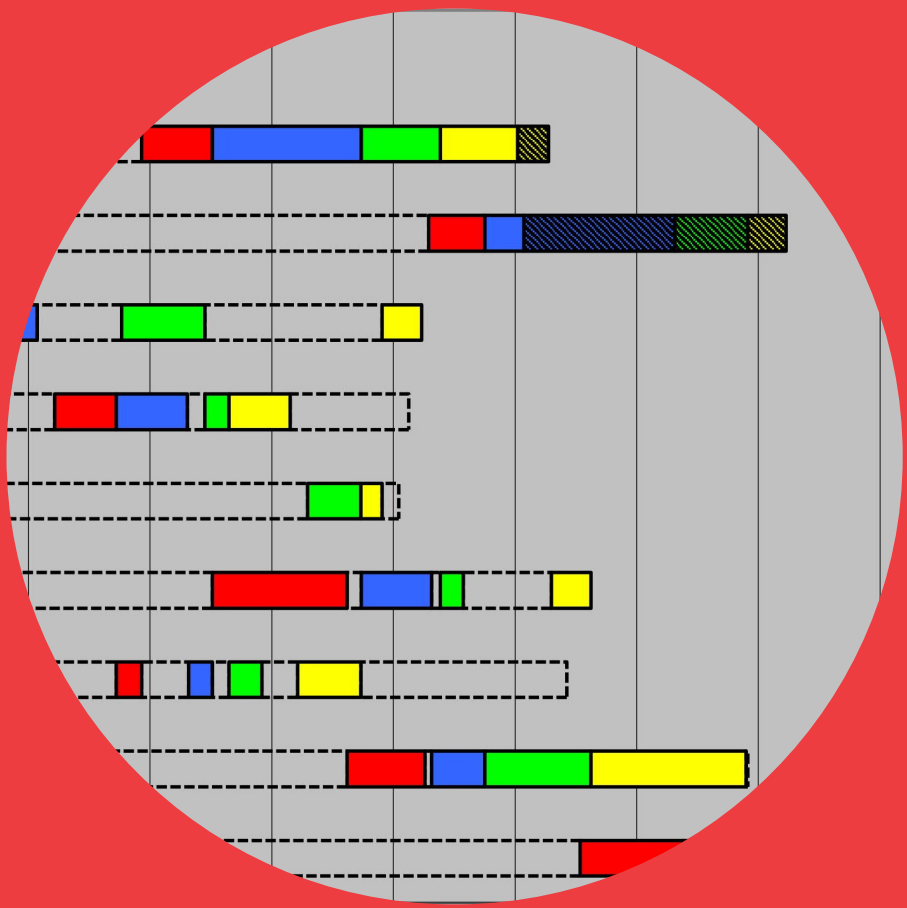
Department of Engineering Design and Production

Online Optimisation Models in Short-term Production Planning

Henri Tokola



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 162/2015

Online Optimisation Models in Short-term Production Planning

Henri Tokola

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Engineering, at a public examination held at the lecture hall E of the university (Otakaari 1, Espoo, Finland) on the 27th of November 2015 at 12 noon.

Aalto University
School of Engineering
Department of Engineering Design and Production
Production Engineering

Supervising professor
Prof. Esko Niemi

Thesis advisor
Prof. Esko Niemi

Preliminary examiners
Prof. Jose L. Martinez Lastra, Tampere University of Technology, Finland
Prof. João Carlos Espíndola Ferreira, Universidade Federal de Santa Catarina, Brazil

Opponent
Prof. Juha Varis, Lappeenranta University of Technology, Finland

Aalto University publication series
DOCTORAL DISSERTATIONS 162/2015

ISBN 978-952-60-6485-7 (printed)
ISBN 978-952-60-6486-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6486-4

Images: 12

Unigrafia Oy
Helsinki 2015

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Henri Tokola

Name of the doctoral dissertation
Online Optimisation Models in Short-term Production Planning

Publisher School of Engineering

Unit Department of Engineering Design and Production

Series Aalto University publication series DOCTORAL DISSERTATIONS 162/2015

Field of research Production Engineering

Manuscript submitted 7 August 2015 **Date of the defence** 27 November 2015

Permission to publish granted (date) 8 October 2015 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

In today's manufacturing world, real-time information is available or soon will be. Manufacturing companies can use it as a part of their information systems (including e.g. material requirements planning (MRP), enterprise resource planning (ERP), and manufacturing execution systems (MES)) for controlling the production system, i.e. for the adjustment of their inventory level, defining the capacity, and even scheduling the starting times of the jobs. A method that can be used to improve the decisions dynamically is online optimisation. Online optimisation repeatedly optimises and adjusts the decisions when there are changes or disturbances in the system. Online optimisation differs from traditional optimisation. First, the data is uncertain but, luckily, the uncertainty might decrease as time goes by. Second, quick decisions are needed but they should be made carefully as the decisions will also affect future decisions.

This thesis studies online optimisation from the production planning perspective. In practice, many real factors, such as rescheduling intervals and material deliveries, affect online optimisation. The six publications of the thesis focus, first, on finding different short-term planning problems in manufacturing companies, second, on worker reactive coordination in parallel stations, third, on periodical rescheduling, fourth, on the effect of disturbances on assemble-to-order systems, fifth, on production line rescheduling, and, sixth, on the erection of the hull in shipbuilding in the event of material delays. The methods that are used in the publications of the thesis include survey, Markov models, simulation, mixed-integer-linear programming, and stochastic modelling.

The modelling of online solutions to practical short-term planning problems is complex because of the large number of variables, most of which have to be considered. As the variables cannot be aggregated in the short term, a single variation in a variable, even a small one, can have significant consequences for the system. The thesis shows that online optimisation gives an advantage in certain short-term situations, such as in the cases of rush jobs or delays of components.

Keywords Optimisation, Production planning

ISBN (printed) 978-952-60-6485-7 **ISBN (pdf)** 978-952-60-6486-4

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2015

Pages 140 **urn** http://urn.fi/URN:ISBN:978-952-60-6486-4

Tekijä
Henri Tokola

Väitöskirjan nimi
Online-optimointimallit lyhyen aikavälin tuotannonohjauksessa

Julkaisija Insinööritieteiden korkeakoulu

Yksikkö Koneenrakennustekniikka

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 162/2015

Tutkimusala Tuotantotekniikka

Käsikirjoituksen pvm 07.08.2015 **Väitöspäivä** 27.11.2015

Julkaisuluvan myöntämispäivä 08.10.2015 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Nykyaikaisessa tuotannossa saatavilla oleva reaaliaikainen tieto lisääntyy jatkuvasti. Valmistava teollisuus voisi käyttää reaaliaikaista tietoa tietojärjestelmissään (joita ovat mm. materiaalinhallintajärjestelmät, toiminnanohjausjärjestelmät ja tuotannon-ohjausohjelmistot) ohjaamaan tuotantoa esimerkiksi säätämällä varastotasoja, määrittelemällä kapasiteetit tai jopa aikatauluttamalla töiden aloitusajat. Yksi menetelmä, jolla reaaliaikaista tietoa voi käyttää dynaamisesti päätöksenteossa, on online-optimointi. Online-optimoinnissa toistuvasti optimoidaan ja säädetään päätöksiä silloin kun järjestelmään tulee muutoksia. Online-optimointi eroaa perinteisestä optimoinnista monin tavoin. Online-optimoinnissa tieto on epävarmaa, mutta epävarmuus voi vähetä kun aika menee eteenpäin. Lisäksi online-optimoinnissa tarvitaan nopeita päätöksiä, jotka täytyy tehdä huolellisesti, koska ne vaikuttavat usein tulevaisuuden päätöksiin.

Väitöskirja tutkii online-optimointia tuotannonohjauksen näkökulmasta. Käytännössä monet tekijät kuten aikataulutuksen aikavälit ja materiaalin saapumiset vaikuttavat online-optimointiin. Työn kuusi julkaisua keskittyvät eri yritysten tuotannonohjauksen ongelmiin, työvoiman reaktiiviseen ohjaukseen, periodiseen uudelleen-aikataulutukseen, häiriöiden vaikuttavuuteen tilausohjautuvassa kokoonpanossa, tuotantolinjan uudelleenaikataulutukseen ja laivanrakennuksen rungonkoonnin aikataulutukseen. Työssä käytettyjä menetelmiä ovat haastattelututkimus, Markov-mallit, simulointi, kokonaisluku-optimointi ja stokastinen mallintaminen.

Online-ratkaisujen mallintaminen käytännön lyhyen aikavälin tuotannonohjauksen ongelmiin on haasteellista johtuen isosta määrästä muuttujia, joista useimmat pitää ottaa huomioon. Koska lyhyellä aikavälillä muuttujia ei usein voi summata, yksittäiset muutokset muuttujissa voivat aiheuttaa merkittäviä seurauksia. Työ näyttää kuitenkin, että online-optimoinnilla voidaan saavuttaa etua tietyissä lyhyen aikavälin tilanteissa, kuten kiiretilauksissa tai komponenttipuutteissa.

Avainsanat Optimointi, Tuotannonohjaus

ISBN (painettu) 978-952-60-6485-7 **ISBN (pdf)** 978-952-60-6486-4

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2015

Sivumäärä 140 **urn** http://urn.fi/URN:ISBN:978-952-60-6486-4

Preface

This work was carried out at the Production Engineering Laboratory at the Department of Engineering Design and Production at the Aalto University School of Engineering. First, I wish to thank my instructor, Professor Esko Niemi, for his support and giving me the opportunity to do my postgraduate studies. I am grateful to him for showing me the worlds of manufacturing and optimisation. Especially the professor's advice about the use of the short processing time rule in the queue in a restaurant comes to my mind: regardless of common rules of morality, on average it is more efficient if priority is given to the people who are quicker at ordering.

I also want to thank my co-author Lauri Ahlroth; we started our academic careers together and undertook a hard but rewarding industrial project at the beginning of them. I wish to thank Sampsa V. Laakso, who shared a workroom with me, for helping me to gain the mindset of a mechanical engineer. From him I learned that doing things wisely is usually the same as doing things simply. Jaakko Peltokorpi and Heikki Remes were the co-authors of some papers written during my doctoral studies and they gave me practical knowledge of how industry operates. I would also like to thank my colleagues Pekka Kyrenius, Juha Huuki, Andrew Roiko, and Mirko Ruokokoski for our many fruitful discussions. When I told them about the great ideas I had, they typically found several weaknesses in them, sending me either back to ground level or to strengthening my ideas. And to the whole staff of the Production Engineering Laboratory, I express my thanks for the good working atmosphere.

I would also like to thank the various sources that gave me the resources I needed to feed my family. These include the Jenny and Antti Wihuri Foundation, Suomen konepajainsinööriyhdistys, the Auramo Foundation, the FIMECC and the Loura project.

I am thankful for my parents and family for their support during the thesis work. Most of all, I especially wish to thank my wife, Maria Tokola. Together with my children, she has been a good motivator for me to do the research well. My family has also provided me with hands-on experience in online, real-time optimisation. There is variation in the world, especially in the short term and in the details, and it affects things.

Espoo, October 2015

Henri Tokola

Contents

List of publications

The thesis consists of an overview and the following publications, which are referred to in the text by their Roman numerals.

I Tokola, H., Järvenpää, E., Salonen, T., Lanz, M., Koho, M., & Niemi, E. (2015). Shop Floor-Level Control of Manufacturing Companies: An Interview Study in Finland. *Management and Production Engineering Review*, *6*(1), 51-58.

II Peltokorpi, J., Tokola, H., & Niemi, E. (2015). Worker coordination policies in parallel station systems: performance models for a set of jobs and for continuous arrival of jobs. *International Journal of Production Research*, *53*(6), 1625-1641.

III Tokola H. & Niemi E. (2011). Combined Periodical and Reactive Control in Multi-item Production-inventory System. *Proceedings, 44th CIRP Conference on Manufacturing Systems* in Madison, Wisconsin, USA on 1-3 June 2011.

IV Tokola H. & Niemi E. (2012). Robustness of Assemble-to-Order Systems against Unexpected Events. *Proceedings, the 2012 IEEE International Conference on Industrial Engineering and Engineering Management* in Hong Kong, China on 10-13 December 2012.

V Tokola, H., Ahlroth, L., & Niemi, E. (2014). A comparison of rescheduling policies for online flow shops to minimize tardiness. *Engineering Optimisation*, *46*(2), 165-180.

VI Tokola, H., Niemi, E., & Remes, H. (2013). Block Erection Sequencing in Shipbuilding With General Lifting and Joining Times. *Journal of Ship Production and Design*, *29*(2), 49-56. The publication has been selected by SNAME Featured Papers Committee as a Significant Paper of 2013. As such, it was reproduced in the 2013 volume of the SNAME Transactions.

Author's contribution

Publication I: "Shop Floor-Level Control of Manufacturing Companies: An Interview Study in Finland"

The survey presented in the paper was conducted jointly by the first five authors of the paper. The further analysis of the shop-floor-level control issues and writing of the paper was done by the author. Esko Niemi contributed to the writing.

Publication II: "Worker coordination policies in parallel station systems: performance models for a set of jobs and for continuous arrival of jobs"

The author contributed significantly to the ideas of the paper and the writing of the mathematical models. The writing of the other parts of paper and the numerical experiments were done by Jaakko Peltokorpi. Esko Niemi contributed to the writing.

Publication III: "Combined Periodical and Reactive Control in Multi-item Production-inventory System"

Most of the work was done by the author. The idea of the paper came to the author after a visit to a company where the production planning was done periodically. Esko Niemi contributed to the writing.

Publication IV: "Robustness of Assemble-to-Order Systems against Unexpected Events"

The background for the paper was invented by the author during a visit to the University of Minnesota, under the supervision of Professor Saif Benjaafar, who had studied assemble-to-order systems earlier. The work was done by the author, but Esko Niemi contributed to the writing of the paper.

Publication V: "A Comparison of Rescheduling Policies for Online Flow Shops to Minimise Tardiness"

Most of the work was done by the author. The initial idea of studying online flow shops came from Professor Esko Niemi. The program used to generate the results was implemented together by Lauri Ahlroth and the author. Professor Esko Niemi contributed to the writing of the paper.

Publication VI: "Block erection sequencing in shipbuilding with general lifting and joining times"

The idea of studying block erection sequencing came after discussions with Esko Niemi, Heikki Remes, and personnel from a Finnish shipyard. After the idea, most of the work was done by the author. Esko Niemi and Heikki Remes contributed to the details and writing of the paper.

List of abbreviations

MRP	Material Requirements Planning
ERP	Enterprise Resource Planning
MES	Manufacturing Execution System
LP	Linear Programming
MILP	Mixed-Integer Linear Programming
WIP	Work in Process
MTO	Make to Order
MTS	Make to Stock
ETO	Engineer to Order
ATO	Assemble to Order
TPS	Toyota Production System
JIT	Just in Time
TOC	Theory of Constraints
CONWIP	Constant Work in Process
CV	Coefficient of Variation
CT	Cycle Time
FIFO	First In First Out
SPT	Shortest Processing Time First
EDD	Earliest Due Date First
EOQ	Economical Order Quantity
LS	List Scheduling

1. Introduction

"It is a bad plan that admits of no modification."
Publilius Syrus, Maxims (100 BC) [1]

It is easy to understand that the above quote is true of long-term planning, in which forecasts are said to be always wrong. But in the short term too, during the daily life of the factory floor, modifications to plans are so natural that it is hard to even think of them as modifications. Short-term plans have to be modified online, during the execution, e.g. as a result of the disruptions caused by breakdowns, late materials, or new, urgent jobs. In the event of these disruptions, it would be interesting to know the ways in which the plans can be modified and what the optimal solution is. Part of this knowledge is in your hands.

1.1 Background

In today's manufacturing industry, ***optimisation*** is a tool that has to be used in order to be competitive. In a mathematical sense, optimisation means that the problem is constructed by defining variables and their constraints, and it is solved in such a way as to get the best possible solution, which is also proved to be the best, optimal one. In production, optimisation is usually used in the decision-making phase in long-term production planning, which typically solves investment problems such as the location of the factory or the number of employees.

Online optimisation *problems* are defined as those problems where the optimisation and decisions based on it have to be performed before all the data is available. The data is later revealed in a certain order or at specific times [2] and some decisions can be changed further. Decisions that were made earlier may be hard to change and thus they can have critical effects on future decisions. This all means that the problem evolves over time. To deal with this kind of problem, online optimisation [3], stochastic programming [4], and dynamic process control [5] have been studied earlier in the scientific literature. The characteristics of the problem include the fact that data is uncertain, knowledge increases as time goes on, current decisions affect future decisions, and, at the same time, quick decisions are needed. The solutions differ, depending on how the above characteristics are weighted.

Therefore, in this thesis, online optimisation problems that appear in ***short-term production planning*** are studied. Short-term planning handles daily

operations covering the control and scheduling of processes, jobs and inventories. Contrary to long-term planning, short-term planning has to handle exceptions resulting from machines, workers, subcontractors, and customers. These exceptions can occur in many phases, including order handling, production, and shipment. The precise forecasting of the above matters is impossible. Therefore, short-term planning is a field where online optimisation using real-time information is very promising; the information is indeed real-time and more information will be available later, and the problems cannot be solved beforehand.

Although the research on optimisation has been quite active in recent decades, optimisation methods are rarely used in practice for short-term planning. The main problems appear to be the dynamic situation and the high number of variables involved. The practice appears to lean on common sense, on best practices, and on simplified, locally determined rules. Although simple rules can be effective in terms of their speed of operation and application, there is still room for further improvements. The current advances in information technology, including e.g. ERP (enterprise resource planning) and MES (manufacturing execution system) systems, have enabled optimisation to be used in more complex systems. It is also possible to optimise the systems faster, by using even real-time information, and ultimately solving the problem at the moment it appears.

1.2 Optimisation and online problems

Optimisation means finding the best possible, optimal solution. However, there are several different optimisation methods that can be used to find out optimal or near-optimal solutions.

1.2.1 Optimisation methods

In the literature, optimisation is usually understood as mathematical optimisation, but it can also mean heuristics, and in practice even choosing the best priority rule. In some cases, the best possible solution is not even needed. It is enough to have a feasible solution. But even finding a feasible solution requires optimisation-related techniques.

Mathematical optimisation aims at getting the optimal solution for a problem. In mathematical optimisation, a mathematical model is formed from variables, objective functions, and constraints, and solving it determines the best possible solution, where the objective is either minimal or maximal. ***Linear programming*** (LP) is a special case of such problems. In it, the objective function and constraints are linear and the variables are real numbers. Mathematical optimisation as a form of LP was first invented by Kantotovich in 1939 [6] to optimise various organisational problems. Dantzig [7] later extended the results and made LP a common tool. A general LP problem is defined as follows:

$$min\ \mathbf{c}x,$$

$$\mathbf{A}x <= b$$

$$x >= 0,\ x \in R$$

In the problem x denotes a vector of variables, c is the coefficient vector, and both the A matrix and b vector define constraints. A computer can be used to find a solution to the LP problem efficiently. The Simplex method, first described by Dantzig [8], is a well-known tool for calculating the optimal solution for any LP problem. Although it usually solves the problems efficiently, ellipsoid algorithms [9] and other interior point methods can be used to solve the problem in polynomial time, even in the worst case. This means that the problem is always easy and that even a large LP problem can be solved fast by a computer.

The LP problem requires linear constraints between variables; however, in many practical problems, non-linear constraints are needed. One common approach to dealing with non-linear constraints is to use binary or integer variables, as is done if the problem is formulated as a ***Mixed-integer linear programming*** **(MILP)** model. The difference from the LP model shown above is that some of the variables x are integers in the MILP model.

In practice, all the problems that appear in production can be modelled with a certain level of precision as MILP problems using piecewise linear functions. Every non-linear function can be approximated as a piecewise function of linear functions, which makes the use of MILP applicable to any real problem that does not need extreme accuracy. MILP problems are solved optimally, e.g. by using branch and bound techniques [10], but as the problem in general is NP (Non-deterministic polynomial-time) hard, the solving time grows exponentially when the number of variables and constraints increases. Because of this, MILP can be used to solve only problems with a small number of variables and constraints. However, MILP is a valuable tool, and when the problem is small but complex, it is a method that finds the optimal solution.

In general, the best possible solution is not needed; it is enough to have a solution that is good. In that case hard problems, such as a large MILP model, can be solved using ***heuristics***. A heuristic algorithm just tries to find a good solution and it does not guarantee the optimal solution as pure mathematical optimisation does. Typical heuristics are based on a local search. A local search tests small local modifications to the current solution and makes the modifications that improve the objective function of the solution. A problem with a local search is that it sticks to the local optimum, a solution that is the best in the neighbourhood of the solution, but not the best one for the whole problem.

The most commonly used heuristic methods are tabu search [11, 12, 13], simulated annealing [14, 15] and genetic algorithms [16]. They are all ***metaheuristics***, which means that they typically lie over the other, problem-specific heuristics and are independent of them. In tabu search, the solution cannot return to the values that have recently been visited. Simulated annealing

makes it possible sometimes to move to a worse solution. In the genetic algorithm, different solutions are generated and combined. Although these heuristics give good solutions fast, the fact is that they also require problem-specific implementation effort and fine tuning. Simulation annealing is the most general approach to the problems and it has been shown to find the optimal solution for all MILP problems, at least with certain parameters and given infinite time [17].

In practice, one can also rely on simple, myopic solutions, such as ***priorities***. In those, the solution is usually constructed in a straightforward manner by selecting the best-looking candidate first, then repeating the selection to pick the second and continuing this until all the candidates are ordered. It should be noted that in many cases the solution can also be optimal or near-optimal. Although myopic solutions might not give the best solution, their major advantage is their easy and understandable implementation.

1.2.2 Characteristics of online problems

Contrary to the classical optimisation problems, online optimisation problems evolve, even in real time. One gets more data later or the current data may change, but the solution for the current stage is needed immediately. Later, when new data arrives, it may be possible to update the solution. The natural characteristics of online problems are:

1. incomplete and uncertain data;
2. knowledge increases as time passes;
3. the constraints on the problem are usually caused by earlier decisions, and
4. quick decisions are needed because of the real-time nature of the problem.

First, in online problems the data is incomplete or it has uncertainty. To take into account the knowledge subject to uncertainty, stochastic models including e.g. probability distribution have to be understood and used. Systems where the data is uncertain are generally difficult to understand [18]. From the scientific perspective, this uncertainty is often seen as the main challenge [19]. Uncertainty can also be seen as an advantage, as the problems can be easier to handle as a result of the low amount of information.

The second characteristic is that knowledge increases when time passes. Simply, this happens because the outcome of uncertain things is revealed. When knowledge increases, it also means that the previous optimal solution may not be optimal any more. Even knowing that something has not happened yet is knowledge. For example, in some cases waiting for a new job, even when there is work available at the moment, might be profitable [20]. Forecasts are a way to increase knowledge without waiting. Forecasts can be seen as assumptions about the future, as is the case e.g. when linear extrapolation is used. A forecast can also be stochastic when distribution models are used.

Third, online problems have constraints and penalties that are typically evolving. The constraints depend on the earlier solution or on the time. As an example, it is generally typical that a job that has been started is not easy to stop.

Freezing of schedules is a practical method that can be used with uncertain planning. It fixes the schedule for a certain time period, during which the schedule cannot be modified.

The fourth characteristic of online problems is that quick decisions are needed. How quick depends on the application being considered [21], but the practical reality is that everyone needs the solution immediately, not later. This characteristic forces the problems to be analysed beforehand and a method to obtain a solution should be at hand. It is not realistic to prepare a mathematical model on demand. So if an online optimisation solution is used, some kind of standardised decision system is needed. The problems that deal with e.g. computational complexity have to be solved and the information flow has to be fast.

Although quick solutions are required, ***Computational complexity***, which makes problems hard to solve, is not a big problem from the production perspective. If the solution is hard, there are usually simple heuristics that can find a good, non-optimal solution to the problem in reasonable time. If the heuristic is simple, it makes the solution easy to understand, implement, and use. This can be one reason why simple solutions, such as priority rules, are typically used in practice. ***A real-time optimisation problem*** can be thought of as an online problem in which the speed of decision is weighed [19]. But compared to traditional real-time optimisation, in this thesis it is assumed that there is time for the computation. Therefore, the most crucial problem is just uncertainty or the lack of information.

1.3 Problem statement and objectives

While focusing on online optimisation and short-term production planning, this thesis covers the following research question(s):

> ***"Why, where, and how should reactive online optimisation be used in short-term production planning with problems such as rush jobs and delayed component deliveries?"***

This overview and the six publications included in the thesis answer this question as follows.

Why should online optimisation be used? As stated above, the main driver for the use is that real-time information will soon become available in production environments. Online optimisation can use the information in order to optimise the system repeatedly and at least partly automatically.

Where should online optimisation be used? This question is answered in Publications I, IV and VI. Publication I contains the results of a survey where the control problems of different companies were discussed. Publication IV studies a general assemble-to-order problem from the online perspective, whereas Publication VI studies the optimisation of one-of-a-kind production. In general, online optimisation should especially be used in those cases where there is a need to modify the existing plans, e.g. as a result of breakdowns of machines.

How should online optimisation be used? Online optimisation can be used to perform production planning in real time. It can give the best possible plan even in cases where there are disturbances or significant uncertainty. Publications II, III, V study different production planning problems. The different online solutions are compared using online optimisation and the results give suggestions about their usefulness against the different disruptions.

1.4 Initial hypotheses

The initial hypotheses of the work are that online optimisation is a way to increase the use of information, which is growing and growing. It is also initially thought that short-term production planning is a fruitful area for online optimisation because of its nature. Online optimisation can especially bring benefits in short-term production planning in the event of disruptions. Thus, in order to validate the value of online optimisation, different production problems have to be collected, modelled and optimised. This can be done using surveys, mathematical modelling, numerical experiments and statistical analysis.

1.5 Structure

The remainder of the thesis is organised as follows. Section 2 describes the state of the art concerning practical short-term production planning and online optimisation. First, it goes through different planning periods, different production systems, and different performance criteria. Then it discusses the complexity of short-term planning and gives an overview of short-term planning methods that are used in practice. Another part of Section 2 discusses different approaches to handling online problems. Section 3 studies online optimisation in the case of short-term production planning and discusses the results of all the publications included in this thesis. It describes how the publications study online problems in different production environments. The end of Section 3 also draws some main conclusions from the publications. Finally, overall general findings and ideas for further research are listed in Section 4.

2. Literature and best practices review

In this chapter, first, production planning, production systems and then online problems are discussed. The end of the chapter contains short assessment where the aspects of production planning and online optimisation are discussed together.

2.1 Short-term production planning

Production planning allocates orders and organises resources in such a way that the production is efficient. Here, efficient means that the utilisation of the resources is high, lead times short, and inventories low. The problem is that these targets are often contradictory and one has to find a balance between them. Short-term planning is put on detailed investigation, because it is a natural application area for the topic of this thesis, online optimisation.

2.1.1 Planning hierarchy

Production planning is typically divided hierarchically into three levels, long-term, medium-term, and short-term planning [22, 23]. The different planning levels are shown in Table 1. To sum up the table briefly, long-term planning considers strategic issues, medium-term tactical ones, and short-term daily operational issues.

Table 1: Different planning levels and their typical aggregation levels, decisions, and time frames.

Planning category	*Aggregation level*	*Typical decisions*	*Time frame*
Long-term planning (strategic)	Product families, forecast	Products, factories	Months to years
Medium-term planning (tactical)	Products, known orders, and orders forecast from long-term planning	Capacity (e.g. workers), inventory targets	Days to months
Short-term planning (control)	Single order or job, worker	Scheduling, inventories	Minutes to days

Long-term planning defines the strategy of the whole company: products, factories, equipment, production lines, and logistics systems. A typical objective is to provide capacity for future products and money is the most important issue in terms of profit and costs. The time frame for the decisions in long-term planning is from months to years.

Medium-term planning defines the tactical decisions on the factory level, including the number of workers, inventory accumulation for seasonal products, and transportation issues. A major objective is to use available resources efficiently, and thus resource utilisation is an important issue. The time frame for these decisions is from days to months. The decisions that are made in long-term planning are usually fixed, i.e. they are the constraints for medium-term planning.

Short-term planning schedules individual products, machines, and workers. Most of the decisions resulting from the long- and medium-term planning are fixed and they act as constraints. There are still many things to be decided: the sequence of the production, arrangement of the workers, inventory control, and vehicle scheduling. The objective is typically to provide products in time or as fast as possible, which is in contrast to the high utilisation objective of medium-term planning and to the cost minimisation objective of long-term planning.

In the remainder of this thesis, it is mostly short-term production planning that is considered. It is the area where there is a need for dynamic online solutions: the data relevant to the problems appear at a specific point in time, a solution may be needed at that time, and the current decisions affect future decisions.

2.1.2 Production systems

Short-term production planning controls the processes of material, machines, tools, and workers, which together form the production system. To simplify the number of processes, in this thesis, workers are omitted; it is assumed that there will always be enough workers to operate the machines or use tools. Traditionally, it has been argued, most noticeably by Henry Ford [24], that in general, if multiple items are produced, it is more efficient to move material than workers and machines [25]. A further simplification is that machines and tools are considered together as single processors at a station. An important remaining problem is how one arranges and plans the work at the stations.

Production systems can be categorised by the way of arrangements of the stations. The arrangement can be a station, parallel stations, a flow shop, or a job shop. Examples of these are shown in Figure 1. The smallest structure is a ***station***, which can mean a machine or a workstation. A station produces a single item or batch of items at a time, during which other available products cannot be produced at all and they have to wait. The process at a station can have different characteristics. The process can have a setup, i.e. non-productive time which occurs before the station starts to produce different kinds of items. If the process at the station is pre-emptive, the process of a current job can be

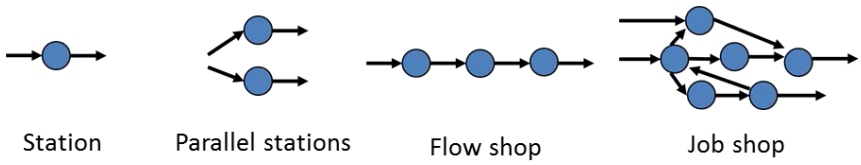


Figure 1. Production systems: Station, Parallel stations, Flow Shop, Job Shop. The arrows describe the flow of items.

stopped and continued again after another job has been processed. There might be multiple similar stations that are called **parallel stations**.

In a ***flow shop***, the stations form a line, and the jobs move sequentially through the shop in the order determined by the line. Compared to single stations, in a flow shop different stations can specialise in different phases of the process. A flow shop is a good way to organise mass production as a result of the easier material delivery and flow of the jobs. The drawback of a production line is the way it handles variation in the processing time. If the queue lengths between the stations are limited, variation in the processing time can lead to the stations becoming starved or blocked. Starving means that a station does not have a job and blocking means that a station cannot move the finished work forward because the next station is still occupied. In automotive industries, blocking is eliminated by stopping the whole assembly line, which is a special type of flow shop, in the event of problems [26]. Stopping the whole line forces engineers, first, to fix the problem quickly and, second, to think how the problem can be totally eliminated in the future.

In a ***job shop***, the jobs in production follow more complex paths than a simple sequential line. Different jobs can use different resources in a different order. Jobs can even return to a station where they have already been earlier. It is easy to see that a job shop is more difficult to understand, plan, and control than a station or a flow shop. Lead times are typically long in a job shop, where one cannot easily predict the availability and workload of the machines well. For example, a rush job may have high priority on a highly utilised machine, which will consequently increase the waiting times of all the jobs that are in the queue for the machine, and of the jobs which will join the queue before the queue is drained empty. However, although a job shop might be complex in terms of production planning, it is a common way to arrange the production because of its ability to maximise the utilisation of the machines in the case of different types of products.

The arrangement of the machines can be selected on the basis of the number of product types and production volumes. Hayes and Wheelwright [27] illustrated this in their product-process matrix. If the production volume is low, single stations are usually enough. If different products must be produced, a job shop may be used, but if the production volume of a product is high, a flow shop in the form of an assembly line should be preferred. On the basis of the author's experience, it is common that production companies see problems in their competitiveness if the above rules are not followed, especially if products that

have high demand are processed at single stations, or if products that have low demand are processed in flow shops together with high-demand products.

After the arrangement of the machines is fixed, products must be produced, but inventories can make the production more efficient in terms of utilisation or they can speed up the lead time that is promised to the customer. Inventories can be in the form of material inventory, work-in-process (WIP) inventory and finished products inventory. Clearly, the more work that is done on the products that are in the inventory, the more valuable the inventory is. To make the use of a bottleneck machine efficient, an inventory can be placed before the machine. For standard products, the finished products inventories can be used to make lead times shorter for the customer, but their costs are higher than the costs of raw material inventories.

Alternative approaches can be used in organising the arrivals of customers' orders in the production system. The extreme ones are ***make-to-order*** (MTO) and ***make-to-stock*** (MTS). In an MTO system, when an order arrives, production of the item in question is started from the beginning of the production. In an MTS system, there is an inventory of end products, and when an order arrives, a product is taken from the inventory. If needed, an inventory replenishment order is forwarded to the production. MTS systems are clearly more efficient in terms of customer lead time, but they do not work well for customised products. If the product has to be designed before production, production can be said to be an ***engineer-to-order*** (ETO) system. ***One-of-a-kind*** projects are typically performed in an ETO system. For a mass customised product, there is also the **assemble-to-order** (ATO) system, where products are assembled from components after being ordered, i.e. the components are made using make-to-stock control but the final products are assembled from components using make-to-order control.

Production systems can also be characterised in terms of how orders are released into production, i.e. whether they are pushed into production according to a schedule, or pulled into production when there is enough free capacity. This is the main difference between ***push-*** and ***pull-type production***. Push production follows the schedule at all stations. The advantages of push production include the fact that the finite schedule enables full utilisation to take place and its schedule is easy to understand. These advantages are especially useful if multiple constrained resources are needed. The main problem in push production is the consequence of delays. As a result of high utilisation even a small disturbance can delay all the jobs that are scheduled for many days ahead. Material requirements planning (MRP), a computer-based scheduling and component ordering software, is a widely used example of push production [28].

Pull-type production takes new orders into production only when there is enough free capacity. In this way, pull production simply restricts overproduction by limiting the work-in-process (WIP). As WIP is limited, the cycle time inside production is also easy to estimate. The problem in pull production is that it is hard to control when the demand, orders, or jobs have a high known variation. On the other hand, pull production balances the

production in the case of unexpected variations, but it does so by idling the resources. This makes it a natural incentive for pull systems to both improve the flexibility of workers and reduce the variation so that the pull system can operate as efficiently as possible.

Pull-type control is an important part of many production philosophies, such as the Toyota production system (TPS) [29], Just-in-Time (JIT) [30], Lean [31], or the Theory of Constraints (TOC) [32]. Pull-type production is used on production lines in the automotive industry, where the pace of production is so fast that overproduction would be catastrophic as the finished products would fill the whole factory. Kanban [33], Constant WIP (CONWIP) [34], and bucket brigade [35] are well-known factory floor implementations of pull production.

2.1.3 Production performance measures

In short-term planning, performance is typically related to utilisation and the lead time. These measures are further affected by work-in-process inventories, quality, and due dates.

The different measures are connected to each other. There is usually a substantial amount of money invested in the resources and therefore the production should be as fully utilised as possible. On the other hand, customers want fast deliveries of their orders, which makes short lead times another target. In an uncertain world, it is not easy to achieve both high utilisation and a short lead time, and one has to either balance between them or focus on one or the other. An easy way to adjust between utilisation and the lead time is inventories in the form of work-in-process (WIP) or batches, but it does not come without costs, as the extra inventory must be paid for by someone and it increases the lead time. However, as a result of the flexibility of inventories, WIP occupies a central role in production planning. Another important daily factor is the quality. It is usually given, but production planning can improve it by shortening the detection time of problems. Bad quality itself also affects the planning of the system dramatically. In some production systems, especially in low-volume production, setups are critical and an increase in the batch size can be used to reduce utilisation and increase the lead time. Figure 2 illustrates the effects of utilisation, quality, lead time, work-in-process (WIP), and the size of batches on each other. These individual relations are discussed below in detail.

The effect of ***utilisation*** on the ***lead time*** is typically the most important relationship in production planning. If the utilisation is high, new orders will have to wait, and thus their lead times are increased by the waiting time. The utilisation percentage can be thought of as being nearly equal to the probability that an order has to wait when it arrives in the system at a random time. On the other hand, an increased lead time permits higher utilisation, as the orders can be queued before processing.

Further, the relationship between utilisation and the lead time can be examined with the help of ***queuing theory***. Queuing theory is based on the work by Erlang [36] and, as the term 'queuing theory' suggests, it gives estimations about the lengths of queues. For exponentially distributed

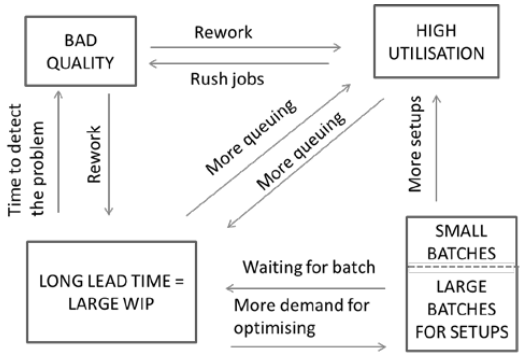


Figure 2. Relations between bad quality, utilisation, lead time, WIP, and batch sizes in short-term planning. An arrow means that an increase in the value of the subject also increases the value of the object. Increases may not be linear. The text above the arrow describes the reason for the increase.

interarrivals and processing times, queuing theory formulae give exact average results in a way that is easy to calculate. With exponentially distributed times, the Markov property holds, i.e. the process is memoryless [37], which means that the current state of the process is not dependent on the past states; only the current state matters. This makes it easy to calculate exact results. For different kinds of queuing systems, different models have been developed. Kendall's notation is typically used to name different models [38]. In a basic queuing model, M/M/1, it is assumed that there is one station and the processing and interarrival times are exponentially distributed.

By using queuing models, it is possible to calculate the average performance results in terms of the utilisation rate of the system. In naive thinking, if one assumes that arrivals take place at fixed intervals and processing times are constant, it is sufficient to have the capacity that is just enough for the processing, i.e. utilisation is 100%. However, when the arrivals and processing times are stochastic, as is the case in reality, one will have long queues in front of the processing station. In this case, the M/M/1 model gives us an estimation of the length of the queue. If the utilisation in the long term is ρ, the average queue length Q in the M/M/1 model is

$$Q = \frac{\rho}{1-\rho} \tag{1}$$

The equation gives us basic information about the effect of utilisation on queuing. When the utilisation of the system, ρ, approaches one, the denominator of the equation approaches zero, and, subsequently, the average queue length approaches infinity. It should be noted that applying the above formula to practice is not straightforward as the capacity can be adjusted by working overtime during periods of peak demand, which reduces queue lengths.

Equation 1 gives us the relationship between the utilisation and the queue length in the case where the processing and arrival times are exponentially distributed. However, in practice, the variation may differ from the exponential distribution. If this is the case, the distributions of the times can be modelled e.g. by using the coefficient of variation, CV, which is the standard deviation

divided by the mean. If the processing time is t and the CVs related to the arrivals and processing times are CV_a and CV_p and the utilisation, ρ, is relatively high, then Kingman's [39] equation (2) can be used as an approximation for the cycle time, which consists of the processing and waiting time. In Kingman's equation the cycle time for a job, CT, is the following.

$$CT = \left(\frac{CV_a^2 + CV_p^2}{2}\right)\frac{\rho}{1-\rho}\, t + t \tag{2}$$

Kingman's equation is an extension of the P-K formula developed by Pollaczek and Khinchin [40][41]. Hopp and Spearman [42] further extended the equation for parallel machines. For a review of queuing models and their applicability to manufacturing, see e.g. [43].

The effect of the lead time on work-in-process is shown by Little's law [44]. It gives the relation between the averages of the throughput rate (TH), cycle time (CT) (which is related to the lead time), and work-in-process (WIP).

$$WIP = TH * CT \tag{3}$$

The main insight from Equation (3) is that when the throughput or the cycle time is increased, then WIP increases as well. In the event that there is a bottleneck which limits the throughput TH, an increase in WIP only increases CT. The law considers only average values, but, on the other hand, it is independent of variation and probability distributions.

Batches are important in terms of utilisation and the lead time as well. Increasing batch sizes in production will increase utilisation as a result of combining multiple setups into one, but on the other hand it increases the lead time [45, p. 83] because waiting for a batch, processing the whole batch, and combining a batch later takes time. Thus, if the utilisation is important, e.g. because of a bottleneck machine, batches should be used. However, in the case of transfers and non-bottleneck machines, the batches should be as small as possible so that all the available capacity is taken into use to reduce the batch sizes and thus shorten the lead times [45, p. 84].

Inventories can be used as buffers to get items immediately when needed. As an example, work-in-process inventories are often used to balance the production system in the event of uncertainty and disruptions. When inventories are considered, important measurements are ***inventory turnover*** and ***service level*** [45, p. 22][45 p. 262]. For a product that has a normally distributed demand with an average of D and standard deviation σ, and that should have a 99.9% service level to cover the demand three-sigma above the mean, the inventory turnover time T depends on the constant replacement lead time L as follows:

$$T = (L + 3\sqrt{L}\sigma)/D \tag{4}$$

This model directly suggests that the cost of inventory turnover (T) correlates closely with the lead time of the items and the cost of the service level (the

second term on the right-hand side of the equation) with the inverse and variation of the demand.

When an individual order is considered and it has a deadline, **tardiness** is an important aspect. A deadline usually comes from the customer, but it is usually agreed by the production. In a general sense, tardiness is related to the production lead time and variation in the production lead time. Especially variations in the production lead time can be critical in terms of the tardiness aspects. Rush jobs, varied jobs, and high utilisation are aspects that can make jobs tardy.

Quality matters in short-term planning as well. From the production planning perspective quality might be just one additional step, the inspection. However, if the quality target is set high, the production control aspect also has an effect. For example, a simple process sequence makes it easier to find flawed products if the problems are e.g. due to systematic problems in the machines [26]. The lead time of the process also has a significant effect on the time between the occurrence of the defect and its detection. This time is sometimes called the information turnaround time [45, p. 171].

The effect of bad quality has a significant impact on utilisation and the lead time. Simply put, as Deming outlines, bad quality means rework [46]. If the scrap percentage from a process is f, then the scrapped products have to be produced again, which increases the seen demand and needed capacity by $1/(1-f)-1$. However, what is more significant is that the reworking will reduce customer satisfaction if the delivery is delayed because of it. Even with low fault probabilities, reworking leads to a situation where the production lead times can be multiples of the promised lead times, as the process has to be repeated.

2.1.4 Complexity of short-term planning

Short-term, daily planning can be considered to be relatively easy when compared to medium- and long-term planning. However, short-term planning is complex because interrelated factors have to be considered in detail. For example, instead of playing with aggregated numbers, each single order, job, and resource has to be considered. Additionally, because of the high number of process steps, there are numerous possible variations and exceptions that may cause problems.

Short-term planning has to consider many ***variables***, and the problem is that these variables are actually different, although they could be considered similarly in medium-term planning. As an example, the demands of the different products commonly follow the Pareto principle, which in production means that around 20% of the products typically account for around 80% of the demand [47]. If this holds true, it might be sufficient for long- and medium-term planning to consider a small share of products, those that account for most of the demand. However, in the short term every product has to be considered and delivered in time. The products are different, and small products tend to have shorter lead times and large products longer ones. If the different products are produced on the same line or machine, which is often the case, the products

with longer processing times cause significant problems for the lead time of the products with shorter processing times.

The variables cannot be aggregated. In many cases, products have to be considered as single pieces, instead of being considered as aggregated numbers by product families, as is often done in medium-term calculations. The effect of small, rush jobs on utilisation might be negligible in medium-term planning, but in short-term planning, their effect is critical, especially if utilisation is high and the system has several bottlenecks. This can be understood by studying the availability of a machine. If the availability of a machine is only 90%, because of breakdowns, for 10% of the year it will be down. If the machine breaks down, it matters for the current job whether the downtime is split into multiple separate days or if the periods of downtime take place in one block.

Problems are typically caused by ***exceptions***, i.e. special cases. Typical exceptions arise from the breakdowns of the machines, material shortcomings, and urgent orders. When a lengthy manufacturing process is repeated day after day, occurrences that have a small probability will occur, or at least something will occur in some part of the production. This is sometimes called Murphy's Law, but the original idea might have come from Augustus De Morgan's publication [48], where he simply states that "whatever can happen will happen if we make enough trials." Risks with low probabilities and a low impact are not relevant in medium-term planning, but in short-term planning they might have consequences even for several days if they occur at a critical moment; for instance, a one-day breakdown of a machine in a year is not large if we consider the whole year, but on the day the breakdown occurs it will delay everything that has been planned to be done by the machine. If there are multiple steps in the production, all the steps are affected by the breakdown. Thus, the way the exceptions are handled is important in short-term planning.

In addition to exceptions, short-term planning has a lot of ***normal variation*** as well. The key thing in normal variation is that while it cannot be accurately forecast, there are certain limits on the variation and it is not completely unknown. This kind of variation is usually encountered when the quality of the process is analysed. Normal variation also appears in the time needed for manual assembly or in the level of daily demand. In the medium and long term, as a result of the law of large numbers, the aggregation of some values tends to be near predictable averages and they can be mistakenly thought of as being constant in the short term as well. But in the short term, they are not constant; they just have a normal variation around a constant average.

Normal variation in the short term can be understood by studying probability laws, as follows. Let us assume that the average daily demand is D and the standard deviation, which describes the variation, in it is σ. If there are no significant trends in the demand, according to the laws of variation, the weekly deviation is $\sqrt{7}\sigma$. For an arbitrary number of time periods X, the coefficient of variation (CV), i.e. the standard deviation divided by the average demand, can be used to describe the relative variation. It is

$$CV = \frac{\sqrt{X}\sigma}{XD} = \frac{1}{\sqrt{X}}\frac{\sigma}{D}. \quad (5)$$

The above equation shows that the relative variation, *CV*, increases when the number of time periods *X* decreases. This is what occurs in short-term planning. On the other hand, the relative variation decreases when the number of time periods *X* increases, as happens, for example, in medium-term planning, where multiple days are aggregated. This kind of idea appears e.g. in the paper by Tokola and Niemi [49] and in the book by Cachon and Terwiesch [45, p. 252].

Uncertainty in long-term planning is different from the variation in the short term. The uncertainties in the long term are actually changes in the average demand *D*. This uncertainty could be simplified as a random walk, as done e.g. by Graves [50]. The average demand "walks" in a random direction one day after another. The changes in the average demand are small, but as they accumulate, the difference after a year might be large. So during the time *t*, the demand might be D_t and the small random daily change could be ε_t. After a day, the next demand would be

$$D_{t+1} = D_t + \varepsilon_t \tag{6}$$

If the average random daily change ε is normally distributed with a zero mean, the average demand after X days will be distributed as follows:

$$D_{t+X} \sim N\left(D_t, \sqrt{X\ var(\varepsilon)}\right) \tag{7}$$

Although the demand seems to stay near D_t, the total sum of demands will have much greater variations as the small changes will accumulate.

Figure 3 illustrates Equations 5 and 7. In the short term, when the planning period is small, the average demand can be known quite accurately, but it has significant standard deviation around the average. On the other hand, in the long term, the average value is unknown, but if it could be forecast, the standard deviation from the average value would be small as a result of the law of large numbers. This can all mean, for example, that the daily demand cannot be forecast, but the weekly demand is quite stable. The uncertainty is different, depending on whether we are talking about the short term, medium term, or long term.

2.1.5 Short-term methods in industry practice

Because of the high number of variables, large variation, and exceptions, it is impossible to forecast everything deterministically in short-term planning and the planning has to be dynamic. Situations where the original plan does not work are faced all the time. New solutions have to be found quickly, as the problems occur and affect the current situation. These matters are the reason why short-term planning is usually performed using simple priorities for daily operation and the exceptions are handled by a quick decision from the manager. Easily applicable methods such as an extended lead time or dispatching rules are commonly used.

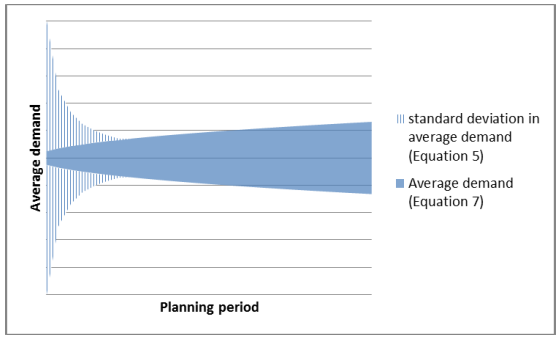


Figure 3. Illustrations of Equations (5) and (7). This shows how the demand uncertainty is different in different planning periods and how the deviation in the short term increases exponentially as the planning period is shortened.

Managers usually perform their daily planning manually. ***Manual planning*** has a great advantage compared to e.g. automated computer-based solutions. It permits a variety of solutions; it does not fix the solution. The well-known "***Law of Requisite Variety***" by Ashby [51] states that only variety can destroy variety. It means that if there are different problems, it is best to use different solutions for them. (Another well-known version of the same idea is "There is no silver bullet which solves all problems." [52]) This is true in short-term planning too. Another advantage of manual planning is that the logic of the planning is clearly known and the plan is easy to adjust in the event of emergencies.

It is often thought that manual planning might be far from optimal. This might be the case in a few complex situations, but in general, people have quite good common sense. If the planning problem deals only with a few items, a good solution can easily be determined. And if there are multiple items, people are able to find the most important ones and optimise them. However, on the basis of the author's experience, it is hard to get the incentives right, especially so that the production as a whole is optimal. Hopp and Spearman have summarised the above as follows: people, not organisations, are self-optimising [42].

Although manual solutions tend to be quite good, they typically have difficulties when there is significant variation in the demand or in the production. This is seen by Graves, who states that most of the planning systems do not explicitly account for uncertainty [53]. He also notes that the dynamic nature of uncertainty is typically handled by methods including increasing safety stocks, using extended lead times, using frozen time periods, time fencing, managing a backlog, doing rescheduling, or having a flexible capacity. On the basis of the author's experience, the parameters for the above methods are often obtained using trial-and-error methods.

The different methods work differently. Some methods make planning easier by relaxing its constraints. Both increasing ***safety stocks*** and using ***extended lead times*** do this. They reduce the performance of the system from the customer's point of view. Increasing the safety stock increases inventories as a defence against variation. Safety stocks at the beginning of production are much more inexpensive than safety stocks in the final inventory, as no work has been

done on them. However, safety stocks work only for common parts or products and thus in the case of a make-to-order system, an extended lead time may be the only usable method. Because it affects the end of the schedule, it is a safety factor against variability in the whole supply chain.

Flexible capacity can both make planning easier by relaxing capacity limits and at the same time make it more complex by adding one extra dimension to planning. It is a way to remove the need for an extended lead time or safety stocks, but it comes at a cost as the flexibility is an investment for the company, e.g. in the form of overtime or extra workers. In the short term, there are typically some strict limits on flexible capacity, e.g. the skill levels of workers and overtime regulations.

Some methods can respond to problems dynamically. Rescheduling, managing a backlog, and priority rules change the plan according to the current situation. ***Rescheduling*** is a reactive online approach to normal scheduling. It can use a static offline scheduling algorithm by repeating it. A good summary of rescheduling and the terms related to it was written by Viera et al. [54]. The summary is written from a scientific point of view and a practical view of different approaches was later discussed in a paper by Herrmann [55]. Reactive rescheduling was also studied by Sabuncuogly et al. [56]. Reactive rescheduling can be combined with robust solutions. From this point of view, planning could have an estimation for processing time variation and use stochastic scheduling to get the best possible average results, as was done by Pinedo [57]. In another study, by Branke and Mattfeld, a robust schedule that adjusted well to the arrival of new jobs was used [58].

To control the production, ***priority*** or ***dispatching rules*** are typically used in practice instead of concrete scheduling. They are one way to do the rescheduling, as they actually do the replanning repeatedly and at every time and in every place they are applied. According to some authors, this might be the reason why dispatching rules are good in the case of rush jobs (See e.g. [59]). The most common priority rules are FIFO (First In First Out), SPT (Shortest Processing Time First), and EDD (Earliest Due Date First). FCFS is not economical, but it is a “fair” rule. It is also the easiest to implement as the only information it needs is the order in which the jobs have arrived [60]. Most importantly, it reduces the flow time variation, which can be taken as increasing quality and reducing the tardiness of jobs in the system. SPT is optimal when the sum of the queuing and processing time is minimised in single-machine cases. The EDD priority rule is optimal when tardiness is minimised in a single-machine case with all the jobs having equal production times. It is also optimal in the single-machine case if all the products can be made before the deadline. Priority rules can take into account the weight of a job, i.e. the importance of the job. For example, in production there are typically ***urgent orders*** that should be completed as quickly as possible. An easy, common way to implement this is to use ***rush jobs*** that go ahead of other jobs.

Reactions to different problems cause disturbances to production plans, and some methods are needed to restrict the effects of these disturbances. A ***frozen time period*** can be used to reduce the disruptive effect of re-planning. During

a frozen time period, the plans cannot be changed any more. A frozen time period also works like an extended lead time, as new orders cannot be added during the frozen time period, and thus it can increase the lead time by the frozen time. If there is a small backlog in the system, then the frozen time is the same as the extended lead time, while in a high-backlog case the effect on the lead time is low. ***Time fencing*** works like a frozen time period, but it allows some limited changes during the period.

2.2 Online problems

Online problems are daily problems that are faced commonly in practice in many areas. It can be said that online problems appear everywhere where time runs and problems have to be solved reactively.

2.2.1 Online problems in the literature

In the literature, online problems are studied in the areas of computer science, operations research, and process engineering. In all these areas, the methods are applied in quite many different types of systems and the ideas behind the methods overlap between the areas [61]. Some matters are considered in all of them but there is a certain emphasis in every sector: the efficiency and novelty of the algorithm in computer science, probabilities or stochastics in operations research, and real-time control in process engineering.

In the computer science literature, the typical approach is to find an online algorithm that solves the problem using only online information and then compare that algorithm to the result of an optimal algorithm that knows all the information, i.e. an ***offline algorithm*** [3]. Generally, the solution for an online problem is worse than an optimal solution to an offline solution. This also means that the method that gives the optimal solution to an offline problem might not give the optimal solution for an online problem. A common target in computer science has been to find an algorithm that gives the best worst-case performance. The worst-case performance is a proof that the algorithm always gives at least the promised performance. Proofs for the worst-case performances are usually developed analytically for simple, easy-to-understand algorithms. The worst-case analysis is easier than the average case analysis, because one does not need any probability distribution. However, it is rather a pessimistic approach that is rarely usable in practice. For instance, it is uncertain whether the worst situations occur frequently, as is assumed in the traditional computer science approach.

In operations research, dynamic problems are related to online problems. The literature in this research field typically studies stochastic versions of the problem, where the probabilities are taken into account. The target is often to understand the dynamics of the system related to variation, not just the optimisation of the system [21]. Markov models can be thought of as simple stochastic models of reality where the goal is just to understand the effects of variation. In recourse models, first used by Dantzig in 1955 [62], there is an uncertain part of the solution and expected costs are minimised or expected

profits maximised. Such models are typically multistage models, where the first phase is optimised on the basis of the expected information and in the following phases the new information becomes known on the basis of the solution to the first phase. Two-stage problems are common [4]. In change constraints (Charnes et al., 1958 [63]) some probability of the system can be outside the given constraints. This is a way to avoid the problems with the worst-case approach used in computer science.

In chemistry and automation systems, the problems are closer to real-time control, where the control parameters have to be adjusted adaptively on the basis of some other continuously changing inputs. A typical example could be the adjustment of the pressure by taking the temperature into account. In addition to the impact on output, a change in the control parameter might have a feedback effect on the inputs [5]. Because of the delays in the effects, some kinds of repeated disturbances, sometimes called bull-whip effects, can appear in the system, and damping them is a challenge.

Table 2 combines the insights presented above into the ways in which the characteristics of online problems are handled in different areas. The main differences between them are that the uncertainty is usually modelled in a stochastic manner in the operational research area. In online optimisation, the information is revealed gradually and stochastic models are rarer. In process control, the dynamic local solutions are weighted as the modelling of the whole system may be a problem.

2.2.2 Categories: Reactive or robust

Online optimisation solutions can be roughly divided into two categories, into ***reactive*** and ***robust solutions***. If a solution is reactive, plans are modified when new information becomes available. Typical reactive solutions include replanning or rescheduling of the current solution in the event of disruptions. In robust solutions, the possibilities of disruptions are taken into account beforehand by e.g. forecasting them. In practice, the solutions tend to have characteristics from both of these categories. [54] [64]

Reactive online solutions modify the current plan after a disturbance to the current plan takes place or when new information appears. This kind of solution is especially useful when one does not know about the future, or when something unexpected and unpredictable happens. As an example, a reactive solution to an unexpected machine breakdown may be that the current orders on that machine are rescheduled to be processed on another machine.

Reactive solutions can be applied ***on demand***, when information appears. A reactive approach is applied if and only if there is a need. This makes it useful for those situations that occur rarely. If the situations need to be addressed frequently, reactive solutions can also be applied ***periodically***.

A reactive approach needs ***flexible resources***; something has to be able to change, or otherwise a reactive approach is not feasible. The more flexible the resource, the easier the rescheduling is. In practice, flexible resources could be

Table 2. Typical characteristics of online problem-related problems in the literature

	1. Incomplete data	*2. Knowledge increases when time passes*	*3. Need for quick solutions*	*4. Restriction caused by other decisions*
Online optimisation in computer science	-	After a decision, information is revealed. The worst possible response occurs.	Easily applicable algorithms	Changes to the earlier solutions are possible
Stochastic optimisation in operations research	Uncertainty in the variables	-	-	Complex dependence of the available resources
Process control in chemistry and automation	Dynamics of the system are not well understood	-	Processes must be dynamically changed	Dynamic, local decisions affecting each other

understood as free production capacity and thus used to shorten lead times for new orders. A drawback with a reactive approach is that it can cause perturbations to the existing plans. This is a big problem if resources are non-flexible or planned resources have dependencies. The perturbations caused by the disruptions can be avoided by using a robust solution, by taking the problem into account before it appears.

A robust solution is a solution that works even in the case of an unexpected event. An example of a robust solution to a machine breakdown can be that some slack is left in the schedule so that the breakdowns cause perturbations only to small parts of the plan.

According to Herrmann [55], there are two parts to robust solutions. One is to reduce the probability of something happening. This can be achieved e.g. by ***forecasting*** to find out potential problems and by using more reliable machines for critical resources. The other part is to reduce the impact if something unexpected happens. This is usually realised by using **a *slack***, i.e. having backup resources or having an extended promised completion time. Enabling reactive approaches to be adopted, e.g. by using duplicate resources and flexibility, can be another alternative.

Robust approaches need some kind of forecasting. Because of forecasts, robust solutions need more complex calculation than reactive solutions. Stochastic optimisation is a method that is used to deal with robust situations. In it, variables are assumed to be somehow distributed, and the objective is to find a solution that gives the best possible expected outcome. However, in production, it might be reasonable to use approximations of distributions. The simplest approximation typically uses an average case as an approximation. Another typical approach is to build a finite number of scenarios and calculate how our solution works for them. This is called a ***scenario-based approach***. If

scenarios are tested one by one, scenarios can be called ***what-if scenarios*** as well.

In practice the solutions are combinations of both reactive and robust approaches. They can be called a ***predictive-reactive*** approach. An example of a predictive-reactive approach is flexible capacity. The proactive part of the solution is the decision to have extra capacity, but the work has to be assigned to the capacity using a reactive approach.

Robustness and reactivity have been studied in project management by Herroelen and Leus [65]. They conclude that robustness is needed e.g. if there are lots of dependencies between different jobs and all of them have high variability. They also state that if variability alone is the problem, but the jobs are independent, reactive priority rules are enough. Without variation the problem becomes deterministic as no disruptions occur. The dependence between jobs increases the need for robustness as a result of the fact that small disruptions can have severe consequences that need excessive use of reactive solutions, if they are used alone. So, variation and dependency in a system define whether robust or reactive methods are used. Variation is the reason for both types of methods, but dependency creates the need for robust solutions.

2.3 Short assessment

Short-term production planning is a potential area for online optimisation. However, in order to fulfil this potential it is important to understand the challenges behind online problems and find out the relationships between short-term production planning and online optimisation.

In the introduction, it was stated that the properties of online problems are incomplete data, the passing of time, constraints, and quick decision making. Figure 4 illustrates these properties, points out what the main challenges are, and describes potential solutions. Incomplete data, the passing of time, and constraints make the modelling itself complex and thus hard. Natural solutions to this are simple, local solutions such as priority rules. The passing of time and constraints increase the need for flexibility. In practice, this could appear in the forms of extra capacity and short lead times. Incomplete data and quick decision making together increase the challenge of getting usable data, which obliges the system to be standardised. This can be achieved by automation and modularisation.

Figure 5 shows how the short-term planning methods fit into the reactive-robust categories that were discussed in the previous chapter. Reactive methods are methods that act after the disturbance, while robust methods are applied before the disturbance. Rescheduling and priority rules are generally reactive methods; they can be applied when problems appear. Flexible capacity is both robust and reactive. Its use typically has to be decided beforehand, but its actual use can be reactive. Time fencing allows certain changes to the schedule, but restricts other changes. Safety stocks and an extended lead time are robust ways to solve uncertainty. The frozen time period method is neither completely

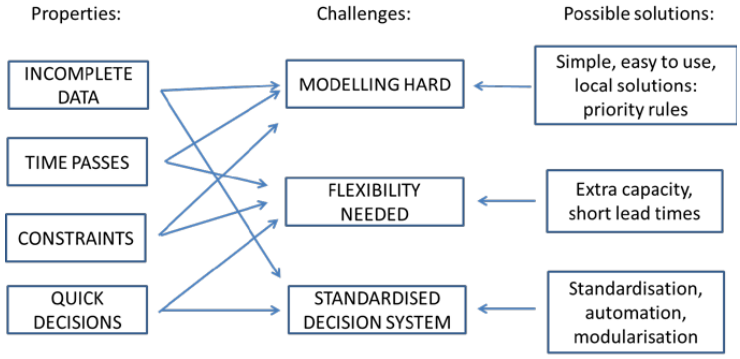


Figure 4. Properties and challenges in online problems, their relation to the problems in online optimisation and solutions to the challenges

reactive nor robust. It is a method to limit the problems that appear as a result of e.g. the extensive use of rescheduling.

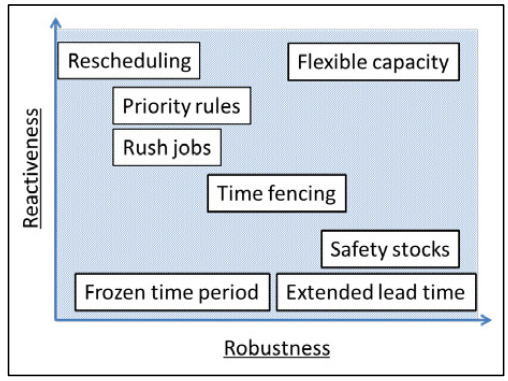


Figure 5. Different practical methods to deal with uncertainty in production control and their relation to reactiveness and robustness

3. Online optimisation models in production planning

Six publications in this thesis study online planning from different perspectives. First, Publication I contains a survey of manufacturing companies where the typical problems, including online problems, are revealed. The other publications in this thesis study five online planning problems in five different production systems. The systems that are studied are parallel stations, make-to-stock (MTS), assemble-to-order (ATO), make-to-order (MTO), and one-of-a-kind production. Each system features different daily online planning challenges. In the case of parallel machine scheduling, reactive worker coordination is studied. In a make-to-stock (MTS) system, the planning system has an effect on the efficiency as the production capacity is limited. In the ATO system, the component-product structure has an important effect on the impact different online problems have. In MTO systems, new orders can be the main issue and it is essential to schedule them efficiently into the production schedule. In one-of-a-kind production, a special case of engineer-to-order (ETO) systems, tardy material deliveries are a significant problem. This chapter discusses the above challenges and describes the models used in the publications of this thesis to study the challenges.

3.1 Practices and problems in companies (Publication I)

Different types of companies face different problems and challenges. For example, visibility in the supply chain, production type, the number of product types, and volumes affect the operating conditions of companies.

The bullwhip effect (also known as the Forrester effect) increases demand variation seen by companies upstream in a supply chain [66]. The reason for this is that visibility through the supply chain is lacking and forecasts are made individually by each company. It seems that companies upstream could benefit from online optimisation more because their demand varies a lot and thus they have to reactively reschedule their schedules.

The type of production also affects problem characteristics. Production volumes typically define the type of production [27]. Mass production has different problems from flexible small companies. Mass production has a more process-like environment, whereas small companies tend to produce items in batches. Batch production should have gain benefits e.g. from rescheduling.

Tokola and Niemi developed a model that suggests that companies with a small number of products often have problems with capacity, whereas companies with a large number of products probably have problems with inventory levels [49]. If there are only a few products, changes in the demand for each product have a high impact on capacity. However, if the number of products is large, it is quite certain that there is also a large number of products that have volatile demand, which makes it hard to have good inventory levels. Both cases benefit from online optimisation. In the case of few products, it is crucial to have extra capacity and coordinate the capacity in an online fashion. In the case of multiple products, capacity has to be flexible, but there does not necessarily have to be extra capacity.

Publication I further studies current problems and practices in different manufacturing companies by publishing the results of an interview study. Figure 6 illustrates the flow and main results of the publication. Three types of companies are studied: small companies, subcontractors, and large companies. The interview reveals common restrictions and problems regarding the online optimisation of production planning. The following insights regarding the online optimisation are found.

The small companies have extra machines and workers are cross-trained. Both help in balancing or rescheduling production in the event of disturbances. The WIP on the factory floor is often tackled using pull-based production planning. The small companies also tend to have more problems in forecasting, which could be tackled by using more dynamic scheduling.

Subcontractors seem to have the greatest number of problems in their daily production. Rush jobs are common (10-20% of orders are rush jobs). Process time variation is also significant. The problem is that workers' know-how is often limited. Rescheduling is performed daily or weekly.

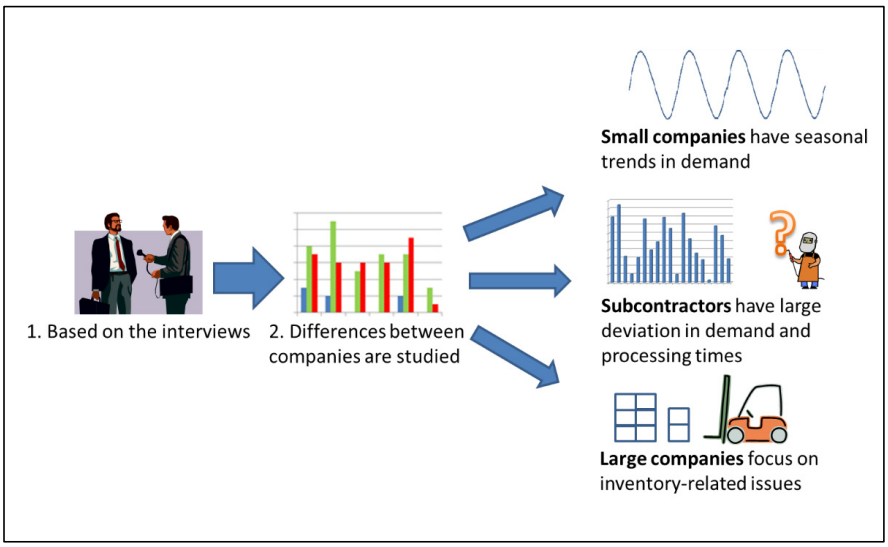


Figure 6. Main results of the interview regarding shop-floor-level control.

In large companies inventories are used to deal with rush jobs. Workers are often cross-trained, which helps balance the production. It is typical that large companies are mass producers and variation is reduced using constant demand, frozen schedules, and bottleneck thinking. Generally, it seems that rescheduling and other online optimisation methods are not as relevant for large companies as for small companies and subcontractors.

3.2 Worker coordination on parallel stations (Publication II)

As discussed above in Section 3.1, the companies tend to have cross-trained workers who can move between work tasks to balance the situation in the factory. This needs coordination. If a job at a station is completed, it might be beneficial to move a cross-trained worker to help a worker at another station. Different types of restrictions can inhibit the movement of workers.

In a parallel station environment it is possible to do work efficiently even with a varying workload. The jobs at different parallel stations do not have to wait for the completion of other jobs if they are completed before the estimated time and they do not block other jobs if they are not completed. This kind of environment commonly appears with subcontractors where e.g. welders are distributed among parallel stations.

As a part of their paper Hopp and Oyen review the literature regarding workforce coordination [67]. They state that a common way to implement coordination is to assign cross-trained workers to tasks on the basis of clock times or predetermined sequence of jobs. This scheduling is typically performed by using mathematical approaches. Floaters are also studied in the case of production lines. Zoned worksharing is also a common target of study in the case of production lines. However, few papers study different coordination policies in the case of parallel stations.

Publication II compares four different coordination policies in parallel station systems. The coordination policies are no helping, floater, pairs, and complete helping. The worker movements in the different policies are shown in Figure 7. In the no-helping policy, moving to help is not possible. In the floater policy, there is one floater worker who can move to assist others, but others cannot assist each other. In the pairs policy, workers form pairs so that they can only assist their pair partners.

For the coordination policies described above, Publication II constructs continuous-time Markov process models to compare the different coordination policies. As an example, the model for a floater is constructed as follows. First, N denotes the number of stations. Next, each state in the model can be defined as a set of (n, S, P), where n is the number of jobs, S is the number of single workers, and P is the number of pairs. In the case of a floater, there can only be one pair ($0 \leq P \leq 1$). Processing rates are defined using the processing rate of a single worker μ and the collaborative efficiency α, which indicates the performance of a work pair in relation to the performance of two single workers. Now, using the previous notations, in the case of a floater, the processing rates

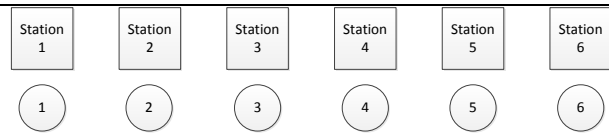
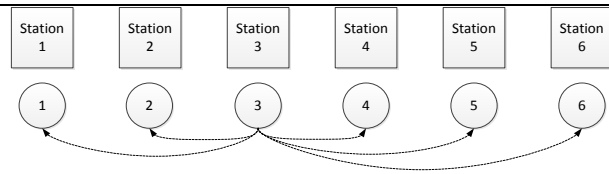
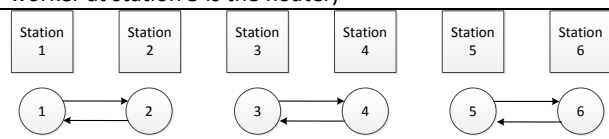
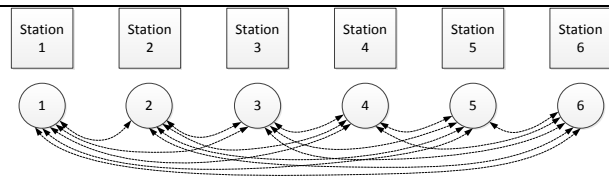
Policy	The station to move to
No helping	 No moving
Floater	 Randomly selected station with work left (in this example, the worker at station 3 is the floater)
Pairs	 Fixed neighbouring station
Complete helping	 Randomly selected station with work left

Figure 7. Studied coordination policies [Publication II].

are the sum of the processing rates of single workers ($S\mu$) and the processing rate of the pair with a floater ($\alpha P2\mu$).

Valid states are defined as follows.

$(n, N, 0)$, if $N \leq n$
$(n, n, 0)$, if $0 \leq n < N$
$(n, n-1, 1)$, if $1 \leq n \leq N-1$

When the job is processed in a state, the movements between states are defined as follows.

$(n, N, 0) \rightarrow (n-1, N, 0)$, if $N < n$
$(n, n, 0) \rightarrow (n-1, n-1, 0)$, if $2 \leq n \leq N$, with a probability of $(n-1)/n$
$(n, n, 0) \rightarrow (n-1, n-2, 1)$, if $2 \leq n \leq N$, with a probability of $1/n$
$(n, n-1, 1) \rightarrow (n-1, n-2, 1)$, if $2 \leq n < N$
$(1, 0, 1) \rightarrow (0, 0, 0)$
$(1, 1, 0) \rightarrow (0, 0, 0)$

New jobs are assumed to arrive at the rate λ. If a new job arrives, the movements between states are defined as follows.

$$(n, n, 0) \rightarrow (n+1, n+1, 0), \text{ if } 1 \leq n < N$$
$$(n, N, 0) \rightarrow (n+1, N, 0), \text{ if } N \leq n$$
$$(n, n-1, 1) \rightarrow (n+1, n, 1), \text{ if } 1 \leq n < N-1$$
$$(n, n-1, 1) \rightarrow (n+1, n+1, 0), \text{ if } n = N-1$$
$$(0, 0, 0) \rightarrow (1, 0, 1)$$

It should be noted that a new job in the case of $(n, n-1, 1)$ is given to a non-floater if there are available non-floaters. It is optimal to keep the floater helping others. Figure 8 shows an example of a diagram of the states in the case where there are six parallel stations ($N = 6$). In the publication, similar models to those above are created for all other coordination policies as well.

Using the above model, it is possible to calculate average cycle times for the jobs by using steady-state equations. This is omitted here but it can be found in Publication II, where average cycle times are calculated for the case of a given set of jobs and for the case of the continuous arrival of jobs.

Publication II reveals certain results that are interesting from the online point of view. In the case of an online setting where new jobs arrive continuously, the pairs policy gives similar results to complete helping, although complete helping is much more complex. However, in an offline setting where a fixed number of jobs is processed, complete helping should be used instead of the pairs policy.

3.3 Urgent orders in make-to-stock production (Publication III)

Urgent orders are often the key online optimisation challenge because, as they are urgent, they have to be produced faster than normal orders. Publication III studies the planning practices of make-to-stock (MTS) production in the case of urgent orders. The publication is based on an industry practice by which the new schedules were constructed weekly, but in the event of shortages, urgent orders (i.e. rush jobs) were dealt with reactively. In the publication, this practice is compared to a more simple periodical planning and to a more demanding complete online replanning.

In the MTS production that was studied, the same production capacity is used to produce N different products to add to the stock. Each product i has a current stock β_i and an average demand λ_i. The planning has to decide what product to

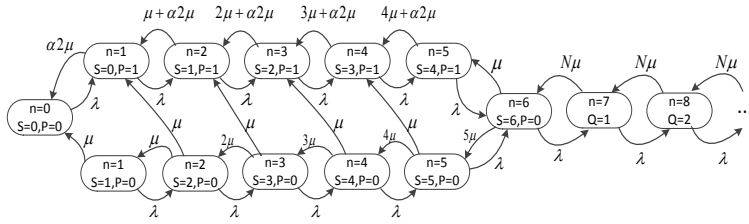


Figure 8. State diagram for six parallel stations with the floater policy [Publication II]

produce, how much to produce, and in which order. This is a common setting in MTS-type productions.

Order-up-to models can be used to calculate the production amounts that are needed, i.e. to specify what to produce and how much. In the order-up-to model, the inventory is updated back to the given target level T_i as quickly as possible when demand reduces inventory levels. The optimal parameter values for the order-up-to model can be solved using newsvendor and economical order quantity models. In the classic newsvendor model by Edgeworth [68], the demand distribution has to be estimated and after that the stock costs, C_U, and stockout costs, C_O, have to be balanced. The newsvendor model gives the optimal ordering amount Q, which is equal to the probability *P* of a stockout as follows:

$$P(Q) \geq \frac{C_U}{C_O + C_U} \tag{8}$$

This newsvendor model can be used to achieve the target level T_i for the product inventory. However, in Publication III, these target levels were set to the probability *P* of 95%. In the order-up-to model, there are no batch ordering costs and thus e.g. an economic order quantity (EOQ), formulated by Harris [69] but also known as Wilson's EOQ formula, is not needed. It could be used if there were significant setup costs for individual products.

Classical models, including the above newsvendor model and EOQ models, typically assume unlimited capacity, e.g. in the form of a constant lead time. The reason for the assumption of unlimited capacity is that the problem is easier to solve than with limited capacity. However, the production always has a limited capacity and in some cases this should be taken into account. Zipkin [70, p. 256] outlined two key insights into limited production capacity. The first insight is that limited capacity and variation (in demand) will cause congestion in production. It will increase the lead time and thus also the inventory. The second insight is that in a system with variation in input or in processing, excess capacity is necessary for stability. If there is no excess capacity, the performance of the system will explode as the service times will become worse without any limit when more jobs arrive than the system can process. These insights are similar to the insights that are obtained from the queuing models introduced in Chapter 3.

In Publication III, the production order is constructed using product weights set to the current inventory divided by the average demand, i.e. β_i/λ_i, and assuming that average demand occurs during the period. The production queue can be constructed as follows. The product with the lowest weight will be produced first. After the item with the lowest weight has been scheduled, its inventory can be assumed to have been increased by one, and the new item with the lowest weight is scheduled. The average demand is assumed to reduce the inventory after the production limit for the day has been reached. This whole process is repeated until the schedule for the given number of days has been constructed. The process can be applied both online and periodically.

In Publication III, the benefits of online scheduling were explored. This was done by comparing the performance of the planning system described above in three planning situations: in the cases of periodical planning, an urgent job approach, and complete replanning. Periodical planning creates a plan for a certain period, e.g. for a week, and the plan is strictly followed during that period. In the urgent job approach, the periodical approach was followed but if the inventory fell below a threshold value r_i, an urgent order for that product was issued. That urgent order bypassed the normal schedule. In complete replanning, the scheduling was done completely in real time, online, without any periodical planning or urgent orders.

In the numerical experiments of Publication III, simulation was used to compare the periodical planning, urgent order approach, and complete replanning. The comparison used stockouts, i.e. the number of unfilled orders, as an objective. In the results, plain periodical planning gives higher numbers of stockouts than the urgent order approach and complete replanning. However, the level of difference between the urgent orders approach and complete replanning depends on the target levels and urgent order thresholds. The urgent orders approach gives higher numbers of stockouts, but if the levels and thresholds are high, both methods give similar levels of stockouts. Urgent orders give significantly higher numbers of stockouts only if the ordering threshold is near zero or the target levels are low.

The above results in Publication III suggest that urgent orders are enough if one does not have too many of them. Otherwise, e.g. in a situation where there are many highly urgent orders, it may be worthwhile to use complete replanning, the online optimisation approach.

3.4 Flexibility of assemble-to-order production (Publication IV)

The flexibility of the production has a critical effect on the way in which short-term planning problems occur and on the way in which online methods can be applied. A concrete example of flexible production is the ATO (assemble-to-order) system, which is presented in Figure 9. In an ATO system, the products are assembled after an order is received and the same components can be used for different products in a flexible way. This enables inventories to be reduced while still avoiding long lead times resulting from the component orders. The flexibility of the ATO system was studied theoretically in Publication IV.

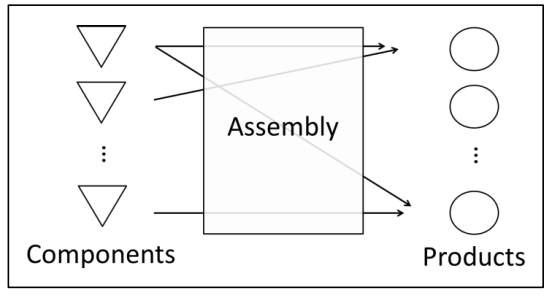


Figure 9. In an ATO system, the same components are used to assemble different products.

In general, the ATO system deals with variation. Variation in a production process is created by input, the processing times of machines, resource problems, and routing problems (see e.g. [45, p. 105]). Variation can be handled using extra capacity, reducing the need for setups, and reducing unnecessary variation. Online optimisation can improve the utilisation of the system with variability, but it requires flexible capacity, which can be considered as a form of extra capacity.

In ATO systems, the extra capacity comes from the flexibility of assembly. In general, extra capacity can be achieved without direct excess capacity by pooling the uncertainties. Examples of pooling include shared capacity among multiple products or a central inventory [45]. Pooling has been studied recently in the literature. See e.g. a general paper by Sobel about pooling [71] and the paper by Benjaafar et al. about the utilisation of pooling in a production-inventory system [72]. Pooling, flexibility [73], and component commonality [74] are related terms and they basically mean the same thing, but from different perspectives: from the management, resource, or product perspectives.

The basic problems of the scheduling of ATO production can be found in the review by Song and Zipkin [75]. In a general ATO system, there are I products, each of which is assembled from a subset of common J components. A matrix can be used to represent these component-product relations so that the (i,j) cell in the matrix defines the need for a component i for the product j. There is a shortage cost for products ($\boldsymbol{s}$) and holding costs for components ($\boldsymbol{h}$) and the objective is to minimise their total costs.

In Publication IV, the basic single period form is considered. It has the following three phases:

> P1: components are ordered on the basis of the forecasts
> P2: demand for products becomes known
> P3: demand is fulfilled by assembling the products from the components that have been ordered

Publication IV studies how unexpected changes affect the ATO problem and how well they can be tackled in a different ATO system. This is done by assuming that the demand is completely known in the P1 phase and unexpected changes have to be handled in the P3 phase. The problem in the P1 phase is easy; the number of components ordered, $\boldsymbol{y}$, is exactly the same as the known demand $\boldsymbol{d}$ and in P3, if everything goes according to the plan, all the components would be used for assembling the products. However, after P2, the following unexpected changes might occur:

> CL: component loss of $\Delta\boldsymbol{c}$ units
> DD: demand decrease of $-\Delta\boldsymbol{d}$ units
> DI: demand increase of $\Delta\boldsymbol{d}$ units
> PL: production limit of Δl units

When one of the above unexpected changes occurs, the components planned for one product can be used for other products. However, it is hard to meet all the demand or use all the available components. Balancing this is the challenge.

To find a solution to the challenge, the following optimisation can be solved.

$$\text{Minimize } G = \boldsymbol{hx} + \boldsymbol{sw} \tag{9}$$

Subject to

$$\boldsymbol{Az} + \boldsymbol{x} = \boldsymbol{Ad} + \Delta\boldsymbol{c} \tag{10}$$

$$\boldsymbol{z} + \boldsymbol{w} = \boldsymbol{d} + \Delta\boldsymbol{d} \tag{11}$$

$$[1,1,\dots,1]\boldsymbol{z} \leq [1,1,..,1]\boldsymbol{d} - \Delta l, \text{ if } \Delta l > 0 \tag{12}$$

$$\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z} \geq 0 \tag{13}$$

The decision variables $\boldsymbol{x}$, $\boldsymbol{w}$ and $\boldsymbol{z}$ denote excess numbers of components, unfilled demands of products and produced amounts of products, respectively. The objective function in Equation (9) minimises the holding and stockout costs. Equation (10) defines that used components plus excess components equals the components in hand. Equation (11) defines that produced products plus unfilled demand is equal to the demand. Equation (12) forces production to be equal to or less than production limit (if there is such a limit).

Further, Publication IV studies the optimisation problem in Equations (9-13) in the case of a single product that has a high shortage cost (corresponding element in $\boldsymbol{s}$ is high) and in the case of a single component with a high holding cost (Corresponding element in $\boldsymbol{h}$ is high). This assumption can be thought of as following the reality as in practice there are differences between products and components and one product (or customer) is always more important than others and one component costs more than others.

Using the above models and assumptions, the probability formulae for facing shortage or holding costs are constructed for all four unexpected changes. As an example, during a component loss (CL), the shortage cost will be faced if a product with a high shortage cost needs the CL component and there are no other products using this component. The probability of a CL occurring for a specific product is p and the probability that no other products are using this component is $(1-\mathrm{p})^{\mathrm{I}-1}$. By combining both previous probabilities, the probability of facing high shortage costs is

$$p(1-\mathrm{p})^{\mathrm{I}-1} \tag{14}$$

All the formulae in Publication IV follow the idea behind the above example.

The formulae constructed in Publication IV can be used to estimate and compare the impacts of different unexpected changes. As expected, shortage costs are the main problem in the case of loose ATO networks (with few common components) and holding costs in the case of dense ATO networks (with multiple common components). Naturally, the decreasing demand is the problem in the case of high holding costs in a dense network. Equally, the increasing demand is the problem in the case of a high shortage cost of a product in a sparse network. Generally, component loss (CL) was shown to have the highest probability of increasing the costs, regardless of the density of the ATO network. The production limit (PL) causes problems only in high-density product-component networks. The production limit is the case in which online optimisation could be most useful as in that case there is the greatest number of

possibilities of the production being changed. This suggests that online solutions are most beneficial if the production is limited but at the same time there is variability in the ATO network.

3.5 Rescheduling in make-to-order production (Publication V)

In make-to-order (MTO) production demand is uncertain but high utilisation has to be achieved. Publication V studies the online scheduling of make-to-order production. In such production different types of jobs arrive and they have to be scheduled for production in such a way that jobs are completed before their due dates or with as little tardiness as possible. When the jobs are scheduled, the material requirements planning (MRP) system is then typically used to order the material. The ordering of the material is quite straightforward and thus the complexity in the MTO system typically relates to the demand and the details of the demand. On the factory floor, this uncertainty is seen as new orders and it is important to schedule them in such a way that high utilisation is achieved, but the jobs are completed before their due dates.

In general, scheduling allocates the tasks to be done by resources at a specific time. A Gantt chart, named after Henry Gantt, is a typical way to describe scheduling. Figure 10 illustrates the scheduling in the form of a Gantt chart. In a Gantt chart, different areas on the y-axis are used for different resources or jobs and the actual start and finish times for the processing of the tasks can be read on the x-axis.

The scheduling of MTO production in cases of known orders has been studied before. Gupta and Stafford [76] reviewed scheduling in the past 50 years. Although scheduling has been studied quite intensively, papers about online scheduling are rarer. This might be due to the modelling complexity, which is hard in normal scheduling and will become even harder when new orders and disruptions are taken into account.

The way in which these new orders are scheduled has been a classical optimisation problem in the online literature. Simple cases, such as machine scheduling and lot sizing with a single machine and parallel machines, have been studied in papers on computer science. Such simple cases typically appear

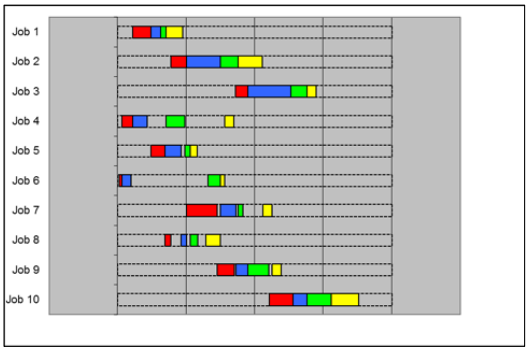


Figure 10. Gantt chart of 10 jobs scheduled for a 4-machine flow shop.

with computers, and with them the scheduling can be applied effectively as the information is easily available.

In production, instead of complex optimisation, priority rules are used for scheduling single machines. Priority rules can be seen as an online solution for the problem. It is well known that for single machine scheduling the shortest processing time (SPT) priority rule gives the optimal average throughput and the earliest due date (EDD) priority rule minimises the maximum tardiness. In an online setting, these results do not apply any more. The jobs have different arrival times, which alone is enough to make the problem hard, and the priority rules give a non-optimal solution. In addition, the details of the new orders are seldom accurately known, which makes the problem even harder. However, if items are arriving in a continuous manner, First-In-First-Out (FIFO) might be a fair algorithm, as it minimises the maximum waiting time and the variation in the waiting time [60] and schedules the jobs similarly in both the online and offline settings.

For parallel machines, list scheduling (LS) was the first notable online algorithm to be studied [77]. In list scheduling, there is a list of jobs from which the first one is selected, removed from the list, and allocated to the machines when they become free. For m machines, LS is shown to give a schedule which has a length which is at most $2\text{-}1/m$ times the optimal solution for m machines in the worst case. In an average situation, it is more efficient. In practice, people naturally use an LS kind of approach to schedule parallel machines.

The scheduling of machines becomes more complex in flow shops and job shops and simple priority rules do not work as efficiently any more as with single or parallel machines. The online scheduling of flow shops and job shops focuses on recovery from disturbances such as machine breakdowns. For the purpose of stability there is match-up rescheduling [78], the purpose of which is to try to fix a schedule that is broken by breakdowns within a certain time frame. It is typical to restrict the number of moves in rescheduling or restrict the moving distance in the processing time of jobs in rescheduling [79, 80]. From the point of view of easiness, there is the Right Shift rule, in which the scheduled jobs are moved forward so that the schedule is feasible again [81]. Another restriction, called affect operation rescheduling [82], focuses on the operations that are delayed, and restricts the changes to the initial schedule. A summary of the above rescheduling methods can be found in [83]. Priority rules can be seen as a way to implement rescheduling. However, they require the supply of materials and components not to be a problem.

Our contribution in Publication V is in rescheduling in a case where jobs that have already been scheduled cannot be advanced because of material orders. The publication considers the scheduling of a flow shop. When a flow shop is studied in its static form, in offline settings, it has J jobs and M machines. Each job j has a release time r_j, after which it can be released to production. On each machine m, the job has the processing time p_{jm}. In order to minimise the total tardiness, the schedule for all the jobs on the machines, t_{jm}s, is solved with the following optimisation model:

$$\text{Minimize } \sum_{j=1}^{J} \max(0, t_{jM} + p_{jM} - d_j) \tag{15}$$

Subject to

$$\forall j \in [1,2,\dots,J], m \in [1,2,..M]: t_{j(m+1)} - t_{j(m)} \geq p_{jm} \tag{16}$$

$$\forall j_1 \in [1,2,\dots,J], j_2 \in [1,2,\dots,J], j_1 \neq j_2, m \in [1,2,\dots,M]:$$
$$t_{j_1 m} - t_{j_2 m} \geq p_{j_2 m} \; or \; t_{j_2 m} - t_{j_1 m} \geq p_{j_1 m} \tag{17}$$

$$\forall j \in [1,2,\dots,J]: t_{j(1)} \geq r_j \tag{18}$$

Publication V studies the above model in an online setting, in which new jobs arrive and the schedule has to be updated, e.g. by repeating the above offline optimisation problem. It is quite clear that if a job has already started or been completed on a machine, its schedule cannot be modified. However, jobs that have been planned but have not yet started can be moved but it may be hard because of other dependencies, e.g. as a result of material orders. With this in mind, Publication V compares rescheduling policies that have different restrictions on the moving direction of jobs. The policies that are compared are these:

FIXED – Existing jobs cannot be moved at all
FORWARD – Existing jobs can only be moved forward in time, not advanced
FREE – No restrictions on moving existing jobs
OFFLINE – Scheduling the problem offline without any restriction, by knowing all jobs beforehand

Restricting the scheduling makes online scheduling easier to understand as the complexity of rescheduling decreases. Publication V presents the complexity in the form of a number of different schedules in each policy. With the FIXED policy, when a new job arrives, the maximum number of orders is n. With the FORWARD policy, the corresponding number is 2^{n-1}, whereas with the FREE policy, there are $n!$ possibilities. When n is large, i.e. when there is a large number of concurrent jobs, the differences between the cases are large. This increases the computational challenge as well. The numerical results in Publication V show that the FORWARD policy, which restricts scheduling backwards, is almost as good as the FREE policy, the non-restricted one.

3.6 Delays in one-of-a-kind production (Publication VI)

In the production planning of one-of-a-kind products, tardy material deliveries are often the biggest daily challenge. This is due to the fact that such products contain multiple non-standard parts and these parts have complex interdependencies. If there is a large number of parts and there is a probability of most of the parts being delayed, it means that almost certainly at least one part will be delayed. One-of-a-kind shipbuilding is studied in Publications VI and in a manuscript by Tokola et al. [84]. Publication VI models the block erection scheduling in shipbuilding and the manuscript [84] extends it with a scenario-based case where delays in the block arrival dates are handled.

In practice, complex one-of-a-kind projects can be handled using the critical path method of Walker and Kelley [85, 86]. The main idea of the critical path method is to find out the time and precedence constraints for different jobs and identify the critical path, i.e. the jobs that are directly critical in terms of the completion time of the project. After that the focus of planning should be on the completion of the jobs that are on the critical path. For example, flexible workers, buffers, or inventories could be used to make sure they finish on time. This is important as the number of jobs on the critical path is typically low compared to the total number of jobs. Kelley [83] notes that the share of critical jobs might generally be less than 10% of all the jobs.

Publications VI and the manuscript [84] study the block erection scheduling of shipbuilding. Shipbuilding is a good example of a one-of-a-kind project that has a large number of components and complex interdependencies between the components. Ships are combined from large blocks, which are erected into a ship one by one. The erection scheduling defines the order and schedule for the blocks. This is not simple as the block erection capacity is limited and there are complex interdependencies between the blocks. Figure 11 illustrates the block structure of the ship.

The complexity of block erection is preliminarily studied in Publication VI, where the following optimisation problem is constructed:

$$\text{Minimise} \max_{h\epsilon[1,2,\dots,H],l\epsilon[1,2,\dots,L]} A_{hl} \quad \text{(Objective function)} \quad (19)$$

Subject to

$$A_{h_1 l_1} \geq A_{h_2 l_2} + p \| A_{h_1 l_1} + p \leq A_{h_2 l_2}$$

$$\forall h_1 \in [1,2,\dots,H],\ h_2 \in [1,2,\dots,H],\ l_1 \in [1,2,\dots,L],$$

$$l_2 \in [1,2,\dots,L],\ h_1 \neq h_2,\ l_1 \neq l_2 \quad \text{(Lifting capacity)} \quad (20)$$

$$A_{h(l+1)} > A_{hl} + D_L \| A_{h(l+1)} + D_L < A_{hl},$$

$$\forall h \in [1,2,\dots,H], l \in [1,2,\dots,L-1] \quad \text{(Horizontal joining)} \quad (21)$$

$$A_{(h+1)l} \geq A_{hl} + D_H, \forall h \in [1,2,\dots,H-1], l \in [1,2,\dots,L] \quad \text{(Vertical joining)} \quad (22)$$

$$A_{(h+1)l} \geq A_{h(l+1)}, \forall h \in [1,2,\dots,H-1], l \in [1,2,\dots,L-1] \quad \text{(Structural stability)} \quad (23)$$

$$A_{(h+1)l} \geq A_{h(l-1)}, \forall h \in [1,2,\dots,H-1], l \in [1,2,\dots,L] \quad \text{(Structural stability)} \quad (24)$$

$$A_{hl} > A_{h(l+1)} > A_{h(l+2)} \| A_{hl} < A_{h(l+1)} < A_{h(l+2)} \| A_{hl} > A_{h(l+1)} < A_{h(l+2)},$$

$$\forall h \in [1,2,\dots,H], l \in [1,2,\dots,L-2] \quad \text{(No skipped blocks)} \quad (25)$$

$$A_{hl} \geq 0, \forall h\epsilon[1,2,\dots,H], l\epsilon[1,2,\dots,L] \quad \text{(Starting time)} \quad (26)$$

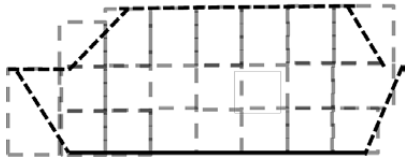


Figure 11. The large ship is divided into blocks, which are erected into the ship being constructed one by one.

As shown in the above model, the scheduling of block erection has several constraints. The objective in Equation (19) minimises the maximum erection time of the blocks, i.e. the time of the erection of the last block. The lifting capacity constraints in Equation (20) force one block to be erected at a time. Horizontal and vertical joining delays are introduced in Equations (21) and (22). The structural stability constraints in Equations (23-24) make sure that there are enough blocks in the levels below before the erection of the topmost blocks. The no-skipped-blocks constraints in Equation (25) define that there cannot be a space between two vertically erected blocks at any time.

Publication VI studies the above model and finds out the ways in which the model can be solved in different erection capacity and joining time conditions. Publication VI shows that pyramid building is optimal in a case where the joining time is significantly large. In addition, the publication also shows that in some cases the problem can be solved efficiently.

A large number of blocks and tight interdependencies between the blocks make it important to study the block erection in the dynamic situation in which the blocks can be delayed in comparison to the planned times. Thus, unpublished manuscript [84] uses a scenario-based MILP model to study the effects of the delays. This scenario-based MILP model is based on the above MILP model introduced in Publication VI. Figure 12 illustrates the steps of the scenario-based model.

In the scenario-based approach, the different delay scenarios are randomly generated, and the average results with the scenarios are studied. When a delay occurs, the delay has to be dealt with or the erection times of blocks have to be

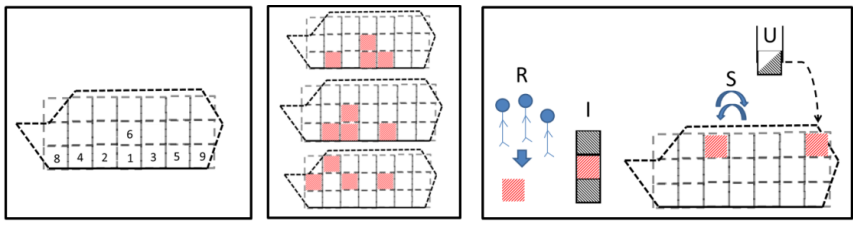


Figure 12. The scenario-based MILP model [84]

advanced from the times that were planned. A delay in the erection of one block typically has consequences for the erection of other blocks and therefore it has to be dealt with efficiently. To deal with the delays, manuscript [84] studies five different planning methods: erecting blocks unfinished, use of inventories, use of rush jobs, sequence changes, and, if the delays cannot be dealt with, delaying the completion of the ship. The use of inventories is the same for all these scenarios but all the other methods can be used in a scenario-wise fashion.

The results in manuscript [84] show how the optimal use of the methods is different for different delay types and for different ship types. Inventories or erecting unfinished blocks can advance the production of blocks without any limit (at least when compared to other methods) and thus they are used if the cost of delaying the whole ship increases. If a significant probability of delays in the blocks exists, there should be a buffer for the items instead of erecting them unfinished. Sequence changes are useful only when there are loose constraints between the blocks. Rush jobs seem to be more applicable for small ships than for large ships.

3.7 Summary

Table 3 combines the online aspects from the publications of the thesis. It categorises the methods used into reactive and robust solutions and outlines the cases where online optimisation is beneficial and the factors that limit the use of online methods.

The conditions where the possibility of online optimisation would have the biggest impact suggest that online optimisation should be used carefully. All of the cases where online optimisation gives significant benefits, i.e. moving workers, small rush jobs, rescheduling, a dense component-product network, and changing critical parameters, are special cases in production. Also, according to traditional thinking and quality issues, online methods such as rescheduling, rush jobs, and other changes should be avoided and the production should be as easy and boring as possible (see e.g. [87 p. 27]). Additionally, even if the online methods are used, they are hard to control and finding their parameters is not easy. In conclusion, as the usefulness of online methods increases, the challenges of applying them also increase as well.

Table 3. The methods in the papers

	Reactive methods	***Robust methods***	***Situation where online approach is beneficial***	***Factors that limit use of online approach***
Survey of control [I]	Rescheduling	Extra machines, overtime hours, inventories, cross-trained workers	Subcontractors, missing parts	Know-how of workers
Worker coordination in parallel station system [II]	Moving workers between stations	Allowing moving and helping.	Pairs and complete helping	Moving limitations (e.g. because of know-how of workers)
Make-to-stock, paper [III]	Reactive sequencing, urgent orders	Shared capacity, periodic sequencing	When rescheduling is often needed	Simple scheduling rules
Assembly-to-order [IV]	Using the components ordered for other products	Sharing components and capacity	Dense component-product network	Few shared components
Make-to-order [V]	Rescheduling	Flexibility for rescheduling	Rush jobs that are small compared to other jobs in the factory	Material deliveries, rescheduling possibilities
One-of-a-kind make-to-order [VI, 84]	Using higher costs to complete the jobs faster(i.e. erecting unfinished blocks), rush jobs, sequence changes	Inventories, capacity	Changing parameters (the penalty parameters can lead to different use of the methods)	Schedule tightness, small number of jobs (i.e. blocks)

4. Conclusions and recommendations for further work

This thesis consists of an introductory part and five papers that all focus on online optimisation in production planning. The methods used in these studies are mostly optimisation models, but also included simulation and stochastic models.

The introductory part gives a general overview of online optimisation problems and short-term production planning. Online problems have been studied earlier in computer science, operation research, and process engineering. Two common approaches to the problems are reactive and robust approaches. In production planning, short-term production scheduling is a natural application area of online optimisation. Modelling short-term planning is a challenge as there is a high number of variables, exceptions, and normal variation. Simple manual solutions are often used in practice as a solution for online problems.

4.1 Conclusions

The papers of the thesis consider the online optimisation of short-term production planning. First, Publication I contains the results of a survey which focused on the short-term production planning problems in different companies. It seems that workforce coordination is one of the common online problems in the industry.

Bearing this in mind, Publication II studies reactive workforce coordination in the case of parallel stations and arriving jobs. The results suggest that pair working might be enough to reduce the effects of arriving jobs. Complete helping where all other workers help all workers is not needed.

Publication III focuses on combined periodical and reactive control of multi-item production. This control approach is used in the industry practice, and the paper concludes that it works quite well if the inventory control settings are correct and the urgent job approach is not used often. On the other hand, the results show that if one uses the urgent job approach often, it can lead to inefficiency.

Publication IV considers an assemble-to-order system and studies how the component-product structures affect small changes in the demand, component deliveries, and production. The results show that it is better to use an online

scheduling approach when dealing with changes in demand and production limits than with component losses.

Publication V considers a purely dynamic and reactive situation on a production line with online optimisation and restrictions on how rescheduling occurs. The paper shows that the clever implementation of online scheduling can easily save computational power and the stability of the production while still improving the results.

Publication VI study scheduling in the context of shipbuilding, an example of one-of-a-kind production. Publication VI creates a scheduling model for shipbuilding and [84] extends it to an online scheduling model, to a case where the blocks are late. The results show how the use of different planning methods in the case of delays depends heavily on the costs of delaying the ship and the structure of the ship.

The publications in this thesis show that short-term planning is a potential area for optimisation. The general approach used in practice, the use of simple priority rules, is not optimal, especially if it is used often. Modelling the practical online optimisation problems also helps in gaining an understanding of the situation and finding the best solution for the current situation.

4.2 Further work

This thesis cannot cover every aspect of online optimisation in production planning. There are many fruitful areas that are omitted. One of the biggest problems in practice is how the data can be obtained so that it is clean and without any errors. Without correct data online optimisation cannot be fully executed. It might be that it is impossible to make the data completely clean and thus it could also be interesting to study how errors in the data affect the performance of online optimisation. How can a robust solution be implemented that behaves well even in the case of erroneous data?

In this thesis, different systems are studied theoretically, without any application that has been tried on the shop floor. Although this theoretical approach shows the value of online optimisation, practical applications should be used to find out if the value of online optimisation can be fulfilled in manufacturing companies.

Theoretical studies could also be extended. Robust scheduling solutions are needed. It could be interesting e.g. to study situations where processing times are just estimates. Another potential area for theoretical study could be the use of simple stochastic distributions that enable analytical results.

References

[1] Syrus, P. (1856). *The Moral Sayings of Publius Syrus: a Roman Slave. From the Latin.* LE Bernard & Company.

[2] Grötschel, M., Krumke, S. O., Rambau, J., Winter, T., & Zimmermann, U. T. (2001). Combinatorial online optimisation in real time. In M. Grötschel, S. O. Krumke & J. Rambau (Eds.), *Online Optimisation of Large Scale Systems*. Berlin, Springer-Verlag.

[3] Borodin, A. & El-Yaniv, R. (1998). *Online computation and competitive analysis* (Vol. 53). Cambridge, Cambridge University Press.

[4] Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2009). Lectures on stochastic programming: modeling and theory (Vol. 9). *Society for Industrial and Applied Mathematics*.

[5] Marlin, T. E. (1995). *Process control: designing processes and control systems for dynamic performance* (Vol. 2). New York, McGraw-Hill.

[6] Kantorovich, L. V. (1939). Mathematical methods in the organisation and planning of production. *Management Science*, 6(4, Jul., 1960), 366-422.

[7] Dantzig, G. B. (1948). *Programming in a linear structure.* Washington, D.C, USAF.

[8] Dantzig, G. B. (1951). Maximization of a linear function of variables subject to linear inequalities. Chapter 21 in T.C. Koopmans (Eds.), *Activity Analysis of Production and Allocation.* New York, John Wiley and Sons.

[9] Khachiian, L. G. (1979, February). Polynomial algorithm in linear programming. *In Akademiia Nauk SSSR, Doklady* , *244*, 1093-1096.

[10] Land, A. H. & Doig, A. G. (1960). An automatic method of solving discrete programming problems. *Econometrica*, *28*(3), 497-520.

[11] Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, *13*(5), 533-549.

[12] Glover, F. (1989). Tabu search—part I. *ORSA Journal on Computing*, *1*(3), 190-206.

[13] Glover, F. (1990). Tabu search—part II. *ORSA Journal on Computing*, *2*(1), 4-32.

[14] Kirkpatrick, S. & Vecchi, M. P. (1983). Optimisation by simulated annealing. *Science*, *220*(4598), 671-680.

[15] Cerny, V. (1985). Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm. *Journal of Optimisation, Theory and Applications*, *45*(1), 41-51.

[16] Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor, MI, University of Michigan Press.

[17] Granville, V., Krivánek, M., & Rasson, J. P. (1994). Simulated annealing: A proof of convergence. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, *16*(6), 652-656.

[18] Massotte, P. (1997). Analysis and approaches for the management of complex production systems. In A. Artiba & E.E. Elmaghtby (Eds.), *The Planning and Scheduling of Production Systems*, London, Chapman & Hill.

[19] Grötschel, M., Krumke, S. O., & Rambau, J. (2001). *Online Optimisation of Large Scale Systems: State of the Art*. Berlin, Springer-Verlag.

[20] Ausiello, G., Allulli, L., Bonifaci, V., & Laura, L. (2006). On-line algorithms, real time, the virtue of laziness, and the power of clairvoyance. In J. Cai, S.B. Cooper, A. Li (Eds.), *Proc. 3rd Int. Conf. on Theory and Applications of Models of Computation, Lecture Notes in Computer Science* (vol. 3959), Berlin, Springer.

[21] Ascheuer, N., Grötschel, M., Krumke, S. O., & Rambau, J. (1999). Combinatorial online optimisation. *In Operations Research Proceedings 1998*. Springer, Berlin.

[22] Bitran, G. R. & Tirupati, D. (1993). Hierarchical production planning. *Handbooks in Operations Research and Management Science*, *4*, 523-568.

[23] Hax, A. C. & Meal H. C. (1975). Hierarchical Integration of Production Planning and Scheduling. In M. A. Geisler (Eds.), *Studies in Management Sciences* (Vol. I). New York, North Holland-American Elsevier.

[24] Ford, H. (1923). *My Life and Work*, London, William Heinemann.

[25] Muther, R. (1944). *Production-Line Technique*, New York, McGraw-Hill Book Company.

[26] Baudin, M. (2002). *Lean assembly: the nuts and bolts of making assembly operations flow*. New York, Productivity Press.

[27] Hayes, R. H. & Wheelwright, S. C. (1979). The dynamics of process-product life cycles. *Harvard Business Review*, *57*(2), 127-136.

[28] Orlicky, J. (1975). MRP: Material Requirements Planning. *The New Way of Life in Production and Inventory Management*, New York, McGraw Hill.

[29] Ohno, T. (1988). *Toyota production system: beyond large-scale production*. Cambridge, Productivity Press.

[30] Ohno, T. & Mito, S. (1988). *Just-in-time for today and tomorrow*. Cambridge, Productivity Press.

[31] Womack, J. P., Jones, D. T., & Roos, D. (1991). *The machine that changed the world: The story of lean production*. New York: HarperPerennial.

[32] Goldratt, E. M. (1990). *Theory of constraints*. New York, North River.

[33] Deleersnyder, J. L., Hodgson, T. J., Muller-Malek, H., & O'Grady, P. J. (1989). Kanban controlled pull systems: an analytic approach. *Management Science*, *35*(9), 1079-1091.

[34] Spearman, M. L., Woodruff, D. L., & Hopp, W. J. (1990). CONWIP: a pull alternative to kanban. *The International Journal of Production Research*, *28*(5), 879-894.

[35] Holland, J. H. (1985). Properties of the bucket brigade. *In Proceedings of the 1st International Conference on Genetic Algorithms*. Hillsdale, NJ, Lawrence Erlbaum.

[36] Erlang, A. K. (1909). The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B*, *20*(6), 87-98.

[37] Markov, A. A. & Nagornyi, N. M. (1984). *Theory of Algorithms* (in Russian). Moscow.

[38] Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, *24*(3), 338-354.

[39] Kingman, J. F. C. (1961, October). The single server queue in heavy traffic. *In Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 57, No. 04). Canbridge, University Press.

[40] Pollaczek, F. (1930). Über eine Aufgabe der Wahrscheinlichkeitstheorie. *I. Mathematische Zeitschrift*, 32(1), 64-100.

[41] Khintchine, A. Y. (1932). Mathematical Theory of a Stationary Queue. *In Matematicheskii Sbornik*, Vol. 39 (1932) Nr. 4, S. 73-84.

[42] Hopp, W. J. & Spearman, M. L. (2008). *Factory physics* (Vol. 2). New York: McGraw-Hill/Irwin.

[43] Shanthikumar, J. G., Shengwei, D., & Zhang, M. T. (2007). Queueing theory for semiconductor manufacturing systems: a survey and open problems. *IEEE Transactions on Automation Science and Engineering*, *4*(4), 513-522.

[44] Little, J. D. (1961). A proof for the queuing formula: L= λW. *Operations Research*, *9*(3), 383-387.

[45] Cachon, G. & Terwiesch, C. (2006). *Matching supply with demand* (Vol. 2). New York: McGraw-Hill.

[46] Deming, W. E. (1986). *Out of the Crisis*. Cambridge, Massachusetts Institute of Technology, Center for Advanced Engineering Study.

[47] Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*, *46* (5), 323-351.

[48] De Morgan, A. (1872). *A budget of paradoxes*. London, Longmans, Green.

[49] Tokola H. & Niemi E. (2013). Estimating Short-term Production Planning Challenges in Multi-item Production*, In Proceedings of the 22nd International Conference of Production Research*, Iguassu Falls, Brazil.

[50] Graves, S. C. (1999). A single-item inventory model for a nonstationary demand process. *Manufacturing & Service Operations Management*, *1*(1), 50-61.

[51] Ashby, W. R. (1955). *An introduction to cybernetics*. London, Chapman and Hall.

[52] Brooks Jr, F. P. (1987). No silver bullet-essence and accidents of software engineering. *IEEE computer*, 20(4), 10-19.

[53] Graves, S. C. (2011). Uncertainty and production planning. *Planning Production and Inventories in the Extended Enterprise*, 151, 83-101.

[54] Vieira, G. E., Herrmann, J. W., & Lin, E. (2003). Rescheduling manufacturing systems: a framework of strategies, policies, and methods. *Journal of Scheduling*, *6*(1), 39-62.

[55] Herrmann, J. W. (2006). Rescheduling strategies, policies, and methods. In J.W. Herrmann (Eds.) *Handbook of Production Scheduling*. New York, Springer Science+Business Media, Inc.

[56] Sabuncuogly, I. & Bayiz, M. (2000). Analysis of reactive scheduling problems in a job shop environment. *European Journal of Operational Research*, *126*(3): 567-586.

[57] Pinedo, M. (1985). A note on stochastic shop models in which jobs have the same processing requirements on each machine. *Management Science*, *31*(7), 840-846.

[58] Branke, J. & Mattfeld, D. (2005). Anticipation and flexibility in dynamic scheduling. *International Journal of Production Research, 43* (15), 3103-3129.

[59] Matsuura, H., Tsubone, H., & Kanezashi, M. (1993). Sequencing, dispatching, and switching in a dynamic manufacturing environment. *International Journal of Production Research*, *31*(7), 1671-1688.

[60] Baudin, M. (1990). *Manufacturing systems analysis with application to production scheduling*. Englewood Cliffs, NJ, Yourdon Press.

[61] Verderame, P. M., Elia, J. A., Li, J., & Floudas, C. A. (2010). Planning and scheduling under uncertainty: a review across multiple sectors. *Industrial & Engineering Chemistry Research*, *49*(9), 3993-4017.

[62] Dantzig, G. B. (1955). Linear programming under uncertainty. *Management Science*, *1*(3-4), 197-206.

[63] Charnes, A., Cooper, W. W., & Symonds, G. H. (1958). Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. *Management Science*, *4*(3), 235-263.

[64] Aytug, H., Lawley, M. A., McKay, K., Mohan, S., & Uzsoy, R. (2005). Executing production schedules in the face of uncertainties: A review and some future directions. *European Journal of Operational Research*, *161*(1), 86-110.

[65] Herroelen, W. & Leus, R. (2004). Robust and reactive project scheduling: a review and classification of procedures. *International Journal of Production Research*, *42*(8), 1599-1620.

[66] Forrester, J. W. (1961). *Industrial Dynamics* (Vol. 2), Cambridge, MA: MIT Press.

[67] Hopp, W. J., & Oyen, M. P. (2004). Agile workforce evaluation: a framework for cross-training and coordination. *IIE Transactions*, *36*(10), 919-940.

[68] Edgeworth, F. Y. (1888). The mathematical theory of banking. *Journal of the Royal Statistical Society*, *51*(1), 113-127.

[69] Harris, F. W. (1915). Operations and cost. *Factory management series*, 48-52.

[70] Zipkin, P. H. (2000). *Foundations of inventory management* (Vol. 2). New York, McGraw-Hill.

[71] Sobel, M. J. (2008). Risk pooling. In Chhajed, D. & Lowe, T. J. (Eds.), *Building Intuition*. New York, Springer.

[72] Benjaafar, S., Cooper, W. L., & Kim, J. S. (2005). On the benefits of pooling in production-inventory systems. *Management Science*, *51*(4), 548-565.

[73] Taymaz, E. (1989). Types of flexibility in a single-machine production system. *The International Journal of Production Research*, 27(11), 1891-1899.

[74] Song, J. S. & Zhao, Y. (2009). The value of component commonality in a dynamic inventory system with lead times. *Manufacturing & Service Operations Management*, *11*(3), 493-508.

[75] Song, J.-S. & Zipkin, P. (2003). Supply chain operations: Assemble-to-order systems. *Handbooks in Operations Research and Management Science*, 11, 561-596.

[76] Gupta, J. N. & Stafford Jr, E. F. (2006). Flowshop scheduling research after five decades. *European Journal of Operational Research*, *169*(3), 699-711.

[77] Graham, R. L. (1966). Bounds for certain multiprocessing anomalies. *Bell System Technical Journal*, *45*(9), 1563-1581.

[78] Bean, J. C., Birge, J. R., Mittenthal, J., & Noon, C. E. (1991). Matchup scheduling with multiple resources, release dates and disruptions. *Operations Research*, *39*(3), 470-483.

[79] Zweben, M., Davis, E., Daun, B., & Deale, M. J. (1993). Scheduling and rescheduling with iterative repair. *IEEE Transactions on Systems, Man and Cybernetics*, *23*(6), 1588-1596.

[80] Masuchun, R. & Ferrell Jr, W. G. (2004). Dynamic rescheduling with stability. *In Proceedings of the 5th Asian Control Conference* (Vol. 3). IEEE.

[81] Sadeh, N., Otsuka, S., & Schnelbach, R. (1993, August). Predictive and reactive scheduling with the Micro-Boss production scheduling and control system. *In Proceedings of the IJCAI-93 Workshop on Knowledge-Based Production Planning, Scheduling and Control.*

[82] Abumaizar, R. J. & Svestka, J. A. (1997). Rescheduling job shops under random disruptions. *International Journal of Production Research*, *35*(7), 2065-2082.

[83] Raheja, A. S. & Subramaniam, V. (2002). Reactive recovery of job shop schedules–a review. *The International Journal of Advanced Manufacturing Technology*, *19*(10), 756-763.

[84] Tokola, H., Niemi, E., & Remes, H. (2014). Block erection in the event of delays in shipbuilding: a scenario-based approach. *Submitted manuscript*, 42 pages.

[85] Kelley Jr, J. E. & Walker, M. R. (1959, December). Critical-path planning and scheduling. In *Papers presented at the December 1-3, 1959, Eastern Joint IRE-AIEE-ACM Computer*. ACM.

[86] Kelley, J. E. (1961). Critical-path planning and scheduling: Mathematical basis. *Operations Research*, *9*(3), 296-320.

[87] McKay, K. N. & Wiers, V. C. (2004). *Practical production control: a survival guide for planners and schedulers*. Boca Raton, Florida, J. Ross Publishing.

Publications

Additional comments on the publications

Publication I:

1. Page 52, column 1, last paragraph. Subcontractors are such small companies that have less than 250 workers. This was the case with all studied subcontractors. Only original equipment manufacturers had more than 250 workers.
2. Page 55, column 1, fourth paragraph: In engineer-to-order companies, the production of parts is sometimes started before the whole product is designed. This is done to get the product finished before deadline.
3. Page 57, column 1, second paragraph: Results of the paper are mainly expected and many of them appear in the other publications as well. However, it is worthwhile to validate that practice follows expectations.
4. Page 57, column 2, last paragraph: In LeanMES project, the aim is to develop MES tools for small and medium sized companies. Based on the results of the paper, new production tools for them should include forecast methods to couple seasonal trends, tools to schedule extra machines and plans for cross training of workers.

Publication II:

1. In the future, the models can be modified so that non-identical workers are considered.
2. In practice, floater is often implemented so that he is an extra worker without own station. Different types of floaters could be studied in the future.

Publication III:

1. Page 2, Column 2, last paragraph: In pull systems, the idling of the factory can be organised using a visual system such as Kanban.
2. Section 3.1: The way different strategies work is described in detail in Chapter 2. In general, a sequence of production is created at the beginning of the simulation so that the priority of $(\beta_i + J_i)/\lambda_i$ is used to select the next job in the sequence. J_i denotes the number of i product in the existing sequence. In the simulation, the production follows the sequence and produces items to the inventory. Poisson distribution is used to pull products from the inventory. If $\beta_i < r_i$, both an urgent job approach and complete replanning strategies change the sequence. In

the urgent job approach, if the inventory of a product falls below its threshold, an order for that job is added into the sequence. In the case of complete replanning, the whole sequence is constructed again for the current situation if the inventory of a product falls below its threshold.

3. Section 5: In future, the possibility of different distributions for planning times could be studied.

Publication IV:

1. Page 1, Abstract: The fact that "The combination of component loss with demand increase gives the highest costs" is usually expected.
2. Page 3, column 1, line 6: X defines the cost type and Y specifies the unexpected even types.
3. Page 5, column 2, line 5: Regarding "Demand increase (DI) yields low costs in both sparse and dense ATO networks (DI)", it does not hold if the demand increase is combined e.g. with production limit.

Publication V:

1. Page 2, Equation (1): j should be equal to 1 instead of 0.
2. Page 4, after the sentence "the second task of job1 is scheduled to start after the second task of job0": This schedule appears in Figure 1c, but in other figures the situation has been rescheduled.
3. Page 6: The paper considers similar processing times in each of the machines. However, in practice, it is common that there is e.g. a bottleneck machine, which has longer processing times than other machines.
4. The maximum number of machines in this publication is four. It would be interesting to study the situation with a larger number of machines.

Publication VI:

1. In relation to the publication, objectives such as lead time and tardiness could be studied in the future.

In today's manufacturing world, real-time information is available or soon will be. Manufacturing companies can use it as a part of their information systems (including e.g. material requirements planning (MRP), enterprise resource planning (ERP), and manufacturing execution systems (MES)) for controlling the production system, i.e. for the adjustment of their inventory level, defining the capacity, and even scheduling the starting times of the jobs. A method that can be used to improve the decisions dynamically is online optimisation. Online optimisation repeatedly optimises and adjusts the decisions when there are changes or disturbances in the system. Online optimisation differs from traditional optimisation. First, the data is uncertain but, luckily, the uncertainty might decrease as time goes by. Second, quick decisions are needed but they should be made carefully as the decisions will also affect future decisions.

Aalto-DD 162/2015

ISBN 978-952-60-6485-7 (printed)
ISBN 978-952-60-6486-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Engineering
Department of Engineering Design and Production
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS