{Seq,2}$	(6,0) $\Delta_{Seq,3}$	(6,0) g^*	(3,1) $\Delta_{Seq,1}$	(3,1) $\Delta_{Seq,2}$	(3,1) $\Delta_{Seq,3}$	(3,1) g^*	(0,2) $\Delta_{Seq,1}$	(0,2) $\Delta_{Seq,2}$	(0,2) $\Delta_{Seq,3}$	(0,2) g^*
	1	0.2	0.05	1.58	1.64	0.94	0.764	0.06	0.06	0.01	0.764	0.00	0.00	0.00	0.779
	2		0.5	2.16	2.24	1.31	0.876	0.19	0.17	0.07	0.861	0.02	0.01	0.01	0.870
	3		0.95	2.57	2.58	1.22	0.990	0.16	0.15	0.02	0.960	0.00	0.00	0.00	0.964
	4	0.6	0.05	1.71	1.85	0.50	0.842	0.35	0.23	0.12	0.829	0.21	0.17	0.17	0.830
1	5		0.5	4.52	4.72	0.67	1.317	0.78	**0.64**	**0.27**	1.235	0.34	0.30	0.30	1.225
	6		0.95	**7.21**	**7.21**	**4.33**	1.759	0.23	0.22	0.01	1.615	0.01	0.01	0.01	1.602
	7	1	0.05	3.11	3.15	1.51	1.132	0.79	0.22	0.04	1.093	0.63	0.33	0.33	1.087
	8		0.5	6.87	6.44	2.98	1.795	**0.93**	0.50	0.23	1.671	**0.68**	**0.56**	**0.56**	1.660
	9		0.95	6.20	6.19	3.67	2.293	0.15	0.15	0.01	2.147	0.01	0.01	0.01	2.143
	1	0.2	0.05	1.56	1.56	1.01	0.740	0.09	0.09	0.00	0.733	0.00	0.00	0.00	0.742
	2		0.5	2.23	2.23	1.29	0.866	0.25	0.25	0.03	0.845	0.00	0.00	0.00	0.851
	3		0.95	2.59	2.59	1.22	0.989	0.31	0.31	0.00	0.958	0.00	0.00	0.00	0.961
	4	0.6	0.05	1.32	1.41	0.71	0.712	0.19	0.12	0.00	0.708	0.05	0.03	0.03	0.713
2	5		0.5	3.57	3.67	0.41	1.258	0.29	0.28	0.14	1.189	0.02	0.02	0.02	1.187
	6		0.95	7.66	7.66	3.20	1.756	0.23	0.23	0.21	1.603	0.00	0.00	0.00	1.594
	7	1	0.05	3.15	3.27	1.41	1.131	0.79	0.52	0.38	1.090	**0.62**	0.20	0.20	1.082
	8		0.5	**8.42**	**8.26**	5.71	1.781	**0.94**	**0.91**	**0.87**	1.616	0.45	**0.40**	**0.40**	1.599
	9		0.95	7.45	7.45	**7.12**	2.207	0.08	0.08	0.01	2.025	0.01	0.01	0.01	2.015
	1	0.2	0.05	1.07	1.08	0.66	0.579	0.05	0.05	0.00	0.580	0.00	0.00	0.00	0.587
	2		0.5	1.05	1.07	0.81	0.583	0.05	0.05	0.02	0.586	0.01	0.01	0.01	0.598
	3		0.95	0.74	0.75	0.63	0.592	0.02	0.02	0.00	0.602	0.00	0.00	0.00	0.616
	4	0.6	0.05	2.36	2.37	1.07	0.731	0.24	0.07	0.02	0.716	0.19	0.11	0.11	0.718
3	5		0.5	3.26	3.17	2.05	0.856	1.03	0.26	0.04	0.825	0.79	0.41	0.41	0.824
	6		0.95	2.82	2.81	1.40	0.988	0.31	0.27	0.17	0.954	0.10	0.08	0.08	0.956
	7	1	0.05	3.22	3.22	1.40	0.963	**1.16**	0.29	0.10	0.929	**1.07**	**0.62**	**0.62**	0.923
	8		0.5	**7.19**	**6.74**	2.94	1.715	0.91	**0.88**	**0.62**	1.587	0.48	0.44	0.44	1.577
	9		0.95	6.21	6.21	**3.69**	2.290	0.14	0.14	0.00	2.144	0.00	0.00	0.00	2.140
	1	0.2	0.05	0.23	0.23	0.18	0.500	0.00	0.00	0.00	0.515	0.00	0.00	0.00	0.532
	2		0.5	0.52	0.52	0.41	0.540	0.01	0.01	0.00	0.548	0.00	0.00	0.00	0.563
	3		0.95	0.80	0.80	0.55	0.580	0.01	0.01	0.00	0.585	0.00	0.00	0.00	0.597
	4	0.6	0.05	1.10	1.13	0.58	0.599	0.15	0.07	0.01	0.598	0.11	0.10	0.10	0.603
4	5		0.5	2.23	2.33	0.65	0.787	0.54	0.26	0.00	0.767	0.31	0.27	0.27	0.766
	6		0.95	2.67	2.67	1.31	0.981	0.34	0.31	0.00	0.948	0.03	0.03	0.03	0.951
	7	1	0.05	3.22	3.34	1.23	0.963	**1.31**	**0.90**	**0.75**	0.925	**1.08**	**0.45**	**0.45**	0.918
	8		0.5	**8.56**	**8.64**	1.16	1.704	0.88	0.83	0.30	1.539	0.27	0.26	0.26	1.522
	9		0.95	7.49	7.49	**7.17**	2.205	0.07	0.07	0.00	2.022	0.00	0.00	0.00	2.012

Table 4.2 Sub-optimality of sequential approaches (Case 1: $\xi_L = 0.75$ and $\xi_p = 0.75$, Case 2: $\xi_L = 1.5$ and $\xi_p = 0.75$, Case 3: $\xi_L = 0.75$ and $\xi_p = 1.5$, Case 4: $\xi_L = 1.5$ and $\xi_p = 1.5$)

price/leadtime quotation decisions with the dispatching of orders lead to a considerable sub-optimality for all sequential approaches when only a fixed cost is incurred in case of tardiness and the contrast in the customer sensitivities to price and leadtime is high. Under the fixed penalty setting, when the system receives equally likely arrivals of two totally different customer types (sub-case 8), the first two sequential approaches perform poorly. The situation improves if the order processing sequence is optimized in the second step of a simultaneous approach. However, the third approach provides less benefit if 95% of arriving customers are LS type customers (sub-case 9). The effects observed in sub-cases 8 and 9 are intensified when leadtime sensitivity is large (cases 2 and 4).

Note that, there is value in optimizing the sequence in which orders are processed, although FCFS is assumed when making (p, L) quotes. This can be seen from the sub-optimality of sequential approach number three being considerably smaller than the sub-optimality of the other two approaches.

4.4.3. Analysis of the impact of simultaneous optimization on KPI's

In order to answer the second research question, we observe the levels of related KPI's. Table 4.3 gives the system utilization (ρ), the percentage of LS (ϕ_{LS}) and PS (ϕ_{PS}) customers that accept the quote under a sequential approach (Seq) and the simultaneous approach (Sim). Since in all three sequential approaches the (p, L) quotes are made by marketing in the same way, ρ, ϕ_{LS} and ϕ_{PS} are the same for all. The on-time probability of orders (η) differs between three approaches that differ in the dispatching policy manufacturing follows. Table 4.4 gives η under the sequential approaches ($Seq, i,\ i = 1, 2, 3$) and the simultaneous approach (Sim). The tables present the results for the base case ($\xi_L = 0.75$ and $\xi_p = 0.75$) in which ζ and κ are varied in the respective order. The arrows on the right of the values in column Sim indicate the direction of change compared to the values in columns Seq and $Seq, 3$ in Tables 4.3 and 4.4 respectively. Tables for cases 2, 3 and 4 are not provided in the manuscript, however our conclusions hold for all.

For a profit maximizing MTO firm, the typical trade-off is between obtaining more orders to increase revenue and being able to attain the promised leadtimes to avoid tardiness costs. A joint optimization enables the evaluation of this trade-off in the most informed way. In other words, it enables accurate judgment whether the benefits due to obtaining more orders outweighs the loss due to more frequently failing to complete them within leadtimes or vice versa.

Sub-case	κ	ζ		(6,0) Seq	(6,0) Sim		(3,1) Seq	(3,1) Sim		(0,2) Seq	(0,2) Sim	
1	0.2	0.05	ρ	0.53	0.55	↑	0.50	0.50	↑	0.50	0.50	↑
			ϕ_{LS}	0.45	0.45	–	0.44	0.45	↑	0.40	0.40	↑
			ϕ_{PS}	0.43	0.44	↑	0.40	0.40	↓	0.40	0.40	–
2	0.2	0.5	ρ	0.52	0.55	↑	0.52	0.52	↑	0.50	0.50	↑
			ϕ_{LS}	0.44	0.45	↑	0.43	0.44	↑	0.40	0.40	–
			ϕ_{PS}	0.39	0.44	↑	0.39	0.39	↓	0.40	0.40	–
3	0.2	0.95	ρ	0.47	0.56	↑	0.44	0.44	↑	0.44	0.44	↑
			ϕ_{LS}	0.38	0.45	↑	0.35	0.35	↑	0.34	0.34	–
			ϕ_{PS}	0.41	0.43	↑	0.43	0.43	↑	0.43	0.43	↑
4	0.6	0.05	ρ	0.57	0.60	↑	0.57	0.57	↑	0.57	0.57	↑
			ϕ_{LS}	0.55	0.55	–	0.47	0.52	↑	0.44	0.49	↑
			ϕ_{PS}	0.45	0.48	↑	0.45	0.45	↓	0.45	0.45	↓
5	0.6	0.5	ρ	0.58	0.61	↑	0.54	0.55	↑	0.53	0.54	↑
			ϕ_{LS}	0.55	0.55	–	0.49	0.49	↑	0.47	0.47	↑
			ϕ_{PS}	0.39	0.44	↑	0.38	0.39	↑	0.38	0.40	↑
6	0.6	0.95	ρ	0.56	0.64	↑	0.54	0.54	↑	0.53	0.53	↑
			ϕ_{LS}	0.46	0.54	↑	0.44	0.44	–	0.44	0.44	↑
			ϕ_{PS}	0.22	0.34	↑	0.23	0.23	–	0.24	0.24	–
7	1	0.05	ρ	0.67	0.70	↑	0.67	0.67	↑	0.66	0.66	↑
			ϕ_{LS}	0.85	0.91	↑	0.75	0.78	↑	0.60	0.69	↑
			ϕ_{PS}	0.52	0.55	↑	0.52	0.52	↓	0.53	0.52	↓
8	1	0.5	ρ	0.67	0.74	↑	0.64	0.65	↑	0.63	0.64	↑
			ϕ_{LS}	0.71	0.86	↑	0.65	0.66	↑	0.61	0.63	↑
			ϕ_{PS}	0.36	0.39	↑	0.39	0.38	↓	0.40	0.39	↓
9	1	0.95	ρ	0.70	0.79	↑	0.68	0.68	↑	0.68	0.68	↑
			ϕ_{LS}	0.59	0.71	↑	0.56	0.56	–	0.56	0.56	–
			ϕ_{PS}	0.14	0.13	↓	0.15	0.15	–	0.15	0.15	–

Table 4.3 ρ, ϕ_{LS} and ϕ_{PS} under a sequential and the simultaneous approach (Case 1: $\xi_L = 0.75$, $\xi_p = 0.75$)

Table 4.3 shows that the simultaneous approach increases the system utilization typically by increasing both ϕ_{PS} and ϕ_{LS} or either one of them without decreasing the other. In some cases where the system receives equally likely arrivals of two dissimilar customer types, e.g. sub-case 5, the simultaneous optimization enables the firm to attract more customers, as well as to improve its service level (see Tables 4.3 and 4.4).

Similar results are observed in cases 2.5, 3.8 and 4.8. One can notice that, in all of these cases, the simultaneous approach prefers PS customers, i.e. either increases ϕ_{PS} more than it increases ϕ_{LS} or increases ϕ_{PS} while decreasing ϕ_{LS}. Although the number of cases where the simultaneous approach prefers LS customers is higher, this clearly indicates that higher profits are not necessarily a result of being able to attract more high margin customers. In most of the cases, on the other hand, the simultaneous approach suggests that lowering firm's service level (η) is a better way towards achieving higher profits.

Typically, the highest service levels are obtained when the order processing sequence is optimized in a second step after (p, L) quotes are made by marketing under the FCFS assumption ($Seq, 3$ in Table 4.4). This can be explained as follows. Although

			(6,0)				(3,1)				(0,2)			
Sub-case	κ	ζ	$Seq,1$	$Seq,2$	$Seq,3$	Sim	$Seq,1$	$Seq,2$	$Seq,3$	Sim	$Seq,1$	$Seq,2$	$Seq,3$	Sim
1	0.2	0.05	0.738	0.738	0.742	0.735↓	0.771	0.771	0.772	0.772↓	0.774	0.774	0.774	0.774↓
2	0.2	0.5	0.699	0.699	0.705	0.674↓	0.710	0.710	0.712	0.708↓	0.734	0.734	0.734	0.734↓
3	0.2	0.95	0.632	0.632	0.644	0.617↓	0.690	0.690	0.693	0.688↓	0.695	0.695	0.695	0.695↓
4	0.6	0.05	0.874	0.873	0.882	0.862↓	0.885	0.885	0.888	0.883↓	0.888	0.888	0.888	0.883↓
5	0.6	0.5	0.644	0.642	0.681	0.678↓	0.712	0.712	0.723	0.729↑	0.732	0.732	0.733	0.737↑
6	0.6	0.95	0.561	0.561	0.598	0.511↓	0.603	0.603	0.613	0.612↓	0.614	0.614	0.615	0.614↓
7	1	0.05	0.902	0.902	0.914	0.896↓	0.912	0.914	0.918	0.916↓	0.932	0.933	0.933	0.927↓
8	1	0.5	0.686	0.691	0.730	0.623↓	0.758	0.761	0.770	0.759↓	0.786	0.786	0.787	0.772↓
9	1	0.95	0.616	0.616	0.650	0.553↓	0.659	0.659	0.666	0.666↓	0.660	0.660	0.661	0.659↓

Table 4.4 On-time probability of orders (η) under the sequential approaches and the simultaneous approach (Case 1: $\xi_L = 0.75$, $\xi_p = 0.75$)

the firm has the flexibility to prioritize orders optimally, e.g. to give a higher priority to processing an order of a high margin customer, this is not taken into account when making quotes to customers. This leads to making cautious leadtime quotes and to a lower percentage of acceptances than it can potentially be. The failure to efficiently use the potential of extracting a higher revenue from (e.g. LS) customers by making lower leadtime quotes might also lead to more cautious price quotes (e.g. to PS customers). A lower success in obtaining customers makes it less challenging to achieve high service levels.

4.5. Conclusions

The joint optimization problem of a profit maximizing firm that quotes price/leadtime pairs to two types of prospective customers who differ in their sensitivities to price and leadtime and dispatches placed orders based on the due-dates was investigated. The problem was modeled as a Markov decision process (MDP) and the optimal pricing, leadtime quotation and dispatching policy was numerically obtained. The optimal policy was compared with different sequential optimization approaches. The results showed that it is close to optimal to process orders based on FCFS when customers are similar in their sensitivity to price and leadtime and the tardiness penalty is proportional to the tardiness. On the other hand, when the tardiness penalty is a fixed cost and the customers are dissimilar in their sensitivity to price and leadtime, considerable inefficiencies result from not harmonizing the price/leadtime quotation decisions with dispatching. We showed that, by considering a joint decision-making approach, the firm can make better (p, L) quotes, i.e. quotes with a higher acceptance rate, to both customer types and attain higher service levels. The results also highlighted the value of dispatching orders, even if

the flexibility to dispatch orders is not initially taken into account when making (p, L) quotes.

5. Efficiency of Paced and Unpaced Assembly Lines under Consideration of Worker Variability – A Simulation Study

Incorporating recent findings from behavioral operations, we compare paced and unpaced assembly lines with respect to their steady-state efficiency via simulation. In particular, workers can speed-up their service times when needed to feed downstream workers or to unblock upstream workers. It is found that unpaced lines are superior to paced lines for many real-world settings, i.e. in mixed-model production environments with a long line length. However, the benefit they provide has been overestimated in previous studies because of simplifying assumptions such as the disregarding of state-dependent behavior or worker fatigue. With an inhomogeneous workforce, the efficiency is also sensitive to worker placement. In unpaced conditions, an inexperienced worker should be placed in the middle of the line, while in paced conditions, he should be placed to the first workstation. Workers capable of speed-up should be placed in the middle of the line in both line types.

5.1. Introduction

This chapter compares two types of assembly lines that are widely found in customization industries such as the automobile industry, a paced line with an automated transportation system and an unpaced line with a manual transportation system. It is motivated by the question: what type is superior in which production environment and how do human characteristics influence this comparison? The proposed simulation model extends mathematical approaches for analyzing the impact of human behavior on assembly lines. It also enables the generation of insights into the management of an inhomogeneous workforce.

The chapter is organized as follows. In Section 5.2, the simulation models are introduced, while Section 5.3 presents the results of the simulation study. Section 5.4 concludes and lists directions for further research.

5.2. Simulation models

We build two simulation models for imitating a paced and an unpaced line, using AnyLogic 7, a Java-based simulation software. We first present the general assumptions that apply to both models in Section 5.2.1 and provide further details about

the models of each line type in Sections 5.2.2 and 5.2.3.

5.2.1. General assumptions

We consider an assembly line that consists of n sequential workstations connected by an automatically or manually operated transportation system along the material flow. The workstations are operated by human workers, i.e. there are no automated machine stations. For the unpaced line, inter-station buffers are installed in order to hold a finite number of intermediate items. For the paced line, workers can float into the downstream workstation that shows effects similar to those of buffers. Both lines are saturated (Dallery and Gershwin, 1992), i.e. the first station is never starved and the last station is never blocked.

Different products are produced in facultative sequence on one assembly line. For each product k, the task times of items follow a truncated normal distribution with mean μ_{task}^k and standard deviation σ_{task}^k to simulate the fact that each product has a minimum (min_{task}^k) and a maximum task time (max_{task}^k) but depending on the configuration, i.e. the options chosen by the customer, it is not the same for any two items. AnyLogic performs truncation by discarding every realization outside the interval $[min_{task}^k, max_{task}^k]$ and generating another random value.

The tasks and workers are introduced as two separate sources of variability. The variability due to workers is modeled as varying work times to complete items with the same task time. Therefore, the realized task time for an item, which is denoted $tasktime$, is taken as the mean of the time a worker needs to complete its processing at his station, i.e. $\mu_{work} = tasktime$. For workers with a theoretical coefficient of variation (CV_{work}) of zero, the work time would be identical to the task time.

Empirical studies such as Dudley (1963), Slack (1990), and Sury (1964) find that unpaced work shows a positively skewed, and paced work a normally shaped work time distribution. Thus, the Normal distribution with mean μ_{work}, standard deviation σ_{work} is used for the paced line to model work times, and the Johnson distribution for the unpaced. This distribution family developed by Johnson (1949) possesses four parameters with which it is possible to set the first four moments mean μ_{work}, standard deviation σ_{work}, skewness β_1, and kurtosis β_2.

Based on Powell and Schultz (2004), a state-dependent behavior is introduced. We assume that workers make downwards adjustments to their mean work time with a factor f that determines the maximum level of adjustment. In other words, workers

can go down to a mean work time of $\mu_{work} = tasktime \cdot (1 - f)$ if needed. Further, unlike Powell and Schultz (2004) and Heimbach et al. (2012), fatigue of workers is integrated into the model. Fatigue describes the diminishing ability of the worker to increase his work pace, i.e. speed-up by f. This understanding of fatigue is different from the one introduced, for example, in the work of Jaber and Neumann (2010), Dode et al. (2016) and Carnahan et al. (2001) who investigate the effects of physiological exhaustion during regular execution of work. We assume a limited number of speed-up cycles (worker achieves more than 100% of his average), while Jaber and Neumann (2010) study the decreasing overall energy level of workers until a break is needed. Inspired by learning curve models with learning and forgetting, the following exponential relationship between speed-up and fatigue is introduced.

$$f(c) = f(0)e^{-cr}, \tag{5.1}$$

where $f(c)$ represents the speed-up factor after c consecutive speed-up cycles, $f(0)$ represents the maximum speed-up of a worker, and r represents the scaling parameter of exhaustion. In the model, c is increased by one if the worker shows a speed-up in the current cycle and otherwise, i.e. if he recovers in that cycle and gains speed-up potential for upcoming cycles, decreased by one.

Schultz et al. (1998) empirically identified 20% as a plausible value for the speed-up parameter $f(0)$, but make no analysis of functional relationships such as exponentially decreasing speed-ups due to fatigue. Setting $f(0)$ to zero resembles state-independent behavior, i.e. the traditional assumption of independent and identically distributed service times. In our simulation, the scale parameter r is assumed to be 0.3 and the maximum number of subsequent speed-up cycles c is assumed to be 3, i.e. a worker cannot decrease his mean processing time after the third subsequent speed-up cycle. The resulting speed-up potential is shown in Table 5.1.

c	0	1	2	≥ 3
$f(c)$	20%	15%	11%	0%

Table 5.1 Overview worker speed-up potential

5.2.2. Model of an unpaced line

In an unpaced line, inter-station buffer spaces allow work-in-process (WIP) to accumulate between process steps and counteract variability. In our model, the maximum capacity of all buffers are set equal (B_{max}). Furthermore, the time it takes to transfer items between workstations and buffers is non-negligible. This is in line with Commault and Semery (1990) who showed that transfer times can have a

significant impact on line performance. Although they assume an automated transfer line, their arguments also apply to unpaced conditions. The consideration of transfer times is important when unpaced and paced lines are compared, because transfer times are (indirectly) considered in the latter case. The transfer times (tr) are deterministic and equal throughout the line.

When a worker is finished with processing an item, he transfers it to the downstream buffer (assuming not being blocked), takes the new item from the upstream buffer (assuming not being starved) and at this moment, before transferring, observes the content of the up- and downstream buffers. The state-dependent behavior is modeled as follows. The workers react to the observed state of buffer sizes by lowering their mean processing time accordingly. The total speed-up f is equally attributed to the up- and downstream buffers, respectively, assuming an additive nature. Using Heimbach et al. (2012), this gives the following relationship between the adjusted mean processing time μ_{work} of a worker in a particular cycle and the current level of the adjacent buffers:

$$\mu_{work} = tasktime \cdot \left(1 - \left(\frac{f}{2} \cdot \frac{s_{i-1,i}}{B_{max}} + \frac{f}{2} \cdot \frac{B_{max} - s_{i,i+1}}{B_{max}}\right)\right) \tag{5.2}$$

where $s_{i-1,i}$ denotes the current buffer level of the buffer between workstations $i-1$ and i.

Note that, in the unpaced model, the total time each item spends in a workstation is composed of the time for transferring the item from the upstream buffer, the realized work time ($worktime$) and the time for transferring it to the downstream buffer. Thus, the cycle time $\mathrm{CT} = worktime + 2 \cdot tr$.

In addition, it is important to consider that an unbalancing of the workload allocation provides performance improvement potential for an unpaced line with stochastic processing times (see e.g. Hudson et al. (2015)). Therefore, we consider an unbalanced allocation scenario in addition to a balanced allocation scenario. The model described above is used for determining the optimal workload allocation for a 3-station line via simulation optimization, assuming a single model setting with normally distributed work times, experienced workers and no speed-up. The optimal workload to be allocated to workstation i is denoted l_i^*. The total amount of work to be completed by the assembly line is normalized to equal the total number of stations. Thus, in the balanced allocation scenario $l_i = 1$ for $i = 1, 2, 3$. Table 5.2 shows the optimal allocations for different buffer size and CV_{work} combinations. In the numerical study, considering the unbalanced scenario means using

the corresponding allocation from the table below.

Buffer Size	CV_{work}	l_1^*	l_2^*	l_3^*
1	0.25	1.011	0.977	1.011
1	0.5	1.027	0.947	1.027
2	0.25	1.006	0.988	1.006
2	0.5	1.016	0.969	1.016

Table 5.2 Cases of line imbalance in unpaced conditions

Since $l_i^* > l_j^*$ means that more of the total work is completed by workstation i than by workstation j, the workload allocation has a direct link to the expected task times. For product k, the mean task time of items is equal to $l_i^* \cdot \mu_{task}^k$ for workstation i, $i = 1, 2, 3$.

5.2.3. Model of a paced line

In a paced line, the items are attached to a conveyor, which moves at a certain speed that allows each worker to work on an item for the duration of a predetermined takt time (TT). Absorption of variability is achieved by allowing the station length to be $(100 + x)\%$ of the TT. By that, the worker is given the possibility to work up to the station length in case he needs more time than TT for completing an item (overtime). If he cannot complete the item within the station length, the line has to be stopped, which will cause all other workers to become idle. This set up has two implications: first, the time available to the worker in the subsequent cycle is reduced by the time he additionally needed in the previous cycle, and second, the time available to the downstream worker is also reduced by the same amount. This flexibility enables a smoothing of variability in task times.

Although such a phenomenon has not yet been reported in the literature with state-dependent behavior, it is reasonable to assume that the speed-up effect is also present in paced conditions. Analogously to the speed-up effect in unpaced conditions, a worker under paced conditions is believed to lower his mean service time if a) he needed more time than the takt time (TT) in the previous cycle, or b) his upstream worker needed more time than the TT. In both cases, the time left to finish the subsequent task is shorter than the TT, which puts the same pressure on him as if the upstream buffer is full and the downstream buffer is empty i.e. $\mu_{work} = tasktime \cdot (1 - f)$.

5.2.4. Experimental design

To analyze different scenarios of product variety, the simulation is run for a single model setting ($k = 1$ with probability 1) and a mixed-model setting with 5 different

products ($k = 1, ..., 5$ with equal probabilities). In the single model scenario $\mu^1_{task} =$ 1 minute and in the mixed-model scenario $\mu^1_{task} = 0.8$ minutes, $\mu^2_{task} = 0.9$ minutes, $\mu^3_{task} = 1$ minute, $\mu^4_{task} = 1.1$ minutes, and $\mu^5_{task} = 1.2$ minutes. The task time coefficient of variation (CV_{task}) is 0.2 for all products. The minimum task time (low specified variant of the product) is assumed to be 50% and the maximum task time (high end specification) is assumed to be 250% of the mean task time.

The low coefficient of variation values for paced and unpaced work times (CV_{work}) are determined as the average of the empirical values found in Franks and Sury (1966), Doerr and Arreola-Risa (2000), Knott and Sury (1987), Murrell (1961), Slack (1990) and Sury (1964), a high coefficient of variation (CV) is assumed to be double that value. The skewness and kurtosis values for the Johnson distribution are based on Knott and Sury (1987). They find an average Pearson's 2^{nd} Coefficient of Skewness of 1.5 in their analysis of 26 unpaced work tasks. In state-dependent behavior, workers show a speed-up of 20%. Unskilled workers show a mean service time approximately 20% above that of skilled workers while the standard deviation remains approximately constant (Franks and Sury, 1966).

For the unpaced line, the buffer size is set to be either one or two. This acknowledges the efforts companies make to introduce lean production methods and reduce buffer capacity for intermediate goods. The transfer time (tr) is chosen in such a way that the expected cycle time (CT) is equal to the takt time (TT) of the paced line. A TT of 1.2 minutes is assumed and an expected work time of 1 minute results in a transfer time of 0.1 minutes. The station length of the paced line is set at 120% of the TT, which means that a worker can use 20% of the downstream workstation to finish the task at hand.

For the paced line, we consider different scenarios regarding the case of the last workstation: the flexibility to float into an extended area is (i) also available (ii) not available to the last workstation, since in practice there are examples of lines designed with and without this flexibility. Whenever it is not varied, the former scenario applies.

Furthermore, when comparing paced and unpaced lines in section 5.3.2, a varying line length i.e. three and 8-station lines are considered as examples for small and more realistic lines. The parameter values used in this chapter are summarized in Table 5.3.

The line efficiency E, the expected value of inter-completion times ($E[IC]$) and the

Parameter	Instance		Value	Unit
Task time distribution	Paced, Unpaced		truncated Normal $N(\mu_{task}, \sigma_{task})$	-
Coefficient of Variation (CV_{task})			0.2	-
Model mix	Single	μ^1_{task}	1	[min]
	Mixed	μ^k_{task} $k = 1, ..., 5$	0.8, 0.9, 1, 1.1, 1.2	[min]
		product share $k = 1, ..., 5$	20, 20, 20, 20, 20	[%]
Work time distribution	Paced line		Normal $N(\mu_{work}, \sigma_{work})$	-
	Unpaced line		Johnson $J(\mu_{work}, \sigma_{work}, \sqrt{\beta_1}, \beta_2)$	
Coefficient of Variation (CV_{work})	Low		0.20 (paced), 0.25 (unpaced)	-
	High		0.40 (paced), 0.50 (unpaced)	
Pearson's 2^{nd} Coefficient of Skewness	0.5		$\sqrt{\beta_1} = 3$, $\beta_2 = 36$	-
	1.5		$\sqrt{\beta_1} = 2$, $\beta_2 = 6$	
Speed-up $f(0)$	State-independent		0	[%]
	State-dependent		20	
Buffer size B_{max}	Low		1	[items]
	High		2	
Line length	Short		3	[workstations]
	Long		8	

Table 5.3 Experimental design parameters

standard deviation of inter-completion times (σ_{IC}) are considered key performance indicators of the lines. E is defined as the expected throughput of the line in relation to the throughput in the deterministic case. Note that, due to considering positive transfer times in the unpaced system, both types of lines produce 1 item every 1.2 minutes in the deterministic case. Thus, E is determined by the expected number of completed items within 1.2 minutes. The inter-completion time is defined as the time between two items completed by the last workstation. A line operating with an efficiency of 1 has $E[IC] = 1.2$.

The warm-up period is the time until the first $2,000$ items are produced and the simulation is stopped after the next $10,000$ items are produced. For all scenarios, 50 replications were run.

5.3. Results

5.3.1. Impact of human behavior

The results for the paced assembly line are reported in Table 5.4, the results for the unpaced assembly line in Table 5.5. We vary the following system parameters when analyzing the effects of state-dependent behavior and fatigue: coefficient of variation of the work time distribution and model mix. Column $f = 0$ gives the

line efficiency E for state-independent behavior while column $f = 0.2$ gives the line efficiency E for state-dependent behavior. Furthermore, since the parameters such as; available buffer spaces, skewness/kurtosis of the work time distribution and workload allocation are additional parameters of the unpaced line that affect its performance, they are also varied in Table 5.5.

CV_{work}	Model Mix	$f = 0$	$f = 0.2$	
			with fatigue	without fatigue
0.2	sm	0.954	0.977	0.979
	mm	0.928	0.960	0.966
0.4	sm	0.863	0.905	0.919
	mm	0.846	0.888	0.905

* sm: single model, mm: mixed-model

Half width of 95% confidence interval ≤ 0.001 in all instances.

Table 5.4 Impact of human behavior for a 3-station paced line

CV_{work}	Buffer Size	Line Balance	Model Mix	$\sqrt{\beta_1} = 0$			$\sqrt{\beta_1}/\beta_2 = 3/36$			$\sqrt{\beta_1}/\beta_2 = 2/6$		
					$f = 0.2$			$f = 0.2$			$f = 0.2$	
				$f = 0$	with fatigue	without fatigue	$f = 0$	with fatigue	without fatigue	$f = 0$	with fatigue	without fatigue
0.25	1	unb	sm	0.933	0.955	0.963	0.935	0.955	0.963	0.932	0.954	0.961
			mm	0.927	0.949	0.957	0.929	0.951	0.957	0.927	0.949	0.955
		bal	sm	0.932	0.954	0.963	0.934	0.954	0.963	0.931	0.953	0.961
			mm	0.927	0.949	0.957	0.928	0.951	0.957	0.926	0.948	0.956
	2	unb	sm	0.958	0.975	1.005	0.958	0.974	1.004	0.957	0.974	1.004
			mm	0.956	0.972	1.002	0.956	0.973	1.001	0.955	0.971	1.001
		bal	sm	0.958	0.975	1.005	0.959	0.975	1.005	0.957	0.974	1.005
			mm	0.955	0.972	1.003	0.956	0.974	1.002	0.954	0.972	1.001
0.5	1	unb	sm	0.859	0.885	0.897	0.867	0.890	0.900	0.858	0.881	0.893
			mm	0.853	0.878	0.890	0.861	0.884	0.895	0.853	0.877	0.887
		bal	sm	0.858	0.883	0.897	0.866	0.889	0.902	0.857	0.882	0.893
			mm	0.852	0.877	0.891	0.860	0.884	0.896	0.853	0.877	0.887
	2	unb	sm	0.904	0.922	0.953	0.907	0.925	0.955	0.903	0.922	0.953
			mm	0.900	0.918	0.949	0.902	0.922	0.952	0.899	0.918	0.948
		bal	sm	0.903	0.922	0.955	0.907	0.926	0.957	0.902	0.921	0.954
			mm	0.900	0.919	0.951	0.904	0.923	0.954	0.899	0.918	0.949

* bal: balanced, unb: unbalanced, sm: single model, mm: mixed-model

Half width of 95% confidence interval ≤ 0.001 in all instances.

Table 5.5 Impact of human behavior for a 3-station unpaced line

5.3.1.1. Impact of state-dependent behavior

The results support the previous findings that the speed-up effect has a positive influence on line performance. For all lines in every setting, E is higher for state-dependent behavior than for state-independent behavior. Moreover, note that in

paced conditions, the state-dependent behavior is more valuable in environments with higher task and human variability. Similarly in unpaced conditions, the state-dependent behavior is more valuable in environments with low buffers and a high work time variability. Thus, speed-ups counteract variability increases and buffer reductions but cannot compensate for them.

While previous studies (Powell and Schultz (2004); Heimbach et al. (2012)) report that the output of unpaced lines with a speed-up parameter of 0.2 is approximately 10% higher than that of lines with state-independent behavior, we only find a difference of around $3-5\%$. One reason for this difference is that existing studies disregard worker fatigue. The consideration of worker speed-up in behavioral studies so far resembles an infinitely possible, rather than a temporarily possible ability. When fatigue is ignored, our results show that the difference can reach up to $6-7\%$. Therefore, consideration of fatigue alone does not explain this discrepancy. As opposed to this work, the aforementioned authors assume that the workers observe the content of their adjacent buffers at all times and make adjustments to their pace as soon as the state of buffer contents changes. Furthermore, exponential service time distributions are considered. Even Powell and Schultz (2004) raise the question whether a coefficient of variation (CV) of 1 resembles realistic conditions. They also test their model for normally distributed service times with a CV of 0.35. In this case, they find a line efficiency, which they define as the total throughput in relation to that of the state-independent line with an unlimited buffer, of around 94% for state-independent and of around 105% for state-dependent behavior.

5.3.1.2. Impact of fatigue

The consideration of fatigue has a minor influence on paced lines with low work time variability (less than 1%) and on unpaced lines with small buffers (around 1%). The reason for this observation in the paced case is that, whenever a worker can fit into the takt time for finishing an item, he has no pressure and therefore does not speed-up for the next item provided that the upstream worker does not float into his area. In the unpaced case, this happens due to the way workers observe the content of adjacent buffers. They first put the completed item to the downstream buffer, take the new item from the upstream buffer and then observe their sizes. When the maximum buffer size is 1, workers typically see no reason for speeding up their pace because the upstream buffer is empty and the downstream buffer is full. Thus, in both cases, workers naturally have a higher chance of recovering from accumulated fatigue.

This effect increases with task and work time variability in a paced line and with

buffer size and work time variability in an unpaced line. The difference between E with fatigue and E without fatigue appears to be around 2% for paced lines with a higher variability ($CV_{work} = 0.4$) and around 3% for unpaced lines with larger buffers. This means that, when line performance is estimated in such environments, ignorance of fatigue would lead to larger errors.

5.3.1.3. Statistical analysis of influencing factors

We conduct an analysis of variance (ANOVA) using the software R for measuring the statistical significance of investigated effects and the influence of system parameters. The results are presented in Table 5.6.

	Df	Sum Sq	Mean Sq	F value	$Pr(>F)$	
Paced Line						
The effect of state-dependent behavior (assuming fatigue)						
Speed-up	1	0.1198	0.1198	5910	$< 2e-16$	***
Model Mix	1	0.0361	0.0361	1780	$< 2e-16$	***
CV_{work}	1	0.6246	0.6246	30823	$< 2e-16$	***
Residuals	396	0.008	0			
The effect of fatigue ($f = 0.2$)						
Fatigue	1	0.0094	0.0094	976.5	$< 2e-16$	***
Model Mix	1	0.0232	0.0232	2400.3	$< 2e-16$	***
CV_{work}	1	0.4362	0.4362	45223.1	$< 2e-16$	***
Residuals	396	0.0038	0			
Unpaced Line						
The effect of state-dependent behavior (assuming fatigue)						
Speed-up	1	0.498	0.498	15650	$< 2e-16$	***
Skewness and Kurtosis	2	0.013	0.007	212.1	$< 2e-16$	***
Workload Allocation	1	0	0	3.729	0.0535	.
Model Mix	1	0.02	0.02	643.5	$< 2e-16$	***
Buffer Size	1	1.302	1.302	40950	$< 2e-16$	***
CV_{work}	1	4.628	4.628	145500	$< 2e-16$	***
Residuals	4792	0.152	0			
The effect of fatigue ($f = 0.2$)						
Fatigue	1	0.482	0.482	8942.729	$< 2e-16$	***
Skewness and Kurtosis	2	0.01	0.005	94.177	$< 2e-16$	***
Workload Allocation	1	0	0	3.499	0.0615	.
Model Mix	1	0.022	0.022	408.396	$< 2e-16$	***
Buffer Size	1	1.99	1.99	36941.71	$< 2e-16$	***
CV_{work}	1	4.245	4.245	78793.34	$< 2e-16$	***
Residuals	4792	0.258	0			

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 5.6 Summary of ANOVA results

The results support the statistical significance of the effects of state-dependent behavior and fatigue on the performance of both line types. In addition, they show that all system parameters, except the workload allocation in an unpaced setting, have statistically significant influences. Although skewness and kurtosis of the work time distribution and the task time variability (i.e. model mix) have an

influence, it is rather small. The differences in E are up-to 1% when different work time distributions are considered. Overall, efficiency is close to the scenario with a normal work time distribution. A single model environment yields less than 1% higher efficiency compared to a mixed-model production.

5.3.1.4. Managing inhomogeneous workforce

So far, we assumed a homogeneous workforce in this article, which is a condition questioned by Schultz et al. (2010) and also practical experience. To test the impact of an inhomogeneous workforce, scenarios where only one worker is capable of speed-up are tested. This worker was subsequently positioned at the first, second, and third station. Both scenarios regarding the flexibility given to the last worker are tested for the paced setting. The results are shown in Table 5.7. In the first column, the lines WS1, WS2 and WS3 indicate at which workstation the worker capable of speed-up is positioned.

The numbers indicate that, if state-dependent behavior is an individual capability, the worker capable of speed-up should be positioned at the middle of the line in an unpaced setting. In a balanced line, this creates a situation similar to a bowl shape because the middle station can work faster than the other stations. In an already unbalanced line, the effect is reinforced by such a strategy. In a paced setting, the best position for the worker capable of speed-up depends on whether the flexibility to float into an extended area is also given to the last worker or not. If flexibility is not given, the best position for this worker is clearly the last workstation. The lack of flexibility in line design for mitigating stoppages is compensated by the worker's state-dependent behavior. If flexibility is given, the intermediate worker is the one that should be capable of speed-up so that he compensates for the first worker's overtimes and also floats into the area of the last workstation less frequently.

Furthermore, the literature findings suggest that experience has a significant influence on line performance. To test this, analogously to the placing of state-dependent behavior, one inexperienced worker was placed subsequently at the first, second, and third station with two experienced workers at the other two stations. This placement is done under two conditions: (1) when no worker is capable of speed-up ($f = 0$) and (2) when all workers are capable of speed-up ($f = 0.2$). The results are reported in the second column of Table 5.7.

The guidelines for the best position of the inexperienced worker depend on whether one takes the state-dependent behavior into account or not. The results for the paced line with flexibility suggest placing the inexperienced worker at the last work-

				Positioning of speed-up		Positioning of inexperience			
						f = 0		*f = 0.2*	
				E	Δ to best	E	Δ to best	E	Δ to best
Paced (No Flexibility)			WS1	0.929	-0.010	0.874	-	0.934	-
		sm	WS2	0.937	-0.002	0.861	-0.013	0.933	-0.001
			WS3	0.939	-	0.860	-0.014	0.901	-0.033
			WS1	0.899	-0.017	0.843	-	0.912	-
		mm	WS2	0.911	-0.004	0.834	-0.009	0.910	-0.002
			WS3	0.916	-	0.830	-0.013	0.884	-0.027
Paced (With Flexibility)			WS1	0.960	-0.004	0.907	-0.002	0.958	-
		sm	WS2	0.964	-	0.905	-0.003	0.956	-0.002
			WS3	0.963	-0.001	0.908	-	0.957	-0.001
			WS1	0.935	-0.007	0.881	-0.002	0.938	-
		mm	WS2	0.943	-	0.880	-0.003	0.936	-0.003
			WS3	0.942	-0.001	0.883	-	0.937	-0.001
Unpaced	buffer size=1 unbalanced		WS1	0.940	-0.003	0.841	-0.016	0.863	-0.023
		sm	WS2	0.942	-	0.857	-	0.886	-
			WS3	0.940	-0.002	0.841	-0.016	0.864	-0.022
			WS1	0.935	-0.003	0.840	-0.015	0.862	-0.020
		mm	WS2	0.938	-	0.854	-	0.882	-
			WS3	0.935	-0.003	0.840	-0.014	0.861	-0.020
	buffer size=1 balanced		WS1	0.938	-0.006	0.847	0.000	0.869	-0.005
		sm	WS2	0.944	-	0.844	-0.004	0.874	-
			WS3	0.938	-0.006	0.848	-	0.869	-0.005
			WS1	0.933	-0.007	0.846	0.000	0.867	-0.004
		mm	WS2	0.940	-	0.841	-0.004	0.871	-
			WS3	0.933	-0.006	0.846	-	0.867	-0.004
	buffer size=2 unbalanced		WS1	0.963	-0.001	0.849	-0.012	0.865	-0.023
		sm	WS2	0.964	-	0.861	-	0.888	-
			WS3	0.963	-0.001	0.849	-0.012	0.865	-0.023
			WS1	0.960	-0.002	0.849	-0.011	0.865	-0.022
		mm	WS2	0.962	-	0.860	-	0.887	-
			WS3	0.960	-0.002	0.849	-0.011	0.865	-0.022
	buffer size=2 balanced		WS1	0.962	-0.004	0.854	-	0.869	-0.012
		sm	WS2	0.966	-	0.853	-0.001	0.881	-
			WS3	0.961	-0.005	0.853	-0.001	0.869	-0.012
			WS1	0.960	-0.004	0.853	0.000	0.869	-0.011
		mm	WS2	0.964	-	0.852	-0.001	0.880	-
			WS3	0.959	-0.004	0.853	-	0.868	-0.011

* Positioning of speed-up: only experienced workers,
Positioning of inexperience: all workers capable of speed-up,
CV_{work}(paced) = 0.2, CV_{work}(unpaced) = 0.25, $\sqrt{\beta_1}/\beta_2 = 3/36$.
sm: single model, mm: mixed-model
Half width of 95% confidence interval ≤ 0.001 in all instances.

Table 5.7 Positioning of speed-up and inexperience

station when $f = 0$, because if an overtime is needed, the worker floats into an extended area without causing any other worker to lose time. The guidelines are reversed when $f = 0.2$. The inexperienced worker is placed at the beginning of the line. Because, whenever the first worker floats into the area of the next workstation, a speed-up behavior is triggered, possibly not only for the second but also for the last workstation. This creates the highest chances that longer service times can be compensated for by the end of the line. On the other hand in case of a paced line without flexibility, the inexperienced worker is placed at the beginning of the line also when $f = 0$. Because, solely due to the random nature of task and work times, there is a chance that the downstream workers compensate for overtimes of the first worker.

In unpaced lines, the inexperienced worker should be placed at the middle station independent of whether or not the workload allocation is balanced when $f = 0.2$. This counter-intuitive result is explained by state-dependent behavior. The disadvantage of a higher mean service time of unskilled workers is compensated by the speed-up they show based on the up- and downstream buffers. The results differ when a state-independent behavior is assumed. In line with the bowl phenomenon, which suggests decoupling station interference through lower station loads in the center of the line, the inexperienced worker should *not* be placed at the middle station when the workload allocation is balanced, because this results in a reversed-bowl shaped workload allocation. The impact of a relatively higher mean service time is (partly) offset when the workload allocation is already bowl-shaped. Therefore, in the unbalanced case, the best position of the inexperienced worker is the middle station.

5.3.2. Comparison of paced and unpaced assembly lines

We compare the paced and unpaced lines based on the efficiency E, the expected value of inter-completion times ($E[IC]$) and the standard deviation of inter-completion times (σ_{IC}) under varying line length, model mix and available buffer spaces in the unpaced line. For the paced line, we consider the scenario where the flexibility to float into an additional area is also given to the last workstation. Tables 5.8 and 5.9 give the results. Column Δ in Table 5.8 shows the relative difference of the efficiency of the unpaced line to the one of the paced line ($(\frac{E_{unpaced}}{E_{paced}} - 1) \cdot 100\%$). Since the efficiency of the paced and unpaced lines with three stations are already reported in Tables 5.4 and 5.5, we directly give the Δ for this case.

First, we consider the line efficiency E given in Table 5.8 and the expected value of

Buffer Size	Model Mix	f		**3 stations**	**8 stations**		
				Δ in %	**Paced**	**Unpaced**	Δ in %
1	sm	0		-2.09	0.893	0.903	1.05
		0.2	with fatigue	-2.29	0.944	0.931	-1.37
			without fatigue	-1.62	0.951	0.946	-0.50
	mm	0		-0.07	0.844	0.895	6.04
		0.2	with fatigue	-0.95	0.909	0.923	1.53
			without fatigue	-0.93	0.924	0.937	1.36
2	sm	0		0.49	0.893	0.940	5.21
		0.2	with fatigue	-0.14	0.944	0.961	1.82
			without fatigue	2.64	0.951	1.004	5.64
	mm	0		2.95	0.844	0.935	10.87
		0.2	with fatigue	1.44	0.909	0.957	5.28
			without fatigue	3.76	0.924	1.000	8.21

*sm: single model, mm: mixed-model
CV_{work}(paced) =0.2, CV_{work}(unpaced) = 0.25
$\sqrt{\beta_1}/\beta_2 = 3/36$, balanced workload allocation in the unpaced line.
Half width of 95% confidence interval ≤ 0.001 in all instances.

Table 5.8 Comparison of paced and unpaced assembly lines based on efficiency

inter-completion times $E[IC]$ given in Table 5.9 and take the environment in which workers exhibit state-dependent behavior ($f = 0.2$) with fatigue as a basis when comparing two types of lines.

The results show that a paced setting is superior to an unpaced setting when production takes place in a single model environment with a short line, regardless of whether the maximum allowed buffer size in the unpaced system is 1 or 2. The performance improvement potential one gains by producing items in an unpaced instead of a paced line increases with line length, volatility of task times, i.e. when moving from a single model to a multi-model environment, and with available buffer spaces in the unpaced line. When a long line and a mixed-model environment is considered, an unpaced setting is superior to a paced setting, regardless of the maximum allowed buffer size in the unpaced system. In all other cases, i.e. when producing in a single model environment with a long line or in a mixed-model environment with a short line, the type of the line that should be preferred depends on the available inter-station buffer space in the unpaced line. If it is low, a paced line should be preferred, otherwise the unpaced system performs better. Figure 5.1 provides a visual summary of these findings.

Buffer Size	Model Mix	f		3 stations				8 stations			
				Paced		Unpaced		Paced		Unpaced	
				$E[IC]$	σ_{IC}	$E[IC]$	σ_{IC}	$E[IC]$	σ_{IC}	$E[IC]$	σ_{IC}
1	sm	0		1.258	0.154	1.285	0.395	1.344	0.200	1.329	0.441
		0.2	with fatigue	1.229	0.118	1.257	0.381	1.271	0.152	1.288	0.413
			without fatigue	1.226	0.113	1.246	0.373	1.262	0.145	1.269	0.394
	mm	0		1.293	0.200	1.293	0.453	1.423	0.257	1.341	0.515
		0.2	with fatigue	1.250	0.152	1.262	0.432	1.320	0.200	1.300	0.480
			without fatigue	1.242	0.140	1.254	0.426	1.298	0.183	1.281	0.464
2	sm	0		1.258	0.154	1.252	0.367	1.344	0.200	1.277	0.394
		0.2	with fatigue	1.229	0.118	1.230	0.356	1.271	0.152	1.248	0.374
			without fatigue	1.226	0.113	1.194	0.337	1.262	0.145	1.195	0.339
	mm	0		1.293	0.200	1.255	0.415	1.423	0.257	1.283	0.449
		0.2	with fatigue	1.250	0.152	1.232	0.398	1.320	0.200	1.254	0.428
			without fatigue	1.242	0.140	1.197	0.377	1.298	0.183	1.200	0.385

*sm: single model, mm: mixed-model
CV_{work}(paced) = 0.2, CV_{work}(unpaced) = 0.25, $\sqrt{\beta_1}/\beta_2 = 3/36$.
Balanced workload allocation in the unpaced line.
Half width of 95% confidence interval ≤ 0.003 for σ_{IC} in the unpaced case.
It is ≤ 0.001 for all other instances.

Table 5.9 Comparison of paced and unpaced assembly lines based on the expected value and standard deviation of inter-completion times

Model Mix	Buffer Size	Short	Long
Single model	1	PACED	PACED
	2		UNPACED
Mixed model	1	PACED	UNPACED
	2	UNPACED	

Figure 5.1 Conditions under which a paced or an unpaced design is more efficient assuming state-dependent behavior with fatigue

Additionally note that the consideration of state-dependent behavior with fatigue has a prominent impact on the conclusions regarding this comparison. As the results indicate, the performance of an unpaced line would be overestimated if the existence of an adaptive human behavior is ignored, especially for long lines. The overestimation of the unpaced performance would also typically be the case if fatigue is ignored, i.e., a worker is assumed to be able to speed-up whenever the state of the buffers in an unpaced line requires it, regardless of fatigue accumulation due to speeding up in many consecutive cycles. There is one exceptional case where fatigue ignorance leads to an overestimation of the performance of a paced line (8 stations, buffer size=1, mixed-model). This observation, in relation to the effects discussed in Section 5.3.1.2, can be attributed to making a larger error in estimating line performance due to disregarding fatigue in the paced case than in the unpaced case. Recall that this error is larger for paced lines in a mixed-model environment and smaller for unpaced lines when the maximum inter-station buffer size is one.

Note that, as opposed to Powell and Schultz (2004), our results do not indicate an increase in the efficiency of lines when their length increases. In other words, state-dependent behavior as modeled in this work is not strong enough to fully compansate for the inefficiencies resulting from a longer line length.

Subsequently, we compare paced and unpaced lines based on the output stability, i.e. based on the standard deviation of inter-completion times (σ_{IC}). The results in Table 5.9 show that the standard deviation of inter-completion times is lower when the line design is paced rather than unpaced in all cases. In other words, a paced line leads to a more stable output process than an unpaced line. Furthermore, this stability is intensified with adaptive human behavior. This is an important result from the practical point of view, because for firms serving in just-in-time production environments, i.e. for firms contractually obliged to provide a certain number of products within a certain time-window with no lateness, less variability in the throughput might be more important than having a slightly higher throughput.

5.4. Discussion and conclusion

In this simulation study, paced and unpaced assembly lines were compared under the consideration of adaptive human behavior. The existing research was extended in two directions.

First, fatigue was introduced as a complement to the speed-up effect to model a temporary decrease in mean service time as opposed to a permanently possible

one. Our results clearly show that the benefits of the speed-up effect can be heavily overestimated when fatigue is disregarded. While Heimbach et al. (2012) and Powell and Schultz (2004) find a 10% higher line efficiency in state-dependent behavior without fatigue for unpaced lines, our simulation only shows an increase of $3-5\%$ compared to state-independent behavior.

Second, paced and unpaced assembly lines are compared to identify guidelines that suggest which line configuration is best under which production circumstances. For this, state-dependent behavior has also been modeled for paced assembly lines. Assuming the possibility to float into the downstream workstation, we expect workers to speed-up when their remaining time to finish an item has been reduced by an upstream worker floating into their workstation. With a homogeneous workforce and state-dependent behavior, assuming a speed-up parameter of 0.2 as found by Schultz et al. (1998), an unpaced line is superior to a paced line in realistic conditions, but to a lesser degree than would be estimated by models ignoring either state-dependent behavior or fatigue. Additionally, we found that a paced configuration is useful for mitigating the volatility in the output process, which is further supported by the state-dependent behavior. With an inhomogeneous workforce, the different worker (in this work: in-experienced or capable of speed-up) should be placed in the middle of an unpaced line. In a paced line, the inexperienced worker should be placed to the first station of the line to minimize efficiency losses. The worker absorbing variability (in this work: capable of speed-up) should be placed at the end of the line if the flexibility to float into an extended area is not given to the last worker and in the middle of the line if the last worker has this flexibility.

However, there are limitations to a generalization of these findings. The results depend on the accuracy of the model and on how human behavior is modeled. More experiments under controlled conditions should be conducted to better understand the triggers and reactions of workers to the production environment. First, empirical evidence on fatigue behavior within state-dependent behavior seems to be necessary to improve the modeling of workers. Especially the exact signals that trigger speed-up and the limits of this behavior are still unclear. More empirical studies are necessary to identify and better describe the relationships between worker behavior and the work environment.

6. Designing Unpaced Production Lines with Human Operators - The Bowl Phenomenon Revisited

In this chapter, the design of unpaced production lines is studied by introducing a model that accounts for the state-dependent adjustments in processing rates of workers in combination with fatigue. We find that the workload allocation that minimizes the expected inter-completion time takes on the shape of a bowl, however the degree of unbalance is notably smaller than that would be predicted using traditional assumptions. Whenever workers exhibit a state-dependent behavior, the performance based on the service level and output process variability is optimized by allocating both available buffer spaces and the total workload in a decreasing pattern. Furthermore, when considering one of these measures as the objective, optimizing the buffer allocation can be more effective than optimizing the workload allocation.

6.1. Introduction

Motivated by the assembly line design problems faced by a first-tier automotive equipment supplier which largely uses human operators in the lines and adopts a just-in-time (JIT) production policy, we investigate the design of an unpaced production line considering the work behavior of human operators and using objective functions that are more appropriate for a firm that follows a JIT principle.

The production line design problem is the problem of optimally allocating system resources such as a total amount of workload and a pool of buffer spaces (Papadopoulos et al. (2009)). A pure workload allocation problem considers the amount of work assigned to workstations as a decision variable and takes the allocation of all other system resources (e.g. the pool of buffer spaces) as given. A pure buffer allocation problem considers the buffer space between workstations as a decision variable and takes the allocation of all other system resources (e.g. the total workload) as given. While these pure optimization problems are interesting for a production line designer whenever there are physical and/or technical constraints regarding reallocation of certain system resources, a simultaneous optimization of their allocation provides more flexibility and a higher performance improvement potential. Therefore, we investigate the following three problems: workload allocation problem, buffer allocation problem and simultaneous workload and buffer

allocation problem.

The studies in the production line design literature typically take the maximization of throughput per time unit as the objective (see Hudson et al. (2015) or Papadopoulos et al. (2009)). There also exists a large number of studies that consider the expected work-in-process inventory in the objective function (e.g. Hillier (2013)). On the other hand, for a firm following the JIT principle, the minimization of the variance of the output process as well as the maximization of the service level are more of an interest. However, the number of studies in the literature is not as large when it comes to analyzing the distribution of the output process or its higher moments than the first (see Lagershausen and Tan (2015)). This is due to the complexity involved in their analyses. In this chapter, we investigate various objectives such as; the minimization of the expected inter-completion time, the minimization of the variance of inter-completion times, the maximization of the service level (the probability that a given number of units are completed within a given time window) separately.

When seeking for the optimal allocations, the consideration of more realistic models of human work behavior is important since it may alter the traditional trade-offs. While the most literature ignores the difference between the behavior of human and machine operators when processing items (Bendoly et al. (2006)), we model a production line with human operators, whose processing times are not independent of the pace of their co-workers. This behavior is supported by existing findings e.g. Falk and Ichino (2006), Gould and Winter (2009), Mas and Moretti (2009) and Schultz et al. (2010). We particularly assume that the workers speed-up when they might otherwise cause idleness as suggested by Doerr et al. (1996) and Schultz et al. (1998). Furthermore, we consider that they get tired when showing extra effort to avoid idleness.

We model the production line as a continuous time Markov chain (CTMC). Although an exact model is utilized, the performance measures of interest -the expected value, variance of inter-completion times and the service level- can only be obtained numerically. As a result, the optimization problems that we consider lack a closed-form objective function. Thus, the optimal solution is obtained by utilizing a procedure that computes the performance measure from the CTMC under a given solution and seeks for the solution that optimizes its value.

The CTMC model is described in Section 6.2. The computation of the performance measures are explained in Section 6.3. Section 6.4 introduces the investigated op-

timization problems while Section 6.5 presents the numerical study. Section 6.6 summarizes findings.

6.2. Continuous time Markov model

6.2.1. Assumptions

We consider the classical assumptions about the production system, i.e. the model of Hunt (1956), as considered by Hillier and Boling (1966) and many others following it (see e.g. Hillier and Boling (1979), Hillier et al. (1993), Hillier and Hillier (2006) and Hillier (2013)), for ensuring comparability of our results.

- There are N single server workstations (W) in series with $N-1$ finite inter-station buffers (B) as shown in the following figure.

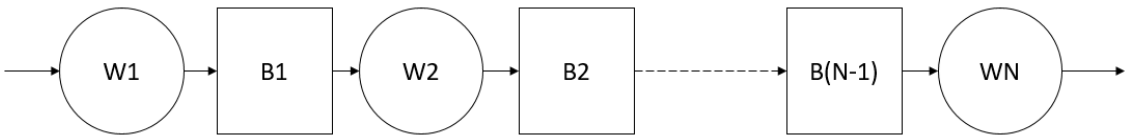

Figure 6.1 Production system

- Every item must be processed through each of the N workstations in the same sequence: $1, 2, ..., N$.
- The first workstation never starves and the last workstation is never blocked.
- The blocking mechanism is BAS (blocking-after-service). At the instant of completion of an item on a workstation while the downstream buffer is full, the item stays in the workstation until a space becomes available on the buffer. During this time, the worker can not process a new item (blocked). This implies that the workstation is also used as an additional buffer space.
- The capacity of buffer i is denoted B_i while the number of items it contains at a point in time is denoted b_i, $i = 1, ..., N-1$.

We specifically assume that the processes in workstations are conducted by human operators. Thus, the processing times do not purely stochastically evolve but also involve effects of human reactions to the state of the line. Similar to Powell and Schultz (2004), in which the performance of a serial line is analyzed under the state-dependent behavior, we make the following assumptions.

- There is an underlying probability distribution to the processing times of human

operators. This is the exponential distribution for all workers and w_i^{nom} denotes the nominal mean processing time of the worker in workstation i.

- The workers are motivated to avoid causing idleness for their co-workers. They continuously observe their immediate environment and adjust the mean of the processing time distribution. The immediate environment means in this context, the size of the adjacent buffers.
- A speed-up parameter f determines the maximum level of adjustment. This means that the minimum value the mean processing time of a worker can take is: $w_i^{nom} - w_i^{nom} \cdot f$.
- The workers react *symmetrically* to their up- and downstream buffers. In other words, they weight these buffers equally when adjusting their processing times.
- The workers *add* their processing time adjustments for up- and downstream buffers. In other words, the processing time adjustment a worker makes at a given system state is determined by the sum of the adjustments with regard to the size of the upstream and downstream buffers.

Additionally, we assume that the workers' motivation for increasing their work pace depends on their physical state. A worker might not be willing to (or able to) lower his mean processing time, if he already kept doing this for a number of consecutive items he previously processed. So, even though the state of the buffers gives him a feedback in the direction of a need for decreasing the mean processing time, an adjustment does not take place when the worker is not motivated to (or able to) exhibit such a behavior.

Modeling the state-dependent behavior independent of the number of consecutive times a worker needs to increase his pace is problematic when the optimal allocation of system resources is sought. As mentioned earlier, the results of Heimbach et al. (2012) suggest under some conditions that it is optimal to assign nearly all the workload of a three-station line to the worker in the middle. However, even if a pace increase is possible for all items a worker processes during the day, the suggested allocation would never be feasible in real-life. Such an allocation possibly leads to health damages for the worker. Otto and Scholl (2011) underline the importance of incorporating this aspect into the models when optimizing task assignment to workstations.

Worker fatigue is a concept that is related to a number of diverse disciplines. Ergonomics literature, e.g. Enoka and Duchateau (2008) defines fatigue as a decline in the maximal force or power capacity of a muscle. To the best of our knowledge,

Öner Közen et al. (2016) is the only study in the literature that considers worker fatigue in relation to the state-dependent behavior. They model the maximum level of processing time adjustment, i.e. the speed-up parameter, as a decreasing function of the number of consecutive times a worker needs to work with lowered processing time. Following a similar line of thought, we assume that:

- The nominal workload can be maintained with no fatigue (e.g. during a regular shift). Fatigue accumulates when workers lower their processing times, due to extracting extra effort to avoid idleness.
- The variable c_i indicates the accumulated fatigue for worker i. Its value increases by one when the worker completes an item using a lower mean processing time than nominal and decreases by one upon completion of an item using the nominal mean processing time. Furthermore, the accumulated fatigue for an idle worker also decreases by one, upon completion of an item by one of the co-workers.
- There is a maximum to the value of this variable ($C_{max,i}$, for worker i). This assumption is in line with the maximum endurance time (MET) measurement in the ergonomics literature which defines the maximum time during which individuals can exert the required force level (Rohmert (1973), El ahrache et al. (2006), Jaber and Neumann (2010)). If the state of buffer contents implies this, a worker i can decrease his mean processing time when $c_i = 0, 1, 2, ..., C_{max,i} - 1$, but not when $c_i = C_{max,i}$.
- The accumulated fatigue indicator c_i is linked to the speed-up parameter f assuming a linear relationship. The studies by Jaber and Neumann (2010) and Soo et al. (2009), for example, support a linear relationship between fatigue and time.

$$f(c_i) = f^0 \cdot \left(1 - \frac{c_i}{C_{max,i}}\right) \tag{6.1}$$

  where f^0 is the value of the speed-up parameter associated with the state where there is no fatigue accumulated. In other words, when the worker is completely recovered, he is able to speed-up with f^0. Higher values of f^0 result in a faster decline of $f(c_i)$ that imitates a sharper effect of fatigue when worker forces himself to put a higher effort (Figure 6.2).

6.2.2. State space

The system with assumptions listed in Section 6.2.1 is modeled as a continuous time Markov chain (CTMC). The state of the Markov chain is denoted by $(n_1, ..., n_{N-1}, c_1, ..., c_N)$. n_i is the number of items completed by workstation i but not yet by workstation $i + 1$, $i = 1, ..., N - 1$. n_i can take values between 0

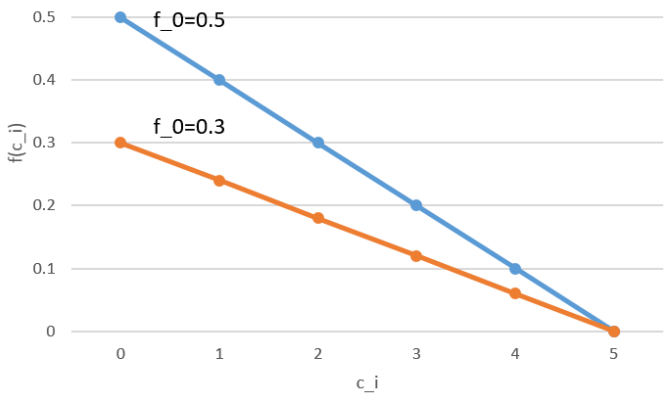

Figure 6.2 Speed-up parameter as a function of the accumulated fatigue ($C_{max} = 5$)

and K_i where $K_i = B_i + 2$, because the workstations i and $i+1$ provide additional spaces. However, when n_i reaches K_i, workstation i is blocked and occupied by the completed item. In this case, the maximum value of n_{i-1} is $K_{i-1} - 1$ and the maximum value of n_{i+1} is $K_{i+1} - 1$. c_j is the fatigue accumulation indicator for worker j, $\forall j = 1, ..., N$. It can take values between 0 and $C_{max,j}$. The set of all possible states, i.e. state space, is denoted by S.

6.2.3. Model of state-dependent behavior

We consider the following model of mean processing time adjustments.

$$w_{i,s} = w_i^{nom} \cdot \left(1 - \left(\frac{f(c_i)}{2} \cdot R_u(i) + \frac{f(c_i)}{2} \cdot R_d(i)\right)\right), i = 1, 2, ..., N \qquad (6.2)$$

where $w_{i,s}$ is the mean processing time of worker i in state s, $R_u(.)$ and $R_d(.)$ are the functions that define workers' reaction to the contents of the upstream and the downstream buffers, respectively. Due to the symmetrical and additive reaction assumptions, the two terms have equal weights and they are summed up. $R_u(.)$ returns a value of one when the upstream buffer is at its capacity while $R_d(.)$ returns a value of one when the downstream buffer is empty. Whenever both take the value one, the minimum mean processing time is reached ($w_{i,s} = w_i^{nom} \cdot (1 - f)$).

Powell and Schultz (2004) assume that these are linear functions, e.g., $R_u(i) = b_{i-1}/B_{i-1}$ and $R_d(i) = (B_i - b_i)/B_i$, in their base model for human reactions. The authors also investigate the case where a reaction takes place only when a buffer is in an extreme state, e.g., $R_u(i) = 1$ if $b_{i-1} = B_{i-1}$ and 0 otherwise, and, $R_d(i) = 1$ if $b_i = 0$ and 0 otherwise. We believe that this assumption is also realistic and interesting from the perspective of the workload and buffer allocation problems. Therefore, we define $R_u(.)$ and $R_d(.)$ such that a parameter

r determines the type of reaction.

$$R_u(i) = \begin{cases} \dfrac{e^{-r(B_{i-1}-b_{i-1})}b_{i-1}}{B_{i-1}} & i = 2, ..., N \\ 0 & i = 1 \end{cases} \tag{6.3}$$

$$R_d(i) = \begin{cases} \dfrac{e^{-r(b_i)}(B_i - b_i)}{B_i} & i = 1, 2, ..., N-1 \\ 0 & i = N \end{cases} \tag{6.4}$$

$r = 0$ ensures linear reactions while a sufficiently large value of r leads to extreme state reactions. Note that for workstation 1, (6.3) returns zero. This is because we assume that the first workstation never starves and thus the first worker actually has only the downstream buffer to get feedback about his likeliness to cause idleness. A similar logic applies for the last workstation and $R_d(N)$.

6.2.4. Transition rate matrix

Let $\mathbf{Q}$ be the transition rate matrix of the above described process. Then,

$$\mathbf{Q} = [q_{s,s'}], \qquad \forall s \neq s', s \in S, s' \in S \tag{6.5}$$

where $q_{s,s'}$ denotes the transition rate from state s to state s'.

The transition rates are determined by the mean processing times of workers, which are state-dependent. Then, the processing rate of worker i in state s ($\mu_{i,s}$) is equal to $1/w_{i,s}$. Moreover, define $I_{i,s}^{idle}$ as a variable indicating that worker i is idle (starved or blocked) in state s.

$$I_{i,s}^{idle} = \begin{cases} 1 & \text{worker } i \text{ is starved or blocked} \\ 0 & \text{otherwise} \end{cases} \tag{6.6}$$

The value of this variable is determined based on the values of state variables n_i. For example, $I_{i,s}^{idle} = 1$ due to starvation if $n_{i-1} = 0$. $I_{i,s}^{idle} = 1$ due to blocking if $n_i = K_i$ or if $n_i = K_i - 1$ and workstation $i + 1$ is also blocked.

Consider an arbitrary state $s = (n_1, ..., n_{N-1}, c_1, ..., c_N)$. The transition rate from

state s to state s' is given as follows.

$$q_{s,s'} = \begin{cases} \mu_{i,s} & \mu_{i,s} > \frac{1}{w_i^{nom}} \quad s' = (n_1, ..., n_{i-1} - 1, n_i + 1, ..., n_{N-1}, \\ & \qquad \max\left\{c_1 - I_{1,s}^{idle}, 0\right\}, ..., c_i + 1, ..., \max\left\{c_N - I_{N,s}^{idle}, 0\right\}). \\ \mu_{i,s} & \mu_{i,s} = \frac{1}{w_i^{nom}} \quad s' = (n_1, ..., n_{i-1} - 1, n_i + 1, ..., n_{N-1}, \\ & \qquad \max\left\{c_1 - I_{1,s}^{idle}, 0\right\}, ..., c_i - 1, ..., \max\left\{c_N - I_{N,s}^{idle}, 0\right\}). \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in \{1, ..., N\}$$

In words, if an item completion occurs from workstation i when the current state is s, the system moves to s' with rate $\mu_{i,s}$ and when going from s to s':

- The number of items to be completed by workstation i decreases by one $(n_{i-1}-1)$ while the number of items completed by workstation i increases by one $(n_i + 1)$.
- If, in state s, worker i uses a higher rate than nominal $(\mu_{i,s} > \frac{1}{w_i^{nom}})$, the accumulated fatigue indicator (c_i) increases by one, since the worker is more tired in state s'. On the other hand, if the worker uses the nominal speed during the time spent in state s $(\mu_{i,s} = \frac{1}{w_i^{nom}})$, c_i decreases by one since the worker is less tired in state s'.
- For all other workers $(j \in \{1, ..., N\} \setminus \{i\})$, the accumulated fatigue indicator decreases by one if the worker was idle during the time spent in state s and $c_j > 0$, otherwise remains unchanged.

Note that, if $c_i = C_{max,i}$ in state s, $\mu_{i,s} = \frac{1}{w_i^{nom}}$.

The row sums of $\mathbf{Q}$ are zero, i.e. $q_{s,s} = -\sum_{s' \neq s} q_{s,s'}$ $\forall s \in S$, by definition. Let π_s be the steady-state fraction of time the process is in state s. We can solve the following system of equations for obtaining π_s, $\forall s \in S$.

$$\mathbf{PQ} = \mathbf{0} \tag{6.7}$$

$$\mathbf{P}\underline{\mathbf{1}} = 1 \tag{6.8}$$

where $\mathbf{P}$ is the row vector formed by π_s, $s \in S$, $\mathbf{0}$ is the row vector of zeros and $\underline{\mathbf{1}}$ is the column vector of 1's.

6.3. Computing performance measures

As mentioned earlier, we investigate production line design problems considering a firm that provides service according to a just-in-time (JIT) fashion. For such a firm, it is more appropriate to evaluate the production line performance based on the stability of the output process or based on the ability to produce a given number

of products within a given time-window. We define the inter-completion time as the time between completions of two consecutive items by the last workstation and analyze the following performance measures: the expected inter-completion time ($E[T_{comp}]$), the variance of inter-completion times ($Var(T_{comp})$) and the probability that the time to complete the next M items following a completion (T_{compM}) is not longer than a given time-window ($F_{T_{compM}}(T_{window})$). The last measure is referred to as the service level (SL) of the production system.

Lagershausen and Tan (2015) provide a state-space-based methodology for determining the distribution of inter-event times (e.g. inter-start times or inter-departure times) in a queuing network that can be modeled as a CTMC. It is based on the idea of identifying the transitions that lead to the event of interest by duplicating the original state space and then conducting a first-passage-time analysis. We use this methodology for computing $E[T_{comp}]$, $Var(T_{comp})$ and $F_{T_{compM}}(T_{window})$.

Since we focus on the inter-completion times, we are interested in the transitions that lead to a completion of an item by the last workstation. Suppose that the state space is duplicated. On the duplicated state space, we should define a new process which evolves as follows. In one of the duplicates, the process moves from one state to the other with occurrences of all *other* transitions than the transitions that lead to an item completion by the last workstation. These states are called "transient states". Whenever transitions that lead to an item completion by the last workstation occur, states in the other duplicate are reached. The states in this group are referred to as "absorbing states". The distribution of the time until absorption by one of the absorbing states would then give the distribution of inter-completion times.

The content of Sections 6.3.1 and 6.3.2 is based on the methodology given by Lagershausen and Tan (2015) and follows this main logic.

6.3.1. Expected value and variance of inter-completion times

We start with duplicating the state space S. The duplicated state space is denoted S^*. In order to capture the inter-completion cycles, a new process is defined on $S \cup S^*$, which evolves as follows. The process starts at one of the transient states ($s \in S$) that is reached immediately after completion of an item by the last workstation and ends at one of the absorbing states ($s^* \in S^*$) upon completion of an item by the last workstation. Once an absorbing state is reached, transitions to other states are not possible. The transition rate matrix of the new process ($\mathbf{Q'_{comp}}$) can be

written in the following form when $s \in S \cup S^*$ are ordered so that transient states ($s \in S$) come first.

$$\mathbf{Q'_{comp}} = \left[\begin{array}{c|c} \mathbf{Q_{comp}} & \mathbf{R_{comp}} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \tag{6.9}$$

$\mathbf{Q_{comp}}$ is an $|S| \times |S|$ matrix that includes rates of transitions from transient states to transient states. In other words, its elements are equal to the rates of transitions that move the system from a state $s \in S$ to a state $s' \in S$ that is reachable from s by an item completion in a workstation *other* than the last workstation, otherwise to zero.

$\mathbf{R_{comp}}$ is also an $|S| \times |S|$ matrix. It includes rates of transitions from transient states to absorbing states. In other words, its elements are equal to the rate of transitions that move the system from a state $s \in S$ to a state $s^* \in S^*$ that is reachable from s by an item completion in the last workstation, otherwise to zero.

Then, $\mathbf{R_{comp}}$ can be obtained based on the transition rate matrix of the original process ($\mathbf{Q}$) as follows.

$$\mathbf{R_{comp}} = [r_{s,s^*}], \qquad s \in S, s^* \in S^* \tag{6.10}$$

where

$$r_{s,s^*} = \begin{cases} \mu_{N,s} & q_{s,s'} = \mu_{N,s} \\ 0 & \text{otherwise} \end{cases} \tag{6.11}$$

where $s' \in S$ is the state in which the state variables take the same values as in state $s^* \in S^*$. The transition rate to an absorbing state is always equal to the processing rate of the last worker ($\mu_{N,s}$) since an inter-completion cycle ends whenever this worker finishes processing an item.

Tables 6.1, 6.2 and 6.3 give $\mathbf{Q}$, $\mathbf{R_{comp}}$ and $\mathbf{Q_{comp}}$ respectively for an example Markov model that considers a three-stage line with no buffer spaces and that ignores the physical state of workers. The state space of the example model consists of eight states.

	$s' \in S$	1	2	3	4	5	6	7	8
$s \in S$	(n_1, n_2)	$(0,0)$	$(1,0)$	$(2,0)$	$(0,1)$	$(1,1)$	$(2,1)$	$(0,2)$	$(1,2)$
1	$(0,0)$	$-\mu_{1,1}$	$\mu_{1,1}$	0	0	0	0	0	0
2	$(1,0)$	0	$-\mu_{1,2}-\mu_{2,2}$	$\mu_{1,2}$	$\mu_{2,2}$	0	0	0	0
3	$(2,0)$	0	0	$-\mu_{2,3}$	0	$\mu_{2,3}$	0	0	0
4	$(0,1)$	$\boldsymbol{\mu_{3,4}}$	0	0	$-\mu_{1,4}-\mu_{3,4}$	$\mu_{1,4}$	0	0	0
5	$(1,1)$	0	$\boldsymbol{\mu_{3,5}}$	0	0	$-\mu_{1,5}-\mu_{2,5}-\mu_{3,5}$	$\mu_{1,5}$	$\mu_{2,5}$	0
6	$(2,1)$	0	0	$\boldsymbol{\mu_{3,6}}$	0	0	$-\mu_{2,6}-\mu_{3,6}$	0	$\mu_{2,6}$
7	$(0,2)$	0	0	0	$\boldsymbol{\mu_{3,7}}$	0	0	$-\mu_{1,7}-\mu_{3,7}$	$\mu_{1,7}$
8	$(1,2)$	0	0	0	0	$\boldsymbol{\mu_{3,8}}$	0	0	$-\mu_{3,8}$

Table 6.1 Transition rate matrix $\mathbf{Q}$ for the Markov model of a three-stage line with no buffer spaces and with no fatigue consideration

	$s^* \in S^*$	9	10	11	12	13	14	15	16
$s \in S$	(n_1, n_2)	$(0,0)$	$(1,0)$	$(2,0)$	$(0,1)$	$(1,1)$	$(2,1)$	$(0,2)$	$(1,2)$
1	$(0,0)$	0	0	0	0	0	0	0	0
2	$(1,0)$	0	0	0	0	0	0	0	0
3	$(2,0)$	0	0	0	0	0	0	0	0
4	$(0,1)$	$\boldsymbol{\mu_{3,4}}$	0	0	0	0	0	0	0
5	$(1,1)$	0	$\boldsymbol{\mu_{3,5}}$	0	0	0	0	0	0
6	$(2,1)$	0	0	$\boldsymbol{\mu_{3,6}}$	0	0	0	0	0
7	$(0,2)$	0	0	0	$\boldsymbol{\mu_{3,7}}$	0	0	0	0
8	$(1,2)$	0	0	0	0	$\boldsymbol{\mu_{3,8}}$	0	0	0

Table 6.2 Matrix $\mathbf{R_{comp}}$ for the Markov model of a three-stage line with no buffer spaces and with no fatigue consideration

Since $\mathbf{Q_{comp}}$ should include all state transitions except the ones covered by $\mathbf{R_{comp}}$, it is determined by

$$\mathbf{Q_{comp}} = \mathbf{Q} - \mathbf{R_{comp}}. \tag{6.12}$$

	$s' \in S$	1	2	3	4	5	6	7	8
$s \in S$	(n_1, n_2)	$(0,0)$	$(1,0)$	$(2,0)$	$(0,1)$	$(1,1)$	$(2,1)$	$(0,2)$	$(1,2)$
1	$(0,0)$	$-\mu_{1,1}$	$\mu_{1,1}$	0	0	0	0	0	0
2	$(1,0)$	0	$-\mu_{1,2}-\mu_{2,2}$	$\mu_{1,2}$	$\mu_{2,2}$	0	0	0	0
3	$(2,0)$	0	0	$-\mu_{2,3}$	0	$\mu_{2,3}$	0	0	0
4	$(0,1)$	0	0	0	$-\mu_{1,4}-\mu_{3,4}$	$\mu_{1,4}$	0	0	0
5	$(1,1)$	0	0	0	0	$-\mu_{1,5}-\mu_{2,5}-\mu_{3,5}$	$\mu_{1,5}$	$\mu_{2,5}$	0
6	$(2,1)$	0	0	0	0	0	$-\mu_{2,6}-\mu_{3,6}$	0	$\mu_{2,6}$
7	$(0,2)$	0	0	0	0	0	0	$-\mu_{1,7}-\mu_{3,7}$	$\mu_{1,7}$
8	$(1,2)$	0	0	0	0	0	0	0	$-\mu_{3,8}$

Table 6.3 Matrix $\mathbf{Q_{comp}}$ for the Markov model of a three-stage line with no buffer spaces and with no fatigue consideration

By conducting a first passage time analysis and assuming steady-state, Lagershausen and Tan (2015) obtain the following relationship between the steady state probabilities for an inter-completion cycle to start at state s (π_s^{entry}) for all $s \in S$ and the steady state probabilities for the process to be absorbed at state s^* ($\pi_{s^*}^{exit}$) for all $s^* \in S^*$.

$$\mathbf{P_{comp}^{exit}} = -\mathbf{P_{comp}^{entry}} \mathbf{Q_{comp}^{-1}} \mathbf{R_{comp}} \tag{6.13}$$

where $\mathbf{P_{comp}^{exit}}$ is the row vector formed by $\pi_{s^*}^{exit}$, $s^* \in S^*$ and $\mathbf{P_{comp}^{entry}}$ is the row vector formed by π_s^{entry}, $s \in S$.

Note that, the absorbing states are reached upon completion of an item by the last workstation and this is exactly the point where a new inter-completion cycle starts. Thus, the absorbing state of one inter-completion cycle is the starting transient state of the next cycle. This means that $\mathbf{P_{comp}^{enter}} = \mathbf{P_{comp}^{exit}}$ and that $\mathbf{P_{comp}^{enter}}$ can be solved from the following system of equations.

$$\mathbf{P_{comp}^{entry}} \left(\mathbf{I} + \mathbf{Q_{comp}^{-1}} \mathbf{R_{comp}} \right) = \mathbf{0} \tag{6.14}$$

$$\mathbf{P_{comp}^{entry}} \underline{\mathbf{1}} = 1 \tag{6.15}$$

Finally, the expected value and the variance of the inter-completion times can be computed using the following relationships respectively.

$$E[T_{comp}] = -\mathbf{P_{comp}^{entry}} \mathbf{Q_{comp}^{-1}} \underline{\mathbf{1}} \tag{6.16}$$

$$Var(T_{comp}) = 2\mathbf{P_{comp}^{entry}} \mathbf{Q_{comp}^{-2}} \underline{\mathbf{1}} - \left(E[T_{comp}] \right)^2 \tag{6.17}$$

6.3.2. Service level: on-time probability

The third performance measure requires the distribution of the time until the next M^{th} item completion following a completion event. Thus, M many duplicates of the state space are necessary this time.

Similarly to the previous case, we investigate the transition rate matrix of the new process defined on the extended (M times duplicated) state space ($\mathbf{Q'_{compM}}$). It can be written in the following form.

$$\mathbf{Q'_{compM}} = \left[\begin{array}{c|c} \mathbf{Q_{compM}} & \mathbf{R_{compM}} \\ \hline \mathbf{0} & \mathbf{0} \end{array} \right] \tag{6.18}$$

where

$$\mathbf{Q_{compM}} = \begin{bmatrix} \mathbf{Q_{comp}} & \mathbf{R_{comp}} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ & \mathbf{Q_{comp}} & \mathbf{R_{comp}} & \dots & \mathbf{0} & \mathbf{0} \\ . & . & . & \dots & . & . \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{Q_{comp}} & \mathbf{R_{comp}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{Q_{comp}} \end{bmatrix} \tag{6.19}$$

and

$$\mathbf{R_{compM}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ . & . & . & \dots & . & . \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{R_{comp}} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{6.20}$$

$\mathbf{Q_{compM}}$ and $\mathbf{R_{compM}}$ are $M \cdot |S| \times M \cdot |S|$ matrices.

In words, whenever one of the absorbing states is visited for the k^{th} time, the system moves to the $k+1^{st}$ group of transient states ($k = 1, 2, ..., M-1$). Finally, whenever a transition moves the system from the M^{th} group of transient states to one of the absorbing states, the process ends since this happens when M^{th} completion from the last workstation occurs.

One can obtain the steady state probability that a new inter-M^{th}-completion cycle starts at state s, $\forall s \in S$ (the probability vector $\mathbf{P^{entry}_{compM}}$), using the following system of equations.

$$\mathbf{P^{entry}_{compM}} \left(\mathbf{I} + \mathbf{Q^{-1}_{compM}} \mathbf{R_{compM}} \right) = \mathbf{0} \tag{6.21}$$

$$\mathbf{P^{entry}_{compM}} \underline{\mathbf{1}} = 1 \tag{6.22}$$

Finally, the service level of the production line ($SL = F_{T_{compM}}(T_{window})$) can be obtained from the following relationship (Lagershausen and Tan, 2015).

$$F_{T_{compM}}(t) = 1 - \mathbf{P^{entry}_{compM}} e^{\mathbf{Q_{compM}} t} \underline{\mathbf{1}} \tag{6.23}$$

6.4. Design problems

The design of unpaced production lines typically involves two optimization problems: the workload allocation and the buffer allocation problem. The workload allocation problem deals with the optimal allocation of a total amount of workload among different workstations. The buffer allocation problem deals with the optimal allocation of available storage spaces between workstations. We investigate these problems separately and simultaneously. In each case, we consider the three performance measures introduced in Section 6.3, separately, as the objective function when determining the optimal solution. In the following we mathematically describe these problems.

6.4.1. Workload allocation problem

We are interested in determining the optimal allocation of the nominal workload (w_i^{nom}, $\forall i \in \{1, ..., N\}$) for the system described in Section 6.2 under a balanced allocation of buffer spaces ($B_1 = \cdots = B_{N-1}$). We consider three different objective functions (separately) when solving the following optimization problem.

$$\textbf{Objective 1:} \text{ minimize} \quad E\left[T_{comp}\left(\mathbf{w}\right)\right] \tag{6.24}$$

$$\textbf{Objective 2:} \text{ minimize} \quad Var\left(T_{comp}\left(\mathbf{w}\right)\right) \tag{6.25}$$

$$\textbf{Objective 3:} \text{ maximize} \quad F_{T_{compM}(\mathbf{w})}(T_{window}) \tag{6.26}$$

subject to

$$\sum_{i=1}^{N} w_i^{nom} = N \tag{6.27}$$

$$w_i^{nom} > 0 \quad \forall i \in \{1, ..., N\} \tag{6.28}$$

This problem lacks a closed-form objective function. The objective value is only numerically obtainable in all three cases. Therefore, to the CTMC model, we attach a procedure that computes the value of the objective function for a given workload allocation and seeks for the optimal allocation.

6.4.2. Buffer allocation problem

We are interested in determining the allocation of the available buffer spaces (B_i, $\forall i \in \{1, ..., N-1\}$) such that the objective function is optimized for the system

described in Section 6.2 under a balanced allocation of nominal workload ($w_i^{nom} = 1$, $\forall i \in \{1, ..., N\}$). We consider three different objective functions (separately) when solving the following optimization problem.

$$\textbf{Objective 1:} \text{ minimize} \quad E\left[T_{comp}\left(\mathbf{B}\right)\right] \tag{6.29}$$

$$\textbf{Objective 2:} \text{ minimize} \quad Var\left(T_{comp}\left(\mathbf{B}\right)\right) \tag{6.30}$$

$$\textbf{Objective 3:} \text{ maximize} \quad F_{T_{compM}(\mathbf{B})}(T_{window}) \tag{6.31}$$

subject to

$$\sum_{i=1}^{N-1} B_i = B_{tot} \tag{6.32}$$

$$B_i \geq 1 \text{ and integer} \qquad \forall i \in \{1, ..., N-1\} \tag{6.33}$$

The optimization model is not allowed to cancel an inter-station buffer ($B_i \geq 1$) because we assume that buffer sizes provide workers with feedback regarding their relative speed. In addition, since this feedback is provided only in low-inventory conditions, the total number of available buffer spaces (B_{tot}) can not be very large. Due to the lack of closed-form expressions for the objective functions, the integrity of the decision variables and due to rather small B_{tot} values being considered, it is appropriate to use complete enumeration for determining the optimal solution.

6.4.3. Simultaneous workload and buffer allocation problem

In this problem, we are interested in determining the optimal allocation of the nominal workload (w_i^{nom}, $\forall i \in \{1, ..., N\}$) and the available buffer spaces (B_i, $\forall i \in \{1, ..., N-1\}$). Thus, the flexibility in allocating both of the system resources is simultaneously used.

$$\textbf{Objective 1:} \text{ minimize} \quad E\left[T_{comp}\left(\mathbf{w,B}\right)\right] \tag{6.34}$$

$$\textbf{Objective 2:} \text{ minimize} \quad Var\left(T_{comp}\left(\mathbf{w,B}\right)\right) \tag{6.35}$$

$$\textbf{Objective 3:} \text{ maximize} \quad F_{T_{compM}(\mathbf{w,B})}(T_{window}) \tag{6.36}$$

subject to

$$\sum_{i=1}^{N} w_i^{nom} = N \tag{6.37}$$

$$\sum_{i=1}^{N-1} B_i = B_{tot} \tag{6.38}$$

$$B_i \geq 1 \text{ and integer} \qquad \forall i \in \{1, ..., N-1\} \tag{6.39}$$

$$w_i^{nom} > 0 \quad \forall i \in \{1, ..., N\} \tag{6.40}$$

We obtain the optimal solution by using a combination of the approaches discussed for solving the two problems introduced in 6.4.1 and 6.4.2. We first determine all possible allocations of buffer spaces and then we compute the optimal workload allocation for each buffer space allocation, using the optimum-seeking algorithm. The buffer and workload allocation that returns the optimal objective value is selected to be the solution to this problem.

6.5. Numerical results

In the numerical study, we investigate solutions that optimize objectives 1, 2 and 3 for workload allocation, buffer allocation and simultaneous workload and buffer allocation problems assuming different environments in which the workers exhibit:

i. a state-independent behavior as commonly assumed in the operations management literature

ii. a state-dependent behavior with no fatigue as in Powell and Schultz (2004) and Heimbach et al. (2012)

iii. a state-dependent behavior with fatigue.

Comparing environments (i) and (iii) will provide the main insights on the impact of considering a more realistic human behavior model (Section 6.5.1), while comparing (ii) and (iii) will underline the importance of taking fatigue into account (Section 6.5.2).

We consider a three-stage production line ($N = 3$) and a varying total number of available buffer spaces $B_{tot} \in \{4, 8\}$. The speed-up parameter is varied in $f^0 \in \{0, 0.2, 0.4, 0.6\}$. $f^0 = 0$ resembles state-independence assumption (environment (i)). We assume that the maximum fatigue accumulation is the same for all workers ($C_{max,i} = C_{max}, \forall i = 1, ..., N$) and $C_{max} = 3$. The service level ($F_{T_{compM}}(T_{window})$) is calculated by taking $M = 1$ and $T_{window} = 3$. The reaction parameter of the mean processing time adjustments is varied in $r \in \{0, 200\}$. $r = 0$ leads to a linear worker reaction to the changes in buffer sizes while $r = 200$ means that workers only react to the extreme states of buffers (full or empty). While we present results that are mostly obtained under $r = 0$, discussions also cover the ones for $r = 200$.

In the presented tables (6.4-6.10), w_i^{nom*} denotes the optimum nominal workload allocated to worker i for $i = 1, 2, 3$ and B_i^* denotes the optimal buffer space to be placed after workstation i. The values of the following performance measures:

$E[T_{comp}]$, $1/E[T_{comp}]$ (throughput per time unit), $Var(T_{comp})$ and SL under the optimal allocation, as well as, the value of the objective function under a balanced allocation, denoted $E[T_{comp}]_b$, $Var(T_{comp})_b$, SL_b respectively for objectives 1, 2 and 3, are provided. The column "Imp" gives improvement of the objective value when the optimal allocation, rather than a balanced allocation, is used. Mathematically, it is; $1-E[T_{comp}]/E[T_{comp}]_b$, $1-Var(T_{comp})/Var(T_{comp})_b$ and SL/SL_b-1 for objectives 1, 2 and 3 respectively. The variable "fatigue" indicates whether environment (ii) or (iii) is considered. If fatigue=0, $f(c_i)$ is equal to f^0 regardless of fatigue accumulation (c_i), otherwise environment (iii) applies and $f(c_i)$ decreases in c_i. The degree of unbalance in a workload allocation is measured by the mean absolute deviation from the balanced allocation:

$$w_{unb} = \sum_{i=1}^{N} \frac{|w_i^{nom} - 1|}{N} \tag{6.41}$$

Similarly, the degree of unbalance in a buffer allocation is measured by

$$B_{unb} = \sum_{i=1}^{N-1} \frac{|B_i - \frac{B_{tot}}{N-1}|}{B_{tot}}. \tag{6.42}$$

We refer to models that consider environment (i) and the maximization of line throughput (minimization of the expected inter-completion time ($E[T_{comp}]$)) for investigating the optimal workload allocation (e.g. Hillier and Boling (1966)), buffer allocation (e.g. Hillier et al. (1993)) and simultaneous workload and buffer allocation (e.g. Hillier and So (1995)) as *traditional models* and the shape of allocations suggested by these models as *traditional guidelines* when discussing our findings in relation to those.

6.5.1. The effect of state-dependent behavior on the optimal design guidelines

First we analyze how the guidelines for the design of production lines change when a more realistic model of human pace, rather than a state-independent model, is considered. To this end, we compare the shapes of the optimal workload and buffer allocations when these are simultaneously optimized in environments (i) and (iii).

6.5.1.1. The effect on guidelines for maximizing throughput

The traditional models suggest a balanced buffer allocation in combination with a symmetrically bowl-shaped workload allocation for both $B_{tot} = 4$ and $B_{tot} = 8$ (Tables 6.4 and 6.5 for $f^0 = 0$). In a line with limited buffer spaces, the productivity

loss due to the interference between workers with stochastic processing times is best mitigated by a bowl-shaped workload allocation. The degree of unbalance in the workload allocation is 6% and 4% for $B_{tot} = 4$ and $B_{tot} = 8$ respectively.

fatigue	f^0	w_1^{nom*}	w_2^{nom*}	w_3^{nom*}	B_1^*	B_2^*	$E[T_{comp}]$	$1/E[T_{comp}]$	$Var(T_{comp})$	SL	$E[T_{comp}]_b$	Imp	w_{unb}	B_{unb}
Workload Allocation														
0	0	1.044	0.912	1.044	2	2	1.357	0.737	1.681	0.893	1.362	0.39%	6%	0%
0	0.2	0.996	1.009	0.995	2	2	1.211	0.826	1.312	0.921	1.211	0.00%	1%	0%
0	0.4	0.913	1.175	0.912	2	2	1.056	0.947	0.968	0.949	1.066	0.95%	12%	0%
0	0.6	0.728	1.547	0.725	2	2	0.874	1.144	0.636	0.976	0.927	5.70%	36%	0%
1	0.2	1.030	0.943	1.027	2	2	1.301	0.769	1.510	0.905	1.304	0.19%	4%	0%
1	0.4	1.016	0.972	1.011	2	2	1.248	0.801	1.361	0.916	1.248	0.05%	2%	0%
1	0.6	1.003	1.001	0.995	2	2	1.197	0.836	1.236	0.926	1.197	0.00%	0%	0%
Buffer Allocation														
0	0	1.0	1.0	1.0	2	2	1.362	0.734	1.686	0.893	1.362	0.00%	0%	0%
0	0.2	1.0	1.0	1.0	2	2	1.211	0.826	1.313	0.921	1.211	0.00%	0%	0%
0	0.4	1.0	1.0	1.0	2	2	1.066	0.938	1.003	0.947	1.066	0.00%	0%	0%
0	0.6	1.0	1.0	1.0	2	2	0.927	1.079	0.748	0.968	0.927	0.00%	0%	0%
1	0.2	1.0	1.0	1.0	2	2	1.304	0.767	1.509	0.904	1.304	0.00%	0%	0%
1	0.4	1.0	1.0	1.0	2	2	1.248	0.801	1.360	0.916	1.248	0.00%	0%	0%
1	0.6	1.0	1.0	1.0	2	2	1.197	0.836	1.237	0.926	1.197	0.00%	0%	0%
Simultaneous Workload and Buffer Allocation														
0	0	1.044	0.912	1.044	2	2	1.357	0.737	1.681	0.893	1.362	0.39%	6%	0%
0	0.2	0.996	1.009	0.995	2	2	1.211	0.826	1.312	0.921	1.211	0.00%	1%	0%
0	0.4	0.913	1.175	0.912	2	2	1.056	0.947	0.968	0.949	1.066	0.95%	12%	0%
0	0.6	0.728	1.547	0.725	2	2	0.874	1.144	0.636	0.976	0.927	5.70%	36%	0%
1	0.2	1.030	0.943	1.027	2	2	1.301	0.769	1.510	0.905	1.304	0.19%	4%	0%
1	0.4	1.016	0.972	1.011	2	2	1.248	0.801	1.361	0.916	1.248	0.05%	2%	0%
1	0.6	1.003	1.001	0.995	2	2	1.197	0.836	1.236	0.926	1.197	0.00%	0%	0%

Table 6.4 Optimal allocation results when minimizing the expected inter-completion time ($B_{tot} = 4$, $r = 0$)

These guidelines also hold for the optimization of objective 1 when the state-dependent behavior of human operators is considered in combination with fatigue. However, the degree of unbalance in the workload allocation is notably smaller than what is traditionally suggested. The reason is that, the need for unbalancing the workload allocation -to mitigate the productivity loss due to the interference between workers- is partially or fully compensated by the adaptive behavior of human workers, depending on the level of processing time adjustments (f^0). For a speed-up parameter of $f^0 = 0.6$, we observe that the optimal workload allocation is nearly balanced.

6.5.1.2. The effect on guidelines for minimizing the variability of output

When the objective is to minimize the variability of inter-completion times (Tables 6.6 and 6.7), a state-independent model ($f^0 = 0$) suggests that a larger space is allocated to the upstream buffer and a higher workload is allocated to the first station than the last station while maintaining a bowl-shape in the workload allocation. These guidelines also hold when the operators exhibit a state-dependent behavior in combination with fatigue, with a speed-up factor of $f^0 = 0.2$. How-

fatigue	f^0	w_1^{nom*}	w_2^{nom*}	w_3^{nom*}	B_1^*	B_2^*	$E[T_{comp}]$	$1/E[T_{comp}]$	$Var(T_{comp})$	SL	$E[T_{comp}]_b$	Imp	w_{unb}	B_{unb}
Workload Allocation														
0	0	1.028	0.943	1.028	4	4	1.235	0.810	1.458	0.914	1.238	0.27%	4%	0%
0	0.2	0.989	1.023	0.988	4	4	1.108	0.902	1.156	0.936	1.109	0.03%	2%	0%
0	0.4	0.923	1.156	0.922	4	4	0.977	1.023	0.878	0.958	0.988	1.07%	10%	0%
0	0.6	0.767	1.468	0.765	4	4	0.831	1.204	0.610	0.979	0.875	5.06%	31%	0%
1	0.2	1.020	0.960	1.020	4	4	1.201	0.832	1.359	0.920	1.203	0.14%	3%	0%
1	0.4	1.012	0.977	1.010	4	4	1.169	0.856	1.271	0.926	1.169	0.05%	2%	0%
1	0.6	1.004	0.994	1.001	4	4	1.137	0.879	1.192	0.932	1.137	0.00%	0%	0%
Buffer Allocation														
0	0	1.0	1.0	1.0	4	4	1.238	0.807	1.463	0.913	1.238	0.00%	0%	0%
0	0.2	1.0	1.0	1.0	4	4	1.109	0.902	1.158	0.936	1.109	0.00%	0%	0%
0	0.4	1.0	1.0	1.0	4	4	0.988	1.012	0.908	0.956	0.988	0.00%	0%	0%
0	0.6	1.0	1.0	1.0	4	4	0.875	1.143	0.702	0.972	0.875	0.00%	0%	0%
1	0.2	1.0	1.0	1.0	4	4	1.203	0.831	1.361	0.920	1.203	0.00%	0%	0%
1	0.4	1.0	1.0	1.0	4	4	1.169	0.855	1.271	0.926	1.169	0.00%	0%	0%
1	0.6	1.0	1.0	1.0	4	4	1.137	0.879	1.192	0.932	1.137	0.00%	0%	0%
Simultaneous Workload and Buffer Allocation														
0	0	1.028	0.943	1.028	4	4	1.235	0.810	1.458	0.914	1.238	0.27%	4%	0%
0	0.2	0.989	1.023	0.988	4	4	1.108	0.902	1.156	0.936	1.109	0.03%	2%	0%
0	0.4	0.923	1.156	0.922	4	4	0.977	1.023	0.878	0.958	0.988	1.07%	10%	0%
0	0.6	0.767	1.468	0.765	4	4	0.831	1.204	0.610	0.979	0.875	5.06%	31%	0%
1	0.2	1.020	0.960	1.020	4	4	1.201	0.832	1.359	0.920	1.203	0.14%	3%	0%
1	0.4	1.012	0.977	1.010	4	4	1.169	0.856	1.271	0.926	1.169	0.05%	2%	0%
1	0.6	1.004	0.994	1.001	4	4	1.137	0.879	1.192	0.932	1.137	0.00%	0%	0%

Table 6.5 Optimal allocation results when minimizing the expected inter-completion time ($B_{tot} = 8$, $r = 0$)

ever, the degree of unbalance in the workload allocation is smaller than the degree a state-independent model indicates. This is due to the self-balancing mechanism brought into the system by the adaptive behavior of human workers.

For $f^0 > 0.2$, we observe that allocating even more spaces to the upstream buffer is preferred ($B_1^* = 7$, $B_2^* = 1$ instead of $B_1^* = 5$ and $B_2^* = 3$) when $B_{tot} = 8$. The existence or the strength of the state-dependent behavior does not affect the optimal buffer allocation when the total number of available buffer spaces is low ($B_{tot} = 4$, $B_1^* = 3$ and $B_2^* = 1$), because in this case, it is not possible to shift more spaces from the downstream to the upstream buffer due to constraint (6.39). On the other hand, the optimal workload allocation for $f^0 > 0.2$ changes to follow a decreasing pattern both for $B_{tot} = 4$ and $B_{tot} = 8$.

Hendricks (1992) states that, although the buffer allocation that maximizes the throughput is not always the same as the one that minimizes output variability, it leads to a near optimal performance in terms of output variability. Our results for the (pure) buffer allocation problem under $f^0 = 0$ agree with his findings. A balanced buffer allocation optimizes both objective 1 and objective 2 (e.g. see Tables 6.4 and 6.6). Therefore, the system performs optimally in terms of $Var(T_{comp})$ in both cases. Differently to his work, we also consider the optimization of buffer allocation *simultaneously* with optimization of workload allocation. In this case, the

fatigue	f^0	w_1^{nom*}	w_2^{nom*}	w_3^{nom*}	B_1^*	B_2^*	$E[T_{comp}]$	$1/E[T_{comp}]$	$Var(T_{comp})$	SL	$Var(T_{comp})_b$	Imp	w_{unb}	B_{unb}
Workload Allocation														
0	0	1.027	0.943	1.029	2	2	1.358	0.737	1.679	0.893	1.686	0.42%	4%	0%
0	0.2	0.979	1.042	0.979	2	2	1.212	0.825	1.311	0.921	1.313	0.20%	3%	0%
0	0.4	0.899	1.205	0.896	2	2	1.056	0.947	0.967	0.949	1.003	3.50%	14%	0%
0	0.6	0.739	1.530	0.732	2	2	0.874	1.144	0.636	0.976	0.748	15.05%	35%	0%
1	0.2	1.021	0.973	1.006	2	2	1.302	0.768	1.507	0.905	1.509	0.13%	2%	0%
1	0.4	1.017	0.999	0.984	2	2	1.248	0.801	1.359	0.916	1.360	0.08%	1%	0%
1	0.6	1.013	1.022	0.965	2	2	1.198	0.835	1.233	0.926	1.237	0.28%	2%	0%
Buffer Allocation														
0	0	1.0	1.0	1.0	2	2	1.362	0.734	1.686	0.893	1.686	0.00%	0%	0%
0	0.2	1.0	1.0	1.0	3	1	1.235	0.810	1.308	0.920	1.313	0.41%	0%	50%
0	0.4	1.0	1.0	1.0	3	1	1.082	0.924	0.989	0.947	1.003	1.32%	0%	50%
0	0.6	1.0	1.0	1.0	3	1	0.936	1.069	0.732	0.969	0.748	2.25%	0%	50%
1	0.2	1.0	1.0	1.0	3	1	1.325	0.754	1.498	0.903	1.509	0.74%	0%	50%
1	0.4	1.0	1.0	1.0	3	1	1.261	0.793	1.334	0.916	1.360	1.87%	0%	50%
1	0.6	1.0	1.0	1.0	3	1	1.202	0.832	1.201	0.928	1.237	2.90%	0%	50%
Simultaneous Workload and Buffer Allocation														
0	0	1.088	0.933	0.979	3	1	1.378	0.726	1.666	0.892	1.686	1.22%	6%	50%
0	0.2	1.037	1.032	0.932	3	1	1.229	0.813	1.298	0.921	1.313	1.16%	5%	50%
0	0.4	0.952	1.192	0.856	3	1	1.070	0.934	0.955	0.949	1.003	4.73%	13%	50%
0	0.6	0.773	1.511	0.715	3	1	0.883	1.133	0.624	0.977	0.748	16.59%	34%	50%
1	0.2	1.065	0.964	0.972	3	1	1.317	0.759	1.484	0.905	1.509	1.67%	4%	50%
1	0.4	1.045	0.990	0.965	3	1	1.258	0.795	1.327	0.917	1.360	2.41%	3%	50%
1	0.6	1.028	1.014	0.958	3	1	1.201	0.833	1.195	0.928	1.237	3.33%	3%	50%

Table 6.6 Optimal allocation results when minimizing the variance of the inter-completion times ($B_{tot} = 4$, $r = 0$)

fatigue	f^0	w_1^{nom*}	w_2^{nom*}	w_3^{nom*}	B_1^*	B_2^*	$E[T_{comp}]$	$1/E[T_{comp}]$	$Var(T_{comp})$	SL	$Var(T_{comp})_b$	Imp	w_{unb}	B_{unb}
Workload Allocation														
0	0	1.020	0.957	1.022	4	4	1.235	0.809	1.458	0.913	1.463	0.35%	3%	0%
0	0.2	0.979	1.042	0.979	4	4	1.109	0.902	1.155	0.936	1.158	0.26%	3%	0%
0	0.4	0.911	1.181	0.908	4	4	0.978	1.023	0.877	0.958	0.908	3.38%	12%	0%
0	0.6	0.761	1.491	0.748	4	4	0.831	1.204	0.610	0.979	0.702	13.20%	33%	0%
1	0.2	1.019	0.975	1.006	4	4	1.202	0.832	1.359	0.920	1.361	0.14%	2%	0%
1	0.4	1.017	0.992	0.990	4	4	1.169	0.855	1.270	0.926	1.271	0.09%	1%	0%
1	0.6	1.016	1.008	0.976	4	4	1.138	0.879	1.190	0.932	1.192	0.18%	2%	0%
Buffer Allocation														
0	0	1.0	1.0	1.0	4	4	1.238	0.807	1.463	0.913	1.463	0.00%	0%	0%
0	0.2	1.0	1.0	1.0	5	3	1.115	0.897	1.155	0.936	1.158	0.23%	0%	25%
0	0.4	1.0	1.0	1.0	5	3	0.992	1.008	0.902	0.956	0.908	0.69%	0%	25%
0	0.6	1.0	1.0	1.0	7	1	0.904	1.107	0.683	0.973	0.702	2.76%	0%	75%
1	0.2	1.0	1.0	1.0	5	3	1.210	0.827	1.358	0.919	1.361	0.16%	0%	25%
1	0.4	1.0	1.0	1.0	7	1	1.220	0.820	1.263	0.923	1.271	0.63%	0%	75%
1	0.6	1.0	1.0	1.0	7	1	1.165	0.858	1.146	0.933	1.192	3.92%	0%	75%
Simultaneous Workload and Buffer Allocation														
0	0	1.048	0.954	0.998	5	3	1.241	0.806	1.455	0.913	1.463	0.53%	3%	25%
0	0.2	1.005	1.038	0.957	5	3	1.114	0.898	1.151	0.936	1.158	0.63%	3%	25%
0	0.4	0.935	1.176	0.889	5	3	0.982	1.018	0.872	0.958	0.908	4.00%	12%	25%
0	0.6	0.813	1.468	0.719	7	1	0.860	1.163	0.599	0.979	0.702	14.69%	31%	75%
1	0.2	1.041	0.972	0.987	5	3	1.206	0.829	1.352	0.920	1.361	0.66%	3%	25%
1	0.4	1.082	0.970	0.948	7	1	1.204	0.831	1.233	0.926	1.271	3.01%	5%	75%
1	0.6	1.060	0.991	0.949	7	1	1.158	0.864	1.128	0.935	1.192	5.37%	4%	75%

Table 6.7 Optimal allocation results when minimizing the variance of the inter-completion times ($B_{tot} = 8$, $r = 0$)

optimal solution for objective 2 differs from the traditional guidelines and it provides a slightly lower $Var(T_{comp})$ (1.666 and 1.681 in Tables 6.6 and 6.4 respectively).

On the other hand, when a state-dependent behavior is in place, the difference between the buffer allocation schemes obtained using objective 1 and objective 2 increases, in both cases where these allocations are optimized separately and simultaneously with the allocation of workload. As a result, the system performance in terms of $Var(T_{comp})$ shows a larger difference between these two allocation alternatives. The variance of the output process is approximately 6% lower in Table 6.7 ($Var(T_{comp}) = 1.128$) than it is in Table 6.5 ($Var(T_{comp}) = 1.192$), when buffer and workload allocations are simultaneously optimized in an environment where $f^0 = 0.6$.

6.5.1.3. The effect on guidelines for maximizing the service level

Following the traditional guidelines, i.e. using a symmetrically bowl-shaped workload allocation and a balanced buffer allocation, seems appropriate for maximizing the service level in environment (i) (see Tables 6.8 and 6.9).

fatigue	f^0	w_1^{nom*}	w_2^{nom*}	w_3^{nom*}	B_1^*	B_2^*	$E[T_{comp}]$	$1/E[T_{comp}]$	$Var(T_{comp})$	SL	SL_b	Imp	w_{unb}	B_{unb}
Workload Allocation														
0	0	1.040	0.919	1.041	2	2	1.357	0.737	1.680	0.894	0.893	0.10%	5%	0%
0	0.2	0.988	1.024	0.988	2	2	1.211	0.826	1.311	0.921	0.921	0.01%	2%	0%
0	0.4	0.903	1.197	0.900	2	2	1.056	0.947	0.967	0.949	0.947	0.26%	13%	0%
0	0.6	0.738	1.532	0.730	2	2	0.874	1.144	0.636	0.976	0.968	0.91%	35%	0%
1	0.2	1.028	0.953	1.019	2	2	1.301	0.768	1.508	0.905	0.904	0.04%	3%	0%
1	0.4	1.019	0.985	0.996	2	2	1.248	0.801	1.359	0.916	0.916	0.01%	1%	0%
1	0.6	1.011	1.016	0.973	2	2	1.197	0.835	1.233	0.926	0.926	0.02%	2%	0%
Buffer Allocation														
0	0	1.0	1.0	1.0	2	2	1.362	0.734	1.686	0.893	0.893	0.00%	0%	0%
0	0.2	1.0	1.0	1.0	2	2	1.211	0.826	1.313	0.921	0.921	0.00%	0%	0%
0	0.4	1.0	1.0	1.0	3	1	1.082	0.924	0.989	0.947	0.947	0.02%	0%	50%
0	0.6	1.0	1.0	1.0	3	1	0.936	1.069	0.732	0.969	0.968	0.10%	0%	50%
1	0.2	1.0	1.0	1.0	2	2	1.304	0.767	1.509	0.904	0.904	0.00%	0%	0%
1	0.4	1.0	1.0	1.0	3	1	1.261	0.793	1.334	0.916	0.916	0.06%	0%	50%
1	0.6	1.0	1.0	1.0	3	1	1.202	0.832	1.201	0.928	0.926	0.22%	0%	50%
Simultaneous Workload and Buffer Allocation														
0	0	1.040	0.919	1.041	2	2	1.357	0.737	1.680	0.894	0.893	0.10%	5%	0%
0	0.2	0.988	1.024	0.988	2	2	1.211	0.826	1.311	0.921	0.921	0.01%	2%	0%
0	0.4	0.961	1.186	0.852	3	1	1.070	0.935	0.955	0.949	0.947	0.29%	12%	50%
0	0.6	0.773	1.512	0.715	3	1	0.883	1.133	0.624	0.977	0.968	0.98%	34%	50%
1	0.2	1.084	0.944	0.972	3	1	1.316	0.760	1.485	0.905	0.904	0.04%	6%	50%
1	0.4	1.056	0.979	0.965	3	1	1.257	0.795	1.328	0.917	0.916	0.15%	4%	50%
1	0.6	1.032	1.011	0.957	3	1	1.201	0.833	1.195	0.928	0.926	0.27%	3%	50%

Table 6.8 Optimal allocation results when maximizing the service level ($B_{tot} = 4$, $r = 0$)

The results show that, although the allocation guidelines are shape-wise similar, a smaller degree of unbalance in the workload allocation is needed in environment (iii) when $f^0 = 0.2$ for $B_{tot} = 8$ and for $B_{tot} = 4$ when $f^0 = 0.2$ and $r = 200$. Note that, the state-dependent behavior is weaker when there are larger buffers and when

the amount of pace increase (f^0) is smaller. Furthermore, workers speed-up less frequently when they react only to the extreme states of buffers ($r = 200$).

In all other cases, where the effect of state-dependent behavior is stronger, the service level maximizing allocations in environment (iii) are different from the traditional ones also shape-wise. Larger buffer spaces are needed in the upstream part of the line. The optimal allocation of buffers change from a balanced one to $B_1^* = 3$, $B_2^* = 1$ when $B_{tot} = 4$ and from balanced to $B_1^* = 5$, $B_2^* = 3$ or to $B_1^* = 7$, $B_2^* = 1$ when $B_{tot} = 8$. The workload allocation for $f^0 > 0.2$ follows a decreasing pattern for $B_{tot} = 4$. For $B_{tot} = 8$, it follows a decreasing pattern under $r = 0$ and a bowl pattern with a higher workload allocated to the first station under $r = 200$. In the latter case, having more work allocated to the last worker than the middle worker helps triggering speed-ups of the last worker, since workers react only to extreme states of buffers and this type of allocation increases the chance that the second buffer becomes "full".

fatigue	f^0	w_1^{nom*}	w_2^{nom*}	w_3^{nom*}	B_1^*	B_2^*	$E[T_{comp}]$	$1/E[T_{comp}]$	$Var(T_{comp})$	SL	SL_b	Imp	w_{unb}	B_{unb}
Workload Allocation														
0	0	1.026	0.948	1.027	4	4	1.235	0.810	1.458	0.914	0.913	0.06%	3%	0%
0	0.2	0.984	1.032	0.984	4	4	1.109	0.902	1.155	0.936	0.936	0.01%	2%	0%
0	0.4	0.915	1.174	0.912	4	4	0.978	1.023	0.877	0.958	0.956	0.22%	12%	0%
0	0.6	0.760	1.493	0.747	4	4	0.831	1.204	0.610	0.979	0.972	0.71%	33%	0%
1	0.2	1.020	0.966	1.014	4	4	1.201	0.832	1.359	0.920	0.920	0.02%	2%	0%
1	0.4	1.015	0.984	1.001	4	4	1.169	0.855	1.270	0.926	0.926	0.01%	1%	0%
1	0.6	1.011	1.003	0.986	4	4	1.138	0.879	1.191	0.932	0.932	0.01%	1%	0%
Buffer Allocation														
0	0	1.0	1.0	1.0	4	4	1.238	0.807	1.463	0.913	0.913	0.00%	0%	0%
0	0.2	1.0	1.0	1.0	4	4	1.109	0.902	1.158	0.936	0.936	0.00%	0%	0%
0	0.4	1.0	1.0	1.0	5	3	0.992	1.008	0.902	0.956	0.956	0.03%	0%	25%
0	0.6	1.0	1.0	1.0	7	1	0.904	1.107	0.683	0.973	0.972	0.10%	0%	75%
1	0.2	1.0	1.0	1.0	4	4	1.203	0.831	1.361	0.920	0.920	0.00%	0%	0%
1	0.4	1.0	1.0	1.0	4	4	1.169	0.855	1.271	0.926	0.926	0.00%	0%	0%
1	0.6	1.0	1.0	1.0	7	1	1.165	0.858	1.146	0.933	0.932	0.15%	0%	75%
Simultaneous Workload and Buffer Allocation														
0	0	1.026	0.948	1.027	4	4	1.235	0.810	1.458	0.914	0.913	0.06%	3%	0%
0	0.2	0.984	1.032	0.984	4	4	1.109	0.902	1.155	0.936	0.936	0.01%	2%	0%
0	0.4	0.940	1.170	0.890	5	3	0.982	1.018	0.872	0.958	0.956	0.25%	11%	25%
0	0.6	0.792	1.474	0.734	6	2	0.842	1.187	0.601	0.979	0.972	0.77%	32%	50%
1	0.2	1.020	0.966	1.014	4	4	1.201	0.832	1.359	0.920	0.920	0.02%	2%	0%
1	0.4	1.036	0.983	0.981	5	3	1.172	0.853	1.259	0.926	0.926	0.04%	2%	25%
1	0.6	1.064	0.992	0.944	7	1	1.158	0.864	1.128	0.935	0.932	0.31%	4%	75%

Table 6.9 Optimal allocation results when maximizing the service level ($B_{tot} = 8$, $r = 0$)

Recall that evaluation of the service level of a line requires obtaining the distribution of the output process, which is computationally more expensive than analyzing its mean and variance. From this perspective it is important to note that, following the variance minimizing guidelines is favorable for a firm which has a small space for buffers ($B_{tot} = 4$), operates in environment (iii) where $f^0 > 0.2$ and targets the maximum service level.

6.5.1.4. The effect on the value of pure buffer allocation

Another interesting question to be answered is whether the optimization of the workload allocation alone or the optimization of the buffer allocation alone provides more benefit over its balanced setting. The traditional guidelines suggest that the optimization of the buffer allocation provides no benefit over its balanced setting. Moreover, Hillier (2013) and Hillier and Hillier (2006), who consider objective functions that account for the inventory costs in addition to the revenue from throughput, find a significantly larger impact of the pure workload allocation than the pure buffer allocation.

Similarly, the results we obtain under the traditional setting (Tables 6.4 and 6.5 for $f^0 = 0$) and when maximizing the service level in environment (i) (Tables 6.8 and 6.9) suggest that the unbalancing/optimization of the workload is sufficient for obtaining the optimal performance. When one considers minimization of the variance of inter-completion times in environment (i), optimizing the buffer allocation provides value only when considered simultaneously with the workload allocation (Tables 6.6 and 6.7).

On the other hand, the resource for which an optimal unbalance provides a larger improvement over its balanced allocation might show differences in the existence of a state-dependent behavior. In case of the first objective, optimization of the workload allocation remains sufficient for obtaining the optimal performance. Under the second objective, optimizing the buffer allocation alone provides benefit. Furthermore, the improvement provided by unbalancing of buffers is larger than the improvement provided by unbalancing the workload allocation. Only in one instance where the speed-up parameter is small ($f^0 = 0.2$), $B_{tot} = 8$ and workers react only to the extreme cases of buffers ($r = 200$), we find a larger impact of workload allocation. This result also explains why the difference between the allocation schemes obtained using objective 1 and objective 2 increases in the existence of a state-dependent behavior (discussed in Section 6.5.1.2). When maximizing the service level in an environment where $f^0 = 0.6$ the benefit of shifting a space from the downstream- to the upstream buffer (e.g. setting $B_1 = 3$ and $B_2 = 1$ instead of $B_1 = 2$ and $B_2 = 2$) is larger than the benefit of optimizing the workload allocation.

6.5.2. The importance of taking fatigue into account

If workers are assumed to be able to speed-up whenever the state of the buffers feeds them back with the information that they are likely to be the cause of idleness,

an optimization model would suggest taking the most benefit out of this ability for bringing more capacity (higher work pace) into the system. Our results show that an optimal workload allocation in environment (ii) typically has the shape of a reversed bowl under all objectives. A reversed-bowl shape leads the highest amount of work to be allocated to the intermediate worker. The system benefits from this kind of an allocation because the worker in the middle of the line is assumed to be able to increase his speed with a factor of f^0 whenever needed, while the other workers can only reach up to a speed-up factor of $f^0/2$. As the speed-up parameter f^0 increases, the optimal solution is to allocate more and more workload to this worker. The degree of unbalance is typically very large. In Table 6.4 we see that the degree of unbalance in workload allocation is 36% when $f^0 = 0.6$. As a result, the middle worker is expected to handle more than the half of the total workload. The workload allocation guidelines reported by Heimbach et al. (2012), who straightforwardly incorporate the behavioral model suggested by Powell and Schultz (2004) into a simulation model, are also in this direction.

It is also interesting to observe that, the effect of the worker reaction type ($r = 0$ or $r = 200$) on the system performance depends on whether or not one takes fatigue into account. Assuming that the number of consecutive items a worker can increase his pace is unlimited leads to observing a better system performance (based on all three measures both for $B_{tot} = 4$ and $B_{tot} = 8$) under the linear reaction ($r = 0$) than under an extreme state reaction assumption ($r = 200$). The conclusion regarding this effect is the opposite when the number of consecutive items a worker can speed-up is limited: The system performance under $r = 200$ is better than it is under $r = 0$. This is because when workers only react to the extreme states of buffers they naturally have a higher chance for recovery. As a result, whenever they speed-up, they do this with a higher adjustment factor. In other words, the ignorance of the need for a recovery leads to estimating a better system performance when workers speed-up more frequency (e.g. see Tables 6.4 and 6.10).

6.6. Discussion and conclusion

This chapter investigated the design of unpaced production lines with human operators. It utilized a Markovian model adhering to the classical assumptions except state-independence and stationarity (i.e. no tiredness) in service process. The differences in the optimal workload and buffer allocation guidelines, when they are separately and simultaneously optimized, were investigated. In addition to the typically investigated objective of throughput maximization (minimization of the expected inter-completion time), we considered the minimization of the output

fatigue	f^0	w_1^{nom*}	w_2^{nom*}	w_3^{nom*}	B_1^*	B_2^*	$E[T_{comp}]$	$1/E[T_{comp}]$	$Var(T_{comp})$	SL	$E[T_{comp}]_b$	Imp	w_{unb}	B_{unb}
Workload Allocation														
0	0	1.044	0.912	1.044	2	2	1.357	0.737	1.681	0.893	1.362	0.39%	6%	0%
0	0.2	1.002	0.996	1.001	2	2	1.230	0.813	1.343	0.918	1.230	0.00%	0%	0%
0	0.4	0.926	1.151	0.923	2	2	1.098	0.911	1.031	0.943	1.105	0.59%	10%	0%
0	0.6	0.685	1.633	0.682	2	2	0.941	1.063	0.720	0.970	0.987	4.66%	42%	0%
1	0.2	1.031	0.940	1.029	2	2	1.300	0.769	1.503	0.905	1.302	0.20%	4%	0%
1	0.4	1.019	0.967	1.015	2	2	1.244	0.804	1.347	0.917	1.245	0.07%	2%	0%
1	0.6	1.007	0.992	1.001	2	2	1.190	0.840	1.214	0.927	1.190	0.01%	1%	0%
Buffer Allocation														
0	0	1.0	1.0	1.0	2	2	1.362	0.734	1.686	0.893	1.362	0.00%	0%	0%
0	0.2	1.0	1.0	1.0	2	2	1.230	0.813	1.342	0.918	1.230	0.00%	0%	0%
0	0.4	1.0	1.0	1.0	2	2	1.105	0.905	1.058	0.941	1.105	0.00%	0%	0%
0	0.6	1.0	1.0	1.0	1	3	0.984	1.016	0.844	0.960	0.987	0.32%	0%	25%
1	0.2	1.0	1.0	1.0	2	2	1.302	0.768	1.503	0.905	1.302	0.00%	0%	0%
1	0.4	1.0	1.0	1.0	2	2	1.245	0.803	1.346	0.916	1.245	0.00%	0%	0%
1	0.6	1.0	1.0	1.0	2	2	1.190	0.840	1.214	0.927	1.190	0.00%	0%	0%
Simultaneous Workload and Buffer Allocation														
0	0	1.044	0.912	1.044	2	2	1.357	0.737	1.681	0.893	1.362	0.39%	6%	0%
0	0.2	1.002	0.996	1.001	2	2	1.230	0.813	1.343	0.918	1.230	0.00%	0%	0%
0	0.4	0.926	1.151	0.923	2	2	1.098	0.911	1.031	0.943	1.105	0.59%	10%	0%
0	0.6	0.717	1.661	0.622	1	3	0.933	1.072	0.718	0.970	0.987	5.50%	44%	25%
1	0.2	1.031	0.940	1.029	2	2	1.300	0.769	1.503	0.905	1.302	0.20%	4%	0%
1	0.4	1.019	0.967	1.015	2	2	1.244	0.804	1.347	0.917	1.245	0.07%	2%	0%
1	0.6	1.007	0.992	1.001	2	2	1.190	0.840	1.214	0.927	1.190	0.01%	1%	0%

Table 6.10 Optimal allocation results when minimizing the expected inter-completion time ($B_{tot} = 4$, $r = 200$)

variability and the maximization of service level, focusing on the interests of a firm providing service according to a JIT fashion.

Our results showed that:

- For the throughput maximization of an unpaced production line with human operators, the traditional guidelines (a balanced buffer allocation in combination with a symmetrically bowl-shaped workload allocation) hold. However, the optimal degree of unbalance in the workload allocation is overestimated by models ignoring the adaptive behavior of workers.
- The minimum variability in the output process of a line in which workers exhibit a state-dependent behavior, can be achieved by following a decreasing pattern both in buffer and workload allocations. Furthermore, when it comes to minimizing the output variability, the state-dependent behavior of human operators alters the existing opinion that optimization of the buffer allocation is less effective than optimization of the workload allocation.
- Although a state-independent model would suggest following very similar guidelines to the traditional ones for achieving the maximum service level, we observe that both the total workload and the total buffer spaces should typically be allocated in a decreasing pattern in a line with human operators. The service level is

closer to its maximum level under the variability minimizing allocation than it is under the throughput maximizing allocation, when the total number of available buffer spaces to allocate is small.

These findings are important because they show that the state-dependent behavior of human operators in production lines requires deviating from the traditional guidelines for the optimal allocation of workload and buffer spaces which are known for decades. In addition to showing that the optimal allocation guidelines change, they also show in which direction this happens.

7. Conclusions

7.1. Summary

The methodological contribution presented in Chapter 3 is the modeling of the priority sequencing problem as a Markov decision process (MDP) for numerically obtaining the optimal priority sequencing policy with respect to the long-run average cost of tardiness and its implication for two customer-related measures. The MDP model can also serve as a performance evaluation tool for a queuing system that works under simple due-date-based priority sequencing rules such as the earliest-due-date-first rule (EDD), when the action space of MDP is restricted according to the rule. Since, to the best of our knowledge, there is no exact analytical characterization of the on-time probability and expected tardiness measures for a queuing system that receives arrivals with stochastic customer-required leadtimes and works under a due-date-based rule, our model, even when used only as a performance evaluation tool, enables an exact evaluation of these measures numerically. Its topical contributions are the assessment of the impact of priority sequencing decisions on the two performance measures and the benchmarking of simple priority sequencing rules against the optimal as opposed to the existing studies that compare priority sequencing rules with each other. It has a practical value if a simple priority sequencing rule that is easy to understand and implement is known to provide a system performance close to the optimal because modeling a real system as an MDP and elaborating the optimal solution would be a difficult task facing the curse of dimensionality for larger problems.

The model presented in Chapter 4 is the first to tackle the joint pricing, leadtime quotation and due-date-based order dispatching problem using the theory for Markov decision processes. It enables one to numerically obtain the optimal policy which constitutes a benchmark for simpler approaches. It uses an S-shaped logistical response function as the model for decisions customers make about placing an order or not. Using this model, we shed light on the benefit a full collaboration between marketing and manufacturing departments provides, by comparing the sequential approaches with the simultaneous approach. This benefit can be traded off against the necessary effort to ensure such a collaboration.

Chapter 5 presents a model that extends the literature by considering the state-dependent behavior of assembly line workers in combination with fatigue for an accurate assessment of the performance comparison between a paced and an un-

paced line design. The model accounts for the task time variability that comes into play due to offering a variety of products, in addition to the variability in processing times. As the findings of empirical studies suggest, the distribution of processing times is modeled using a positively-skewed Johnson distribution in an unpaced line while it is modeled using a Normal distribution in a paced setting. The effect of worker heterogeneity in the ability to adapt to the system state as well as in the work experience is investigated and guidelines are provided for the heterogeneous workforce placement.

Methodologically, the continuous time Markov model presented in Chapter 6 contributes to literature by enabling an exact evaluation of the performance of an unpaced production line with human workers who adjust their expected processing times according to the state of the up- and downstream buffers as well as their physical state. The model enables investigation of the workload and the buffer allocation problems with various performance measures in the objective function. Although the literature typically considers the first moment of the output process, e.g. the expected throughput, we additionally investigate the variability as well as the distribution of the output process. Using this model, we revisit the traditional guidelines for the maximization of expected throughput i.e. the bowl phenomenon and highlight the directions in which these guidelines change when human aspect is specifically accounted for and when other performance measures -more relevant for JIT production- are considered.

The findings of this thesis can be summarized as follows.

- By employing simple due-date-based priority sequencing rules, it is possible to achieve near optimal performance. When a fixed cost is part of the tardiness penalty, (i) adhering to the EDD rule -which is known to be optimal under a proportional penalty- might result in substantially larger costs than the optimal and (ii) postponing the priority sequencing decisions to order completion rather than making them upon arrival provides improvement potential.
- The sub-optimality of a firm's performance due to failing to make the marketing and operations departments collaborate for developing a joint quotation and dispatching policy might be considerable when the tardiness penalty is a fixed cost and the customers are highly heterogeneous in their price and leadtime sensitivities. On the other hand, it is negligible when a proportional tardiness penalty applies and the customers are rather homogeneous in their sensitivities. Furthermore, by developing a joint quotation and dispatching policy, it is possible

for a firm to attract more customers while improving the service level.

- In many real-world settings, a higher efficiency can be obtained for lines with human operators when an unpaced design, rather than a paced design, is adopted. However the efficiency improvement is smaller than that would be estimated by models ignoring human behavior or fatigue. Furthermore, paced lines lead to a more stable output process than unpaced lines and this stability is intensified in lines with human operators.
- A firm that uses human workforce in an unpaced production line can maximize the throughput by using a balanced buffer allocation and a bowl-shaped workload allocation with a smaller degree of unbalance than that is suggested by the traditional models. The firm can minimize the variability of the output process or maximize the service level by using a decreasing pattern in the buffer and in the workload allocations. As opposed to the existing opinion, the optimization of the buffer allocation can provide more benefit than the optimization of the workload allocation.

7.2. Limitations and future research directions

This thesis mainly introduces exact numerical models for the analysis of short and long-term decisions related to management of manufacturing systems. The major limitation of this work is the exponential increase of the state space in the problem size. The required computational time can quickly become prohibitively large.

For the analysis of large problem sizes, meta-modeling of the numerically obtainable optimal policy, the development of useful simple rules and heuristics and the investigation of accurate approximation methods are useful future research directions for making a step towards a real-world application.

Another important future research direction is the investigation of details regarding the state-dependent behavior of workers via the use of controlled behavioral experiments. Finding answers to the following open questions about this behavior will lead to obtaining more precise guidelines for the performance improvement and the optimal design of serial production lines with human operators: What is the maximum achievable speed adjustment factor for a fully recovered worker? What is the maximum number of consecutive items for which a worker can achieve an increased paced if he needs to repeat this behavior? What is the functional relationship between the achievable speed adjustment factor and the accumulated fatigue? Furthermore, an attempt towards a better understanding of the order

placement decisions of customers who are given a price and leadtime quote, via the use of real data, is valuable.

A. Appendix

Let Y represent the processing time of an arbitrary order and $\mathcal{G}_Y(z)$ be its generating function. Since Y is geometrically distributed with parameter β, $\mathcal{G}_Y(z) = \frac{\beta z}{1-(1-\beta)z}$. Define $\mathcal{G}_{N_{system}|N_{system}>0}(z)$ as the generating function of the number of orders in the system given that it is not idle. $\mathcal{G}_{W^f_{queue}}(z)$ and $\mathcal{G}_{W^f}(z)$ are the generating functions of the waiting time in the queue and production leadtime, respectively.

$$\mathcal{G}_{N_{system}|N_{system}>0}(z) = \sum_{n=1}^{K} z^n \cdot \frac{\pi_n}{1-\pi_0} = \frac{z(\gamma-\beta)(1-(z\gamma/\beta)^K)}{(z\gamma-\beta)(1-(\gamma/\beta)^K)} \tag{A.1}$$

$$\mathcal{G}_{W^f}(z) = \mathcal{G}_{N_{system}|N_{system}>0}(\mathcal{G}_Y(z)) = \frac{z(\gamma-\beta)\left(1-\left(\frac{z\gamma}{1-(1-\beta)z}\right)^K\right)}{(z(1-\beta+\gamma)-1)(1-(\gamma/\beta)^K)} \tag{A.2}$$

$$E\left[W^f\right] = \mathcal{G}'_{W^f}(1) = \frac{1}{\beta}\left(\frac{1}{1-\gamma/\beta} - \frac{K\left(\gamma/\beta\right)^K}{1-\left(\gamma/\beta\right)^K}\right) \tag{A.3}$$

This is given as equation (3.22) in the thesis.

$$\mathcal{G}_{N_{system}-1|N_{system}>0}(z) = \sum_{n=1}^{K} z^{n-1} \cdot \frac{\pi_n}{1-\pi_0} = \frac{(\gamma-\beta)\left(1-(z\gamma/\beta)^K\right)}{(z\gamma-\beta)(1-(\gamma/\beta)^K)} \tag{A.4}$$

$$\mathcal{G}_{W^f_{queue}}(z) = \mathcal{G}_{N_{system}-1|N_{system}>0}(\mathcal{G}_Y(z)) = \frac{(\gamma-\beta)\left(1-\left(\frac{z\gamma}{1-(1-\beta)z}\right)^K\right)}{\left(\frac{z\beta\gamma}{1-(1-\beta)z}-\beta\right)(1-(\gamma/\beta)^K)} \tag{A.5}$$

$$E\left[W^f_{queue}\right] = \mathcal{G}'_{W^f_{queue}}(1) = \frac{1}{\beta}\left(\frac{\gamma/\beta}{1-\gamma/\beta} - \frac{K\left(\gamma/\beta\right)^K}{1-\left(\gamma/\beta\right)^K}\right) \tag{A.6}$$

This is given as equation (3.20) in the thesis.

Bibliography

Altendorfer, K., Jodlbauer, H., 2011. An analytical model for service level and tardiness in a single machine MTO production system. International Journal of Production Research 49 (7), 1827–1850.

Altendorfer, K., Minner, S., 2015. Influence of order acceptance policies on optimal capacity investment with stochastic customer required lead times. European Journal of Operational Research 243 (2), 555–565.

Altıok, T., 1985. Production lines with phase-type operation and repair times and finite buffers. International Journal of Production Research 23 (3), 489–498.

Artalejo, J. R., Economou, A., Gómez-Corral, A., 2008. Algorithmic analysis of the Geo/Geo/c retrial queue. European Journal of Operational Research 189 (3), 1042–1056.

Ata, B., 2006. Dynamic control of a multiclass queue with thin arrival streams. Operations Research 54 (5), 876–892.

Ata, B., Olsen, T. L., 2009. Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs. Operations Research 57 (3), 753–768.

Ata, B., Olsen, T. L., 2013. Congestion-based leadtime quotation and pricing for revenue maximization with heterogeneous customers. Queueing Systems 73 (1), 35–78.

Baker, K. R., Bertrand, J., 1982. A dynamic priority rule for scheduling against due-dates. Journal of Operations Management 3 (1), 37–42.

Barman, S., 1998. The impact of priority rule combinations on lateness and tardiness. IIE Transactions 30 (5), 495–504.

Battini, D., Delorme, X., Dolgui, A., Persona, A., Sgarbossa, F., 2016. Ergonomics in assembly line balancing based on energy expenditure: A multi-objective model. International Journal of Production Research 54 (3), 824–845.

Battini, D., Faccio, M., Persona, A., Sgarbossa, F., 2011. New methodological framework to improve productivity and ergonomics in assembly system design. International Journal of Industrial Ergonomics 41 (1), 30–42.

Bendoly, E., Donohue, K., Schultz, K. L., 2006. Behavior in operations management: Assessing recent findings and revisiting old assumptions. Journal of Operations Management 24 (6), 737–752.

Berman, O., 1982. Efficiency and production rate of a transfer line with two machines and a finite storage buffer. European Journal of Operational Research 9 (3), 295–308.

Blumenfeld, D. E., 1990. A simple formula for estimating throughput of serial production lines with variable processing times and limited buffer capacity. International Journal of Production Research 28 (6), 1163–1182.

Boudreau, J., Hopp, W., McClain, J. O., Thomas, L. J., 2003. On the interface between operations and human resources management. Manufacturing & Service Operations Management 5 (3), 179–202.

Bramson, M., 2001. Stability of earliest-due-date, first-served queueing networks. Queueing Systems 39 (1), 79–102.

Buzacott, J., Kostelski, D., 1987. Matrix-geometric and recursive algorithm solution of a two-stage unreliable how line. IIE Transactions 19 (4), 429–438.

Carnahan, B. J., Norman, B. A., Redfern, M. S., 2001. Incorporating physical demand criteria into assembly line balancing. IIE Transactions 33 (10), 875–887.

Çelik, S., Maglaras, C., 2008. Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. Management Science 54 (6), 1689–1699.

Charnsirisakskul, K., Griffin, P. M., Keskinocak, P., 2006. Pricing and scheduling decisions with leadtime flexibility. European Journal of Operational Research 171 (1), 153–169.

Chatterjee, S., Slotnick, S. A., Sobel, M. J., 2002. Delivery guarantees and the interdependence of marketing and operations. Production and Operations Management 11 (3), 393–410.

Commault, C., Semery, A., 1990. Taking into account delays in buffers for analytical performance evaluation of transfer lines. IIE Transactions 22 (2), 133–142.

Conway, R., Maxwell, W., McClain, J. O., Thomas, L. J., 1988. The role of work-in-process inventory in serial production lines. Operations Research 36 (2), 229–241.

Cross, R. G., Dixit, A., 2005. Customer-centric pricing: The surprising secret for profitability. Business Horizons 48 (6), 483–491.

Dallery, Y., Gershwin, S. B., 1992. Manufacturing flow line systems: A review of models and analytical results. Queueing Systems 12 (1-2), 3–94.

Davis, L., 1965. Pacing effects on manned assembly lines. International Journal of Production Research 4 (3), 171–184.

Defregger, F., Kuhn, H., 2007. Revenue management for a make-to-order company with limited inventory capacity. OR Spectrum 29 (1), 137–156.

Dode, P., Greig, M., Zolfaghari, S., Neumann, W. P., 2016. Integrating human factors into discrete event simulation: A proactive approach to simultaneously design for system performance and employees' well being. International Journal of Production Research 54 (10), 3105–3117.

Doerr, K. H., Arreola-Risa, A., 2000. A worker-based approach for modeling variability in task completion times. IIE Transactions 32 (7), 625–636.

Doerr, K. H., Mitchell, T. R., Klastorin, T. D., Brown, K. A., 1996. Impact of material flow policies and goals on job outcomes. Journal of Applied Psychology 81 (2), 142–152.

Doytchinov, B., Lehoczky, J., Shreve, S., 2001. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. Annals of Applied Probability 11 (2), 332–378.

Dudley, N., 1963. Work-time distributions. International Journal of Production Research 2 (2), 137–144.

Duenyas, I., 1995. Single facility due date setting with multiple customer classes. Management Science 41 (4), 608–619.

Duenyas, I., Hopp, W. J., 1995. Quoting customer lead times. Management Science 41 (1), 43–57.

Easton, F. F., Moodie, D. R., 1999. Pricing and lead time decisions for make-to-order firms with contingent orders. European Journal of Operational Research 116 (2), 305–318.

Ebben, M., Hans, E., Weghuis, F. O., 2005. Workload based order acceptance in job shop environments. OR Spectrum 27 (1), 107–122.

Edie, L. C., 1954. Traffic delays at toll booths. Journal of the Operations Research Society of America 2 (2), 107–138.

El ahrache, K., Imbeau, D., Farbos, B., 2006. Percentile values for determining maximum endurance times for static muscular work. International Journal of Industrial Ergonomics 36 (2), 99–108.

Enoka, R. M., Duchateau, J., 2008. Muscle fatigue: What, why and how it influences muscle function. The Journal of Physiology 586 (1), 11–23.

Falk, A., Ichino, A., 2006. Clean evidence on peer effects. Journal of Labor Economics 24 (1), 39–57.

Folgado, R., Pecas, P., Henriques, E., 2015. Mapping workers' performance to analyse workers heterogeneity under different workflow policies. Journal of Manufacturing Systems 36, 27–34.

Franks, I., Sury, R., 1966. The performance of operators in conveyor-paced work. International Journal of Production Research 5 (2), 97–112.

Germs, R., Van Foreest, N. D., 2011. Admission policies for the customized stochastic lot scheduling problem with strict due-dates. European Journal of Operational Research 213 (2), 375–383.

Goldberg, H. M., 1977. Analysis of the earliest due date scheduling rule in queueing systems. Mathematics of Operations Research 2 (2), 145–154.

Goldberg, H. M., 1980. Jackson's conjecture on earliest due date scheduling. Mathematics of Operations Research 5 (3), 460–466.

Gould, E. D., Winter, E., 2009. Interactions between workers and the technology of production: Evidence from professional baseball. The Review of Economics and Statistics 91 (1), 188–200.

Gross, D., Shortie, J. F., Thompson, J. M., Harris, C. M., 2008. Fundamentals of Queueing Theory. John Wiley & Sons, Inc.

Guhlich, H., Fleischmann, M., Stolletz, R., 2015. Revenue management approach to due date quoting and scheduling in an assemble-to-order production system. OR Spectrum 37 (4), 951–982.

Heimbach, D.-K. I., Grahl, J., Rothlauf, F., 2012. The effects of state-dependent human behavior on the design of a serial line. Zeitschrift für Betriebswirtschaft 82 (7-8), 745–762.

Hendricks, K. B., 1992. The output processes of serial production lines of exponential machines with finite buffers. Operations Research 40 (6), 1139–1147.

Hertel, G., Kerr, N. L., Messé, L. a., 2000. Motivation gains in performance groups: paradigmatic and theoretical developments on the Köhler effect. Journal of Personality and Social Psychology 79 (4), 580–601.

Hillier, F. S., Boling, R. W., 1966. The effect of some design factors on efficiency of production lines with variable operation times. Journal of Industrial Engineering 17 (12), 651–658.

Hillier, F. S., Boling, R. W., 1967. Finite queues in series with exponential or erlang service times—a numerical approach. Operations Research 15 (2), 286–303.

Hillier, F. S., Boling, R. W., 1977. Toward characterizing the optimal allocation of work in production line systems with variable operation times. Advances in Operations Research, 109–119.

Hillier, F. S., Boling, R. W., 1979. On the optimal allocation of work in symmetrically unbalanced production line systems with variable operation times. Management Science 25 (8), 721–728.

Hillier, F. S., So, K. C., 1995. On the optimal design of tandem queueing systems with finite buffers. Queueing Systems 21 (3-4), 245–266.

Hillier, F. S., So, K. C., 1996. On the robustness of the bowl phenomenon. European Journal of Operational Research 89 (3), 496–515.

Hillier, F. S., So, K. C., Boling, R. W., 1993. Notes: Toward characterizing the optimal allocation of storage space in production line systems with variable processing times. Management Science 39 (1), 126–133.

Hillier, M., 2013. Designing unpaced production lines to optimize throughput and work-in-process inventory. IIE Transactions 45 (5), 516–527.

Hillier, M. S., Hillier, F. S., 2006. Simultaneous optimization of work and buffer space in unpaced production lines with random processing times. IIE Transactions 38 (1), 39–51.

Hopp, W. J., Spearman, M. L., 2001. Factory Physics: Foundations of Manufacturing Management. Irwin/McGraw-Hill.

Hudson, S., McNamara, T., Shaaban, S., 2015. Unbalanced lines: Where are we now? International Journal of Production Research 53 (6), 1895–1911.

Hunt, G. C., 1956. Sequential arrays of waiting lines. Operations Research 4 (6), 674–683.

Jaber, M., Neumann, W., 2010. Modelling worker fatigue and recovery in dual-resource constrained systems. Computers & Industrial Engineering 59 (1), 75–84.

Jackson, J. R., 1961. Queues with dynamic priority discipline. Management Science 8 (1), 18–34.

Jayamohan, M., Rajendran, C., 2000. New dispatching rules for shop scheduling: A step forward. International Journal of Production Research 38 (3), 563–586.

Johnson, N. L., 1949. Systems of frequency curves generated by methods of translation. Biometrika 36 (1/2), 149–176.

Jolai, F., Asadzadeh, S., Taghizadeh, M., 2008. Performance estimation of an email contact center by a finite source discrete time Geo/Geo/1 queue with disasters. Computers & Industrial Engineering 55 (3), 543–556.

Kc, D. S., Terwiesch, C., 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. Management Science 55 (9), 1486–1498.

Keskinocak, P., Tayur, S., 2004. Due date management policies. In: Simchi-Levi, D., Wu, S., Shen, Z.-J. (Eds.), Handbook of Quantitative Supply Chain Analysis. Vol. 74 of International Series in Operations Research & Management Science. Springer US, pp. 485–554.

Kleinrock, L., Finkelstein, R. P., 1967. Time dependent priority queues. Operations Research 15 (1), 104–116.

Knott, A., 1970. The inefficiency of a series of work stations—a simple formula. The International Journal of Production Research 8 (2), 109–119.

Knott, K., Sury, R. J., 1987. A study of work-time distributions on unpaced tasks. IIE Transactions 19 (1), 50–55.

Kruk, Ł., Lehoczky, J., Ramanan, K., Shreve, S., 2011. Heavy traffic analysis for EDF queues with reneging. The Annals of Applied Probability 21 (2), 484–545.

Lagershausen, S., Tan, B., 2015. On the exact inter-departure, inter-start, and cycle time distribution of closed queueing networks subject to blocking. IIE Transactions 47 (7), 673–692.

Lau, H.-S., 1986a. A directly-coupled two-stage unpaced line. IIE Transactions 18 (3), 304–312.

Lau, H.-S., 1986b. The production rate of a two-stage system with stochastic processing times. International Journal of Production Research 24 (2), 401–412.

Littman, M., 1997. Probabilistic propositional planning: Representations and complexity. Proceedings of the National Conference on Artificial Intelligence, 748–754.

Maglaras, C., Van Mieghem, J. A., 2005. Queueing systems with leadtime constraints: A fluid-model approach for admission and sequencing control. European Journal of Operational Research 167 (1), 179–207.

Mas, A., Moretti, E., 2009. Peers at work. The American Economic Review 99 (1), 112–145.

Moreira, M. R. A., Alves, R. A. F., 2009. A methodology for planning and controlling workload in a job-shop: A four-way decision-making problem. International Journal of Production Research 47 (10), 2805–2821.

Murrell, K., 1961. Operator variability and its industrial consequences. International Journal of Production Research 1 (3), 39–55.

Murrell, K., 1972. Laboratory studies of repetitive work 1: Paced work and its relationship to unpaced work. International Journal of Production Research 10 (3), 169–185.

Muth, E. J., 1973. The production rate of a series of work stations with variable service times. International Journal of Production Research 11 (2), 155–169.

Muth, E. J., Alkaff, A., 1987. The bowl phenomenon revisited. International Journal of Production Research 25 (2), 161–173.

Nelson, R., 1995. Probability, Stochastic Processes, and Queueing Theory. Springer New York.

Neumann, W. P., Medbo, P., 2016. Simulating operator learning during production ramp-up in parallel vs. serial flow production. International Journal of Production Research, 1–13.
URL http://dx.doi.org/10.1080/00207543.2016.1217362

Öner Közen, M., Minner, S., 2016a. Designing unpaced production lines with human operators - the bowl phenomenon revisited. Working Paper, TUM School of Management, Technische Universität München.

Öner Közen, M., Minner, S., 2016b. Dynamic pricing, leadtime quotation and duedate based priority dispatching. Working Paper, TUM School of Management, Technische Universität München.

Öner Közen, M., Minner, S., 2016c. Impact of priority sequencing decisions on on-time probability and expected tardiness of orders in MTO production systems with external due-dates. Working Paper, TUM School of Management, Technische Universität München.

Öner Közen, M., Minner, S., Steinthaler, F., 2016. Effciency of paced and unpaced assembly lines under consideration of worker variability - a simulation study. Working Paper, TUM School of Management, Technische Universität München.

Otto, A., Scholl, A., 2011. Incorporating ergonomic risks into assembly line balancing. European Journal of Operational Research 212 (2), 277–286.

Palaka, K., Erlebacher, S., Kropp, D. H., 1998. Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. IIE Transactions 30 (2), 151–163.

Papadopoulos, C. T., O'Kelly, M. E., Vidalis, M. J., Spinellis, D., 2009. Analysis and design of discrete part production lines. Springer.

Papadopoulos, H., Heavey, C., 1996. Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines. European Journal of Operational Research 92 (1), 1–27.

Papadopoulos, H. T., 1996. An analytic formula for the mean throughput of K-station production lines with no intermediate buffers. European Journal of Operational Research 91 (3), 481–494.

Pekgün, P., Griffin, P. M., Keskinocak, P., 2008. Coordination of marketing and production for price and leadtime decisions. IIE Transactions 40 (1), 12–30.

Pike, R., Martinj, G., 1994. The bowl phenomenon in unpaced lines. International Journal of Production Research 32 (3), 483–499.

Plambeck, E. L., 2004. Optimal leadtime differentiation via diffusion approximations. Operations Research 52 (2), 213–228.

Powell, S. G., Schultz, K. L., 2004. Throughput in serial lines with state-dependent behavior. Management Science 50 (8), 1095–1105.

Rajendran, C., Holthaus, O., 1999. A comparative study of dispatching rules in dynamic flowshops and jobshops. European Journal of Operational Research 116 (1), 156–170.

Rao, N. P., 1975a. On the mean production rate of a two-stage production system of the tandem type. International Journal of Production Research 13 (2), 207–217.

Rao, N. P., 1975b. Two-stage production systems with intermediate storage. AIIE Transactions 7 (4), 414–421.

Rao, N. P., 1976. A generalization of the 'bowl phenomenon' in series production systems. The International Journal of Production Research 14 (4), 437–443.

Ray, S., Jewkes, E. M., 2004. Customer lead time management when both demand and price are lead time sensitive. European Journal of Operational Research 153 (3), 769–781.

Rogers, P., Nandi, A., 2007. Judicious order acceptance and order release in make-to-order manufacturing systems. Production Planning and Control 18 (7), 610–625.

Rohmert, W., 1973. Problems of determination of rest allowances part 2: Determining rest allowances in different human tasks. Applied Ergonomics 4 (3), 158–162.

Savaşaneril, S., Griffin, P. M., Keskinocak, P., 2010. Dynamic lead-time quotation for an M/M/1 base-stock inventory queue. Operations research 58 (2), 383–395.

Schleyer, M., Furmans, K., 2007. An analytical method for the calculation of the waiting time distribution of a discrete time G/G/1-queueing system with batch arrivals. OR Spectrum 29 (4), 745–763.

Schultz, K. L., Juran, D. C., Boudreau, J. W., 1999. The effects of low inventory on the development of productivity norms. Management Science 45 (12), 1664–1678.

Schultz, K. L., Juran, D. C., Boudreau, J. W., McClain, J. O., Thomas, L. J., 1998. Modeling and worker motivation in jit production systems. Management Science 44 (12-part-1), 1595–1607.

Schultz, K. L., McClain, J. O., Thomas, L. J., 2003. Overcoming the dark side of worker flexibility. Journal of Operations Management 21 (1), 81–92.

Schultz, K. L., Schoenherr, T., Nembhard, D., 2010. An example and a proposal concerning the correlation of worker processing times in parallel tasks. Management Science 56 (1), 176–191.

Siemsen, E., Balasubramanian, S., Roth, A. V., 2007. Incentives that induce task-related effort, helping, and knowledge sharing in workgroups. Management Science 53 (10), 1533–1550.

Slack, N., 1990. Work time distributions in production system modelling. Oxford Centre for Management Studies.

Slotnick, S. A., 2011a. Optimal and heuristic lead-time quotation for an integrated steel mill with a minimum batch size. European Journal of Operational Research 210 (3), 527–536.

Slotnick, S. A., 2011b. Order acceptance and scheduling: A taxonomy and review. European Journal of Operational Research 212 (1), 1–11.

Slotnick, S. A., 2014. Lead-time quotation when customers are sensitive to reputation. International Journal of Production Research 52 (3), 713–726.

Soo, Y., Nishino, M., Sugi, M., Yokoi, H., Arai, T., Kato, R., Nakamura, T., Ota, J., 2009. Evaluation of frequency band technique in estimating muscle fatigue during dynamic contraction task. In: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. IEEE, pp. 933–938.

Sury, R., 1964. An industrial study of paced and unpaced operator performance in a single stage work task. International Journal of Production Research 3 (2), 91–102.

Sury, R., 1965. The simulation of a paced single stage work task. International Journal of Production Research 4 (2), 125–140.

Sury, R., 1971. Aspects of assembly line balancing. International Journal of Production Research 9 (4), 501–512.

Tang, C. S., 2010. A review of marketing–operations interface models: From coexistence to coordination and collaboration. International Journal of Production Economics 125 (1), 22–40.

Tijms, H. C., 2003. A first course in stochastic models. John Wiley & Sons.

Van Foreest, N. D., Wijngaard, J., Van der Vaart, T., 2010. Scheduling and order acceptance for the customised stochastic lot scheduling problem. International Journal of Production Research 48 (12), 3561–3578.

Watanapa, B., Techanitisawad, A., 2005. Simultaneous price and due date settings for multiple customer classes. European Journal of Operational Research 166 (2), 351–368.

Wein, L. M., 1991. Due-date setting and priority sequencing in a multiclass M/G/1 queue. Management Science 37 (7), 834–850.

Zhao, X., Stecke, K. E., Prasad, A., 2012. Lead time and price quotation mode selection: Uniform or differentiated? Production and Operations Management 21 (1), 177–193.