# New approaches for large scale multi-step production planning problems

Sébastien Beraudy

N°d'ordre NNT : 2020LYSEM015

# THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de
**l'Ecole des Mines de Saint-Etienne**

**Ecole Doctorale** N° 488
**Sciences, Ingénierie, Santé**

**Spécialité de doctorat** : Génie Industriel

Soutenue publiquement le 29/09/2020, par :
**Sébastien Beraudy**

---

# NEW APPROACHES FOR LARGE SCALE MULTI-STEP PRODUCTION PLANNING PROBLEMS

# NOUVELLES APPROCHES POUR LES PROBLEMES DE PLANIFICATION DE LA PRODUCTION MULTI-ETAPES DE GRANDE DIMENSION

---

Devant le jury composé de :

| | | |
|---|---|---|
| Kedad-Sidhoum, Safia | Professeur, CNAM | Présidente |
| Billaut, Jean-Charles | Professeur, Université de Tours | Rapporteur |
| Labadie, Nacima | Professeur, Université de Technologie de Troyes | Rapporteuse |
| Aggoune, Riad | Docteur, Luxembourg Institute of Science and Technology | Examinateur |
| Gicquel, Céline | Maître de conférence, Université Paris Sud | Examinatrice |
| Dauzère-Pérès, Stéphane | Professeur, Mines Saint-Etienne | Directeur de thèse |
| Absi, Nabil | Professeur, Mines Saint-Etienne | Co-directeur de thèse |
| Jouvenot, Patrick | Ingénieur, STMicroelectronics | Invité |
| Vermarien, Léon | Ingénieur, STMicroelectronics | Invité |

| Spécialités doctorales | Responsables : | Spécialités doctorales | Responsables |
|---|---|---|---|
| SCIENCES ET GENIE DES MATERIAUX | K. Wolski Directeur de recherche | MATHEMATIQUES APPLIQUEES | O. Roustant, Maître-assistant |
| MECANIQUE ET INGENIERIE | S. Drapier, professeur | INFORMATIQUE | O. Boissier, Professeur |
| GENIE DES PROCEDES | F. Gruy, Maître de recherche | SCIENCES DES IMAGES ET DES FORMES | JC. Pinoli, Professeur |
| SCIENCES DE LA TERRE | B. Guy, Directeur de recherche | GENIE INDUSTRIEL | N. Absi, Maitre de recherche |
| SCIENCES ET GENIE DE L'ENVIRONNEMENT | D. Graillot, Directeur de recherche | MICROELECTRONIQUE | Ph. Lalevée, Professeur |

## EMSE : Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'Etat ou d'une HDR)

| | | | | |
|---|---|---|---|---|
| ABSI | Nabil | MR | Génie industriel | CMP |
| AUGUSTO | Vincent | CR | Image, Vision, Signal | CIS |
| AVRIL | Stéphane | PR2 | Mécanique et ingénierie | CIS |
| BADEL | Pierre | MA(MDC) | Mécanique et ingénierie | CIS |
| BALBO | Flavien | PR2 | Informatique | FAYOL |
| BASSEREAU | Jean-François | PR | Sciences et génie des matériaux | SMS |
| BATTON-HUBERT | Mireille | PR2 | Sciences et génie de l'environnement | FAYOL |
| BEIGBEDER | Michel | MA(MDC) | Informatique | FAYOL |
| BLAYAC | Sylvain | MA(MDC) | Microélectronique | CMP |
| BOISSIER | Olivier | PR1 | Informatique | FAYOL |
| BONNEFOY | Olivier | PR | Génie des Procédés | SPIN |
| BORBELY | Andras | MR(DR2) | Sciences et génie des matériaux | SMS |
| BOUCHER | Xavier | PR2 | Génie Industriel | FAYOL |
| BRODHAG | Christian | DR | Sciences et génie de l'environnement | FAYOL |
| BRUCHON | Julien | MA(MDC) | Mécanique et ingénierie | SMS |
| CAMEIRAO | Ana | MA(MDC) | Génie des Procédés | SPIN |
| CHRISTIEN | Frédéric | PR | Science et génie des matériaux | SMS |
| DAUZERE-PERES | Stéphane | PR1 | Génie Industriel | CMP |
| DEBAYLE | Johan | MR | Sciences des Images et des Formes | SPIN |
| DEGEORGE | Jean-Michel | MA(MDC) | Génie industriel | Fayol |
| DELAFOSSE | David | PR0 | Sciences et génie des matériaux | SMS |
| DELORME | Xavier | MA(MDC) | Génie industriel | FAYOL |
| DESRAYAUD | Christophe | PR1 | Mécanique et ingénierie | SMS |
| DJENIZIAN | Thierry | PR | Science et génie des matériaux | CMP |
| BERGER-DOUCE | Sandrine | PR1 | Sciences de gestion | FAYOL |
| DRAPIER | Sylvain | PR1 | Mécanique et ingénierie | SMS |
| DUTERTRE | Jean-Max | MA(MDC) | | CMP |
| EL MRABET | Nadia | MA(MDC) | | CMP |
| FAUCHEU | Jenny | MA(MDC) | Sciences et génie des matériaux | SMS |
| FAVERGEON | Loïc | CR | Génie des Procédés | SPIN |
| FEILLET | Dominique | PR1 | Génie Industriel | CMP |
| FOREST | Valérie | MA(MDC) | Génie des Procédés | CIS |
| FRACZKIEWICZ | Anna | DR | Sciences et génie des matériaux | SMS |
| GARCIA | Daniel | MR(DR2) | Sciences de la Terre | SPIN |
| GAVET | Yann | MA(MDC) | Sciences des Images et des Formes | SPIN |
| GERINGER | Jean | MA(MDC) | Sciences et génie des matériaux | CIS |
| GOEURIOT | Dominique | DR | Sciences et génie des matériaux | SMS |
| GONDRAN | Natacha | MA(MDC) | Sciences et génie de l'environnement | FAYOL |
| GONZALEZ FELIU | Jesus | MA(MDC) | Sciences économiques | FAYOL |
| GRAILLOT | Didier | DR | Sciences et génie de l'environnement | SPIN |
| GROSSEAU | Philippe | DR | Génie des Procédés | SPIN |
| GRUY | Frédéric | PR1 | Génie des Procédés | SPIN |
| HAN | Woo-Suck | MR | Mécanique et ingénierie | SMS |
| HERRI | Jean Michel | PR1 | Génie des Procédés | SPIN |
| KERMOUCHE | Guillaume | PR2 | Mécanique et Ingénierie | SMS |
| KLOCKER | Helmut | DR | Sciences et génie des matériaux | SMS |
| LAFOREST | Valérie | MR(DR2) | Sciences et génie de l'environnement | FAYOL |
| LERICHE | Rodolphe | CR | Mécanique et ingénierie | FAYOL |
| MALLIARAS | Georges | PR1 | Microélectronique | CMP |
| MOLIMARD | Jérôme | PR2 | Mécanique et ingénierie | CIS |
| MOUTTE | Jacques | CR | Génie des Procédés | SPIN |
| NAVARRO | Laurent | CR | | CIS |
| NEUBERT | Gilles | | | FAYOL |
| NIKOLOVSKI | Jean-Pierre | Ingénieur de recherche | Mécanique et ingénierie | CMP |
| NORTIER | Patrice | PR1 | Génie des Procédés | SPIN |
| O CONNOR | Rodney Philip | MA(MDC) | Microélectronique | CMP |
| PICARD | Gauthier | MA(MDC) | Informatique | FAYOL |
| PINOLI | Jean Charles | PR0 | Sciences des Images et des Formes | SPIN |
| POURCHEZ | Jérémy | MR | Génie des Procédés | CIS |
| ROUSSY | Agnès | MA(MDC) | Microélectronique | CMP |
| ROUSTANT | Olivier | MA(MDC) | Mathématiques appliquées | FAYOL |
| SANAUR | Sébastien | MA(MDC) | Microélectronique | CMP |
| SERRIS | Eric | IRD | | FAYOL |
| STOLARZ | Jacques | CR | Sciences et génie des matériaux | SMS |
| TRIA | Assia | Ingénieur de recherche | Microélectronique | CMP |
| VALDIVIESO | François | PR2 | Sciences et génie des matériaux | SMS |
| VIRICELLE | Jean Paul | DR | Génie des Procédés | SPIN |
| WOLSKI | Krzystof | DR | Sciences et génie des matériaux | SMS |
| XIE | Xiaolan | PR0 | Génie industriel | CIS |
| YUGMA | Gallian | CR | Génie industriel | CMP |

Mise à jour : 30/08/2018

# Acknowledgments

PhD thesis is not a solitary exercise. Without the implication of the many people quoted below (explicitly or not) this work would not be.

My first thanks are for Stéphane Dauzère-Pérès and Nabil Absi who are contributors of this thesis rather than "simple" supervisors. Thank you Stéphane, firstly, for proposing me the thesis subject, but mostly for pointing out promising path for research and for letting us leech on your vast knowledge. You are one of the strongest hard-worker I have ever seen, able to juggle with many subjects but also to be completely focused. Thank you Nabil, for the help on the algorithms and implementations issues, for your many advises on methodology and also for driving us to the restaurant when the canteen closed (an hungry student would not have been able to work correctly).

In a second time, I would like to thank all the members of my jury. Thank you all for your good questions that make me realize the value of this thesis, that may be a cornerstone for breakthrough research. I specifically thank Jean-Charles Billaut and Nacima Labadie for accepting to report my manuscript.

I am also grateful to the people of STMicroelectronics for your cooperation and the many discussions we had. They were well represented in my jury and have shed light on my work and directions to follow with their industrial experience. I particularly thank Quentin for being always available for any questions (no matter if it was unimportant or not). The collaboration with STMicroelectronics was part of the project Productive4.0. I was also glad to see the development of such big project with successful collaborations. Inside our team dedicated to this project, I should thank Hamideh and Karim for their support role.

Not last and certainly not least, I am thankful to the CMP staff I have interacted with, they are the reason why this work environment rocks, truly friendly and humane (I admit the region and the sun are also helpful). I have particularly appreciated the ambiance in SFL where interactions are not only work-related (the cake contest day was clearly one of the most unproductive day) and with many smiles if not laughs. I would like to thank in particular Margaux, Sean and Elodie for being my "coinche" partners and much more.

Finally, I will thank my family and the rest of my friends, that would have prefer I came back more often. Instead they had to go in the south to see me (not sure if it is a loose or a win).

# Contents

# General introduction

## General context

Semiconductor is a highly competitive industry which has grown strongly in recent decades due to the increasing demand for microelectronic in all devices (computers, cars, production machines,...). In semiconductor manufacturing, the fierce competition between companies to meet the increasing demands and to develop new products clashes with the investment cost in semiconductor manufacturing facilities (also called fabs). Thus, the management of companies seeks to maximize the use of the available capacity in fabs. To add insult to injury, semiconductor manufacturing probably includes the most complex industrial processes, involving hundreds of operations for each product and cycle times between two and three months. For all these reasons, production planning is critical and should take into account the congestion that can occur in production flows. This is where the thesis contributions lie: Proposing new production planning models relevant for the semiconductor industry, and new approaches to solve these models in short computational times.

## Structure of the thesis

The manuscript is organized as follows. Chapter 1 details the characteristics of the semiconductor industry. A generic perspective of the industry is first given. Then, the discussion focuses on the first part of the semiconductor manufacturing process (also called "Front-end"), where the different issues and factors of complexity are presented. In the last section of the chapter, the data framework designed in the European project Productive 4.0 is introduced and used to describe the data sets considered in our numerical experiments.

Chapter 2 provides a literature review on production planning for the semiconductor industry. This literature review consists of a generic introduction to production planning problems, the main features of production planning in semiconductor manufacturing, a review of the main techniques to model congestion in semiconductor manufacturing and some extensions of production planning problems in semiconductor manufacturing.

Chapter 3 presents a generic production planning model based on the semiconductor manufacturing literature, and studies the question of which is the best objective to achieve from an industrial point of view. Two alternative objectives are proposed: A productivity objective which maximizes the number of operations that are performed and a financial objective which maximizes the profit using the Net Present Value (NPV). The two objectives are compared using computational experiments on academic and industrial instances.

Chapter 4 proposes to explore new flow constraints to allow more flexibility in the production plan. Classical fixed lead time constraints are criticized, because they prevent any flexibility on the internal production flows, i.e. the number of periods that products spend

in each operation is fixed. Thus, flexible lead time constraints are proposed that enable production quantities to wait in an operation. Computational experiments with different lead time profiles on industrial instances are performed to illustrate the relevance of flexible lead times.

Chapter 5 introduces a new formulation of the production planning problem with fixed and flexible lead times, that relies on the concept of timed routes. In a timed route, each operation is allocated to a fixed period. The timed route formulation can be solved by a column generation approach when the number of timed routes becomes too large, which is the case when flexible lead times are considered. Computational experiments in industrial instances show that the computational times are greatly reduced compared to the classical models of Chapter 4. New ways of analyzing the production flows allowed by the timed route formulation are also investigated.

Chapter 6 extends the work of chapter 5, since the new timed route model offers many possible extensions, many of which cannot be taken into, in the classical models of Chapter 4. Minimum and maximum cycle times are modeled, but also alternative objective functions with penalties on production flows whose cycle times differ from a cycle time target or on production flows with too long lead times in an operation.

Finally, in Chapter 7, conclusions are drawn and perspectives for future work are open, in particular extensions of the new timed route formulation introduced in this thesis, such as a more detailed modeling of the demand or modeling the entire supply chain instead of a single factory.

# Chapter 1

# Industrial context

## 1.1 Introduction

The semiconductor industry is probably the most complex and competitive industry. This industry must deal with continuously increasing demand for very complex products, Integrated Circuits (ICs), while continuously improving the performance of its products and manufacturing. This industry is described in Section 1.2, with a global view of its supply chain. Then, Section 1.3 focuses on the manufacturing part of the supply chain known as the "Front-End" where most of the complexity and added values lie. Section 1.4 introduces a formalization of the generic data model developed in the framework of the European project Productive4.0, and describes the format and characteristics of the data sets used in the numerical experiments of this thesis.

## 1.2 Semiconductor industry

### 1.2.1 Semiconductor market

The semiconductor industry refers to the making of electronic components based on semiconductor materials. It mainly concerns the fabrication of Integrated Circuits, but it can be extended to products such as LEDs. Integrated Circuits (ICs) or chips are products heavily used in our daily life (e.g. smartphones, computers, televisions, cars), but also within industrial machines used in most factories or even in medical equipment. The current trending application is the Internet of Things (IoT), connected devices which share information through the Internet. The boom of IC demands directly depends on the increase of the electronic market and the use of electronic devices in our everyday life. To sense the economic impact of the semiconductor industry, Figure 1.1 shows the annual worldwide semiconductor revenues (in US dollars) from 1990 to 2019. Starting at the beginning of the century, the global yearly revenue largely increased from 150 billion to 400 billion. Of course, the semiconductor market does not always increase, and can suffer from economical hazards. But still, one of the major issues of the semiconductor industry is to answer a constantly increasing demand.

A critical characteristic of the semiconductor industry is the continuous technological improvement. As prophesied by Gordon Moore (former CEO of Intel), the number of transistors in integrated circuits approximately doubles every two years. In fact, it is a self-realizing prophecy because it has served to R&D departments as a goal to reach. However, nowadays,

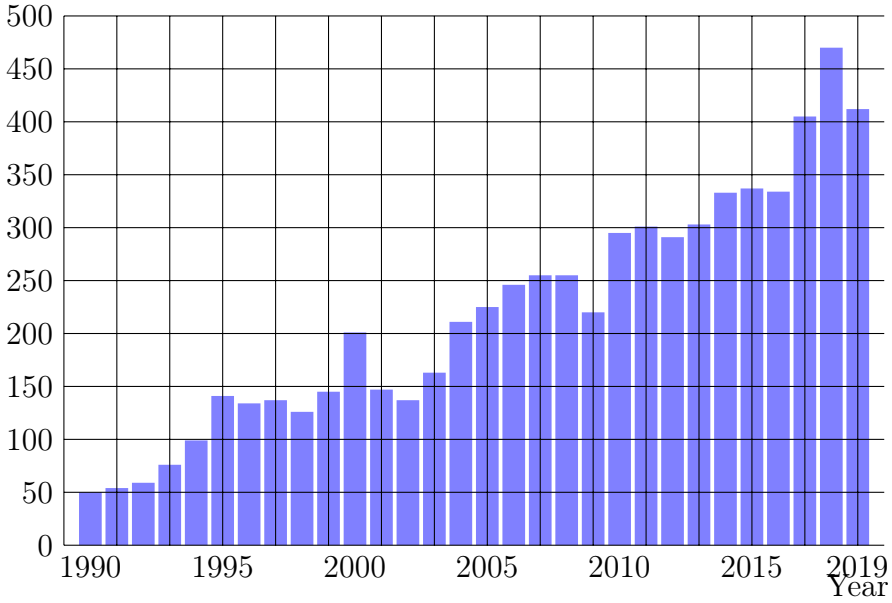



Figure 1.1: Worldwide revenue (in billions of US dollars) of semiconductor companies (data from Semiconductor Industry Association)

the technological improvement can no longer follow Moore's law. In fact, it is since the last decade that the industry has acknowledged that Moore's law is no longer applicable due to the scale side effects on transistors (around 10 nanometers). But other ways have been found to circumvent the issue, for example by developing multi-core chips.

Another major point impacting the semiconductor industry is the cost of its equipment. Machines are worth between tens of thousands to several million dollars. Machines are rarely fully stopped, and thus impose significant fixed electricity costs. Furthermore, integrated circuits are manufactured in clean rooms (with very low density of dust particles, to avoid altering of the product), which adds a significant cost to the costs of operating a factory. Due to these fixed costs, semiconductor companies aim at running their equipment at full capacity. Another way to cope with these very large fixed costs is the economy of scale. The semiconductor industry has evolved to produce more integrated circuits by using larger wafers of silicon (basic components on which hundreds of ICs are produced). The last way to reduce costs (or to improve benefits) is in the management of operations, to which this thesis intends to contribute. Another consequence of having very expensive machines is the coexistence of old facilities, called "legacy fabs" (usually with wafers of 200 mm), and facilities with the most recent machines, called "modern fabs" (usually with wafers of 300 mm).

Finally, semiconductor manufacturers are commonly divided into two categories. The first one, called High Volume/Low Mix (HVLM) and usually corresponding to large companies such as Intel and Samsung, mass-produce a small variety of products, and thus naturally benefit from economies of scale. On the opposite, in the second category called Low Volume/High Mix (LVHM), companies are manufacturing a large variety of specialized products, many in small quantities. The major European semiconductor manufacturing companies (STMicroelectronics, Bosch and Infineon) are in the second category, and thus have to deal with very complex challenges. The goal of advanced production management approaches, such as the ones proposed in this thesis in the context of the European project

Productive4.0, is to allow LVHM companies to be as effective and efficient than HVLM companies.

### 1.2.2 Semiconductor supply chain

Let us consider the semiconductor supply chain process shown schematically in Figure 1.2. It starts with silicone ingots, which are cut into raw wafers. These raw wafers are processed in a manufacturing facility, called "wafer fab", where layers of resistors and transistors on a wafer form hundreds or thousands of ICs. Finished wafers are tested to make an electronic map of the defective dies. The "probing", as it is called, can be performed in the same wafer fab or in another facility. Finally, wafers are diced, and the smaller parts are called "dies". The defective dies are discarded and the good ones are either sent to a die bank or directly dispatched to assembly facilities, where the dies are packaged to become the final chips. A final test is performed to determine the quality and the performance of the chips before sending them to customers. The average cycle time for an integrated circuit is 2.5 months, most of the time being spent in the wafer fab.

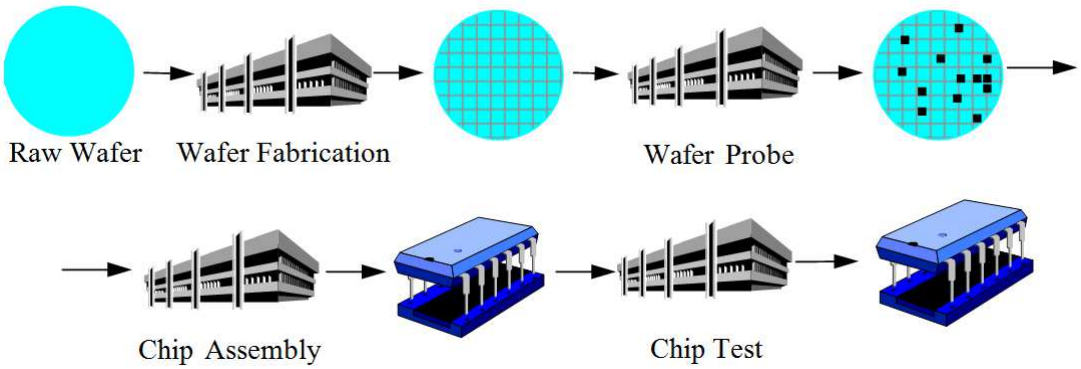



Figure 1.2: Overview of a semiconductor manufacturing Supply Chain (Schömig and Fowler (2000))

The wafer manufacturing part of the supply chain is called "Front End". The second part of the supply chain is called "Back End". Note that, nowadays, the frontier between Front End and Back End is more blurred than defining as the "probing" stage. To add more complexity in the vocabulary, even within the wafer manufacturing process, the first part of the process is called the Frond End Of the Line (FEOL) and the second part is called the Back End Of the Lines (BEOL). European semiconductor manufacturing companies have Front End facilities still located in Europe, contrary to most Back End facilities which are located in countries with low labor rates (e.g. Marocco, Singapore and China). For a semiconductor manufacturing company, the cost of transporting the wafers is very small in the overall cost of the products. Wafers are often transported by plane to decrease cycle times, which can result in products, from raw wafers to finished integrated circuits, traveling multiple times between continents.

In a semiconductor supply chain, contracting with external facilities is usual to meet the demands on non-critical technologies. Because the most saturated capacities are those of wafer fabs, it is generally the front end process which is outsourced. However, outsourcing can also be done in the Back End, where intellectual property and technology are not as critical as in the Front End.

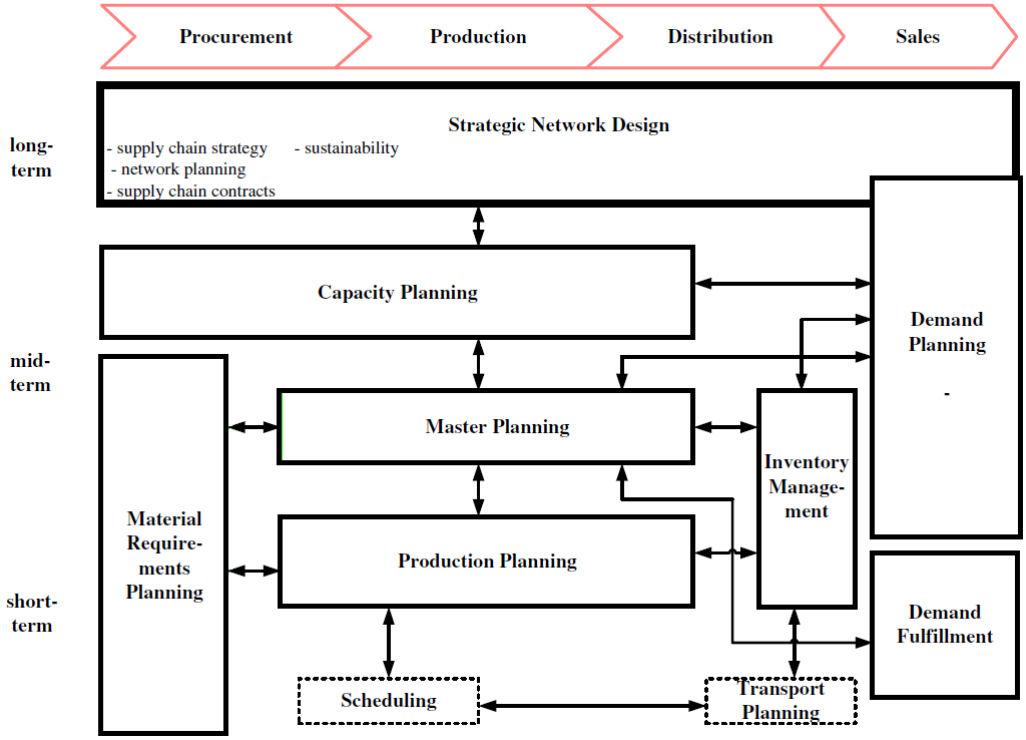



Figure 1.3: Planning decisions in a semiconductor manufacturing supply chain

In such a supply chain, multiple levels of planning decisions are required to manage the entire industry. In Figure 1.3, a flowchart from Mönch et al. (2018a) describes the planning process at the operational, tactical and strategic levels of the semiconductor industry hierarchy. At the strategic level, there is the design of the supply chain network (with decisions such as the acquisition of new factories). Then, by interfacing the strategic and tactical levels, there is the capacity planning where the decision to buy new machines to increase the capacities of the installations is made. Master planning and production planning are quite linked as indicated by Mönch et al. (2018b) but, while production planning aims to determine the quantities to be released in order to meet demand at wafer fab level, master planning does the same with several installations taking into account the entire supply chain. Master planning can be located between the strategic and tactical levels. Production planning is considered to be at the tactical level, but needs relevant inputs from the operational level (as it will be shown in Section 1.3). Sometimes, semiconductor manufacturing companies introduce an intermediate planning stage, called "operational production planning" between production planning and production scheduling, to ensure the feasibility of production plans (see for example Christ et al. (2018)). At the operational level, decisions on the scheduling of the product and on the machines assignment are taken, and also the transportation of the products between machines. Material requirements and demand planning are linked to several levels of the hierarchy, because they can be used with several levels of aggregation. This thesis, in particular, focuses on production planning at the tactical level.

Within this organization, the many junctions between the different levels are critical. While it is generally "easy" to disseminate decisions from the upper level to the lower level (the hard part is ultimately that the lower level meets expectations), getting the information back to the higher level could be difficult. However, it is essential to ensure that the higher-level plans are realistic. Communication between different organizations and integration of

the models used to take decisions are essential to improve the coordination of the whole company.

## 1.3 Semiconductor manufacturing issues

This section focuses on the Front End part of the supply chain. A short description of the processes involved completing a raw wafer is given. Next, the environment in which these processes take place is described, as well as the various factors that complicate the management of the wafer fab and a non-exhaustive list of important questions to be addressed.

### 1.3.1 Characteristics of a wafer manufacturing facility

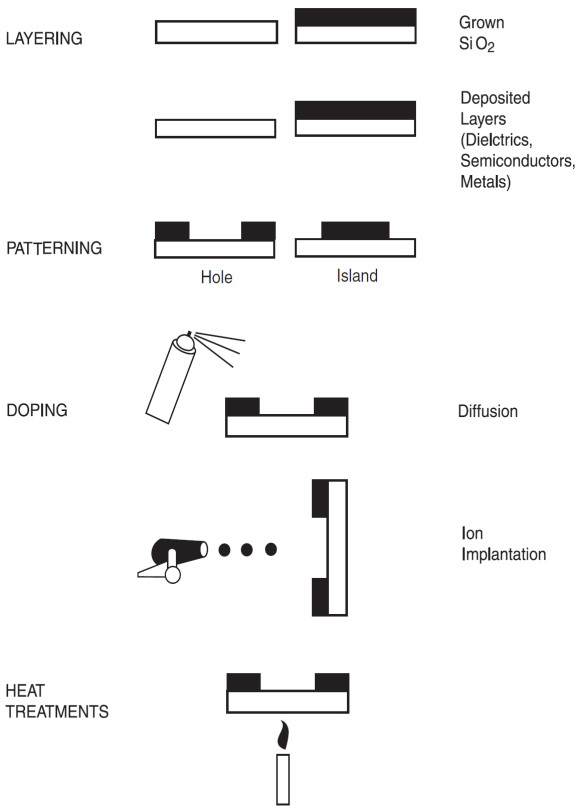



Figure 1.4: Basic operations in wafer manufacturing (from Van Zant (2004))

Integrated circuits are made up of nanoscopic transistors and resistors and, to build such small components, precise operations are carried out. Figure 1.4 shows in a simple way the basic operations to build a chip. Layering and patterning are means of superimposing different materials and creating electronic patterns, while doping and ion implantation changes the electronic structure of the material. These operations are repeated many times with a specific recipe (how to perform the operation) for each processing step. Readers wishing to learn more about the physical processes involved in wafer manufacturing are referred to Van Zant (2004).

Figure 1.5 illustrates the structure and the flows in a wafer manufacturing facility. The main "objects" used to characterize wafer manufacturing are defined below.

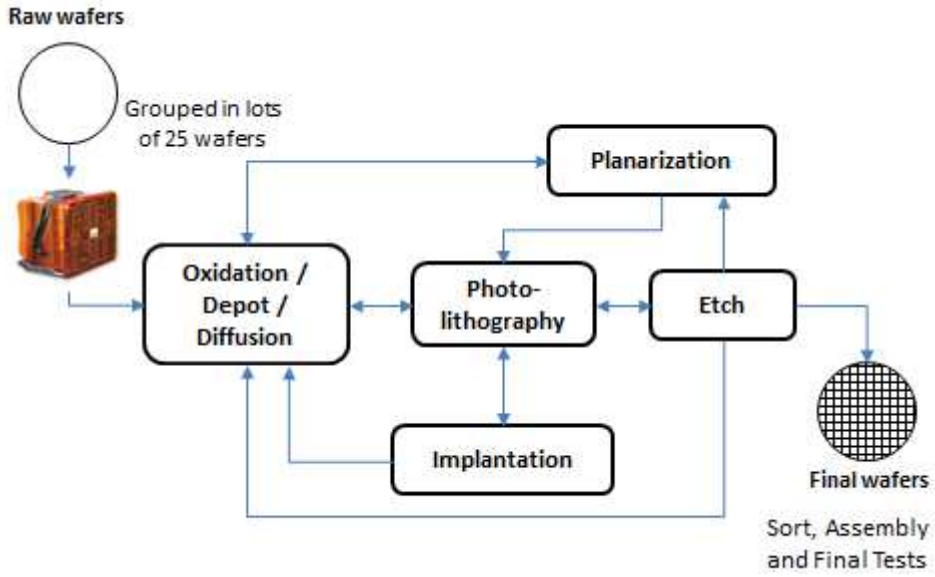



Figure 1.5: Workshops and flows in a wafer manufacturing facility (Dauzère-Pérès, 2011)

- **Product**. It refers, at the lowest level, to a single type of wafers, associated with a single production route. It can be aggregated with other products of the same "family" or "technology", depending on the level of aggregation. The physical representation of a product is the wafer, which is usually grouped in a FOUP (Front Opening Unified Pod, the red box in figure 1.5) of 25 wafers of the same product.

- **Route.** It corresponds to a sequence of operations, or processing steps needed to complete a wafer. Note that, if routes are usually linear, some parts of a route may have alternative paths leading to similar products. An expected cycle time (time to transform a raw wafer into a finished product) is usually associated with each route. The design of routes and operations is where most of the intellectual property resides in semiconductor manufacturing.

- **Operation.** Operations or processing steps are the basic components of production routes. To process an operation, a specific recipe should be followed, which details the process parameters (such as gas pressure, temperature, time), the type of machines that are required and sometimes auxiliary resources. With an operation is also associated an expected Lead Time (which include the transportation time, waiting time and processing time). Processing times range from a few minutes to several hours.

- **Machine.** Because of their costs, machines are very important in semiconductor manufacturing. They are usually flexible, i.e. they can process numerous kinds of operations. They can require a human operator or be fully automated. Some machines can process products in batches (batching machines) or in different chambers (cluster machines). With a machine is associated a dynamic list of operations which can be processed on the machine, i.e. operations that the machine is "qualified" to process. Qualifying a new operation on a machine might require a significant unique set-up time.
  Machines of the same type are aggregated into workshops, as shown in Figure 1.5 that does not include some workshops such as metrology and test.

### 1.3.2 Complexity factors

This section lists the major factors explaining the complexity associated with semiconductor manufacturing.

1. **Large problem dimensions.** A European Front End manufacturing facility usually has a portfolio of thousands of products. For each product, hundreds of process operations are needed in a facility that includes hundreds of machines. The weekly output of a facility is larger than several thousands of completed wafers. In this context, solving discrete optimization problems where the entire facility is considered is usually unrealistic.

2. **Re-entrant flows.** The very large number of process operations is due to the fact that wafers go through the same workshop many times (e.g. more than 40 times in the photolithography workshop). Hence, not only different products at the same stage of their manufacturing process compete for the same capacity, but also products that are at different stages. Managing this competition is critical and difficult, and modeling the associated congestion is important in production planning and in this thesis.

3. **Aggressive production targets.** Another reason for congestion is related to the huge cost of machines which should thus work at full capacity. When capacity is saturated, any incident can cause delays. Management also encourages shortening cycle times, which are quite long (from two to three months). But due to Little's Law, these two objectives can be contradictory because one way to reduce cycle times is to reduce Work In Process (WIP) inside the facility.

4. **Continuous complexity increase in Front End manufacturing.** The complexity of Front End manufacturing is continuously increasing due to the introduction of more complex products that require more operations to be completed. Moreover, some products are personalized in their routes, or are merged with other products. This leads to a more complex Bill-Of-Materials (BOM) than just sequences of operations. Also, the number of time constraints between non-consecutive operations in the route of new products is increasing, up to several thousand altogether.

5. **Auxiliary resources.** The need for critical auxiliary resources to perform some operations imposes the planning to take into account some additional delays. The primary example is the photolithography workshop where masks (or reticles) are needed to make specific patterns on the wafers. Masks are not always readily available and must be brought from different storage areas, which can sometimes be time consuming. Other critical auxiliary resources are operators that are required for various operations in the oldest semiconductor manufacturing factories.

All these features of semiconductor manufacturing are depicted in Mönch et al. (2012), with other features such as the sequence-dependent set-up times.

### 1.3.3 Issues arising in Front End manufacturing

After discussing major complexity factors, let us present a non-exhaustive list of issues in semiconductor manufacturing.

- **Modernization of facilities.** The main concern of legacy fabs (oldest fabs, mainly 200mm fabs) is to keep state-of-the-art processes. This is in particular done through modernization projects such as adding AHMS (Automated Material Handling Systems). Modernization projects can be lengthy due to the desire of minimizing the disruption of manufacturing while making significant changes to the production system. Current modernization processes are often under the umbrella of what is called "Industry 4.0".
- **Balancing machine capacity.** An important issue in semiconductor manufacturing is to balance the capacity of machines between different activities: Production, maintenance (to keep the machines up), engineering (to improve process performance) and R&D (to develop future products). Machine capacity should also be balanced by efficiently planning the qualifications of operation to perform on machines.
- **Global fab management and local scheduling.** Scheduling lots of wafers, even at the workshop level, can be difficult, and finding feasible schedules might already greatly help at the operational level. Furthermore, even if production planning determines the quantities to be released in the facility, ensuring consistency between global fab management decisions and local scheduling decisions at the workshop level is challenging because of the multiple objectives to optimize. To cope with this issue, various systems are used. A common one is to assign priorities to lots and use priority-based dispatching rules to schedule lots. Another system consists in defining at the global level production targets, in terms of quantity of each product to produce in a workshop in a period, and to schedule the lots at the local level so that the production targets are satisfied.
- **Production planning.** Semiconductor manufacturing facilities aim at obtaining realistic production plans that take into account the full complexity of the manufacturing processes. As it will be shown in Chapter 2, it is a relevant research field for academics and industries. One reason is that the manufacturing facility must commit to orders it can meet. This requires fast production planning approaches that rely on an effective modeling of the manufacturing system that usually does not consider the internal flows in detail. This is the subject of this thesis.

Most of these issues were raised as points needing research in Chien et al. (2011). The only point not discussed in this paper is the balancing of machine capacity, but works of Rowshannahad et al. (2015), Perraudat et al. (2019) and Ziarnetzky, Mönch, Ponsignon and Ehm (2019) show the importance of such balancing, even if they do not consider the whole sources of capacity consumption.

## 1.4 Generic Data Model and industrial instances

During the European project Productive 4.0, a generic data model for the semiconductor industry was established thanks to the cooperation of three European semiconductor manufacturing companies and Mines Saint-Etienne. In section 1.4.1, the relevant structures for production planning problems are presented. Industrial instances, provided by the site of Crolles of STMicroelectronics in the format of the generic data model, are then detailed in Section 1.4.2.

### 1.4.1 Generic Data Model

Productive 4.0 is an ECSEL European project that started in May 2017. Within this project, numerous industrial and academic partners are involved (more than 100). Inside the work package in charge of the simulation and production control of the supply chain (to which this thesis is associated), a generic data model for the semiconductor supply chain was developed. Although version 1.1 of the model has been released, in the following, only the first release (version 1.0) will be discussed, since the latest improvements did not change the core of the data model. Version 1.0 of the generic data model is described in Laipple et al. (2018).

An important characteristic of the generic data model is that, to be used for many use cases of the supply chain, it includes several aggregation levels. As it can be seen in Figure 1.6, five major entities are considered.

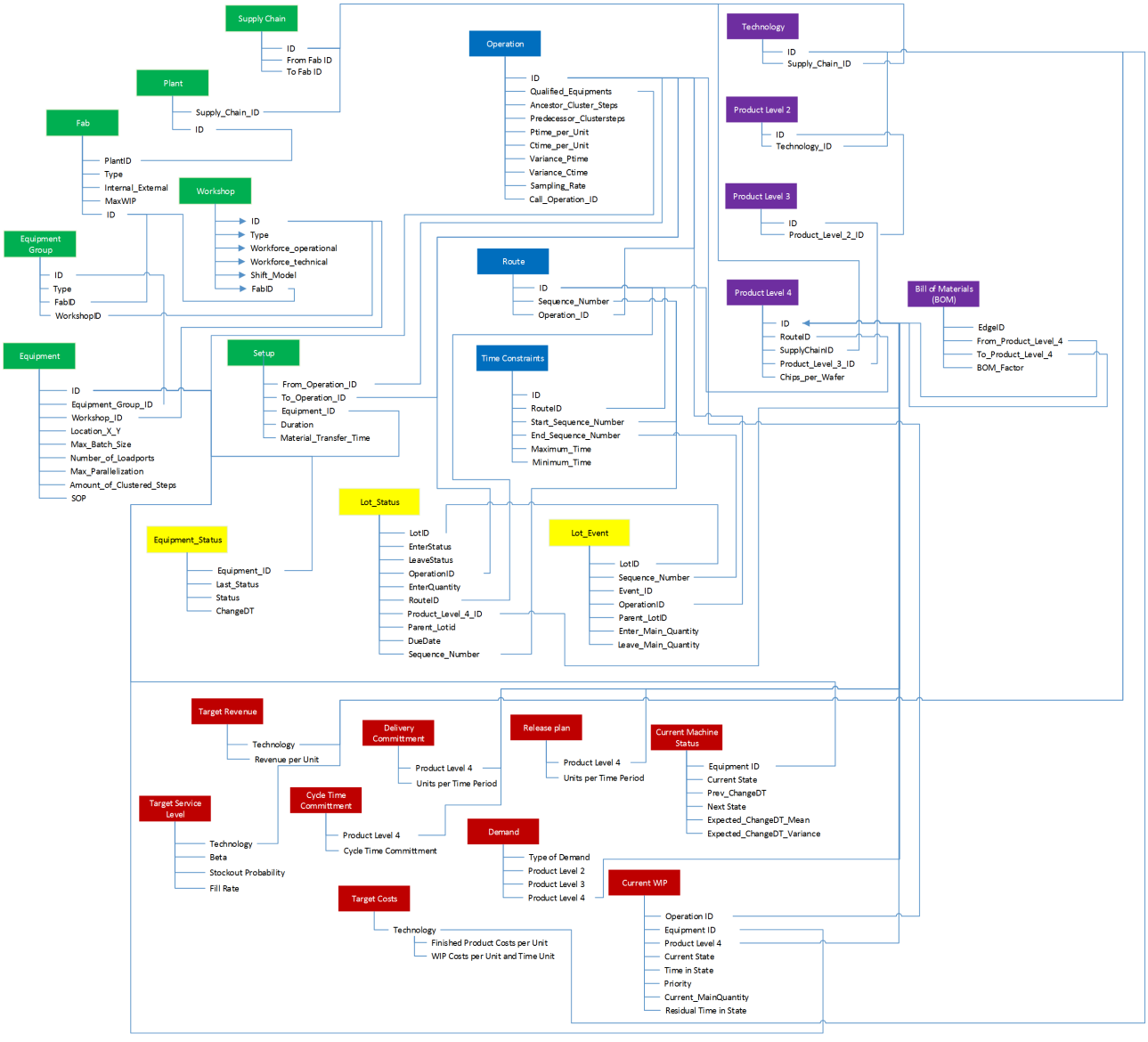



Figure 1.6: Generic Data Model for semiconductor Supply Chain from Georg Laipple's poster in Productive 4.0 Athens conference

– Major entities in green, blue and purple are called "Master entities", and form the backbone of the supply chain.

  - Green entities correspond to the physical system of the supply chain from the roots (the equipment) to the top of aggregation (the plant). The supply chain entity which links plants completes this major entity.
  - Blue entities refer to the processes in semiconductor manufacturing. It contains the description of operations, routes and time constraints occurring on the production routes. Note that the set-up matrix which describes the set-up time needed on a machine to process a new operation is colored in green, but could eventually be colored in blue.
  - Purple entities correspond to the products. Even if more levels are actually used in practice, the number of product levels in the generic model is equal to four. The highest level is called "Technology", where products are grouped by application. Product Level 4 is the lowest level where each product refers to one route. The Bill-Of-Materials (BOM) table helps to determine the link between products at different stages of the supply chain.

- Entities in yellow are “tracing entities”, which include the historical data, using foreign keys to connect the data to the master entities.

- Entities in red are “snapshot data”. They represent a one-time situation with the current WIP, the current demand, the current status of machines, etc.

If a lot of the data can easily be shared with academics, this is not the case of the production routes and most of the snapshot data (e.g. costs and profits, demands).

In the production planning use case, which is used in the whole manuscript, we use the following tables: Workshop, product level 4, route, operation, equipment capability, current WIP and demand. The equipment capacity is aggregated at the workshop level because, in the available data, some operations can be processed on several equipment group, and it is not in the scope of this thesis to decide on the assignment of operations to equipment groups. Products are considered at their lowest aggregation level, but could have been aggregated. An important issue with product aggregation is how to aggregate the different initial WIP of products. Table “equipment capability” allows an operation to be associated with the workshop on which the operation must be processed. Note that we consider that machines are up on the planning horizon, thus we do not use the snapshot of the equipment state. However, the available capacity of machines can be reduced to take into account the average time they are down.

### 1.4.2 Academic and industrial instances

The first instance used in our computational experiments was presented in Kayton et al. (1997) and is not industrial, but based on a reduced model of a manufacturing facility. The instance was helpful to evaluate the impacts of our work. The second instance is an industrial instance provided by STMicroelectronics. It corresponds to the full representation of a manufacturing facility and was generated from a highly used capacity planning software. The data set is a snapshot taken in 2018, from a 200mm fab. A research engineer worked on the industrial data, so that they fit the generic data model. The characteristics of the instances can be found in Table 1.1.

The small instance is depicted in Figure 1.7. Note that the first product has the longest route. The second product shares a large part of its route with the first product. Finally,

| | Reduced fab | 200mm fab |
|---|---|---|
| Horizon length | 61 | {91, 119, 147} |
| Number of workshops | 11 | 10 |
| Number of machines | 11 | 550 |
| Number of active products | 3 | 400 |
| Actual number of products | 3 | {15, 40, 75} |
| Maximum route length | 23 | 500 |
| Demand scenarios | {Low, Medium, High} | |

Table 1.1: Characteristics of the instances used in our computational experiments

the last product has a short route, and some of the workshops are not shared with the two other products.

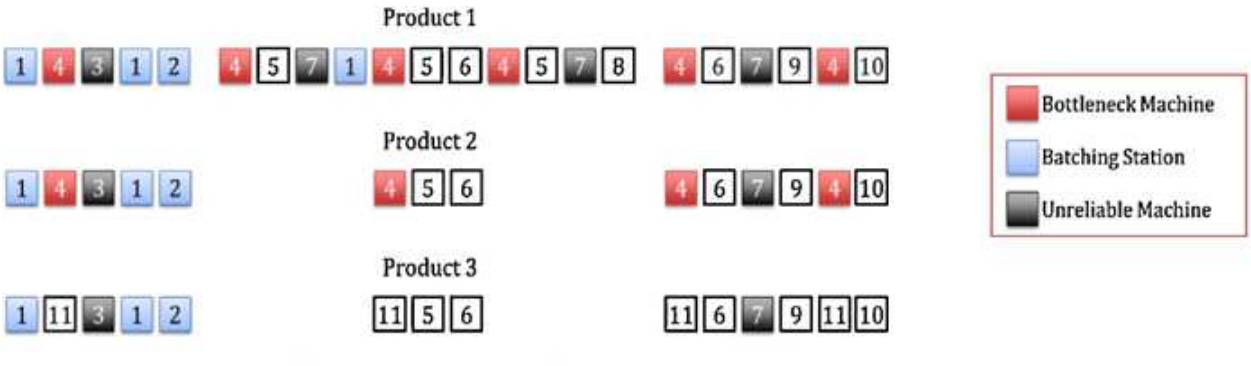



Figure 1.7: Reduced instance for production planning

Note that, for the industrial instance, the horizon should be long, because products have cycle times that range between 40 and 80 periods. To generate demands, the historical output over 6 months was considered. With these historical data, a frequency of orders, average demand and standard deviation for each product were estimated. Then, a demand scenario was randomly generated based on these characteristics. We only consider the products with the highest demands. To be able to analyze the influence of the number of products, 3 sets of products are considered. Among these top products, there are products with very small routes (fewer than 20 operations), and we assume that they are R&D products or products transferred from other fabs for few operations and are thus out of our scope. This is why these products were not used in our experiments. Products without demands generated on the horizon were also not considered. For the three instances, each demand scenario, related to the number of products, is adjusted by a factor on the generated demand to produce three scenarios with, respectively, a low and feasible demand, a medium demand that stresses the capacity and a high demand that cannot be fully met.

Because costs were not provided in the industrial data set (too sensitive information), the costs used in our computational experiment, and given in Table 1.2, were defined based on Kacar et al. (2012).

| | |
|---|---|
| Profit | 60 |
| Backlog | 50 |
| Inventory | 15 |
| WIP management | 0.001 |

Table 1.2: Unitary costs in our computational experiments

### 1.4.3 Conclusions

In this chapter, the context of the semiconductor industry, and more precisely of semiconductor manufacturing, was introduced. The reasons explaining the causes of complexity of the decision problems in semiconductor manufacturing were discussed, and also some of the associated challenges. An important feature is the large dimensions of the problems to solve (in terms of machines, products and operations). Because demand for Integrated Circuits is continuously growing, production planning, which decides the quantities to be produced over a planning horizon, is critical. This thesis aims at proposing novel approaches to optimize production plans that are effective and efficient to deal with semiconductor manufacturing characteristics. The literature on production planning, and more specifically in semiconductor manufacturing, is surveyed in Chapter 2.

# Chapter 2

# Literature review

## 2.1 Introduction

In this chapter, a review of the literature on production planning in semiconductor manufacturing and related topics is given. In Section 2.2, the production planning problem is defined with some history on the problem definition and a discussion on solution methods. Section 2.3 discusses the literature on production planning in semiconductor manufacturing. Section 2.4 focuses on congestion modeling in the semiconductor manufacturing literature, while Section 2.5 explores extensions of the production planning problem in semiconductor manufacturing.

## 2.2 Introduction on production planning

Our main concern in this thesis is the planning of the production of a wafer manufacturing facility. Going back to the roots of the production planning problem, this section provides some definitions and commonly used solution approaches to solve production planning problems.

### 2.2.1 Definition of production planning

Production planning aims to provide a tactical plan for production activities within one or several production facilities. It is frequently extended to the ordering of raw materials. Production plans are generally established on a mid-term horizon (several weeks to several months). Classical objectives in production planning are to meet the demands while minimizing total costs/maximizing total profits. Common decision variables on the production are the quantities ordered and produced at specific periods and inventory levels, but it could also include backlog levels, safety lead times on orders, etc. The facility environment can add many specific constraints that must be satisfied such as limited capacity on stocks or on production, setup costs and times in production periods, batches of products, etc.

### 2.2.2 A brief history of heuristics and exacts methods to solve production planning problem

Looking on the academic side, the production planning problem has only been studied since the second half of the 20th century, with the study of Modigliani and Hohn (1955) which

analyzes the trade-off between the inventory cost and the production cost. This study has greatly influenced the research in the field of production planning. A second major contribution of this decade is the paper of Wagner and Whitin (1958), in which the uncapacitated single product lot-sizing problem (LS-U) is described and solved using a dynamic program that runs in a polynomial time. LS-U is a production planning problem that considers the minimization of production, inventory and setup costs while deciding when and how much to produce. Even if LS-U does not seem realistic, it is used in many decomposition methods for more complex lot-sizing problems. The case when production capacity limit is considered in LS-U is coined the capacitated single product lot-sizing problem. In the general case where capacity varies over the time horizon, the problem is NP-Hard (Florian et al.; 1980). But, in the specific case where capacity is time invariant, a polynomial algorithm is proposed by Van Hoesel and Wagelmans (1996). Many extensions to these basic lot-sizing problems have been considered: Lot sizing with backlog or lost sales decisions, multi-item lot sizing with multiple products and various constraints on setups, multi-echelon multi-item lot sizing where several production steps are considered, etc.

Exact methods are not commonly used in industrial applications. Many heuristics have been successfully implemented and have led to better results for companies (compared to their previous state without any optimization of their production plans) that may let some companies in no need of advanced Operation Research methods. If methods to improve the efficiency of a facility can be tracked down to Taylor (1911) and its Method-Time-Measurement, methods to determine production plans in multi-level systems started with Materiel Requirements Planning (MRP) in the 70s. MRP is still widely used in many companies. However, MRP has a major drawback, it does not take into account production capacities, or only indirectly through lead times. Between the 70s and the 90s, methods to manage the bottleneck operations in factories arose, such as Theory of Constraints (Goldratt; 1990) or Just in Time (known for its implementation in the Toyota company with the Kanban method (Sugimori et al.; 1977)). The 90s saw two main changes in companies due to the democratization of IT systems. The first change is the improvement of MRP, which evolved to Manufacturing Resource Planning (MRP II). MRP II (Wight; 1995) integrates limited production capacities and more financial aspects of the business. The second change is the introduction of Enterprise Resource Planning (ERP) systems in all transactions occurring within the company are centralized. With a centralized access to the database, it is easier to accurately plan production and, behind the planning functions of an ERP software, there is often an MRP II module. Furthermore, since the beginning of the century, additional decision modules, called Advanced Planning System (APS), can be plugged in ERP systems in order to optimally plan at different stages and levels of the supply chain. The name APS covers the use of Artificial Intelligence or Operations Research techniques, and it is not simple to know which algorithms are running in an APS and what can be expected from the solutions. However, in semiconductor manufacturing, ERP systems are mostly used for order management and not so much for production planning (Mönch et al.; 2012). It lets place for advanced Operations Research methods that integrate the specificities of semiconductor manufacturing compared to the (too much) generic MRP II.

## 2.3 Production planning in semiconductor manufacturing

This section reviews the common characteristics of production planning problems in semiconductor manufacturing. Section 2.3.1 focuses on the modeling, while Section 2.3.2 covers the methods used to solve the problem and the instances used. Finally, in Section 2.3.3, we try to position the semiconductor manufacturing production planning problem among the lot-sizing and production planning literature.

### 2.3.1 Modeling

#### Objective functions and decision variables

In semiconductor manufacturing, the production planning problem usually aims at minimizing inventory and backlog costs, and sometimes Work In Process (WIP) management or production costs are also considered (Kim and Kim; 2001; Asmundsson et al.; 2009; Kacar et al.; 2012; Albey et al.; 2015; Ziarnetzky, Mönch and Uzsoy; 2019). Few articles (e.g. Hung and Leachman (1996); Habla and Mönch (2008); Albey, Bilge and Uzsoy (2017)) propose to maximize the total profit. However, to the best of our knowledge, only Chou and Hong (2000) propose to study various objective functions for the production planning problem in semiconductor manufacturing. In Chou and Hong (2000), profit maximization, machine use maximization, and output maximization are studied (as well as some hybridization between profit maximization and the two other objectives). Three main decision types of variables are commonly used in semiconductor manufacturing production planning. Inventory variables are always present in semiconductor manufacturing production planning models. Other common decision variables are the backlogged quantities, which model the possibility of delaying the fulfillment of the demand. Other classical variables that can be found are the quantities of finished products, manufactured at each period. But to capture and act on the production flows in the manufacturing process, additional decision variables are considered to model the quantities that are processed in each workshop (or machine) and to model (depending on the approach) the WIP (Work In Process) quantities.

#### Domain of the variables

Due to the large size of the production planning instances that should be solved in semiconductor manufacturing, the induced complexity of discrete variables can hardly be sustained. This is why many articles experimenting on realistic data sets such as Hung and Leachman (1996), Chen et al. (2010) and Bard et al. (2010) do not use integer variables. Mixed-Integer Linear Programming (MILP) models can be found in Habla et al. (2007), Hwang and Chang (2003) and Chou and Hong (2000), where integer variables model, respectively, integer quantities of products, binary variables accounting for the processing of a lot at each operation and period, and batching constraints.

#### Capacity constraints

In semiconductor manufacturing, the vast majority of research considers limited production capacities that are shared between workshops. Due to the very large number of machines in a facility, aggregating capacity (i.e. machines) at workshop level is very common. Some papers

assume that capacity constraints can be relaxed on non-bottleneck workshops (e.g. Habla et al. (2007)). An important effort to precisely model the capacity of the photolithography workshop can be found in (Bang and Kim; 2010). This is because photolithography concentrates the most expensive machines and is a common bottleneck in Front End facilities. However, only considering a single bottleneck workshop may be a too strong assumption. In fact, bottleneck workshops are usually not fixed and dynamically arise in the facility, as stated for instance by Chou and Hong (2000).

#### Failing and idle machines

A second matter when considering the capacity of workshops is the states of the machines. Some machines might be on planned maintenance, others might be idle due to large setup times to change their qualifications and some might be down due to failure. To model these effects, the naive way is to consider a ratio of up/down machines within every workshop. Another way is to consider capacity variation over the time horizon, but it can only model planned or anticipated maintenance operations and can only take machine failures into account on average. Finally, the effect of machine failures can be captured by lead time parameters or Clearing Functions (see Section 2.4) if failures frequently happen.

#### Setup times and costs

Due to the difficulty of solving models with integer variables, setup costs and times are commonly not explicitly modeled even if they often occur in reality. Most of the time, setup times are included in the lead times (this will be discussed in Section 2.4), and thus the fact that setup times might be sequence-dependent is not taken into account. The same remark holds for batching constraints. Batching operations are considered either by adding capacity to the machines or by dividing the process time of the operations. However, these assumptions can lead to unfeasible plans. Only Chou and Hong (2000) consider batching constraints in production planning within their MILP model.

### 2.3.2 Solution methods and common production planning instances in semiconductor manufacturing

In the semiconductor manufacturing literature, most solution methods are using commercial linear programming solvers such as IBM ILOG CPLEX, SAS OR or Fico Xpress. In some papers, such as Albey et al. (2015), Kriett et al. (2017) and Ziarnetzky, Mönch and Uzsoy (2019), a rolling horizon is used to cope with large time horizons. Genetic algorithms are also used in Liu et al. (2011) to solve a multi-objective problem (minimizing the mean and variance of the total cost under uncertainties). Three papers, (Hwang and Chang; 2003), (Habla et al.; 2007) and (Lim et al.; 2014), use Lagrangian relaxation to solve MILP models. Chou and Hong (2000) propose a heuristic to solve their MILP model in a relatively short computational time. The heuristic consists in determining with an iterative procedure the bottleneck workshops, and then in simplifying the model to only consider the bottleneck workshops. To cope with quadratic programs, (Albey et al.; 2014) and (Albey, Bilge and Uzsoy; 2017) use solvers such as KNITRO and BARON. Few instances were shared in the academic semiconductor manufacturing community. Various works exploit industrial data without making the instances available. Due to industrial privacy, only the dimensions of the studied data are provided. Some papers acknowledge that the instances come from

industrial data without giving the name of the company. However, we can observe that semiconductor companies partnering with academics on production planning are numerous: Micron, Xfab, Infineon, Texas Instrument, Vanguard, ... According to Ewen et al. (2017), there are only three instances shared between academics. They were all developed in the nineties and have at most 3 products. The first one is the MIMAC instance (Fowler and Robinson; 1995), which was developed based on SEMATEC data. A couple of years after, Spier and Kempf (1995) came with the MiniFab data set while Kayton et al. (1997) propose another reduced model of a fab approved by industrial partners (which was mainly used to assert the potential of clearing functions). Some characteristics of these instances are shown in Table 2.1.

| **Name** | **Nb. machines** | **Nb. workshops** | **Nb. products** |
|---|---|---|---|
| MIMAC I | 215 | 70 | 2 |
| MiniFab | 5 | 3 | 2 |
| Kayton's | 12 | 11 | 3 |

Table 2.1: Characteristics of common data sets

Note that some authors choose to generate their own data sets, (Kim et al.; 2014) and (Lim et al.; 2014), but their data sets are not often used. Note also that a new realistic instance called SMT2020 is proposed by Hassoun et al. (2019).

Time horizons are typically of several months (between 2 and 5 months), because smaller horizons could lead to products unable to finish their processing routes if the horizon is smaller than the cycle times of products. Usually, a period corresponds to one day or one week depending on the considered horizon. In many articles, the period length is not clearly defined. Note that the variability of the planning horizon is also due to the fact that production planning in semiconductor manufacturing is at the interface between the tactical and operational levels.

Table 2.2 identifies the data and the horizon used in the recent literature (note that the list is not exhaustive). When the instance is created, a comment on the size of the instance is given in parentheses.

### 2.3.3 Positioning production planning in semiconductor manufacturing in the production planning literature

In this section, the semiconductor manufacturing production planning problem is compared to the multi-level dynamic lot-sizing problem, and to integrated lot-sizing and scheduling problems. As a matter of fact, semiconductor production planning problems can be described as multi-product multi-step production planning problems.

#### Multi-level dynamic lot-sizing

Dynamic lot-sizing problems are classical problems in production management, introduced by Wagner and Whitin (1958). A generalization of this family of problems is the multi-level capacitated lot-sizing problem (MLCLSP), proposed by Billington et al. (1983). Multi-level lot sizing is generally separated into three branches depending on the Bill of Materials (BOM) structure (note that other specific BOMs may occur). If each product has at most one predecessor product and at most one successor product, it is called production in series.

| **Article** | **Instance** | **Period** | **Horizon** |
|---|---|---|---|
| Hung and Leachman (1996) | Industrial: Micron | month | 12 month |
| Chou and Hong (2000) | Industrial | ? | ? |
| Kim and Kim (2001) | Created (reduced size) | ? | 3 periods |
| Hwang and Chang (2003) | Industrial: Vanguard International Semi. Corp. | day | 60 days |
| Asmundsson et al. (2006) | Kayton's | ? | 70 periods |
| Habla et al. (2007) | Industrial: X-Fab Semi. Foundries AG | ? | ? |
| Habla and Mönch (2008) | Industrial: Infineon Technologies AG | week | 50 weeks |
| Asmundsson et al. (2009) | Kayton's | 2/4 hours | 70 days |
| Chen et al. (2010) | Industrial | day | 60 days |
| Irdem et al. (2010) | Kayton's | week/day | 14 weeks/91 days |
| Bang and Kim (2010) | Industrial | ? | 4 months |
| Bard et al. (2010) | Industrial: Texas Inst. | hour | 4-13 weeks |
| Kacar et al. (2012) | Kayton's | ? | 30 periods |
| Kacar et al. (2013) | MIMAC I | week | 15 weeks |
| Lim et al. (2014) | created (realistic) | week | 13 weeks |
| Albey et al. (2014) | created (reduced size) | 5 hours | 60 hours |
| Kim et al. (2014) | created (medium size) | day | 30 days |
| Albey et al. (2015) | Industrial | ? | 3 periods |
| Kacar et al. (2016) | MIMAC I | ? | 18 periods |
| Kim and Lee (2016) | created (medium size) | day | 30 days |
| Albey, Bilge and Uzsoy (2017) | created (reduced size) | 5 hours | 60 hours |
| Kriett et al. (2017) | MIMAC I | week | 20 weeks |
| Albey et al. (2019) | Kayton's | week | 12 weeks |
| Ziarnetzky, Mönch and Uzsoy (2019) | MIMAC I | week | 7 weeks |
| Zhang et al. (2020) | Kayton's | week | 26 weeks |

Table 2.2: Characteristics of instances used in the semiconductor manufacturing literature

In an assembly structure, each product has at most one successor, whereas each product has at most one predecessor in a divergent structure. Lead times can be used at every level, not only to model capacity (which is already considered in the capacitated case), but also to consider that process times are not null. The MLCLSP was solved using various heuristics, from Lagrangian heuristics to metaheuristics. The reader is referred to Buschkühl et al. (2010) for a literature review on dynamic capacitated lot-sizing problems.

Semiconductor manufacturing production planning problems are in the family of serial multi-level capacitated lot-sizing problems with lead times without setup costs or times. Another characteristic of semiconductor production planning problems is the shared capacity between levels.

#### Integrated lot sizing and scheduling

Integrated lot-sizing and scheduling problems have often been studied, according to the review of Copil et al. (2017). While lot sizing aims at meeting the demand at the lowest cost, scheduling corresponds to assigning and scheduling products on machines while minimizing the makespan or other objective functions. Such a junction between tactical and operational planning problems can lead to high complexity. Solving the full integrated mathematical model with multiple machines and multiple operations is often unrealistic. For this reason, solution methods can be separated into three kinds: Heuristics which solve the full problem (e.g. Gómez Urrutia et al. (2014)), hierarchical methods which solve the problems sequentially (e.g. Liberatore and Miller (1985)), and iterative methods (e.g. Dauzère-Pérès and Lasserre (1994)). In the literature, the focus may vary between scheduling oriented modeling and lot sizing oriented modeling. Furthermore, various types of heuristics are proposed depending on the problem to solve.

In semiconductor manufacturing production planning, the full integration of production planning and detailed scheduling decisions would cause a great computational burden, but some articles refer to the hierarchical methods except that the scheduling part is obtained through a simulation model (these articles will be detailed in Section 2.4). Anyway, data from the scheduling level is useful to plan production quantities in a smart way. Observed lead times may not only denote the overuse of capacity from a workshop, but can also include information of associated processes which might be limiting. To sum up, having a bottleneck machine does not always mean it works at full capacity, and the production planner needs to consider this information to improve his plans.

## 2.4 Congestion modeling

With the complexity of production flows in semiconductor manufacturing and the incentives to produce more and more (to optimize the yield of high investment machines), congestion is doomed to occur. Congestion is when a large workload induces longer lead times. However, as stated by Pahl et al. (2005), the relationship between the lead times and the workload is not linear. This non-linear effect in production planning is hard to model with only the classical capacity constraints. This is why most of the semiconductor manufacturing literature focuses on congestion modeling. In this section, the three most studied ways to model congestion are presented. For a deeper review of the used methods, but not up to date due to the recent progress of the use of clearing functions, the reader is referred to Armbruster and Uzsoy (2012). For a complete review of production planning with congestion with a less limited scope, the reader is referred to the book of Missbauer and Uzsoy (2020).

### 2.4.1 Fixed lead times

The first and most straightforward way to model congestion is to use fixed lead times, which are frequently assumed in production planning (Spitter et al. (2005) and Pahl (2012)). It consists in assuming a fixed delay, called Lead Time (LT), at each operation. Thus, a product entering in the waiting queue for an operation will be processed only LT periods after. Lead times are usually determined based on historical data and include waiting times, process times and also transportation times. Lead Times can account for machine failure or unavailability of auxiliary resources with the waiting time. Fixed Lead Times are easy to model and introduce low complexity, but they have several drawbacks. In particular, the workload is usually not balanced on all periods of the lead time, but is only counted in either the last period or the first period. In addition, production flows are not flexible and the values of the lead times are critical. If the lead times are too short, production flows must be strongly reduced in order to satisfy capacity constraints. Nevertheless, the fixed lead times are convenient and can be improved by using non-integer lead times as shown in Kacar et al. (2016).

### 2.4.2 Iterative procedures

An important fact is that lead times are not exogenous parameters. In fact, they directly depend on the production flows of products which compete for the same resource in a period. In short, lead times depend on the production plan. To address this circularity between production planning and operational level execution, Hung and Leachman (1996) propose an iterative procedure using both linear programming and discrete event simulation. The linear programming model is used to find a production plan that takes into account the lead times given by the simulation model, while the simulation model takes as inputs the production plan and evaluate it. These two steps are repeated until convergence.

Since 1996, several articles have used iterative processes with similar mathematical models but different simulation models (e.g. Bang and Kim (2010) and Zhang et al. (2020)). In fact, simulation is one of the major advantages of the iterative process because it allows complex behaviors to be integrated such as batching constraints, uncertainties on machine failures and repairs which are hardly tractable in an optimization model. However, as noted by Irdem et al. (2010) or Missbauer (2020), the reason behind this convergence does not have theoretical background except for the Kim and Kim (2001) method and other works based on it. Furthermore, the experiments of Bang and Kim (2010) show that iterative procedures are affected by the choice of the simulation model. In addition, a major drawback of iterative procedures is their computational burden, which can be seen in Bang and Kim (2010) where computational times range between 12 hours and 40 hours.

### 2.4.3 Clearing functions

The last main way to tackle congestion is the use of so-called Clearing Functions (CFs). Initially introduced by Graves (1986), Clearing Functions give the expected output of machines (or workshops) as a function of the workload. In their current shape (since the paper of Asmundsson et al. (2006)), CFs are non-linear functions that are estimated using simulation or historical data. CF constraints are generally linearized and included in a single linear programming model. In recent works, Albey, Bilge and Uzsoy (2017) study a CF that can deal with multiple products and multiple stages and even can be used with robust optimization

(Albey, Yanıkoğlu and Uzsoy; 2017). One of the main advantages of using CFs is the short computational times compared to using iterative procedures, because the burden is moved to the pre-processing phase (i.e. establishing CFs). But that is not the case with up-to-date clearing functions which need a quadratic solver or even a conic solver, Gopalswamy (2019). And for all CFs, when the structure of the facility changes, e.g. new machines are added, the Clearing Functions need to be re-evaluated.

### 2.4.4 Other ways

In Albey et al. (2019), congestion is tackled with a fixed lead time under uncertainties. They develop a robust optimization model to solve the problem. In Kriett et al. (2017), lead times are not directly considered, but cycle times are controlled by maintaining WIP levels close to targeted WIP levels.

## 2.5 Extension of production planning in semiconductor manufacturing

If production planning is a critical matter in semiconductor manufacturing and not easily solved, it does not mean that extensions of the problem were not studied. Even in the semiconductor industry where energy consumption of a facility can hardly be reduced (because machines are never totally stopped and operating the clear room consumes the most energy), the question of sustainability arises (Villarreal et al. (2012) and Hamed et al. (2018)). Ziarnetzky et al. (2017) propose a production planning problem that takes into account the sourcing of energy by considering a facility endowed with solar photovoltaics and wind turbines. In their model, they introduce penalties when using nonrenewable energy and a cost reduction when producing too much renewable energy.

In another extension of production planning problems, (Ziarnetzky, Mönch, Ponsignon and Ehm; 2019) consider the positive effect of engineering operations on the production efficiency, but also workload balancing between engineering operations and production operations. Note that the model was extended to the supply chain level with multiple facilities and additional penalties when the workload is not balanced between similar facilities.

In Lim et al. (2014), the authors propose to integrate the assignment of lots to orders in production planning. The objective is to minimize the total tardiness of lots. The MILP program (with binary variables for lot assignments) is solved using a Lagrangian relaxation approach.

Finally, master planning in the semiconductor industry is an extension of production planning to all facilities of the semiconductor supply chain. Due to the change of scale, more aggregations are needed, and also a longer time horizon to consider the life cycle of products (Bansal et al.; 2020), and more particularly of new products. But, as stated by Mönch et al. (2018b), master planning and production planning are quite similar. Up to now, master planning problems have not been studied much in semiconductor manufacturing, (Ponsignon and Eng; 2012), (Lowe and Mason; 2016) and (Ziarnetzky and Mönch; 2016).

## 2.6 Conclusions

Production planning in semiconductor manufacturing has been frequently studied, with a major focus on modeling production flows and congestion, aka phenomena deeply rooted at the scheduling decision level. Due to this particularity, coupling optimization and simulation models in a dynamic way (hierarchical methods) or only to assess the efficiency of the model is common. However, it comes with an additional computational burden if the simulation model is detailed. A first point of improvement is to continue the discussion on the objective functions opened by Chou and Hong (2000). A good objective function should take into account industrial needs, but should also ease the explanation of decisions to the decision makers. A second point of improvement is to keep digging the possibilities to model congestion with smaller computational times and/or more detailed decisions. Another gap to fill in the literature is the design of heuristics or other solution methods dedicated to production planning problems that could either reduce the computational times or cope with more complex models.

# Chapter 3

# Maximization of productivity and profit

## 3.1 Introduction

One of the research priorities in semiconductor manufacturing, as presented in Chapter 2, is congestion modeling. The objective functions are usually minimizing a combination of inventory and backlog costs. However, one of the most important objectives in semiconductor manufacturing is also to maintain a high productivity level due to the high investment in each facility. In the industry, productivity is measured by the *number of operations* performed in the planning horizon. This performance indicator is also called the number of "moves" in the industry jargon. To our knowledge, only Chou and Hong (2000) have proposed several objective functions: (1) Total profit maximization, (2) Maximization of the number of products, (3) Minimization of the residual capacity of machines and (4) Some hybrid objective functions mixing two of the three previous objectives. However, when maximizing the total profit, profit and costs are integrated in the same function and it is not easy to analyze the source of the additional profit (or the cost reduction). Chou and Hong (2000) lack a comparative analysis of the different objectives to highlight the side effects of each objective function. In this chapter, we propose new objective functions and integrate them in a classical production planning model with fixed lead times in order to enhance productivity and maximize profit. We use the so-called actualization rate to model the Net Present Value (NPV) of the profit. In Section 3.2, we generalize the classical semiconductor manufacturing production planning model with fixed lead times (based on Kacar et al. (2013)). Contrary to most studies, two timescales are considered. The first timescale (macro-periods) is used to model the satisfaction of demands, while the second timescale (micro-periods) is used to model the production process. In Section 3.3, a first alternative objective function, where the number of "moves" is maximized, is presented. In Section 3.4, a second objective function is proposed that considers the NPV of the total profit. Numerical experiments are conducted in Section 3.5. The models are compared and analyzed using a data set from the literature and an industrial data set with different demand profiles and actualization rates. In Section 3.6, the end of horizon effect is briefly studied and, to cope with this effect, a limit on the inventory of the final macro-period is proposed.

## 3.2 Generic model

In this section, a compact formulation based on the literature is presented for planning the production of $P$ products over a discrete time horizon that has two timescales. The time

horizon is decomposed into $T$ micro-periods (typically days) and $S$ macro-periods (typically weeks). Demands $D_{ps}$ are given per product $p$ and per macro-period $s$. Each product $p$ needs a sequence of operations $\mathcal{L}_p$ to be processed on $K$ workshops. Each workshop $k$ can process a finite set of operations $\mathcal{L}_p^k$ for each product $p$ and has a finite capacity $C_k$.

The plan is determined by optimizing internal production flows. The goal is to decide the quantities $X_{plt}$ of product $p$ to be processed at operation $l$ and period $t$. The set of operations of product $p$ and their resource consumption $\alpha_{pl}$ provide the timing of operations. In order to trace production flows, a variable $W_{plt}$ that model the work in process of product $p$ at operation $l$ and period $t$ is introduced. A unit work in process cost $w_{pl}$ is associated with each product $p$ and operation $l$.

The goal is to satisfy demands while minimizing inventory, backlog and work in process costs. We introduce a unit inventory cost $h_{ps}$ and a unit backlog cost $b_{ps}$ for each product $p$ and each period $s$ (typically a week). Let us also introduce two decision variables $I_{ps}$ and $B_{ps}$, that respectively model the inventory and the backlog of product $p$ at time period $s$.

Capacity congestion is first modeled with a fixed lead time $LT_{pl}$ for product $p$ at operation $l$. In this model, we assume that transportation times and costs between two workshops are negligible or constant. Products that complete a given operation are placed in a waiting queue for the next operation (the waiting queue is supposed to be uncapacitated, i.e. no limited storage). We also assume that the processing time of each operation is lower than one day (this assumption is justified since the longest operation usually needs less than half a day). All lead times are expressed in full micro-periods.

Due to the large industrial data sets we address, only continuous variables are considered in our models. For simplification reasons, batches are not explicitly taken into account, but batching operations have their processing time divided by the number of products in a typical batch of the operation. All sets, parameters and decision variables are summarized below.

### Notations

The following parameters are considered:

$P$: Number of products,
$K$: Number of workshops,
$\mathcal{L}_p$: Sorted list of operations of product $p$,
$\mathcal{L}_p^k$: Set of operations of product $p$ processed in workshop $k$,
$T$: Number of micro-periods in the planning horizon,
$S$: Number of macro-periods in the planning horizon,
$ts_s$: First micro-period in $\{1, \ldots, T\}$ of macro-period $s$ in $\{1, \ldots, S\}$,
$tf_s$: Last micro-period in $\{1, \ldots, T\}$ of macro-period $s$ in $\{1, \ldots, S\}$,
$\alpha_{pl}$: Unit resource consumption of operation $l$ of product $p$,
$C_k$: Daily available resource capacity of workshop $k$,
$LT_{pl}$: Lead time of operation $l$ in $\mathcal{L}_{(p)}$ of product $p$,
$D_{ps}$: Demand of product $p$ at the end of macro-period $s$,
$h_{ps}$: Unit inventory cost of product $p$ at the end of macro-period $s$,
$b_{ps}$: Unit backlog cost of product $p$ at the end of macro-period $s$,
$w_{pl}$: Unit work in process cost of product $p$ at operation $l$,
$B_{p0}$: Initial backlog of product $p$,
$I_{p0}$: Initial inventory of product $p$,
$W_{pl0}$: Initial work in process of product $p$ at operation $l$.

Two types of variables are used: Variables related to the internal production flows ($X_{plt}$, $Y_{plt}$ and $W_{plt}$), and variables related to the demand of the final product ($I_{pt}$ and $B_{pt}$). Variables $Y_{pt}^{\text{out}}$ are linking both sets of variables. The variables are formally defined below:

$X_{plt}$: Quantity of product $p$ to be released in micro-period $t$ to operation $l \in \mathcal{L}_p$,
$X_{pt}^{\text{in}} = X_{p1t}$: Quantity of product $p$ released in micro-period $t$,
$Y_{plt}$: Quantity of product $p$ completing its operation $l \in \mathcal{L}_p$ in micro-period $t$,
$Y_{pt}^{\text{out}} = Y_{p|\mathcal{L}_p|t}$: Output quantity of product $p$ in micro-period $t$,
$W_{plt}$: Work in process of product $p$ at operation $l \in \mathcal{L}_p$ at the end of micro-period $t$,
$I_{ps}$: Inventory level of product $p$ at the end of macro-period $s$,
$B_{ps}$: Backlog level of product $p$ at the end of macro-period $s$.

Note that variables such as inventory, backlog or WIP, have by extension their initial value called respectively $I_{p0}$, $B_{p0}$ and $W_{pl0}$. They are given as parameter.

**Mathematical formulation**

The mathematical model with fixed lead times is formalized below.

$$\min \quad \sum_{p=1}^{P} \sum_{l \in \mathcal{L}_p} \sum_{t=1}^{T} w_{pl} W_{plt} + \sum_{p=1}^{P} \sum_{s=1}^{S} (h_{ps} I_{ps} + b_{ps} B_{ps}) \tag{3.1}$$

$$\text{s.t.} \quad Y_{plt} = X_{p(l+1)(t)} \qquad \forall p \in \{1,\dots,P\},\ \forall l \in \{1,\dots,|\mathcal{L}_p|-1\},\ \forall t \in \{1,\dots,T\} \tag{3.2}$$

$$W_{plt} = W_{pl(t-1)} + X_{plt} - Y_{plt} \quad \forall p \in \{1,\dots,P\},\ \forall l \in \mathcal{L}_p,\ \forall t \in \{1,\dots,T\} \tag{3.3}$$

$$X_{plt} = Y_{pl(t+LT_{pl})} \qquad \forall p \in \{1,\dots,P\},\ \forall l \in \mathcal{L}_p,\ \forall t \in \{1,...,T - LT_{pl}\} \tag{3.4}$$

$$D_{ps} + B_{p(s-1)} = \sum_{\tau=ts_s}^{tf_s} Y_{p\tau}^{\text{out}} + I_{p(s-1)} - I_{ps} + B_{ps} \qquad \forall p \in \{1,\dots,P\},\ \forall s \in \{1,\dots,S\} \tag{3.5}$$

$$\sum_{p=1}^{P} \sum_{l \in \mathcal{L}_p^k} \alpha_{pl} Y_{plt} \leq C_k \qquad \forall k \in \{1,\dots,K\},\ \forall t \in \{1,\dots,T\} \tag{3.6}$$

$$X_{plt}, Y_{plt}, W_{plt} \geq 0 \qquad \forall p \in \{1,\dots,P\},\ \forall l \in \mathcal{L}_p,\ \forall t \in \{1,\dots,T\} \tag{3.7}$$

$$I_{ps}, B_{ps} \geq 0 \qquad \forall p \in \{1,\dots,P\},\ \forall s \in \{1,\dots,S\} \tag{3.8}$$

The objective function (3.1) minimizes the total inventory, backlog and work in process cost. Constraints (3.2)-(3.5) model flow conservation. Constraints (3.2) ensure the link between the output of a given operation $Y_{plt}$ and the input of the next operation $X_{p(l+1)t}$. Constraints (3.3) balance the work in process over the time horizon for each operation. Constraints (3.4) guarantee the fixed lead time for each operation of each product. Constraints (3.5) are the flow conservation constraints of the products, ensuring the satisfaction of demands through the inventory and the production in the current period or their backlog to subsequent periods. The capacity constraints in each workshop are modeled through Constraints (3.6). Constraints (3.7) and (3.8) ensure the non-negativity of decision variables.

## 3.3 Maximization of a productivity Key Performance Indicator (KPI)

In the semiconductor industry, an important indicator of productivity is the number of performed operations, also called "moves". It corresponds to the number of completed operations multiplied by the number of products processed per tool, workshop and plant. For example, if we have 8 machines and if each machine processes 100 units, then the number of "moves" is equal to $8 \times 100 = 800$. We propose to include this indicator in the previously defined objective function (3.1) with a scaling factor $E$. The new objective function is given below that maximizes the number of moves while minimizing the objective function (3.1).

$$\max \quad E \sum_{p=1}^{P} \sum_{l \in \mathcal{L}_p} \sum_{t=1}^{T} Y_{plt} - \sum_{p=1}^{P} \sum_{s=1}^{S} (h_{ps} I_{ps} + b_{ps} B_{ps}) \tag{3.9}$$

## 3.4 Profit maximization using an actualization rate

Mixing the minimization of the costs and the maximization of the number of "moves" is not the most natural way to improve productivity. In the following, we replace the maximization of the number of "moves" by the maximization of the profit generated by the products. This leads to a homogeneous objective function expressed in monetary units. Let us introduce $G_p$, the profit per unit of product $p$. In addition, using this new objective function, it is possible to model the fact that future profits and their associated decisions are less important than the current profits. This is done by introducing the notion of NPV, which is often used in Economics to calculate the return on investment taking into account the time value of money (one monetary unit today is larger than the same monetary unit tomorrow). All future financial flows are included in a single function with an actualization rate $\beta_s \in (0, 1]$ that depends on the macro-period $s$, in order to emphasize the importance of the financial results in the first macro-periods. Some articles (Hung and Leachman; 1996; Albey, Bilge and Uzsoy; 2017) also consider profit maximization with time discount, but the discount function is not given. In our model, this actualization rate is applied each week, which means that the present value of the profit in macro-period $s$ reduces as $s$ increases. More precisely, in the following, we use $\beta_s = \beta_0^{(s-1)}$. The discount rate applied week by week is constant, thus we denoted it $\beta$ . Equation (3.10) below models the new objective function.

$$\max \quad \sum_{p=1}^{P} \sum_{s=1}^{S} \beta^{(s-1)} (-h_{ps} I_{ps} - b_{ps} B_{ps} + \sum_{t=ts_s}^{tf_s} G_p Y_{pt}^{out}) \tag{3.10}$$

## 3.5 Numerical experiments

In this section, we conduct numerical experiments using IBM ILOG CPLEX to solve the models with the objective functions presented above. The quality of the solutions is assessed by considering the total costs/profits, the productivity (represented by the number of "moves") and the total output of products. In Section 3.5.1, the data sets used for the numerical experiments and presented in Chapter 1 are summarized. Then, in Section 3.5.2, we analyze the impact of the scaling parameter $E$ on the objective function (3.9). Finally, in Section 3.5.3, the impact of the NPV on the objective function (3.10) is analyzed.

### 3.5.1 Data sets

Experiments are first conducted on the reduced data set from Kayton et al. (1997), that includes three products with routes of 14 to 23 operations and 11 workshops. The first product has the longest route, while the second product shares a large part of its route with the first product. The third product has a short route, and some of its workshops are not shared with the two other products.

According to the characteristics of the instance given by Kacar et al. (2012), lead times are fixed as follows:

Operations on bottleneck machines have a lead time of 5 micro-periods,
Operations on unreliable machines have a lead time of 3 micro-periods,
Operations on batching machines have a lead time of 1 micro-period,
Operations on the remaining machines are given a lead time of 0 micro-period.

The horizon is divided into 61 micro-periods, i.e. 9 macro-periods. Three profiles of static (not time-dependent) demands are considered.

Scenario 1. High infeasible demands: {45,15,15},
Scenario 2. Medium feasible demands: {33,11,11},
Scenario 3. Low feasible demands: {15,5,5}.

Then, the models are tested on the industrial data set of one the manufacturing facility of ST Crolles, "C200", that includes 10 workshops. We only consider the demand of the 75 most produced products. To avoid an over-capacitated factory (due to the restricted number of products considered), demands are slightly increased. As for the Kayton's instance, three scenarios of demand are considered. Experiments are conducted with a time horizon of 119 micro-periods (i.e. 17 macro-periods, which is equivalent to 4 months). Cycle times are between 40 and 80 days in the data set. We tried to experiment our models with an initial WIP based on a snapshot of the production facility, but the Fixed Lead Time model raises an infeasibility error due to capacity constraint, while trying to process the initial WIP in the first period. Thus, experiments are done with a reduced initial WIP (divided by two).

In this chapter, the work in process costs are set to zero in all the numerical experiments. As in Kacar et al. (2012), the unit backlog cost is set to 50, the unit inventory cost to 15 and the profit per unit of product to 60.

### 3.5.2 Analysis of productivity maximization

Let us first analyze the impact of the scaling factor $E$ on the total cost $(\sum_{p=1}^{P}\sum_{s=1}^{S}(h_{ps}I_{ps} + b_{ps}B_{ps})$, the number of moves $(\sum_{p=1}^{P}\sum_{l\in\mathcal{L}_p}\sum_{t=1}^{T}Y_{plt})$ and the total output $(\sum_{p=1}^{P}\sum_{t=1}^{T}Y_{pt}^{out})$ (denoted respectively "Total Cost", "#Moves" and "Total Output" in the following tables) when using the objective function (3.9). The scaling factor $E$ is fixed to 0, 1, 5 and 10. Tables 3.1, 3.2 and 3.3 summarize the results for the three demand profiles (resp. high demand, medium demand and low demand). Column "$E = 0$" corresponds to the generic model (3.1)-(3.8) and provides reference values. The deviation in percentage from these reference values is given between brackets.

Tables 3.1, 3.2 and 3.3 show that the productivity can significantly be improved. Even with a small scaling factor $E$, the number of "moves" increases from 7% for high demands up

Table 3.1: Impacts of the scaling factor ($E$) on the number of "Moves" for high demands on the Kayton's instance.

| | $E = 0$ | $E = 1$ | $E = 5$ | $E = 10$ |
|---|---|---|---|---|
| **Total Outputs** | 565 | 565 | 628 (+11.1%) | 691 (+22.2%) |
| **Total Costs** | 22,515 | 22,534 (+0.1%) | 25,715 (+14.2%) | 31,457 (+39.7%) |
| **#Moves** | 9,750 | 10,473 (+7.4%) | 11,464 (+17.6%) | 12,341 (+26.6%) |

Table 3.2: Impacts of the scaling factor ($E$) on the number of "Moves" for medium demands on the Kayton's instance.

| | $E = 0$ | $E = 1$ | $E = 5$ | $E = 10$ |
|---|---|---|---|---|
| **Total Output** | 495 | 495 | 576 (+16.4%) | 649 (+31.2%) |
| **Total Costs** | 0 | 0 | 4,020 | 10,849 |
| **#Moves** | 8,940 | 10,003 (+11.9%) | 11,251 (+25.9%) | 12,295 (+37.5%) |

to 82% for low demands. With larger values of $E$, larger improvements are obtained on the number of moves. This is done at the expense of the total cost. There is a trade-off between productivity and inventory and backlog costs. Note also from Tables 3.1, 3.2 and 3.3 that the total output increases when the productivity is improved. Through these results, we cannot find a correlation factor between $E$ and $\#Moves$.

The results on the industrial instances are summarized in Tables 3.4, 3.5 and 3.6. Here again, the trade-off between productivity and inventory and backlog cost is quite visible. Note that, with an increase of the total cost by 2 %, it is possible to increase the number of "moves" by 50% (with $E = 1$ and high demand). Note also that the number of "moves" obtained with a fixed scaling factor does not seem strongly dependent on the demand profile (except for low demands where there is a lot of capacity left after the production of demand).

However, this first objective function is a naive way to improve productivity. This is why we explore, in the following, the impact of considering the objective function that maximizes the total profit.

### 3.5.3 Impact of using a financial objective

The profit maximization objective function helps to move from a pure cost-driven model to a profit-driven model. As shown in the following, the NPV model also improves productivity.

The NPV model with different actualization rates $\beta$ is compared to the results of the Generic model (Column "Generic"), where $\beta$ is fixed to 1 (i.e. no depreciation), 0.95 and 0.8. The indicators "Total output" and "#Moves" are kept. Two other indicators are introduced: The total profit considering no depreciation and the total profit considering an actualization rate of 0.95. Tables 3.7, 3.8 and 3.9 summarize the results on the Kayton's instance

Table 3.3: Impacts of the scaling factor ($E$) on the number of "Moves" for low demands on the Kayton's instance.

| | $E = 0$ | $E = 1$ | $E = 5$ | $E = 10$ |
|---|---|---|---|---|
| **Total Output** | 225 | 227 (+0.7 %) | 446 (+98.4%) | 564 (+150.6%) |
| **Total Costs** | 0 | 636 | 8,595 | 21,713 |
| **#Moves** | 4,138 | 7,525 (+81.9%) | 9,927 (+139.9%) | 11,900 (+187.6%) |

Table 3.4: Impacts of the scaling factor ($E$) on the number of "Moves" for high demands on the industrial instance.

| | $E = 0$ | $E = 1$ | $E = 5$ | $E = 10$ |
|---|---|---|---|---|
| **Total Outputs** | 3,294 | 3,224 (-2.1%) | 3,685 (+11.9%) | 4,499 (+36.6%) |
| **Total Costs** | 241,517 | 247,604 (+2.5%) | 321,753 (+33.2%) | 504,083 (+108.7%) |
| **#Moves** | 603,539 | 884,945 (+46.6%) | 913,338 (+51.3%) | 937,423 (+55.3%) |

Table 3.5: Impacts of the scaling factor ($E$) on the number of "Moves" for medium demands on the industrial instance.

| | $E = 0$ | $E = 1$ | $E = 5$ | $E = 10$ |
|---|---|---|---|---|
| **Total Output** | 3,113 | 3,074 (-1.3%) | 3,773 (+21.2%) | 4,575 (+47.0%) |
| **Total Costs** | 199,097 | 209,562 (+5.3%) | 302,302 (+51.8%) | 468,836 (+135.5%) |
| **#Moves** | 567,849 | 885,105 (+55.9%) | 916,721 (+61.4%) | 939,582 (+65.5%) |

considering high, medium and low demand profiles, respectively.

First, note that the objectives with the actualization rates $\beta = 1$, $\beta = 0.95$ and $\beta = 0.8$ are different by definition. From Tables 3.7 to 3.9, note that the actualization rates $\beta = 1$ and $\beta = 0.95$ provide similar results, while the actualization rate $\beta = 0.8$ provides a larger total output which induces a larger number of moves (#Moves). By comparing the generic model to models with a profit per product, the total output increases by 12% for instances with high demands and up to 193% for instances with low demands. The total profit increase is also significant, varying from 11% to 63%. A detailed analysis of the results shows that, even if the total profit increases for high demands are greater in percentage compared to the profit increases for medium demands, the absolute increase of the total profit for medium demands (3,370) is larger than the absolute increase of the total profit for high demands (2,164). This larger total profit can be explained by the fact that the demand of instances with medium demands can be met while the demand of instances with high demands cannot be met.

Tables 3.10, 3.11 and 3.12 summarize the results on the industrial instance, considering high, medium and low demand profiles, respectively. First, note that, due to the inadequate initial WIP, backlog and inventory costs in the first weeks are unavoidable and may lead to negative profits. This helps understanding the negative profits when the weekly actualization rate is fixed to 0.95. These negative values are due to the high impact of the first periods on the objective function. Note that, when no actualization rate is applied ($\beta = 1$), the profits can be positive. We also observe that the number of "moves" and the total output of products are clearly boosted by maximizing the profit. However, although the quantity of products increases when reducing the actualization rate, the number of "moves" slightly decreases. It can be assumed that, with a lower actualization rate, more products with short

Table 3.6: Impacts of the scaling factor ($E$) on the number of "Moves" for low demands on the industrial instance.

| | $E = 0$ | $E = 1$ | $E = 5$ | $E = 10$ |
|---|---|---|---|---|
| **Total Output** | 2,890 | 2,947 (+2.0 %) | 3,850 (+33.2%) | 4,727 (+63.6%) |
| **Total Costs** | 178,737 | 199,438 (+11.6%) | 298,160 (+66.8%) | 453,647 (+153.8%) |
| **#Moves** | 520,279 | 889,179 (+70.9%) | 921,855 (+77.2%) | 943232 (+81.3%) |

Table 3.7: Variations of the actualization rate $\beta$ for high demands on the Kayton's instance.

| **Models** | **Generic** | **NPV** $\beta$=1 | **NPV** $\beta$=0.95 | **NPV** $\beta$=0.8 |
|---|---|---|---|---|
| **Total output** | 565 | 633 (+12.0 %) | 643 (+13.8%) | 713 (+26.0%) |
| **#Moves** | 9,750 | 10,645 (+9.2%) | 10,787 (+10.6%) | 11,727 (+20.3%) |
| **Total profit with $\beta$=1** | 11,407 | 13,571 (+19.0%) | 13,563 (+18.9%) | 11,361 (-0.4%) |
| **Total profit with $\beta$=0.95** | 11,249 | 12,761 (+13.4%) | 12,793 (+13.7%) | 11,588 (+3.0%) |

Table 3.8: Variations of the actualization rate $\beta$ for medium demands on the Kayton's instance.

| **Models** | **Generic** | **NPV** $\beta$=1 | **NPV** $\beta$=0.95 | **NPV** $\beta$=0.8 |
|---|---|---|---|---|
| **Total output** | 495 | 595 (+20.2 %) | 610 (+23.2%) | 697 (+40.7%) |
| **#Moves** | 8,940 | 10,266 (+14.8%) | 10,458 (+17.0%) | 11,672 (+30.6%) |
| **Total profit with** $\beta$=1 | 29,700 | 33,070 (+11.3%) | 33,065 (+11.3%) | 30,564 (+2.9%) |
| **Total profit with** $\beta$=0.95 | 24,404 | 26,747 (+9.6%) | 26,795 (+9.8%) | 25,421 (+4.1%) |

routes and thus short cycle times are produced in the last periods, leading to a reduced number of "moves".

Because profits with the generic model are always negative, it is not possible to compute the increase ratio. But an important point is that, even if when minimizing costs leads to negative profits, considering a profit per product leads to manufacture products that are not ordered on the planning horizon and allows positive profits to be achieved.

## 3.6 End of horizon effect

With profit maximization, secondary objectives such as maximizing the total number of products or maximizing the productivity (the number of "moves") are also improved. In order to analyze in depth the production plans obtained with the profit maximization, we focus on the throughput of products.

Figures 3.1a, 3.1b and 3.1c show the weekly total outputs for each experiment on the Kayton's instance compared to the demands for respectively high, medium and low demands. Note that the generic model is not depicted in Figures 3.1b and 3.1c since it fits the demands.

From Figures 3.1b and 3.1c, note that all solution of the different models start by producing the required demand in the first period, an then the solution of model with the lowest actualization rate produces more than the demand. The solutions of the other models start overproducing at the end of the horizon since there is no inventory cost. In our experiments, we assume that the profit of a product is lost if it stays more than four weeks in the inventory. The same remarks can be drawn for Figure 3.1a, where the NPV model with $\beta = 1$ and $\beta = 0.95$ follows the behavior of the generic model until the end of the horizon where the NPV model starts overproducing.

Similar results are observed with the industrial instances. Because the demand in the industrial instance is dynamic, Figures (3.2a), (3.2b) and (3.2c) show only the produced quantities minus the demand (which is equivalent to remove the backlog quantities from the inventories) on the horizon. If, for the first 5 weeks, the production plan is exactly the same (due to the same initial WIP and fixed lead time constraints), note that there is a large inventory in the last weeks when maximizing the profit. Note that production plans

Table 3.9: Variations of the actualization rate $\beta$ for low demands on the Kayton's instance.

| **Models** | **Generic** | **NPV** $\beta$=1 | **NPV** $\beta$=0.95 | **NPV** $\beta$=0.8 |
|---|---|---|---|---|
| **Total output** | 225 | 469 (+108.4 %) | 487 (+116.7%) | 660 (+193.2%) |
| **#Moves** | 4,138 | 8,060 (+94.8%) | 8,316 (+101.0%) | 10,243 (+147.6%) |
| **Total profit with** $\beta$=1 | 13,500 | 22,105 (+63.7%) | 22,104 (+63.7%) | 17,543 (+29.9%) |
| **Total profit with** $\beta$=0.95 | 11,093 | 17,035 (+53.6%) | 17,097 (+54.1%) | 14,687 (+32.4%) |

Table 3.10: Variations of the actualization rate $\beta$ for high demands on the industrial instance.

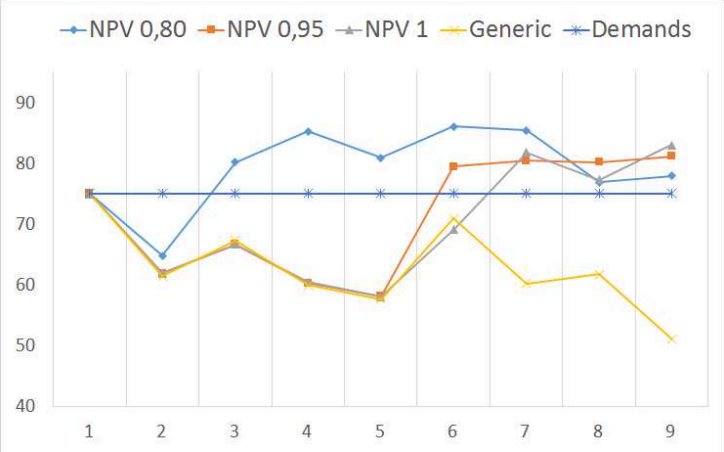
| **Models** | **Generic** | **NPV** $\beta$=1 | **NPV** $\beta$=0.95 | **NPV** $\beta$=0.8 |
|---|---|---|---|---|
| **Total output** | 3,294 | 4,824 (+46.4 %) | 4,841 (+47.0%) | 5,161 (+56.7%) |
| **#Moves** | 603,539 | 784,199 (+29.9%) | 780,573 (+29.3%) | 760,756 (+26.0%) |
| **Total profit with** $\beta$=1 | -43,875 | 12,255 | 11,577 | -14,221 |
| **Total profit with** $\beta$=0.95 | -51,619 | -26,691 (+48.3%) | -26,404 (+48.8%) | -35,060 (+32.1%) |

Table 3.11: Variations of the actualization rate $\beta$ for medium demands on the industrial instance.

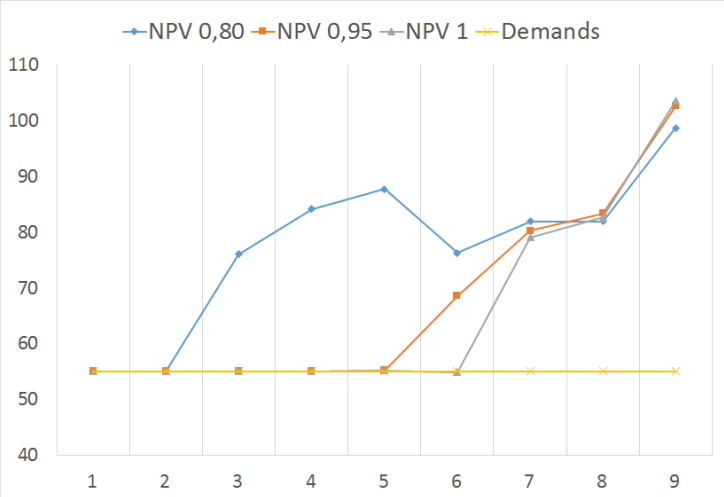
| **Models** | **Generic** | **NPV** $\beta$=1 | **NPV** $\beta$=0.95 | **NPV** $\beta$=0.8 |
|---|---|---|---|---|
| **Total output** | 3,113 | 4,893 (+57.2 %) | 4,913 (+57.8%) | 5,185 (+66.6%) |
| **#Moves** | 567,849 | 782,364 (+37.8%) | 776,045 (+36.7%) | 760,721 (+34.0%) |
| **Total profit with** $\beta$=1 | -12,301 | 55,375 | 54,743 | 33,746 |
| **Total profit with** $\beta$=0.95 | -30,857 | -1,816 (+94.1%) | -1,482 (+95.2%) | -8,755 (+71.6%) |

Table 3.12: Variations of the actualization rate $\beta$ for low demands on the industrial instance.

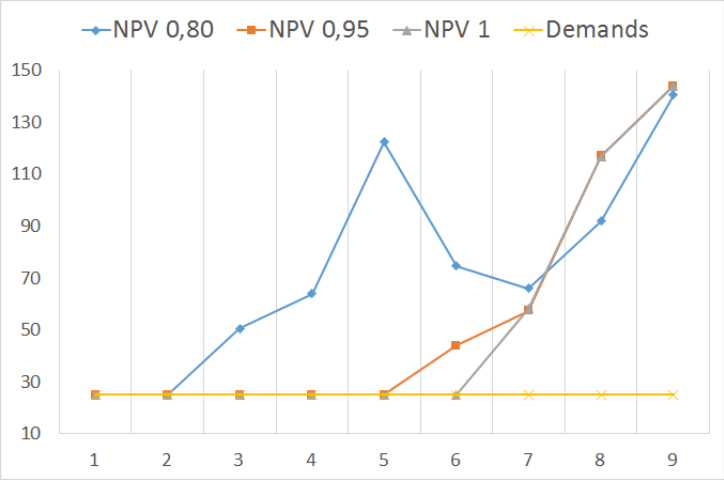
| **Models** | **Generic** | **NPV** $\beta$=1 | **NPV** $\beta$=0.95 | **NPV** $\beta$=0.8 |
|---|---|---|---|---|
| **Total output** | 2,890 | 5,126 (+77.4 %) | 5,127 (+77.4%) | 5,316 (+83.9%) |
| **#Moves** | 520,279 | 783,747 (+50.6%) | 774,674 (+48.9%) | 763,824 (+46.8%) |
| **Total profit with** $\beta$=1 | -5,360 | 78,667 | 78,146 | 63,428 |
| **Total profit with** $\beta$=0.95 | -23,864 | 12,366 | 12,645 | 7,546 |




(a) High demand



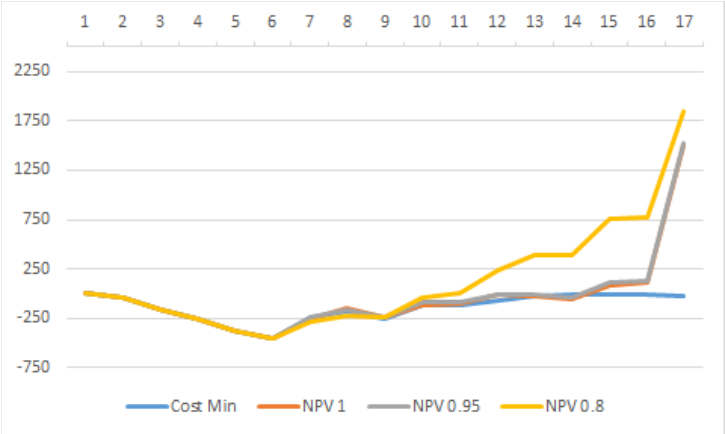

(b) Medium demand



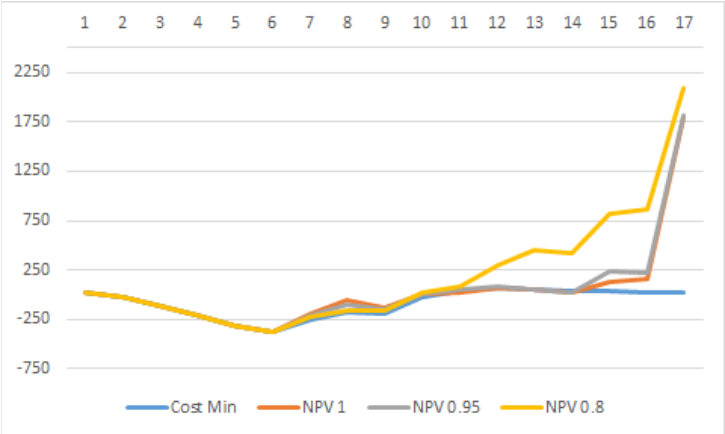

(c) Low demand

Figure 3.1: Weekly outputs (Kayton's instance).

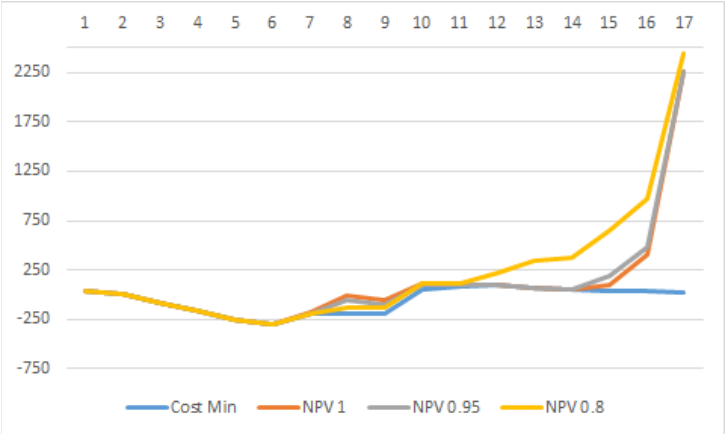
optimized with actualization rates of 1 and 0.95 are very close, contrary to the production plans optimized with an actualization rate of 0.8.




(a) High demand




(b) Medium demand




(c) Low demand

Figure 3.2: Production quantities for different demand profiles (industrial instance).

These end-of-horizon effects were expected. In fact, financial results are increased at the expense of meeting demands. The end-of-horizon overstock can be too large, and this anticipated production may have to be limited if demands after the end of the horizon are not expected to be large enough. A in-depth analysis of the production plan with the Kayton's

instance shows that two of the three products are anticipated. These products are the ones with the shortest routes, i.e. requiring less capacity to manufacture.

### 3.6.1 Limiting excessive production

End of horizon effects are quite common in production planning. With cost minimization, the system tends to empty the work in process and the final inventories at the end of the horizon. This is why Habla and Mönch (2008) or Kriett et al. (2017) extend the time horizon and assign demands to the additional periods. When maximizing the total profit, the work in process also tends to be empty, but the inventories in the last period of some products increase.

One way to limit the end-of-horizon effect is to limit the inventory at the end of the planning horizon for some specific products. In the following, we only consider the case with medium demands and the NPV model with an actualization rate of 0.95 for the Kayton's instance. Figure 3.3a details the results obtained in Section 3.5.3 by depicting the weekly outputs of the three products. This figure shows that the first product follows the demand profile, while Products 2 and 3 are overproduced at the end of the horizon.

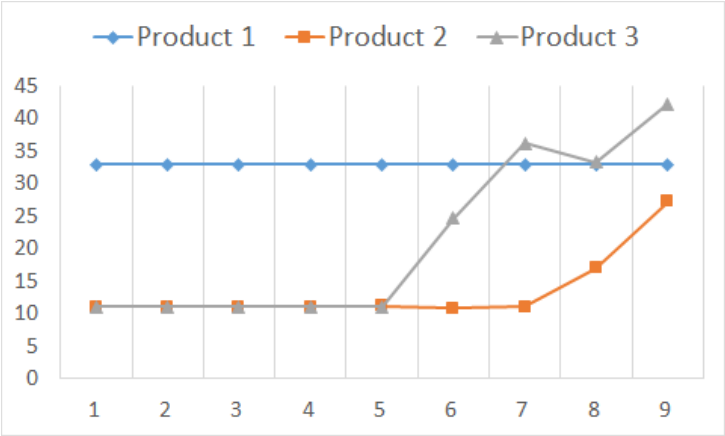
First, in Figure 3.3b, we start by limiting the inventory of the last period of Product 3 to four times the demand of the last period. Recall that Product 3 shares few non-critical machines with other routes. Limiting its last period inventory has almost no impact on the production plan of other products. Note that this additional constraint reduces the total profit by 2% (considering an actualization rate of 0.95).

In Figure 3.3c, the inventory of the last period of Product 2 is limited to the demand of the last period. This new constraint causes a transfer of production from Product 2 to Product 1. This is because the routes of Products 1 and 2 share several critical machines. Note that, in this case, adding a limit at the end of the horizon for Product 2 does not significantly impact the total profit (a reduction of 0.3%).

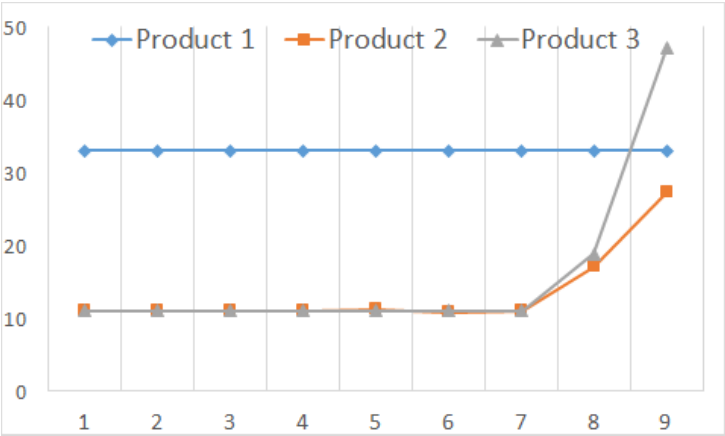
## 3.7 Conclusion and perspectives

In this chapter, we introduced models with new objective functions that aim at optimizing productivity and financial objectives for wafer manufacturing. These models were tested on a data set of the literature and on a large-scale industrial data set. First, we proposed a model that considers a classical industrial indicator (number of "moves"). The experiments showed that there is a trade-off between productivity and classical costs (inventory costs and backlog costs). Second, we developed a profit-driven model by introducing the NPV in a profit function. The experiments illustrated that the profit-driven model ensures a better productivity than a pure cost-driven model, but it can lead to overproduction (in particular at the end of the horizon). Thus, we proposed to limit the inventory level of some products at the end of the horizon. These limits are important for products that share capacities with non-overproduced products.

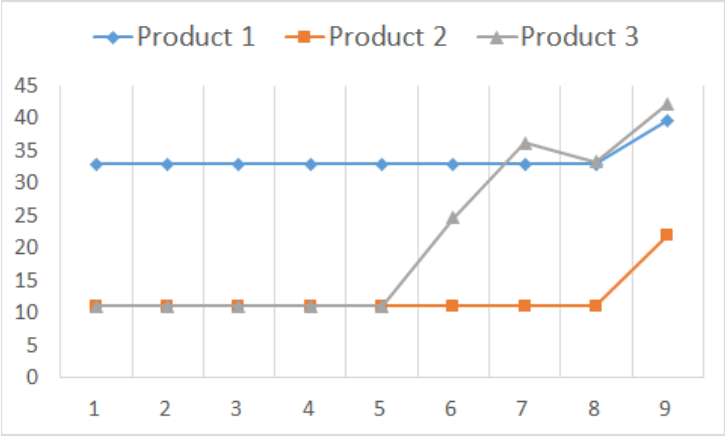
In the future, it will be interesting to propose an approach to link these limits to the demand forecasts (after the planning horizon) because, as stated in van den Heuvel and Wagelmans (2005), knowing the future demand can considerably improve the solutions. Another way to limit the extra inventory is to use a nonlinear function, i.e. a piecewise linear function, such that the profit per unit of product decreases with the number of products in the inventory. The profit per unit must be lower than the unit backlog cost to avoid favoring




(a) No inventory limit




(b) Inventory limit on Product 3




(c) Inventory limit on Product 2

Figure 3.3: Weekly outputs by product (NPV model with $\beta$=0.95 and medium demand).

the surplus of a product to meet the demand of another product. Figure 3.4 shows an example of a piecewise linear profit function. The break point of the profit function could be the target inventory of the product.

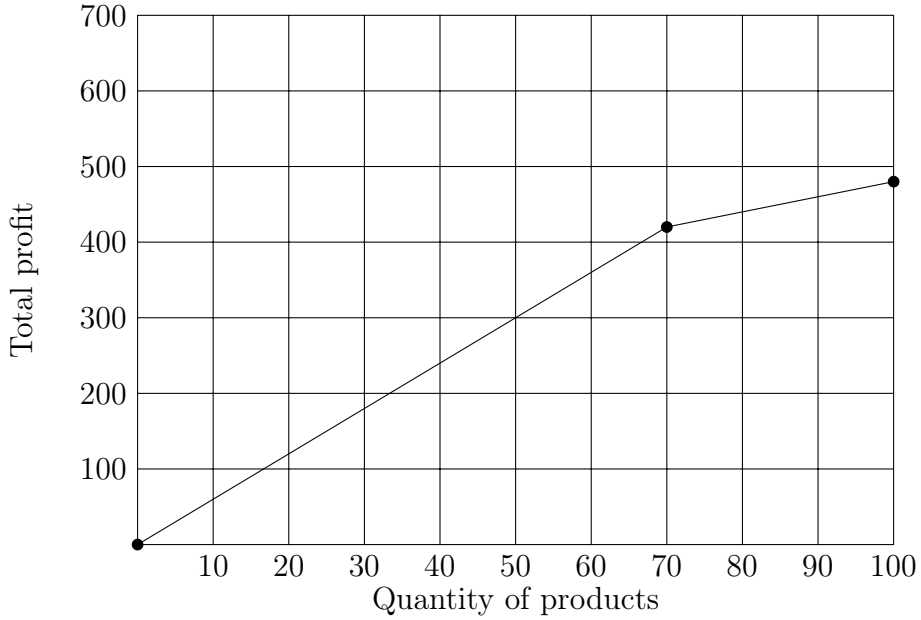



Figure 3.4: Piecewise linear profit function

It will also be interesting to study the NPV objective function with financial closure dates (e.g. quarters) to analyze the impact of closure dates and the end-of-horizon effects. In an NPV objective function with closure dates, instead of having a depreciation at every macro-period, the discount on profits occurs at some important financial macro-periods $t_j \in \mathcal{J}$, milestones such as the end of the month or end of the quarter, where facilities generally have financial commitment. Let us assume that there cannot be two milestones in the same macro-period. Thus, the objective can be written as follows:

$$\max \quad \sum_{p\in\mathcal{P}} \sum_{t_j\in\mathcal{J}} \overline{\beta_j} \Big( \sum_{t_{j-1}+1}^{t_j} G_p \times Y_{pt}^{out} + \sum_{\substack{s\in\mathcal{S} \\ tf_s \le t_{j+1} \\ ts_s \ge t_j}} h_{ps} \times I_{ps} + b_{ps} \times B_{ps} \Big) \tag{3.11}$$

where $\overline{\beta_j} = \prod_{k=1}^{j} \beta_k$ and $t_{j=0} = 0 \in \mathcal{J}$.

An alternative to a NPV objective function with closure dates, that reproduces the increase of inventory just before the closure dates, is to embed the NPV with a weekly actualization on a rolling horizon where the last periods considered are the closure dates. However, due to the length of the cycle times, it means that only a quarter could be used as a closure date.

Note that other objective functions could be considered. For example, let us consider an indicator which takes into account that machines are expensive and that managers want to use them as much as possible in order to get a good return on investment. The indicator could be the {Utilization time of Machine over the planning horizon} / {Machine price}. By maximizing this indicator in the objective function, it may increase the productivity but

the resulting plans could be different from the ones of the objective functions studied in this chapter.

It can be noticed in this chapter, that the different models are compared using indicators which are not specifically optimized. Even if profits, number of finished products and number of "moves" seems strongly correlated, other solutions (with the same objective value) could show better indicators. To clarify the advantages and drawbacks of the different objective functions, lexicographic optimization could be used to determine the best possible value for each indicator.

Finally, a strong limit of our models is the use of Fixed Lead Time constraints. The limits of Fixed Lead Times are discussed in Chapter 4, and alternative lead time constraints are proposed.

# Chapter 4

# Flexible lead time in production planning

## 4.1 Introduction

Unlike Chapter 3, where different objectives have been studied, this chapter focuses on the constraints used to model productions flows. Fixed lead times are certainly the most common and easiest way to model congestion, but they suffer from some well-known drawbacks that have pushed research towards iterative approaches and Clearing Functions. In this chapter, we present another more flexible way to model lead times. In Section 4.2, fixed lead time constraints and their drawbacks are discussed. In Section 4.3, new constraints called flexible lead time constraints which replace fixed lead time constraints, are proposed. In Section 4.4, computational experiments are discussed, where different lead time profiles (fixed and flexible) are compared on productivity, financial and cycle time indicators.

## 4.2 Drawback of fixed lead times

In Chapter 2.4, it was pointed out that the simplest way to model congestion is to use fixed lead times, and that linear programming models with fixed lead time constraints can be solved quickly. However, fixed lead times fail to take into account the circularity between the quantities to be processed over a planning horizon and the induced capacity congestion. Indeed, the relationship between the machine workloads and the actual lead times is not explicitly modeled. These are not the only drawbacks of fixed lead time constraints (3.4).

First, note that the speed of resolution of models with fixed lead time constraints (developed in Chapter 3) is due to the fact that all production flows are determined by the production releases. Due to Constraints (3.2) and (3.4), if a product is released in the facility in period $t$, then operation $l$ of the product will precisely be processed in period $t + \sum_{\lambda=1}^{l} LT_\lambda$. Hence, there are "only" $TP$ important decision variables. However, this limited number of decision variables considerably limits the operational decisions at the scheduling level. Linear programming models with fixed lead time constraints can hardly take into account changes in the product mix. More precisely, the fixed lead times usually remain the same whether many products or very few products are being processed in the factory. This is a classical drawback of MRP (Material Requirements Planning).

Second, the workshop capacity $C_k$ is only consumed in period $t + LT_l$, where $l \in \mathcal{L}^k$ is an operation processed in workshop $k$ and $t$ is the period when a quantity of products arrived

in the waiting queue for operation $l$. This means that, during $LT - 1$ periods (from $t$ to $t + LT_l - 1$), not a single product is assumed to be processed. This could be realistic in some cases where products are being transported and waiting before being processed, but this is irrelevant for most operations. The workload should be spread over the lead time, i.e. from $t$ to $t + LT_l$, rather than stressing the workshop capacity on a single period.

The combination of the two preceding effects can lead to an unnecessary reduction of the production volumes, in particular if short fixed lead times are imposed. Because fixed lead times must be satisfied, i.e. fixed lead time constraints are hard constraints, the only adjustment variables are the quantities to produce.

Note also, that, if no history of past inputs is given, the first periods cannot be constrained with fixed lead times larger than one period.

## 4.3 Flexible lead times

In this section, we introduce flexible lead time constraints that allow more flexibility in production planning models than fixed lead time constraints. The principle of flexible lead times is first discussed in Section 4.3.1. Then, in Section 4.3.2, flexible lead times are modeled using flexible lead time constraints, formerly called WIP penetration constraints.

### 4.3.1 Principle

To answer most of their biases, fixed lead times are relaxed by considering minimum lead times. With minimum lead times, products are not anymore forced to be processed and go directly to the following operation after the end of the lead time, but can wait in the Work-In-Process (WIP).

With flexible lead times, at any time, production capacity can focus on one product (with high demand) during one period and on other products in subsequent periods. This is not possible with fixed lead times, where each product must strictly follow the production pace imposed by the lead times. Thus, using flexible lead times, it is possible to release as many products as necessary in the factory while, when fixed lead times are imposed, the release quantities are highly constrained by the production flows and how the lead times are determined. In addition, only imposing minimum lead times means that products released in a period may be processed during multiple periods after the minimum lead time. Hence, the machine workload can be distributed over time, and be smoothed over the planning horizon.

To sum up, ensuring a minimum lead time without constraining the maximum lead time to be identical allows the lead time to be flexible. Flexible lead times will be easier to satisfy without disrupting the production flows.

### 4.3.2 Modeling

To the best of our knowledge, only two papers, (Hwang and Chang; 2003) and (Chen et al.; 2010), use constraints similar to the minimum lead time constraints discussed in Section 4.3.1. These constraints, called WIP penetration constraints, limit the number of operations of a product that can be performed in a single period. With the right parameters, WIP penetration constraints can model the minimum lead times discussed earlier, but can also model minimum lead times on several consecutive operations.

The first aim of WIP penetration constraints is to limit the flow of a product, by limiting the number of operations that can be performed in a single period. In the following, these constraints are called "flexible lead time constraints". Let us introduce $o_{\max}(l)$, which is the maximum number of operations before operation $l$ ($l$ included) which can be completed in the same period as $l$. If there is no such limit, $o_{\max}(l)$ is set to $+\infty$. Constraints (4.1) below are the flexible lead time constraints.

$$Y_{plt} \leq \sum_{j=l-o_{\max}(l)}^{l} W_{pj(t-1)} \quad \forall t \in \{1,...,T\} \quad \forall p \in \{1,...,P\} \quad \forall l \in \mathcal{L}_p \quad \text{s.t. } o_{\max}(l) \neq +\infty \tag{4.1}$$

Constraints (4.1) bind the output of operation $l$ with the Work-In-Process of the previous operations, i.e. products which have not yet completed operation $l - o_{\max}(l)$ cannot be processed in operation $l$.

If $o_{\max}(l) = 0$, Constraints (4.1) ensure that only products already in the WIP of operation $l$ can be processed, i.e. products will have to wait at least one period in the WIP of $l$, which is a relaxation of the fixed lead time when $LT = 1$ (according to our industrial data, a large majority of operations have a lead time smaller than 1).

In the model with fixed lead times (3.1)-(3.8), Constraints (3.4) are replaced by Constraints (4.1).

**Property 1.** *For a product $p$, an operation $l \in \mathcal{L}_p$ and a period $t$, Constraint 4.1 with $o_{max} = 0$ is a relaxation of Constraint 3.4 with $LT(l) = 1$.*

*Proof.* By mathematical induction on period $t$, it is possible to prove that, for product $p$ at operation $l$, Constraint (4.1) when $o^{\max}(l) = 0$ is a relaxation of Constraint (3.4) when $LT = 1$.
*Base case:* When $t = 1$, because $LT = 1$, the initial WIP should be processed at period $t = 1$ but the first release products cannot. Thus $Y_{plt} = W_{pl(t-1)}$, which implies that $Y_{plt} \leq W_{pl(t-1)}$.
*Step case:* Let $t = \tau$ and assume the induction hypothesis is true for $t = \tau - 1$. Because of Constraint (3.4), $Y_{plt} = X_{pl(t-1)}$. Then, using Constraint (3.3), $X_{pl(t-1)} = Y_{pl(t-1)} - W_{pl(t-2)} + W_{pl(t-1)}$, thus $Y_{plt} = Y_{pl(t-1)} - W_{pl(t-2)} + W_{pl(t-1)}$. By the induction hypothesis, $Y_{pl(t-1)} \leq W_{pl(t-2)}$, and thus $Y_{plt} \leq W_{pl(t-1)}$. This concludes the proof. □

When $LT = 0$, there are two possible relaxations. The first one is to set $o_{\max}(l) = +\infty$, and the second one is to remove the constraint. To reduce the size of the model, the second option is chosen.

Moreover, Constraints (4.1) provide additional flexibility. By adjusting the parameter $o_{\max}(l)$, the limits on production flows can be relaxed to include the WIP of previous operations. Figure 4.1 shows the possible production flows that can be processed in operation $l$ in a single period when $o_{\max} > 0$, i.e. the products that are in the WIP of the $o_{\max}$ operations before $l$ ($l$ included). Products in the WIP of operation $l - o_{\max} - 1$ and of operations earlier in the route cannot be processed in operation $l$ in a single period. A study of this additional flexibility can be found in Section 4.4.3.

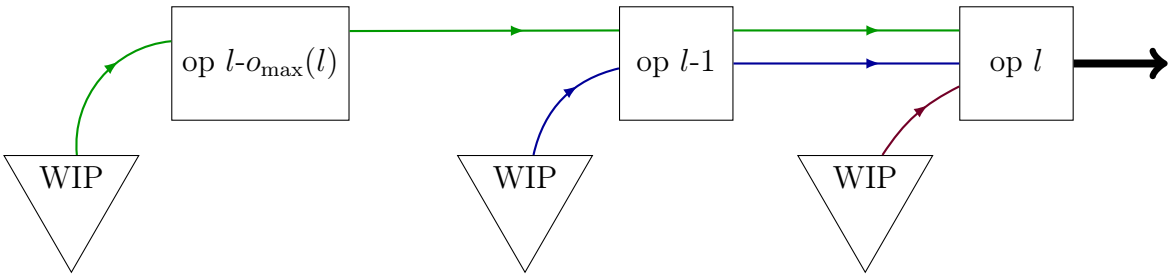



Figure 4.1: Possible production flows processed in operation $l$ in a single period

With flexible lead time constraints, many different models can be designed according to the lead time profile used. In the following, 1 fixed lead time profile and 4 flexible lead time profiles have been studied.

1. The first profile, $\mathcal{P}_{LT}^{\text{fixed}}$, corresponds to the classical fixed lead times of Chapter 3.

2. The second profile, $\mathcal{P}_{LT}^{\text{flex}}$, corresponds to the flexible lead times with parameters based on profile $\mathcal{P}_{LT}^{\text{fixed}}$, where products can wait in every operation as many periods as necessary. The lead time constraints are relaxed as follows: If the fixed lead time for operation $l$ is equal to 0, then no constraint is used, otherwise a flexible lead time constraint is introduced with $o_{\max}(l) = 0$.

3. Profile $\mathcal{P}_{LT}^{\text{flex}(+1)}$, resp. $\mathcal{P}_{LT}^{\text{flex}(+2)}$, is similar to profile $\mathcal{P}_{LT}^{\text{flex}}$ except that, rather than having $o_{\max}(l) = 0$, $o_{\max}(l)$ is set to 1, resp. 2. This means that only the WIP of the last two, resp. three, operations can be processed in the operation where the lead time was (strictly) positive. The impact of varying this parameter is studied in Section 4.4.3.

4. The last profile, $\mathcal{P}_{PT}^{\text{flex}}$, corresponds to flexible lead times, but it is based on the actual processing times, i.e. it is not related to the four other lead time profiles. With profile $\mathcal{P}_{PT}^{\text{flex}}$, production flows are only limited by the maximum number of operations for a product that can be completed in a period, according to the cumulative process times of these operations. In a sense, it is a relaxation of the previous model where delays are not induced by exogenous parameters, but only by physical constraints. It is used in Section 4.4.2 as the perfect plan where no congestion occurs.

## 4.4 Computational experiments

The design of our computational experiments is first introduced in Section 4.4.1. Then, the results obtained with fixed lead time constraints and flexible lead time constraints are compared in Section 4.4.2. Finally, in Section 4.4.3, we study, in a model with flexible lead time constraints, the impact of the parameter $o_{\max}(l)$, which introduces more flexibility in the production flows but also change the cycle times.

### 4.4.1 Design of experiments

The five lead time profiles previously discussed are studied in this part.

The objective function (3.10) maximizing profit is considered with an actualization rate $\beta = 1$ to mitigate the costs in the first periods induced by the lack of initial WIP. The costs

are the ones in Table 1.2. Note that WIP management costs are needed to prevent models with flexible lead times from introducing unnecessary products that may impact indicators such as cycle times and machine utilization rates.

Note that we decided to only use the industrial instances in the experiments of this section.

As indicated in Chapter 3, the initial WIP of the industrial data set induces unfeasible solutions with the fixed lead time profile (which is not the case with the flexible lead time profiles). For this reason, we first study the different lead time profiles without an initial WIP (leading to large backlog costs). Then, the lead times profiles are studied with an initial WIP that may not match the demand of the first periods. The planning horizon includes 91 micro-periods (equivalent to 13 macro-periods, approximately three months).

Note that a "real cycle time" (rCT) indicator is introduced. The real cycle time is computed as the sum of the average "real lead times" computed for each operation. The "real lead time" is the number of periods that are needed, in the optimized production plans, to process a quantity of products that is released and the current WIP.

Note that the inherent complexity of each lead time profile should not be overlooked. The average computational times of the solver for profiles $\mathcal{P}_{LT}^{\text{fixed}}$, $\mathcal{P}_{LT}^{\text{flex}}$ and $\mathcal{P}_{PT}^{\text{flex}}$, summarized in Tables 4.1 and 4.2, are quite different. There is at least a factor 10 between the computational times of each profile, making $\mathcal{P}_{PT}^{\text{flex}}$ the hardest profile to solve.

Table 4.1: Computational times (in seconds) for lead time profiles without initial WIP

| | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|
| **Avg. comput. time** | 75 | 1,391 | 18,926 |

When considering an initial WIP, the computational time slightly increases for profiles $\mathcal{P}_{LT}^{\text{flex}}$ and $\mathcal{P}_{PT}^{\text{flex}}$, with an increase of respectively 29% and 9%. Regarding fixed lead times, no increase in computational time is observed.

Table 4.2: Computational times (in seconds) for lead time profiles with initial WIP

| | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|
| **Avg. comput. time** | 75 | 1,789 | 20,723 |

### 4.4.2 Comparison of fixed and flexible lead times

#### Production indicators

Profiles $\mathcal{P}_{LT}^{\text{fixed}}$, $\mathcal{P}_{LT}^{\text{flex}}$ and $\mathcal{P}_{PT}^{\text{flex}}$ are compared using the industrial instances in the different demand scenarios. Tables 4.3, 4.4 and 4.5 show, for the three lead time profiles and without initial WIP: The total profit, the total output of finished products, the number of "moves", the average and the standard deviation of the workshop utilization rates. Tables 4.6, 4.7 and 4.8 are equivalent to Tables 4.3, 4.4 and 4.5, but when there is an initial WIP.

First, note that, even if profile $\mathcal{P}_{LT}^{\text{fixed}}$ always leads to negative profits without initial WIP, in a perfect world represented by $\mathcal{P}_{PT}^{\text{flex}}$, large profits can be reached. A notable remark is that the flexible relaxation of profile $\mathcal{P}_{LT}^{\text{flex}}$ allows large cost reduction (at least divided by a factor 2) and even positive profits in the case of low demand when there is no initial WIP. When an initial WIP is considered, the profit increase is huge, multiplied by more than 4

Table 4.3: Comparison of fixed and flexible lead times without initial WIP and high demand

| | | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|---|
| **Total profit** | | -205,459 | -126,070 | 216,270 |
| **Total output** | | 3,130 | 4,044 (+29%) | 4,539 (+45%) |
| # **moves** | | 535,688 | 597,441 | 672,525 |
| **Utilization rate** | **Mean** | 46.6% | 52.5% | 58.8% |
| **of workshops** | **Std. dev.** | 21.8% | 24.2% | 17.8% |

Table 4.4: Comparison of fixed and flexible lead times without initial WIP and medium demand

| | | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|---|
| **Total profit** | | -172,483 | -93,128 | 228,400 |
| **Total output** | | 3,091 | 4,026 (+30%) | 4,812 (+56%) |
| # **moves** | | 534,546 | 597,856 | 679,491 |
| **Utilization rate** | **Mean** | 46.5% | 52.5% | 59.6% |
| **of workshops** | **Std. dev.** | 21.8% | 24.5% | 18.0% |

Table 4.5: Comparison of fixed and flexible lead times without initial WIP and low demand

| | | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|---|
| **Total profit** | | -133,077 | -52,554 | 243,481 |
| **Total output** | | 3,064 | 4,055 (+32%) | 5,137 (+68%) |
| # **moves** | | 534,709 | 598,673 | 688,112 |
| **Utilization rate** | **Mean** | 46.4% | 52.6% | 60.5% |
| **of workshops** | **Std. dev.** | 21.9% | 24.6% | 17.3% |

Table 4.6: Comparison of fixed and flexible lead times with initial WIP and high demand

| | | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|---|
| **Total profit** | | -16,069 | 73,955 | 285,543 |
| **Total output** | | 3,626 | 4,763 (+31%) | 5,960 (+64%) |
| # **moves** | | 554,338 | 610,158 | 699,992 |
| **Utilization rate** | **Mean** | 48.5% | 54.1% | 62.1% |
| **of workshops** | **Std. dev.** | 17.8% | 21.6% | 15.6% |

Table 4.7: Comparison of fixed and flexible lead times with initial WIP and medium demand

| | | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|---|
| **Total profit** | | 2,602 | 97,551 (+3,649%) | 296,886 (+11,310%) |
| **Total output** | | 3,740 | 4,958 (+33%) | 6,200 (+66%) |
| # **moves** | | 556,639 | 617,218 | 706,005 |
| **Utilization rate** | **Mean** | 48.7% | 54.8% | 62.8% |
| **of workshops** | **Std. dev.** | 17.9% | 21.1% | 14.9% |

Table 4.8: Comparison of fixed and flexible lead times with initial WIP and low demand

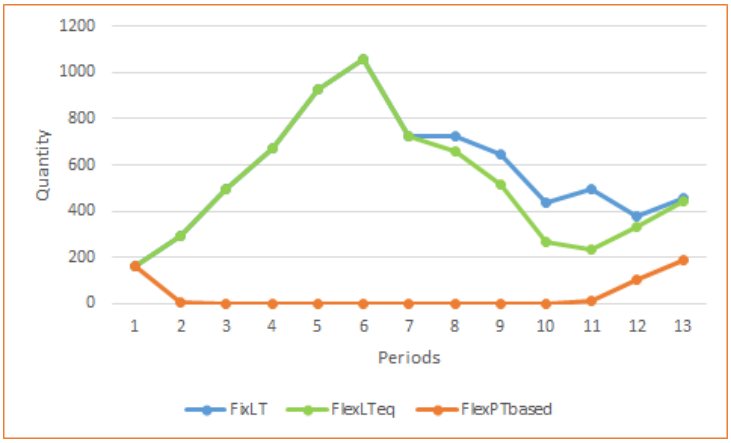
| | | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|---|
| **Total profit** | | 22,224 | 126,414 (+469%) | 310,844 (+1299%) |
| **Total output** | | 3,888 | 5,308 (+37%) | 6,553 (+68%) |
| **# moves** | | 555,150 | 627,702 | 713,628 |
| **Utilization rate of workshops** | **Mean** | 48.8% | 55.9% | 63.7% |
| | **Std. dev.** | 18.2% | 20.8% | 12.5% |

when they are positive. This increase can be linked to the increase of the number of finished products (close to 30%), which is half of the maximal number of finished products observed with $\mathcal{P}_{PT}^{\text{flex}}$. The increase of the number of finished products is followed by an increase of the number of "moves" and of the workshop utilization rates. Note that the small average utilization rate of workshops (about 50% of the capacity) hides the fact that the utilization rate significantly varies from one workshop to another. Critical workshops are very much used, and some others very little. For example, in the scenario with medium demand, without initial WIP and when considering profile $\mathcal{P}_{LT}^{\text{fixed}}$, the most used workshop has an average utilization rate of 84%, while the average utilization rate of the least used workshop is only 9%. The average utilization rates of other workshops range between 40% and 70%. The standard deviation only reflects the variation of the utilization rate on the horizon and not between the workshops. Profile $\mathcal{P}_{LT}^{\text{flex}}$ also leads to a slightly larger variability in the use of the workshops, which can be attributed to the flexibility of the production plan.

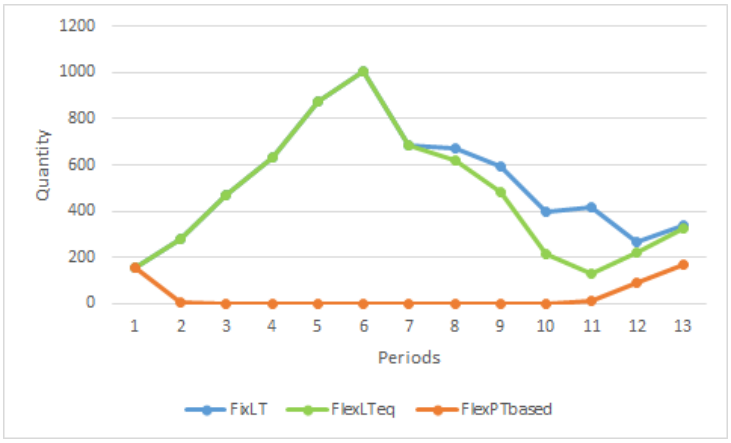
To analyze how the costs are decreasing, let us focus on the backlogged quantities. The inventory is very low, except in the last period with every lead time profile when there is no initial WIP. In fact, with flexible lead times, because WIP management costs are much smaller than inventory costs, it is almost always better to stop products a few operations before they are completed, in order to avoid the inventory costs. The only reason to pay inventory costs is if the workshops used by the last operations are saturated during the macro-period of the demand. The only large inventories are observed for $\mathcal{P}_{LT}^{\text{fixed}}$ when there is an initial WIP, but these inventories are mainly due to the initial WIP that does not match with the demand. Note that this is not a problem with flexible lead time profiles, since there is no limit on how much the WIP can wait at an operation. However, the main cause of the profit increase is the increase of the number of finished products that was already discussed, although the backlog reduction is not negligible. Figures 4.2a, 4.2b and 4.2c show the backlogged quantities on the planning horizon when there is no initial WIP, where "FixLT" corresponds to profile $\mathcal{P}_{LT}^{\text{fixed}}$, "FlexLTeq" to profile $\mathcal{P}_{LT}^{\text{flex}}$ and "FlexPTbased" to profile $\mathcal{P}_{PT}^{\text{flex}}$. Figures 4.3a, 4.3b and 4.3c are equivalent when there is an initial WIP. In each figure, the backlog linearly increases until the sixth macro-period for profiles $\mathcal{P}_{LT}^{\text{fixed}}$ and $\mathcal{P}_{LT}^{\text{flex}}$. The largest backlog ends on the scenario, ranging from 900 (Low demand) to 1100 (High demand) without an initial WIP and from 500 to 600 with an initial WIP. Note that, with profile $\mathcal{P}_{LT}^{\text{flex}}$, the backlog begins to decrease in exactly the same macro-period as profile $\mathcal{P}_{LT}^{\text{fixed}}$, but the decrease is much more impressive. About $\mathcal{P}_{PT}^{\text{flex}}$, the backlog is only in the first period, which explains why the profits are much higher.

### Cycle times

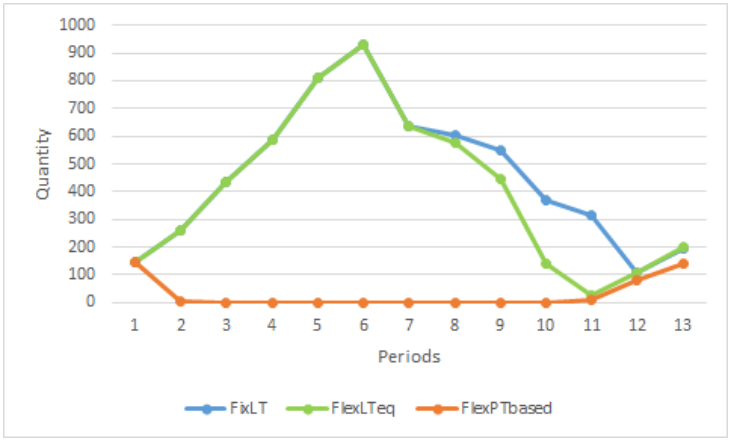
Finally, the production flows and their cycle times have not yet been discussed. Tables 4.9, 4.10 and 4.11 show, for every product, the mean of the real cycle times and the number of




(a) High demands

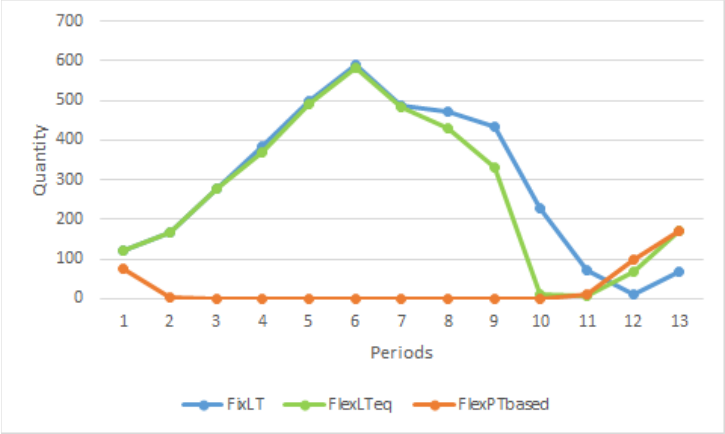



(b) Medium demands

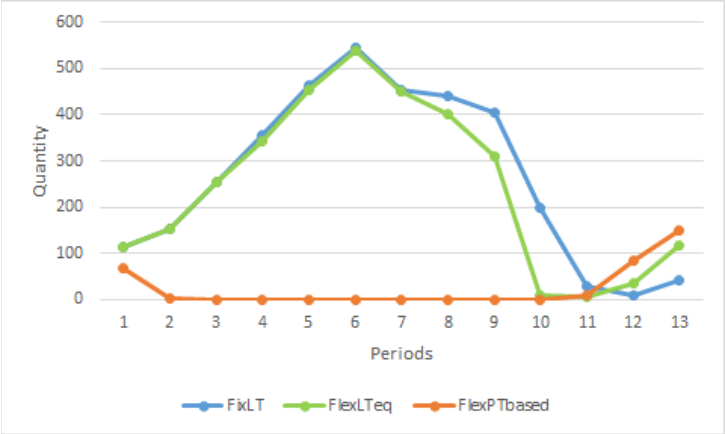



(c) Low demands

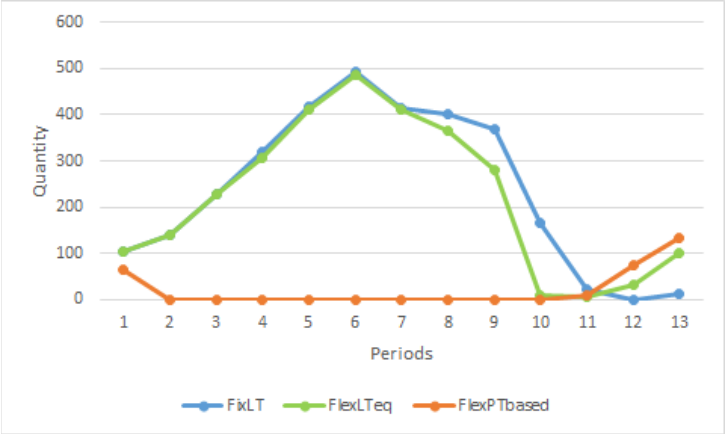
Figure 4.2: Weekly backlogged quantities for profiles $\mathcal{P}_{LT}^{\text{fixed}}$, $\mathcal{P}_{LT}^{\text{flex}}$ and $\mathcal{P}_{PT}^{\text{flex}}$ with industrial instance C200 without initial WIP.




(a) High demands




(b) Medium demands




(c) Low demands

Figure 4.3: Weekly backlogged quantities for profiles $\mathcal{P}_{LT}^{\text{fixed}}$, $\mathcal{P}_{LT}^{\text{flex}}$ and $\mathcal{P}_{PT}^{\text{flex}}$ with industrial instance C200 with initial WIP.

products for which the real cycle time matches with the cycle time observed with fixed lead times (with a tolerated gap of one micro-period). Note that the cycle times of products for $\mathcal{P}_{PT}^{\text{flex}}$ can be smaller than the cycles times of products when fixed lead times are considered, i.e. for $\mathcal{P}_{LT}^{\text{flex}}$. Tables 4.12, 4.13 and 4.14 are equivalent to Tables 4.9, 4.10 and 4.11 with an initial WIP.

Note that, with an initial WIP, it is not possible to compute the cycle time of several products because the initial WIP is enough to meet the demand. Another issue occurring with or without initial WIP is that some products are not even produced, thus, depending on the profile of lead times, the mean of the cycle times does not exactly include the same products. Note also that the average real cycle time may not be the most accurate indicators of the changes in the production flows, because it does not take into account the quantities of products associated with each cycle time. A weighted average of real cycle times (weighted by the quantities of finished products) would be a more relevant indicator.

Table 4.9: Comparison of cycle times for fixed and flexible lead times without initial WIP and high demand

| | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|
| **Average rCT** | 54.7 | 72.8 | 11.3 |
| **Nb products with minimum CT** | 70/70 | 9/70 | 0/70 |

Table 4.10: Comparison of cycle times for fixed and flexible lead times without initial WIP and medium demand

| | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|
| **Average rCT** | 54.9 | 71.4 | 12.1 |
| **Nb products with minimum CT** | 70/70 | 9/70 | 0/70 |

Table 4.11: Comparison of cycle times for fixed and flexible lead times without initial WIP and low demand

| | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|
| **Average rCT** | 54.9 | 74.5 | 12.9 |
| **Nb products with minimum CT** | 70/70 | 8/70 | 0/70 |

Table 4.12: Comparison of cycle times for fixed and flexible lead times with initial WIP and high demand

| | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|
| **Average rCT** | 54.2 | 71.4 | 23.1 |
| **Nb products with minimum CT** | 70/70 | 5/70 | 0/70 |

The average cycle time for profile $\mathcal{P}_{LT}^{\text{fixed}}$ is approximately equal to 55 micro-periods, while it is equal to 70 micro-periods for profile $\mathcal{P}_{LT}^{\text{flex}}$. Fixed lead time constraints ensure that the expected cycle times are exactly met. The average cycle time significantly increases with profile $\mathcal{P}_{LT}^{\text{flex}}$, and only 10% of the products reach the cycle time of fixed lead times. This

Table 4.13: Comparison of cycle times for fixed and flexible lead times with initial WIP and medium demand

| | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|
| **Average rCT** | 54.1 | 70.8 | 24.8 |
| **Nb products with minimum CT** | 70/70 | 4/70 | 0/70 |

Table 4.14: Comparison of cycle times for fixed and flexible lead times with initial WIP and low demand

| | $\mathcal{P}_{LT}^{\text{fixed}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
|---|---|---|---|
| **Average rCT** | 54.8 | 68.9 | 24.9 |
| **Nb products with minimum CT** | 70/70 | 5/70 | 1/70 |

means that, by lengthening the cycle time of specific products, it is possible to meet more demands. This is because the additional flexibility in the production flows helps to better use the production capacity.

It can also be seen that, with profile $\mathcal{P}_{PT}^{\text{flex}}$, the average cycle time ranges between 25% (without initial WIP) and 50% (with initial WIP) of the average cycle time of profile $\mathcal{P}_{LT}^{\text{fixed}}$. This is what could be expected if only process times are considered. In a sense, it matches the industrial reality where the ratio $\frac{\text{Cycle time}}{\sum \text{process times}}$, called Xfactor, is always larger than 3.

We can conclude that the flexible lead times better handle the backlogs and inventories that are due to the initial WIP that is inadequate to meet the demand. Thus, flexible lead times are more appropriate than fixed lead times when demands are changing (or the forecasts are updated), which makes it relevant in semiconductor manufacturing.

### 4.4.3 Analysis of the impact of parameter $o_{\max}$

In this section, we compare the results of profiles $\mathcal{P}_{LT}^{\text{flex}}$, $\mathcal{P}_{LT}^{\text{flex}(+1)}$ and $\mathcal{P}_{LT}^{\text{flex}(+2)}$ when the parameter $o_{\max}$ is modified. The goal is to study the impact of the flexibility associated with allowing more operations in a period on total profit and productivity.

#### Profit and productivity indicators

Tables 4.15, 4.16 and 4.17 (resp. Tables 4.18, 4.19 and 4.20) show the profit and productivity indicators for the numerical experiments without an initial WIP (resp. with an initial WIP).

Table 4.15: Variation of parameter $o_{\max}(l)$ without initial WIP and high demand

| | | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{LT}^{\text{flex}(+1)}$ | $\mathcal{P}_{LT}^{\text{flex}(+2)}$ |
|---|---|---|---|---|
| **Total profit** | | -126,070 | -82,913 | -22,281 |
| **Total output** | | 4,044 | 4,265 (+5%) | 4,184 (+3%) |
| **# moves** | | 597,441 | 608,832 | 625,143 |
| **Utilization rate of workshops** | **Mean** | 52.5% | 53.4% | 54.7% |
| | **Std. dev.** | 24.2% | 23.2% | 24.0% |

Table 4.16: Variation of parameter $o_{\max}(l)$ without initial WIP and medium demand

| | | $\mathcal{P}_{LT}^{\mathbf{flex}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+2)}}$ |
|---|---|---|---|---|
| **Total profit** | | -93,128 | -51,070 | 9,166 |
| **Total output** | | 4,026 | 4,200 (+4%) | 4,202 (+4%) |
| **# moves** | | 597,856 | 604,880 | 621,706 |
| **Utilization rate of workshops** | **Mean** | 52.5% | 53.1% | 54.4% |
| | **Std. dev.** | 24.5% | 23.8% | 24.0% |

Table 4.17: Variation of parameter $o_{\max}(l)$ without initial WIP and low demand

| | | $\mathcal{P}_{LT}^{\mathbf{flex}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+2)}}$ |
|---|---|---|---|---|
| **Total profit** | | -52,554 | -14,476 | 43,451 |
| **Total output** | | 4,055 | 4,155 (+2%) | 4,263 (+5%) |
| **# moves** | | 598,673 | 602,142 | 623,102 |
| **Utilization rate of workshops** | **Mean** | 52.6% | 53.0% | 54.6% |
| | **Std. dev.** | 24.6% | 23.9% | 24.1% |

Table 4.18: Variation of parameter $o_{\max}(l)$ with initial WIP and high demand

| | | $\mathcal{P}_{LT}^{\mathbf{flex}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+2)}}$ |
|---|---|---|---|---|
| **Total profit** | | 73,955 | 109,186 (+48%) | 147,980 (+100%) |
| **Total output** | | 4,763 | 4,835 (+2%) | 5,103 (+7%) |
| **# moves** | | 610,158 | 614,091 | 636,942 |
| **Utilization rate of workshops** | **Mean** | 54,1 | 54.5% | 56.4% |
| | **Std. dev.** | 21.6 | 20.8% | 20.8% |

Table 4.19: Variation of parameter $o_{\max}(l)$ with initial WIP and medium demand

| | | $\mathcal{P}_{LT}^{\mathbf{flex}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+2)}}$ |
|---|---|---|---|---|
| **Total profit** | | 97,551 | 131,558 (+35%) | 168,063 (+72%) |
| **Total output** | | 4,958 | 5,123 (+3%) | 5,359 (+8%) |
| **# moves** | | 617,218 | 622,279 | 645,117 |
| **Utilization rate of workshops** | **Mean** | 54.8% | 55.3% | 57.2% |
| | **Std. dev.** | 21.1% | 20.2% | 20.6% |

Table 4.20: Variation of parameter $o_{\max}(l)$ with initial WIP and low demand

| | | $\mathcal{P}_{LT}^{\mathbf{flex}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+2)}}$ |
|---|---|---|---|---|
| **Total profit** | | 126,414 | 158,919 (+26%) | 192,820 (+53%) |
| **Total output** | | 5,308 | 5,489 (+3%) | 5,660 (+7%) |
| **# moves** | | 627,702 | 632,433 | 652,575 |
| **Utilization rate of workshops** | **Mean** | 55.9% | 56.4% | 58.1% |
| | **Std. dev.** | 20.8% | 19.6% | 19.7% |

By allowing more flexibility on the lead time constraints, the total profit increases sharply. For example, with an initial WIP and a high demand and compared to $\mathcal{P}_{LT}^{\text{flex}}$, the total profit increases by nearly 50% with $\mathcal{P}_{LT}^{\text{flex}(+1)}$ and doubles with $\mathcal{P}_{LT}^{\text{flex}(+2)}$. However, the increase of the number of finished products is more limited (at most 8% with $\mathcal{P}_{LT}^{\text{flex}(+2)}$, medium demand and an initial WIP) and cannot explain the significant increase of the total profit. Again, the main reason is the reduction in backlog as shown in Figures 4.4a, 4.4b, 4.4c, 4.5a, 4.5b and 4.5c, where "FlexLTeq" stands for $\mathcal{P}_{LT}^{\text{flex}}$, "FlexLTeq+1" for $\mathcal{P}_{LT}^{\text{flex}(+1)}$ and "FlexLTeq+2" for $\mathcal{P}_{LT}^{\text{flex}(+2)}$. All backlog curves show the same linear increase in the first macro-period, except for $\mathcal{P}_{LT}^{\text{flex}(+2)}$ where the decrease starts one macro-period before the other profiles. The curves are somehow nested. While allowing more flexibility on the lead times, backlogs can decrease more quickly.

The number of "moves" and the average workshop utilization rate are correlated to the total output of finished products. When considering the standard deviation of the workshop utilization rates, nothing can be concluded.

### Cycle times

Tables 4.21, 4.22, 4.23, 4.24, 4.25 and 4.26 show the cycle times observed with flexible lead times and the number of products that have a cycle time close to the cycle time observed with fixed lead times. When considering an initial WIP, the average cycle time is shorter with profiles $\mathcal{P}_{LT}^{\text{flex}(+1)}$ and $\mathcal{P}_{LT}^{\text{flex}(+2)}$ than the average cycle time with $\mathcal{P}_{LT}^{\text{flex}}$. To explain these smaller cycle times, the tables also give the number of products with an average cycle time that is shorter than the cycle times given by the fixed lead times. Even without an initial WIP, where the cycle times are larger than those with profile $\mathcal{P}_{LT}^{\text{flex}}$, an increasing number of products have a shorter cycle time than the cycle time obtained with fixed lead times. This number is between 4 and 13 with profile $\mathcal{P}_{LT}^{\text{flex}(+1)}$ and doubles with profile $\mathcal{P}_{LT}^{\text{flex}(+2)}$. As expected, there is not such reduced cycle times with profile $\mathcal{P}_{LT}^{\text{flex}}$.

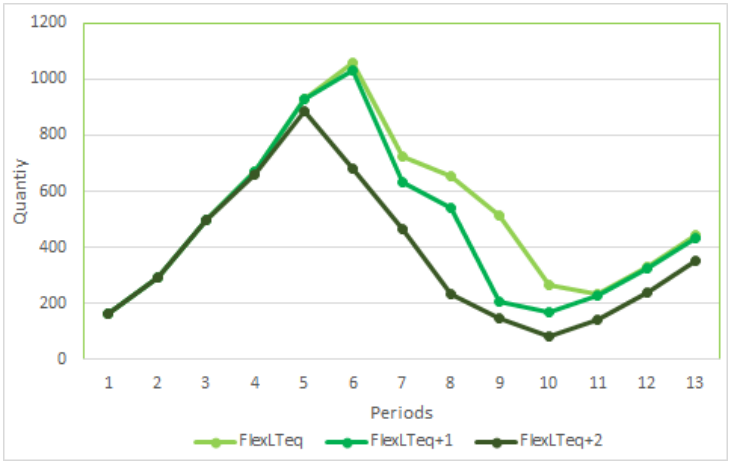
Table 4.21: Cycle times observed when varying $o_{\max}(l)$ without initial WIP and high demand

| | $\mathcal{P}_{LT}^{\textbf{flex}}$ | $\mathcal{P}_{LT}^{\textbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\textbf{flex(+2)}}$ |
|---|---|---|---|
| **Avg rCT** | 72.8 | 75.6 | 74.3 |
| **Nb products with minimum CT** | 9/70 | 7/70 | 3/70 |
| **Nb products with CT under minimum CT** | 0/70 | 7/70 | 13/70 |

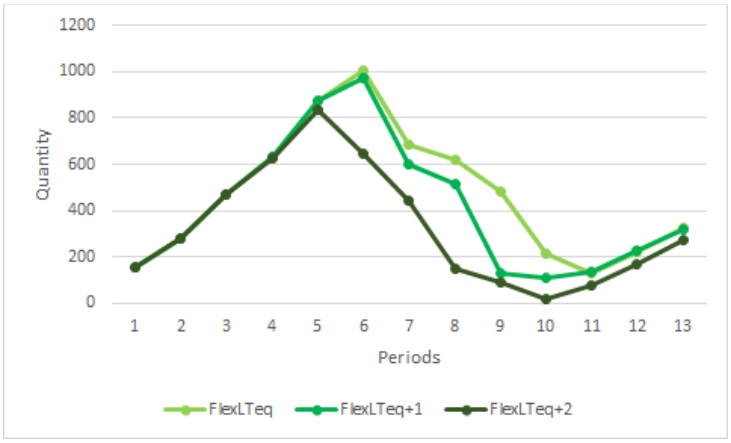
Table 4.22: Cycle times observed when varying $o_{\max}(l)$ without initial WIP and medium demand

| | $\mathcal{P}_{LT}^{\textbf{flex}}$ | $\mathcal{P}_{LT}^{\textbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\textbf{flex(+2)}}$ |
|---|---|---|---|
| **Avg rCT** | 71.4 | 79.9 | 72.3 |
| **Nb products with minimum CT** | 9/70 | 10/70 | 3/70 |
| **Nb products with CT under minimum CT** | 0/70 | 6/70 | 16/70 |

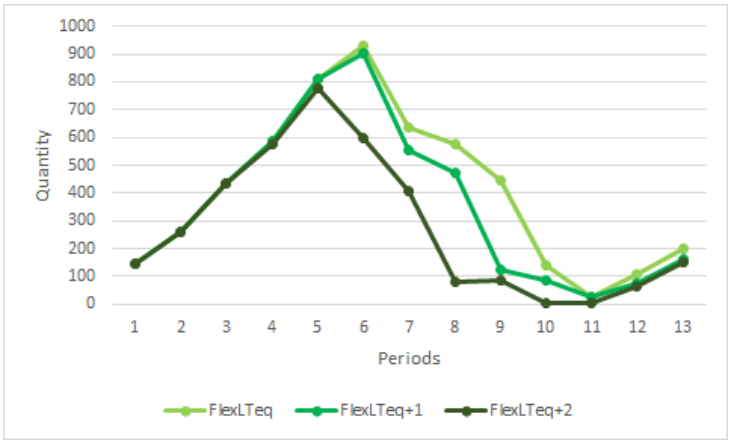
If the average cycle time decreases when $o_{\max}(l)$ increases, it is not only due to the added flexibility, but also to the lead time constraints that are too relaxed, thus leading to a




(a) High demands

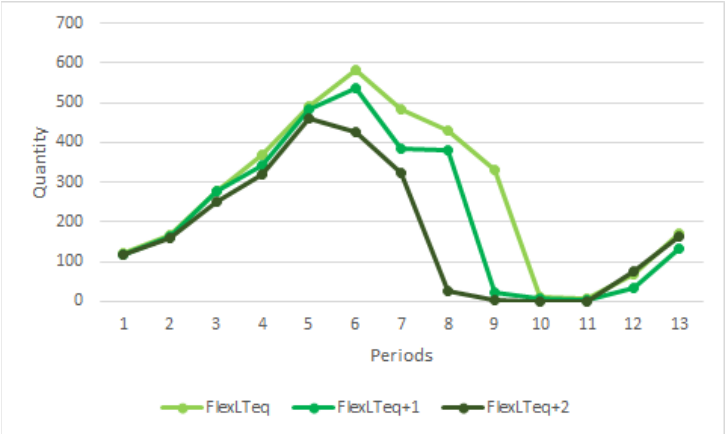



(b) Medium demands

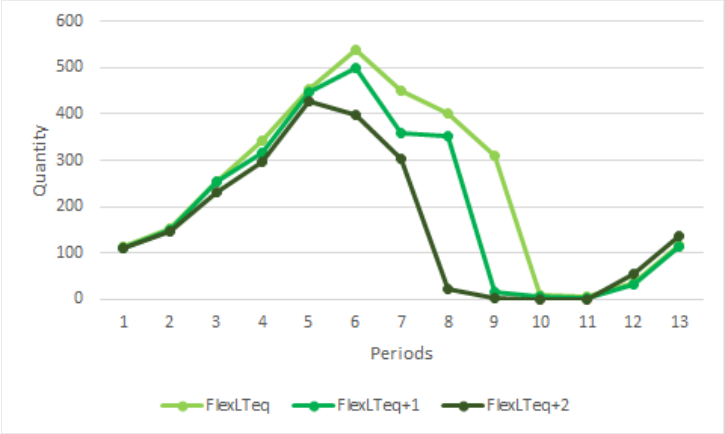



(c) Low demands

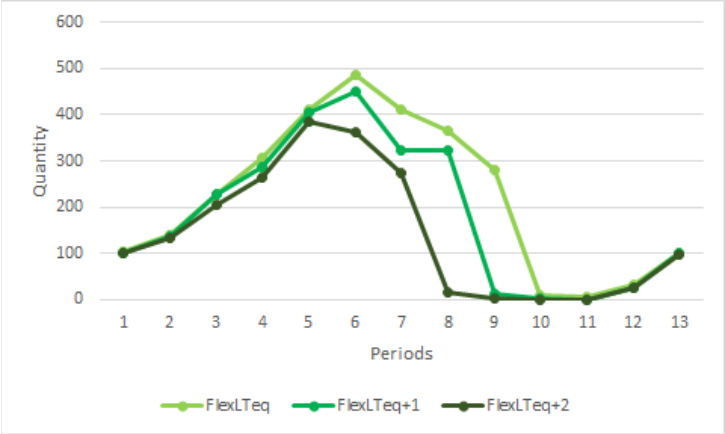
Figure 4.4: Weekly backlogged quantities for profiles $\mathcal{P}_{LT}^{\text{flex}}$, $\mathcal{P}_{LT}^{\text{flex}(+1)}$ and $\mathcal{P}_{LT}^{\text{flex}(+2)}$ with industrial instance C200 without initial WIP.




(a) High demands




(b) Medium demands




(c) Low demands

Figure 4.5: Weekly backlogged quantities for profiles $\mathcal{P}_{LT}^{\text{flex}}$, $\mathcal{P}_{LT}^{\text{flex}(+1)}$ and $\mathcal{P}_{LT}^{\text{flex}(+2)}$ with industrial instance C200 with initial WIP.

Table 4.23: Cycle times observed when varying $o_{\max}(l)$ without initial WIP and low demand

| | $\mathcal{P}_{LT}^{\mathbf{flex}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+2)}}$ |
|---|---|---|---|
| **Avg rCT** | 74.5 | 75.7 | 78.6 |
| **Nb products with minimum CT** | 8/70 | 11/70 | 6/70 |
| **Nb products with CT under minimum CT** | 0/70 | 4/70 | 12/70 |

Table 4.24: Cycle times observed when varying $o_{\max}(l)$ with initial WIP and high demand

| | $\mathcal{P}_{LT}^{\mathbf{flex}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+2)}}$ |
|---|---|---|---|
| **Avg rCT** | 71.4 | 61.2 | 59.9 |
| **Nb products with minimum CT** | 5/70 | 11/70 | 7/70 |
| **Nb products with CT under minimum CT** | 0/70 | 11/70 | 23/70 |

Table 4.25: Cycle times observed when varying $o_{\max}(l)$ with initial WIP and medium demand

| | $\mathcal{P}_{LT}^{\mathbf{flex}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+2)}}$ |
|---|---|---|---|
| **Avg rCT** | 70.8 | 63.8 | 60.3 |
| **Nb products with minimum CT** | 4/70 | 13/70 | 3/70 |
| **Nb products with CT under minimum CT** | 0/70 | 10/70 | 24/70 |

Table 4.26: Cycle times observed when varying $o_{\max}(l)$ with initial WIP and low demand

| | $\mathcal{P}_{LT}^{\mathbf{flex}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+1)}}$ | $\mathcal{P}_{LT}^{\mathbf{flex(+2)}}$ |
|---|---|---|---|
| **Avg rCT** | 68.9 | 64.1 | 61.1 |
| **Nb products with minimum CT** | 5/70 | 13/70 | 4/70 |
| **Nb products with CT under minimum CT** | 0/70 | 13/70 | 24/70 |

shorter minimum cycle time. The number of products with a cycle time shorter than expected increases with $o_{\max}(l)$. The reason might be that there are occurrences of overlapping flexible lead times. Figures 4.6 and 4.7 give an example of such phenomena. Figure 4.6 illustrates the production flows when two successive operations $l$ and $l+1$ have a positive lead time. A product that completes operation $l-1$ in period $t$ will, at the earliest, be processed in operation $l$ in period $t+1$, and therefore be processed in operation $l+1$ in period $t+2$. However, as shown in Figure 4.7, by increasing $o_{\max}$ and allowing one additional operation, production quantities might not wait one period in the WIP of operation $l+1$, leading to a reduction of one period of the minimum cycle time. More generally, the flexible lead times of two operations $l$ and $l'$, $l \neq l'$ overlap if their ranges $[l - o_{\max}(l), l]$ and $[l' - o_{\max}(l'), l']$ overlap.

The larger $o_{\max}(l)$, the larger the number of overlapping lead times. Thus, when establishing a flexible lead time profile, we should be careful of the minimum cycle time it induces.

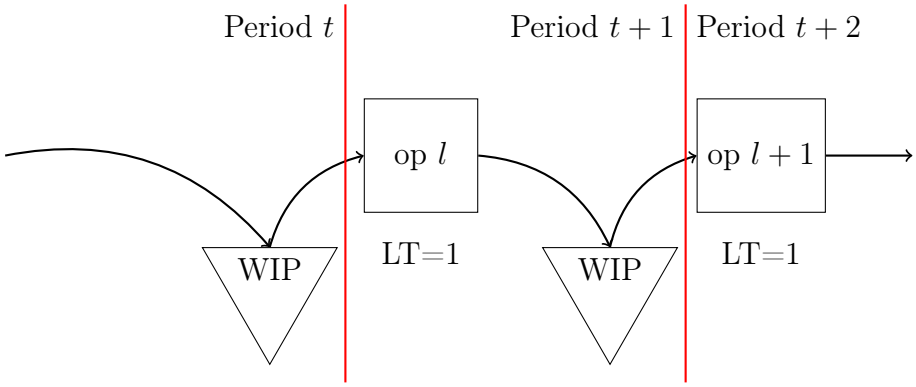



Figure 4.6: Production flows with a fixed lead time of 1 period for two consecutive operations (or equivalently with flexible lead times and $o_{\max} = 0$)

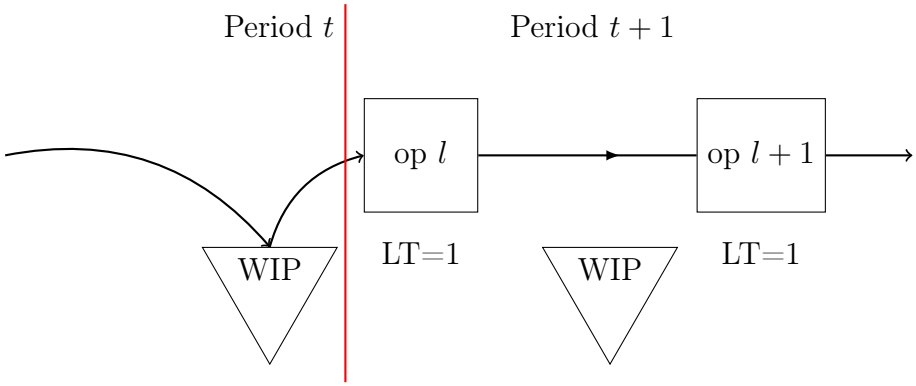



Figure 4.7: Possible production flows with flexible lead times and $o_{\max} = 1$

## 4.5 Conclusions and perspectives

In this chapter, we have introduced flexible lead times, which are an alternative and a relaxation of fixed lead times and which allow production flows to be more flexible. Flexible lead times help to increase productivity but also reduce costs or increase profits, although longer cycle times may be incurred. They make it possible to complete more finished products by better using production capacity. Another modeling advantage of considering flexible lead time constraints is that it is possible to adjust the degree of flexibility through the parameters $o_{\max}(l)$.

Depending on the setting of the parameters $o_{\max}(l)$ in the flexible lead time constraints, various flexible lead time profiles can be proposed. However, a first obstacle to the use of flexible lead times is the sharp increase in the computational times of linear programs with flexible lead time constraints. Another critical point to investigate is that the cycle times of products subject to flexible lead time constraints are not limited, and can therefore be very long and vary considerably for the same product. This may not be acceptable in an industrial context. Indeed, with the flexible lead time constraints proposed in this chapter, production flows cannot be traced and easily controlled as it is the case with fixed lead times. These problems are solved by the reformulation of our production planning problem proposed and solved in Chapter 5.

# Chapter 5

# Timed routes approaches for production planning

## 5.1 Introduction

As shown in Chapter 4, mathematical models with flexible lead times can take hours to be solved. To address the semiconductor production planning problem in reasonable computational times, we propose the concept of timed routes. In the timed route of a product, each production operation of the route of the product is assigned to a period in the planning horizon. Timed routes are designed so that it is easy to use a column generation approach when the number of timed routes is exponential in the mathematical model. Another drawback of allowing flexible lead times is that, because lead times have no upper bounds, products may remain in the manufacturing system too long and have very long cycle times. However, a major advantage of timed routes is that the cycle time of products can be limited, by only considering timed routes with a maximum number of periods between the period assigned to the first operation of the route and the period assigned to the last operation of the route.

This chapter is structured as follows. Section 5.2 completes the literature review of Chapter 2 with an overview of column generation approaches for production planning. Then, in Section 5.3, a reformulation based on the new concept of timed routes is proposed. When modeling flexible lead times instead of fixed lead times, the number of timed routes becomes exponential. Hence, a column generation approach is presented in Section 5.4 to solve the problem with flexible lead times. Computational experiments on industrial data are conducted in Section 5.5, that shows the efficiency of the timed route reformulation. Conclusions are drawn and future research directions are provided in Section 5.6.

## 5.2 Literature review on column generation for production planning

Column generation has been successfully applied to various optimization problems such as vehicle routing problems (e.g. Ceselli et al.; 2009), airplane crew scheduling problems (e.g. Gamache et al.; 1999) or machine scheduling problems (e.g. van Den Akker et al.; 1999). Column generation was introduced by Dantzig and Wolfe (1960), and consists in separating the original problem into a master problem and a pricing problem that generates useful columns for the master problem. At first, a Restricted Master Problem (RPM) with a limited number of columns is solved. Then, using reduced costs, the pricing problem is

solved to find one or several columns to add to the RPM. The process is iterated until no new column is found. To better understand column generation, the reader can refer to Barnhart et al. (1998), where different strategies of generation are discussed (in a branch and price framework) or the extensive tutorial of Desrosiers and Lübbecke (2005).

In production planning and lot sizing, the first work on column generation was published by Manne (1958), two years prior to the seminal paper of Dantzig and Wolfe (1960). Manne's paper is partially deficient and was corrected and implemented in Degraeve and Jans (2007). Column generation was applied to solve lot-sizing problems in several kinds of industries such as the tire industry (Jans and Degraeve; 2004), the paper industry (Bredström et al.; 2004) and the steel industry (Yi et al.; 2019). The most commonly used column type is a production plan column that specifies the production periods. However, in terms of production planning without set-up costs, the production periods are not critical. That is why the formulation proposed in our study is significantly different. As far as we know, column generation was never applied to solve a multi-product multi-step lot-sizing problem.

In semiconductor manufacturing, to the best of our knowledge, column generation was never used to solve production planning problems. Even in the entire semiconductor manufacturing literature, only four articles using column generation were spotted: On lot allocation to customers (Ng et al.; 2010), on cutting wafers (Nisted et al.; 2011), on capacity expansion (Kim and Uzsoy; 2008) and on scheduling (Jampani and Mason; 2010).

## 5.3 A novel formulation using timed routes

In this section, a reformulation of the mathematical models introduced in Chapters 3 and 4 is proposed. The new model is based on the new concept of "timed route" which is formalized in Section 5.3.1. Timed routes allow production flows to be fully modeled. The mathematical model using timed routes is introduced in Section 5.3.2. In Section 5.3.3, a polynomial time algorithm to generate all possible timed routes with fixed lead times is presented.

### 5.3.1 Concept of timed route

Note that Leachman and Carmon (1992) were the first to discuss a route based formulation of the semiconductor manufacturing production problem. Unfortunately, the authors discarded the idea due to the large number of decision variables required by the model.

A production route is the sequence of operations that a product needs to follow to be completed (see Figure 5.1). A timed route is a production route for which a processing period is assigned to each operation (see Figure 5.2). More formally, in a timed route $r$, a period $t(p, r, l)$ is assigned to each operation $l$ in the route of product $p$. For example, in Figure 5.2, the timed route starts at period $t$ and is completed at period $t(p, r, |\mathcal{L}_p|)$. Thus, the cycle time of a timed route $r$ of product $p$ is:

$$CT(p, r) = t(p, r, |\mathcal{L}_p|) - t(p, r, 1) + 1$$

With timed routes, it is possible to detail the production flows, and to know exactly where and when capacity is consumed. The cycle time related to a timed route is explicit, contrary to the classical lead time formulations of Chapter 3, where determining the cycle time means looking at the set of lead time constraints on the operation of the route to extract the total cycle time (and this is even more true with the flexible lead times of Chapter 4, where only the minimum cycle time can be computed). With the full view of possible production

flows, inconsistent or useless timed routes can be discarded. The timed routes could be validated based on industrial knowledge. Moreover, new constraints on production flows can be introduced such as minimum and maximum cycle times.

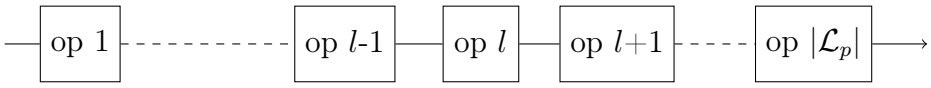



Figure 5.1: A production route

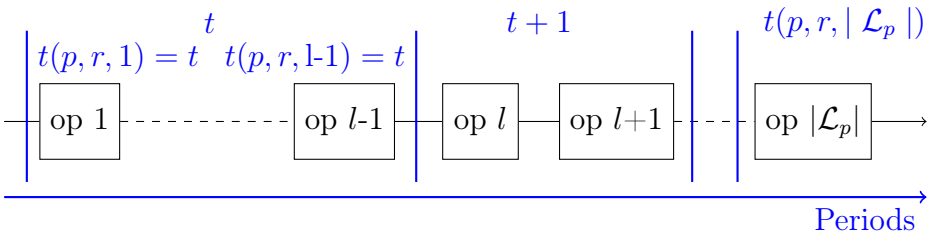



Figure 5.2: A timed route

### 5.3.2 Mathematical model

In the following, the timed route model is formalized. Let us denote $\mathcal{R}_p$ the set of timed routes of product $p$. With each timed route $r \in \mathcal{R}_p$, a Work in Process (WIP) management unitary cost $w_{pr}$ is associated. The WIP cost of a timed route is equivalent to the sum of the WIP costs of the different operations on the time horizon. Only the first operation of each period (except for the first period) carries a WIP cost. This WIP cost can be counted several times if no operation takes place in the subsequent periods. Let us write the total WIP cost of a given timed route $r$, $\sum_{l \in \mathcal{L}_p} b_l^{pr} w_{pl}$, where $b_l^{pr}$ is the number of periods between the processing periods of operation $l-1$ and operation $l$ in timed route $r$, i.e. $b_l^{pr} = t(p,r,l) - t(p,r,l-1)$. Note that waiting before the first operation of a route is not allowed, i.e. $b_1^{pr} = 0$.

Let $a_{lt}^{pr}$ be a binary parameter which is equal to 1 if, in timed route $r \in \mathcal{R}_p$ of product $p$, operation $l$ is processed in period $t$, and is equal to 0 otherwise. $Z_{pr}$ is the decision variable that corresponds to the quantity released on timed route $r$. Recall in table 5.1, the previous notation of parameters and variables that remain used.

The timed route formulation is given below.

| Variables | |
|---|---|
| $I_{ps}$ | Inventory level of product $p$ at the end of macro-period $s$ |
| $B_{ps}$ | Backlog level of product $p$ at the end of macro-period $s$ |
| Parameters | |
| $h_{ps}$ | Unit inventory cost of product $p$ at the end of macro-period $s$ |
| $b_{ps}$ | Unit backlog cost of product $p$ at the end of macro-period $s$ |
| $C_k$ | is the daily capacity of workshop $k$ |
| $\alpha_{pl}$ | Unit resource consumption of operation $l$ of product $p$ |
| $D_{ps}$ | Demand of product $p$ at the end of macro-period $s$ |

Table 5.1: Parameters and variables previously used

$$\min \quad \sum_{p=1}^{P} \sum_{r \in \mathcal{R}_p} w_{pr} Z_{pr} + \sum_{p=1}^{P} \sum_{s=1}^{S} (h_{ps} I_{ps} + b_{ps} B_{ps}) \tag{5.1}$$

$$\text{s.t.} \quad \sum_{p=1}^{P} \sum_{r \in \mathcal{R}_p} \sum_{l \in \mathcal{L}^k} a_{lt}^{pr} \alpha_{pl} Z_{pr} \leq C_k \qquad \forall k \in \{1, \ldots, K\} \quad \forall t \in \{1, \ldots, T\} \tag{5.2}$$

$$I_{ps} \geq \sum_{r \in \mathcal{R}_p} \sum_{\tau=1}^{tf_s} a_{|\mathcal{L}_p|\tau}^{pr} Z_{pr} - \sum_{\sigma=1}^{s} D_{p\sigma} \qquad \forall p \in \{1, \ldots, P\} \quad \forall s \in \{1, \ldots, S\} \tag{5.3}$$

$$B_{ps} \geq - \sum_{r \in \mathcal{R}_p} \sum_{\tau=1}^{tf_s} a_{|\mathcal{L}_p|\tau}^{pr} Z_{pr} + \sum_{\sigma=1}^{s} D_{p\sigma} \qquad \forall p \in \{1, \ldots, P\} \quad \forall s \in \{1, \ldots, S\} \tag{5.4}$$

$$Z_{pr}, I_{ps}, B_{ps} \geq 0 \qquad \forall p \in \{1, \ldots, P\} \quad \forall r \in \mathcal{R}_p \quad \forall s \in \{1, \ldots, S\} \tag{5.5}$$

The objective function (5.1) minimizes the total backlog, inventory and WIP management cost induced by the selected timed routes, which is equivalent to the objective function (3.1). Constraints (5.2) model the limit on capacity consumption in each workshop at every period, and correspond to Constraints (3.6). Constraints (5.3) and (5.4) ensure the inventory balance. They are equivalent to Constraints (3.5) but are written separately to simplify the writing of the dual problem. This formulation can be seen as a set covering problem.

### 5.3.3 Generation of timed routes associated with fixed lead times

Let us show how the set of timed routes is determined when fixed lead times are considered. Due to Constraints (3.2) and (3.4) in the model with fixed lead times, all production flows on a route follow the same pattern. If $t$ is the first period of the route and $|\mathcal{L}_p|$ the number of operations of product $p$, then the pattern can be designed as the timed route in Figure 5.3. The pattern is used for every period $t$ with $t \leq T - \sum_{l=2}^{|\mathcal{L}_p|} LT_l$. The complexity of an algorithm creating all these timed routes is $O(P|\overline{\mathcal{L}}|T)$, where $|\overline{\mathcal{L}}|$ is the average number of operations in a route.

## 5.4 A column generation approach for flexible lead times

Because, as shown in this section, the number of timed routes with flexible lead times is exponential, we propose a column generation approach to solve the timed route formulation.

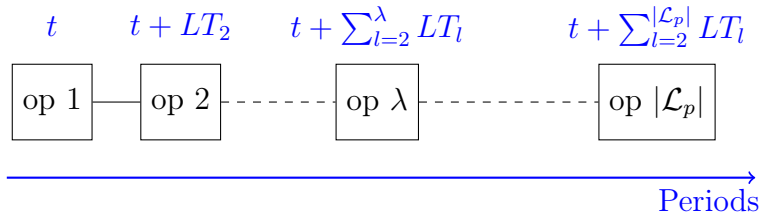



Figure 5.3: Pattern of timed routes with Fixed Lead Times

In Section 5.4.1, a dynamic programming algorithm that generates all the timed routes when considering flexible lead times is described. The column generation approach is introduced in Section 5.4.2, where reduced costs associated with timed routes are evaluated and used to implement a dominance rule to strengthen the algorithm of Section 5.4.1.

### 5.4.1 Dynamic program to generate timed routes for flexible lead times

Using timed routes, all production flows can be described and traced. Thus, we can consider other production flows than the ones generated using fixed lead times. Considering several timed routes with different lead times for one operation leads to more flexibility. This is the case with the flexible lead times presented in Chapter 4. Furthermore, when using the timed route formulation with flexible lead times, it is possible to avoid products with too large cycle times.

To establish a timed route, each operation needs to be assigned to a period in the horizon. Representing this assignment by a graph, nodes are labeled ($s$,$c$,$t$,$l$) where $s$ is the index of the current partial route, $c$ the current partial cost, $t$ the period and $l$ the last operation of the partial timed route $s$. The directed edges are the possible sequences of nodes. Due to the structure of a route, the graph can be seen as a graph with levels. Figure 5.4 provides an example of such graph, with 2 operations and 3 periods. Using this kind of graphs, an algorithm generating dynamically the edges and new vertices level by level will work well.

Rather than exploring the total space of possible states, the number of vertices is reduced by using $o_{max}(l)$, the maximum number of operations processable after operation $l$ in the same period than $l$. The vertices and edges which can be used when $o_{max}(l) = 1$ for every operation are traced with plain arrows and in blue in Figure 5.4. Even with this reduction, the total number of timed routes for product $p$ is still in $O(|\mathcal{L}_p|^T)$ because, at each operation of the route, a period between 1 and T can be assigned.

A dynamic program can be implemented as described in Algorithms 5.1 and 5.2. The main algorithm (Algorithm 5.1) generates all timed routes. It starts with a set of partial timed routes only containing the partial timed route with no period assigned, labeled (0,0,0). For each period, the algorithm tries to extend the set of partial timed routes by looking for the children nodes of each partial timed route. This procedure is developed in Algorithm 5.2.

In Algorithm 5.2, the partial time routes are returned, which extend the input partial timed route in period $t$. Extending a partial timed route means looking for each outgoing edge from the last node in the graph depicted earlier. The number of partial timed routes generated is $o_{max}(l)$ where $l$ is the last operation assigned in the input partial timed route. The information on the last operation is updated in the new partial timed routes.

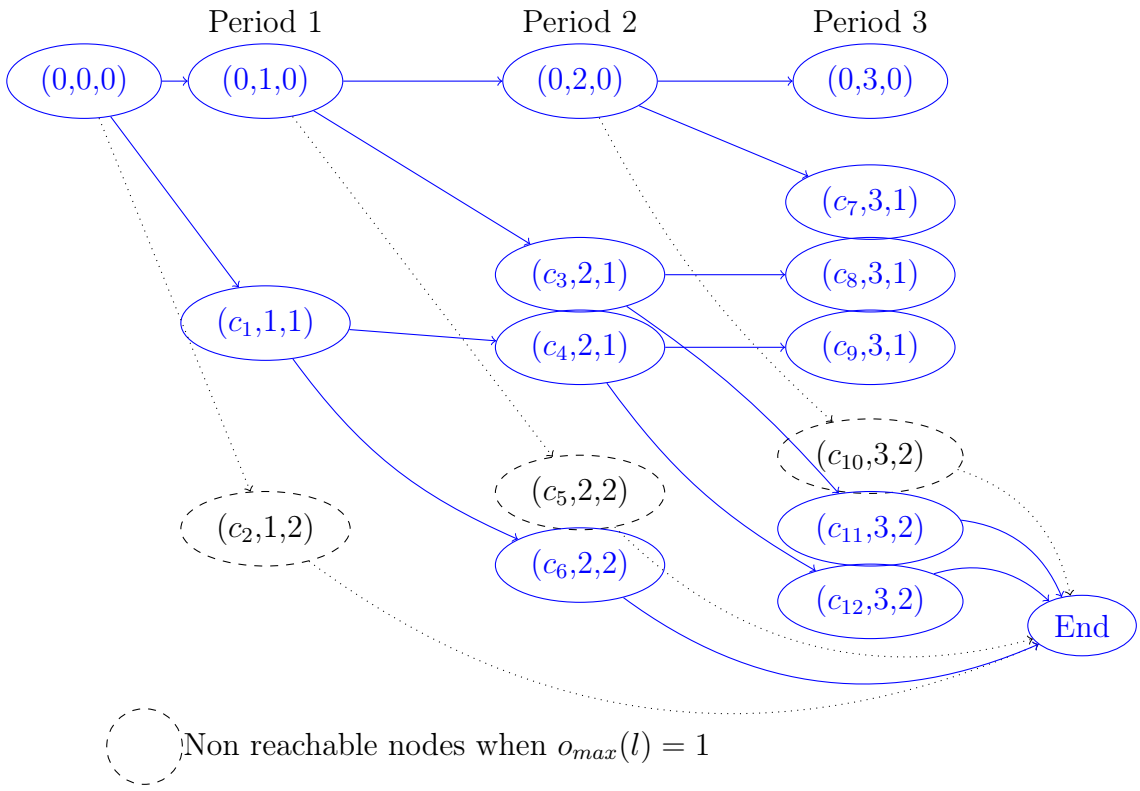



Figure 5.4: Graph of states: Example with 2 operations and 3 periods

The dynamic program explores all possibilities, which leads to an exponential number of routes. Assuming that the complexity of generating a new partial timed route is in $O(1)$, the total complexity of the algorithm is $O(|\mathcal{L}_p|^T)$ for product p.

**Algorithm 5.1** Generation of timed routes

$CTR = \emptyset$ // $CTR$: Set of complete timed routes
$ir$ // $ir$: Initial partial timed route
lastop($ir$)= 0 // No operation allocated in $ir$
$PTR = \{ir\}$ // $PTR$: Set of current partial timed routes
**for** $t = 1$ to T **do**
  **for all** $s \in PTR$ **do**
    CreateExtensions($s$,$t$)
  **end for**
**end for**
**return** $CTR$

**Algorithm 5.2** CreateExtensions($s$,$t$)

$l = \text{lastop}(s) + 1$
**for** $e = 0$ to $o_{max}(l)$ **do**
  $sr = s$
  **for** $i = 0$ to $e$ **do**
    $\text{op}(sr, l + i) = t$
  **end for**
  $\text{lastop}(sr) = l + e$
  **if** $l + e = |\mathcal{L}_p|$ **then**
    $CTR = CTR \cup \{sr\}$
  **else**
    $PTR = PTR \cup \{sr\}$
  **end if**
**end for**

### 5.4.2 Column generation approach

The set of timed routes for flexible lead times is exponential, as shown by the complexity of the dynamic program. To handle this issue, we propose a column generation approach, in which timed routes are generated dynamically. The framework of the approach can be found in Figure 5.5.

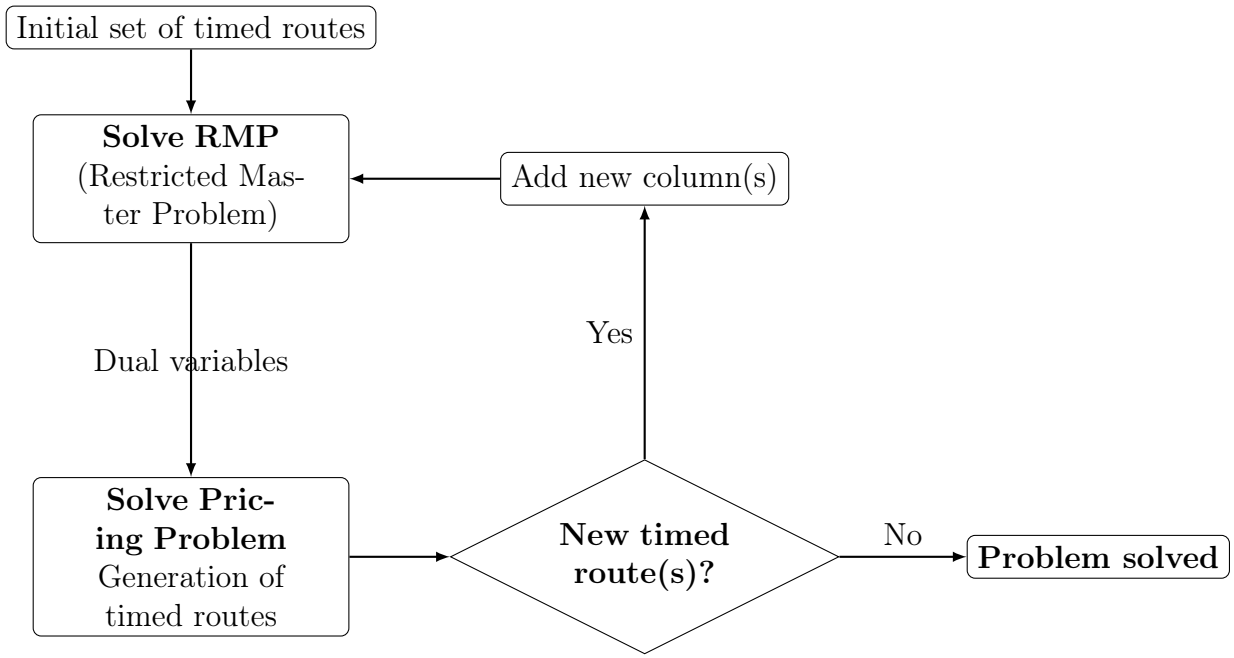



Figure 5.5: Framework of column generation approach for production planning

The master problem corresponds to the model in Section 5.3.2. Thus, the Restricted Master Problem (RMP) is written with a restricted set of timed routes for each product. The restricted set of timed routes is initialized with the timed routes generated with fixed lead times. A fast resolution of the pricing problem, that generates new improving timed routes, is critical to the success of the column generation approach. An efficient algorithm is proposed in the following section.

**Solving the pricing problem**

To determine the timed routes to insert in the RMP, we consider the reduced costs associated with timed routes. The dual problem associated with the timed route formulation corresponds to (5.6)-(5.9), where $\lambda^*_{kt}$ denotes the dual variables associated with Constraints (5.2), and $\beta^+_{ps}$ (resp. $\beta^-_{ps}$) denotes the dual variables associated with Constraints (5.3) (resp. Constraints (5.4)).

$$\max \quad -\sum_{t=1}^{T}\sum_{k=1}^{K} C_k\lambda^*_{kt} - \sum_{p=1}^{P}\sum_{s=1}^{S}(\sum_{\sigma=1}^{s} D_{p\sigma})\beta^+_{ps} + \sum_{p=1}^{P}\sum_{s=1}^{S}(\sum_{\sigma=1}^{s} D_{p\sigma})\beta^-_{ps} \tag{5.6}$$

$$\text{s.t.} \quad -\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}^k} a^{pr}_{lt}\alpha_{pl}\lambda^*_{kt} - \sum_{s=1}^{S}\sum_{\tau=1}^{tf_s} a^{pr}_{|\mathcal{L}_p|\tau}\beta^+_{ps} + \sum_{s=1}^{S}\sum_{\tau=1}^{tf_s} a^{pr}_{|\mathcal{L}_p|\tau}\beta^-_{ps} \leq w_{pr}$$

$$\forall p \in \{1,\dots,P\} \quad \forall r \in \mathcal{R}_p \tag{5.7}$$

$$\beta^+_{ps} \leq h_{ps} \qquad \forall p \in \{1,\dots,P\} \quad \forall s \in \{1,\dots,S\} \tag{5.8}$$

$$\beta^-_{ps} \leq b_{ps} \qquad \forall p \in \{1,\dots,P\} \quad \forall s \in \{1,\dots,S\} \tag{5.9}$$

In the dual problem, only Constraints (5.7) are related to timed routes. Thus, in the column generation approach, we only need to look for timed routes which violate the most Constraints (5.7), i.e. timed routes with reduced cost $w_{pr} + \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{l\in\mathcal{L}^k} a^{pr}_{lt}\alpha_{pl}\lambda^*_{kt} + \sum_{s=1}^{S}\sum_{\tau=1}^{tf_s} a^{pr}_{|\mathcal{L}_p|\tau}\beta^+_{ps} - \sum_{s=1}^{S}\sum_{\tau=1}^{tf_s} a^{pr}_{|\mathcal{L}_p|\tau}\beta^-_{ps} \leq 0$. Note that, since there is no constraint linking the products in the pricing problem, timed routes can be generated separately for each product.

In order to define a route, we need to assign each operation $l$ to a period $t$, i.e. to determine $a^{pr}_{lt}$. The reduced cost can be decomposed into three parts.

1. A period assignment cost which is denoted $\alpha_{pl}\lambda^*_{kt}$,
2. The WIP cost of the route, which can be decomposed into the WIP cost at each period,
3. Inventory and backlog costs. If the period of the last operation (i.e. when the product is completed) is $s^*$, then the inventory and backlog costs are equal to $\sum_{s=s^*}^{S}(\beta^+_{ps} - \beta^-_{ps})$.

**Dominance rule**

With such a complexity, the dynamic program can hardly be used in practice. In order to keep the computational times under control, we consider a dominance rule that relies on Property 2.

**Property 2.** *For product p at a period t, if two partial timed routes $s_1$ and $s_2$ have achieved the same number of operations l, then the route with the lowest partial reduced cost dominates the other. In other words, for $s_1$=(1,$rc_1$,t,l) and $s_2$ =(2,$rc_2$,t,l), then $s_1$ dominates $s_2$ if and only if $rc_1 \leq rc_2$.*

*Proof.* It can be shown by contradiction that, if the periods or the last operations are different, then an arbitrary large negative reduced cost can be introduced in the complete and

dominated timed route. Thus, we can introduce $s_3$, the optimal part to complete $s_1$ and $s_2$ to form a complete timed route. We denote $rc_3$ the reduced cost associated with $s_3$ and $s_1 \oplus s_3$ (respectively $s_2 \oplus s_3$) the complete timed route associated with $s_1$ (resp. $s_2$) and its total reduced cost $rc_{1\oplus 3}$ (resp. $rc_{2\oplus 3}$). Because $rc_{1\oplus 3} = rc_1 + rc_3$ and $rc_{2\oplus 3} = rc_2 + rc_3$, comparing the total reduced costs $rc_{1\oplus 3}$ and $rc_{2\oplus 3}$ is equivalent to comparing the partial reduced costs $rc_1$ and $rc_2$. □

Note that, if constraints on the duration of cycle times are introduced, some conditions on the start period of partial timed routes are needed to apply this dominance rule.

By applying this dominance rule in the dynamic program, the number of new partial timed routes at the end of each iteration/period is at most equal to the number of operations for a product. Thus, at iteration $t$ of the algorithm for a given product $p$, the number of partial timed routes before dominance is smaller than $|\mathcal{L}_p|^2$. It reduces the complexity of Algorithms 5.1 and 5.2 to $O(|\mathcal{L}_p|^2 T)$ for each product. To implement the dominance rule, we use an array that contains the dominant partial timed routes (at the currently explored period) for each operation of the route (except for the final step). The size of this array, denoted $ND[]$, is $|\mathcal{L}_p|$. Thus, the overall complexity is in $O(T \sum_{p=1}^{P} |\mathcal{L}_p|^2)$.

**Algorithm 5.3** CreateNonDominatedExtension($s$,$t$,$ND[]$)

```
// ND[]: Array (of size |L_p| for product p) of dominant partial timed routes up to period
t-1 indexed by the last operation reached.
l = lastop(s) + 1
for e = 0 to o_max(l) do
  sr = s   // Extend timed route s by e operations to perform at period t
  for i = 0 to e do
    op(sr, l + i) = t
    UpdateReducedCost(sr)
  end for
  lastop(sr)= l + e
  if l + e = |L_p| then
    CTR = CTR ∪ {sr}
  else
    // Dominance check
    if ReduceCost(sr) > ReducedCost(ND[l + e]) then
      // sr dominates the former dominant partial timed route, which ends at period t
      with operation l + e
      PTR = PTR ∪ {sr}
      PTR = PTR \ {ND[l + e]}
      ND[l + e] = sr
    end if
  end if
end for
```

## 5.5 Computational experiments

Computational experiments have been conducted on industrial data to show the efficiency of the timed route formulation and of our column generation approach. In Section 5.5.1, the design of the computational experiments is detailed. In Section 5.5.2, the compact formulation (3.1)-(3.8) and the timed route reformulation (5.1)-(5.5) are compared for fixed lead times. Section 5.5.3 compares the column generation approach with flexible lead times and the compact formulation proposed in Chapter 4). Finally, Section 5.5.4 explores the new ways of analyzing the production flows that the timed route formulation allows.

### 5.5.1 Design of experiments

Experiments are conducted on the first industrial data set of the 200mm semiconductor manufacturing facility. The main characteristics of the instances are recalled in Table 5.2. Crossing all choices of the characteristics, 27 scenarios are considered.

We only consider the most produced products. To study the influence of the number of products, we consider 3 sets of products. Each demand scenario, related to the number of products, is then adjusted by a factor on the generated demand to produce 3 scenarios where respectively demand is low and feasible, demand is medium but stresses the facility capacity and demand is high and cannot be fully met.

| | |
|---|---|
| Horizon length | {91, 119, 147} |
| Number of workshops | 10 (aggregating about 500 machines) |
| Number of products | {15, 40, 75} |
| Demand scenario | {Low, Medium, High} |

Table 5.2: Characteristics of the industrial instances

Furthermore, three profiles of lead times presented in Chapter 4, are studied by solving the compact models and using the column generation approach.

1. The first profile, $\mathcal{P}_{LT}^{\text{fixed}}$, corresponds to the classical fixed lead times.

2. The second profile, $\mathcal{P}_{LT}^{\text{flex}}$, corresponds to flexible lead times and is based on $\mathcal{P}_{LT}^{\text{fixed}}$, but products can wait in every operation as many periods as necessary. This lead time profile reduces the backlog and inventory costs by allowing more flexible production flows.

3. The third profile, $\mathcal{P}_{PT}^{\text{flex}}$, also corresponds to flexible lead times, but is based on the actual processing times, i.e. it is not related to the two other lead time profiles. With profile $\mathcal{P}_{PT}^{\text{flex}}$, production flows are only limited by the maximum number of operations for a product that can be completed in a period, according to the cumulative process times of these operations. In a sense, it is a relaxation of the previous model where delays are not induced by exogenous parameters. Note that, contrary to $\mathcal{P}_{LT}^{\text{flex}}$ where Constraint (4.1) is not written, when $LT(l) = 0$ for an operation $l$, with $\mathcal{P}_{PT}^{\text{flex}}$, Constraint (4.1) is written for every operation.

As shown in the computational results of section 5.5.3, $\mathcal{P}_{PT}^{\text{flex}}$ leads to the most difficult problems in terms of computational time. For example, with the compact formulation, on

scenarios with medium or large dimensions, there is at least a factor of ten between the computational times for $\mathcal{P}_{LT}^{\text{flex}}$ and $\mathcal{P}_{PT}^{\text{flex}}$.

All numerical experiments were executed on a computer with a processor Intel(R) Xeon(R) CPU W3550 and 16 Go of RAM Memory, using a JAVA program (JRE 1.8) and IBM ILOG CPLEX (version 12.6).

### 5.5.2 Comparison between the compact formulation and the timed route reformulation with fixed lead times

Due to the polynomial number of timed routes with fixed lead times, all timed routes are generated and included in the model. Table 5.3 shows the computational times spent by IBM ILOG CPLEX for several scenarios. First, note that the computational times do not seem to change much with the demand level. Thus, only looking at the medium scenarios, it can be seen that the timed route model performs significantly better than the compact one. On average, the computational time is decreased by 94%, with a minimum decrease of 88%. When considering the impact of the horizon length, the results show that the timed route formulation is more sensitive to the horizon length than the compact model. The gap between the computational times of both models reduces as the horizon length increases. For all these scenarios, the computational times of the timed route formulation are always smaller than the smallest computational time with the compact formulation. For fixed lead times, the timed route formulation is efficient when all the timed routes are generated.

| **Number of products** | **Horizon length** | **Low demand** | | **Medium demand** | | **High Demand** | |
|---|---|---|---|---|---|---|---|
| | | **C** | **TR** | **C** | **TR** | **C** | **TR** |
| **Low (15)** | **91** | 15 | 0 | 15 | 0 | 14 | 0 |
| | **119** | 20 | 1 | 20 | 1 | 19 | 1 |
| | **147** | 25 | 5 | 25 | 3 | 25 | 4 |
| **Medium (40)** | **91** | 40 | 1 | 41 | 1 | 40 | 1 |
| | **119** | 58 | 3 | 57 | 4 | 58 | 4 |
| | **147** | 68 | 7 | 70 | 7 | 71 | 7 |
| **Large (75)** | **91** | 84 | 1 | 83 | 1 | 83 | 2 |
| | **119** | 113 | 4 | 114 | 8 | 114 | 5 |
| | **147** | 145 | 16 | 147 | 15 | 145 | 16 |

Table 5.3: Computational times (in seconds) for profile $\mathcal{P}_{LT}^{\text{fixed}}$ (C: Compact formulation; TR: Timed Route formulation)

### 5.5.3 Column generation approach for flexible lead times

In this section, the compact formulation and the timed route formulation with flexible lead time profiles are compared. The first flexible lead time profile studied is $\mathcal{P}_{LT}^{\text{flex}}$. The associated compact model has a lower number of lead time constraints compared to the compact model with fixed lead times. This is due to the fact that lead time constraints are only introduced for positive lead times. The second flexible lead time profile is $\mathcal{P}_{PT}^{\text{flex}}$. Its compact formulation has about the same number of constraints as the compact formulation with fixed lead times, but production flows are less constrained. The associated flexible lead time constraints are based on the actual processing times of operations.

As shown in Section 5.4.1, the timed route formulation with flexible lead times requires an exponential number of timed routes. To get a feeling of the resulting complexity, we generate all the timed routes for the Kayton's instance. With 8 Gb of RAM and when the horizon is larger than 15 periods, it is not possible to generate all timed routes for profile $\mathcal{P}_{LT}^{\text{flex}}$ and a memory error arises.

In Section 5.5.3, the parameters and strategies used in the column generation approach are detailed. The experimental results for profile $\mathcal{P}_{LT}^{\text{flex}}$ are presented in Section 5.5.3, while the results for profile $\mathcal{P}_{PT}^{\text{flex}}$ are analyzed in Section 5.5.3.

### Column generation strategy

Dominance rules are used to reduce computational times. To warm up the column generation approach, all timed routes from $\mathcal{P}_{LT}^{\text{fixed}}$ are included in the model. Due to the light use of the processor during the timed route generation, parallelism is enabled while generating timed routes for each product. Note that, when disabling the parallelism, reduction of computation time is weaker but still is important: the average reduction is of 70% for profile $\mathcal{P}_{LT}^{\text{flex}}$ and 84% for $\mathcal{P}_{PT}^{\text{flex}}$.

The last parameter to choose is how many timed routes are selected for each product at each iteration. This parameter is tuned with the case of Medium demand, with profile $\mathcal{P}_{LT}^{\text{flex}}$. Figure 5.6 shows the average decrease of the computational time compared to the case in which only one timed route is generated by product. This figure is completed with the maximum decrease and the minimum decrease obtained among the 27 scenarios. Note that the average time spent to solve the timed route formulation limited to one new timed route by product at each iteration is 239 seconds. It can be seen that, when the parameter varies between 4 and 10, the decrease of the computational time is quite stable and the lowest. With up to 150 timed routes by product (which is an upper bound to the number of non-dominated timed routes generated by the dynamic program when $T < 150$), it can be seen that the decrease of the computational time is similar to when the parameter is set to 2. This figure shows the trade-off between generating numerous columns to converge with fewer iterations and generating only the best columns to accelerate the resolution of the restricted master problem.

In the following experiments, the number of timed routes by product at each iteration is set to 5. This choice might not be the best in every scenario, but is relevant enough to show the strength of our approach.

### Comparison of computational times for profile $\mathcal{P}_{LT}^{\text{flex}}$

Contrary to fixed lead times, the computational times for flexible lead time profiles depend on the demand scenario. Table 5.4 shows the computational times to solve $\mathcal{P}_{LT}^{\text{flex}}$. No simple rule can be deduced (for both formulations) from the different scenarios because the complexity of the problem depends on several parameters. Computational times to solve the timed route model are quite close with medium and high demands, and are always larger than the computational times with low demands.

The main result of the experiment is that the column generation approach always significantly performs better. On average, the computational time is reduced by 87.5% while the solution time for the compact model ranges from 2 minutes to 79 minutes. The least impressive case is 73.3% when the time spent by the compact formulation is the lowest (120 seconds). Unlike fixed lead times, we cannot conclude anything about the behavior of the

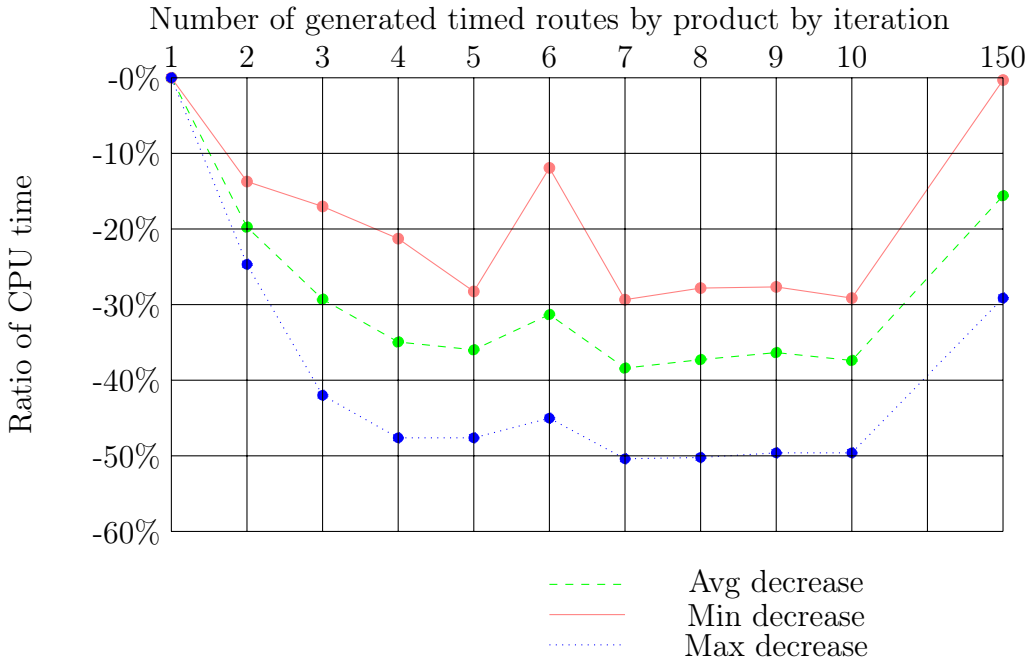



Figure 5.6: Number of timed routes by product at each iteration vs. ratio of CPU time

| Number of products | Horizon length | Low demand | | Medium demand | | High Demand | |
|---|---|---|---|---|---|---|---|
| | | C | TR | C | TR | C | TR |
| **Low (15)** | **91** | 192 | 14 | 142 | 31 | 120 | 32 |
| | **119** | 409 | 59 | 266 | 67 | 284 | 66 |
| | **147** | 580 | 95 | 595 | 121 | 580 | 124 |
| **Medium (40)** | **91** | 648 | 34 | 469 | 66 | 563 | 63 |
| | **119** | 1,190 | 109 | 1,086 | 134 | 1,254 | 153 |
| | **147** | 1,674 | 180 | 2,034 | 218 | 2,174 | 218 |
| **Large (75)** | **91** | 1,620 | 23 | 1,578 | 107 | 1,363 | 109 |
| | **119** | 3,693 | 74 | 3,000 | 236 | 2,145 | 242 |
| | **147** | 4,277 | 316 | 3,797 | 401 | 4,619 | 395 |

Table 5.4: Computational times (in seconds) for profile $\mathcal{P}_{LT}^{\text{flex}}$ (C: Compact formulation; TR: Timed Route formulation)

compact model when the horizon increases, only that the computational times increase with the length of the horizon (which is expected due to the algorithm complexity).

### Comparison of computational times for profile $\mathcal{P}_{PT}^{\text{flex}}$

Considering profile $\mathcal{P}_{PT}^{\text{flex}}$ whose computational results can be found in Table 5.5, some conclusions are shared with $\mathcal{P}_{LT}^{\text{flex}}$. For example, the computational times vary depending on the demand scenario, but in the case of $\mathcal{P}_{PT}^{\text{flex}}$, it can also be noted that the larger the demand, the larger the CPU time to solve the problem, and the increase depends on the scenario. The computational times are again highly reduced by the column generation approach on the timed route formulation. On average, they are reduced by 95.8%. The computational time for the compact model ranges from 3 minutes to more than 6 days (with a median of 2.5 hours).

| Number of products | Horizon length | Low demand | | Medium demand | | High Demand | |
|---|---|---|---|---|---|---|---|
| | | C | TR | C | TR | C | TR |
| **Low (15)** | **91** | 183 | 33 | 1,549 | 37 | 1,760 | 51 |
| | **119** | 272 | 48 | 2,677 | 83 | 2,836 | 120 |
| | **147** | 4,100 | 161 | 4,211 | 255 | 4,962 | 407 |
| **Medium (40)** | **91** | 5,587 | 84 | 6,254 | 156 | 6,486 | 233 |
| | **119** | 8,939 | 167 | 10,092 | 298 | 10,277 | 429 |
| | **147** | 13,014 | 291 | 14,407 | 587 | 16,152 | 793 |
| **Large (75)** | **91** | 979 | 95 | 10,516 | 179 | 11,902 | 246 |
| | **119** | 18,862 | 193 | 18,498 | 346 | 20,472 | 460 |
| | **147** | 599,443 | 404 | 323,891 | 678 | 546,596 | 1,273 |

Table 5.5: Computational times (in seconds) for profile $\mathcal{P}_{PT}^{\text{flex}}$ (C: Compact formulation; TR: Timed Route formulation)

With the compact formulation, there is a huge gap in the computational times for the three lead time profiles. Due to the extreme computational time in the scenarios with a large number of products and a long horizon, the average computational time is a biased indicator. Therefore, we prefer to analyze the median computational time. Over all the scenarios, the median computational times are 58 seconds for $\mathcal{P}_{LT}^{\text{fixed}}$, 1,190 seconds for $\mathcal{P}_{LT}^{\text{flex}}$ and 8,939 seconds for $\mathcal{P}_{PT}^{\text{flex}}$. When using the timed route formulation and the column generation approach, the computational times also increase as the lead time profile becomes more complex, but the increase is much more limited. The overall medians of the computational times for the compact formulation are equal to 3 seconds for $\mathcal{P}_{LT}^{\text{fixed}}$, 109 seconds for $\mathcal{P}_{LT}^{\text{flex}}$ and 233 seconds for $\mathcal{P}_{PT}^{\text{flex}}$. One reason which can explain why computational times for the timed route formulation with $\mathcal{P}_{PT}^{\text{flex}}$ is close to $\mathcal{P}_{LT}^{\text{flex}}$, might be the difference of these two lead time profiles. It can be seen in Table 5.6 that, in most scenarios (except when the demand is high and the horizon is long, in red, in Table 5.6), $\mathcal{P}_{PT}^{\text{flex}}$ needs fewer iterations of the column generation approach to converge to the optimal solution.

The reason is probably that, while the compact formulation struggles with a huge number of constraints, many useful timed routes are quickly generated in the column generation approach, thus fewer iterations are needed before converging. It could be interesting to tune the maximum number of timed routes by product at each iteration.

| Number of products | Horizon length | Low demand | | Medium demand | | High Demand | |
|---|---|---|---|---|---|---|---|
| | | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ | $\mathcal{P}_{LT}^{\text{flex}}$ | $\mathcal{P}_{PT}^{\text{flex}}$ |
| **Low (15)** | **91** | 25 | 11 | 46 | 12 | 47 | 17 |
| | **119** | 57 | 12 | 61 | 21 | 60 | 31 |
| | **147** | 60 | 31 | 72 | 50 | 71 | 80 |
| **Medium (40)** | **91** | 22 | 12 | 37 | 23 | 36 | 35 |
| | **119** | 38 | 16 | 43 | 30 | 48 | 46 |
| | **147** | 47 | 23 | 46 | 45 | 47 | 66 |
| **Large (75)** | **91** | 8 | 7 | 32 | 14 | 35 | 19 |
| | **119** | 16 | 11 | 40 | 18 | 43 | 25 |
| | **147** | 42 | 17 | 46 | 25 | 46 | 54 |

Table 5.6: Number of iterations in the column generation approach with flexible lead time profiles

### 5.5.4 Detailed analysis of cycle times using timed routes

If a major advantage of the timed route formulation is the decrease of computational times, another significant advantage is that production flows are explicitly characterized by the timed routes that are used and the quantity allocated to each timed route. In this section, we show how cycle times can be analyzed by looking at the selected timed routes of products. We chose the scenario with flexible lead times ($\mathcal{P}_{LT}^{\text{flex}}$), a medium horizon length (119 micro-period), a medium number of products (40) and medium demand. Table 5.7 provides, for each product, its theoretical minimum cycle time induced by the lead time constraints, its minimum, maximum and mean cycle times, and finally the maximum lead time at an operation of the timed routes of the product. The mean cycle time is computed as the mean of the cycle times of the timed routes of the product, weighted by the quantity allocated to each timed route. Only timed routes with a strictly positive product quantity are considered when computing the minimum and maximum cycle times and the maximum lead time. All indicators are expressed in number of micro-periods.

Note that the mean cycle time is always strictly larger than the theoretical minimum cycle time, see for example product 3 with a minimum cycle time of 40 micro-periods and a mean cycle time of 52.4 micro-periods. The difference between the theoretical minimum cycle time and the mean cycle time is always larger than 10 micro-periods in this instance. However, looking at the minimum and maximum cycle times of timed routes that are used (i.e. with positive production quantity), it can be seen that, for most products, the cycle time ranges between the theoretical minimum cycle time and the number of micro-periods in the planning horizon. Many products have a maximum cycle time that is larger than 110 micro-periods, see for example products 4, 15, 24 and 33. However, there are exceptions such as product 19 with a maximum cycle time of 75 micro-periods or product 35 with a maximum cycle time of 69 and a minimum cycle time of 68.

Note that, if a wide range between the minimum and maximum cycle times shows that flexibility is exploited, too large cycle times are not welcome or even acceptable in practice. In this instance, the large cycle times are mainly caused by one large lead time that occurs a few operations after the start of the timed route. For example, products 1, 11 or 21 have a maximum lead time which is larger than the minimum cycle time, so the timed route in which these lead times occur is the one with the largest cycle times. We believe that the

| Product | Cycle Time | | | | Max Lead Time |
|---|---|---|---|---|---|
| | Theor. min | Min | Max | Mean | |
| **1** | 49 | 51 | 103 | 66.2 | 52 |
| **2** | 37 | 42 | 96 | 55.6 | 36 |
| **3** | 40 | 41 | 103 | 52.4 | 60 |
| **4** | 59 | 61 | 110 | 75.5 | 23 |
| **5** | 47 | 49 | 103 | 72.3 | 52 |
| **6** | 47 | 47 | 110 | 74.1 | 59 |
| **7** | 77 | 77 | 97 | 81.9 | 14 |
| **8** | 55 | 55 | 117 | 67.7 | 60 |
| **9** | 68 | 96 | 103 | 96.4 | 12 |
| **10** | 65 | 68 | 110 | 80.8 | 22 |
| **11** | 48 | 49 | 110 | 66.7 | 59 |
| **12** | 58 | 62 | 97 | 73.9 | 15 |
| **13** | 41 | 42 | 103 | 63.7 | 60 |
| **14** | 69 | 69 | 96 | 72.1 | 10 |
| **15** | 45 | 47 | 110 | 70.3 | 61 |
| **16** | 47 | 49 | 103 | 67.9 | 52 |
| **17** | 69 | 69 | 103 | 81.6 | 13 |
| **18** | 63 | 95 | 96 | 96.0 | 13 |
| **19** | 47 | 49 | 75 | 56.6 | 18 |
| **20** | 61 | 63 | 110 | 72.4 | 24 |
| **21** | 48 | 49 | 110 | 61.8 | 59 |
| **22** | 45 | 48 | 111 | 76.2 | 61 |
| **23** | 68 | 69 | 110 | 81.3 | 16 |
| **24** | 59 | 97 | 118 | 99.8 | 57 |
| **25** | 40 | 40 | 96 | 55.8 | 52 |
| **26** | 65 | 67 | 110 | 79.6 | 23 |
| **27** | 61 | 62 | 117 | 86.6 | 34 |
| **28** | 75 | 76 | 105 | 81.6 | 13 |
| **29** | 55 | 55 | 110 | 63.2 | 52 |
| **30** | 55 | 55 | 117 | 82.1 | 47 |
| **31** | 45 | 47 | 104 | 63.1 | 58 |
| **32** | 47 | 48 | 103 | 60.4 | 52 |
| **33** | 45 | 47 | 111 | 64.2 | 61 |
| **34** | 63 | 89 | 110 | 97.3 | 45 |
| **35** | 67 | 68 | 69 | 68.9 | 3 |
| **36** | 79 | 84 | 84 | 84.0 | 4 |
| **37** | 59 | 62 | 118 | 79.1 | 57 |
| **38** | 57 | 63 | 117 | 82.8 | 50 |
| **39** | 68 | 89 | 103 | 100.3 | 13 |
| **40** | 45 | 49 | 103 | 63.8 | 56 |

Table 5.7: Analysis of the cycle times of products in the scenario with profile $\mathcal{P}_{LT}^{\text{flex}}$, medium horizon, medium number of products and medium demand

large lead times at the beginning of timed routes are due to the instance, where the lack of initial WIP leads to more capacity being available to process the first operations of a product in the first periods. This capacity is no longer available in later periods, once the manufacturing facility is full. Nevertheless, constraints to avoid such large cycle times or lead times can be implemented with the timed route formulation, whereas these constraints will be very difficult to consider in the compact formulation with flexible lead times.

Some information is not shown in Table 5.7. In particular, each product uses several timed routes (at least two) and the quantity allocated to each timed route may vary significantly. A production quantity that is too small does not make sense, and our future research aims to consider a minimum quantity per timed route in the timed route formulation and, more importantly, in the column generation approach. Note that, again, a minimum production quantity is much easier to model with the timed route formulation than with the compact formulation.

## 5.6 Conclusions and perspectives

In this chapter, we introduced the novel concept of timed route that enables a new model for multi-product multi-step production planning problems to be introduced. The timed route approach was validated on industrial data, and experimental results show that the new formulation significantly outperforms compact formulations for various lead time profiles. To achieve such performance and because considering flexible lead times induces an exponential number of columns, a column generation approach was presented with a polynomial dynamic program that generates the timed routes in the pricing problem.

Many research opportunities are offered using timed routes and timed route formulations. An interesting point to investigate is the various industrial rules that could only be developed for mathematical models based on timed routes. As already discussed and by definition, timed routes allow production flows and their cycle times to be explicitly modeled. On the opposite, flexible lead time constraints in a compact mathematical model do not easily allow cycle times to be limited and production flows to be explicitly managed. Hence, many relevant industrial constraints can be taken into account through timed routes. For example, timed routes could be generated by considering minimum or maximum cycle times of products, or minimum or maximum lead times between two non-consecutive production operations. Also, a cycle time for each product could be targeted in the objective function, by introducing new costs on timed routes instead of the somehow artificial WIP management costs. These costs could be associated with the deviation to the target cycle time. In addition, costs based on the duration of the lead time in a production operation could be proposed, that would be non-linear in compact models but linear in timed route models. Some of these aspects are explored in Chapter 6.

Moreover, the computational times of the column generation approach could be accelerated by using smart column generation heuristics. Another research perspective is to consider initial inventories in the product routes. Shorter timed routes will be required to flush the initial inventories. Finally, we would like to study whether timed routes could be used in other contexts, e.g. when modeling product flows in supply chains where the notion of "route" is also relevant.

# Chapter 6

# Extensions of the timed route approaches

## 6.1 Introduction

As discussed in the conclusion of Chapter 5, the concept of timed routes opens multiples perspectives. First, it is possible to enrich the problem by only changing the construction rules of the timed routes. This will make the pricing problem more complex without affecting the structure of the master problem. In addition, the timed route formulation offers the possibility of pricing the production flows knowing target lead times and target cycle times. Due to the separation between the master problem and the pricing problem that generates the timed routes, it is also possible to consider non-linear cost structures. Furthermore, the analysis of the timed routes generated by the flexible lead times in Chapter 5 shows some issues regarding the length of cycle times and even some abnormally long lead times. Thus, the goal of this chapter is to control the production flows using the possibilities allowed by the timed routes.

In Section 6.2, we first study the simultaneous integration of minimum and maximum cycle time constraints. Then, we study the integration of both constraints separately. In Section 6.3, we study the integration of target cycle times and maximum lead times. These new constraints only impact the pricing problem by adapting the timed route generation algorithm. Note that this chapter does not study all possibilities offered by the timed route formulation. Other potential extensions are discussed in Section 6.4.

## 6.2 Controlling the cycle time

One of the flaws of flexible lead times is that cycle times can be as long as the time horizon length (as shown in the numerical experiments of Chapter 5). As already mentioned, the timed route formulation allows integrating minimum and maximum cycle times. In fact, to handle these two constraints it is sufficient to only generate routes that satisfy these constraints. One may claim that flexible lead times induce a minimum cycle time, but this minimum cycle time can be very short and may affect the production processes. Indeed, it might be preferable to avoid too fast production time routes, and possibly unrealistic, by introducing minimum cycle times. This is mainly useful when trying to avoid that a product is prioritized without considering the overall production process. In semiconductor manufacturing, a useful indicator is called "X-factor". The X-factor is the ratio of the cycle time to the total cumulative processing time of a route ( $\frac{\sum_{l \in \mathcal{L}_p} LT_{pl}}{\sum_{l \in \mathcal{L}_p} pt_{pl}}$, where $pt_{pl}$ is the process

time for operation $l$). Except for some products without demand, most of the production lots have their specific X-factor. Note that the X-factor makes sense only at the tactical level because, at the operational level, the priority and speed of lots can vary along the time horizon.

It is important to note that controlling cycle times using the compact formulation is a hard task. First, because we have shown the limits of this formulation in previous sections. Second, because controlling the cycle time for each product requires to follow the production flows that leads to the introduction of other time indices in variables that make the formulation even harder to handle. Additional constraints on the final inventory and the initial input are also required. However, with the timed route formulation, we "only" need to adapt the pricing problem, to better control the complexity and the computational time.

In this section, Algorithms 5.1 and 5.2 are adapted in order to handle minimum and maximum cycle times. New dominance rules are proposed. The consideration of cycle time bounds leads to a higher complexity, but both algorithms remain polynomial.

### 6.2.1 Considering minimum and maximum cycle times

In this section, we adapt the dynamic programming algorithms of Chapter 5 in order to consider both minimum and maximum cycle times. We adapt Algorithms 5.1 and 5.3 to generate only timed routes satisfying minimum and maximum cycle time constraints. In order to adapt these algorithms, it is important to calculate the current cycle time at each period. This calculation is possible since the start period of each partial time route is known, which allows computing the cycle time of each partial timed route.

Algorithm 5.1 can be adapted by checking the feasibility of a timed route before inserting it in the set of complete timed routes. A feasibility check is also introduced during the construction of a timed route to check whether it can lead to a route that overpasses the maximum allowed cycle time. This check is performed by summing the cycle time of the current partial timed route and the minimum cycle time needed to complete the remaining operations. If the resulting cycle time is higher than the maximum cycle time, this current partial timed route can be discarded. A similar feasibility check cannot be performed for minimum cycle times. In fact, the last operation can always occur far enough in order to satisfy the needed minimum cycle time. With these feasibility checks, Algorithm 5.1 is replaced by Algorithm 6.1 without impacting the complexity of the algorithm. The main impact on complexity is when modifying the algorithm that extends partial timed routes.

**Algorithm 6.1** Generation of timed routes(CTmin, CTmax)

// remainingCT() is precalculated for every operation based on the lead time profile, else it is equal to 0
$CTR = \emptyset$ // $CTR$: Set of complete timed routes
$ir$ // $ir$: Initial partial timed route
lastop($ir$)= 0 // No operation allocated in $ir$
$PTR = \{ir\}$ // $PTR$: Set of current partial timed routes
**for** $t$ = 1 to T **do**
  **for all** $s \in PTR$ **do**
    **if** CycleTime($s$)+remainingCT(lastop($s$)) $\leq$ CTmax **then**
      CreateNonDominatedExtensions($s$,$t$,CTmin,CTmax,ND[][])
    **else**
      $PTR = PTR - \{s\}$
    **end if**
  **end for**
**end for**
**return** $CTR$

In fact, the dominance rule stated in Proposition 2 in Section 5.4.2 can no longer be used. Recall that this dominance rule states that, if two partial timed routes complete the same operation in the same period, one route dominates the other one only based on the reduced costs. This dominance rule is not longer valid when considering constraints of cycle times since one needs to know the period of the first operation. A partial timed route cannot be dominated by another timed route that starts earlier even if its reduced cost is worse. In fact, the first route can have more opportunities to improve its reduced cost.

Since, in this section, we consider both minimum and maximum cycle times, the following dominance rule (Property 3) only considers partial timed routes starting at the same period.

Let us recall the following notation from Chapter 5: $t(p, s, l)$ is the period assigned to operation $l$ in timed route $s$ of product $p$.

**Property 3.** *For a product $p$ at a period $t$, if two partial timed routes $s_1$ and $s_2$, starting at the same period, have achieved the same number of operations $l$, then the route with the lowest partial reduced cost dominates the other. In other words, for $s_1$=(1,$rc_1$,$t$,$l$) and $s_2$ =(2,$rc_2$,$t$,$l$), then $s_1$ dominates $s_2$ if and only if $t(p, s_1, 1) = t(p, s_2, 1)$ and $rc_1 \leq rc_2$.*

*Proof.* For the argument on the same operation, refer to the proof of Property 2 in Section 5.4.2. The rest of the proof is done by contradiction. Suppose that $s_1$=(1,$rc_1$,$t$,$l$) and $s_2$ =(2,$rc_2$,$t$,$l$) are two partial timed routes such that $t(p, s_1, 1) = t(p, s_2, 1)$, $rc_1 \leq rc_2$ and $s_1$ does not dominate $s_2$. This means that there are two partial timed routes $s_3$ and $s_4$ which complete respectively $s_1$ and $s_2$ to obtain the corresponding optimal routes $s_1 \oplus s_3$ and $s_2 \oplus s_4$. $s_1$ does not dominate $s_2$ means that $rc_1 + rc_3 \geq rc_2 + rc_4$ and thus $rc_3 \geq rc_4$ since $rc_1 \leq rc_2$. Since $s_1$ and $s_2$ start at the same period, $s_4$ is also a valid extension of $s_1$. Then $rc_1 + rc_4 \leq rc_2 + rc_4$. This means that $s_1$ dominates $s_2$ which contradicts the assumption that $s_1$ does not dominate $s_2$. □

Using the dominance rule of Property 3, Algorithm 5.3 is replaced by Algorithm 6.2. Algorithm 6.2 uses an array of dominant partial timed routes with a size of $T|\mathcal{L}_p|$. In our forward dynamic programming algorithm, at most $T|\mathcal{L}_p|$ non-dominated labels are stored since we also consider the start period to analyze the dominance. At each iteration of

the algorithm, at most $T|\mathcal{L}_p|^2$ partial timed routes are generated. Ultimately, the overall complexity is in $O(T^2 \sum_{p=1}^{P} |\mathcal{L}_p|^2)$.

**Algorithm 6.2** CreateNonDominatedExtension($s$, $t$, CTmin, CTmax, $ND[][]$)

```
// ND[][]: Array (of size |L_p|*T for product p) of dominant partial timed routes up to
period t-1 indexed by the last operation reached and the starting period.
l = lastop(s) + 1
for e = 0 to o_max(l) do
  sr = s    // Extend timed route s by e operations to perform at period t
  for i = 0 to e do
    op(sr, l + i) = t
    UpdateReducedCost(sr)
  end for
  lastop(sr)= l + e
  if (l + e = |L_p|) AND (CTmin≤CycleTime(sr)≤CTmax) then
    CTR = CTR ∪ {sr}
  else
    // Dominance check
    if ReduceCost(sr) > ReducedCost(ND[l + e][startperiod(sr)]) then
      // sr dominates the former dominant partial timed route, which ends at period t
      with operation l + e and starting at the same period
      PTR = PTR ∪ {sr}
      PTR = PTR \ {ND[l + e][startperiod(sr)]}
      ND[l + e][startperiod(sr)]= sr
    end if
  end if
end for
```

### 6.2.2 Minimum or maximum cycle times

In this section, we adapt the above developed algorithms in order to consider only minimum or maximum cycle times. In these cases, the dominance rule of Proposition 3 and the proposed algorithms 6.1 and 6.2 remain valid, but more effective dominance rules can be proposed. These dominance rules are given by Proposition 4 (respectively Proposition 5) when only maximum cycle times (respectively only minimum cycle times) are considered.

**Property 4.** *For product $p$ at a period $t$ considering only maximum cycle time constraints, if two partial timed routes $s_1$ and $s_2$, such that the starting period of $s_1$ is larger than the starting period of $s_2$, and $s_1$ and $s_2$ have achieved the same number of operations $l$ until period $t$, then $s_1$ dominates $s_2$ if its partial reduced cost is lower than the partial reduced cost of $s_2$. In other words, for $s_1$=(1,$rc_1$,$t$,$l$) and $s_2$ =(2,$rc_2$,$t$,$l$), then $s_1$ dominates $s_2$ if and only if $t(p, s_1, 1) \geq t(p, s_2, 1)$ and $rc_1 \leq rc_2$.*

*Proof.* We use the same argument as in the proof of property 3. The proof is done by contradiction. Suppose that $s_1$=(1,$rc_1$,$t$,$l$) and $s_2$ =(2,$rc_2$,$t$,$l$) are two partial timed routes such that $t(p, s_1, 1) \geq t(p, s_2, 1)$, $rc_1 \leq rc_2$ and $s_1$ does not dominate $s_2$. This means that there are two partial timed routes $s_3$ and $s_4$ which complete respectively $s_1$ and $s_2$ to obtain

the corresponding optimal routes $s_1 \oplus s_3$ and $s_2 \oplus s_4$. $s_1$ does not dominate $s_2$ means that $rc_1 + rc_3 \geq rc_2 + rc_4$ and thus $rc_3 \geq rc_4$ since $rc_1 \leq rc_2$. Since $s_1$ starts after $s_2$, $s_4$ is also a valid extension of $s_1$. Then $rc_1 + rc_4 \leq rc_2 + rc_4$. This means that $s_1$ dominates $s_2$, which contradicts the assumption that $s_1$ does not dominate $s_2$.

□

**Property 5.** *For product p at a period t considering only minimum cycle time constraints, if two partial timed routes $s_1$ and $s_2$, such that the starting period of $s_1$ is before starting period of $s_2$, and $s_1$ and $s_2$ have achieved the same number of operations l until period t, then $s_1$ dominates $s_2$ if its partial reduced cost is lower than the partial reduced cost of $s_2$. In other words, for $s_1$=(1,$rc_1$,t,l) and $s_2$ =(2,$rc_2$,t,l), then $s_1$ dominates $s_2$ if and only if $t(p, s_1, 1) \leq t(p, s_2, 1)$ and $rc_1 \leq rc_2$.*

*Proof.* This proof is similar to the proof of property 4 □

Note that to certify these dominance rules, more operations are needed than the rule of Property 3 and a more complex algorithm is given by Algorithm 6.3 that replaces Algorithm 6.1. In Algorithm 6.3, all partial timed routes for a given start period are stored in an array ND, indexed by the last operation reached and the start period. Contrary to Algorithm 6.2, every partial timed route that satisfies the cycle time constraint is included in the array ND, regardless of its reduced cost. The filtering of dominated partial timed routes is done at the end of each iteration in Algorithm 6.3, and therefore only non-dominated timed routes can be extended in the next iteration.

The additional complexity introduced by the filtering is in $O(T^2|\mathcal{L}_p|)$ for a given product $p$. Thus, it does not affect the overall complexity.

Note that, when only considering minimum cycle time constraints, the algorithm can be improved without reducing the worst-case complexity. In fact, the dominance rule of Property 5 can be applied only when the minimum cycle time is not reached for a given set of partial timed routes. Once the minimum cycle time is reached, the dominance rule of Property 2 can be applied for these partial timed routes.

### 6.2.3 Numerical experiments

The experiments of this chapter are conducted on a single instance. In order to build on the analysis of Chapter 5, we use the medium size instance with medium demand. Recall that this instance is characterized by 119 micro-periods and 40 products. Note that the experiments of this chapter are only conducted on a reference instance and provide first insights. In order to validate all these insights, experiments on a large variety of instances are needed.

In this section, we first analyze the impact of introducing minimum and maximum cycle time constraints on the objective function and on the computational time. Second, we analyze the relationship between the computational time and the allowed cycle time lengths.

As introduced previously, minimum and maximum cycle times are calculated based on the X-factor. Recall that the X-factor is the ratio of the cycle time to the total cumulative processing time of a route. The X-factor is identified by an interval. For our numerical experiments, the minimum value is fixed to 3.5 and the maximum value to 7. Based on these two values, the minimum cycle time is set to 3.5 times the cumulative processing time and the maximum cycle time to 7 times the cumulative processing time.

**Algorithm 6.3** Generation of timed routes(CTmax)

$CTR = \emptyset$ // $CTR$: Set of complete timed routes
$ir$ // $ir$: Initial partial timed route
lastop($ir$)= 0 // No operation allocated in $ir$
$PTR = \{ir\}$ // $PTR$: Set of current partial timed routes
**for** $t$ = 1 to T **do**
  **for all** $s \in PTR$ **do**
    **if** CycleTime($s$)+leftoverCT(lastop($s$)) $\leq$ CTmax **then**
      CreateExtensions($s$,$t$,0,CTmax,ND[][])
    **else**
      $PTR = PTR - \{s\}$
    **end if**
  **end for**
  // Dominance check
  **for** op = 1 to $|\mathcal{L}_p|$ **do**
    CurrentTimedRoute = null
    CurrentReducedCost = $+\infty$
    **for** $\tau$ = T to 1 **do**
      **if** CurrentReducedCost>ReducedCost(ND[op][$\tau$]) **then**
        **if** CurrentTimedRoute != null **then**
          $PTR \cup$ {CurrentTimedRoute}
        **end if**
        CurrentTimedRoute = ND[op][$\tau$]
        CurrentReducedCost = ReducedCost(ND[op][$\tau$])
      **end if**
    **end for**
  **end for**
**end for**
**return** $CTR$

### Impact of cycle time constraints

In the following, we compare four variants of the flexible lead time model. All models are solved using the column generation algorithm of Chapter 5 with the specific dynamic programming algorithms introduced in this chapter to solve the associated subproblems.

- **NoCT:** The flexible lead time model without any constraint on cycle times. This model is presented in Chapter 5,
- **CTminmax:** The flexible lead time model with minimum and maximum cycle time constraints. The cycle time of a route ranges between 3.5 and 7 times its cumulative processing time,
- **CTmax:** The flexible lead time model with only maximum cycle time constraints,
- **CTmin:** The flexible lead time model with only minimum cycle time constraints.

| | **NoCT** | **CTmin** | **CTmax** | **CTminmax** |
|---|---|---|---|---|
| **Total cost** | 395,025 | 395,025 | 447,570 (+13%) | 461,029 (+17%) |
| **CPU time (s)** | 134 | 2,265 | 253 | 475 |
| **# of iterations** | 43 | 54 | 23 | 26 |
| **Avg CPU time by iteration (s)** | 1.2 | 39.2 | 9.4 | 16.8 |

Table 6.1: Analysis of the impacts of cycle time constraints

From Table 6.1, we can first notice that, as expected, more constraints on cycle times leads to larger total costs. First, we observe that adding only minimum cycle time constraints has no impact on the objective function. Then, the total cost increases by 13% when considering only maximum cycle time constraints. Finally, when introducing both minimum and maximum cycle time constraints, the objective function increases by 17% compared to the model without constraints on cycle times. The impact of adding constraints on cycle times was expected but it is interesting to quantify this impact. We also notice from the same table that the computational time is highly impacted when introducing cycle time constraints. When introducing maximum cycle time constraints, the computational time is almost doubled. When introducing minimum and maximum cycle time constraints, the computational time is multiplied by 3.5. Surprisingly, the impact of introducing minimum cycle time constraints on the computational time is very high since it is multiplied by almost 17. This increase can be explained by more iterations in the column generation algorithm and the high average CPU time spent by each iteration of the column generation algorithm (multiplied by 32). Note that, even if the theoretical complexity of the pricing algorithms for CTmin and CTmax is the same, the computational time for CTmin is 4 times larger. This difference can be explained by the feasibility check that can be performed for CTmax but not for CTmin. Considering the average computational time needed to solve the subproblems for CTmax and CTminmax, it also highly increases compared to the average computational time of NoCT. It is multiplied by 8 when introducing maximum cycle time constraints and by 14 when considering minimum and maximum cycle time constraints.

In order to better understand the results of Table 6.1, in the following we perform an analysis per product. Table 6.2 provides the minimum, the maximum and the mean cycle

time per product and per case. In this table, NoCT is abridged in "no", CTmin in "min", Ctmax in "max" and CTminmax in "minmax".

First, let us analyze the results when introducing only minimum cycle time constraints. We can notice that the difference from the results without any constraint on cycle times is very small. 32 products out of 40 keep the same minimum, maximum and mean cycle times. Note that to obtain the same results, CTmin generates and uses more routes than NoCT in order to obtain the same total inventory and backlog costs. The impact on the cycle times for the remaining 8 products is relatively minor. As expected, the major changes are on minimum cycle times and thus on mean cycle times. A deeper analysis shows that products 31 and 33 have an increase in the minimum cycle time and products 2 and 26 have an increase in the mean cycle time. The other products have a minor increase in the minimum cycle time. We can conclude from these results that the impact of adding minimum cycle time constraints is limited. These results can be explained by the characteristics of the studied instance for which the demand is medium and thus the capacity is not as tight as when considering fixed lead times. Note also that the minimum cycle time constraint might already be respected with the minimum cycle time induced by the flexible lead times.

We can also notice that the impact of considering maximum cycle time constraints is more significant. The introduction of maximum cycle time constraints highly decreases the maximum cycle time (as expected) but also decreases the minimum cycle time. For example, the maximum cycle time of product 24 decreases from 118 to 65 micro-periods and its minimum cycle time decreases from 97 to 62 micro-periods. Only two products (products 35 and 36) out of 40 do not show a decrease in the maximum cycle time. 26 products have seen their minimum cycle time decreasing. We also notice that the decrease of maximum cycle times leads to the decrease of the mean cycle time. Only three products (products 14, 35 and 36) have seen their mean cycle time increasing. Note that product 30 has no timed route, this is due to the strong constraint that forces the cycle time to be lower than 28 micro-periods. The same result can be found with the CTminmax model.

When introducing minimum and maximum cycle time constraints, one could expect that the results will be similar to those obtained when introducing only maximum cycle time constraints. In fact, we have noticed that the impact of minimum cycle constraints is negligible. At best, we could expect that minimum cycle time constraints will limit the decrease of the minimum cycle time observed with the CTmax model. The experiment shows a more complex behavior. The results with CTmax are a good base of comparison for CTminmax, the majority of products have minimum and maximum cycle times that are close. However, the mean cycle time is very different for the majority of products. This means that the allocation of the quantity of products to the timed routes is impacted by the minimum cycle time constraints. Sixteen products have shown an increase in the minimum cycle time. Sometimes, this increase goes beyond the minimum cycle time constraint. This is the case for example for product 36, with a minimum cycle time that reaches 96 micro-periods while, with CTmin or CTmax, it is equal respectively to 84 or 83 micro-periods.

Through this experiment, we can notice that the use of minimum cycle times does not highly impact the results and does not provide the expected behavior. In fact, it looks more interesting to integrate industrial knowledge on operations and their lead times rather than fixing minimum cycle times. This will induce minimum cycle times when constructing the timed routes.

| Product | Minimum cycle time | | | | Maximum cycle time | | | | Mean cycle time | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | no | min | max | min max | no | min | max | min max | no | min | max | min max |
| **1** | 51 | 51 | 50 | 50 | 103 | 103 | 58 | 58 | 66.2 | 66.2 | 53.8 | 53.7 |
| **2** | 42 | 42 | 39 | 38 | 96 | 96 | 54 | 54 | 55.6 | 54.6 | 43.6 | 43.0 |
| **3** | 41 | 41 | 41 | 41 | 103 | 103 | 56 | 56 | 52.4 | 52.4 | 44.7 | 44.6 |
| **4** | 61 | 61 | 59 | 61 | 110 | 110 | 65 | 63 | 75.5 | 75.5 | 61.8 | 62.1 |
| **5** | 49 | 49 | 48 | 48 | 103 | 103 | 62 | 62 | 72.3 | 72.3 | 52.8 | 52.4 |
| **6** | 47 | 47 | 47 | 48 | 110 | 110 | 59 | 58 | 74.1 | 74.1 | 50.2 | 49.1 |
| **7** | 77 | 77 | 77 | 77 | 97 | 97 | 83 | 83 | 81.9 | 81.9 | 80.6 | 79.9 |
| **8** | 55 | 55 | 55 | 55 | 117 | 117 | 59 | 59 | 67.7 | 67.7 | 55.6 | 55.7 |
| **9** | 96 | 96 | 68 | 70 | 103 | 103 | 74 | 70 | 96.4 | 96.4 | 72.8 | 70.0 |
| **10** | 68 | 67 | 65 | 68 | 110 | 110 | 71 | 73 | 80.8 | 80.5 | 68.3 | 69.8 |
| **11** | 49 | 49 | 49 | 49 | 110 | 110 | 57 | 57 | 66.7 | 66.7 | 51.5 | 51.1 |
| **12** | 62 | 62 | 57 | 59 | 97 | 97 | 65 | 65 | 73.9 | 73.9 | 60.8 | 60.4 |
| **13** | 42 | 42 | 42 | 42 | 103 | 103 | 57 | 57 | 63.7 | 63.7 | 42.4 | 42.4 |
| **14** | 69 | 69 | 73 | 70 | 96 | 96 | 74 | 72 | 72.1 | 72.1 | 73.6 | 70.2 |
| **15** | 47 | 47 | 45 | 45 | 110 | 110 | 55 | 55 | 70.3 | 70.3 | 48.8 | 47.6 |
| **16** | 49 | 49 | 47 | 48 | 103 | 103 | 56 | 56 | 67.9 | 67.9 | 48.8 | 49.0 |
| **17** | 69 | 69 | 69 | 70 | 103 | 103 | 75 | 74 | 81.6 | 81.6 | 73.2 | 70.3 |
| **18** | 95 | 89 | 63 | 68 | 96 | 96 | 63 | 68 | 96.0 | 95.6 | 63.0 | 68.0 |
| **19** | 49 | 49 | 48 | 48 | 75 | 75 | 60 | 60 | 56.6 | 56.6 | 53.4 | 53.5 |
| **20** | 63 | 63 | 62 | 62 | 110 | 110 | 68 | 68 | 72.4 | 72.4 | 63.3 | 63.9 |
| **21** | 49 | 49 | 49 | 49 | 110 | 110 | 54 | 54 | 61.8 | 61.8 | 49.9 | 49.7 |
| **22** | 48 | 48 | 46 | 46 | 111 | 110 | 59 | 54 | 76.2 | 75.8 | 50.7 | 47.3 |
| **23** | 69 | 68 | 68 | 69 | 110 | 110 | 72 | 73 | 81.3 | 81.3 | 71.1 | 71.4 |
| **24** | 97 | 97 | 62 | 60 | 118 | 118 | 65 | 63 | 99.8 | 99.8 | 63.1 | 62.0 |
| **25** | 40 | 40 | 40 | 40 | 96 | 96 | 56 | 56 | 55.8 | 55.8 | 44.3 | 43.7 |
| **26** | 67 | 67 | 65 | 69 | 110 | 110 | 72 | 73 | 79.6 | 80.5 | 68.7 | 69.9 |
| **27** | 62 | 62 | 61 | 62 | 117 | 117 | 74 | 74 | 86.6 | 86.6 | 66.3 | 63.8 |
| **28** | 76 | 76 | 75 | 76 | 105 | 105 | 85 | 85 | 81.6 | 81.6 | 79.1 | 79.9 |
| **29** | 55 | 55 | 55 | 55 | 110 | 110 | 58 | 59 | 63.2 | 63.2 | 55.6 | 55.7 |
| **30** | 55 | 55 | - | - | 117 | 117 | - | - | 82.1 | 82.1 | - | - |
| **31** | 47 | 48 | 46 | 46 | 104 | 103 | 59 | 59 | 63.1 | 63.1 | 49.7 | 48.8 |
| **32** | 48 | 48 | 48 | 48 | 103 | 103 | 57 | 57 | 60.4 | 60.4 | 50.6 | 50.4 |
| **33** | 47 | 48 | 46 | 46 | 111 | 110 | 59 | 59 | 64.2 | 64.5 | 49.6 | 48.8 |
| **34** | 89 | 89 | 63 | 63 | 110 | 110 | 68 | 68 | 97.3 | 97.3 | 64.4 | 66.0 |
| **35** | 68 | 68 | 70 | 69 | 69 | 69 | 70 | 75 | 68.9 | 68.9 | 70 | 73.7 |
| **36** | 84 | 84 | 83 | 96 | 84 | 84 | 96 | 96 | 84.0 | 84.0 | 87.1 | 96.0 |
| **37** | 62 | 62 | 59 | 61 | 118 | 118 | 64 | 63 | 79.1 | 79.1 | 62.1 | 61.7 |
| **38** | 63 | 63 | 57 | 58 | 117 | 117 | 69 | 69 | 82.8 | 82.8 | 60.2 | 60.3 |
| **39** | 89 | 89 | 68 | 70 | 103 | 103 | 73 | 73 | 100.3 | 100.3 | 70.0 | 70.3 |
| **40** | 49 | 49 | 45 | 45 | 103 | 103 | 62 | 63 | 63.8 | 63.8 | 50.9 | 51.8 |

Table 6.2: Detailed results on the impact of cycle time constraints

**Analysis of the relation between computational time and cycle time range**

To analyze the impact of the allowed cycle time range, we once again use the flexible lead time model with minimum and maximum cycle time constraints. The minimal value of the cycle time range is fixed to the minimum theoretical cycle time. The maximal value of the cycle time range is calculated as a fraction of the maximum theoretical cycle time which is equal to the length of the time horizon. In our experiments, four values are tested for the maximal value: 100%, 75%, 50% and 25% of the length of the time horizon.

Table 6.3 reports the total inventory and backlogging costs ("Total cost") , the total computational time ("CPU time (s)"), the number of iterations in the column generation ("# of iterations") and the average computational time by iteration ("Avg CPU time by iteration (s)").

| | **Range of cycle times** | | | |
|---|---|---|---|---|
| | **100%** | **75%** | **50%** | **25%** |
| **Total cost** | 395,414 | 397,410 | 403,714 | 419,551 |
| **CPU time (s)** | 2,572 | 2,304 | 1,608 | 1,012 |
| **# of iterations** | 46 | 48 | 44 | 42 |
| **Avg CPU time by iteration (s)** | 52.8 | 45.2 (-14.4%) | 34.3 (-35.0%) | 22.3 (-57.8%) |

Table 6.3: Variation of the cycle time range with CTminmax

From Table 6.3, it can be seen that the decrease of the range of cycle times slightly increases the total cost. We notice that, by dividing the range of the cycle time by four, the increase in the total cost is only 6.1%. Note that this observation does not generalize the previously drawn observations when introducing the minimum and maximum cycle times (see Table 6.1). As expected, we can notice from Table 6.3 that the total computational time decreases as the range of cycle times decreases. Since there is no relationship between the cycle time range and the number of iterations, let us focus on the average computational by iteration of the column generation approach, which highly decreases as the range of cycle times decreases. This is certainly due to the number of generated labels in the dynamic programming algorithm when solving the subproblems. In fact, when the range of cycle times is lower, less labels are generated.

To analyze the computational times when only maximum cycle time constraints are considered, we experimented the same scenarios without considering minimum cycle time constraints. Table 6.4 provides the same indicators as Table 6.3, when using only maximum cycle time constraints.

The same conclusion can be drawn from Table 6.4 regarding the impact of the cycle time range on the total cost. Regarding the computational time, we can notice that compared to the computational time when considering minimum and maximum cycle times (Table 6.4), the computational time when integrating only maximum cycle times is twice shorter. This is mainly due to the strength of the dominance rule that generates less timed routes. If we analyze the average computational time per iteration, we also observe a strong reduction of the computational time when the maximum cycle time reduces. However this reduction is less impressive than when introducing both minimum and maximum cycle times (Table 6.4).

| | Range of cycle times | | | |
|---|---|---|---|---|
| | **100%** | **75%** | **50%** | **25%** |
| **Total cost** | 395,082 | 395,551 | 397,264 | 409,319 |
| **CPU time (s)** | 982 | 925 | 795 | 442 |
| **# of iterations** | 50 | 50 | 48 | 35 |
| **Avg CPU time by iteration (s)** | 17.3 | 16.2 (-6.4%) | 14.4 (-16.8%) | 10.9 (-37.0%) |

Table 6.4: Variation of the cycle time range with CTmax

## 6.3 Alternative costs for timed routes

In Section 6.2, we mainly focused on strong constraints on cycle times. This is one way to control cycle times. In this section, we model these constraints differently. In fact, usually in industry, decision makers have in mind target cycle times or want to challenge their production teams by fixing a target cycle time for some products. In previous chapters, in order to control the cycle time, we introduced a WIP management cost for each product. This WIP management cost has only an impact on the length of the cycle time but does not model the objective to reach target cycle times. With the timed route formulation, it is now possible to consider more explicit strategies to reach target cycle times. In Section 6.3.1, we first introduce target cycle times with different penalties (linear or quadratic). We also generalize the column generation approach to integrate these penalties. In Section 6.3.2, we introduce constraints on intermediate lead times. These sections are supported by numerical experiments to show the impact of these new strategies on the objective function and the structure of the generated timed routes.

### 6.3.1 Target cycle times

As mentioned above, instead of fixing limits to cycle times, it is often relevant to target a cycle time for some products. This can be achieved by penalizing the distance between the effective cycle time and the target cycle time. This penalization can be counted only when the partial timed route is complete. It penalizes the time routes whose effective cycle time is longer or shorter than the target cycle time. Target cycle times are usually used in semiconductor manufacturing in order to control cycle times. Target cycle times are also used to challenge the actual reference cycle times of a factory in order to achieve shorter cycle times.

To implement target cycle times, we introduce a penalty based on the difference between the effective cycle time and the target cycle time. First, we considered the linear cost given by the absolute value of their difference $\Delta_{pr}^{|CT|} = |CT_{\text{target}} - CT_r|$. Then, we considered a quadratic cost to balance cycle time difference $\Delta_{pr}^{CT^2} = (CT_{\text{target}} - CT_r)^2$.

$\Delta_{pr}^{|CT|}$ and $\Delta_{pr}^{CT^2}$ are penalized in the objective function by a small factor to limit the impact of targeting cycle times compared to inventory and backlogging costs. These costs replace the WIP management costs $w_{pr}$ in the objective function (5.1). Contrary to the WIP management cost which is counted at each period, the difference from the target cycle time is only counted at the end of the timed route. The dominance rule on partial timed routes can only be applied on the routes that have the same start period. Thus, the dominance

rule of Property 3 in Section 6.2.1 is applied when penalizing the distance to target cycle times. A stronger dominance rule is applied at the end of the calculation of the final reduced costs when integrating the target cycle time costs. Note that the master problem does not change.

To show the efficiency of introducing target cycle times, we present numerical experiments with four pricing strategies:

- **WIP:** Stationary WIP costs,
- **Linear:** Linear costs to target cycle times,
- **Quad:** Quadratic costs to target cycle times,
- **Quad with tolerance** Quadratic cost to target cycle times, with a tolerance of 4 micro-periods with no cost when exceeding the target cycle time.

Table 6.5 summarizes the experiments of the four above mentioned strategies on the instance used in Section 6.2.3. Recall that this instance is characterized by 119 micro-periods and 40 products. As in the previous tables, the columns "Total cost", "CPU time (s)" and "# of iterations" correspond respectively to the sum of inventory and backlogging costs, the total computational time in seconds and the number of iterations performed in the column generation algorithm.

Note that the results are based on the use of the dominance rule of Property 2 in Section 5.4.2 rather than the dominance rule of Property 3. It is important to mention that the use of the dominance rule of Property 2 is not sufficient to ensure the optimality of the solution. However, its computational time is 10 times lower than the one when using the dominance rule of Property 3. Our experiments showed that the gap to optimality when using the dominance rule of Property 2 is very small (0.005%).

Let us now show why the dominance rule of Property 2 is not sufficient to obtain the optimal solution. Suppose that we have two routes $s_1$ and $s_2$ with a target cycle time of 1 period. For timed route $s_1$, the first operation is processed in period 1, the second one in period 2 and the last one in period 3. For timed route $s_2$, operations 1 and 2 are processed in period 2 and the last operation in period 3. If we use the dominance rule of Property 2 at period 2, $s_1$ dominates $s_2$ if $\lambda_{2,1} > \lambda_{2,2}$. However, $s_2$ can lead to a better cost than $s_1$ if $\Delta_2^{CT} - \Delta_1^{CT} > \lambda_{2,1} - \lambda_{2,2}$. This shows that the dominance rule of Property 2 is not sufficient. Using the same arguments as for the proof of Property 2, we can show that Property 3 is valid for our problem.

| | **Costs** | | | |
|---|---|---|---|---|
| | **WIP** | **Linear** | **Quad** | **Quad with tolerance** |
| **Total costs** | 395,025 | 395,025 | 395,261 | 395,249 |
| **CPU time (s)** | 134 | 114 | 109 | 114 |
| **# of iterations** | 43 | 46 | 46 | 51 |

Table 6.5: Comparison of cost functions on the difference with target cycle times

From Table 6.5, we can notice that using a linear cost on target lead times does not impact the total inventory and backlogging costs. However, we notice that using quadratic

cost on target lead times slightly increases the total inventory and backlogging costs (0.06 %). This increase is smaller when a tolerance of a given number of periods with no cost is allowed (0.057%). We can observe that the computational times are very close for the four strategies. The number of iterations of the column generation algorithm slightly increases when using a quadratic penalty with tolerance but the computational time is not impacted.

To analyze the impact of using target lead times on cycle times, a detailed analysis is needed. Note that, in our case, the target lead times are fixed in order to converge toward the theoretical minimum cycle times. Table 6.6 provides the maximum and mean cycle times for each product and for each strategy.

From Table 6.6, we can observe that maximum cycle times do not change when using the linear cost to the target cycle time. If we analyze in details the mean cycle times, we observe very small changes. Over 40 products, the mean cycle time of four products (products 25, 33, 34 and 35) slightly increase while the mean cycle time of five products (products 14, 18, 23, 28 and 31) decrease. These variations are very small and correspond to a difference of less than 0.1 micro-periods. The largest decrease of the mean cycle time is observed for product 18, where the mean cycle time decreases from 96 to 93.1 micro-periods.

When analyzing the impact of the quadratic cost of the difference to the target cycle times, we can observe more variations. The maximum cycle time of 15 products decrease by at least one micro-period. This decrease can be very high for some products. For example, the mean cycle time of product 21 decreased by more than 28 micro-periods. We can also notice that the mean cycle time is also impacted by the quadratic strategy. A large decrease of the mean cycle time can be observed for the products whose maximum cycle times have decreased. For example, for product 5, the mean cycle time decrease from 72.3 to 58.1 micro-periods. Note that some products have seen their mean cycle times unchanged, even if the maximum cycle times decrease. For products with unchanged maximum cycle times, a small increase of the mean cycle time can be spotted (e.g. products 7 and 17). The largest increase of the mean cycle time can be observed for product 14, with an increase by 5.6 micro-periods.

For the quadratic cost strategy with a tolerance of four micro-periods, the results are close to those with the quadratic cost strategy without any tolerance. Only three products (products 11, 14 and 22) have longer maximum cycle times than with the quadratic cost strategy. Note that these maximum cycle times are smaller than those obtained with the WIP cost strategy. These longer maximum cycle times are associated to some changes of the mean cycle times. For products 11 and 22, the mean cycle time increases respectively by 2.1 and 0.5 micro-periods. For product 14, the mean cycle time decreases by 3.2 micro-periods. Very small changes occur in the mean cycle time of 11 other products.

From these preliminary experiments, we can observe that introducing a tolerance in the quadratic cost strategy does not significantly impact the results. Through these experiments, we have shown that it is possible to easily integrate new constraints and objectives on cycle times without significantly modifying the original column generation algorithm.

### 6.3.2 Smoothing lead times

An issue that might still occur, even when controlling cycle times, is that the cycle time is not balanced along the route and some operations have large lead times. In order to balance the lead times on route, we introduce a linear cost on the maximum lead time between two consecutive operations. The maximum lead time of route $r$ for product $p$ is denoted $LT_{pr}^{\max} = \max_l(t(p,r,l) - t(p,r,l-1))$. This maximum lead time is penalized and replaces the

| Product | Maximum cycle time | | | | Mean cycle time | | | |
|---|---|---|---|---|---|---|---|---|
| | WIP | Linear | Quad | Quad with tolerance | WIP | Linear | Quad | Quad with tolerance |
| **1** | 103 | 103 | 88 | 88 | 66.2 | 66.2 | 66.2 | 62.2 |
| **2** | 96 | 96 | 83 | 83 | 55.6 | 54.6 | 52.9 | 52.9 |
| **3** | 103 | 103 | 82 | 82 | 52.4 | 52.4 | 47.6 | 47.5 |
| **4** | 110 | 110 | 110 | 110 | 75.5 | 75.5 | 75.5 | 75.5 |
| **5** | 103 | 103 | 96 | 96 | 72.3 | 72.3 | 58.1 | 58.1 |
| **6** | 110 | 110 | 103 | 103 | 74.1 | 74.1 | 67.5 | 67.5 |
| **7** | 97 | 97 | 97 | 97 | 81.9 | 81.9 | 82.0 | 82.0 |
| **8** | 117 | 117 | 117 | 117 | 67.7 | 67.7 | 67.7 | 67.7 |
| **9** | 103 | 103 | 103 | 103 | 96.4 | 96.4 | 96.4 | 96.4 |
| **10** | 110 | 110 | 110 | 110 | 80.8 | 80.8 | 81.1 | 80.6 |
| **11** | 110 | 110 | 89 | 96 | 66.7 | 66.7 | 54.1 | 56.2 |
| **12** | 97 | 97 | 97 | 97 | 73.9 | 73.9 | 74.0 | 74.0 |
| **13** | 103 | 103 | 74 | 74 | 63.7 | 63.7 | 47.7 | 47.7 |
| **14** | 96 | 96 | 95 | 96 | 72.1 | 72.0 | 77.7 | 74.5 |
| **15** | 110 | 110 | 110 | 110 | 70.3 | 70.3 | 70.0 | 69.9 |
| **16** | 103 | 103 | 103 | 103 | 67.9 | 67.9 | 67.8 | 67.8 |
| **17** | 103 | 103 | 103 | 103 | 81.6 | 81.8 | 82.1 | 82.6 |
| **18** | 96 | 96 | 96 | 96 | 96.0 | 93.1 | 96.0 | 96.0 |
| **19** | 75 | 75 | 75 | 75 | 56.6 | 56.6 | 56.6 | 56.6 |
| **20** | 110 | 110 | 110 | 110 | 72.4 | 72.4 | 72.4 | 72.4 |
| **21** | 110 | 110 | 82 | 82 | 61.8 | 61.8 | 56.4 | 56.4 |
| **22** | 111 | 111 | 75 | 90 | 76.2 | 76.2 | 64.3 | 64.8 |
| **23** | 110 | 110 | 110 | 110 | 81.3 | 81.2 | 78.9 | 80.2 |
| **24** | 118 | 118 | 118 | 118 | 99.8 | 99.8 | 99.8 | 99.8 |
| **25** | 96 | 96 | 96 | 96 | 55.8 | 55.9 | 55.8 | 56.4 |
| **26** | 110 | 110 | 110 | 110 | 79.6 | 79.6 | 80.2 | 80.7 |
| **27** | 117 | 117 | 117 | 117 | 86.6 | 86.6 | 86.6 | 86.6 |
| **28** | 105 | 105 | 105 | 105 | 81.6 | 82.1 | 81.3 | 81.1 |
| **29** | 110 | 110 | 110 | 110 | 63.2 | 63.2 | 63.2 | 63.2 |
| **30** | 117 | 117 | 117 | 117 | 82.1 | 82.1 | 82.1 | 82.1 |
| **31** | 104 | 104 | 103 | 103 | 63.1 | 63.0 | 62.9 | 63.0 |
| **32** | 103 | 103 | 82 | 82 | 60.4 | 60.4 | 55.5 | 55.5 |
| **33** | 111 | 111 | 97 | 97 | 64.2 | 64.4 | 62.8 | 62.4 |
| **34** | 110 | 110 | 103 | 103 | 97.3 | 97.8 | 95.8 | 95.7 |
| **35** | 69 | 69 | 69 | 69 | 68.9 | 69.0 | 69.0 | 69.0 |
| **36** | 84 | 84 | 84 | 84 | 84.0 | 84.0 | 84.0 | 84.0 |
| **37** | 118 | 118 | 118 | 118 | 79.1 | 79.1 | 79.1 | 79.1 |
| **38** | 117 | 117 | 105 | 105 | 82.8 | 82.8 | 80.3 | 80.3 |
| **39** | 103 | 103 | 103 | 103 | 100.3 | 100.3 | 101.3 | 101.2 |
| **40** | 103 | 103 | 103 | 103 | 63.8 | 63.8 | 62.6 | 62.3 |

Table 6.6: Detailed results for target cycle times

WIP management cost $w_{pr}$ in the objective function (5.1). $LT_{pr}^{\max}$ can be computed during the construction of the associated timed route. Note that we can also use a quadratic cost as done with the target cycle time. As in the previous section, we compare the linear and the quadratic cost of the maximum lead time to the classical WIP management cost strategy.

Note that the dominance rules in Property 2 and Property 3 are not sufficient. The argument is given through the following counterexample. Suppose that we have two timed routes $s_1$ and $s_2$ composed of four operations. $s_1$ and $s_2$ can be represented respectively by the following arrays of periods where the four operations take place: [1][2][3][5] and [1][1][3][5].

We suppose that $s_2$ dominates $s_1$, i.e. that the reduced cost of $s_1$ is larger than the reduced cost of $s_2$ since both routes have the same maximum lead time (equal to 2). We can notice that the difference between the two routes occurs at operation 2 and the maximum lead time is equal to 2 for both routes. $s_2$ dominates $s_1$ means that $\lambda_{2,2} > \lambda_{2,1}$. Recall that $\lambda_{o,t}$ is the dual cost of operation $o$ at period $t$.

However, if we apply the dominance rule of Property 2 on the partial timed routes at period 3, $s_1$ dominates $s_2$ if $\lambda_{2,2} - \lambda_{2,1} < maxLTcost(2,3) - maxLTcost(1,3) =$, where $maxLTcost(r,t)$ is a cost function that provides cost associated with the maximum lead time of timed route $r$ observed at period $t$. This means that the timed route $s_2$ is dominated by route $s_1$ at period 3, which contradicts the assumption that $s_2$ dominates $s_1$.

Table 6.7 summarizes the experiments of the three strategies: "WIP", "Linear" and "Quadratic". As in the previous tables, the columns "Total cost", "CPU time (s)" and "# of iterations" correspond respectively to the sum of the inventory and backlogging costs, the total computational time in seconds and the number of iterations performed in the column generation algorithm. We use the dominance rule of Property 2 in the pricing problem.

| | **Costs** | | |
|---|---|---|---|
| | **WIP** | **Linear** | **Quadratic** |
| **Total costs** | 395,026 | 395,025 | 395,179 |
| **CPU time (s)** | 134 | 114 | 107 |
| **# of iterations** | 43 | 47 | 44 |

Table 6.7: Comparison of various cost functions based on the maximum lead time

Table 6.7 shows the impact of the pricing strategy (WIP, Linear and Quadratic) is negligible. We can notice that, as in the previous section, the quadratic cost has a relatively higher impact than the linear cost. Concerning the computational times, because of the use of the dominance rule of Property 2, there is a very small difference between the three strategies. The same remark can be drawn for the number of iterations for the column generation algorithm.

Table 6.8 presents a detailed analysis by comparing the maximum lead time and the mean cycle time for each product and for each pricing strategy: "WIP", "Linear" and "Quad".

From Table 6.8, we can notice that the linear cost of the maximum lead time does not impact the mean cycle time of the products. Only 12 products out of 40 have seen changes in their mean cycle time when using the linear cost instead of a WIP management cost. The largest difference occurs for product 2 where the mean cycle time decreases by 1 micro-period, but in most cases the differences are lower than 0.2 micro-periods. However, when analyzing the maximum lead times, changes are more frequent and larger. The maximum lead time remains stable for only 11 products. The changes are mostly a decrease of the

| Product | Max lead time | | | Mean cycle time | | |
|---|---|---|---|---|---|---|
| | WIP | Linear | Quad | WIP | Linear | Quad |
| **1** | 52 | 28 | 24 | 66.2 | 66.2 | 66.2 |
| **2** | 36 | 34 | 34 | 55.6 | 54.6 | 54.6 |
| **3** | 60 | 36 | 20 | 52.4 | 52.4 | 47.6 |
| **4** | 23 | 23 | 23 | 75.5 | 75.5 | 75.5 |
| **5** | 52 | 28 | 24 | 72.3 | 72.3 | 72.4 |
| **6** | 59 | 35 | 23 | 74.1 | 74.1 | 67.6 |
| **7** | 14 | 13 | 13 | 81.9 | 82.0 | 81.9 |
| **8** | 60 | 36 | 13 | 67.7 | 67.7 | 65.4 |
| **9** | 12 | 11 | 11 | 96.4 | 96.4 | 96.4 |
| **10** | 22 | 22 | 20 | 80.8 | 80.6 | 80.9 |
| **11** | 59 | 35 | 23 | 66.7 | 66.7 | 61.9 |
| **12** | 15 | 15 | 13 | 73.9 | 73.9 | 74.0 |
| **13** | 60 | 36 | 20 | 63.7 | 63.8 | 54.1 |
| **14** | 10 | 11 | 9 | 72.1 | 72.0 | 81.7 |
| **15** | 61 | 37 | 24 | 70.3 | 70.3 | 63.8 |
| **16** | 52 | 28 | 24 | 67.9 | 67.9 | 67.8 |
| **17** | 13 | 13 | 12 | 81.6 | 81.9 | 82.6 |
| **18** | 13 | 16 | 15 | 96.0 | 96.0 | 95.9 |
| **19** | 18 | 18 | 15 | 56.6 | 56.6 | 56.6 |
| **20** | 24 | 24 | 16 | 72.4 | 72.4 | 72.4 |
| **21** | 59 | 35 | 24 | 61.8 | 61.8 | 60.4 |
| **22** | 61 | 36 | 36 | 76.2 | 76.0 | 75.8 |
| **23** | 16 | 16 | 16 | 81.3 | 81.3 | 77.1 |
| **24** | 57 | 33 | 24 | 99.8 | 99.8 | 99.8 |
| **25** | 52 | 28 | 24 | 55.8 | 55.9 | 55.9 |
| **26** | 23 | 22 | 20 | 79.6 | 80.2 | 79.6 |
| **27** | 34 | 33 | 23 | 86.6 | 86.6 | 86.6 |
| **28** | 13 | 13 | 13 | 81.6 | 81.7 | 80.9 |
| **29** | 52 | 29 | 24 | 63.2 | 63.2 | 63.2 |
| **30** | 47 | 40 | 45 | 82.1 | 82.2 | 82.4 |
| **31** | 58 | 33 | 24 | 63.1 | 63.2 | 63.0 |
| **32** | 52 | 29 | 24 | 60.4 | 60.4 | 60.6 |
| **33** | 61 | 36 | 37 | 64.2 | 64.3 | 64.4 |
| **34** | 45 | 23 | 23 | 97.3 | 97.3 | 96.6 |
| **35** | 3 | 3 | 3 | 68.9 | 69.0 | 69.0 |
| **36** | 4 | 4 | 4 | 84.0 | 84.0 | 84.0 |
| **37** | 57 | 33 | 23 | 79.1 | 79.1 | 79.1 |
| **38** | 50 | 35 | 24 | 82.8 | 82.8 | 80.3 |
| **39** | 13 | 13 | 12 | 100.3 | 100.4 | 100.9 |
| **40** | 56 | 34 | 35 | 63.8 | 63.8 | 63.8 |

Table 6.8: Detailed results for maximum lead time strategies

maximum lead time as with product 3, where the maximal lead time decreases from 60 to 36 micro-periods. We can observe that the larger the maximum lead time when using the WIP management cost, the larger the reduction when using the maximum lead time strategies. We can also notice that, for some products, maximum lead times are reduced. Products 14 and 18 which have originally "short" maximum lead times, present small increases of their maximum lead times when using the linear cost of the maximum lead time.

When focusing on the quadratic cost of the maximum lead time, we can notice that the impact on cycle times is more significant. The maximum lead times of 36 products is reduced. Moreover, the decrease is larger than when using the linear cost of the maximum lead time (for 25 products). For example, the maximum lead time of product 15 is reduced from 61 to 24 micro-periods. The mean cycle time is also highly reduced for 13 products. For example, the mean cycle time of product 13 reduces by 9.6 micro-periods. We also notice a very small increase (of less than 0.1 micro-periods) of the mean cycle time for 11 products. Only the mean cycle time of product 14 increases from 72.1 to 81.7 micro-periods.

Through these results, we can claim that the maximum lead time is reduced as expected. Note that some maximum lead times are relatively long since they can reach 45 micro-periods. However, the maximum cycle time is not highly impacted. We can also note that the production quantity allocated to routes with long maximum lead times are relatively small. In order to avoid such timed routes, hard constraints can be used in order to smooth the lead times over the cycle time, and thus to generate timed routes with shorted lead times.

We have also observed that the impact of the linear cost on the violation of target cycle times or maximum lead times on the solution is negligble. However, the quadratic cost has more impact on the resulting solutions and tends to reach the objectives of shorter and balanced timed routes. Still, the use of unitary penalty costs for the violation of the targets cycle times does not completely forbid long timed routes. It only allocates less production quantities to these routes in order to reduce the total cost. The same remark can be drawn for maximum lead times. They only tend to balance the lead time over the timed route without impacting the maximum cycle time. In order to enforce the control on the timed routes, we can jointly consider hard constraints on cycle times and quadratic costs on the timed routes. In fact, this may help to avoid long cycle times while keeping flexibility within the timed routes.

## 6.4 Conclusion and perspectives

In this chapter, we studied some extensions of the timed route formulation to improve the results obtained in the previous chapter. In fact, one of the drawbacks of using timed routes is the possibility to generate very long cycle times. Thus, our first purpose in this chapter was to introduce new objectives and constraints in order to better control cycle times. First, we studied the construction of the timed routes that respect a minimum cycle time and/or a maximum cycle time. We introduced new dominance rules to adapt the pricing algorithm proposed in Chapter 5 and thus the column generation algorithm. Note that the pricing algorithms remain polynomial when considering the new dominance rules. Note that the proposed extensions can only be considered if they lead to efficient dominance rules or preemptive checks. However, if the dominance rules do not lead to efficient algorithms to solve the pricing problem, heuristics can be considered in order to speedup the column generation algorithm.

Another issue addressed in this chapter is the relevance of using WIP management costs.

In fact, these costs are generally artificial and their goal is to control lead times and cycle times. However, this goal seems not to be reached by the classical formulation. We introduced alternatives to the WIP management costs. The first alternative is to fix a target cycle time per product and to penalize its difference to the actual cycle time. This penalty cost replaces the WIP management cost in the objective function. The second alternative is to penalize the maximum lead time between two operations for each timed route. This cost also replaces the WIP management cost. For the two alternatives, we considered linear and quadratic costs. We have shown that the quadratic cost is more efficient than the linear cost. Note that other penalty functions could be used to obtain more balanced results.

Finally, we believe that it could be interesting to consider more operational constraints within the master problem to deal with industrial needs. For example, the quantities allocated to a route in our approach are continuous. However, in industry, the allocated quantities are discrete or minimum production quantities are needed. This will lead to the introduction of integer variables within the master problem to model minimum production quantities or production using batches. A second extension of our model is to consider more complex Bill of Materials, with production routes sharing common sequences of operations and divergent sequences but leading to equivalent products that satisfy the same demands. Nowadays, in industry in general and in semiconductor manufacturing in particular, the trend is to tend toward the customization of products as late as possible. This means that some products share the same routes up to certain operations and then, when the forecast is accurate, each product follows its own final operations. For all these extensions, the use of binary and/or integer variables is necessary and our column generation algorithm should be adapted to solve the resulting problem in reasonable computational times.

# Chapter 7

# General conclusion and perspectives

## 7.1 General conclusion

The goal of this thesis was to propose novel models and solutions approaches to tackle large-scale production planning problems in semiconductor manufacturing. Our contributions are summarized below.

First, the semiconductor manufacturing context was introduced in Chapter 1, discussing the main issues of probably the most complex industry. Semiconductor manufacturing is mainly characterized by hundreds of operations to complete products on hundreds of machines, long cycle times of several weeks and the high investment cost to build and operate the facility. The semiconductor industry is also characterized with high profits, driven by a growing demand.

Chapter 2 proposes a review of the semiconductor manufacturing literature on production planning. We emphasized two main differences between production planning problems in semiconductor manufacturing and classical production planning problems (in particular lot-sizing problems):

1. The size of the problem instances to solve (hundreds of products, machines and operations), that makes the use of discrete variables (such as the ones used to model set-up times) very difficult,

2. Long cycle times with many operations to perform, which are considered in the literature.

A first gap in the semiconductor manufacturing literature is the discussion on the most relevant objective functions for managers. A second gap is the modeling of congestion. Due to the long cycle times and the complexity of the production flows, congestion cannot be ignored. The semiconductor manufacturing literature proposes three main ways to model congestion (fixed lead times, iterative process updating the lead times and clearing functions). We have tried in this thesis to fill these two gaps, at least partially.

Chapter 3 provided some answers to the first gap on the most relevant objective functions in semiconductor manufacturing production planning. The production planner usually has two main goals: Maximizing the productivity of the facility, but also reaching the financial commitments of the facility. We show that maximizing the profit using the Net Present Value, which relies on a discount rate, helps to achieve both goals. By maximizing the profit, more products are completed, leading to more productivity. However, the end-of-horizon effect needs to be controlled. Otherwise, there is too much inventory of a limited

number of products. To control this effect, we proposed to limit the final inventory to a fraction of the total demand for each product. An alternative is to use a piecewise linear function to model the profit per finished product.

By analyzing the production planning model in Chapter 3, we noticed that fixed lead time constraints were too restrictive, leading to non-smooth production flows. To cope with this issue, we introduced flexible lead time constraints in Chapter 4. Flexible lead times allow products to wait in the queue of an operation as long as needed, while respecting minimum lead times. We have shown that flexible lead times allow larger profits at the cost of longer cycle times. However, the two main flaws of flexible lead time constraints are that (1) They induce significantly larger computational times and (2) Contrary to the fixed lead times, the production flows cannot be controlled, i.e. some production quantities might have very long cycle times.

In Chapter 5, we proposed a reformulation of the semiconductor production planning problem using the new concept of timed route. A timed route is the allocation of each operation of a production route to a period. Using timed routes, we were able to reformulate the production planning problem with fixed lead times but also with flexible lead times. Due to the exponential number of timed routes induced by flexible lead times, a column generation approach was proposed, and a dominance rule established to reduce the time complexity when solving pricing problems. For both fixed and flexible lead times, the timed route formulation (with the column generation approach) performs significantly better in terms of computational times. Moreover, timed routes allow the complete knowledge of the production flows used in the plan.

The timed route formulation can be used in many other situations than just fixed or flexible lead times without modifying the complexity of the master problem. This is what we showed in Chapter 6 with several variations of the pricing problem in the column generation approach. First, we studied timed routes with bounded cycle times, by adapting the algorithm used to solve the pricing problem. Then, we explored alternative costs to WIP management costs such as penalties on the gap to cycle time targets or maximum lead times. These new costs show that, with timed routes, it is also possible to consider nonlinear costs. However, we are far from having fully exploited the potential of the timed route formulation.

## 7.2 Perspectives

There are many perspectives to this work, notably initiated by the introduction of the timed route formulation. In this section, we detail what we think are the most interesting perspectives, to investigate if they can be studied in a short or medium term or if they require long-term research.

### 7.2.1 Polishing the timed route formulation

If the timed route formulation was proven efficient to reduce computational times and explicitly models the production flows (see Chapter 5), the following perspectives would be interesting to investigate:

- Using other objective functions that are more relevant in semiconductor manufacturing than cost minimization, as shown in Chapter 3. The new reduced cost associated with the timed routes should be determined.

- Integrating an initial WIP, that will lead proposing new timed route generation algorithms for truncated timed routes. The initial WIP will be considered through Constraints (7.1), where $\mathcal{R}_{pl}^{\text{init}}$ is the set of truncated timed routes that can be used to process the initial WIP of product $p$ at operation $l$.

$$\sum_{r \in \mathcal{R}_{pl}^{\text{init}}} Z_{pr} = W_{pl0} \qquad \forall p \in \{1, \ldots, P\} \quad \forall l \in \mathcal{L}_p, \tag{7.1}$$

  where $W_{pl0}$ is the initial WIP of product $p$ waiting for operation $l$, and $Z_{pr}$ is the decision variable associated to the quantity of product $p$ allocated to timed route $r$.

  Particular attention must be paid to the complexity induced by the integration of the initial WIP.

- Integrating binary variables in the master problem (for example, with batch constraints, minimum quantities to process or set-up times). If the computational times are too large, dedicated solution approaches should be developed.

- Introducing more complex bills of material. If semiconductor manufacturing's bill of material is always represented as sequential, in practice, it sometimes can be more complex (some products can be merged, others are personalized in the last operations). Moreover, we could consider applying timed route formulation to many different industries if all possible bills of material can be integrated with short computational times.

### 7.2.2 Conclusive comparison of several objective functions

Another perspective opened by this thesis is to compare more efficiently several objective functions for production planning and their hybridization. Rather than simply comparing cost minimization, maximizing the number of "moves" and maximizing profits with Net Present Value as done in Chapter 3, we could also consider maximizing the utilization rates of machines and maximizing the number of finished products. These objective functions can be compared using production planning models with fixed lead times, but also with flexible lead times. Using the timed route formulation, comparisons could be made on cycle times generated by these models.

### 7.2.3 Industrialization of the process

Before industrializing our work, there is at least one large step.

- Our experiments show that the optimized production plans saturate some workshops that are not saturated in reality. We should investigate the reasons behind the categorization of such workshops as bottlenecks. One way would be to simplify our model by ignoring the workshops that are never considered as bottlenecks in the manufacturing facility.

- Detailed procedures to choose the level of aggregation should be proposed. Regarding the product level, we only used the lowest level of products, and thus we ignored part of the demands that are small but not totally insignificant. To consider the full demand, using a higher level of aggregation seems to be a good option. To determine

the capacity consumption at a high level of product aggregation, two solutions can be considered. The simplest is to consider, for each class of products, the most demanded product and take it as the reference product, and only consider its production route. The other alternative is to create a meta production route that, at an operation, can consume resources from several workshops, with a fixed consumption ratio based on the demand for low-level products. A remaining difficulty is the aggregation of the initial WIP of low-level products.

- Rounding heuristics should be studied to provide integer quantities of lots to release. If the rounding heuristics lead to sub-optimal plans due to possibly small production quantities, minimum production quantity constraints should be used in the model.

### 7.2.4 Integrating a better demand qualification

An important aspect in production planning which is not yet very well covered in the literature is the qualification of demands. Some work has been carried out on the integration of the evolution of demand forecast in production planning problems (Heath and Jackson; 1994), and transposed to semiconductor manufacturing (Albey et al.; 2015; Ziarnetzky et al.; 2018). But there are other ways to model a more detailed demand.

There are usually firm demands, for which deadlines are given, and "flexible" demands (typically make-to-stock) with due dates that are not strict. Another dimension of the demand that should not be forgotten is that some demands might be overestimated, while others reflect the exact required quantity.

Considering a fixed demand and an additional demand for a product was discussed in the perspectives of Chapter 3. By using a piecewise linear profit function for the products, products that exceed their base demand can generate lower profits. In financial terms, it makes sense if the quantity of additional products that can be sold is uncertain, and thus the lower profit represents the mean profit between scenarios when products can be sold and scenarios when they cannot.

### 7.2.5 Robustness and consistency of production plans

The main criteria to assess the quality of a production plan are not only the indicators that are calculated directly, but others can be evaluated a posteriori. For example, production plans can be evaluated in various situations that may not exactly match the input data. Furthermore, we need to ensure that the simplifications assumed in the model do not lead to infeasibility at the operational level. In "theory", the best way to test it could be to implement the production plans in the industrial environment. However, due to the criticality of this process and the associated costs/profits at stakes, it is not acceptable.

Usually, detailed simulation is used to assess the quality of production plans. Making a realistic simulation model of an entire semiconductor manufacturing facility is not simpler than establishing a decision model, simplifications are also required to keep a tractable model. However, there are much less simplifications than in a production planning model, because more computational times can be allocated if the simulation is not running in real time (also known as a digital twin), but only offline to assess the consistency of the production plans once and for all. A simulation model of "Front-End" semiconductor manufacturing facilities is developed by Barhebwa-Mushamuka et al. (2019). The authors designed an

optimization-simulation approach that seeks to maintain consistency between the global production schedule and the local schedules in the workshops.

Another way to ensure that our models are resilient to variations of the input data is to use robust optimization. The three main sources of uncertainties in semiconductor manufacturing production planning are: Demands (as expressed in Section 7.2.4), lead times and machine failures, i.e. available production capacities. If a production planning model that considers robustness to lead time variations was established by Albey et al. (2019), it could be interesting to adapt this approach to our timed route formulation. To the best of our knowledge, no paper integrates uncertainties on production capacity in the semiconductor production planning literature. Concerning demand uncertainties, considering the capacity planning literature in semiconductor manufacturing (Hood et al.; 2003) could be useful as a starting point.

### 7.2.6 Extending the models to the supply chain

Another interesting problem in semiconductor manufacturing is related to supply chain planning. It is usually called master planning and can be seen as an intermediate problem between the tactical and strategic decision levels. As stated in Section 2.5, master planning problems in semiconductor manufacturing have been less studied than other planning problems, and only in recent years. Master planning broadly corresponds to production planning on multiple facilities. We could generalize our timed route formulation by considering supply chain routes.

However, master planning also brings new issues. When considering the back-end part of the semiconductor manufacturing supply chain, the allocation of final products to customers should be considered (and also the possible substitution of final products). But a new critical issue in this master planning problem is how to model the life cycle of products (around one year and a half) and the introduction of new products in the supply chain.

# Appendix A

# Extended summary in French

Dans un contexte d'utilisation toujours plus importante d'appareil électroniques que ce soit dans la vie de tous les jours ou dans un cadre industriel voir aussi médical, la demande en semi-conducteur est fortement croissante. Les semi-conducteurs désignent les circuits intégrés que l'on retrouve dans chacun de ses produits. A ce jour, l'industrie des semi-conducteurs est sans doute celle qui présente une production des plus complexes. D'une part, chaque achat de machine est une décision qui relève du domaine stratégique au vue des sommes investis (qui peuvent dépasser le million de dollars). D'autre part il y a la complexité des flux de productions. Les usines européennes de semi-conducteurs ont généralement un portefeuille de plusieurs centaines de produits avec pour chacun des demandes réduites qui empêche une économie d'échelle. Chaque produit nécessite des centaines d'opérations qui sont réalisés par un parc de centaines de machines hétérogènes et polyvalentes. Parmi ces opérations, de nombreuses nécessitent l'utilisation d'un même type de machine ce qui entraine à la fois une compétition pour les capacités de production entre les différents produits mais aussi entre un même produit à différente étapes de sa production. Cette complexité de production est à la racine des congestions qui ont lieu et qui amènent à des temps de cycles de 2 à 3 mois.

Tout cela amène à de nombreuse problématiques au sein des usines de production de semi-conducteurs (plus communément appelées « fab ») comme la cohérence entre les objectifs globaux de production de l'usine et les objectifs locaux au niveau des ateliers ou encore l'équilibrage de l'utilisation de la capacité entre production, maintenance et recherche et développement. La problématique qui nous concerne dans cette thèse est la planification de la production (le choix des quantités de produits et du moment où ils seront lancés pour répondre à la demande), une étape cruciale dans le milieu des semi-conducteurs.

Pour réaliser nos expérimentations numériques nous bénéficions, dans le cadre du projet européen Productive 4.0, d'un modèle générique des données qui formalise la structure des données échangées entre académiques et industriels. C'est sous ce format que sont structurées les données fournit par STMicroelectronics Crolle, partenaire de ce projet. Ces données représentent l'intégralité d'un « fab », que cela soit les ateliers, les produits ou les routes productions. Seul les coûts ne nous ont pas été fourni. Nous avons aussi utilisé une instance de taille réduite présentée dans Kayton et al. (1997).

Dans le Chapitre 2, nous étudions la littérature sur la planification de la production et tout particulièrement dans le cadre de la fabrication de semi-conducteurs. Un rappel est donné concernant les modèles de Lot-Sizing, mais force est de constater que de nombreuses entreprises se privent des méthodes de la recherche opérationnelle, leur préférant un simple

MRP (Materiel Requirement Planning) qui a pour défaut majeur de ne pas prendre en compte la capacité limitée des moyens de productions.

En ce qui concerne les usines de fabrications de semi-conducteurs, le constat est différent. Face à la complexité des flux de production, de nombreuses collaborations entre académiques et industriels ont donné naissance à des modèles linéaires de planification de la production. Ces modèles ont, pour la grande majorité, comme objectif de minimiser les coûts que cela soit d'inventaire ou d'arriéré. Ils s'appuient principalement sur des variables de décisions sur les quantité produites et stockés. A noter qu'à cause des grandes dimensions des instances industrielles, l'emploi de variables entières est quasi inexistant.

Le point d'étude majeur dans ce domaine est la modélisation de la congestion. Trois méthodes majeures sont utilisées : les délais fixes de production, l'ajustement itératif de ces délais de production en utilisant un modèle de simulation et les Clearing Functions qui en fonction de la charge en entrée déterminent quelles quantités de produits pourront être opérées.

Dans le Chapitre 3, nous nous sommes attelés à définir quelle serait la meilleure fonction objective à chercher à satisfaire pour convenir aux besoins de l'entreprise. Dans un premier temps, nous avons introduit toutes les notations nécessaires à la constitution d'un modèle générique de planification de la production pour le semi-conducteur utilisant des contraintes de délais fixes de production, avec pour seul changement par rapport aux classiques de la littérature un horizon à double échelle. L'échelle micro correspond à la production tandis que l'échelle macro correspond aux demandes et aux stocks.

Dans un second temps, afin favoriser la productivité, nous avons étudié une fonction cherchant à maximiser le nombre d'opérations effectuées (tout en prenant en compte les coûts d'inventaire et d'arriéré). Il s'est avéré qu'il y a un arbitrage à faire entre productivité et les coûts classiques.

Il n'est pas chose aisée d'additionner un nombre d'opérations avec des coûts du fait des unités différentes. Afin de parer à cela, la deuxième fonction objective étudiée n'intègre que coûts et profits. Cette fonction vise à maximiser les revenus, mais afin de tenir compte d'une plus grande importance des retours sur investissement rapides, une dévaluation hebdomadaire des profits et coûts est considérée. Cette fonction objective peut être considérée comme la Valeur Actuelle Nette, très connu du monde économique. Les expériences numériques montrent qu'en ne considérant que la minimisation des coûts on passe à côté d'une augmentation non négligeable des revenus. Cette augmentation des revenus est générée par une plus grande production de produits qui ne répondent pas tous à une demande déjà établie. Cela a donc aussi pour effet d'augmenter la productivité comme souhaité avec la précédente fonction objective. Il est à noter que l'effet de fin d'horizon peut amener à générer un très large inventaire d'un seul type de produit. Cet effet n'étant pas désirable il a fallu rajouter des contraintes sur l'inventaire final de chaque produit, ce qui permet de mieux équilibrer l'inventaire final entre tous les produits sans modifier significativement les profits.

Dans le Chapitre 4, la discussion s'est portée sur les contraintes de délais fixes de production qui servent à modéliser les congestions. Dans un premier temps, nous avons exposé les différents inconvénients que génèrent ces contraintes, notamment le manque de lissage des charges sur l'horizon ou encore le fait qu'elles limitent les décisions sur la production uniquement à déterminer les quantités de produits qui seront introduit dans le système à chaque période, les décisions sur les quantités à opérer en chaque atelier étant déjà figées du fait des délais fixes.

Afin de corriger ces défauts, nous avons proposé une contrainte de délai flexible de production. Le principe de cette contrainte est simple : on conserve un délai minimal à respecter mais l'on autorise le produit à patienter plus longuement sans être opéré immédiatement à la fin de son délai minimal comme avec les délais fixes. Ainsi cela permet à la charge d'être lissée sur l'ensemble de l'horizon suivant le délai minimal et laisse des décisions à prendre au niveau de chaque opération.

Les expériences numériques montrent que ces délais flexibles ont pour effet de faire quasiment disparaitre les coûts d'inventaire et de réduire les arriérés grâce à une plus grande flexibilité au niveau de l'utilisation de la capacité. Cependant, cela a aussi un coup en terme de temps de calcul, qui est quasiment multiplié par 20. Un autre inconvénient des délais flexibles est qu'il est beaucoup plus difficile de tracer les flux de production, et il est n'est pas possible de déterminer si un produit va attendre jusqu'à la fin de l'horizon.

Le Chapitre 5 est l'occasion de reformuler intégralement le problème de planification multi-étapes. Dans un premier temps nous introduisons le concept de route temporisée. Cette construction permet d'allouer à chaque opération la période dans laquelle elle sera réalisée. Ainsi on peut déterminer exactement les flux de productions en jeux, leur temps de cycle ainsi que quand sera consumée la capacité. On peut ensuite reformuler le problème de planification sous la forme d'un problème d'allocation de quantités de produits sur les différentes routes temporisées afin de répondre à la demande tout en respectant les contraintes de capacité. Le problème restant était de déterminer quelles routes temporisées utiliser. Dans le cas d'un modèle avec délais fixes de production, elles sont assez simples à déterminer et sont en nombre polynomial. Ainsi il est aisé de les introduire toutes directement dans le modèle.

Cependant, en ce qui concerne le modèle avec délais flexibles cela devient un peu plus complexe, le nombre potentiel de routes temporisées est exponentiel. Pour circonvenir ce problème, une approche par génération de colonne est nécessaire. Un programme dynamique qui permet de générer les routes temporisées est développé ainsi que la règle de dominance suivante « pour deux routes temporisée qui arrivent à la même opération celle qui a le plus fort coût réduit domine l'autre ». Ainsi à chaque itération de la génération de colonne la complexité reste polynomiale. Du côté des expérimentations numériques, le résultat majeur est l'importante réduction des temps de calculs. Pour le modèle avec délais fixes le temps de calcul est réduit d'en moyenne 94%, tandis qu'avec le modèle avec délais flexibles il est réduit d'en moyenne 87%. Un second aspect positif est que l'on a pu étudié précisément les temps de cycles pour une instance du modèle avec délais flexibles. On s'est aperçu que les temps de cycles maximaux étaient proches de la longueur de l'horizon et que régulièrement les routes temporisées contenaient une opération avec un temps d'attente exorbitant qui parfois représente la moitié du temps de cycle.

La formulation à base de routes temporisées recèle d'autres possibilités dont certaines sont explorées dans le Chapitre 6. Afin de répondre à la problématique des temps de cycle trop longs, avec les routes temporisées, on a pu envisager d'attribuer à chacune un coût unitaire qui dépend de longueur du délai maximum entre deux opérations, que cela soit une relation linéaire ou quadratique. Cela est tout simplement impossible avec la formulation classique. De la même manière nous avons testé un coût unitaire qui pénalise l'écart à un temps de cycle cible, avec une pondération linéaire, quadratique et possiblement avec des seuils de tolérances où l'on considère que l'écart est suffisamment faible pour omettre le coût. Dans tous les cas les expérimentations montrent que si l'utilisation des coûts linéaires à un

faible (mais non négligeable) impact sur la solution, lorsque les coûts sont quadratiques, les changements sont majeurs. Cependant même si ces solutions limitent les temps de cycle longs cela ne les bannit pas pour autant. Pour en venir complètement à bout, il faut utiliser des contraintes dures de temps cycle maximal et potentiellement de temps de cycle minimum.

Modifier l'espace des routes temporisées admissibles implique de retravailler les règles de dominances selon si l'on se trouve dans le cas d'une simple contrainte de temps de cycle maximal ou une contrainte qui encadre ce temps de cycle entre deux bornes. Les algorithmes de générations des routes temporisées ont aussi été retravaillé, le tout menant à une complexité plus importante que dans la génération sans contrainte mais toujours polynomiale. Afin réduire un peu cette complexité, on a rajouté une vérification préemptive pour déterminer si la route temporisée en cours de construction va dépasser le temps de cycle maximal. Les résultats numériques montrent qu'imposer de telles limites aux temps de cycle amène à augmenter de façon sensible les coûts totaux. En ce qui concerne le temps de calcul il peut être doublé voire quadruplé dans le cas d'une contrainte en minimum et maximum, mais est très dépendant de l'intervalle entre temps cycle minimal et maximal autorisé.

En ce qui concernent les perspectives des travaux enclenchés dans cette thèse, plusieurs pistes sont envisageables. Tout d'abord il faut perfectionner les routes temporisées, pouvoir intégrer un inventaire intermédiaire initial et d'autres fonctions objectives. L'idéal serait ensuite d'intégrer des variables binaires dans le problème maitre qui pourrait ainsi modéliser des contraintes de quantité minimale à produire.

Une autre tâche concerne l'industrialisation de nos méthodes : il faudra faire encore un peu de travail sur les données en entrée mais surtout sur les données en sortie. En effet des quantités entières seraient plus pratique pour établir des consignes de productions que les quantités en nombre réel déterminées par le programme linéaire. Il faudra donc trouver une bonne heuristique d'arrondissement.

Un enrichissement possible de nos modèles passe par une meilleure compréhension des demandes. Il faut pouvoir gérer des demandes fermes et d'autres plus variables, certaines pouvant admettre un délai tandis que d'autres ne peuvent souffrir le moindre retard.

Enfin il faut s'assurer de la fiabilité des modèles en des conditions réalistes. Pour cela, la première étape est de se confronter à des simulations détaillées d'une usine. Une seconde étape serait d'utiliser une optimisation robuste sur l'un des critères portant à incertitude tel que la demande, les délais, et les capacités de production.

La dernière perspective pour nos travaux serait d'étendre notre formulation à base de routes temporisées au niveau de la Supply Chain. Dans le milieu des semi-conducteurs, les similitudes sont nombreuses entre la planification de la production à l'échelle tactique celle à l'échelle stratégique. Cependant il faudrait prendre en compte l'aspect cycle de vie du produit qui n'est visible qu'à cette échelle de temps là.

## Chapitre 1 : Contexte industriel

**1.1 Introduction**

**1.2 L'industrie des semi-conducteurs**

**1.3 Les challenges de la fabrications de semi-conducteurs**

**1.4 Modèle de données générique et instances industrielles**

## Chapitre 2 : Revue de la littérature

**2.1 Introduction**

**2.2 Introduction à la planification de la production**

**2.3 Planification de la production pour la fabrication de semi-conducteurs**

**2.4 Modélisation de la congestion**

**2.5 Extensions au problème de planification de la production pour la fabrication de semi-conducteurs**

**2.6 Conclusions**

## Chapitre 3 : Maximisation de la productivité et du profit

**3.1 Introduction**

**3.2 Modèle générique**

**3.3 Maximisation d'un indicateur de productivité**

**3.4 Maximization du profit en utilisant un taux d'actualisation**

**3.5 Expérimentations numériques**

**3.6 Effet de bord sur la fin de l'horizon**

**3.7 Conclusions et perspectives**

## Chapitre 4 : Délai de production flexible

**4.1 Introduction**

**4.2 Inconvénients des délais fixes de production**

**4.3 Délais flexibles de production**

**4.4 Expérimentations numériques**

**4.5 Conclusions et perspectives**

## Chapitre 5 : Formulation à base de route temporisées du problème de planification de la production

**5.1 Introduction**

**5.2 Revue de la littérature sur l'utilisation de génération de colonne pour la planification de la production**

**5.3 Une nouvelle formulation à base de routes temporisées**

**5.4 Une approache par génération de colonnes pour résoudre les problèmes avec délais flexibles de production**

**5.5 Expérimentations numériques**

**5.6 Conclusions et perspectives**

## Chapitre 6 : Extensions sur les approches avec routes temporisées

**6.1 Introduction**

**6.2 Le contrôle des temps de cycles**

**6.3 Des coûts alternatifs sur les routes temporisées**

**6.4 Conclusions et perspectives**

# List of Figures

# List of Tables

# List of Algorithms

# Bibliography


Albey, E., Bilge, Ü. and Uzsoy, R. (2014). An exploratory study of disaggregated clearing functions for production systems with multiple products, *International Journal of Production Research* **52**(18): 5301–5322.

Albey, E., Bilge, Ü. and Uzsoy, R. (2017). Multi-dimensional clearing functions for aggregate capacity modelling in multi-stage production systems, *International Journal of Production Research* **55**(14): 4164–4179.

Albey, E., Norouzi, A., Kempf, K. G. and Uzsoy, R. (2015). Demand modeling with forecast evolution: an application to production planning, *IEEE Transactions on Semiconductor Manufacturing* **28**(3): 374–384.

Albey, E., Yanıkoğlu, İ. and Uzsoy, R. (2017). Load dependent lead time modelling: a robust optimization approach, *Proceedings of the 2017 Winter Simulation Conference*.

Albey, E., Yanıkoğlu, İ. and Uzsoy, R. (2019). A robust optimization approach for production planning under exogenous planned lead times, *2019 Winter Simulation Conference (WSC)*, IEEE, pp. 2312–2323.

Armbruster, D. and Uzsoy, R. (2012). Continuous dynamic models, clearing functions, and discrete-event simulation in aggregate production planning, *New Directions in Informatics, Optimization, Logistics, and Production*, INFORMS, pp. 103–126.

Asmundsson, J., Rardin, R. L., Turkseven, C. H. and Uzsoy, R. (2009). Production planning with resources subject to congestion, *Naval Research Logistics (NRL)* **56**(2): 142–157.

Asmundsson, J., Rardin, R. L. and Uzsoy, R. (2006). Tractable nonlinear production planning models for semiconductor wafer fabrication facilities, *IEEE Transactions on Semiconductor Manufacturing* **19**(1): 95–111.

Bang, J.-Y. and Kim, Y.-D. (2010). Hierarchical production planning for semiconductor wafer fabrication based on linear programming and discrete-event simulation, *IEEE Transactions on Automation Science and Engineering* **7**(2): 326–336.

Bansal, A., Uzsoy, R. and Kempf, K. (2020). Iterative combinatorial auctions for managing product transitions in semiconductor manufacturing, *IISE Transactions* **52**(4): 413–431.

Bard, J. F., Deng, Y., Chacon, R. and Stuber, J. (2010). Midterm planning to minimize deviations from daily target outputs in semiconductor manufacturing, *IEEE Transactions on Semiconductor Manufacturing* **23**(3): 456–467.

Barhebwa-Mushamuka, F., Dauzère-Pèrès, S. and Yugma, C. (2019). Work-in-process balancing control in global fab scheduling for semiconductor manufacturing, *2019 Winter Simulation Conference (WSC)*, IEEE, pp. 2257–2268.

Barnhart, C., Johnson, E. L., Nemhauser, G. L., Savelsbergh, M. W. and Vance, P. H. (1998). Branch-and-price: Column generation for solving huge integer programs, *Operations Research* **46**(3): 316–329.

Billington, P. J., McClain, J. O. and Thomas, L. J. (1983). Mathematical programming approaches to capacity-constrained mrp systems: Review, formulation and problem reduction, *Management Science* **29**(10): 1126–1141.

Bredström, D., Lundgren, J. T., Rönnqvist, M., Carlsson, D. and Mason, A. (2004). Supply chain optimization in the pulp mill industry–ip models, column generation and novel constraint branches, *European Journal of Operational Research* **156**(1): 2–22.

Buschkühl, L., Sahling, F., Helber, S. and Tempelmeier, H. (2010). Dynamic capacitated lot-sizing problems: a classification and review of solution approaches, *OR Spectrum* **32**(2): 231–261.

Ceselli, A., Righini, G. and Salani, M. (2009). A column generation algorithm for a rich vehicle-routing problem, *Transportation Science* **43**(1): 56–69.

Chen, M., Sarin, S. and Peake, A. (2010). Integrated lot sizing and dispatching in wafer fabrication, *Production Planning and Control* **21**(5): 485–495.

Chien, C.-F., Dauzère-Pérès, S., Ehm, H., Fowler, J. W., Jiang, Z., Krishnaswamy, S., Lee, T.-E., Moench, L. and Uzsoy, R. (2011). Modelling and analysis of semiconductor manufacturing in a shrinking world: challenges and successes, *European Journal of Industrial Engineering 4* **5**(3): 254–271.

Chou, Y.-C. and Hong, L.-H. (2000). A methodology for product mix planning in semiconductor foundry manufacturing, *IEEE Transactions on Semiconductor Manufacturing* **13**(3): 278–285.

Christ, Q., Dauzère-Pérès, S., Lepelletier, G. and Vialletelle, P. (2018). A multi-purpose operational capacity and production planning tool, *2018 29th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, IEEE, pp. 40–44.

Copil, K., Wörbelauer, M., Meyr, H. and Tempelmeier, H. (2017). Simultaneous lotsizing and scheduling problems: a classification and review of models, *OR Spectrum* **39**(1): 1–64.

Dantzig, G. B. and Wolfe, P. (1960). Decomposition principle for linear programs, *Operations Research* **8**(1): 101–111.

Dauzère-Pérès, S. and Lasserre, J.-B. (1994). Integration of lotsizing and scheduling decisions in a job-shop, *European Journal of Operational Research* **75**(2): 413–426.

Degraeve, Z. and Jans, R. (2007). A new dantzig-wolfe reformulation and branch-and-price algorithm for the capacitated lot-sizing problem with setup times, *Operations Research* **55**(5): 909–920.

Desrosiers, J. and Lübbecke, M. E. (2005). A primer in column generation, *Column generation*, Springer, pp. 1–32.

Ewen, H., Mönch, L., Ehm, H., Ponsignon, T., Fowler, J. W. and Forstner, L. (2017). A testbed for simulating semiconductor supply chains, *IEEE Transactions on Semiconductor Manufacturing* **30**(3): 293–305.

Florian, M., Lenstra, J. K. and Rinnooy Kan, A. (1980). Deterministic production planning: Algorithms and complexity, *Management Science* **26**(7): 669–679.

Fowler, J. and Robinson, J. (1995). Measurement and improvement of manufacturing capacity (mimac) project final report, sematech technology transfer# 95062861a-tr. austin, tx, 1995. also published as: Manufacturing science and technology for ic production, jessi t30c/esprit 8003, theme 3.3, *Technical report*, MST3-AI300-R-NI04-1.

Gamache, M., Soumis, F., Marquis, G. and Desrosiers, J. (1999). A column generation approach for large-scale aircrew rostering problems, *Operations Research* **47**(2): 247–263.

Goldratt, E. M. (1990). *Theory of constraints*, North River Croton-on-Hudson.

Gómez Urrutia, E. D., Aggoune, R. and Dauzère-Pérès, S. (2014). Solving the integrated lot-sizing and job-shop scheduling problem, *International Journal of Production Research* **52**(17): 5236–5254.

Gopalswamy, K. (2019). *Production Planning with Clearing Functions: Data-driven Approaches and Conic Programming.*, PhD thesis.

Graves, S. C. (1986). A tactical planning model for a job shop, *Operations Research* **34**(4): 522–533.

Habla, C. and Mönch, L. (2008). Solving volume and capacity planning problems in semiconductor manufaturing: a computational study, *Proceedings of the 40th Conference on Winter Simulation*, Winter Simulation Conference, pp. 2260–2266.

Habla, C., Monch, L. and Driebel, R. (2007). A finite capacity production planning approach for semiconductor manufacturing, *Automation Science and Engineering, 2007. CASE 2007. IEEE International Conference on*, IEEE, pp. 82–87.

Hamed, A., Ehm, H., Ponsignon, T., Bayer, B. and Kabak, K. E. (2018). Flexibility as an enabler for carbon dioxide reduction in a global supply chain: a case study from the semiconductor industry, *2018 Winter Simulation Conference (WSC)*, IEEE, pp. 3408–3419.

Hassoun, M., Kopp, D., Mönch, L. and Kalir, A. (2019). A new high-volume/low-mix simulation testbed for semiconductor manufacturing, *2019 Winter Simulation Conference (WSC)*, IEEE, pp. 2419–2428.

Heath, D. C. and Jackson, P. L. (1994). Modeling the evolution of demand forecasts ith application to safety stock analysis in production/distribution systems, *IIE transactions* **26**(3): 17–30.

Hood, S. J., Bermon, S. and Barahona, F. (2003). Capacity planning under demand uncertainty for semiconductor manufacturing, *IEEE Transactions on Semiconductor Manufacturing* **16**(2): 273–280.

Hung, Y.-F. and Leachman, R. C. (1996). A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations, *IEEE Transactions on Semiconductor Manufacturing* **9**(2): 257–269.

Hwang, T.-K. and Chang, S.-C. (2003). Design of a lagrangian relaxation-based hierarchical production scheduling environment for semiconductor wafer fabrication, *IEEE Transactions on Robotics and Automation* **19**(4): 566–578.

Irdem, D. F., Kacar, N. B. and Uzsoy, R. (2010). An exploratory analysis of two iterative linear programming—simulation approaches for production planning, *IEEE Transactions on Semiconductor Manufacturing* **23**(3): 442–455.

Jampani, J. and Mason, S. J. (2010). A column generation heuristic for complex job shop multiple orders per job scheduling, *Computers & Industrial Engineering* **58**(1): 108–118.

Jans, R. and Degraeve, Z. (2004). An industrial extension of the discrete lot-sizing and scheduling problem, *IIE Transactions* **36**(1): 47–58.

Kacar, N. B., Irdem, D. F. and Uzsoy, R. (2012). An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms, *IEEE Transactions on Semiconductor Manufacturing* **25**(1): 104–117.

Kacar, N. B., Monch, L. and Uzsoy, R. (2013). Planning wafer starts using nonlinear clearing functions: A large-scale experiment, *IEEE Transactions on Semiconductor Manufacturing* **26**(4): 602–612.

Kacar, N. B., Mönch, L. and Uzsoy, R. (2016). Modeling cycle times in production planning models for wafer fabrication, *IEEE Transactions on Semiconductor Manufacturing* **29**(2): 153–167.

Kayton, D., Teyner, T., Schwartz, C. and Uzsoy, R. (1997). Focusing maintenance improvement efforts in a wafer fabrication facility operating under the theory of constraints, *Production and Inventory Management Journal* **38**(4): 51.

Kim, B. and Kim, S. (2001). Extended model for a hybrid production planning approach, *International Journal of Production Economics* **73**(2): 165–173.

Kim, S. H., Kim, J. W. and Lee, Y. H. (2014). Simulation-based optimal production planning model using dynamic lead time estimation, *The International Journal of Advanced Manufacturing Technology* **75**(9-12): 1381–1391.

Kim, S. H. and Lee, Y. H. (2016). Synchronized production planning and scheduling in semiconductor fabrication, *Computers & Industrial Engineering* **96**: 72–85.

Kim, S. and Uzsoy, R. (2008). Exact and heuristic procedures for capacity expansion problems with congestion, *IIE Transactions* **40**(12): 1185–1197.

Kriett, P. O., Eirich, S. and Grunow, M. (2017). Cycle time-oriented mid-term production planning for semiconductor wafer fabrication, *International Journal of Production Research* **55**(16): 4662–4679.

Laipple, G., Dauzère-Pérès, S., Ponsignon, T. and Vialletelle, P. (2018). Generic data model for semiconductor manufacturing supply chains, *2018 Winter Simulation Conference (WSC)*, IEEE, pp. 3615–3626.

Leachman, R. C. and Carmon, T. F. (1992). On capacity modeling for production planning with alternative machine types, *IIE transactions* **24**(4): 62–72.

Liberatore, M. J. and Miller, T. (1985). A hierarchical production planning system, *Interfaces* **15**(4): 1–11.

Lim, S.-K., Kim, J.-G. and Kim, H.-J. (2014). Simultaneous order-lot pegging and wafer release planning for semiconductor wafer fabrication facilities, *International Journal of Production Research* **52**(12): 3710–3724.

Liu, J., Li, C., Yang, F., Wan, H. and Uzsoy, R. (2011). Production planning for semiconductor manufacturing via simulation optimization, *Proceedings of the 2011 Winter Simulation Conference (WSC)*, IEEE, pp. 3612–3622.

Lowe, J. J. and Mason, S. J. (2016). Integrated semiconductor supply chain production planning, *IEEE Transactions on Semiconductor Manufacturing* **29**(2): 116–126.

Manne, A. S. (1958). Programming of economic lot sizes, *Management Science* **4**(2): 115–135.

Missbauer, H. (2020). Order release planning by iterative simulation and linear programming: Theoretical foundation and analysis of its shortcomings, *European Journal of Operational Research* **280**(2): 495–507.

Missbauer, H. and Uzsoy, R. (2020). *Production Planning with Capacitated Resources and Congestion*, Springer.

Modigliani, F. and Hohn, F. E. (1955). Production planning over time and the nature of the expectation and planning horizon, *Econometrica, Journal of the Econometric Society* pp. 46–66.

Mönch, L., Fowler, J. W. and Mason, S. J. (2012). *Production planning and control for semiconductor wafer fabrication facilities: modeling, analysis, and systems*, Vol. 52, Springer Science & Business Media.

Mönch, L., Uzsoy, R. and Fowler, J. W. (2018a). A survey of semiconductor supply chain models part i: semiconductor supply chains, strategic network design, and supply chain simulation, *International Journal of Production Research* **56**(13): 4524–4545.

Mönch, L., Uzsoy, R. and Fowler, J. W. (2018b). A survey of semiconductor supply chain models part iii: master planning, production planning, and demand fulfilment, *International Journal of Production Research* **56**(13): 4565–4584.

Ng, T. S., Sun, Y. and Fowler, J. (2010). Semiconductor lot allocation using robust optimization, *European Journal of Operational Research* **205**(3): 557–570.

Nisted, L., Pisinger, D. and Altman, A. (2011). Optimal wafer cutting in shuttle layout problems, *Journal of Combinatorial Optimization* **22**(2): 202–216.

Pahl, J. (2012). Production planning with load dependent lead times and sustainability aspects.

Pahl, J., Voß, S. and Woodruff, D. L. (2005). Load dependent lead times—from empirical evidence to mathematical modeling, *Research methodologies in supply chain management*, Springer, pp. 539–554.

Perraudat, A., Dauzère-Pérès, S. and Vialletelle, P. (2019). Evaluating the impact of dynamic qualification management in semiconductor manufacturing, *2019 Winter Simulation Conference (WSC)*, IEEE, pp. 2336–2347.

Ponsignon, T. and Eng, M. (2012). *Modeling and Solving Master Planning Problems in Semiconductor Manufacturing*, PhD thesis, Fernuniversität Hagen.

Rowshannahad, M., Dauzere-Peres, S. and Cassini, B. (2015). Capacitated qualification management in semiconductor manufacturing, *Omega* **54**: 50–59.

Schömig, A. and Fowler, J. (2000). Modeling semiconductor manufacturing operations, *Proceedings of the 9th ASIM dedicated conference simulation in production and logistics*, pp. 55–64.

Spier, J. and Kempf, K. (1995). Simulation of emergent behavior in manufacturing systems, *Proceedings of SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, IEEE, pp. 90–94.

Spitter, J., Hurkens, C. A., De Kok, A., Lenstra, J. K. and Negenman, E. G. (2005). Linear programming models with planned lead times for supply chain operations planning, *European Journal of Operational Research* **163**(3): 706–720.

Sugimori, Y., Kusunoki, K., Cho, F. and UCHIKAWA, S. (1977). Toyota production system and kanban system materialization of just-in-time and respect-for-human system, *International Journal of Production Research* **15**(6): 553–564.

Taylor, F. W. (1911). The principles of scientific management.

van Den Akker, J. M., Hoogeveen, J. A. and van de Velde, S. L. (1999). Parallel machine scheduling by column generation, *Operations Research* **47**(6): 862–872.

van den Heuvel, W. and Wagelmans, A. P. (2005). A comparison of methods for lot-sizing in a rolling horizon environment, *Operations Research Letters* **33**(5): 486–496.

Van Hoesel, C. and Wagelmans, A. P. M. (1996). An o (t 3) algorithm for the economic lot-sizing problem with constant capacities, *Management Science* **42**(1): 142–150.

Van Zant, P. (2004). *Microchip fabrication*, McGraw-Hill, Inc.

Villarreal, S., Jimenez, J. A., Jin, T. and Cabrera-Rios, M. (2012). Designing a sustainable and distributed generation system for semiconductor wafer fabs, *IEEE Transactions on Automation Science and Engineering* **10**(1): 16–26.

Wagner, H. M. and Whitin, T. M. (1958). Dynamic version of the economic lot size model, *Management Science* **5**(1): 89–96.

Wight, O. (1995). *Manufacturing resource planning: MRP II: unlocking America's productivity potential*, John Wiley & Sons.

Yi, J., Jia, S.-j., Du, B. and Liu, Q. (2019). Multi-objective model and optimization algorithm based on column generation for continuous casting production planning, *Journal of Iron and Steel Research International* **26**(3): 242–250.

Zhang, F., Song, J., Dai, Y. and Xu, J. (2020). Semiconductor wafer fabrication production planning using multi-fidelity simulation optimisation, *International Journal of Production Research* pp. 1–16.

Ziarnetzky, T. and Mönch, L. (2016). Simulation-based optimization for integrated production planning and capacity expansion decisions, *Proceedings of the 2016 Winter Simulation Conference*, IEEE Press, pp. 2992–3003.

Ziarnetzky, T., Mönch, L., Kannaian, T. and Jimenez, J. (2017). Incorporating elements of a sustainable and distributed generation system into a production planning model for a wafer fab, *2017 Winter Simulation Conference (WSC)*, IEEE, pp. 3519–3530.

Ziarnetzky, T., Mönch, L., Ponsignon, T. and Ehm, H. (2019). Integrated planning of production and engineering activities in semiconductor supply chains: A simulation study, *2019 Winter Simulation Conference (WSC)*, IEEE, pp. 2324–2335.

Ziarnetzky, T., Mönch, L. and Uzsoy, R. (2018). Rolling horizon, multi-product production planning with chance constraints and forecast evolution for wafer fabs, *International Journal of Production Research* **56**(18): 6112–6134.

Ziarnetzky, T., Mönch, L. and Uzsoy, R. (2019). Simulation-based performance assessment of production planning models with safety stock and forecast evolution in semiconductor wafer fabrication, *IEEE Transactions on Semiconductor Manufacturing* .

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2020LYSEM015

Sébastien BERAUDY

# NEW APPROACHES FOR LARGE SCALE MULTI-STEP PRODUCTION PLANNING PROBLEMS

Speciality: Industrial Engineering

Keywords: Production planning, Semiconductor manufacturing, Lead time, Timed route, Column generation approach

Abstract:

Conjugating highly increasing demands, limited capacity (because each machine costs a lot) and cycle times larger than two months, the semiconductor industry (which produces integrated circuits) is the most complex. In the semiconductor manufacturing, the production planning is critical and should consider various phenomena such as re-entrant flows and congestion. The semiconductor manufacturing literature strongly focuses on the way to model congestion notably using fixed lead times.

We first discuss what goal should pursue the production planner. Only minimizing the costs is not enough. To ensure profits and productivity of the facility, the chosen objective function maximizes the profits under an actualization rate.

Our study was next dedicated to the fixed lead time constraints that are used to model congestion. Acknowledging the drawbacks of fixed lead times, we propose news constraints called flexible lead times. Flexible lead times offer more flexibility in the planning by allowing products to wait more than the minimum lead time before being processed.

However, the flexible lead times lead to both a harder visualization of production flows and large computational times. To answer these two drawbacks, we propose a reformulation of the production planning problem using timed routes. In a timed route, each operation is allocated to a fixed period.

Another advantage of the timed route formulation is that large cycle times and lead times (that are common when using flexible lead times) can be banned or dissuaded.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2020LYSEM015

Sébastien BERAUDY

# NOUVELLES APPROCHES POUR LES PROBLEMES DE PLANIFICATION DE LA PRODUCTION MULTI-ETAPES DE GRANDE DIMENSION

Spécialité: Génie Industriel

Mots clefs : Planification de la production, Fabrication de semi-conducteur, Délai de production, Route temporisée, Génération de colonnes

Résumé:

Entre demande fortement croissante, capacité limitée (car extrêmement coûteuse) et processus longs de plus de deux mois, l'industrie des circuits intégrés (autrement appelés semi-conducteurs) est des plus complexes. La planification de la production y est cruciale et doit prendre en compte de nombreux phénomènes comme les flux réentrants de produits et les congestions. Dans la littérature associée, un important effort a été fait pour modéliser les congestions en utilisant notamment des délais fixes de production.

La discussion s'est d'abord portée sur les objectifs poursuivis par les entreprises qui ne peuvent pas se contenter de minimiser les coûts. Afin de garantir les profits mais aussi la productivité de l'usine, une fonction objective maximisant les profits avec dévaluation temporelle de ces résultats financiers est préconisée. L'étude a ensuite porté sur les contraintes de délais fixes de production qui servent à modéliser les congestions et montrer leurs inconvénients. Nous avons proposé des délais flexibles qui en autorisant de rallonger les délais offrent plus de flexibilité au système de production.

Cependant ces délais flexibles mènent à la fois une moindre visibilité des flux de productions et une augmentation importante du temps de calcul. Pour parer à cela, une reformulation du problème de planification de la production est proposée, en utilisant des routes temporisées i.e. à chaque opération d'un produit est attribuée la période dans laquelle elle sera effectuée. Un autre avantage de cette formulation est que l'on peut interdire ou décourager les temps de cycle et les délais trop longs, qui sont courants avec les modèles utilisant des délais flexibles.