Novel optimization approaches for global fab scheduling in semiconductor manufacturing

Félicien Barhebwa-Mushamuka

N°d'ordre NNT : **2020LYSEM020**

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 20/11/2020, par :

Félicien BARHEBWA-MUSHAMUKA

Novel Optimization Approaches for Global Fab Scheduling in Semiconductor Manufacturing

---

Devant le jury composé de :

Sauer, Nathalie, Professeure, Université de Lorraine	Présidente
Jussien-Guéret, Christelle, Professeure, Université d'Angers	Rapporteure
Amodeo, Lionel, Professeur, Université de Technologie de Troyes	Rapporteur
Lacomme, Philippe, Maître de conférences HDR, Université Clermont Auvergne	Examinateur
Vermarien, Leon, Ingénieur STMicroelectronics, Rousset	Examinateur
Yugma, Claude, Professeur, EMSE, Gardanne	Directeur de thèse
Stéphane Dauzère-Pérès, Professeur, EMSE, Gardanne	Co-directeur de thèse
Levasseur, Sandra, Ingénieur STMicroelectronics, Crolles	Invitée
Vialletelle, Philippe, Ingénieur STMicroelectronics, Crolles	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

This thesis project would not have been possible without the help and support of many people. I would like to express my gratitude to Claude Yugma and Stéphane Dauzère-Pérès, for their constant support, advice and invaluable encouragement throughout these years. They were very kind and generous with their time to oversee this work.

I would like to thank Christelle Jussien-Guéret and Lionel Amodeo for agreeing to be part of this thesis evaluation committee and for reviewing this manuscript. I would also like to thank Nathalie Sauer, Philippe Lacomme, Léon Vermarien, Sandra Levasseur and Philippe Vialletelle for accepting the invitation to be part of the evaluation committee.

Warm thanks to all the professors I met throughout my studies, in particular to Marie-Christine Costa for always supporting me with her advice and encouragement over the past five years, I address these words warm to Anne Pardoen, Valerie Ferrarini and Evgenia Shabalina. A big thank you to my colleagues, friends, especially Catherine Grillot and all my family for their support through this program and a special dedication to my mother who will never be with us physically, but always present in our hearts.

Finally, this thesis was part of the productive 4.0 project, which is a European project included in the Horizon 2020 program. I would also like to thank all the members of the productive 4.0 project, in particular STMicroelectronics for this fruitful collaboration.

Contents

General Introduction

Semiconductor manufacturing systems, as other manufacturing systems, transform raw materials into finished products. The first primary raw material used in semiconductor manufacturing was germanium. The germanium is now replaced by silicon due to its abundance, its resistance to very high temperatures, etc. The finished product is made up of different types of electronic devices depending on the technology used. These devices include resistors, diodes, transistors, integrated circuits, etc.

Semiconductor devices are the foundation of the electronics industries. From small businesses to large businesses such as the automotive, telecommunications and aerospace industries, semiconductors are ubiquitous. The growth in semiconductor manufacturing in recent years is mainly due to the growth in demand for smart phones, cloud computing and other high-level electronic devices. The world is now talking about Industry 4.0, autonomous driving, Artificial Intelligence, the Internet of Things, etc. These emerging technologies will continue to maintain the growth and impact of semiconductor manufacturing in our daily lives. In addition to operating in a rapidly growing market, semiconductor manufacturing presents major challenges which probably makes it the most complex industry. Products are produced on the basis of hundreds of complex operations and they spend on average two to three months in the system. The main concern of semiconductor companies is how to stay competitive in this growing market. Hence, they must find new and robust strategies to produce efficiently and stay competitive.

High-level decisions in an industry such as strategic and tactical decisions not only determine the industry's strategies for finished products, but also define how the industry must stay competitive in the market as well as in its Supply Chain. The operational decision level allows the industry to achieve its objectives. Indeed, at this level of decision, the manufacturing of products is materialized. In semiconductor manufacturing and other complex systems, the effective management of the operational decision level remains the basis for achieving short, medium and long-term objectives, thus enabling the company to remain competitive and viable in a growing market. This thesis is based at the operational decision level, in particular global (i.e, factory-wide) scheduling decisions. The structure of the thesis consists of six chapters described below.

Chapter 1 presents the industrial and scientific context in which this thesis takes place as well as the motivations and main objectives of the thesis. We describe the main components of the semiconductor manufacturing system as well as the main manufacturing processes. We briefly present the different decision levels in semiconductor manufacturing. Next, related work on consistency between decision levels in semiconductor manufacturing is reviewed, followed by related work on Work-In-Process and cycle time management strategies. Finally, related work on simulation in semiconductor manufacturing is reviewed.

Chapter 2 presents the new global scheduling approach that we propose to solve the problem under study as well as the way in which this approach is evaluated. The global

scheduling approach is a mechanism we propose to steer scheduling decisions at work-center (group of machines with same capabilities) level. It adopts two views of the operational decision level:

- The global level (factory level), which uses the global information (Work-In-Process in the whole factory, lot releases, cycle time targets, aggregate resource capacity, etc.) and provides objectives to the local level,
- The local level (work-center level), which uses local information (waiting times of lots, processing times, lots currently in queues, etc.). It receives objectives from the global level and tracks these objectives using dispatching rules or scheduling algorithms in each work-center.

The principle that guides our approach consists in the determination of production targets (objectives) at the global level that should be followed at local level and updated regularly. Production targets are quantities to complete for each product in each operation and each period on a scheduling horizon. The main levers of our approach are global scheduling strategies, implemented in global scheduling models. Global scheduling models determine production targets to optimize different objectives. The main parameters of the global scheduling approach include the scheduling horizon, the length of each period in the scheduling horizon and the time at which the global scheduling model is applied (triggering horizon). These parameters are required for the evaluation of the global scheduling approach. The chapter ends by describing the simulation environment in which the approach is evaluated.

Chapter 3 presents the single objective and multi-objective global scheduling strategies for balancing the Work-In-Process and maximizing the throughput. The chapter begins by introducing the concept of *balancing coefficients*. Balancing coefficients are percentages of the Work-In-Process of each product that should remain in the system at the end of each period on a scheduling horizon. Different ways of determining the balancing coefficients are discussed on the basis of the release scheme, the estimated throughput and Little's law. Next, the chapter discusses the single-objective global scheduling strategy called Work-In-Process balancing control which tries to ensure that the Work-In-Process is properly distributed throughout the whole factory. This strategy aims to control the flow of products to minimize the output variability on cycle times and throughput and to speed up products. Like all the global scheduling strategies discussed in this thesis, this strategy is implemented using a Linear Programming model. The strategy is enforced with a Work-In-Process balancing penalty in the objective function and smoothing constraints. The chapter also discusses a multi-objective global scheduling strategy. This strategy aims to maximize throughput and minimize the output variability on cycle times and throughput. The multi-objective global scheduling strategy is solved using an ϵ-constraint approach.

In Chapter 4, global scheduling strategies for minimizing and controlling cycle times are discussed. Two different global scheduling strategies are compared for cycle time minimization, the *push strategy* and the *time at operation strategy*. The global scheduling strategy for controlling cycle times manages the Work-In-Process to minimize the tardiness (positive gap) on given cycle time targets of products. After grouping the operations of products in subsequences (blocks of operations), cycle times are then controlled through three main parameters:

- The cycle time target of each block, which is derived from the cycle time target of the product,

- The classes of release dates, which are aggregations of release dates of quantities of product released in the factory and,
- The temporal tracing of the Work-In-Process, i.e., the management of the Work-In-Process based on the time the Work-In-Process have already spent in the factory.

Two methods for determining cycle time targets of blocks are considered, a naive method and a simulation-based method.

Chapter 5 presents multi-objective global scheduling strategies for controlling cycle times. These strategies aim to minimize the positive and negative gaps from the cycle time targets. The Work-In-Process is not only managed according to the tardiness, but also according to the earliness on the cycle time targets of blocks. Different combinations of penalty costs on the tardiness and earliness are tested and compared.

We conclude the manuscript with Chapter 6 where general conclusions and short-term and long-term perspectives on the global scheduling approach are given.

Chapter 1

Industrial and Scientific Context

1.1 Indroduction

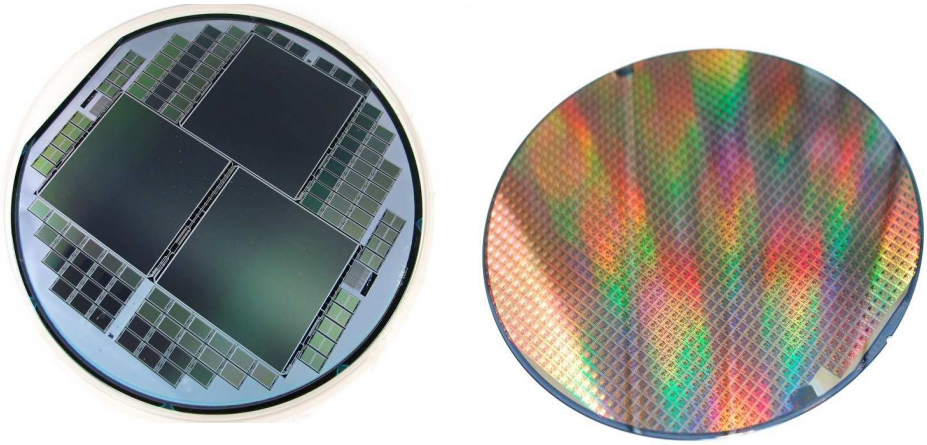


Figure 1.1: Semiconductor row material "Wafer" (source: Flickr, Rob Bulmahn, http://www.flickr.com/photos/ (CC License))

Electronic devices surround us whether in our homes or our workplaces such as radio, telephones, television, computers, advanced medical diagnostic equipment and other high-tech devices. These devices include many electronic components, such as diodes, resistors and transistors. The way these components are manufactured and assembled so that we can have the electronic devices we use daily is the foundation of the semiconductor manufacturing industry.

The factors underlying the main economic challenges facing the semiconductor industry are generally based on increased costs and investment in research and development (R&D). This is due to the increased costs of upgrading existing manufacturing plants and new construction to effectively meet market expectations. Other significant costs in the manufacture of semiconductors are the costs of the machines used for processing jobs, that are extremely expensive, some of them up to the US$ 40 million, and are therefore scarce resources (Mönch et al. (2012)). Equipment costs alone account for more than 70% of the total indirect cost (May and Spanos (2006)).

Aside from economic challenges, semiconductor manufacturing processes are probably the most complex manufacturing processes (Mönch et al. (2012)). In addition to certain

common characteristics that can be found in most manufacturing contexts, the manufacture of semiconductors includes characteristics that make production very complex, such as re-entrant flows induced mainly by scarce and expensive resources, hundreds of operations for each product leading to very long cycle times, different types of scheduling problems, etc.

Section 1.2 introduces the industrial context. In Section 1.3, the scientific context is presented. Section 1.4 presents the motivation and main objectives of the thesis. Finally, Section 1.5 concludes and positions the thesis according to the scientific context.

1.2 Industrial Context

We describe the main parts of the semiconductor manufacturing system as well as the main manufacturing process in Section 1.2.1. Next, we briefly present in Section 1.2.2 different levels of decisions in semiconductor manufacturing.

1.2.1 Semiconductor Manufacturing Processes: An Overview

The process of manufacturing Integrated Circuits can be summarized in two main parts. The first part, semiconductor wafer fabrication (wafer fab) or front-end, corresponds to the long and complex process of manufacturing silicon chips on silicon wafers. The second part, back-end, corresponds to the cutting and packaging of the chips and the final tests. The manufacturing process of an integrated circuit is summarized on Figure 1.2

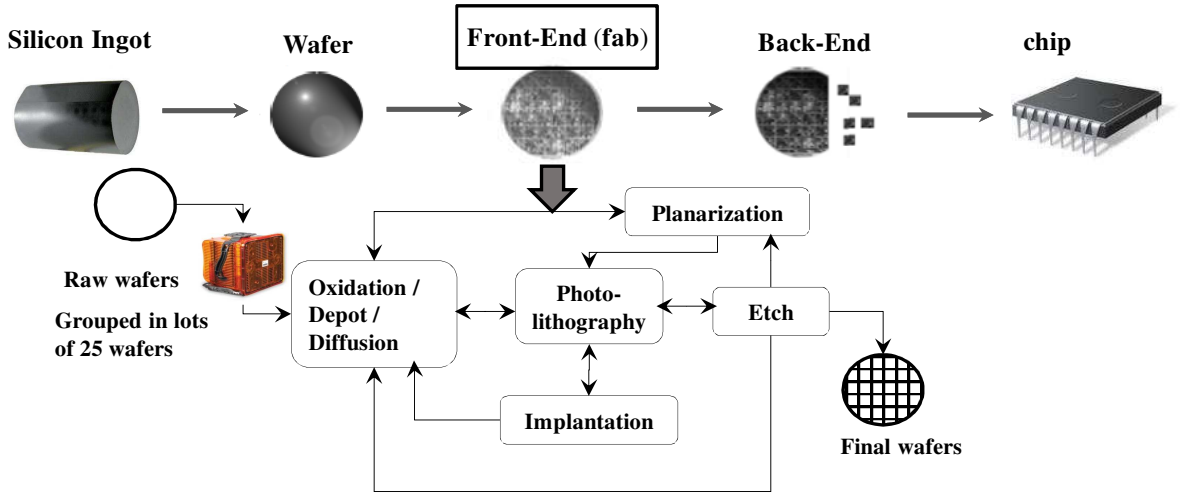


Figure 1.2: Operations in the manufacturing process of integrated circuits (adapted from Mönch et al. (2012))

The manufacturing process in the semiconductor industries begins with the preparation of the raw material. The raw material comes from the silicon ingot extracted from the sand. The silicon ingot is first purified before being cut using specific diameters and finally polished. This thin disk obtained is called a wafer on which several integrated circuits are produced, see figure 1.1. Note that this polished wafer is initially non-conductive. It will only be semiconductor when other substances and operations are applied to it. Integrated

circuits are known as a semiconductor chip. The manufacturing process of the wafer is done outside the manufacturing process of integrated circuits.

Once the wafer is ready to be used as a raw material, many operations are required to produce an integrated circuit. These operations are manufactured on different machines grouped in different work-centers. The main operations include:

- **Oxidation**, in this process, a thin layer composed of various materials is deposited on the wafer. This forms an oxide layer whose role is to protect the surface of the wafer against impurities.
- **Photolithography**, once the wafer is provided with a protective layer on its surface, the circuit design is transferred to the wafer. This task is accomplished by exposing the patterned mask to light. This mask also called reticle is an important auxiliary resource used during photolithography operations in addition to photolithography machines.
- **Etching**, after a photolithography operation, the etching process is used on the wafer to remove unnecessary materials in order to keep only the desired circuit patterns.
- **Deposition**, it consists of depositing different materials on the surface of the wafer. The operation can be applied to different stages of the manufacturing process of the integrated circuit. The additional material on the wafer can act as an insulating layer between the conductive layers or as a new layer which can be used for a new junction.
- **Chemical/mechanical planarization**, the topography of the wafer surface is changed each time processes such as etching, deposition or oxidation is used. This leads to an uneven surface. Chemical Mechanical Planarization (CMP) is used to flatten the surface of the wafer. This is done before adding each new layer. The objective is to reduce the differences in thickness and avoid the accumulation of an uneven topology over several layers.
- **Ion Implantation**, this operation consists in implanting ions and other impurities in the crystal structure of the semiconductor material. The goal is to modify its conductivity to allow the flow of electricity through silicon and make transistors.
- **Diffusion**, this operation consists of a series of atomic movements of the dopant and impurities in the crystal structure of the semiconductor material. Diffusion and ion implantation complement each other. The former can be used for a deep junction and the latter for a shallow junction.

In addition to the processes that ensure that the electronic elements are well connected, there are other important additional operations such as metrology and inspection. These processes are generally applied at critical points in the manufacturing process. The objective is essentially to ensure the quality of the integrated circuit produced. The manufacturing process of an integrated circuit ends at the back-end area, which can be geographically located at the same place or at a different area from the front end area. In the back-end, important operations are carried out before the product is sent to the end customers. These operations are called packaging and packaging testing. Some of the objectives of the back-end area are to test the inter-terminal connection and to provide protection to the integrated circuit against external factors. To learn more about semiconductor manufacturing processes, see Mönch et al. (2012) and May and Spanos (2006).

1.2.2 Management Decision Levels in Semiconductor Manufacturing

Management decisions in an industry are structured according to their scope and their impact on the industry objectives. They are divided into three main categories. (1) Long-term decisions that affect the entire industry belong to the highest level of management. (2) The intermediate level of management consists of decisions based on the use of the resources made available by the highest level of management decisions. Finally, (3) decisions at the lowest level, where production operations are carried out, affect daily operations. All these decisions, whatever the level at which they are situated, directly or indirectly concern the functions of management, direction, supply, planning, organization, staffing, production, control, etc. In semiconductor manufacturing, the decisions at these three levels are described below:

- **Strategic decisions**, generally define the overall strategy of the company. They are generally based on the markets to be covered, the decisions on the supply chains to integrate, the decisions to buy production capacity, the location of factories, etc. These decisions are made over several years. They guarantee the sustainability of the business and the competitiveness of the business in the market. In the semiconductor manufacturing industry, the design of different technologies, the choice of products and the combination of products to be manufactured, as well as the resources necessary to acquire in order to achieve the objectives of the company are studied at the strategic level.

- **Tactical decisions**, give a global vision in the medium term of what the company is capable of producing. After important strategic decisions that give the company the necessary physical resources and a long-term vision, tactical decisions focus on how to use these resources to meet the demand in the market. Tactical decisions span a horizon ranging from weeks to a year. In semiconductor manufacturing, tactical decisions are usually planning decisions. With the given customer demand and production capacity, tactical decisions determine the order release and the level of production to be achieved for a given period (usually a week or a month). Planning at the tactical level also relates to maintenance planning. Maintenance planning ensures that the health of the machine is maintained and avoids sudden stops of the machine during production, which could constitute a loss of capacity. Machine qualification decisions are also made at the tactical level (Johnzén et al. (2011) and Perraudat et al. (2019)). Qualifying a machine means certifying its ability to process an operation. The objective is to guarantee the quality and performance requirements, but also the flexibility of the production system.

- **Operational decisions** are concerned with the production of finished products that meet the specification of customers. These decisions are said to be short term because they are applied over a short horizon of a few hours up to a few days. Due to their short execution time, they are the most detailed of all decisions. In semiconductor manufacturing, scheduling decisions, i.e., determining the allocation of products to machines and the production sequence of products on each machine is one of the main decisions at the operational level. In addition to the scheduling decisions, the operational level is concerned with the machine requalification decisions, the Automated Material Handling System (AMHS) decisions and the measurement decisions, which

ensure the quality and yield of the production system. The operational level consists of optimizing the main Key Performance Indicators that drive the company in order to achieve its long-term objectives.

In order to ensure the consistency of the decisions taken at the three decision levels (strategic, tactical and operational), different approaches are proposed in the literature to integrate or simply to ensure the communication between these decision levels. Some studies on the integration of decision levels can be found in (Dauzère-Péres and Lasserre (2012) and Dauzère-Pérès and Lasserre (2002)). Other approaches that ensure communication between the tactical and operational decisions levels are discussed in Section 1.3.

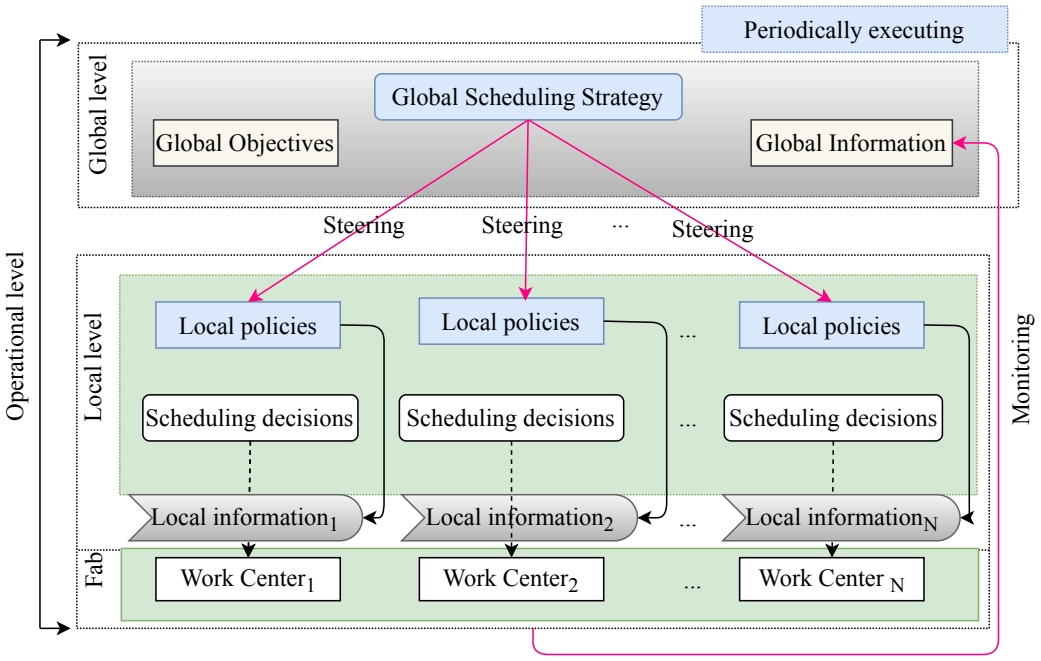


Figure 1.3: The two views of the operational decision level (adapted from Sadeghi (2017))

This thesis is based on management decisions at the operational level. Given the complexity of the manufacturing process of integrated circuits, the operational level is structured into two sub-levels, see Figure 1.3. (1) A global level, defining global objectives using global information of the factory such as the order release, resource capacity, Work-In-Process of the factory, etc. (2) A local level centered on scheduling decisions based on local information at the level of each work-center such as waiting times, processing times, machine queue lengths, etc. The global level determines the global management strategies for all work-centers, while the local level, at the scope of each work-center follows the global strategies as constraints to satisfy in order to optimize the KPIs of the whole plant.

1.3 Scientific Context: Literature Review

This section provides a review of the literature with a focus on three main axes which are the pillars of this thesis:

– Consistency between decision levels, section 1.3.1. This section reviews different approaches used in semiconductor manufacturing to ensure consistency between different decision levels. Different works are discussed on hierarchical approaches, iterative approaches, production targets management-based approaches and priorities-based approaches.

– Work-In-Process and cycle time management, section 1.3.2. This section presents different strategies for Work-In-Process and cycle time management used in the literature. These strategies aim to optimize KPIs such as cycle times, throughput and variability.

– Simulation in semiconductor manufacturing, section 1.3.3. This section reviews the literature on simulation in semiconductor manufacturing. In this thesis, our proposed approach is evaluated with a simulation model.

1.3.1 Consistency between Decisions Levels

As mentioned in Section 1.2.2, the management of production decisions is generally grouped according to the time horizon on which they are applied: Long-term (strategic), medium-term (tactical) and short-term (operational). This decomposition simplifies the decision process. Decisions made at a higher level become constraints to be satisfied or targets to be reached at lower levels. However, decisions at different levels are often made independently, which can lead to inconsistent or unfeasible decisions (Dauzère-Péres and Lasserre (2012), Dauzère-Pérès and Lasserre (2002)). Consistency in semiconductor manufacturing ensures that global decisions defined at the factory level are followed locally. It can be studied at different levels of decision. This section reviews the studies that deal with consistency between the tactical and operational decision levels, and those based only on the consistency of decisions at the operational decision level.

In the literature, hierarchical and iterative approaches are used to ensure the consistency between the tactical and operational decision levels. They create a kind of communication tunnel in which the two levels of decisions exchange information. While priority and production targets management are used to ensure consistency at the operational decisions level.

Hierarchical approaches

In hierarchical approaches, information is exchanged only once. These approaches use an upper layer model (tactical level) which determines daily or weekly targets and a lower layer model (operational level) which aims to reach these targets. Targets are used as input to the lower layer model after being sliced into a very short detailed plan of three or six hours. Consistency is then ensured by additional constraints in the lower layer model in order to coordinate short-term actions to achieve the production objectives provided by the higher level model. Hwang and Chang (2003) suggest a two-level hierarchical production-scheduling model. The two levels of the hierarchy consist of a midterm scheduler and a short-term scheduler. These two schedulers are aimed to achieve coordination between the fab-wide objectives and the local shop-floor operations and they are modeled as Integer Programming models. Liao et al. (1996), propose an Integer Programming model which optimizes the daily scheduling operations at work-centers in a semiconductor manufacturing system. With a given daily production target, the model breaks these daily production

targets into a production schedule in a time scale (of one to three hours) over a day which then serves as guideline to coordinate dispatching decisions.

In addition to Integer Programming models, other modeling approaches such as flow model and Discrete-Event Simulation are used to model the higher or lower level of the hierarchy. Tsakalis et al. (1997) address the problem of controlling re-entrant semiconductor fabrication lines by providing a two-level hierarchical modeling of production flow control. At the highest level, desired per-week yields are determined on the basis of economic factors. These target yields are converted to desire per shift yields. Then, a lower level controller ensures that the appropriate decisions are made at the lower level so that the desired per-shift yields are achieved. The high-level model is based on a flow model over a long planning/scheduling horizon. The lower level is modeled as a nonlinear, discrete-time, discrete variable and dynamic system. El Adl et al. (1996) propose a hierarchical model for a semiconductor manufacturing system. The higher decision level provides long term decisions (tactical decisions) and is supported by a linear flow model. It involves setting realistic objectives for the lower decision level. The latter is represented by a control mechanism which guarantees that these objectives are achieved.

The number of layers in the hierarchy is not limited to two layers as indicated in Vargas-Villamil et al. (2003). In their study, a three-level hierarchical approach for inventory control and production optimization of semiconductor manufacturing is discussed. Two upper layers are formed. The first layer provides aggregated global parameters for the second layer, which is an optimization model in charge of production planning. The latter, in turn, provides inputs for a distributed control policy implemented in a Discrete-Event Simulation at the lower level, with the goal of tracking the target determined by the optimization layer.

Iterative approaches

In iterative approaches, the information is shared between the higher level model and the lower level model in each iteration. The iterative process is stopped when the plan provided by the higher level model is feasible at lower level model.

In some approaches, decisions are based on the release quantities (starting quantities) determined in the higher level model. The release quantities are then evaluated in a lower level model often represented by a simulation model. In most cases, the higher level model is formulated as an Integer Programming model or a Linear Programming model. Hung and Leachman (1996), propose an automated production planning model for semiconductor manufacturing system. An iterative Linear Programming model is coupled to a Discrete-Event simulation model. The optimization model provides release quantities schedules to the simulation model. Flow time statistics are collected in each iteration and are used to improve the optimization model for an updated plan. The process continues until the plan provided by the optimization model is feasible in simulation. Kim et al. (2001) suggest almost the same modeling as in Hung and Leachman (1996) for planning release quantities. The difference lies on the additional data used to improve the updated plan in the optimization model.

In other approaches, decisions are based on quantities to produce for each product in a defined period of time at the higher level model. The feasibility of the resulting production plan is then assessed at the lower level by a simulation model in an iterative scheme. Bang and Kim (2010) address a production planning and scheduling problem in a semiconductor wafer manufacturing facility. The production plan is determined using an Integer Linear Programming model at the aggregate level. This plan is then evaluated in an iterative

scheme by a Discrete-Event Simulation model in which a priority-based rule is implemented. The objective is to guarantee feasible and relevant production plans. Kim and Lee (2016) investigate an iterative approach, which integrates production planning and scheduling in a flexible manufacturing system. To ensure the synchronization of production planning and scheduling decisions, the manufacturing lead-time, the number of setup events, and the available Work-In-Process level are updated via an iterative simulation and optimization approach.

Our approach differs from classical hierarchical and iterative approaches in the literature, which deal with the integration and communication between tactical and operational decision levels. Higher level models are usually based on demand and resource capacity while, the approach proposed in this manuscript deals only with the operational level. Instead of demand as global information, our approach uses the lot release quantities, cycle time targets, resource capacity, Work-In-Process, etc.

Priority management approaches

Priority management approaches go beyond the communication framework between the tactical decision level and operational decision level. In these approaches, the operational decision level is structured into two sub-levels: A global level based on global information (factory level) and a local level based on local information (work-center level). Bureau, Dauzère-Pérès, Yugma, Vermariën and Maria (2007) develop a Work-In-Process framework to meet the need for consistency at the operational decision level. In the same spirit, in Bureau, Dauzère-Pérès, Yugma and Vermarien (2007), an approach for simulating consistent global and local scheduling decisions is developed. The main idea is to speed up or to slow down flows according to priorities. Priorities are used at the global level as a global management strategy and are dynamically updated at the local level. The same strategy has been described in Vialletelle and France (2006). For the same purpose of using priorities as main parameters for global management, Sadeghi et al. (2016) address a flexible multi-method simulation model for semiconductor manufacturing to control the Work-In-Process and to satisfy time constraints, i.e., to ensure that the maximum time between two operations (consecutive or not) is respected.

Priorities can embed several elements such as customer emergency or customer preferences. They are also discussed at the highest level. Thus, managing them at the local level does not seem to be the best way to ensure consistency at the operational decision level. Changing too often the priorities of the Work-In-Process at the local level by speeding up late products or by slowing down early products may cause the priorities to lose their relevance defined at the highest level.

The proposed approach in this thesis enforces consistency at the operational decision level by switching from setting priorities to setting production targets. Production targets are quantities to complete for each product at each operation in each period on a scheduling horizon. An Adapted rules is required to ensure that These production targets are followed at the local level.

Production targets management-based approaches

Small number of studies in the literature use optimization methods with production targets at the operational level. For a noticeable exception, Govind et al. (2008) propose a study on an integrated operation management approach which includes a module to determine production targets based on a Linear Programming model. However, the approach is

only focused on one work-center (photolithography area). The optimization model was not provided, the scheduling horizon is about one week and the decisions in work-centers are rescheduled in less than five minutes.

Studies in the literature that use production targets at the operational level are essentially empirical or based on numerical calculations. Wu et al. (1998) study a daily production target setting system for wafer fabrication. The determination of production targets uses an algorithm based on numerical calculation. In the same spirit, in Kao and Chang (2018), no optimization model is proposed, but numerical calculations provide the production targets. Furthermore, this approach is short-sighted because it is at the machine level instead of being at the fab level. To correct this short-sighted effect, they approximate the variations induced by production targets with a Bernoulli trial model. These variations are included in the computation of production targets for the correction iteration.

Our global scheduling approach differs from the previous studies in the literature which use numerical computations to determine production targets. Our approach uses optimization models to determine production targets, which broaden the scope of the parameters to be used and offer the possibility to include several Work-In-Process management strategies in single-objective optimization models or in multi-objective optimization models. These global scheduling strategies are novel policies used to optimize different objectives. They innovate in the use of novel Work-In-Process management techniques such as the use of balancing coefficients to optimize output variability on cycle times and throughput, see Chapter 3, or in the use of temporal tracing of Work-In-Process to control cycle times, see Chapter 4. Finally, our global scheduling approach is designed and evaluated using a generic multi-method simulation model, where the exchange/communication between the simulation and the global scheduling model is clearly defined.

1.3.2 Management of Work-In-Process and Cycle Times in Semiconductor Manufacturing

This section presents a review of previous works in the literature related to the global scheduling strategies proposed in this thesis. These strategies are mainly based on the management of the Work-In-Process in order to optimize and control Key Performance Indicators (KPIs) such as cycle time, throughput and output variability on cycle times and throughput. Previous studies on Work-In-Process management in semiconductor manufacturing are presented followed by the review of the literature on cycle time management in semiconductor manufacturing.

Work-In-Process management in semiconductor manufacturing

In several manufacturing systems, a Kanban control system or its simplified version, the constant Work-In-Process (CONWIP), among the methods in practice that populate the strategies for Work-In-Process balancing (Spearman et al. (1990)). The principle of the Kanban production control system lies on the limitation of the Work-In-Process in the production line. It uses cards to control the number of products in the factory. Each product released in the factory seizes a card. If all cards are taken, a newly entering product has to wait until a previous product gets out of the factory and releases its card. The CONWIP works in the same manner, but controls the line using a single set of cards (Kalisch et al. (2008)).

The Work-In-Process (WIP) corresponds to the products already in the fab but not yet completed. Balancing the Work-In-Process is of great importance because it allows a good capacity utilization by ensuring that products are properly distributed in the fab. It also ensures that all products steadily move forward to the completion of their operations. An overview of advanced scheduling and dispatching policies for Work-In-Process management for a Make-to-Order wafer fabrication is provided in Sturm et al. (1999).

It is difficult to improve KPIs such as cycle time, throughput, and on-time delivery without a thorough management of the Work-In-Process. Work-In-Process balancing control (ensuring that Work-In-Process is properly distributed throughout the whole manufacturing system) is considered as an efficient method to improve KPIs (Lee and Lee (2003)). In semiconductor manufacturing, strategies to avoid unbalanced Work-In-Process have been studied in different ways:

– Work-In-Process balancing and control strategies based on the operation view point see Dabbas and Fowler (2003); Li et al. (1996); Leachman et al. (2002); Fordyce et al. (1992); Bureau, Dauzère-Pérès, Yugma, Vermariën and Maria (2007). Priorities and scheduling policies are used to balance Work-In-Process on the different operations of products. In Bureau, Dauzère-Pérès, Yugma, Vermariën and Maria (2007), different blocks of operations (sub-sequences of operations) are created. Work-In-Process targets are defined for each block and the balancing is achieved by minimizing the deviation between the current Work-In-Process and the defined Work-In-Process target in each block. In Leachman et al. (2002), different methodologies and algorithms are proposed for Short Cycle Time and Low Inventory Management (SLIM). A continuous-time target output schedule or continuous-time target cycle times are translated into target profiles of Work-In-Process through the sequence of operations for each product. Instead of individual lots, operations are considered as the main scheduling object in SLIM. Fordyce et al. (1992) propose a daily output planning by using Work-In-Process targets for each operation of a product. The goal is to provide quantities of lots that should be processed in each operation at a given period in order to meet immediate demands or to anticipate future demands.

– Work-In-Process balancing strategies based on the work-center view point, see Zhou and Rose (2010); Chung and Jang (2009); Lee and Lee (2003); Miyashita et al. (2004); Chien and Hu (2006). Work-In-Process targets are defined for each work-center, generally bottleneck work-centers. The balancing is achieved by minimizing the deviation between the current Work-In-Process and the defined Work-In-Process target.

Using both the operation and work-center view points, Work-In-Process control is either performed by using targets and/or priorities. Chung and Jang (2009) study a Work-In-Process balancing procedure using production targets for throughput maximization in semiconductor manufacturing. The balancing is achieved by sending detailed target production quantities to bottleneck work-centers. These targets are transformed from production quantities sent from production planning. The same control is implemented in Lee and Lee (2003). Besides production quantities, targets are also based on the Work-In-Process or on the cycle time. The idea is to divide every route (sequence of operations for one product) in blocks which correspond to a logical separation that allows intermediate controls on products during manufacturing. Work-In-Process targets or cycle time targets are estimated for each block, see Lee et al. (2008); Bureau, Dauzère-Pérès, Yugma, Vermariën and Maria (2007). The objective is then to ensure that the difference between the current Work-In-

Process (resp. current cycle time) and the Work-In-Process target (resp. cycle time target) is minimized for each block.

The critical point with Work-In-Process balancing strategies that use Work-In-Process targets/levels is the determination of these Work-In-Process targets/levels. Various strategies have been used in the literature in semiconductor manufacturing to determine Work-In-Process targets/levels. Some of these strategies are based on simulation, see Potti et al. (1994); Miyashita et al. (2004), artificial neural network and/or queuing network, see Lin and Lee (2001); Liu et al. (2006); Lin et al. (2009). Since the exact estimation of the Work-In-Process target/level for each operation or each block is always difficult to perform Lee et al. (2002), dispatching rules can be used to balance the Work-In-Process in the factory. Zhou and Rose (2011) propose a new composite of dispatching rules which combines the Operation Due Date rule, the Shortest Processing Time rule and the Least Work at Next Queue rule (LWNQ) to consider several objectives simultaneously. The LWNQ is a simple workload control rule which looks at WIP balance with the viewpoint of machines. Among the waiting lots, it provides the highest priority to the lot that is to be processed by the next machine with least remaining production hours. Wang et al. (2007) propose a compound priority dispatching rule that takes into account both Work-In-Process management and wafer start lot control. It is shown in their study that the compound priority dispatching rule can reduce the mean total queue time by 50% and increase the throughput rate by 20% compared with the First-In-First-Out (FIFO) and Shortest Remaining Processing Time (SRPT) dispatching rules. In the same spirit, Zhou and Rose (2019) propose a global fab dispatching scheme, which switches from the use of Work-In-Process targets to the use of a workload indicator, whose role is to measure the pull request of work-centers. They conclude that significant improvement is made when dispatching rules based on a workload indicator are used instead of dispatching rules based on Work-In-Process targets.

Besides these methods which use a route subdivision in blocks, another approach based on a so-called Work-In-Process Control Table is discussed in Zhou and Rose (2010). In this approach, each upstream work-center maintains a Work-In-Process control table, which contains the current Work-In-Process information of the downstream work-centers such as the Work-In-Process target, the current Work-In-Process and the difference between the Work-In-Process target and the current Work-In-Process. This Work-In-Process control table is regularly updated and allows the upstream work-centers to optimally supply lots to the downstream work-centers. Those targets are estimated either based on historical data or by simulation.

Work-In-Process balancing strategies can also use priorities to speed up or slow down lots in blocks of operations to smooth the workload in different blocks (Bureau, Dauzère-Pérès, Yugma, Vermariën and Maria (2007)). Depending on the due date and workload information, a priority matrix table can also be used to assign lot priorities to manage the Work-In-Process. The objective is to balance the overall workload of the manufacturing system. Zhou and Rose (2012) provide a priority matrix to control the flow of lots in the system and a Work-In-Process calibration method whose purpose is to recover the Work-In-Process balance due to an event such as an unpredictable machine failure.

The global scheduling approach proposed in this thesis for Work-In-Process balancing differs from strategies in the literature which use production targets, see Chung and Jang (2009) for instance. Our approach does not just focus on **bottleneck work-centers**, but also takes into account the **interaction between work-centers**, thus preventing the short-sightedness of independent scheduling decisions. Instead of imposing Work-In-Process targets/levels at the local level, our approach provides production targets for each product to

be completed at each operation and at each period over a scheduling horizon. Production targets are determined on the basis of global information (fab level) such as lots release quantities, resource capacity, Work-In-Process in the factory, cycle time targets, etc.

Cycle time management in semiconductor manufacturing

Cycle times include processing times, as well as transportation times and the time lots spend waiting in queues. The cycle time is one of the important Key Performance Indicators (KPIs) in semiconductor manufacturing as it impacts several other metrics and KPIs such as throughput, yield and on-time delivery. Controlling cycle times reduces wafers risk contamination, yield loss and the inventory that should be maintained (Lu et al. (1994)).

In the semiconductor manufacturing literature, several studies focus on the understanding of cycle time and the way it can be improved. Bonal et al. (2001) provide a statistical method for cycle time management. The objective of the study is to ensure a quick detection of changes on operation processes that can affect the stability of the cycle time. Pierce and Yost (1995) study cycle time metrics for wafer fabrication in a research and development environment. In Sivakumar (2000), a discrete event simulation model for a semiconductor back-end manufacturing system is proposed to analyze the effect of controllable input parameters on cycle time distribution and other output variables. In the same spirit, a simulation model is provided in Qi et al. (2002) to study the effect of some variables such as job arrival distribution, batch size, downtime pattern and input control on mean cycle time and average Work-In-Process. Chien et al. (2005) study how a learning curve approach can be used to determine empirical rules for cycle time improvement. Strategies based on the analysis of different problems related to cycle time by using data from the manufacturing execution system are studied in Robinson and Chance (2000) and Ab Rahim et al. (2012). Kramer (1989) studies the improvement of cycle time with a focus on the breaking of the product cycle time into elements common to specific tools. The paper argues that the improvement of the cycle time of each element leads to the improvement of the overall cycle time. For more studies on the understanding of cycle time and the way it can be improved, see Nemoto et al. (2000), Brown et al. (1999) and Domaschke et al. (1998).

The relationship between cycle time and other KPIs or parameters has also been investigated. A study based on the relationship between cycle time and yield in semiconductor wafer fabrication can be found in Wein (1992). Tirkel et al. (2009) investigate the relationship between cycle time and yield as affected by in-line metrology inspections of production lots. In Fronckowiak et al. (1996), a discrete event simulation model is used to study the impact of job priorities on cycle time. This study shows the significant impact of hot lots on cycle times. Leachman and Ding (2010) provide analytic formulas to quantify the revenue losses due to excursions not detected until end-of-line testing as a function of manufacturing cycle times, excursion probabilities and kill rates.

The cycle time main challenges in semiconductor manufacturing are still based on how it can be predicted/estimated, controlled and reduced:

- Cycle time prevision and estimation are studied with the purpose to control and plan customer orders in tactical decisions, and further to manage some production factors such as the level of input, the level of Work-in-Process in order to improve KPIs such as on-time delivery, throughput and yield. Different approaches are used for cycle time prediction and estimation: (1) Big data analytic (Wang and Zhang (2016)), (2) Statistical methods, which include techniques such as probability distribution-based

method and regression based method (Tai et al. (2012)), (3) Artificial intelligent techniques based on domain knowledge, machine learning and data mining (Tirkel (2011), Hassoun (2013)), Neural Networks (Chien et al. (2012)), and selective Bayesian classifier based on a selection of minimal, most discriminative key-factor set for cycle time prediction (Meidan et al. (2011)), (4) Simulation for cycle time prediction (Chung and Huang (2002)), (5) Queueing model adapted for semiconductor manufacturing (Akhavan-Tabatabaei et al. (2009)).

– Cycle time reduction refers to the strategy of decreasing the time a product spends in the factory from its release to its last operation. Shorter cycle times drive a better on-time delivery, help to decrease Work-In-Process and ensure good production quality (higher yield), Meyersdorf and Yang (1997). Several strategies have been studied, essentially based on the management of factors that influence the cycle time. Variability is considered as one of the cycle time killers, see (Robinson et al. (2002)). In Majorana and Iuliano (1997), a study on the management of variability is provided for cycle time improvement. Chen (2013) provides a three-step procedure for cycle time reduction: Identification of controllable factors that influence the product cycle time, investigation of the relationship between the controllable factors and product cycle time and finally, based on this relationship, actions should be planned to shorten the product cycle time.
  Other factors that influence cycle times have been used as a lever for cycle time reduction such as batch size (Babbs and Gaskins (2007)), lot size (Zarifoglu et al. (2012), Eberts et al. (2015), Wang and Wang (2007)), Work-In-process management (Chien and Hu (2006)), queue time management (Sada et al. (2001)) and priority management (Schmidt (2007)). Equipment management, essentially the study on preventive maintenance segregation, is proposed in Rozen and Byrne (2016) with the goal to determine the optimum preventive maintenance policy that results in reduced fabrication cycle times. Leachman et al. (2002) provide a set of methodologies and scheduling applications for managing cycle times in semiconductor manufacturing called SLIM (Short cycle time and Low Inventory Manufacturing).
  The minimization of mean, variance and standard deviation of cycle times is also widely studied. Scheduling policies are one of the levers used for mean and variance cycle time reduction in semiconductor manufacturing (Mittler and Schoemig (1999), Lu et al. (1994)). For more information about the minimization of mean and variance of cycle times, see Yoon and Lee (2000), Lu et al. (1993) and Mittler et al. (1995).

Due to the complexity of semiconductor manufacturing, some of the research in semiconductor manufacturing focus on the reduction of cycle times based on the activity of some machines. This is the case for the studies proposed in Swe et al. (2006) for cycle time reduction on cluster tools and in Brown et al. (1998) for the test area. Other works focus on a unique work-center of the factory. For illustration, see the studies proposed in Akcalt et al. (2001) and van der Eerden et al. (2006) for cycle time reduction in the photolithography area or Butterbaugh (2004) in batch cleaning.

In our global scheduling approach for cycle time control (see Chapter 4), cycle times are managed by controlling the competition of products on shared resources using the production targets determined by a global scheduling optimization model. In previous approaches in the literature, the release dates of products were not considered in the control of the Work-In-Process. Our global scheduling approach innovates by using both the release dates and the temporal tracing of the Work-In-Process in the global scheduling model. Temporally

tracing the Work-In-Process is critical to differentiate quantities of the same product and at the same processing stage, but released at different times in the factory.

1.3.3 Simulation in Semiconductor Manufacturing

Production systems transform input materials into final products to be delivered to customers. Customers can be enterprises (Business-to-Business) or final consumers (Business-to-Consumer). The transformation process usually follows a sequence of operations from the inputs materials to the final products.

Modeling and simulation are critical for complex systems such as semiconductor manufacturing systems. The poor understanding of the key dependencies, weaknesses, and bottlenecks in such complex systems can lead to poor decision-making. To tackle these issues, simulation is one of the most powerful tools available to decision makers responsible for the design and operations of complex processes and systems. It makes possible the study, analysis, and evaluation of different situations and behaviors of a complex system which would not be otherwise possible to apprehend (Shannon (1998)).

Even though the need of modeling and simulation becomes extremely important, challenges still need to be addressed, such as the reducing of problem solving cycles, the development of real-time simulation-based problem-solving capability and the need for true plug-and-play interoperability of simulations and supporting software. For more information about the key challenges in modeling and simulation of a complex manufacturing system, the readers are invited to check Fowler and Rose (2004).

The problem of standardization is also another challenge in simulation. Two simulation experts might create quite different simulation models of the same production system, even when using the same language. One of the challenges in simulation is to provide a standard modeling and a framework for implementation. This problem is discussed in Ehm et al. (2009). A general review on simulation for manufacturing systems can be found in Negahban and Smith (2014).

Modeling usually comes before simulation, to obtain an abstraction of the system or an abstraction of the components of the system which should be simulated. The output of the modeling procedure is the models which can be mathematical models, physical, or logical representations of a system, entities, phenomenons, or processes. Simulation represents the system process function which is under study to predict a future state/behavior of the system.

As simulation is used in this thesis to evaluate our approach, this section outlines an overview of the literature on simulation in semiconductor manufacturing. The goal is to highlight the existing approaches as well as research gaps. We propose to classify the literature using three criteria: *The Scope of the Simulation*, *The Simulation Methods*, and *The Simulation Problem Type*.

1.3.3.1 The Scope of the Simulation

The first and most complex production stage in semiconductor manufacturing is the front-end, where a series of process steps (operations) are processed on wafers. Simulation in semiconductor manufacturing is studied using different views. A simulation of the whole front-end fabrication process is studied in Kiba et al. (2009), Arisha and Young (2005), Collins et al. (2001), Fronckowiak et al. (1996) and Kuhl and Laubisch (2004).

To study a particular problem in manufacturing systems, researchers can sometimes use a simplified simulation model which represents essential objects in order to decrease the

complexity and the model development time. El-Khouly et al. (2009) discuss a simplified simulation model with six processing steps and five machines in three work-centers. The purpose is to evaluate the effect of different dispatching rules and lot release policies on some performance measurements such as the mean and standard deviation of cycle times.

Due to the complexity of the front-end stage, simulation studies are also conducted on a particular work-center. Akçali et al. (2000) propose a simulation model of a wafer fabrication facility in order to examine the effects of different loading and dispatching policies for diffusion operations. In Mack (2005), the most popular and useful examples of lithography simulators in a manufacturing environment are reviewed.

For the matter of re-usability and complexity simplification, generic simulation models now have some attention in the literature on semiconductor manufacturing. Papers which discuss generic simulation models can be found in Sadeghi et al. (2016), Arisha et al. (2004), Kim et al. (2009) and Mackulak et al. (1998).

One of the challenges in semiconductor manufacturing is the modeling of complex process tools, such as cluster tools, that need to be simulated. A cluster tool is an integrated, environmentally isolated, wafer-manufacturing system consisting of processing chambers, internal robots to transport wafers, and load locks where the wafer-to-cassette exchange takes place. LeBaron and Pool (1994) address the simulation of cluster tools in order to accurately predict their performance. Simulation in the back-end stage in semiconductor manufacturing is not widely studied in the literature. Some discussion on simulation for semiconductor packaging, testing, and scheduling in back-end can be found in:

- Wang et al. (2017), they propose a simulation model for packaging facility in semiconductor manufacturing. Several strategies are discussed to enhance sustainability of a factory. Instead of observing the continuous application of a factory using a simulation model in the long term, this study identified short-term evidence to estimate the sustainability of a factory simulation model. They conclude that the sustainability of a factory simulation model can only be confirmed if the model is still applicable several years after it is built.

- Lin and Chen (2015), they propose a simulation optimization approach for a hybrid flow-shop scheduling problem in a real-world semiconductor back-end assembly facility. Their approach includes a simulation model for performance evaluation and an optimization strategy with application of a genetic algorithm. They argue that their approach aids in assigning orders optimally to the proper production line and machine types while achieving minimal flow time.

- Werner et al. (2006), they suggest a simulation-based scheduling system for a semiconductor back-end facility. The goal is to develop a Discrete Event Simulation-based approach for the complete back-end, which is suitable for the case of changing bottlenecks and different line scenarios. The study focuses on optimizing the process flow and calculating the exact release dates for lots.

1.3.3.2 The Simulation Methods

To our knowledge, in the literature, three simulation methods are discussed:

- System Dynamics (SD) which is a method for studying dynamic systems. The approach provides an aggregate level of the systems by emphasizing feedback mechanisms and their endogenous nature.

– Discrete Event (DE) modeling in which the main modeling idea is to consider the system as a sequence of operations being performed across entities. The notion of queue line is very well modeled with this method.

– Agent-Based (AB) modeling is the more recent modeling method suitable for modeling the individual behavior of objects of a system and their interactions.

For more details about the three simulation methods, see Borshchev (2013) and Barbosa and Azevedo (2017).

Most of the papers in semiconductor manufacturing discuss simulation models with Discrete Event simulation. It seems that only few papers combining the Discrete Event and the Agent-Based simulation can be found in the literature on semiconductor manufacturing, see Sadeghi et al. (2016) for a noticeable exception.

In general, multi-method simulation models combining different simulation modeling methods are getting more attention from researchers. The readers are invited to see Barbosa and Azevedo (2017).

1.3.3.3 Simulation Problem Type

Simulation in semiconductor manufacturing is studied for different purposes. The most common encountered problems in the literature are related to the operation control decisions, performance of Automated Material Handling Systems, evaluation of production planning, tool performance (LeBaron and Pool (1994)) and strategic decisions related to the factory (Shikalgar et al. (2002)).

Operation control decisions

Various operation controls are executed in a wafer fabrication facility. Simulation models help to assess the effect of different dispatching policies on some key performance indicators such as cycle time, number of wafers produced and on-time delivery (Akçali et al. (2000), El-Khouly et al. (2009), Freitag and Hildebrandt (2016), Kuhl and Laubisch (2004)). Dispatching rules and rework strategies are considered in the set of major operational decisions that affect fab productivity. In several papers, these issues are independently studied. Kuhl and Laubisch (2004) showed that the interrelationship between dispatching rules and rework strategies has a significant effect on the productivity of the Fab.

Other examples of operations control decisions analyzed using simulation in semiconductor manufacturing include Work-In-Progress management (Collins et al. (2001) and Kohn et al. (2009)), lot releases, mask scheduling and batch scheduling (Kim et al. (1998)), production scheduling (Jeong et al. (2006)), consistency between global flow decisions (Fab level) and local flow decisions (work-center level) (Sadeghi et al. (2016)), scheduling evaluation in semiconductor back-end manufacturing (Lin and Chen (2015)), effect of job priorities on cycle times (Fronckowiak et al. (1996)) and effect of the mix of products used on cycle times (Chang (2016)). A discussion of general simulation applications in semiconductor manufacturing can be found in Koo et al. (2016).

Performance of Automated Material Handling Systems (AMHS)

Managing Automated Material Handling Systems (AMHS) is very difficult to study without simulation modeling. These problems include the minimization of the average lot-delivery time, the changes when adding or removing stations, the management of incidents, i.e.,

the breakdown of vehicles on the rail, which can block the mobility of other vehicles and consequently affect the fab productivity, etc.

Most AMHS simulation models assume that the logic of the production processes is given and the AMHS management decisions are made based on this assumption. In Kong (2007), a two-step simulation method for an Automated Material Handling System in semiconductor manufacturing is provided combining a production simulation model and an AMHS simulation model. The objective of the production simulation is to predict the throughput and estimate the capability of the AMHS. After applying the production simulation model, the AMHS simulation model is used to estimate the number of vehicles required and predict delivery times.

In general, machine dispatching rules and vehicle dispatching rules are studied independently. In Christopher et al. (2005), it is shown how the interaction of both machine dispatching rules and vehicle dispatching rules have a significant effect on the fab productivity. For more information on AMHS challenges, see Cardarelli and Pelagagge (1995), Jimenez et al. (2002), Ndiaye et al. (2016a), Ndiaye et al. (2016b) and Ben-Salem et al. (2016).

Evaluation of production planning

In semiconductor manufacturing, hybrid simulation-analytic methods are popular to evaluate production planning approaches, which after corresponds to optimization models used at a higher level (medium/tactical decision levels) and evaluated by a simulation model, which acts as the shop floor. After the simulation is completed, statistics are collected that can be used to improve the optimization model for updated planning. The process continues when the plan provided by the optimization model is feasible in simulation. For a review of simulation-optimization methods in semiconductor manufacturing, see Ghasemi et al. (2018).

Here is a non-exhaustive list of papers which investigate hybrid simulation-analytic methods for production planning evaluation: Byrne and Bakir (1999), Irdem et al. (2010), Bang and Kim (2010), Liu et al. (2011), Hung and Leachman (1996), Hung and Leachman (1996)). A general taxonomy/discussion on hybrid simulation-analytic methods can be found in (Figueira and Almada-Lobo (2014), Shanthikumar and Sargent (1983) and Hsieh (2002).

1.3.3.4 Gap Analysis

Previous sections reviewed articles related to the application of simulation in semiconductor manufacturing. We observe that several articles do not provide the structure of simulation model and the conceptual model. We found two noticeable exceptions. The first one is Mueller et al. (2007). In this study, the authors discuss the automatic generation of a simulation model based on an object-oriented Petri net data structure. The second exception is Lin and Long (2011), where the development of a multi-agent distributed platform for semiconductor manufacturing is addressed.

Only few papers discuss the validation and verification of simulation models, see Nayani and Mollaghasemi (1998), Kong (2007) and Chance et al. (1996). Various verification and validation methods of simulation models can be found in Sargent (2013). Based on the future challenges of simulation stated in Ehm et al. (2009), few papers discuss the matter of simulation framework, see Mönch et al. (2003) for a noticeable exception. In their paper, the authors provide a simulation framework for the performance assessment of shop-floor control systems.

For the implementation of simulation models, a dedicated software is likely to be preferred than a programming language (C, C++, Java, etc.). Popular simulation software programs

include AutoSched AP, ARENA, AnyLogic and AutoMod. For additional software programs used for implementing simulation models, see Shannon (1998).

The simulation model used in this thesis is based on the operation control decisions. It is a generic multi-method model, which combines Discrete Event (DE) and Agent-Based (AB) simulation initially developed in Sadeghi et al. (2016) using the AnyLogic software and improved in this thesis. A conceptual model is provided which allows the model to be verified. The validation of the simulation model is done on industrial data. In addition, an innovation is brought in this thesis compared with previous studies on simulation in the literature, where the interface of exchange/communication has not yet been well defined when simulation is used with mathematical optimization models. In this thesis, the exchange/communication interface is provided and its functioning is clearly defined.

1.4 Motivation and Main Objectives of the Thesis

In a wafer manufacturing plant, different products are produced on different machines which are grouped in different work-centers (machines with the same capabilities). Each work-center includes specific process characteristics such as batch processing, parallel processing, and auxiliary resources. These features increase the complexity of scheduling decisions. In addition to the re-entrant flow characteristic of semiconductor manufacturing, scheduling decisions become very difficult to apply for the entire factory. To cope with this complexity, the commonly used approaches for scheduling decisions in work-centers are as follows:

1. Real-time scheduling using dispatching rules, i.e., every time a resource is available, a decision based on certain rules is made to process the next product. A review on dispatching rules can be found in Varadarajan and Sarin (2006) and Sarin et al. (2011).

2. Optimized scheduling algorithms dedicated to a work-center, for instance, scheduling on parallel machines with auxiliary resources in the photolithography work-center, see Bitar et al. (2016) or on batch machines in the diffusion work-center (Yugma et al. (2012), Jung et al. (2014) and Knopp et al. (2017)).

A general literature survey on scheduling in semiconductor manufacturing can be found in Mönch et al. (2011). The main disadvantage of these approaches is that they are shortsighted. Independent scheduling decisions in each work-center are limited by the information available within the perimeter of the work-center. Work-centers interact when products move from one work-center to another, but this interaction is not considered in individual decisions of each work-center. For example, an upstream work-center can send quantities of a given product to a downstream work-center in a short period of time, which has a limited number of machines qualified to process this product. With a global vision of the system, an unbalanced flow can be observed which can deteriorate global key performance indicators even if the decisions taken locally in the work-center are optimized.

Another motivation of this thesis is that strategies based on the definition of production targets already exist in semiconductor manufacturing plants. For the whole factory (fab), production targets are determined not by relying on optimization methods, but on the basis of the experience of managers as shown in Figure 1.4 or using numerical and empirical calculations, see Kao and Chang (2018), Wu et al. (1998) and Govind and Fronckowiak (2003). In Kao et al. (2014) a study comparing different dispatching rules and an approach based on production targets shows that the approach based on production targets outperforms one

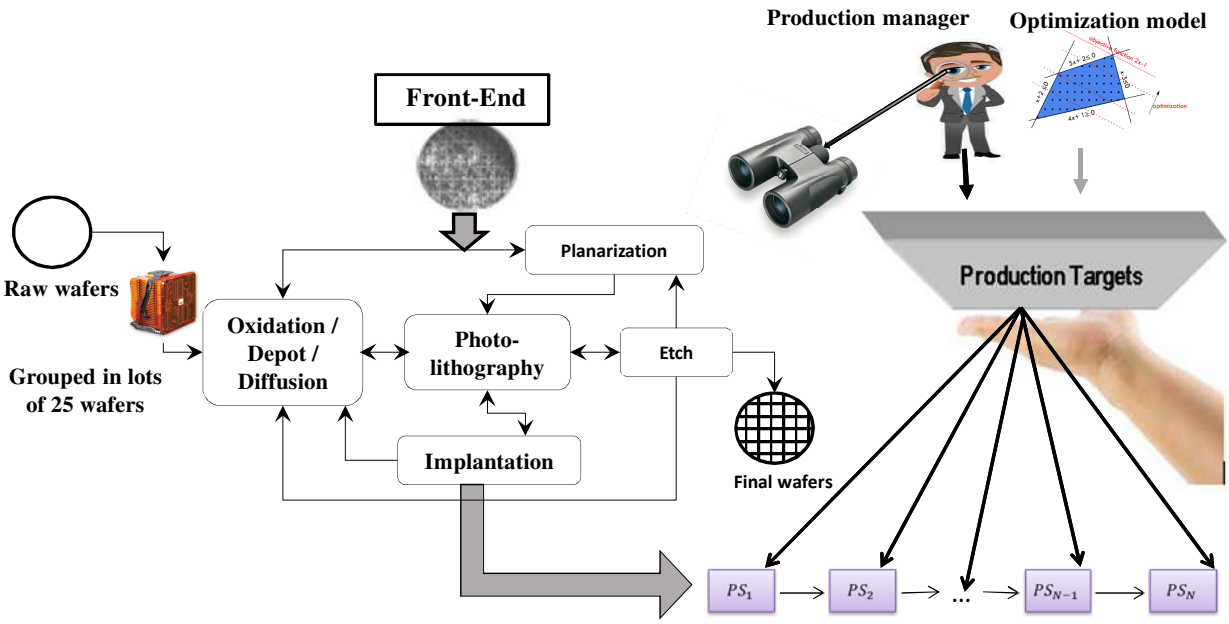


Figure 1.4: Optimization method to determine production targets

that only uses dispatching rules in terms of line balancing and cycle time performance with both high and low variability. In addition, it is shown in Chang (1999) that a right setting of targets leads to an increase of more than 20% of daily moves (number of wafers leaving an operation) and a decrease of 8% of the Work-In-Process. Although the results published in these previous studies have shown that the use of production targets at the work-center level allows a good management of Work-In-Process and contributes to improve KPIs, they are still shortsighted and empirical. They do not consider the interaction between work-centers and production targets are not determined using optimization methods.

The aim of this thesis is to propose a global scheduling approach and to validate it on industrial data by simulation. The main objective of the approach is to determine the right production targets using optimization methods. The proposed global scheduling approach widens the scope of the parameters used and includes several strategies (global scheduling strategies) defined according to the objectives to be optimized. These strategies are implemented as mathematical optimization models (global scheduling models) and they consider the interaction that exists between work-centers using global information such as lot release quantities, cycle time targets of products, resource capacities and the Work-In-Process of the factory.

Our global scheduling approach is evaluated using a generic multi-method data driven simulation model, initially developed in Sadeghi et al. (2016) and extended and improved in this thesis. The global scheduling models are called regularly in the simulation, which represents the factory. In a rolling horizon scheme, the current state of the simulation is collected to feed the global scheduling models. The latter determine production targets, which are used in the simulation as guidelines to local scheduling decisions.

1.5 Conclusions

This chapter presented the industrial and scientific context and the main motivation of this thesis. We described the main parts of semiconductor manufacturing systems as well as the main manufacturing processes. The manufacturing complexity of integrated circuits is presented as well as some main operations. Next, we briefly introduced the different levels of decision making in semiconductor manufacturing. The operational level is at the heart of this thesis, in particular global scheduling decisions in semiconductor manufacturing. An operational decision level structure based on two views was provided. It follows the same structure of two management levels at operational level as in Bureau, Dauzère-Pérès, Yugma and Vermarien (2007). In this structure, the top level view (factory level) is used as the steering mechanism for the bottom level (work-center level). However, instead of setting priorities as global strategies as in Bureau, Dauzère-Pérès, Yugma and Vermarien (2007), our approach sets production targets. Related works in the literature are reviewed and finally, the motivation and the main objectives of the thesis were provided.

Chapter 2

Global Scheduling Approach

2.1 Introduction

The global scheduling approach which is proposed in this thesis adopts two views of the operational decision level as in Bureau, Dauzère-Pérès, Yugma and Vermarien (2007): The global level (factory level) and the local level (work-center level). The global level uses the global information (Work-In-Process in the whole fab, lot releases, cycle time targets, resource capacity, etc.), while the local level uses local information (waiting times of lots, processing times, lots currently in queues, etc.). The global level aims to determine production targets which are regularly updated in a rolling horizon and should be followed at the work-center level. Different strategies can be implemented in the global scheduling approach depending on criteria to be optimized such as cycle times and throughput. In this thesis, these strategies are based on Work-In-Process management techniques and are modeled using global scheduling models written as Linear Programs. The approach includes two key points:

- The determination of production targets, i.e., quantities of each product to be completed in each operation and each period on a scheduling horizon. These production targets should be followed at work-center level and updated regularly in order to integrate the evolution of the factory,
- Strategies depending on criteria to optimize. These strategies are implemented through mathematical programming models (global scheduling models)

2.2 Framework of the Global Scheduling Approach

The front end area in semiconductor manufacturing is generally managed locally at the work-center level with dispatching rules or dedicated scheduling algorithms. This is done independently in each work-center as discussed in Chapter 1.

Local management has certain drawbacks such as a short-sighted view and may create an unbalanced Work-In-Process in the factory. To deal with this problem, global scheduling management is required. In general, priorities are used as global strategies at the global level to steer dispatching or scheduling decisions at the local level. The disadvantage of priority-based management includes dynamically defining and managing priorities. The issue is that priorities can contain several elements such as the importance of the customers. Priorities are also defined at a higher decision level (tactical level). Thus, a priority-based management

approach used as a global scheduling strategy does not seem to be a good approach to drive scheduling decisions at local level, because priorities might lose their relevance when they are often changed at the local level. In addition to priorities, global management is also carried out using production targets which serve as guidelines for scheduling decisions at work-center level. However, as discussed in Chapter 1, the determination of production targets is generally based on the experience of production managers or on the basis of simple calculations.

The global scheduling approach proposed in this thesis uses optimization methods to widen the scope of the parameters that are considered and to offer the possibility of using different strategies integrated in mathematical models for the management of the Work-In-Process. The goal is to determine the production targets to optimize different criteria. The framework in Figure 2.1 summarizes the global scheduling approach and how it is evaluated.

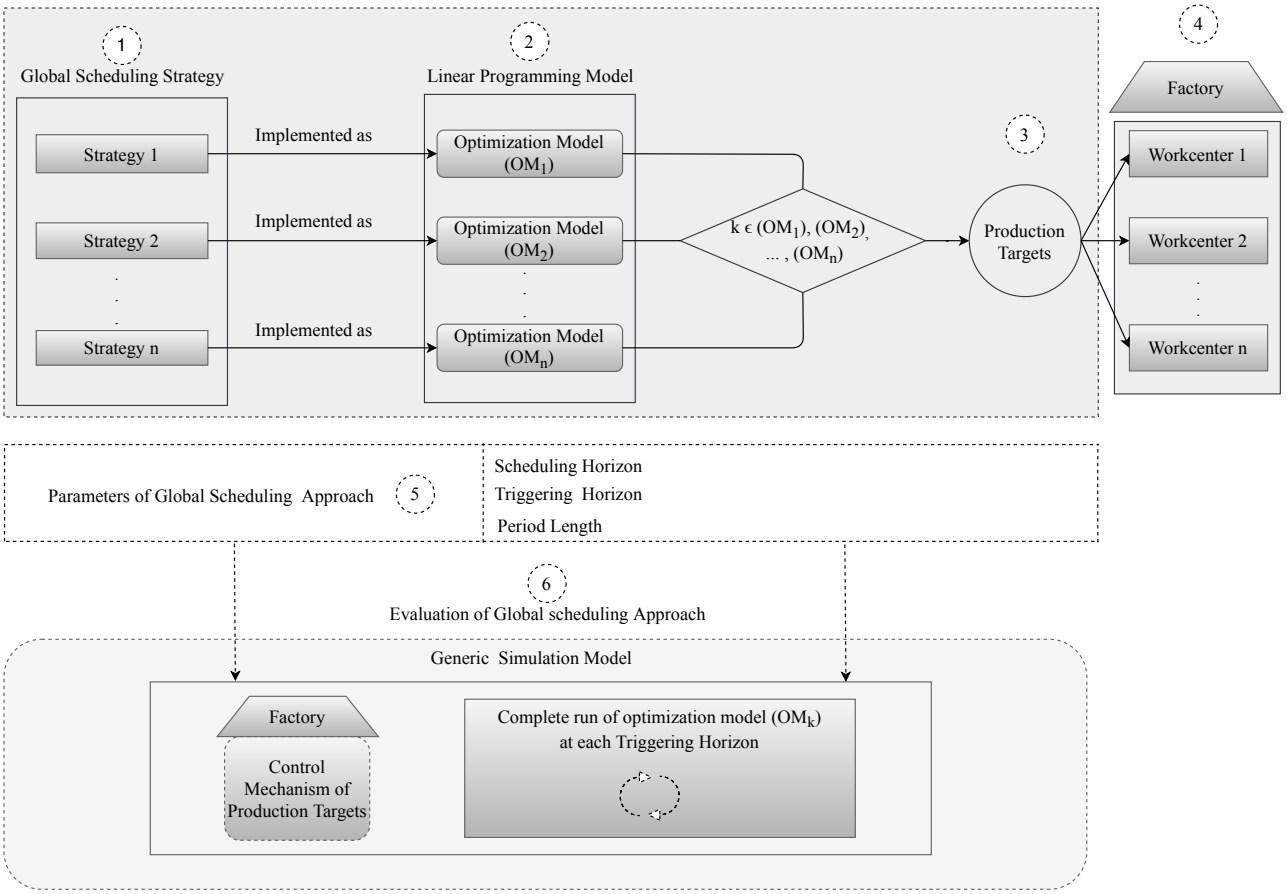


Figure 2.1: Framework of the global scheduling approach

The global scheduling approach includes three main parts:

– Global scheduling strategies *(1)*. In this thesis, a strategy is a Work-In-Process management policy which can be based on the operations of products, different resources, etc. with the objective to optimize one or more criteria. The strategies we defined aim to minimize the variability of the throughput of finished products, to maximize the throughput and minimize the cycle times, and to control the cycle times. They are based on techniques of Work-In-Process management as detailed in Section 2.2.1.

– Global scheduling models *(2)*. They implement the global scheduling strategies as Linear Programming models, where objective functions represent the way strategies are conducted and constraints bound the action of the strategies.

– Production targets *(3)*, i.e., quantities to complete for each product in each operation at each period on a scheduling horizon. For all strategies the production targets are outputs of the global scheduling models. Thus, by sending the quantities to be completed at the work-center level, the global scheduling approach is able to optimize different criteria.

The evaluation of the global scheduling approach *(6)* is carried out using a generic multi-method simulation model which represents the local scheduling level *(4)*. To evaluate the approach, various parameters are required such as the scheduling horizon, the length or duration of each period in the scheduling horizon and the horizon within which the global scheduling strategy is applied, called the triggering horizon in this thesis *(5)*. These parameters are described in Section 2.2.2 and the evaluation of the global scheduling approach is detailed in Section 2.3.

2.2.1 Different Global Scheduling Strategies

Global scheduling strategies are driven by the criteria to optimize. They are implemented as Linear Programming models (global scheduling models) and use global information from the factory such as the Work-In-process, release dates, cycle time targets, resource capacity, etc.

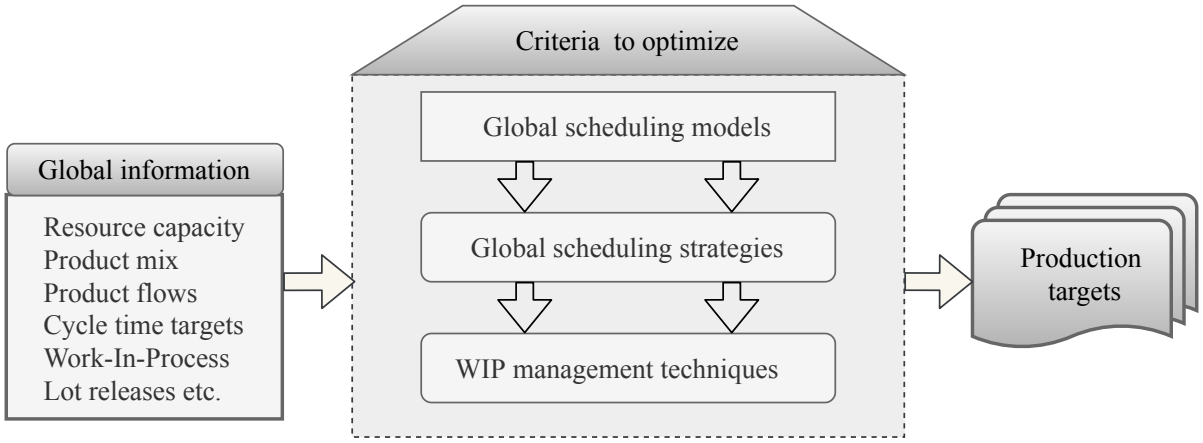


Figure 2.2: Global Scheduling Strategies

As shown in Figure 2.2, global scheduling models implement global scheduling strategies, which are based on Work-In-Process management techniques. Global scheduling strategies aim to determine production targets that should be followed at local level using simple dispatching rules or dedicated scheduling algorithms. The objectives which drive the different global scheduling strategies used in this thesis are described as follows:

- **Minimization** of the variability of cycle times and throughput. This objective is optimized using strategies that determine the percentage of Work-In-Process that each product should maintain at the end of each period over a scheduling horizon. This percentage is called the balancing coefficient.
- **Maximization of throughput**. The optimization of this objective is made possible by the use of a pull strategy. The pull strategy ensures that the more advanced are the products in the factory, the highest their priority.
- **Minimization** of cycle times. The Work-In-Process management strategy used to optimize this objective is essentially focused on minimizing the product waiting times. The strategy ensures that products that are waiting in an operation are processed before those that have arrived later.
- **Control of cycle times**. The associated Work-In-process management strategy is based on the grouping of product operations into subsequences of operations (blocks of operations) to facilitate the control of the time each product spends through each block. Each block has a target cycle time based on the given cycle time target of the product. The aim is to minimize the tardiness and/or the earliness at products in each block in order to meet the product cycle time targets.

2.2.2 Parameters and Decisions

In the global scheduling approach, a Linear Programming model is solved regularly in a rolling horizon setting. Thus, it is crucial to define the key parameters for the global scheduling approach (see Figure 2.3): (1) The duration of each period, (2) the scheduling horizon (number of periods in the horizon) and (3) The number of periods (called triggering horizon in this thesis) before solving again the Linear Programming model. The triggering horizon is

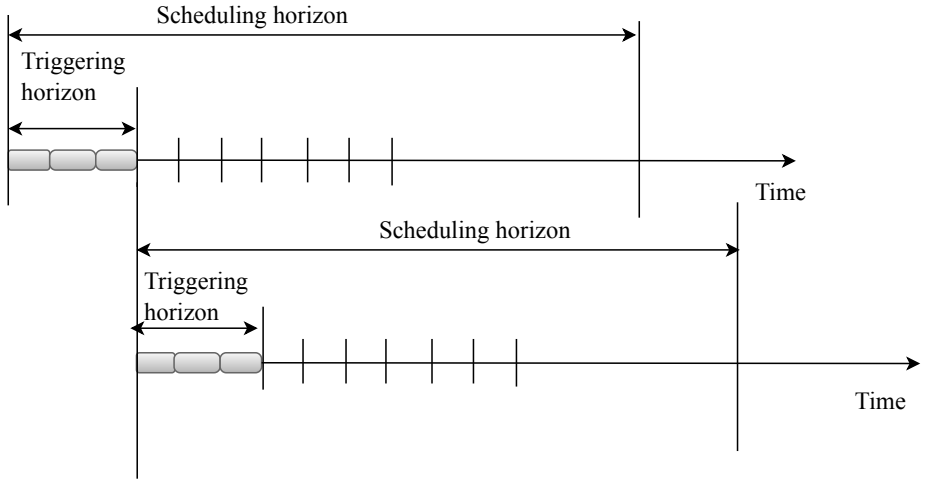


Figure 2.3: Scheduling horizon and triggering horizon in global scheduling strategy approach

important because the global scheduling model does not consider the detailed characteristics of the work-centers. Therefore, the model must be regularly solved to update the decisions by taking into account the events that occurred before the triggering horizon. The triggering horizon should not be too small to avoid changing decisions too often, or too long

not to ignore some critical events. The scheduling horizon is important since it is used to predict the future behavior of the system affected by the scheduling decisions. A sufficiently long scheduling horizon helps the global scheduling model to mitigate the end of the horizon effects. The triggering horizon and the scheduling horizon, but also the duration of each period, strongly depend on the problem and can be determined through computational experiments.

2.3 Evaluation of the Approach using a Simulation Model

Simulation is a widely accepted approach for the design and analysis of manufacturing systems. It can model non-linear and stochastic problems and allow the examination of the likely behavior of a proposed manufacturing system under selected conditions. Through simulation analysis, many details and constraints can be considered in the evaluation of a manufacturing system. Different computational experiments can be performed with a simulation model. These experiments can be the prediction of the effect of mix changes, the extraction of relevant information, etc. The application of new strategies to control the manufacturing system may require additional experiments to analyze KPIs such as cycle times, throughput and output variability on cycle times and throughput.

In semiconductor manufacturing, a front-end manufacturing facility, also called wafer fab, usually processes many products. Each product has a processing route, which contains a sequence of hundreds of operations. Products of the same type are grouped in a lot (a lot contains at most 25 wafers). A typical fab includes several hundred machines, which are grouped in work-centers. Each work-center is dedicated to a specific type of operations. Based on the characteristics of frond-end manufacturing facilities, a generic data-driven simulation model for complex semiconductor manufacturing facilities was initially developed in Sadeghi et al. (2016) with the aim to study the consistency between global flow management objectives/decisions (fab level) and local scheduling/dispatching objectives/decisions (work-center level). This simulation model is being improved in order to take into account new parameters such as the warm-up time (the time when the factory is loading), the mechanism at local level to control objectives sent from the global level, and the strategies of the global scheduling approach. The aim is to ensure that global objectives defined at the fab level are followed at the local level. Details on the exchange/communication interface between the global scheduling approach, which represents the global level, and the simulation model, which represents the local level, is given in Section 2.3.2. The simulation model is used to evaluate the different global scheduling strategies proposed in this thesis.

Section 2.3.1 presents the structure of the data-driven simulation model, input parameters, objectives and Key Performances Indicators to evaluate the performance of the manufacturing system. Section 2.3.2 presents the exchange/communication interface between the simulation model and the global scheduling approach. Finally, Section 2.3.3 presents the design of the computational experiments used to evaluate the performance of the global scheduling approach.

2.3.1 Simulation Model

The conceptual model of the simulation model is presented in Figure 2.5. From the meta-model (first panel of Figure 2.5), the data model and the model structure of the Factory are derived. The simulation model is a multi-method model, which combines Discrete Event

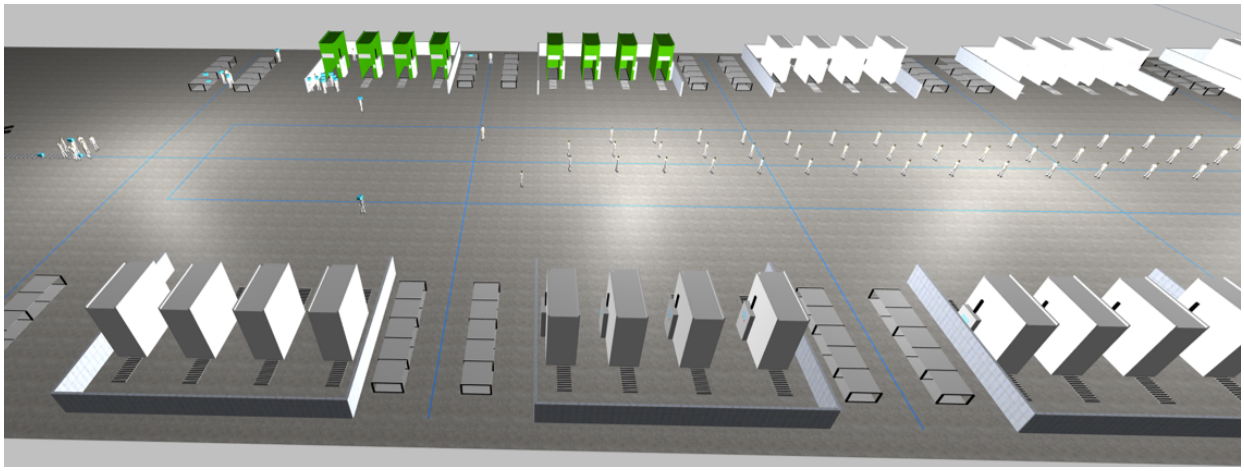


Figure 2.4: Panoramic view of simulation model

(DE) and Agent-Based (AB) simulation methods in order to combine the benefits from both modeling approaches, in particular the notion of queues in Discrete-Event simulation and the flexibility, behavior, and communication of agents in Agent-Based Simulation. The different types of behavior or processes in each entity (agent) are designed with DE simulation modeling. For instance, since lots of different products at different operations compete for the same machines, the notion of queue in front of machines is needed. Therefore, queues are designed with DE simulation where each work-center (group of machines with same capabilities) is modeled as an agent. To complete its production process in the fab, every lot (also modeled as an agent) needs to interact with other agents via different rules in such a way that all interactions generate the overall system behavior, see Sadeghi et al. (2016). The Agent-Based simulation model is used to ensure these interactions. The main types of agents are the lots and work-centers. Secondary agents are non-physical components such as operations and routes. The interaction and behavior of agents are taken into account in the production logic implementation. Input data (work-centers, routes, additional operational parameters, etc.) are supported by an Excel file format.

In the production logic (last panel of Figure 2.5), after releasing a lot in the system, a route identification (Route ID) is associated with the lot depending on the lot type. In the next stage, the first operation is requested and allocated to a qualified work-center. As an operation can be treated in different work-centers, the work-center with the smallest queue is selected. The process continues as in the previous stage until the last operation of the route is completed. Then, the lot exits the system.

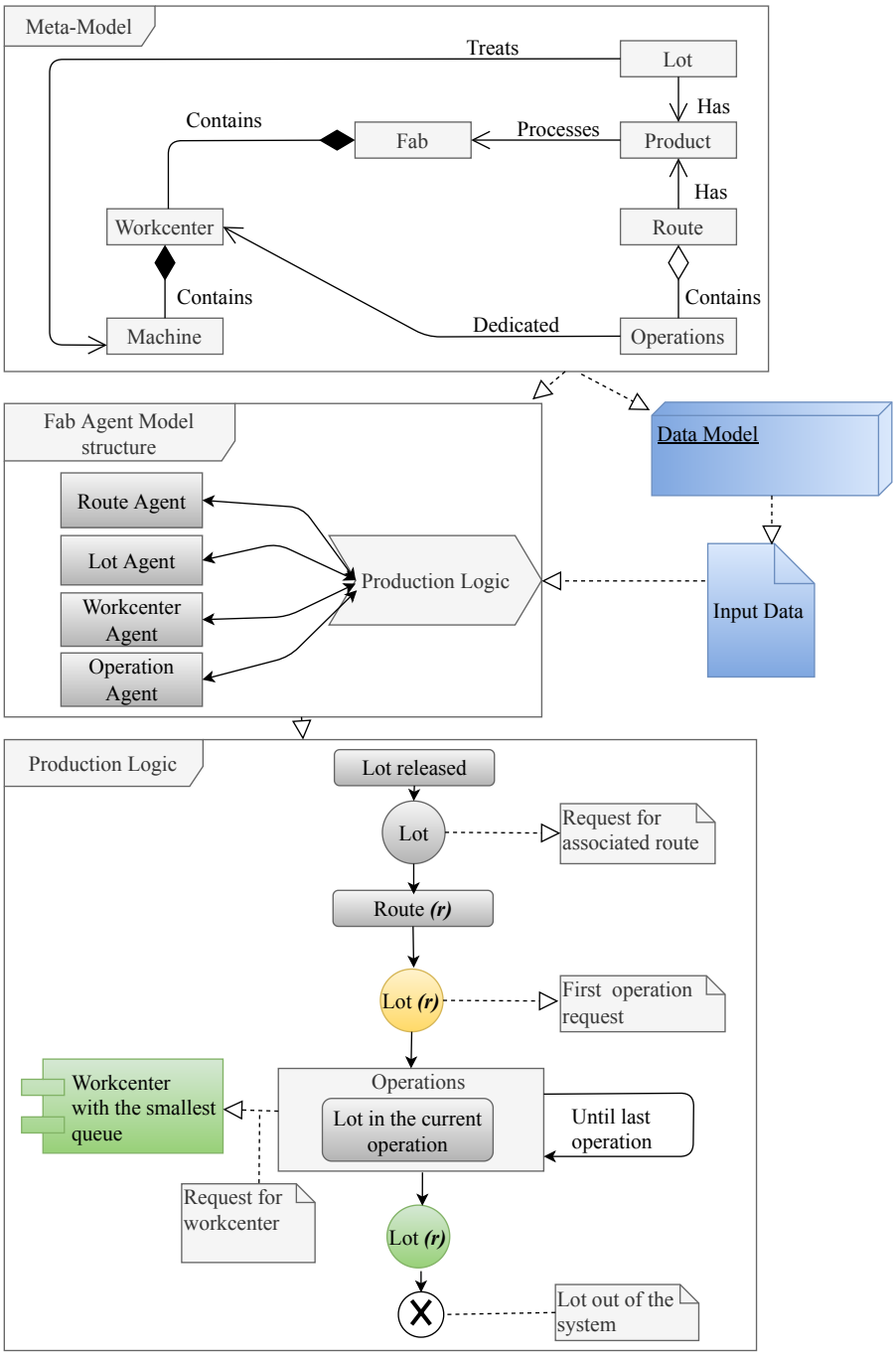


Figure 2.5: Conceptual model of simulation model

System performance evaluation

To analyze the system performance, qualitative evaluations such as *good*, *fair*, *adequate*, and *poor* are vague and difficult to use in any meaningful way. Instead, quantitative performance measures are preferable. The process of choosing appropriate manufacturing performance measures is difficult due to the complexity of these systems. To be successful, production systems must deliver the products with the desired functions, aesthetics, and high-quality to the customers at the right time. To do so, several performance measures exist, such as cycle time, Work-In-Process inventory, throughput, productivity, and service levels. In the

following, we provide an overview of the most commonly used performance and productivity evaluation metrics in the semiconductor industry. The goal is to understand both the impact of the different metrics and the relationship between them.

The selection of input parameters depends on the current status of the fab under study and the use of the simulation model. For instance, parameters such as the warm-up time (the time when the factory is loading) should be taken into account since we assume that we are not working with a new factory. The simulation model is related to operational control, and the main input parameters are:

- The number of lots started in the system per week,
- The total number of lots to produce, this parameter depends on the simulation horizon (the longer the simulation horizon, the larger the total number of lots to produce),
- The number of products in the mix,
- The warm-up time or initial Work-In-Process, and
- The priority of each product.

In order to compare the factory expected and actual realizations, theoretical parameters based on historical data can be used as auxiliary parameters. When considering operational decisions in the simulation model, common auxiliary parameters are the theoretical cycle time of each product, the theoretical throughput, and the theoretical yield. The simulation model is designed based on an objective that drives the modeling. The objective is mainly evaluated via defined Key performance Indicators (KPIs) which outline the achievement of the objective. Thus, Key Performance Indicators are decided at the highest level and support the overall long-term strategic objectives of the factory and they are in turn supported by metrics at local level as shown on 2.6.

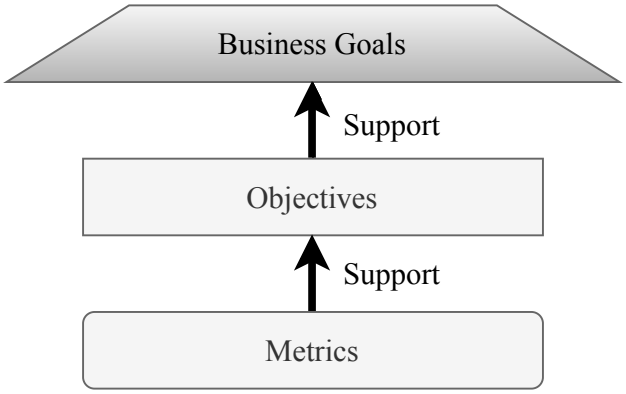


Figure 2.6: Objectives and Metrics

The main objectives to optimize at the operational level are:

- The average and the standard deviation of cycle times,
- The throughput, i.e., the number of wafers produced,
- The output variability on cycle times and throughput,

– The on-time delivery,

– The yield percentage

In general, the metrics that support these objectives are:

– The number of moves (productivity),

– The average Work-In-Process in the system for each product,

– The machine utilization, and

– The Work-In-Process balancing.

The main objectives optimized in this thesis to analyze the effectiveness of the global scheduling approach are the output variability on cycle times and throughput, the throughput (number of completed wafers), and the cycle times. Other objectives are being defined depending on the strategy to follow, such as speed up products to minimize their cycle times, balancing the Work-In-Process to satisfy estimated throughput or satisfying cycle time targets, etc.

2.3.2 Exchange/Communication Interface between Simulation Model and Global Scheduling Approach

The input/output exchange/communication implements the communication strategy between the aggregate global scheduling model and the simulation model. Since the simulation model considers a granularity (process unit is one lot) that is different from the granularity of the global scheduling model which process quantities, this section describes the exchange/communication allowing the simulation and optimization models to feed each other with input/output data, see Figure 2.7.

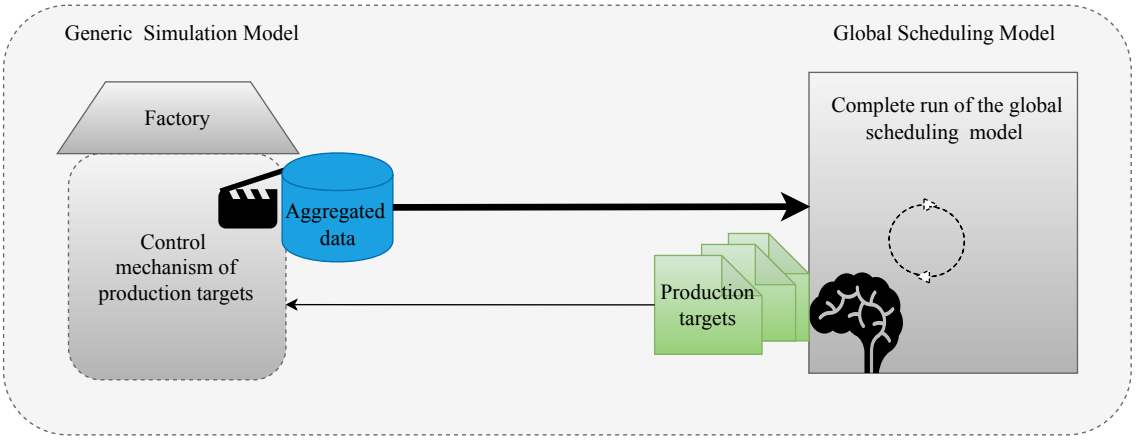


Figure 2.7: Exchange of information between the global scheduling model and the simulation model

The global scheduling model implements the global scheduling strategy, which guides scheduling decisions at the work-center level in the factory (represented by the simulation model) by providing objectives in terms of production targets, i.e., product quantities to complete for each operation and at each period on a scheduling horizon. As the global

scheduling optimization model is actually embedded in the code of the simulation model, the interface of the exchange/communication is of great importance and should be clearly defined.

There is a significant number of papers in the literature where simulation and optimization are combined in semiconductor manufacturing. However, it seems that there is no paper that discusses the interface of the exchange/communication between optimization and simulation. A review of simulation optimization methods with application in semiconductor operational problems can be found in Ghasemi et al. (2018).

Analyzing the system under study before setting up the exchange/communication interface is crucial. Indeed, it is difficult to define the interface for the exchange/communication without deciding whether the combination of the global scheduling model and the simulation model should be a *System* or a *System Of Systems* (SOS). The difference between a *System* and a *System Of Systems* (SOS) lies essentially in its components. A System Of Systems is a System. However, the components of a SOS act as autonomous systems (Boardman and Sauser (2006)). Based on this autonomous property of an SOS, we consider in this thesis that the combination of the simulation model and the global scheduling model is an SOS. The simulation model can be executed correctly without the global scheduling model and the latter can work independently outside the simulation model. For more details on SOS, see Boardman and Sauser (2006) and Gorod et al. (2008).

The interface of the exchange/communication ensures the analysis of the primary correctness function of the global scheduling model. It also ensures that each part of the SOS, the simulation model and the global scheduling optimization model can operate independently. In this thesis, the simulation model is used to evaluate the global scheduling model. The simulation-based optimization framework and the connectivity interface which links the simulation model and the global scheduling model are discussed below.

Simulation-based optimization framework

Figure 2.8 describes the components of the simulation-based optimization framework.

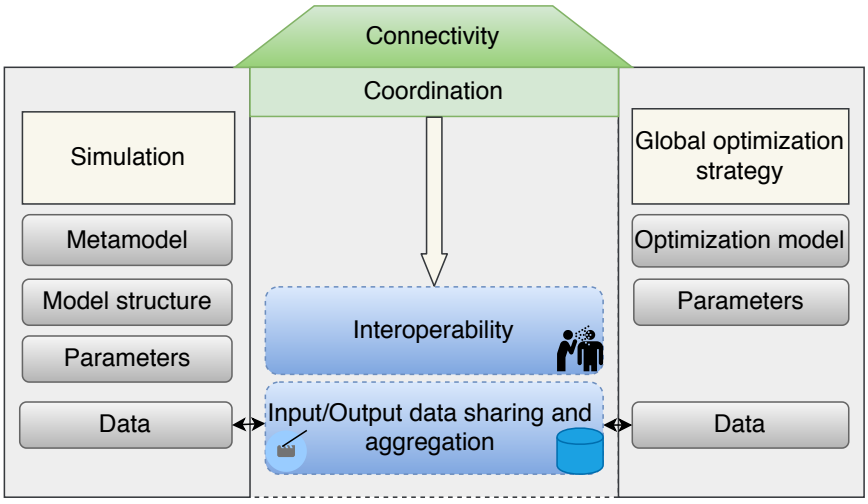


Figure 2.8: Simulation-based Optimization Framework

The simulation-based optimization framework includes both systems that form the SOS (the simulation model and the global scheduling optimization model) and the interface of

the exchange/communication (connectivity interface).

The simulation model in the framework has the following elements:

- A meta-model (see the first panel of Figure 2.5), which describes the entities involved in the simulation model (fab, product, lot, route, machine, operations, etc.) and their relationships,
- Parameters, that represent different adjustable characteristics of the simulation model such as started lots per week, total number of lots to produce, warm-up time, priority of each product, etc.,
- A model structure, which specifies the interactions of entities in the production logic (see the last panel of Figure 2.5).

The global scheduling model includes the following elements:

- A Linear Programming model that materializes global scheduling strategies. It includes an objective function, constraints and variables,
- Parameters associated with the global scheduling strategies such as the scheduling horizon, the duration of the period in the scheduling horizon and the triggering horizon.

The interface of exchange/communication

The interface of exchange/communication (connectivity interface) includes the following elements:

1. *Coordination* represents the way both systems are synchronized in an SOS environment. The main mechanism used in this thesis to make the coordination efficient is based on the running order. The SOS begins with the run of the simulation model on the triggering horizon. Next, the global scheduling optimization model is called in a rolling horizon by a simulation trigger event. After collecting dynamic parameters from the current status of the simulation model, such as current Work-In-Process levels in work-centers, and static parameters, such as future releases and aggregate resource capacities, the global scheduling model is solved to determine production targets. In the meantime, the simulation model is paused. When the optimization is completed the production targets Y_{glp} for product g in operation l and period p determined by the global scheduling model is then imposed as **constraints at the work-center level** in terms of production quantities of each product to complete at each operation in each period. Then, the simulation model resumes and tracks these production quantities.
2. *Interoperability* corresponds to the way the simulation model and the optimization model cooperate in order to achieve the objective of the SOS, i.e., the optimization of different objectives. The simulation aims at satisfying the objectives sent by the global scheduling optimization model in terms of production targets. A mechanism based on a controller variable is used, which indicates whether the production target of a particular product is reached at a given operation in each period. In addition, the interoperability ensures that the future product releases collected as static parameters are properly synchronized with the product release scheme in the simulation. Finally, the interoperability guarantees that, in the simulation model, the representation of the parameters of the global scheduling optimization model (the scheduling horizon,

the triggering horizon and the duration of each period in the scheduling horizon) are transformed from periods units to simulation time units.

3. *Input/output sharing data and aggregation* represent the way static and dynamic data are collected and aggregated. For example, machines are grouped in work-centers (each work-center includes machines with the same capabilities), the sum of the capacities of the machines in the work-centers are aggregated as individual capacity of the work-center, quantities of lots at each operation are used in the global scheduling model instead of individual lots, etc.

The connectivity interface that links the simulation model and the global scheduling model is presented on Figure 2.9.

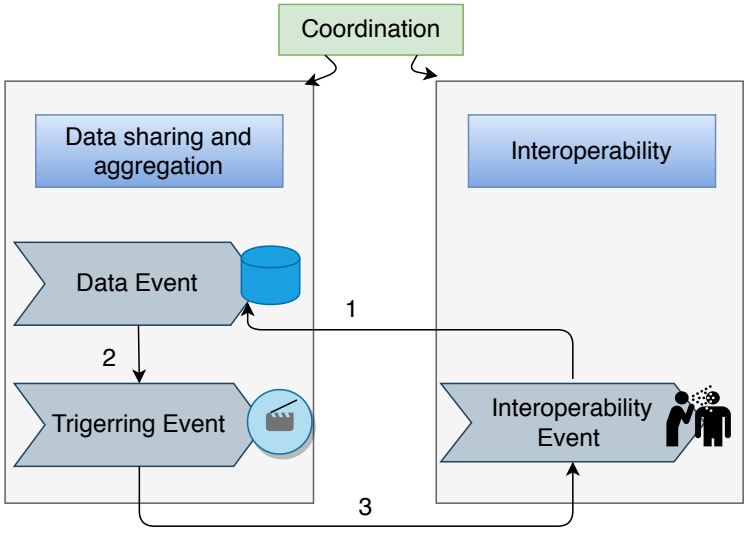


Figure 2.9: Connectivity interface linking the simulation model and the global scheduling model

In the simulation model, three dynamic events are in charge of the communication between the simulation model and the global scheduling optimization model:

– The interoperability event, which ensures that the production targets, defined by the global scheduling optimization model, are followed in the simulation model using the controller variables which track production quantities in each period in the triggering horizon,

– The data event, which collects all static and dynamic parameters needed by the global scheduling model and ensures that data is aggregated properly,

– The triggering event, which calls the global scheduling model and collects its output. It is used to assess the accuracy of the outputs of the global scheduling model before their use in the simulation model.

As shown on Figure 2.9, these three events trigger each other cyclically. Before the first call of the optimization model, the *Interoperability Event* counts the number of periods to ensure that the simulation model is paused at the end of the triggering horizon. Next, it triggers the *Data Event* to collect static and dynamic parameters from the current state of

the simulation model and aggregates this data. Then, the *Data Event* calls the *Triggering Event*, which in turn calls the global scheduling model. After the global scheduling model is solved, the *Triggering Event* triggers the *Interoperability Event* to ensure that production targets are followed in the simulation model. The *Interoperability Event* again starts to count the number of periods before the next call of the global scheduling model and so on. This iterative procedure continues until the end of the simulation horizon.

2.3.3 Design of Computational Experiments

The simulation model starts by creating the required agents such as the routes of products, product operations, work-centers, etc. These agents are then fed with data from Excel files (data related to the fab such as work-centers, number of machines in each work-center, processing times, etc). Finally, parameters of the simulation model and of the global scheduling model are initialized, and lots of products are generated following the product release scheme.

The interoperability provides the controller variables, which are set up to indicate whether each production target Y_{glp} for product g at an operation l in period p is reached.

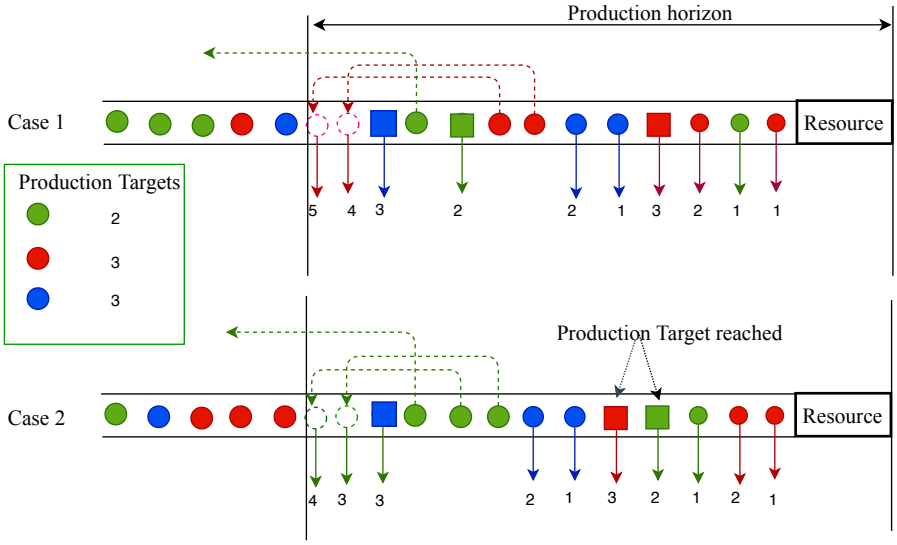


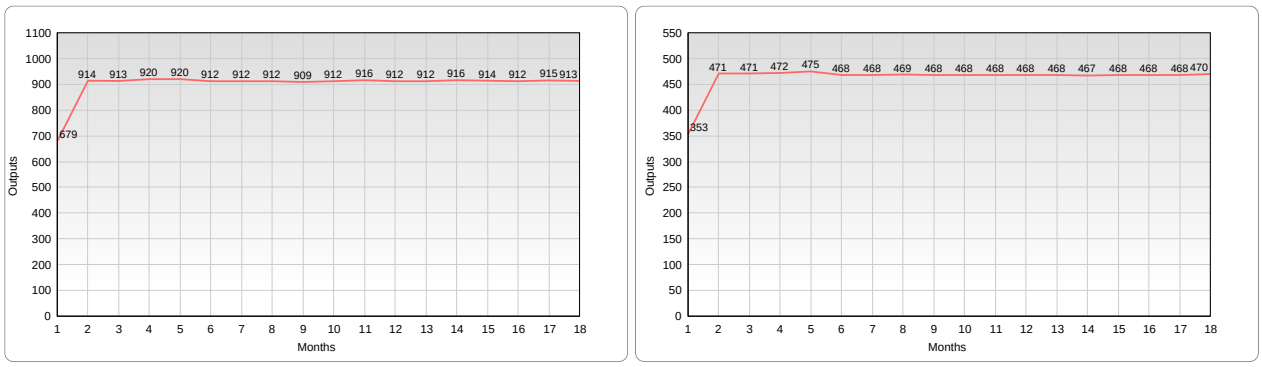
Figure 2.10: Mechanism to track production targets

In some situations, depending on the status of the queue of the resource and the production targets, short cycle times will not always lead to high throughput. A product can have high throughput with long cycle times or short cycle times with low throughput. This is due to the fact that, if a product reaches its production target, then its production is temporally stopped in order to track the production targets of other products. A lot of a stopped product can only be produced when all products reach their respective targets or when it is the only one in the queue of a resource. For illustration, assume processing times of one time unit on resources looking at Figure 2.10, the red product has an average cycle time of 5.8 time units with a throughput of five lots in *Case 1*. While in *Case 2*, the same product has a cycle time of 2.6 time units with a throughput of three lots.

Numerical experiments have been conducted using two industrial data sets associated to two different factories. Five product families are considered for the first industrial data set and ten product families grouped in two different instances for the second industrial data

set. The industrial data set of the first factory includes 449 machines in 203 work-centers, and products have between 352 and 622 operations in their routes. The industrial data set of the second factory includes 570 machines in 329 work-centers, and products have between 104 and 315 operations in their routes. In both data sets, products are continuously released in the system in a uniform scheme. The global scheduling model and the data driven generic multi-method simulation model are implemented using the AnyLogic software (version 8.4) which interacts with the standard solver IBM ILOG CPLEX (version 12.6). Experiments are performed on a computer with Windows 10 as operating system, a processor Intel(R)Xeon(R) CPUE3-1240v5, 2*3.50 GHz and 32 Go of RAM.

As we are not working with a new factory, six months of warm-up time (time to load the factory) are used. Figure 2.11 presents how the warm-up time was determined based on the outputs of each factory, i.e., products that get out of the system each month were collected in the simulation model for eighteen months. Figure 2.11 shows that the steady state of each factory is reached before the sixth month. The first six months of warm-up time are excluded when collecting statistical data.



(a) First factory industrial data

(b) Second factory industrial data

Figure 2.11: Determination of warm-up time based on factory outputs

The period duration in the global scheduling model corresponds to one shift (eight hours), and the global scheduling model is called in the simulation model every three periods (triggering horizon of twenty four hours), i.e., more than 300 times in total. The scheduling horizon in the global scheduling model is fixed to thirty-three days (792 hours). The simulation horizon is fixed to eighteen months except in Chapters 4 and 5, where the simulation horizon is twelve months. This is because of the lack of memory during simulation, essentially when the simulation model is coupled with the global scheduling models.

2.4 Conclusions

This chapter describes a global scheduling approach to steer local scheduling decisions at work-center level. The approach adopts two views of the operational decision level: The global level (factory level) and the local level (work-center level). The approach includes three key points:

1. Production targets that should be followed at work-center level and updated regularly to take into account the evolution of the factory and the objectives to optimize,

2. Strategies based on the objectives to optimize, that are implemented through mathematical programming models (global scheduling models) and,

3. Required parameters for the evaluation of the approach. These parameters include the scheduling horizon (number of period in the horizon), the duration of each period in the scheduling horizon and the number of periods before solving again the global scheduling model (called triggering horizon).

The approach ensures that several objectives can be optimized by only sending production targets to the work-center level. These production targets are determined using different global scheduling strategies. The approach is suitable for semiconductor manufacturing, but can be used for any other complex manufacturing system. A multi-method generic simulation model is used to evaluate the approach.

Chapter 3

Work-In-Process Balancing Control and Throughput Maximization

3.1 Introduction

This chapter presents new strategies for managing Work-In-Process in complex industrial systems, particularly in semiconductor manufacturing. Work-In-Process corresponds to the products already in the factory, but not yet completed. It is certainly difficult to improve Key Performance Indicators (KPIs) such as cycle times, throughput, variability of cycle times and throughput, and on-time delivery without a thorough management of the Work-In-Process. Work-In-Process balancing control, i.e., ensuring that Work-In-Process is properly distributed throughout the whole factory is considered as an efficient method to improve KPIs (Lee and Lee (2003)). Balancing the Work-In-Process allows a good use of production capacity and ensures a good distribution of products in the factory. It also guarantees that all products progress regularly towards the end of their operations.

Work-In-Process control is performed to optimize one or more objectives at a time. The main objective of this chapter is to optimize the output variability on cycle times and throughput of completed products and their throughput. Secondary concerns are the acceleration of products for better cycle times as well as the satisfaction of throughput and cycle time targets. Throughput is known as one of the important KPIs in semiconductor manufacturing, because a high throughput often leads to a factory achieving high revenues (Chung and Jang (2009)). Minimizing the output variability on cycle times and throughput (ensuring that all products are advancing at a regular pace) prevents some non-prioritized products from slowing others. A common example is of a high percentage of hot lots which can substantially increase the average cycle time and/or inventory costs of all other lots (Ehteshami et al. (1992)). It seems that there are not many studies on the output variability on cycle times and throughput in the literature on semiconductor manufacturing. For a noticeable exception, see Chen et al. (2010). In their study, they propose an approach to minimize the output variability not on finished products, but on the deviation between actual production and desired customer demand. For more details about variability in semiconductor manufacturing, see Li et al. (1996), Schoemig (1999) and Dequeant et al. (2016).

In the literature on semiconductor manufacturing, single objective problems are widely studied, but in practice, problems often appear with multiple contradictory objectives. For example, minimizing the output variability on cycle times and throughput can, in certain situations, imply a reduction in throughput. This is due to the slowing of the products with

short cycle times which can otherwise contribute to the increase in total throughput of the factory. Numerical results in Section 3.3, show that when the output variability on cycle times and throughput is optimized alone, some products do not reach 100% satisfaction of their estimated throughput, but with a multi-objective approach, with both the optimization of the output variability on cycle times and throughput and the optimization of throughput, almost all products reach 100% satisfaction of their estimated throughput with an acceptable compromise on the variability compared to the results when only the output variability on cycle times and throughput is optimized. The trade-offs are a maximum difference of 3.3 and 0.1 variability on cycle times respectively for industrial instances one and two presented in Section 3.3. These compromises are acceptable compared with the results when only simulation is used without the global scheduling approach which presents a maximum difference of 10.1 and 0.8 variability on cycle times respectively for industrial instances one and two. The output variability on cycle times and throughput is analyzed using the InterQuartile Range (IQR), which is a robust measure of variability. It indicates the central dispersion of 50% of the values in the data set and is calculated based on the median.

To our knowledge, minimizing the output variability on cycle times and throughput and maximizing throughput have never been studied together. The strategies used in this chapter balance the minimization of the output variability on cycle times and throughput and the maximization of throughput. The single objective strategy aiming at minimizing the output variability on cycle time and throughput of completed products is performed by using balancing coefficients, i.e., the percentage of Work-In-Process for each product that should remain at the end of each period over a scheduling horizon. These balancing coefficients define the flow of each product in the factory. A small balancing coefficient for a product means that the product should significantly contribute to the total throughput of the factory. The multi-objective strategy is formulated using the ϵ-constraint approach.

Section 3.2 presents the Work-In-Process balancing control to minimize the output variability on cycle times and throughput and provides an analysis of the satisfaction of throughput and cycle times. The study in this section was presented at the Winter Simulation Conference 2019 (WSC 2019), see Barhebwa-Mushamuka et al. (2019b). In Section 3.3, the multi-objective optimization for Work-In-Process balancing and throughput maximization is formulated using an ϵ-constraint approach and its adjusted version. The study in this section was presented at the International Conference on Automation Science and Engineering (IEEE CASE 2019), see Barhebwa-Mushamuka et al. (2019a).

3.2 Work-In-Process Balancing Control

This section addresses the problem of Work-In-Process balancing control in semiconductor manufacturing systems. The Work-In-Process balancing strategy is proposed to minimize the output variability on cycle times and throughput. Balancing coefficients are used as leverage to speed up a product while controlling the impact of this acceleration on other products. In addition, the strategy through the balancing coefficients guides the Work-In-Process in order to satisfy the given throughput and cycle time targets. Work-In-Process is controlled so that each product progresses in the factory according to its balancing coefficients.

3.2.1 Controlling Work-In-Process using Balancing Coefficients

The objective of the Work-In-Process balancing strategy is to provide a flow scheme for each product. This is achieved through the use of balancing coefficients which are the percentage of Work-In-Process for each product that should remain at the end of each period over a scheduling horizon. The global scheduling model will determines production targets so that the products advance in the factory according to their defined balancing coefficients. As shown on Figure 3.1, the optimization model controls the Work-In-Process to such an extent that, at the end of each period, no product should have a Work-In-Process greater than a given percentage δ_g of the total current Work-In-Process in the factory. Products with small balancing coefficients are expected to make a large contribution to the overall throughput of the factory.

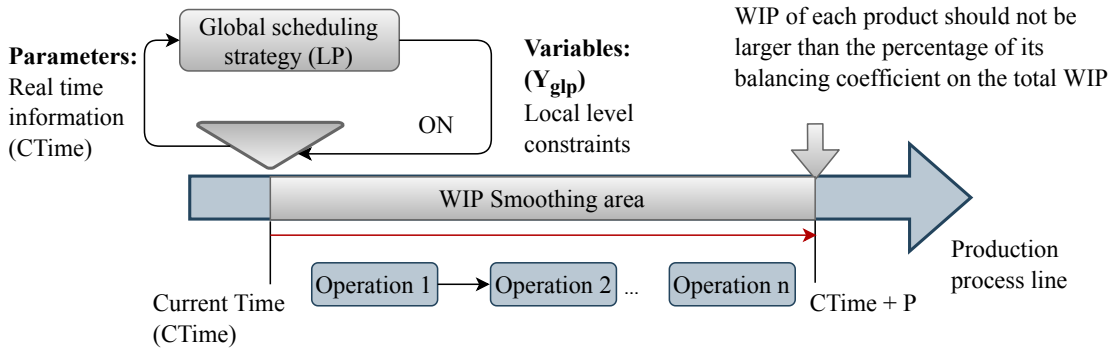


Figure 3.1: Work-In-Process balancing control strategy

Determination of Balancing Coefficients

Balancing coefficients are product-oriented with the objective to control the production flow of each product. Generally, it is difficult to quantify and to manage the impact that an acceleration of one product can have on other products in the mix, see Sadeghi et al. (2016). The balancing coefficients provide not only the way each product should flow in the factory, but they also control the acceleration of products, i.e., a product can be accelerated without drastically deteriorating the cycle times or the throughput of the other products. The balancing coefficients must be determined so as to avoid unrealistic Work-In-Process. For instance, by asking a particular product to keep more Work-In-Process than it has in the factory. Or to force a product with the largest Work-In-Process to remain with the lowest Work-In-Process of the factory, i.e., to ask for production rates which cannot be reached by the factory. Several methods can be used to determine the balancing coefficients. Without being exhaustive, these methods are based on:

1. The way products are started in the factory (release scheme). In semiconductor manufacturing, the manufacturing process of a product goes beyond a hundred operations. Products take weeks to complete all operations before leaving the factory. Therefore, the balancing coefficients based on release scheme are determined in such a way that, products with short intervals between consecutive release dates in the system remain with large Work-In-Process. And products with a long interval between consecutive release dates remain with less Work-In-Process. Balancing coefficients determined on

the basis of release scheme provide a better reduction of the variability of the finished products. This is due to the fact that products with a short interval between consecutive release dates and products with short cycle times will not always be prioritized. To determine properly the balancing coefficients, simulation experiments can be used to observe the contribution to Work-In-Process of each product according to its release dates.

2. The estimated throughput ($\hat{S}_g$). Let us define the estimated throughput as the quantity of products released into the system after the warm-up time (time to load the factory) up to the end of the simulation horizon minus the maximum cycle time of all products. This implies that the estimated throughput should be completed before the end of the simulation horizon. In our computational experiments, we considered a maximum cycle time of all products equal to three months. In the same way as the estimated throughput, the balancing coefficient is based on the demand for products or the throughput provided by historical data or simulation. Let δ_g be the balancing coefficient of product g and C_g the contribution of each product g in the total estimated throughput as shown in equation 3.1.

$$C_g = \frac{\hat{S}_g}{\sum_{g \in \mathcal{G}} \hat{S}_g} \tag{3.1}$$

Then, the balancing coefficient δ_g of each product g is computed as:

$$\delta_g = \frac{1 - C_g}{\sum_{g \in \mathcal{G}} 1 - C_g} \tag{3.2}$$

The balancing coefficients computed in (3.2) guarantee that the estimated throughput is reached for each product. If some products do not reach 100% of the estimated throughput, the balancing coefficients are updated using algorithm 3.1. The algorithm determines the percentages to decrease on the balancing coefficients of products that do not reach 100% of the estimated throughput and propagates the decreased percentages on the balancing coefficients of products that exceed 100%. Steps (1) - (9) define the parameters and the variables used in the algorithm, while step(10) initializes the sum of the percentages higher than 100%. Steps (11) and (12) compute the percentages above target for all the products which have exceeded their estimated throughput and determine the remaining percentages to reach 100% for products that have not met their estimated throughput. The computation of percentages to update balancing coefficients is provided on steps (13) to (20). Steps (21) - (28) update balancing coefficients for all the products.

This algorithm reduces the balancing coefficients of products that have not reached 100% satisfaction of the estimated throughput. The goal is to speed up these products while increasing the balancing coefficients of products that have exceeded 100% satisfaction of the estimated throughput to slow them down.

Algorithm 3.1 Updating balancing coefficients

$\mathcal{G}$, Set of all products
$\mathcal{G}^{\mathcal{O}}$, Set of products that over-satisfy the estimated throughput, $\mathcal{G}^{\mathcal{O}} \subset \mathcal{G}$
$\mathcal{G}^{\mathcal{U}}$, Set of products that under-satisfy the estimated throughput, $\mathcal{G}^{\mathcal{U}} \subset \mathcal{G}$
$Balancing\_Cg$, Balancing coefficient of product g
$Above\_Tg$, Perecentage above 100% for product g
$Remaining\_Tg$, Percentage remaining to reach 100% of achievement for product g
$Percentage\_Ag$, Percentage of the achieved throughput for product g
$Update\_Qgg'$, Quantity to update balancing coefficient of product g
based on the remaining percentage to reach 100% of achievement for product g'
$sum\_Above \; \forall \; g \in \mathcal{G}^{\mathcal{O}}$, sum of percentage above 100% for all products
$sum\_Above \leftarrow 0$
$Above\_Tg \; \forall \; g \in \mathcal{G}^{\mathcal{O}} \leftarrow Percentage\_Ag$ - 100
$Remaining\_Tg \; \forall \; g \in \mathcal{G}^{\mathcal{U}} \leftarrow 100$ - $Percentage\_Ag$
for all $g \in \mathcal{G}^{\mathcal{O}}$ **do**
    $sum\_Above \leftarrow sum\_Above + Above\_Tg$
end for
for all $g' \in \mathcal{G}^{\mathcal{U}}$ **do**
    for all $g \in \mathcal{G}^{\mathcal{O}}$ **do**
        $Update\_Qgg' \leftarrow (Above\_Tg \div sum\_Above)\ Remaining\_Tg'$
    end for
end for
for all $g \in \mathcal{G}^{\mathcal{U}}$ **do**
    $Balancing\_Cg \leftarrow Balancing\_Cg$ - $Remaining\_Tg$
end for
for all $g' \in \mathcal{G}^{\mathcal{U}}$ **do**
    for all $g \in \mathcal{G}^{\mathcal{O}}$ **do**
        $Balancing\_Cg \leftarrow Balancing\_Cg + Update\_Qgg'$
    end for
end for

The drawback of the balancing coefficients computed in equation 3.2 is that they are only focusing on satisfying the estimated throughput without integrating the product cycle times.

3. Little's law, so that the throughput and the cycle times are taken into account in the computation of the balancing coefficients. Little's law states that the Work-In-Process is equal to the Throughput (T) multiplied by Cycle Time (CT). Assume that T_g is the throughput of product g and CT_g, the cycle time of product g provided by historical data or simulation. Then, the balancing coefficient of each product g is computed as using (3.3):

$$\delta_g = \frac{T_g \; CT_g}{\sum\limits_{g \in \mathcal{G}} T_g \; CT_g} \tag{3.3}$$

δ_g is the contribution of product g in the total Work-In-Process of the factory. The balancing coefficients computed in (3.3) guarantee that both the throughput and the cycle time of each product are reached.

Deciding on how the balancing coefficients are computed essentially depends on the expected objective of the factory, i.e., satisfying cycle times, satisfying demands or the estimated throughput, minimizing the output variability on cycle times and throughput, etc. Table 3.1 defines different parameters and decision variables that will be used throughout this chapter in the global scheduling models.

Parameters:	
$\mathcal{G}$	Set of products,
$\mathcal{K}$	Set of work-centers,
$\mathcal{L}_g$	Set of operations in route of product g,
$\mathcal{LK}(k)$	Set of operations and products that must be processed in work-center k, i.e $(g,l) \in \mathcal{L}K(k)$ means that operation l in route of product g must be processed in work-center k,
P	Number of periods in planning horizon,
IW_{gl}	Initial Work-In-Process at operation l of product g,
R_{gp}	Release quantity of product g in period p,
α_{gl}	Unit process time for product g at operation l of product g,
C_{kp}	Capacity of work-center k in period p,
μ	Balancing penalty,
δ_g	Balancing coefficient of product g,
q_{glp}	Unit Work-In-Process cost at operation l of product g in period p.
Decision variables:	
Y_{glp}	Quantity of product g completing operation l in period p,
W_{glp}	Work-In-Process of product g at operation l at the end of period p,
X_{glp}	Quantity of product g arriving at operation l in period p,
Z_{gp}	Maximum Work-In-Process deviation of product g in period p.

Table 3.1: Notations

Global Scheduling Model Without Work-In-Process Balancing Control

The global scheduling model without Work-In-Process balancing control is described in this section. The strategy implemented in the objective function consists in minimizing the Work-In-Progress remaining for product g at operation l during period p. Let $P^{control}_{NoWIP}$ be the Linear program that models the global scheduling strategy without Work-In-Process balancing control.

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{P} W_{glp} \tag{3.4}$$

Subject to :

$$X_{glp} = Y_{g(l-1)p} \qquad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ \forall p \tag{3.5}$$

$$W_{g11} = IW_{g1} + R_{g1} - Y_{g11} \qquad \forall g \in \mathcal{G} \tag{3.6}$$

$$W_{gl1} = IW_{gl} - Y_{gl1} \qquad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2 \tag{3.7}$$

$$W_{g1p} = W_{g1(p-1)} + R_{gp} - Y_{g1p} \qquad \forall g \in \mathcal{G},\ p = 2,\ \ldots,\ P \tag{3.8}$$

$$W_{glp} = W_{gl(p-1)} + X_{glp} - Y_{glp} \qquad \forall g \in \mathcal{G},\ \forall l \geq 2,\ p = 2, ...,\ P \tag{3.9}$$

$$\sum_{(g,l)\in\mathcal{LK}(k)} \alpha_{gl}\, Y_{glp} \leq C_{kp} \qquad \forall k \in \mathcal{K},\ p = 1,\ \ldots,\ P \tag{3.10}$$

$$W_{glp},\ Y_{glp},\ X_{glp} \geq 0 \qquad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ p = 1,\ \ldots,\ P \tag{3.11}$$

The objective function (3.4) ensures that the Work-In-Process at each operation at the end of each period is minimized. Constraints (3.5) tie consecutive operations. Constraints (3.6)-(3.9) are flow constraints linking the Work-In-Process of each product at each operation in each period with the quantity completed in period p (Y variables) and the quantity arriving in period p (X variables). Constraints (3.10) are aggregate resource capacity constraints.

Global Scheduling Model With Work-In-Process Balancing Control

We are now including Work-In-Process balancing control in the global scheduling model. In the new objective function in (3.12), the total deviation of the Work-In-Process of each product is minimized. The penalty μ controls the Work-In-Process balancing strategy, by penalizing the maximal deviation on the Work-In-Process of each product. The maximum balancing deviation Z_{gp} of product g in period p ensures the feasibility of the solution if some products can be completed while others are still in the factory.

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{P} W_{glp} + \mu \sum_{g\in\mathcal{G}} \sum_{p=1}^{P} Z_{gp} \tag{3.12}$$

The Linear programming model denoted $P^{control}_{WIP}$ corresponds to Constraints (3.5)-(3.11), Constraints (3.13)-(3.15) and the objective function (3.12). $P^{control}_{WIP}$ implements the global scheduling strategy with Work-In-Process balancing control. Constraints (3.13) and (3.14) balance the Work-In-Process between products depending on the balancing coefficients δ_g. They ensure that the Work-In-Process in route g at the end of each period cannot be larger

than a percentage of the current total Work-In-Process in the factory. The balancing coefficient δ_g can be set to its trivial lower bound, 100% divided by the number of products if all products have the same demand, or can be set according to the percentage of each product in the set of products considered. In our computational experiments, δ_g is computed based on the release scheme of product in the factory, the estimated throughput and the Little's law.

$$\sum_{l \in \mathcal{L}_g} W_{glp} \;\leq\; \delta_g \sum_{g' \in \mathcal{G}} \sum_{l \in \mathcal{L}_{g'}} W_{g'lp} \;+\; Z_{gp} \quad \forall g \in \mathcal{G},\ \forall p \tag{3.13}$$

$$Z_{gp} \;\leq\; (1 \;-\; \delta_g) \sum_{g' \in \mathcal{G}} \sum_{l \in \mathcal{L}_{g'}} W_{g'lp} \quad \forall g \in \mathcal{G},\ \forall p \tag{3.14}$$

$$Z_{gp} \geq 0 \quad \forall g \in \mathcal{G},\ p = 1,\ \ldots,\ P \tag{3.15}$$

3.2.2 Computational Experiments

Computational results are provided in this section to analyze the performance of the Work-In-Process balancing strategy and its impact on cycle times and throughput. Numerous tests have been conducted on industrial data. The instances include 570 machines in 329 work centers, which are shared between operations of various types of products. Products have between 104 and 315 operations in their routes. Five products are considered and lots are continuously released in the factory in a uniform scheme, i.e., one lot for each product every 280 minutes, 360 minutes, 480 minutes, 480 minutes and 480 minutes for products 1, 2, 3, 4 and 5, respectively. As the ideal situation is to have $Z_{gp} = 0$, a sufficiently high penalty is necessary so that Z_{gp} tends to zero. The balancing penalty μ is fixed to 60,000, which is large enough to penalize Z_{gp}.

In the computational experiments carried out in this thesis, when the simulation model is run without the global scheduling model, the FIFO (First-In-First-Out) rule is the only dispatching rule used in the simulation. When the simulation model is run with the global scheduling model, an additional rule to the FIFO rule, called *Production Target Dispatching Rule* (PTDR), is used to track the production targets. PTDR first allows the processing of quantities of products that do not meet their production targets. This means that, if product g reaches its target quantity, then its production is temporarily stopped to reach the targets of other products. The production of product g only resumes when all products have reached their respective targets or when product g is the only one in the queue of a resource.

In the computational experiments discussed in this section, the performance of the global scheduling approach with and without balancing control on the output variability of cycle times and throughput are compared. The global scheduling approach uses balancing coefficients that are determined based on the release scheme of products in the factory. Next, numerical results on the use of the balancing coefficients to speed up products are discussed. Finally, the satisfaction of the estimated throughput and cycle times is studied. Here, the determination of balancing coefficients are based on the estimated throughput, the throughput from simulation and Little's law.

Comparing Global Scheduling Approach With and Without Balancing Control

This section presents results and analysis of the Work-In-Process management based on the balancing coefficients. Table 3.2 shows the remaining Work-In-Process in the global

	Remaining Work-In-Process in global scheduling model					
Products	75^{th} call		85^{th} call		95^{th}call	
	WIP	Percentage	WIP	Percentage	WIP	Percentage
1	103	14.8%	125	16.0%	151	16.5%
2	118	17.0%	136	17.4%	162	17.7%
3	249	35.8%	273	34.9%	361	39.4%
4	218	31.4%	195	24.9%	178	19.4%
5	7	1.0%	53	6.8%	64	7.0%
Total	695	100%	782	100%	916	100%

Table 3.2: Global scheduling model without Work-In-Process balancing, remaining WIP

scheduling model for each product at the 75^{th}, 85^{th} and 95^{th} calls of the global scheduling model without Work-In-Process balancing control. Table 3.3 shows the remaining Work-In-Process in the global scheduling model for each product at the 75^{th}, 85^{th} and 95^{th} calls of the global scheduling model when using Work-In-Process balancing control. In Table 3.2, Work-In-process is not controlled by the balancing coefficients. However, in Table 3.3, the Work-In-Process of each product is much better controlled. The flow of products depends on its associated balancing coefficients. Note that the smaller the balancing coefficient of

		Remaining Work-In-Process in global scheduling model					
Products	δ_g	75^{th} call		85^{th} call		95^{th}call	
		WIP	Percentage	WIP	Percentage	WIP	Percentage
1	45%	268	30.0%	295	31.0%	340	34.5%
2	16%	181	20.3%	196	20.6%	183	18.5%
3	13%	147	16.5%	160	16.8%	165	16.7%
4	13%	150	16.8%	150	15.8%	149	15.1%
5	13%	147	16.5%	150	15.8%	149	15.1%
Total	100%	893	100%	951	100%	986	100%

Table 3.3: Global scheduling model with Work-In-Process balancing, remaining WIP

a product, the higher the speed of the product's flow. But, in some particular cases, it can be difficult to slow down a product with a very small number of operations or to speed up a product with a large number of operations. Table 3.3 shows the flows of Products 3, 4 and 5 at almost the same rate based on their balancing coefficients.

Throughout this chapter, the Percentage Throughput Achieved for each product is calculated on the basis of the estimated throughput $\hat{S}_g$. The output variability is evaluated using the InterQuartile Range (IQR) measure.

	Products				
	1	2	3	4	5
Average Cycle Time (days)	63.0	63.2	64.5	50.9	50.6
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,398	1,097	816	868	870
Percentage Throughput Achieved	98.4%	99.3%	98.4%	104.7%	104.9%
Product variability IQR (Cycle Time)	25.2	22.9	23.9	19.6	19.8
Overall variability IQR (Cycle Time)	**13.1**				

Table 3.4: Simulation without global scheduling approach.

	Products				
	1	2	3	4	5
Average Cycle Time (days)	52.5	53.9	79.8	30.5	65.9
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,474	1,157	749	1050	803
Percentage Throughput Achieved	103.8%	104.7%	90.3%	126.6%	96.9%
Product variability IQR (Cycle Time)	20.0	15.4	33.7	4.9	25.1
Overall variability IQR (Cycle Time)	**31.3**				

Table 3.5: Global scheduling approach without Work-In-Process balancing control

	Products				
	1	2	3	4	5
Balancing coefficients δ_g	45%	16%	13%	13%	13%
Average Cycle Time (days)	65.1	54.4	53.4	49.9	51.8
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,380	1,142	858	884	855
Percentage Throughput Achieved	97.2%	103.3%	103.4%	106.6%	103.1%
Product variability IQR (Cycle Time)	23.3	21.9	19.4	18.5	20.2
Overall variability IQR (Cycle Time)	**8.9**				

Table 3.6: Global scheduling approach with Work-In-Process balancing control

The comparison of the results in Tables 3.6 and 3.5 shows the benefit on the output variability of using balancing coefficients to control Work-In-process. The results of the simulation model coupled with the global scheduling model with (Table 3.6) and without (Table 3.5) Work-In-Process balancing control are provided as well as the results of the simulation model only, i.e., without the global scheduling approach (Table 3.4). The global scheduling approach using the Work-In-Process balancing control leads to the best output variability. This is demonstrated by the product cycle time variability (variability between cycle times of completed quantities of each product), see Figure 3.2 and the overall variability (variability between products in terms of average cycle times), see Tables 3.4, 3.5 and 3.6. The overall variability (equal to 13.1) when the simulation is used without the global scheduling approach significantly increases (equal to 31.3) with the global scheduling approach with no WIP balancing control (model $P_{NoWIP}^{control}$). However, when there is a WIP balancing control in the global scheduling model (model $P_{WIP}^{control}$), the overall variability significantly decreases (equal to 8.9).

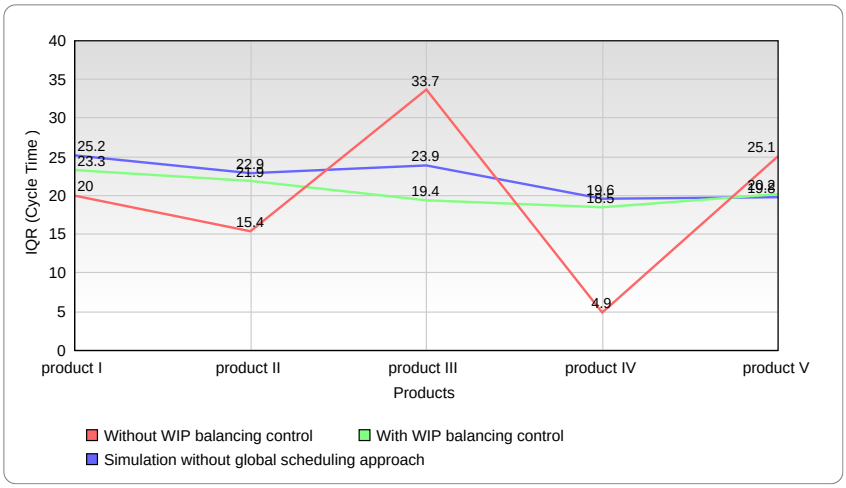


Figure 3.2: Product InterQuartile Ranges (Cycle Times).

Speeding up Products using Balancing Coefficients

To speed up a product, its balancing coefficient should be set to a small value. A small value of the balancing coefficient indicates that the product should have a small remaining Work-In-Process in the fab at the end of the scheduling horizon in the global scheduling model. Tables 3.7 and 3.8 show the remaining Work-In-Process for each product at the 75^{th} , 85^{th} and 95^{th} calls of the global scheduling model for different values of the balancing coefficients. In Table 3.7, Products 1 and 2 are accelerated by changing their balancing coefficients respectively from 45% to 10% for Product 1 and from 16% to 15% for Product 2. In Table 3.8, Products 2 and 3 are accelerated by changing their balancing coefficients respectively from 15% to 14% for Product 2 and from 25% to 14% for Product 3. The impact of the Work-In-Process balancing strategy on the final results is shown in Table 3.9 as well as in Table 3.10 after imposing production targets from the global scheduling model as constraints in the simulation model.

A product can be accelerated using its Work-In-Process balancing coefficient as shown in Table 3.9, and the contribution of each product on the overall throughput follows the

Products	δ_g	Remaining Work-In-Process in global scheduling model					
		75^{th} call		85^{th} call		95^{th}call	
		WIP	Percentage	WIP	Percentage	WIP	Percentage
1	**10%**	149	12.5%	153	12.8%	158	12.8%
2	**15%**	185	15.5%	188	15.6%	192	15.5%
3	25%	309	25.9%	314	26%	320	25.9%
4	25%	272	22.9%	275	22.8%	280	22.7%
5	25%	279	23.4%	278	23.0%	284	23.0%
Total	100%	1194	100%	1208	100%	1234	100%

Table 3.7: Speeding up Products 1 and 2.

Products	δ_g	Remaining Work-In-Process in global scheduling model					
		75^{th} call		85^{th} call		95^{th}call	
		WIP	Percentage	WIP	Percentage	WIP	Percentage
1	32%	373	32.0%	382	32.0%	388	32.3%
2	**14%**	163	14.0%	167	14.0%	170	14.1%
3	**14%**	163	14.0%	167	14.0%	170	14.1%
4	20%	233	20.0%	239	20.0%	242	20.1%
5	20%	233	20.0%	239	20.0%	232	19.3%
Total	100%	737	100%	872	100%	890	100%

Table 3.8: Speeding up Products 2 and 3

	Products				
	1	2	3	4	5
Balancing coefficients δ_g	10%	15%	25%	25%	25%
Average Cycle Time (days)	**52.4**	**54.4**	76.9	51.0	50.5
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,475	1,153	760	867	870
Percentage Throughput Achieved	**103.8%**	**104.3%**	91.7%	104.5%	104.9%
Weighted Total Average Cycle Time	55.9				
Total Throughput	5,125				

Table 3.9: Impact of balancing coefficients on cycle time and throughput, speeding up products 1 and 2

balancing coefficients. In comparison with the results presented in Table 3.6, Table 3.9 shows the acceleration of Products 1 and 2 due to their small Work-In-Process balancing coefficients while other products are slowed down to an acceptable level. Table 3.10 shows how Products 2 and 3 are accelerated while other products are slowed down due to their larger balancing coefficients. Compared to the results in Table 3.9, Product 2 reaches 105.3% of achieved throughput and Product 3 reaches 107.2% of achieved throughput, while Product 1 is slowed down with a change of its balancing coefficient from 10% to 32% but reaches 102.9% of achieved throughput. Note also that there is a limit on the acceleration of a

	Products				
	1	2	3	4	5
Balancing coefficients δ_g	32%	14%	14%	20%	20%
Average Cycle Time (days)	55.9	**53.8**	**47.3**	54.2	53.9
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,462	1,164	889	856	854
Percentage Throughput Achieved	102.9%	**105.3%**	**107.2%**	103.2%	103.0%
Weighted Total Average Cycle Time	53.3				
Total Throughput	5,229				

Table 3.10: Impact of balancing coefficients on cycle time and throughput, speeding up products 2 and 3

product, which is due to the fact that the product is very slow (long process times) or has more operations. Slowing down a product is also limited because the product is very fast (short process times) or has fewer operations. This is illustrated on Product 1 where, even with a balancing coefficient $\delta_g = 32\%$, a throughput of 102.9% is still achieved with a cycle time of 55.9 days.

Throughput and Cycle Time Satisfaction using Balancing Coefficients

These experiments analyze the satisfaction of a given throughput and the satisfaction of the cycle time targets. In general, the throughput and cycle times are provided by historical data or simulation. The estimated throughput for each product is then converted into balancing coefficients, which are used in the global scheduling model. Table 3.11 presents the Work-In-Process balancing coefficients for each product. The results in Tables 3.12 and 3.13 are based on the balancing coefficients computed using the estimated throughput. Table 3.12

Products	1	2	3	4	5
Estimated throughput (lots)	1420	1105	829	829	829
Balancing coefficients	18%	19%	21%	21%	21%

Table 3.11: Estimated throughput and balancing coefficients

presents the results of the satisfaction of the throughput using the balancing coefficients. The throughput achieved for each product reaches 100% except product 3 as shown in Table 3.12. To reach 100% satisfaction of throughput for product 3, balancing coefficients are updated using Algorithm 3.1. To reach 100% of estimated throughput, product 3 needs an additional

	Products				
	1	2	3	4	5
Balancing coefficients δ_g	18%	19%	21%	21%	21%
Average Cycle Time (days)	53.6	57.0	70.4	54.5	53.6
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,463	1,131	790	856	853
Percentage Throughput Achieved	103.0%	102.3%	**95.3**%	103.2%	102.8%

Table 3.12: Throughput satisfaction using balancing coefficients

4.7% on throughput. The sum of the percentages greater than 100% for products which have exceeded their estimated throughput (products 1, 2, 4 and 5) is calculated. Then, the contribution of each of these products to this calculated sum is determined. These contributions provide the percentages on how the 4.7% to be reduced to the balancing coefficient of product 3 (the percentage remaining for product 3 to reach 100% satisfaction of the throughput achieved) will be added to the balancing coefficients of products 1, 2, 4 and 5. The balancing coefficients are updated as indicated in the head of Table 3.13. All the products reach at least 100% of the throughput achieved, but the cycle times for products 4 and 5 are greatly increased compared to the cycle times of the simulation model without the global scheduling approach presented in Table 3.4. Table 3.14 presents the Work-In-Process

	Products				
	1	2	3	4	5
Balancing coefficients δ_g	**19.24%**	**19.96%**	**16.30%**	**22.33%**	**22.16%**
Average Cycle Time (days)	55.5	60.3	57.2	**62.2**	**62.0**
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,445	1,111	848	829	830
Percentage Throughput Achieved	**101.7%**	**100.5%**	**102.2%**	**100.0%**	**100.1%**

Table 3.13: Throughput satisfaction after adjusting balancing coefficients

balancing coefficients calculated based on the Little's law. Using Little's law, the goal is to take into account in the calculation of the balancing coefficients, not only the throughput but also the cycle times of the products. Table 3.15 shows that the throughput and the cycle

Products	1	2	3	4	5
Average Cycle Times (days)	63.0	63.2	64.5	50.9	50.6
Throughput (lots)	1398	1097	816	868	870
Balancing coefficients	29.5%	23.2%	17.6%	14.8%	14.8%

Table 3.14: Balancing coefficients computed based on Little's law

times are satisfied. The cycle times of products 4 and 5 are better managed compared with the results in Table 3.13. In addition, the cycle time and the throughput for all products are improved compared with the results in Table 3.14 of the simulation model without the global scheduling approach. Moreover, the cycle time of product 3 is the largest observed

	Products				
	1	2	3	4	5
Balancing coefficients δ_g	29.5%	23.2%	17.6%	14.8%	14.8%
Average Cycle Time (days)	**55.8**	**52.2**	**58.9**	**48.9**	**48.3**
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,451	1,144	841	879	881
Percentage Throughput Achieved	**103.8%**	**104.3%**	**103.1%**	**101.3%**	**101.3%**

Table 3.15: Throughput and cycle time satisfaction

cycle time as shown in Table 3.15. In order to improve the cycle time of product 3, we compute the new balancing coefficients as shown in Table 3.16. The cycle time of product 3 is decreased by 15% and its throughput is increased by 3.4%. The throughput and cycle time of other products in Table 3.15 are not changed. As shown in Table 3.16, with the new computed balancing coefficients, reducing the cycle time of product 3 deteriorates in an acceptable rate the cycle times of other products. This is due to the effect of resource sharing.

Products	1	2	3	4	5
Average Cycle Times (days)	55.8	52.2	**50.0**	48.9	48.3
Throughput (lots)	1451	1144	**870**	879	881
Balancing coefficients	30.0%	22.1%	16.1%	15.9%	15.7%

Table 3.16: Balancing coefficients computed based on Little's law, reduction of cycle time of product 3

	Products				
	1	2	3	4	5
Balancing coefficients δ_g	30.0%	22.1%	16.1%	15.9%	15.7%
Average Cycle Time (days)	56.3	54.8	**54.7**	50.9	50.3
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,447	1,160	**853**	866	870
Percentage Throughput Achieved	99.7%	101.4%	**98.0%**	98.5%	98.8%

Table 3.17: Improving throughput and cycle time of Product 3

Compared with the results in Table 3.15, Table 3.17 shows a decrease of 7.1% and an increase of 1.4% respectively for the cycle time and the throughput of product 3. The maximal deterioration for other products are of 4.9% and 1.4% respectively on the cycle times and the throughput. The cycle times and the throughput of all the products are further improved compared with the results of the simulation model without global scheduling strategy (see Table 3.14).

3.3 Multi-objective Optimization for Work-In-Process Balancing and Throughput Maximization

Most of the problems we face in real life often multi-objective. However, in mathematical optimization, these problems are not always easy to solve because the objectives considered can be in contradiction with each other. In multi-objective optimization, the goal is no longer to find the optimal solution to each objective, but rather to find compromise solutions. In most cases, and to reduce the complexity of multi-objective problems, the tendency is to transform these problems into single-objective problems. Several approaches have been studied in the literature of multi-objective optimization, depending on whether it is possible to significantly quantify and compare the value of one objective with respect to another or not. If this comparison can take place, the method such that the use of the weighted sum which combines multiple objectives into a single objective can be applied. When it is difficult to compare significantly and quantitatively the value of different objectives, a lexicographic approach for example can be applied (Rentmeesters et al. (1996)). Other approaches have been widely studied in the literature such as goal programming approach, desirability functions, ϵ-constraint, etc. Each approach has advantages and drawbacks and the choice largely depends on the problem which one needs to study. The choice can also be based on the compromises between objectives. For more details about multi-objective approaches, see T'kindt and Billaut (2006), Deb (2001) and Ehrgott (2005).

In this section, a multi-objective optimization strategy for global fab scheduling is formulated using an ϵ-constraint approach and its adjusted version. Before presenting the multi-objective model and the ϵ-constraint approach, we present the lexicographic approach especially since it structures the ϵ-constraint approach.

3.3.1 Lexicographic approach

In the lexicographic approach, the concept of preference is crucial and must be defined before the resolution. The objectives must be ranked according to their importance or significance. Then, single-objective problems are successively solved starting with the most important objective. After the first iteration, the problem is solved for the second most important objective, by adding a constraint specifying that the first objective must be equal to its optimal value. The procedure continues as shown in (3.16). For more details see (Marler and Arora (2004))

$$\min_{x \in X} F_i(x) \tag{3.16}$$

Subject to :

$$F_j(x) \leq F_j^*, \quad j = 1, 2, ..., i-1,$$

Where, i is a function's position in the ordering sequence, and F_j^* is the optimal objective function of the j^{th} objective function, found in the j^{th} iteration. Recall that a multi-objective problem studied in this section the first objective consists of maximizing the throughput (number of produced wafers) using productivity (outputs of operations). This is achieved using a Work-In-Process management technique based on pull strategy. The second objective consists of minimizing the output variability. This is achieved using balancing penalty in

the objective function and smoothing constraints. Maximizing throughput is considered the most important objective, hence it will be optimized first.

A lexicographic approach is not appropriate for our problem because maximizing productivity does not necessarily mean maximizing the total throughput of the factory. Maximum productivity can be achieved with low throughput. This is because the number of operations in the route is not the same for each product. In a lexicographic approach, the second stage, which consists in minimizing the output variability on cycle times with the throughput constrained by a maximum productivity, may not lead to the highest total throughput of the factory. Since the maximum total throughput of the factory is obtained with a particular combination of the throughput of the products, an ϵ-constraint approach seems to be more relevant for our problem. When solving the second objective, which minimizes the output variability of completed products, instead of setting the first objective at the maximum productivity, an ϵ-constraint approach reduces productivity if necessary to find combinations of throughput of products that lead to the highest total throughput of the factory while minimizing the output variability on cycle times and throughput.

The multi-objective global scheduling strategy, as well its implementation as a mathematical programming model, is presented in Section 3.3.2. Section 3.3.3 presents the ϵ-constraint approach and its adjusted version. Finally, in Section 3.3.4, computational results and an analysis are provided.

3.3.2 Multi-objective Global Scheduling Model

The multi-objective global scheduling strategy is described in this section. Different strategies are implemented in the objective function. A pull strategy is considered for maximizing the outputs of operations, objective function f_1 (production quantities of each product completing an operation in a period), and a Work-In-Process balancing control strategy to minimize the output variability on cycle times and throughput, objective function f_2.

$$f_1 = Max \sum_{g \in \mathcal{G}} \sum_{l \in \mathcal{L}_g} \sum_{p=1}^{P} q_{glp}\, Y_{glp} \tag{3.17}$$

$$f_2 = Min \sum_{g \in \mathcal{G}} \sum_{l \in \mathcal{L}_g} \sum_{p=1}^{P} W_{glp} + \mu \sum_{g \in \mathcal{G}} Z_g \tag{3.18}$$

Subject to :

$$X_{glp} = Y_{g(l-1)p} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ \forall p \tag{3.19}$$

$$W_{g11} = IW_{g1} + R_{g1} - Y_{g11} \quad \forall g \in \mathcal{G} \tag{3.20}$$

$$W_{gl1} = IW_{gl} - Y_{gl1} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2 \tag{3.21}$$

$$W_{g1p} = W_{g1(p-1)} + R_{gp} - Y_{g1p} \quad \forall g \in \mathcal{G},\ p = 2,\ \ldots,\ P \tag{3.22}$$

$$W_{glp} = W_{gl(p-1)} + X_{glp} - Y_{glp} \quad \forall g \in \mathcal{G},\ \forall l \geq 2,\ p = 2, ...,\ P \tag{3.23}$$

$$\sum_{l \in \mathcal{L}_g} W_{glp} \leq \delta_g \sum_{g' \in \mathcal{G}} \sum_{l \in \mathcal{L}_{g'}} W_{g'lp} + Z_g \quad \forall g \in \mathcal{G},\ \forall p \tag{3.24}$$

$$Z_g \leq (1 - \delta_g) \sum_{g' \in \mathcal{G}} \sum_{l \in \mathcal{L}_{g'}} W_{g'lp} \quad \forall g \in \mathcal{G},\ \forall p \tag{3.25}$$

$$\sum_{(g,l)\in\mathcal{LK}(k)} \alpha_{gl}\ Y_{glp} \ \leq\ C_{kp} \quad \forall k \in \mathcal{K},\ p = 1,\ \ldots,\ P \tag{3.26}$$

$$W_{glp},\ Y_{glp},\ X_{glp},\ Z_g \geq 0 \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ p = 1,\ \ldots,\ P \tag{3.27}$$

Constraints (3.19) tie consecutive operations. Constraints (3.20)-(3.23) are flow constraints linking the Work-In-Process of each product at each operation in each period with the quantity completed in period p (Y variables) and the quantity arriving in period p (X variables). Constraints (3.24) and (3.25) ensure that the Work-In-Process of product g at each period cannot be larger than $\delta_g\%$ of the current total Work-In-Process in the factory. In our computational experiments, δ_g is set up to its trivial lower bound, 100% divided by the number of products if all products have the same production quantity. Else, δ_g depends on the percentage of the quantity of each product in the set of products considered or on the release scheme. Constraints (3.26) are resource capacity constraints. Let us now discuss the objective functions f_1 and f_2. Objective function f_1 (3.17) maximizes the throughput by maximizing productivity. Let LB be the number of operations starting from the last operation and going backwards to the minimum number of operations from which products are pulled out of the factory to ensure that they are leaving the factory as soon as possible, i.e.,

$$LB = \min_{g\in G} |\mathcal{L}_g|,$$

Also, let UB be the maximum number of operations in the routes of all products, i.e.

$$UB = \max_{g\in G} |\mathcal{L}_g|,$$

UB is decreased backward on the operations of each product as shown in Figure 3.3. The goal is to ensure that products become more important as they are moving in the factory. Quantities of product that are in the end of their production are produced as fast as possible to increase the total throughput of the factory. This describes the pull strategy used in Objective function f_1 to speed up products in their LB last operations, see Figure 3.3. This strategy ensures that products with short cycle times and/or products with fewer operations leave the factory as quickly as possible. The larger LB, the broader the possibility to pull a large amount of Work-In-Process, i.e., to speed up the outputs.

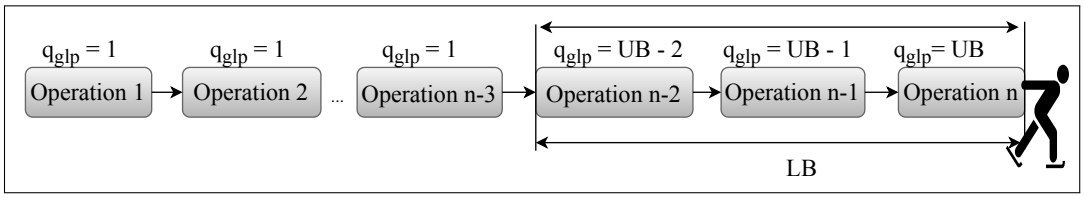


Figure 3.3: Pull strategy on LB last operations

Objective function f_2 (3.18) minimizes the total deviation of the Work-In-Process of each product by using a Work-In-Process balancing control strategy. It also ensures that the Work-In-Process at each operation at the end of each period is minimized. The parameter μ_b controls the balancing strategy, by penalizing the maximal deviation on the Work-In-Process for each product. The maximum balancing deviation Z_g of product g ensures the feasibility of the solution if some products can be completed while others are still in the factory.

3.3.3 ϵ-constraint approach

The ϵ-constraint approach is widely used in the literature to solve problems with multiple objectives. It considers the minimization/maximization of one objective function while setting other objective functions in constraints, see Haimes (1971). In Miettinen (1999), it is demonstrated that, if the solution to the ϵ-constraint approach exists, it is always a weakly Pareto optimal solution. A solution s is weakly Pareto optimal if there is no other solution m such that f $(m) <$ f (s) in minimization. By solving problems for different values of ϵ, a set of weakly Pareto optimal solutions is determined (see Miettinen (1999)). The methodology used in this section includes two main stages using the ϵ-constraint approach.

1. In the first stage, problem (M_1) which consists of objective function f_1 in (3.17) subject to Constraints (3.19) - (3.27), maximizes the throughput by maximizing the productivity. Assume that f_1^* is the optimal objective value of (M_1). f_1^* is not necessarily the optimal overall throughput because all products share the same resources and the maximum throughput is obtained with a particular combination of percentages of throughputs for the products.

2. In the second stage, problem (M_2), which consists of objective function f_2 in (3.18) subject to Constraints (3.19) - (3.27) and (3.28) below, is solved for different values of ϵ. This stage deteriorates f_1^* by some ϵ in order to find a solution that minimizes the Work-In-Process variability while maintaining a high productivity.

$$\sum_{g\in\mathcal{G}}\sum_{l\in\mathcal{L}_g}\sum_{p=1}^{P} q_{glp}\, Y_{glp} \;\geq\; f_1^* - \epsilon \tag{3.28}$$

Besides the advantage that it can be used either for convex objective spaces or non-convex objective spaces, one of the disadvantages of the ϵ-constraint approach is that the search space depends on the chosen values for ϵ, see Deb (2001). In our case, when ϵ becomes too large, the constraint on objective f_1 can easily be satisfied with a productivity that does not corresponds to the highest throughput.

To solve this problem, an adjusted version of the ϵ-constraint approach is proposed. The idea is to keep the information on the maximization of throughput in the objective function f_2. This is achieved by penalizing the objective function f_1 with λ, which is a value that can be chosen as small as possible. In this thesis we use $\lambda = 0.01$.

The same methodology used in the ϵ-constraint approach is applied in the adjusted version of ϵ-constraint approach, but the objective function f_2 in the second stage becomes:

$$f_2 = Min \sum_{g\in\mathcal{G}}\sum_{l\in\mathcal{L}_g}\sum_{p=1}^{P} \lambda\left(\frac{W_{glp}}{\lambda} - q_{glp}\, Y_{glp}\right) + \mu\sum_{g\in\mathcal{G}} Z_g \tag{3.29}$$

The throughput maximization included in objective function f_2 in (3.29) leads the search towards good throughput solutions without deteriorating too much the productivity while minimizing the output variability.

3.3.4 Computational Experiments

This section analyzes the results of the global multi-objective scheduling model coupled with the generic simulation model (with FIFO dispatching rules) for different values of ϵ. Recall that the value of ϵ is not a percentage of the overall throughput, but a percentage of the objective function f_1^* of problem (M_1) which maximizes productivity. Numerical experiments have been conducted on two industrial instances associated to two different factories presented in chapter 2 section 2.3.3. As in section 3.2.2, the output variability is evaluated using InterQuartile Range (IQR) and the balancing penalty μ is fixed to 60,000 which is large enough to penalize the WIP deviation.

ϵ-constraint approach: first instance

The results presented in Tables 3.18, 3.19, 3.20, 3.21 and 3.22 are based on the first factory instance. For different values of ϵ shown in the first column, Columns (2) to (6) present percentages of the achieved throughput for each product. In Column (7), the total throughput of the factory is given and Column (8), shows the output variability on cycle times computed using IQR. Five products are considered, lots are continuously released in the factory in a

	Products					Total	Cycle time
ϵ	1	2	3	4	5	Throughput	IQR
10%	102.6%	97.7%	99.9%	97.8%	95.7%	9,715	11.6
11%	102.0%	94.8%	99.7%	97.9%	95.2%	9,639	10.9
12%	103.7%	96.7%	99.8%	97.2%	95.8%	9,706	11.6
13%	104.2%	99.4%	98.1%	97.1%	97.5%	9,767	8.1
14%	101.6%	99.3%	101.4%	96.9%	94.5%	9,716	10.9
15%	102.8%	99.4%	101.8%	97.6%	94.4%	9,762	10.2
16%	102.6%	97.7%	101.2%	96.8%	96.0%	9,731	11.0
17%	104.9%	98.7%	99.8%	96.3%	97.2%	9,779	10.0
18%	102.6%	98.7%	101.7%	95.6%	97.8%	9,768	11.5
19%	101.1%	101.1%	100.8%	95.2%	97.3%	9,749	11.4
20%	101.9%	101.6%	99.3%	96.7%	95.6%	9,741	11.0
21%	104.3%	98.3%	102.5%	96.3%	95.5%	9,775	10.4
22%	99.9%	102.7%	102.8%	95.8%	99.0%	9,843	11.3
23%	101.6%	101.8%	100.9%	94.2%	94.9%	9,705	10.5
24%	102.2%	97.8%	99.9%	95.3%	97.7%	9,700	11.1
25%	99.6%	100.2%	100.8%	97.2%	96.2%	9,720	10.8
26%	104.8%	96.4%	103.5%	94.8%	97.9%	9,783	11.5
27%	**101.3%**	**101.0%**	**103.1%**	**98.2%**	**100.0%**	**9,910**	**7.2**
28%	101.2%	103.5%	102.7%	96.8%	98.4%	9,889	10.7
29%	103.0%	99.0%	102.9%	97.3%	97.0%	9,817	10.5
30%	104.4%	99.0%	100.8%	95.8%	98.2%	9,804	10.7

Table 3.18: Throughput and variability for different values of ϵ, ϵ-constraint approach, instance 1

uniform scheme, i.e., five lots (1 lot for each product) are released every 202 minutes. The balancing coefficient δ_g is fixed to 20% for all products. Decision makers may choose $\epsilon =$

27%, which increases the throughput percentages for all products and leads to the largest overall throughput. Tables 3.19, 3.20 and 3.21 help to compare the case with $\epsilon = 27\%$ to the cases when the linear programming model is solved using Objective function f_2 only and when the simulation model is only run with the FIFO dispatching rules, i.e., without the global scheduling approach. The results for $\epsilon = 27\%$ dominate in terms of total throughput (9,910) and weighted average cycle time. Moreover, and probably more interesting, the average cycle time and the average throughput are much better balanced between products for $\epsilon = 27\%$.

	Products				
Indicators	1	2	3	4	5
Average Cycle Time (days)	57.1	62.8	56.8	62.4	65.5
Release Quantities	2,566	2,566	2,566	2,566	2,566
Throughput	1993	1988	2,029	1,932	1,968
Percentage Achieved of Throughput	**101.3%**	**101.0%**	**103.1%**	**98.2%**	**100.0%**
Weighted Average Cycle Time	60.0				
Total Throughput	**9,910**				

Table 3.19: Multi-objective global scheduling approach with $\epsilon = 27\%$

	Products				
Indicators	1	2	3	4	5
Average Cycle Time (days)	59.0	63.2	57.0	65.1	65.0
Release Quantities	2,566	2,566	2,566	2,566	2,566
Throughput	2,058	1,914	1,946	1,901	1,915
Percentage Achieved of Throughput	**104.6%**	**97.3%**	**98.9%**	**96.6%**	**97.3%**
Weighted Average Cycle Time	61.1				
Total Throughput	**9,734**				

Table 3.20: Global scheduling approach with Objective function f_2 only

	Products				
Indicators	1	2	3	4	5
Average Cycle Time (days)	54.3	66.0	50.9	73.3	59.7
Release Quantities	2,566	2,566	2,566	2,566	2,566
Throughput	2,030	1,914	2,060	1,846	1,980
Percentage Achieved of Throughput	**103.2%**	**97.3%**	**104.7%**	**93.8%**	**100.6%**
Weighted Average Cycle Time	60.6				
Total Throughput	**9,830**				

Table 3.21: Simulation (FIFO dispatching rules) without global scheduling approach

Table 3.22 provides the measure of the mix product variability on the cycle time using the InterQuartile Range (IQR). Table 3.22 shows that the simulation model without the global scheduling approach, where only the FIFO dispatching rules are used, leads to the highest output variability on the cycle time.

Indicators	Simulation only	Global scheduling approach only with f_2	Multi-objective with $\epsilon = 27\%$
IQR (Cycle Time)	17.1	7.0	**7.2**

Table 3.22: Variability measure on average cycle time based on IQR

ϵ-constraint: second instance

Tables 3.23, 3.24, 3.25, 3.26, and 3.27 present the results on the second instance.

	Products					Total	Cycle time
ϵ	1	2	3	4	5	Throughput	IQR
10%	97.8%	105.0%	102.4%	108.2%	102.8%	5,151	12.5
11%	97.9%	104.8%	104.3%	105.0%	103.8%	5,147	12.8
12%	97.9%	105.1%	104.4%	105.0%	104.5%	5,156	12.9
13%	98.3%	104.8%	103.3%	103.9%	105.5%	5,149	12.8
14%	97.4%	105.1%	103.8%	105.0%	104.5%	5,144	12.6
15%	97.5%	104.7%	103.8%	104.3%	105.6%	5,145	12.8
16%	96.5%	104.7%	104.3%	105.5%	104.2%	5,132	12.6
17%	97.2%	104.7%	104.3%	104.5%	105.0%	5,138	12.7
18%	97.6%	105.0%	104.2%	105.0%	104.1%	5,145	12.4
19%	97.9%	105.3%	102.0%	108.9%	103.0%	5,151	12.9
20%	**101.9%**	**105.2%**	**100.7%**	**111.5%**	**101.3%**	**5,211**	12.7
21%	98.2%	105.2%	100.9%	113.1%	99.7%	5,160	16.9
22%	97.1%	105.4%	100.8%	112.9%	99.7%	5,143	16.8
23%	101.7%	105.7%	101.2%	112.9%	99.6%	5,214	14.7
24%	101.1%	105.2%	99.7%	112.9%	100.7%	5,197	13.5
25%	100.3%	105.7%	100.2%	112.7%	100.8%	5,195	14.8
26%	100.1%	105.4%	100.6%	112.7%	100.0%	5,185	12.6
27%	98.5%	105.0%	100.7%	113.2%	100.1%	5,165	15.9
28%	101.9%	105.7%	99.0%	113.5%	100.9%	5,214	17.6
29%	98.3%	105.2%	100,6%	111.8%	101.2%	5,159	15.7
30%	98.0%	104.7%	100.3%	112.5%	100.7%	5,151	14.6

Table 3.23: Throughput and variability for different values of ϵ, ϵ-constraint approach, instance 2

Five products are considered, lots are continuously released in the factory in a uniform scheme, i.e., 1 lot is released every 280 minutes, 360 minutes, 480 minutes, 480 minutes and 480 minutes for products 1, 2, 3, 4 and 5 respectively. The balancing coefficients are fixed to 45%, 16%, 13%, 13% and 13% for products 1, 2, 3, 4 and 5 respectively. Several values of ϵ lead to good results as shown in Table 3.23. The choice depends essentially on the priorities of products and the importance of the variability and the total throughput. However, $\epsilon = 20\%$ seems to be the best choice for decision makers as it increases the throughput for all products and ensures a good overall throughput. Tables 3.24, 3.25, and 3.26 help to compare the case with $\epsilon = 20\%$ to the cases when the model is solved with Objective function f_2 only and when the simulation model is run with the FIFO dispatching rules, i.e., without the

global scheduling model. The results for $\epsilon = 20\%$ dominate in terms of overall throughput and total weighted average cycle time. Again, and importantly, the average cycle time and the average throughput are much better balanced between products for $\epsilon = 20\%$.

	Products				
Indicators	1	2	3	4	5
Average. Cycle Time (days)	61.4	53.8	58.0	40.2	56.5
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,448	1,163	835	925	840
Percentage Achieved of Throughput	**101.9%**	**105.2%**	**100.7%**	**111.5%**	**101.3%**
Weighted Average Cycle Time	54.6				
Total Throughput	**5,211**				

Table 3.24: Multi-objective global scheduling approach with $\epsilon = 20\%$

	Products				
Indicators	1	2	3	4	5
Average Cycle Time (days)	70.0	49.5	49.0	53.0	53.3
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,382	1,170	865	870	868
Percentage Achieved Throughput	**97.3%**	**105.8%**	**104.3%**	**104.9%**	**104.7%**
Weighted Average. Cycle Time	56.2				
Total Throughput	**5,155**				

Table 3.25: Global scheduling approach with Objective function f_2 only

Table 3.27 provides the measure of the output variability on the cycle time using the

	Products				
Indicators	1	2	3	4	5
Average Cycle Time (days)	63.0	63.2	64.5	50.9	50.6
Release Quantities	1,852	1,441	1,081	1,081	1,081
Throughput	1,398	1,097	816	868	870
Percentage Achieved Throughput	**98.4%**	**99.3%**	**98.4%**	**104.7%**	**104.9%**
Weighted Average Cycle Time	59.1				
Total Throughput	**5,049**				

Table 3.26: Simulation (FIFO dispatching rules) without global scheduling approach

IQR. It shows again, although less significantly than for the first industrial instance, that the simulation model without the global scheduling approach leads to the highest output variability on the cycle times.

Indicators	Simulation only	Global scheduling approach only with f_2	$\epsilon = 20\%$
IQR(Cycle Time)	13.1	12.3	**12.7**

Table 3.27: Variability measure on average cycle time based on IQR

Adjusted ϵ-constraint approach: first instance

Table 3.28, shows the results on the first industrial instance when the adjusted version of ϵ-constraint approach is used. The same instance characteristics are used as in section 3.3.4. In

	Products					Total	Cycle time
ϵ	1	2	3	4	5	Throughput	IQR
10%	112.3%	99.3%	97.9%	99.5%	87.6%	9,769	15.2
11%	110.0%	100.1%	95.6%	98.5%	90.3%	9,727	11.2
12%	109.2%	96.2%	98.6%	99.6%	91.2%	9,735	13.5
13%	108.3%	99.6%	96.5%	97.7%	92.9%	9,740	13.2
14%	110.2%	100.4%	95.3%	98.7%	91.1%	9,755	13.9
15%	108.5%	99.2%	98.9%	98.1%	91.8%	9,770	12.5
16%	109.3%	101.8%	97.8%	98.3%	90.5%	9,793	13.1
17%	113.9%	97.4%	99.3%	101.5%	89.1%	9,861	15.5
18%	109.7%	99.0%	95.0%	97.6%	92.2%	9,708	11.6
19%	113.6%	95.2%	100.4%	100.0%	88.3%	9,788	15.2
20%	114.1%	95.4%	96.6%	99.8%	87.5%	9,705	14.6
21%	112.9%	100.4%	94.3%	99.7%	87.2%	9,730	15.6
22%	110.4%	97.5%	99.4%	98.5%	89.5%	9,744	13.8
23%	111.5%	97.7%	97.4%	98.9%	91.4%	9,771	12.7
24%	113.0%	97.6%	97.7%	99.8%	88.9%	9,781	12.2
25%	107.6%	102.6%	98.7%	98.1%	89.2%	9,761	13.2
26%	110.6%	97.8%	97.4%	99.6%	89.9%	9,743	15.8
27%	110.3%	101.5%	98.3%	99.2%	87.7%	9,778	14.2
28%	**110.4%**	**100.1%**	**100.8%**	**100.1%**	**100.0%**	**10,061**	**10.3**
29%	114.0%	95.2%	98.0%	99.7%	89.7%	9,784	16.6
30%	111.8%	95.9%	101.9%	99.3%	88.3%	9,784	16.6

Table 3.28: Throughput and variability for different values of ϵ, adjusted ϵ-constraint approach, instance 1

Table 3.18, only 9 values of ϵ (out of 21) lead to a total throughput above the total throughput provided when the global scheduling approach is only used with objective function f_2 which minimizes the output variability. In Table 3.28, 16 values of ϵ (out of 21) lead to a total throughput higher than the one of the global scheduling approach when only the objective function f_2 is used. Another interesting aspect is that, in Table 3.18, not all products reach a throughput percentage greater or equal to 100%, but, in Table 3.28, all the products reach a throughput percentage greater than or equal to 100%. In addition, Table 3.28 presents the highest total throughput and at almost the same level of productivity deterioration (almost the same value of ϵ) as in Table 3.18.

Adjusted ϵ-constraint approach: second instance

This section presents and analyzes results in Table 3.29 based on the second industrial instance when the adjusted version of ϵ-constraint approach is used. The same instance

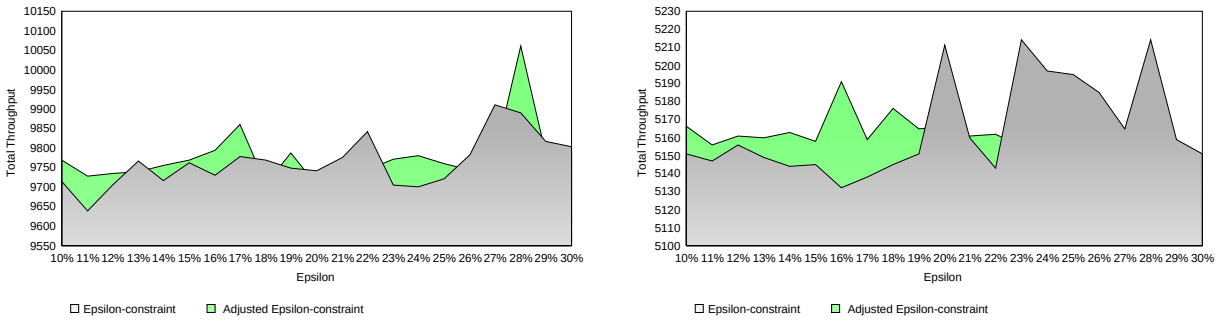
	Products					Total	Cycle time
ϵ	1	2	3	4	5	Throughput	IQR
10%	98.7%	101.5%	102.0%	110.4%	101.6%	5,166	15.1
11%	98.2%	105.1%	102.8%	104.0%	106.9%	5,156	12.4
12%	98.9%	104.6%	101.1%	104.9%	107.7%	5,161	12.7
13%	98.7%	104.9%	100.8%	102.4%	110.4%	5,160	14.4
14%	98.7%	105.1%	101.3%	102.0%	110.2%	5,163	13.2
15%	98.4%	104.8%	101.4%	102.9%	109.5%	5,158	13.6
16%	**101.1%**	**104.7%**	**102.3%**	**102.9%**	**108.2%**	**5,191**	**12.4**
17%	98.7%	105.0%	101.6%	102.3%	109.4%	5,159	12.5
18%	99.6%	105.1%	101.9%	102.6%	109.0%	5,176	12.6
19%	98.7%	104.9%	97.8%	106.7%	109.5%	5,165	16.6
20%	98.9%	105.1%	99.6%	108.4%	105.5%	5,166	15.4
21%	98.7%	104.9%	100.1%	109.4%	104.2%	5,161	13.3
22%	98.8%	104.8%	102.6%	105.9%	105.2%	5,162	12.5
23%	98.5%	105.0%	103.6%	105.5%	104.6%	5,155	12.3
24%	98.6%	104.7%	100.7%	108.6%	104.3%	5,157	14.7
25%	99.0%	104.7%	100.2%	107.5%	106.1%	5,165	12.8
26%	98.6%	104.9%	100.6%	108.3%	104.8%	5,160	13.5
27%	97.3%	105.2%	107.3%	102.9%	103.1%	5,143	8.2
28%	96.5%	105.2%	107.0%	103.6%	102.8%	5,131	9.5
29%	97.0%	105.7%	107,4%	102.8%	103.9%	5,149	8.5
30%	96.0%	104.9%	107.1%	103.1%	103.1%	5,120	10.9

Table 3.29: Throughput and variability for different values of ϵ, instance 2 adjusted ϵ-constraint approach

characteristics are used as in section 3.3.4. In Table 3.18, only 12 values of ϵ (out of 21) lead to a total throughput greater than the total throughput of the global scheduling approach when only the objective function f_2 is used. In Table 3.29, 16 values of ϵ (out of 21) lead to a total throughput higher than the one of the global scheduling approach when only the objective function f_2 is used. In addition, the results that increase the throughput for all the products are reached earlier at $\epsilon = 16\%$ compared to Table 3.18, $\epsilon = 20\%$.

The numerical results show in Figure 3.4 that the adjusted ϵ-constraint approach provides great compromises in terms of throughput and loss of the productivity.

The adjusted ϵ-constraint approach also increases the largest throughput by 1.5% for the first instance with almost the same deterioration in productivity. At $\epsilon = 17\%$, the throughput is larger than the throughput with the global scheduling approach and only objective function f_2 (with an additional throughput of 31 lots) and when the simulation runs without the global scheduling approach (with an additional throughput of 127 lots) and with a variability on cycle times (IQR = 15.5), which remains lower than the one of the simulation model without the global scheduling approach (IQR = 17.1). In the second industrial instance, the deterioration of the productivity is reduced to 16% with the adjusted version of ϵ-constraint



(a) Total throughput instance1

(b) Total throughput instance 2

Figure 3.4: Comparing ϵ-constraint and Adjusted ϵ-constraint approaches on total throughput

approach for its largest throughput instead of 20% with the ϵ-constraint approach. However, this improvement comes with a cost. The ϵ-constraint approach outperforms its adjusted version in terms of output variability on cycle times. The choice of the ϵ-constraint approach or its adjusted version essentially depends on the level of importance that decision makers give to the output variability, productivity and throughput. To summarize all the results, the ϵ-constraint approach opens the solutions space to interesting results in terms of variability, while its adjusted version opens the solutions space to very interesting results in terms of throughput.

For the first industrial instance, result for $\epsilon = 28\%$ provided by the adjusted ϵ-constraint approach seems to be a good compromise for decision makers. It increases the throughput for all the products and the overall throughput of the factory while maintaining an acceptable output variability (IQR = 10.3). This variability is lower than that provided by the simulation only (IQR = 17.1), where First-In-First-Out (FIFO) dispatching rules are used without the global scheduling approach.

For the second industrial instance, the output variability is almost the same for both approaches. If the objective is only to reach 100% of throughput achieved for each product, the result for $\epsilon = 16\%$ provided by the adjusted ϵ-constraint approach seems to be a good compromise for decision makers as it does not much deteriorate productivity.

3.4 Conclusions and Perspectives

This chapter presents strategies for combining Work-In-Process balancing and throughput maximization. The Work-In-Process balancing strategy is formulated as a Linear Programming model (global scheduling model). It aims to control the flow of products in order to minimize the output variability on cycle times and throughput, to speed up products and ensure satisfaction of throughput and cycle times. To achieve these objectives, the strategy is applied with smoothing constraints and a Work-In-Process balancing penalty in the objective function. Different methods based on the estimated throughput, Little's law, and the release quantities, etc. were used to determine the balancing coefficients.

The balancing coefficients calculated on the basis of Little's law provide good results on the satisfaction of cycle times and throughput, but they remain static on the horizon. These balancing coefficients do not integrate dynamic information such as the time the Work-In-

Process has already spent in the factory. Therefore, they are not robust enough to control the product cycle times in case of factory disruption. The control of cycle times through the entire production line using cycle time targets, release dates and temporal tracing of Work-In-Process, i.e., management by the time the Work-In-Process has already spent in the factory is provided in Chapter 4.

As the balancing coefficients remain static they only work well when the factory is in a steady state, but this is rarely the case in practice, due to machine breakdowns, critical preventative maintenance operations, etc. These balancing coefficients will not necessarily give the expected results in these situations. Thus, to handle this drawback, techniques for dynamically updating the balancing coefficients should be studied. This will ensure that the balancing coefficients are adjusted according to the state of the factory at each time when the global scheduling strategy is called.

Strategies to optimize the throughput and the output variability are formulated as a multi-objective programming model using an ϵ-constraint approach. The ϵ-constraint approach has been adjusted so that information on the maximization of throughput is included in the objective function, which minimizes the output variability. The goal is to determine solutions that will maintain high productivity while minimizing the output variability. The effectiveness of this strategy and the impact of the Work-In-Process balancing control have been demonstrated by computational results on industrial data.

It may also be important to integrate new objective functions such as capacity utilization in addition to the minimization of the output variability on cycle times and throughput and the maximization of the throughput in the multi-objective global scheduling strategy. Other approaches can also be studied such as goal programming or the desirability approaches whose results can be compared with those provided by the ϵ-constraint approach and its adjusted version.

Chapter 4

Cycle Time Minimization and Cycle Time Control

4.1 Introduction

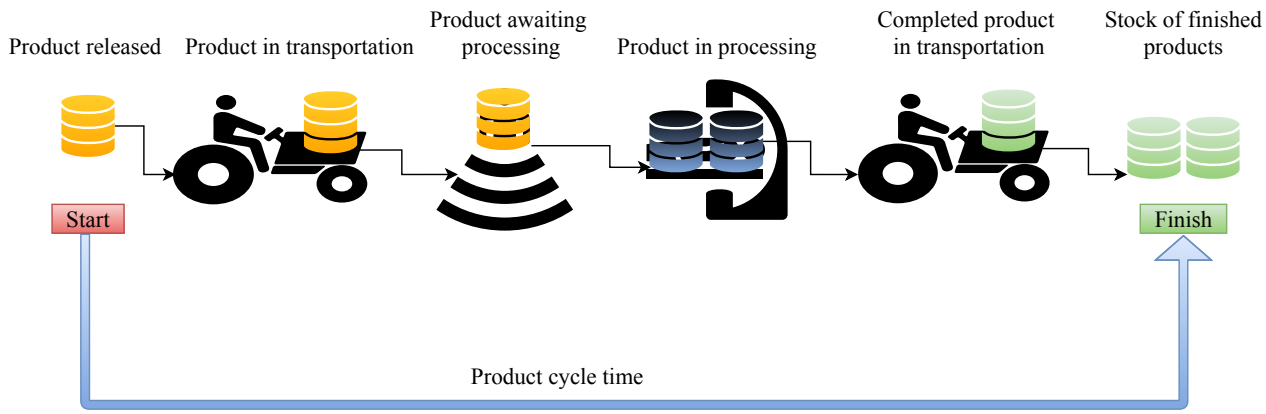


Figure 4.1: Cycle time of product

This chapter discusses different global scheduling strategies to minimize and control cycle times. The cycle time includes the transportation times, the time spent waiting in the queues of resources and the processing times, see Figure 4.1. In scheduling decisions, most of the criteria are derived from the product completion times, which are the main information for calculating cycle times, see Mati et al. (2011). Cycle time is one of the important key performance indicators in semiconductor manufacturing. The minimization and control of cycle times have an impact on several other metrics and key performance indicators such as throughput, yield, on-time delivery, etc. Short cycle times also help to reduce wafer risk contamination, yield loss and the inventory that should be maintained (Lu et al. (1994)).

To minimize cycle times, two strategies are proposed and compared:

1. A push strategy, where products are pushed forward to their last operations by using high Work-In-Process holding costs on the first operations of products.

2. A time at operation strategy, where quantities of Work-In-Process that arrived at different times in an operation are penalized differently, in order to prioritize the processing of Work-In-Process quantities that have spent more time in the operation.

The results of the simulation model without the global scheduling approach are compared to the results of the push strategy and the time at operation strategy. The computational results show that the time at operation strategy outperforms the push strategy, respectively the simulation without the global scheduling approach, with a decrease of 9 days, respectively 4.7 days, on the average cycle time and an increase of 781 lots, respectively 355 lots, for the throughput. The cycle times for all products with the time at operation strategy are lower than the cycle times of products when the simulation is used without the global scheduling approach. In addition, the average cycle time and the throughput are much better balanced between products for the time at operation strategy.

To control the cycle times, the principal levers used in this chapter are the product release dates, the cycle time targets of products used as inputs parameters and the temporal tracing of the Work-In-Process, i.e., the management of the Work-In-Process based on the time the Work-In-Process have already spent in the factory. Product operations in the route of each product are grouped in subsequences (blocks of operations), each subsequence with a cycle time target derived from the product's total cycle time target. The goal is to ensure that the time that quantities of products spend in each subsequence of operations is better controlled to minimize the deviation between the observed cycle time and the product cycle time target. Naive and simulation-based methods are proposed and compared using 12 and 50 blocks of operations. Results show that the simulation-based method outperforms the naive method. In addition, cycle times are better controlled when a large number of blocks of operations are used (50 blocks of operations in our experiments).

Section 4.2 presents a comparison between a push strategy and a time at operation strategy to minimize cycle times. In section 4.3, a novel global scheduling strategy to control cycle times is presented with a focus on the temporal tracing of the Work-In-process to better manage the cycle time of each product.

4.2 Cycle Time Minimization

Cycle time covers the life of a product in a factory, combining the value-added and non-value-added processes. Many parameters influence cycle times in semiconductor manufacturing. Robinson et al. (2002) provide some key factors such as equipment availability, utilization, product mix, variability, hot lots, re-entrant flows, etc. Since products share the same resources at multiples stages of their processes in semiconductor manufacturing, regulating the competition between products on the different shared resources is critical, in this work, this is done by providing production targets that ensures that the waiting time of products in each operation is not preventing cycle time targets to be reached.

4.2.1 Push Strategy versus Time at Operation Strategy

To minimize cycle times, the push strategy and the time at operation strategy are oriented towards the management of the Work-In-Process on at the operations of each product. These strategies are described below.

– The Push strategy, which consists of placing costs in decreasing order from the first operation to the last operation allowing products to advance as quickly as possible towards their last operations. Assume that UB is the maximum number of operations in the product mix. UB is decreased forward on the set of operations of each product,

to ensure that products are pushed forward toward their last operations. Figure 4.2 presents the Push strategy.

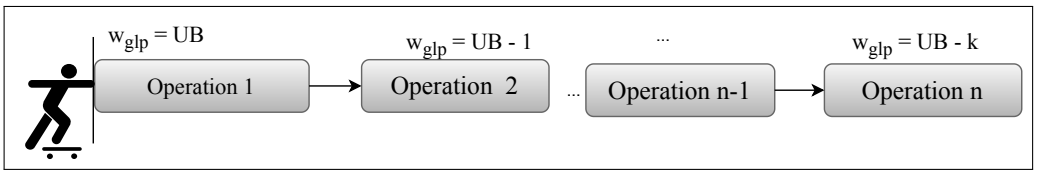


Figure 4.2: Push strategy

Table 4.1 presents the parameters and decision variables used in the global scheduling models for the push strategy and the time at operation strategy.

Parameters:	
$\mathcal{G}$	Set of products,
$\mathcal{K}$	Set of work-centers,
$\mathcal{L}_g$	Set of operations in route of product g,
$\mathcal{LK}(k)$	Set of operations and products that must be processed in work-center k, i.e $(g,l) \in \mathcal{L}K(k)$ means that operation l in route of product g must be processed in work-center k,
P	Number of periods in planning horizon,
IW_{gl}	Initial WIP at operation l of product g,
R_{gp}	Release quantity of product g in period p,
α_{gl}	Unit process time at operation l of product g,
C_{kp}	Capacity of work-center k in period p,
w_{glp}	Unit WIP holding cost at operation l of product in route g in period p.
Decision variables:	
X_{glp}	Quantity of product g arriving in operation l in period p,
Y_{glp}	Quantity of product g completing operation l in period p,
Z_{glp}	WIP of product g at operation l at the end of period p,
Z_{glpt}	WIP of product g at operation l at the end of period p that arrived in period t ($t \leq p$ and $\sum_{t=1}^{p} Z_{glpt} = Z_{glp}$).

Table 4.1: Notations

Below, the Linear Program that models the global scheduling Push strategy is written.

$$Min \sum_{g \in \mathcal{G}} \sum_{l \in \mathcal{L}_g} \sum_{p=1}^{P} w_{glp} Z_{glp} \tag{4.1}$$

Subject to :

$$X_{glp} = Y_{g(l-1)p} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ \forall p \tag{4.2}$$

$$Z_{g11} = IW_{g1} + R_{g1} - Y_{g11} \quad \forall g \in \mathcal{G} \tag{4.3}$$

$$Z_{gl1} = IW_{gl} - Y_{gl1} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2 \tag{4.4}$$

$$Z_{g1p} = Z_{g1(p-1)} + R_{gp} - Y_{g1p} \quad \forall g \in \mathcal{G},\ p = 2,\ \ldots,\ P \tag{4.5}$$

$$Z_{glp} = Z_{gl(p-1)} + X_{glp} - Y_{glp} \quad \forall g \in \mathcal{G},\ \forall l \geq 2,\ p = 2, ...,\ P \tag{4.6}$$

$$\sum_{(g,l)\in\mathcal{LK}(k)} \alpha_{gl} Y_{glp} \leq C_{kp} \quad \forall k \in \mathcal{K},\ p = 1,\ \ldots,\ P \tag{4.7}$$

$$Z_{glp},\ Y_{glp},\ X_{glp} \geq 0 \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ p = 1,\ \ldots,\ P \tag{4.8}$$

The objective function (4.1) ensures that the Work-In-Process is moving forward to the last operations. The costs w_{glp} are chosen in such a way that $w_{glp} \leq w_{gl-1p} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ p = 1,\ \ldots,\ P$. Constraints (4.2) tie consecutive operations. Constraints (4.3)-(4.6) are flow constraints linking the Work-In-Process of each product at each operation in each period with the quantity completed in period p (Y variables) and the quantity arriving in period p (X variables). Constraints (4.7) are resource capacity constraints.

The downside of the Push strategy is that, when production targets are determined, products with a large number of operations are prioritized. This is because the larger the number of operations, the higher the cost for holding Work-In-Process.

Let us consider two products, product P_1 with 5 operations and product P_2 with 3 operations. Assume again that product P_1 in operations 1 and 2 shares the same resource R with product P_2 in all its operations. Resource R has a limited capacity. The unit holding cost for the Work-In-Process is given in Figure 4.3.

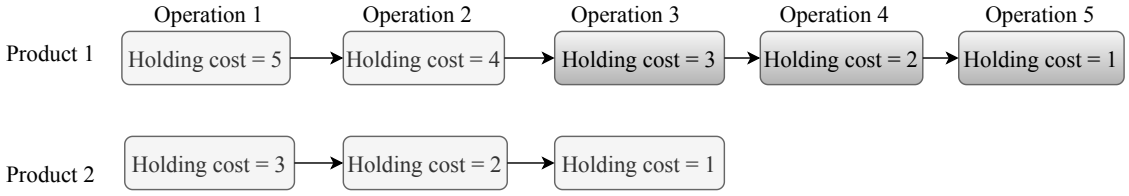


Figure 4.3: Push strategy drawback illustration.

If there is not enough products P_1 to fill the capacity of resource R, then both products P_1 and P_2 will be processed. As product P_1 has a larger priority based on its unit holding cost, if there is enough product P_1 to fill the capacity of resource R, then product P_2 will not be produced. The time at operation strategy is designed to overcome this drawback.

– The time at operation strategy ensures that the Work-In-Process of product g arriving in period p at operation l and which remains at the end of period p does not have the same holding cost β as the Work-In-Process that arrived at period $t < p$ at operation l. Figure 4.4 presents the time at operation strategy.

Numerical results in Section 4.2.2 show that the time at operation strategy reduces the average cycle time compared to the Push strategy by 15% and increases the overall throughput of the factory by 8%. However, this strategy does not take into account the product release dates and therefore not the past temporal trace of the Work-In-Process. Product release dates are considered in Section 4.3, which deals with the problem of controlling cycle times. The Linear Program that models the time at operation global scheduling strategy is

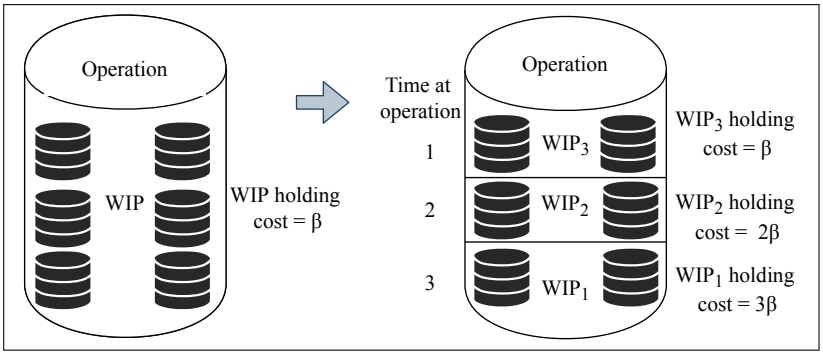


Figure 4.4: Time at operation strategy

written bellow:

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \sum_{t=1}^{p} [(p+1)-t] Z_{glpt} \tag{4.9}$$

Subject to :

$$X_{glp} = Y_{g(l-1)p} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ \forall p \tag{4.10}$$

$$Z_{g111} = IW_{g1} + R_{g1} - Y_{g11} \quad \forall g \in \mathcal{G} \tag{4.11}$$

$$Z_{gl11} = IW_{gl} - Y_{gl1} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2 \tag{4.12}$$

$$\sum_{t=1}^{p} Z_{g1pt} \geq \sum_{t=1}^{p-1} Z_{g1(p-1)t} + R_{gp} - Y_{g1p} \quad \forall g \in \mathcal{G},\ p = 2,\ \ldots,\ H \tag{4.13}$$

$$\sum_{t=1}^{p} Z_{glpt} \geq \sum_{t=1}^{p-1} Z_{gl(p-1)t} + X_{glp} - Y_{glp} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ p = 2,\ \ldots,\ H \tag{4.14}$$

$$\sum_{t=1}^{m} Z_{glpt} \geq \sum_{t=1}^{m} Z_{gl(p-1)t} - Y_{glp} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ p = 2,\ \ldots,\ H,\ m \leq p-1 \tag{4.15}$$

$$\sum_{(g,l)\in\mathcal{LK}(k)} \alpha_{gl} Y_{glp} \leq C_{kp} \quad \forall k \in \mathcal{K},\ p = 1,\ \ldots,\ H \tag{4.16}$$

$$Z_{glpt},\ Y_{glp},\ X_{glp} \geq 0 \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ p = 1,\ \ldots,\ H,\ t \leq p \tag{4.17}$$

The objective function (4.9) ensures that the Work-In-Process is pushed forward to the last operations by taking into account how long a products remain in operations. The holding cost of the Work-In-Process is increasing with the number of periods it remains in an operation. The objective function ensures that the Work-In-Process arriving in period p in an operation and still remain at the end of p does not have the same holding cost as the Work-In-Process which arrived in period $t < p$ in the operation.

Constraints (4.10) tie consecutive operations. Constraints (4.11) model the first operation in the first period upon which initial Work-In-Process and release quantities must be considered. Constraints (4.12) model the Work-In-Process for the remaining operations in the

first period based on the completed quantity (Y variables) in the first period and the initial Work-In-Process. Constraints (4.13) for the first operation in which the release must be considered and based on the completed quantity (Y variables) in the first operation, represent the flow of the Work-In-Process from period t to period $p(t \leq p)$. Constraints (4.14)-(4.15) compute the Work-In-Process which remains in each of the remaining operation and its flow from period t to period p ($t \leq p$) with the quantity completed in period p (Y variables) and the quantity arriving in period p (X variables). Constraints (4.16) are resource capacity constraints.

4.2.2 Computational Experiments

Numerous tests have been conducted on industrial instance with 449 machines in 203 workcenters, which are shared between operations of various types of products. Products have between 352 and 622 operations in their routes. Five products are considered and the release scheme is one lot for each product every 205 minutes.

The global scheduling models which implement the push strategy and the time at operation strategy are coupled with the generic simulation model. The First-In First-Out (FIFO) rule is used as dispatching rule when the simulation runs without the global scheduling approach. This FIFO rule is supplemented with the Production Target Dispatching Rule (PTDR) when the global scheduling approach is used. Recall that the production target dispatching rule ensures that production targets of products are followed in the simulation model.

Tables 4.2, 4.3 and 4.4 present the numerical results obtained with respectively simulation without global scheduling strategy, simulation coupled with global scheduling push strategy (model of push strategy) and simulation coupled with the global scheduling time at operation (model of time at operation strategy).

			Products		
Indicators	1	2	3	4	5
Average Cycle Time (days)	**44.2**	**53.3**	*41.1*	85.4	*59.7*
Release Quantities	2,529	2,529	2,529	2,529	2529
Throughput	**2,099**	**2,009**	*2,125*	1,715	*1,954*
Percentage Achieved Throughput	108.2%	103.6%	109.6%	88.4%	100.8%
Weighted Total Average Cycle Time	**55.6**				
Total Throughput	**9,902**				

Table 4.2: Results of simulation model without global scheduling strategy

Compared to Table 4.2, the results in Table 4.3 show that the throughput of product 3 is significantly reduced by 9.7% and the cycle time increases by 39.9%. The same can be said for product 5 with respectively a decrease of 20.3% and an increase of 36.8% respectively on the throughput and the cycle time. This is because products 3 and 5 have a relatively small number of operations (352 and 415 respectively) and because they compete the same resources than products 1 and 2. Products 1 and 2 have a larger number of operations (501 and 440 respectively), and are thus prioritized by the push strategy. The throughput of products 1 and 2 increase both by 4.8%, and their cycle times decrease respectively by 24.0% and 2.6%. This explain the drawback of the push strategy explained in Section 4.2.1. Product 4 has the largest number of operations (622 operations) and, based on the results

in Tables 4.2, 4.3 and 4.4, is not competing for the same resources than the other products. This is why the cycle times of product 4 do not change much from Table 4.2 to Table 4.3.

	Products				
Indicators	1	2	3	4	5
Average Cycle Time (days)	**33.6**	**51.9**	*57.5*	86.5	*81.7*
Release Quantities	2,529	2,529	2,529	2,529	2529
Throughput	**2,200**	**2,106**	*1,919*	1694	*1557*
Percentage Achieved Throughput	113.5%	108.6%	99.0%	87.4%	80.3%
Weighted Total Average Cycle Time	**59.9**				
Total Throughput	**9,476**				

Table 4.3: Results of simulation model with push strategy

We can observe that, in Table 4.4, the throughput of all products is larger than in Table 4.2. In addition, the total throughput is larger with a lower total average cycle time compared to Tables 4.2 and 4.3. This is because the time at operation strategy manages the Work-In-Process so that product quantities that arrived at different times in the queue of an operation are penalized differently. The Work-In-Process that has been waiting the most is prioritized and not the entire Work-In-Process of a single product. Thus, the Work-In-Process of all products might remain at the end of each periods in the global scheduling model.

	Products				
Indicators	1	2	3	4	5
Average Cycle Time (days)	43.8	44.6	34.4	79.1	58.7
Release Quantities	2,529	2,529	2,529	2,529	2529
Throughput	2,125	2,132	2,232	1,792	1976
Percentage Achieved Throughput	109.6%	110.0%	115.1%	92.4%	101.9%
Weighted Total Average Cycle Time	**50.9**				
Total Throughput	**10,257**				

Table 4.4: Results of simulation model with time at operation strategy

4.3 Cycle Time Control

4.3.1 Introduction and Motivation

Controlling cycle time implies rigorous management of the Work-In-Process in the factory. This management requires having static information such as the cycle time target for each product and the release dates (times when product quantities are released in the factory). In addition, dynamic information is necessary such as the stage of production of each product and the time products have already spent in the factory since their release dates. To our knowledge, cycle time control has never been studied in the literature of semiconductor manufacturing. Cycle times are generally observed in the end of production. Even when it

comes to minimizing cycle times, it is difficult to quantify the extent to which cycle times are minimized. Cycle times are generally outputs of the factory and it seems that they are not often used as inputs parameters, especially in global scheduling models.

This section presents a novel global scheduling strategy to control cycle times in large, complex manufacturing systems consisting of multiple work-centers by using cycle time targets of products as input parameters to the global scheduling model. One of the main innovations of the proposed strategy is that it is based on the use of the temporal tracing of the Work-In-Process. This temporal tracing of Work-In-Process is critical to differentiate lots of the same product and at the same processing stage, but released at different times in the factory. In previous strategies, the release dates of products in the Work-In-Process were not considered. This strategy innovates by using both the release dates and the temporal tracing of the Work-In-Process in the global scheduling model. The control of cycle times is managed by controlling the competition of products on the shared resources using production targets.

Section 4.3.2 presents the way the historical trace of the Work-In-Process is used in the global scheduling model. In Section 4.3.3, the management of product cycle time targets is presented before its use in the global scheduling model. Finally, in Section 4.3.4, computational experiments and analysis of two different approaches (Naive and Simulation-based) are presented and discussed.

4.3.2 Temporal Tracing of the Work-In-Process

A first originality of this strategy is that the WIP is temporarily traced in the global scheduling model. As shown in Figure 4.5, instead of having just one parameter IW_{gl} for the initial Work-In-Process in operation l of product g, parameter IW_{glr} is used for the initial WIP in operation l of product g released at period $-r$ in the past, where $IW_{gl} = \sum_{r=1}^{R} IW_{glr}$ and R is the number of release periods that must be considered in the past.

The variables modeling the Work-In-Process of product g at operation l at the end of period p are also considering the release period. More precisely, Z^P_{glpr} corresponds to the quantity of product g at operation l released in period $-r$ in the past, and Z^F_{glpr} to the quantity of product g at operation l to be released in period r in the future. The optimization model that aims at satisfying given product cycle time targets is formalized in Section 4.3.4. A temporal tracing of the WIP is required to know which products in the WIP should actually be processed in a period. Indeed, if the information on the release periods of products is not available, it is impossible to know the time already spent in the factory by products in the global scheduling model, and thus to ensure the satisfaction of cycle time targets.

When considering Figure 4.5, products in IW_{glr} should be processed before products in $IW_{glr'}$ if $r > r'$ since products in IW_{glr} have been released earlier. To our knowledge, no optimization models in the literature explicitly consider the temporal tracing of the Work-In-Process.

The approach for controlling cycle times starts by building blocks of operations, i.e., subsequences of operations of the products as in Bureau, Dauzère-Pérès, Yugma, Vermariën and Maria (2007). The goal is then to control the cycle times of products by controlling the completion times of products in blocks, which are determined from their release dates. Hence, another important aspect of our approach is to establish a cycle time target for each block of operations (expressed as a parameter T_{gl} for operation l of product g), which is derived from the cycle time target of the product. The objective in the global scheduling

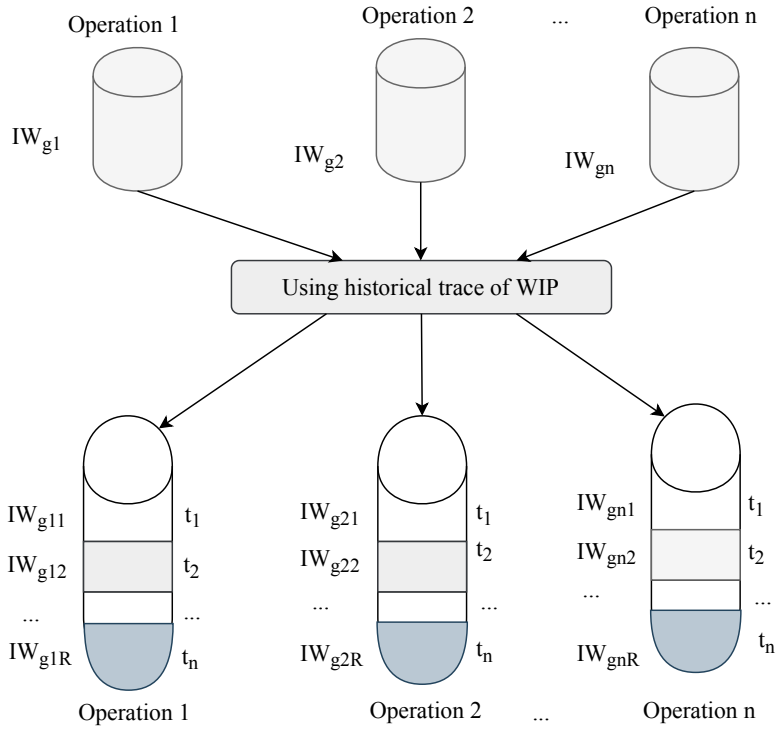


Figure 4.5: Operations with historical trace of initial WIP

model is to prioritize products in a block that are behind their cycle time targets.

4.3.3 Product Cycle Time Targets and Blocks of Operations

To control the cycle time of a product, its cycle time target is distributed over product operations. The number of operations is divided into blocks of operations in such a way that the last operation in the last block of operations should end at the cycle time target of the product. Blocks are defined using two different methods:

- A *naive method*, where operations of a product is divided into blocks (subsequence of operations) with the same number of operations in each block, and the cycle time target is the same in each block. Figure 4.6 illustrates the process of defining blocks of operations. For a product with a cycle time target of 40 days and 5 blocks, blocks of 8 operations are defined in such a way that operations in the first block should end at 20% of the cycle time target of the product, operations in the second block should end at 40% of the cycle time target of the product, etc..

- A method based on simulation (*simulation-based method*), where the operations of a product are divided into blocks (subsequence of operations) with the same number of operations, but with different cycle time targets. The time duration or the cycle time target of a block is determined based on the time each product spends in that block in the simulation. After the warm-up time (times to load the factory), products are traced in the simulation, and the times they spent in each block are collected. These times provide the percentage of the cycle time target of a product in each block, and are used to determine the cycle time target of each block. Based on the example in Figure 4.7, for a product with a cycle time target of 40 days and 5 blocks of operations, if the

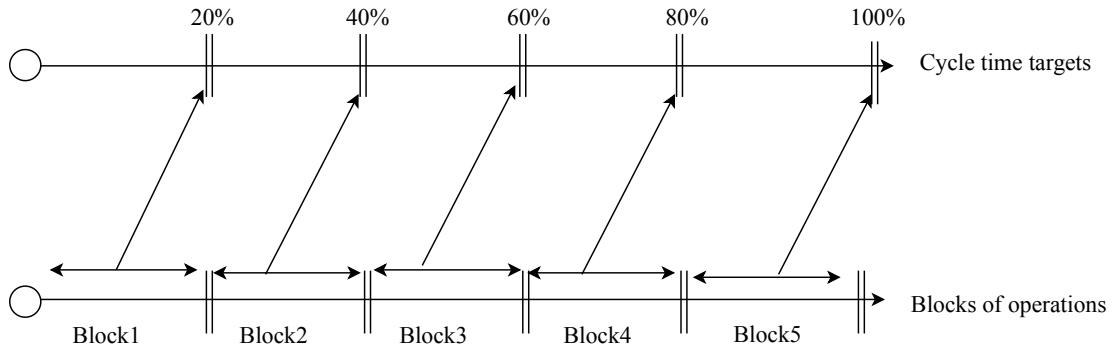


Figure 4.6: Example of the definition of cycle time targets for blocks with the naive method

information collected from the simulation indicates that the product spent on average 2 days, 14 days, 4 days, 16 days and 4 days in blocks 1, 2, 3, 4 and 5 respectively, then operations in the first block should end at 5% of the cycle time target of the product, operations in the second block should end at 40% of the cycle time target of the product, etc.

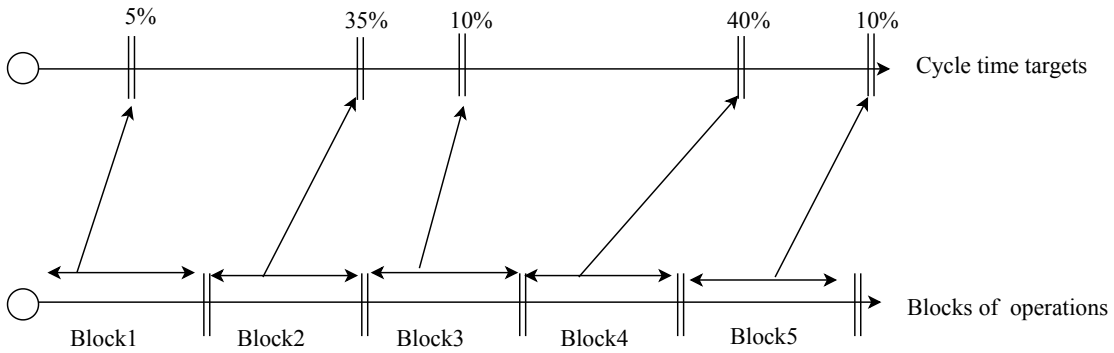


Figure 4.7: Example of definition of cycle time targets for blocks with the simulation-based method

All cycle time targets of blocks are converted into a number of periods and are used in the optimization model. Our computational experiments are conducted and compared with respectively 12 blocks and 50 blocks of operations.

4.3.4 Global Scheduling: Linear Programming Model

Table 4.5 presents the parameters and decisions variables used in the optimization model.

The global scheduling optimization model which implements the strategy for cycle time control is formalized below.

Parameters:	
$\mathcal{G}$	Set of products,
$\mathcal{K}$	Set of work-centers,
$\mathcal{L}_g$	Set of operations in route of product g,
$\mathcal{LK}(k)$	Set of operations and products that must be processed in work-center k, i.e $(g,l) \in \mathcal{L}K(k)$ means that operation l in route of product g must be processed in work-center k,
R	Number of release classes considered in the past,
H	Number of periods in the planning horizon,
T_{gl}	Cycle time target of operation l of product g, which is derived from the cycle time target of the block of the operation, each operation l of product g in that block should be completed at period T_{gl},
Q_{gp}	Release quantity of product g in period p,
IW_{glr}	Initial WIP in operation l of product g released in period $-r$,
α_{gl}	Unit process time for product g at operation l of product g,
C_{kp}	Capacity of work-center k in period p.
Decision variables:	
X^P_{glpr}	Quantity of product g released at period $-r$ in the past, completing operation l in period p, where $r = 1, \ldots, R$,
X^F_{glpr}	Quantity of product g to be released at period r, completing operation l in period p, where $r = 1, \ldots, p$,
Y_{glp}	Total quantity of products in route g to complete in operation l in period p, i.e. the production target,
Z^P_{glpr}	WIP of product g at operation l at the end of period p released at period $-r$ in the past, where $r = 1, \ldots, R$,
Z^F_{glpr}	WIP of product g at operation l at the end of period p to be released at period r, where $r = 1, \ldots, p$.

Table 4.5: Notations

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{R} \max(0, p+r-T_{gl}) Z^P_{glpr} + \sum_{r=1}^{p} \max(0, p-r-T_{gl}) Z^F_{glpr} \right) \quad (4.18)$$

Subject to :

$$Z^P_{g11r} = IW_{g1r} - X^P_{g11r} \quad \forall g \in \mathcal{G},\ r = 1, \ldots, R \quad (4.19)$$

$$Z^F_{g111} = Q_{g1} - X^F_{g111} \quad \forall g \in \mathcal{G} \quad (4.20)$$

$$Z^P_{gl1r} = IW_{glr} + X^P_{g(l-1)1r} - X^P_{gl1r} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ r = 1, \ldots, R \quad (4.21)$$

$$Z^F_{gl11} = X^F_{g(l-1)11} - X^F_{gl11} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2 \quad (4.22)$$

$$Z^P_{g1pr} = Z^P_{g1(p-1)r} - X^P_{g1pr} \quad \forall g \in \mathcal{G},\ p = 2, \ldots,\ H,\ r = 1, \ldots, R \quad (4.23)$$

$$Z^F_{g1pr} = Z^F_{g1(p-1)r} - X^F_{g1pr} \quad \forall g \in \mathcal{G},\ p = 2, \ldots,\ H,\ r = 1, \ldots, p-1 \quad (4.24)$$

$$Z^F_{g1pp} = Q_{gp} - X^F_{g1pp} \quad \forall g \in \mathcal{G},\ p = 2, \ldots,\ H \quad (4.25)$$

$$Z^P_{glpr} = Z^P_{gl(p-1)r} + X^P_{g(l-1)pr} - X^P_{glpr} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ p = 2, \ldots,\ H,\ r = 1, \ldots, R \quad (4.26)$$

$$Z^F_{glpr} = Z^F_{gl(p-1)r} + X^F_{g(l-1)pr} - X^F_{glpr} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ p = 2, \ldots,\ H,\ r = 1, \ldots, p-1 \quad (4.27)$$

$$Z^F_{glpp} = X^F_{g(l-1)pp} - X^F_{glpp} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ l \geq 2,\ p = 2, \ldots,\ H \quad (4.28)$$

$$\sum_{r=1}^{R} X^P_{glpr} + \sum_{r=1}^{p} X^F_{glpr} = Y_{glp} \quad \forall g \in \mathcal{G},\ \forall l \in \mathcal{L}_g,\ p = 1, \ldots,\ H \quad (4.29)$$

$$\sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g;\ (g,l)\in\mathcal{LK}(k)} \alpha_{gl} Y_{glp} \leq C_{kp} \quad \forall k \in K,\ p = 1, \ldots,\ H \quad (4.30)$$

$$Z^P_{glpr},\ X^P_{glpr} \geq 0 \quad \forall g \in \mathcal{G},\ \forall l \in L(g),\ p = 1, \ldots,\ H,\ r = 1, \ldots, R \quad (4.31)$$

$$Z^F_{glpr},\ X^F_{glpr} \geq 0 \quad \forall g \in \mathcal{G},\ \forall l \in L(g),\ p = 1, \ldots,\ H,\ r = 1, \ldots, H \quad (4.32)$$

$$Y_{glp} \geq 0 \quad \forall g \in \mathcal{G},\ \forall l \in L(g),\ p = 1, \ldots,\ H \quad (4.33)$$

The objective function (4.18) aims at satisfying the cycle time target of operations in blocks, by prioritizing the reduction of the Work-In-Process at operations with products that are late the most. The lateness is equal to $\max(0, p+r-T_{gl})$ for products released in past periods ($r = 1, \ldots, R$) and to $\max(0, p-r-T_{gl})$ for products released in future periods ($r = 1, \ldots, p$, for $p = 1, \ldots, H$). Hence, late products are pushed forward to their following operations. Constraints (4.19) and (4.20), resp. Constraints (4.21) and (4.22), determine the remaining Work-In-Process at the end of the first period in the first operation of each product, resp. in the following operations of each product. Constraints (4.19) and (4.21) correspond to the WIP for products released in past periods, while Constraints (4.20) and (4.22) correspond to the WIP for products released in the first period. Constraints (4.23), (4.24) and (4.25), resp. Constraints (4.26), (4.27) and (4.28), determine the remaining WIP

in the first operation, resp. in each operation except the first one, of each product at the end of each period except the first one. Constraints (4.23) and (4.26) correspond to the WIP for products released in past periods, while Constraints (4.24), (4.25), (4.27) and (4.28) correspond to the Work-In-Process for products released in previous periods in the horizon except the first one. Constraints (4.29) ensure that Y_{glp}, the total quantity of product g that completes operation l in period p, is equal to the sum of the quantities of product g released in past periods and of the quantities of product g released in previous periods in the horizon. Constraints (4.30) model the resource capacity constraints. Finally, Constraints (4.31) through (4.33) are the non-negativity constraints.

Note that the initial Work-In-Process of products released in the past (considered in Constraints (4.19) and (4.21)), the products released before p (considered in Constraints (4.20) and (4.25)) or the WIP in previous periods (considered in Constraints (4.23), (4.24), (4.26) and (4.27)) are not considered in Constraints (4.22) and (4.28), since the products released in p are only entering through the first operation in Constraints (4.25).

4.3.5 Aggregating into Classes of Release Periods

As shown in the previous section, past release periods of each product are considered in the global scheduling model to ensure the temporal tracing of the Work-In-Process and to control the product cycle times. However, because cycle times are very long in semiconductor manufacturing, up to 3 months, the number R of release periods considered in the past can be very large, leading to a very large number of variables X^P_{glpr} and Z^P_{glpr} in the linear programming model. This is why, instead of modeling each release period in the past and to make the model tractable, we aggregate past periods into N classes of A consecutive past periods, where $R = AN$.

N should be chosen not too large (too few past periods in each class), to avoid having a very large scheduling model, and not too small (too many past periods in each class), to avoid having in the same class products released at very different periods. There are several ways to build classes of past periods, where the extreme cases correspond either to using each period as a class ($N = R$) or to using a single class ($N = 1$). In the linear programming model, R is replaced by N, and the objective function (4.18) needs to be adjusted accordingly as follows:

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{N} \max(0, p + (r)A - T_{gl}) Z^P_{glpr} + \sum_{r=1}^{p} \max(0, p - r - T_{gl}) Z^F_{glpr} \right) \quad (4.34)$$

In the computational experiments of Section 4.3.6, 12 ($N = 12$) classes of 7 days ($A = 7$) are used, i.e., 84 days, which is much larger than the average cycle time of most products in our experiments. Hence, products in the factory and released in the past week are aggregated into the first release class, products in the factory and released two weeks ago are aggregated into the second release class, etc.

4.3.6 Computational Experiments

Industrial data were used for the experiments below from a factory with about 600 machines distributed in about 300 work-centers. Two instances with 5 products each are used:

1. The first instance includes products numbered 1 through 5, which have between 104 and 315 operations. One unit of product is released every 280 minutes, 360 minutes, 480 minutes, 480 minutes and 480 minutes for products 1, 2, 3, 4 and 5, respectively.

2. The second instance includes products numbered 6 through 10, which have between 153 and 221 operations. One unit of product is released every 460 minutes, 460 minutes, 480 minutes, 480 minutes and 480 minutes for products 6, 7, 8, 9 and 10, respectively.

To avoid starting with an empty factory, six months of warm-up time are used, which are excluded when collecting statistical data. The simulation is then run for 6 months after the warm-up time.

4.3.6.1 Without Global Scheduling Model

This section presents the results obtained with the simulation model without the global scheduling model, i.e. where only First-In-First-Out dispatching rules are used. Table 4.6 shows the results for the first instance. For each product, the average cycle time, the release quantities and the completed quantities (throughput), both in number of products, are given. Note that the release quantities correspond to the number of products released after the warm-up period, and the completed quantities are equal to the number of products completed among the release quantities and that are used to compute the average cycle time. As expected, the average cycle time of a product usually decreases when the number of completed products increases.

The average cycle time of products 1, 2 and 3 are rather close, with a value of about 50 days, while products 4 and 5 are faster with a value of about 40 days.

	Products				
	1	**2**	**3**	**4**	**5**
Average Cycle Times (days)	**48.9**	**50.3**	**51.1**	**40.1**	**39.8**
Release Quantities	926	721	541	541	541
Completed Quantities	617	479	356	394	394

Table 4.6: Simulation without global scheduling approach, instance 1

Table 4.7 shows the results for the second instance. The average cycle time of Products 6 and 9 are the fastest, with a value of about 32 days, while Products 7, 8 and 10 are slower, in particular product 10 with a value of about 78 days.

	Products				
	6	**7**	**8**	**9**	**10**
Average Cycle Times (days)	**31.8**	**59.1**	**46.1**	**32.2**	**77.7**
Release Quantities	563	563	541	541	541
Completed Quantities	441	341	374	423	265

Table 4.7: Simulation without global scheduling approach, instance 2

In the experiments presented in sections 4.3.6.2 and 4.3.6.3, the average cycle times in Tables 4.6 and 4.7 will be used as initial cycle time targets in the global scheduling model.

4.3.6.2 Naive Method to Determine Cycle Time Targets of Blocks

Instance 1

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	42.7	40.8	49.2	36.8	**43.1**
Cycle Time Gaps	-11.0%	-18.4%	-3.5%	-8.0%	**10.5%**
Release Quantities	926	721	541	541	541
Completed Quantities	703	546	373	411	388

Table 4.8: Naive method, Simulation with global scheduling approach, instance 1 with 12 blocks of operations

Table 4.8, resp. Table 4.9, shows the results for instance 1 obtained when the global scheduling approach is applied, i.e., when the simulation model is coupled with the global scheduling model, with 12 blocks of operations, resp. 50 blocks of operations. The first row provides the cycle time targets, defined with the results obtained in Section 4.3.6.1, which are used to derive the cycle time targets of blocks with the naive method. The second row presents the average cycle time obtained with the global scheduling approach, and the third row the gaps between the cycle time targets and the average cycle time. The last two rows provide the release quantities and the completed quantities, both in number of products.

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	39.3	43.7	39.8	**42.5**	**41.0**
Cycle Time Gaps	-18.1%	-12.6%	-22.3%	**6.0%**	**5.1%**
Release Quantities	926	721	541	541	541
Completed Quantities	695	532	402	386	383

Table 4.9: Naive method, simulation with global scheduling approach, instance 1 with 50 blocks of operations

In Table 4.8, the average cycle time of products 1 to 4 are smaller than their cycle time targets, but the average cycle time of product 5 is about 4 days larger than its cycle time target. The results in Table 4.9 show that increasing the number of blocks to 50 helps to improve the results since, although two products have an average cycle time which is larger than their cycle time target, the largest difference is reduced to 2.5 days. However, the negative cycle time gaps of some products are very large, up to -22.3% for product 3 with 50 blocks of operations, which is not wanted when other products have positive cycle time gaps.

To reduce the average cycle time of product 5, its cycle time target has been reduced to 15 days and the numerical results with the global scheduling approach and 12 blocks of operations can be found in Table 4.10. Note that the cycle times targets of the other products have not been changed. Table 4.10 shows that the average cycle times of products

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	**15.0**
Average Cycle Times (days)	42.0	40.7	49.7	37.6	**39.5**
Cycle Time Gaps	-12.5%	-18.6%	-2.5%	-6.0%	163.3%
Release Quantities	926	721	541	541	541
Completed Quantities	698	546	364	408	404

Table 4.10: Naive method, simulation with global scheduling approach, instance 1 with 12 blocks of operations and reduction of cycle time target of product 5

1 to 4 are still smaller than their cycle time targets, and that the average cycle time of product 5 is now very close to its cycle time target in Table 4.8 and smaller than its average cycle time in Table 4.6. However, the cycle time target of product 5 had to be drastically reduced to obtain this result.

Instance 2

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77.0
Average Cycle Times (days)	**44.0**	54.5	37.5	26.0	61.8
Cycle Time Gaps	**41.9%**	-7.6%	-18.5%	-18.7%	-19.7%
Release Quantities	563	563	541	541	541
Completed Quantities	391	324	402	459	355

Table 4.11: Naive method, simulation with global scheduling approach, instance 2 with 12 blocks of operations

Table 4.11, resp. Table 4.12, shows the results for instance 2 obtained with the simulation model coupled with the global scheduling model and with 12 blocks of operations, resp. 50 blocks of operations. The average cycle times of products 7 to 10 are smaller than their cycle time targets, and significantly smaller for products 8, 9 and 10. However, the average cycle time of product 6 is much larger (13 days) than its cycle time target. As for instance 1, increasing the number of blocks to 50 helps to reduce the maximum cycle time gaps, as shown in Table 4.12 since, although two products have now an average cycle time which is larger than their cycle time target, the largest difference is reduced to 7.5 days, which is still quite large. Also as in for instance 1, the negative cycle time gaps of some products are very large, up to -42.9% for product 10 with 50 blocks of operations.

Because of the re-entrant flows and shared resources, trying to satisfy the cycle times of products 7 to 10 leads to a significant slowdown of product 6. Hence, and as in the previous section, the cycle time target of product 6 is decreased to 15 days, while the cycle time targets of the other products remain the same, and the global scheduling approach with 12 blocks of operations is applied again. The associated numerical results are given in Table 4.13. They show that the average cycle times of products 7 to 10 remain smaller than

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77.0
Average Cycle Times (days)	**36.3**	**66.5**	32.4	23.2	44.1
Cycle Time Gaps	**17.1%**	**12.7%**	-13.6%	-27.5%	-42.9%
Release Quantities	563	563	541	541	541
Completed Quantities	427	301	416	450	409

Table 4.12: Naive method, simulation with global scheduling approach, instance 2 with 50 blocks of operations

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	**15.0**	59.0	46.0	32.0	77.0
Average Cycle Times (days)	**42.4**	55.2	37.9	26.7	62.3
Cycle Time Gaps	182.6%	-6.4%	-17.6%	-16.6%	-19.1%
Release Quantities	563	563	541	541	541
Completed Quantities	395	336	400	455	337

Table 4.13: Naive method, simulation with global scheduling approach, instance 2 with with 12 blocks of operations and reduction of cycle time target of product 6

their cycle time targets, and that the average cycle time of product 6 has only been slightly reduced and remains much larger than its average cycle time in Table 4.6.

Reducing the cycle time of product 6

For the average cycle time of product 6 to reach the cycle time target of 31 days, the first step is to understand how the flows of other products impact the flow of product 6. The cycle time of product 6 can be decreased by slowing down other products, i.e., increasing their cycle times targets. Four scenarios are thus considered as shown in Table 4.14, where each of the four last products is alternatively slowed down by increasing its cycle time target to 150 days, and the cycle time target of product 6 is set to 31 days again.

		Products				
		6	**7**	**8**	**9**	**10**
Scenario 1	**Cycle time targets (days)**	31.0	**150.0**	46.0	32.0	61.0
	Average Cycle Times (days)	35.3	100.7	46.8	25.2	49.3
Scenario 2	**Cycle time targets (days)**	31.0	59.0	**150.0**	32.0	61.0
	Average cycle times (days)	42.2	52.5	56.2	26.2	56.0
Scenario 3	**Cycle time targets (days)**	31.0	59.0	46.0	**150.0**	61.0
	Average cycle times (days)	34.3	46.0	37.1	70.1	49.7
Scenario 4	**Cycle time targets (days)**	31.0	59.0	46.0	32.0	**150.0**
	Average cycle times (days)	39.5	47.7	35.2	26.4	109.5

Table 4.14: Impact of slowing down a single product on average cycle times of product 6

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	**31.0**	59.0	46.0	**250.0**	61.0
Average Cycle Times (days)	**26.5**	43.6	36.5	**86.9**	48.5
Cycle Time Gaps	-14.5%	-26.1%	-20.6%	-65.2%	-20.5%
Release Quantities	563	563	541	541	541
Completed Quantities	434	386	406	279	387

Table 4.15: Naive method, simulation with global scheduling approach, slowing down product 9

The results in Table 4.14 show how the flow of each product impacts the cycle time of product 6. Increasing the cycle time target of each product helps to reduce the average cycle time of product 6 from its initial value or 44 days (see Table 4.11). However, the impact of products 7 and 9 is more significant and rather close. This is because products 6, 7 and 9 are often competing for the same machines. Table 4.15 shows how the average cycle time of product 6 can be further reduced by increasing even more the cycle time target of product 9 from 150 days to 250 days. The average cycle time of product 6 is now equal to 26.5 days, i.e., it is finally lower than the cycle time target of 31 days.

		Products				
		6	**7**	**8**	**9**	**10**
Scenario 5	**Cycle time targets (days)**	31.0	**90.0**	46.0	**90.0**	61.0
	Average cycle times (days)	34.5	59.0	45.5	50.5	52.1
Scenario 6	**Cycle time targets (days)**	31.0	59.0	46.0	**90.0**	**90.0**
	Average cycle times (days)	36.0	45.2	37.4	49.9	73.3
Scenario 7	**Cycle time targets (days)**	31.0	**90.0**	46.0	32.0	**90.0**
	Average cycle times (days)	37.6	59.7	41.3	27.1	73.5
Scenario 8	**Cycle time targets (days)**	31.0	**90.0**	46.0	**90.0**	**90.0**
	Average cycle times (days)	**30.4**	59.5	41.9	47.1	72.8

Table 4.16: Impact of slowing down multiple products on average cycle times of product 6

The issue with the results in Table 4.15 is that product 9 is significantly slowed down. An alternative is to slow down multiple products simultaneously and less drastically than in Tables 4.14 and 4.15. Four new scenarios are thus considered as shown in Table 4.16, where two or three products are slowed down by increasing their cycle time target to 90 days, instead of 150 days in Table 4.14 and 250 days in Table 4.15. More precisely, the cycle time targets of products 7 and 9 are increased in scenario 5, of products 9 and 10 in scenario 6, of products 7 and 10 in scenario 7 and of products 7, 9 and 10 in scenario 8. The cycle time target of product 6 remains equal to 31 days.

The results in Table 4.16 show that the average cycle time of product 6 is always significantly reduced from its initial value or 44 days (see Table 4.11) when the cycle time targets of two products are reduced. However, it is when the cycle time targets of three products are reduced (scenario 8) that the cycle time target of product 6 is finally satisfied.

4.3.6.3 Simulation-based Method to Determine Cycle Time Targets of Blocks

The analysis conducted in Section 4.3.6.2 shows that the naive method to determine cycle time targets of blocks is limited, and makes it difficult for the global scheduling approach to ensure that the cycle times of some products are satisfied. The results in this section show that the simulation-based method to determine cycle time targets of block helps to answer these limits. Let us recall that, using the simulation-based method, blocks include the same number of operations but have different cycle time targets.

Instance 1

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	**48.3**	40.3	**51.6**	**40.1**	**42.5**
Cycle Time Gaps	**0.6%**	-19.4%	**1.7%**	**0.3%**	**8.9%**
Release Quantities	926	721	541	541	541
Completed Quantities	675	477	371	386	386

Table 4.17: Simulation-based method, simulation with global scheduling approach, instance 1 with 12 blocks of operations

Table 4.17, resp. Table 4.18, shows the results for instance 1 obtained with the simulation model coupled with the global scheduling model with 12 blocks of operations, resp. 50 blocks of operations. The same cycle time targets for products than in Section 4.3.6.2 are used. In Table 4.17, most products have their average cycle times that are very close to their cycle time targets, except for product 2 with an average cycle time which is 19.4% lower and product 5 with an average cycle time which is 8.9% larger. Using 50 blocks of operations leads to very good results as shown in Table 4.18, where all products have an average cycle time which is lower than their cycle time target.

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	40.7	43.5	47.3	40.0	35.7
Cycle Time Gaps	-15.2%	-13.0%	-7.3%	0.0%	-8.5%
Release Quantities	926	721	541	541	541
Completed Quantities	705	524	377	389	405

Table 4.18: Simulation-based method, simulation with global scheduling approach, instance 1 with 50 blocks of operations

Through the use of optimized production targets and the simple controller variables used in the simulation model, cycle times are under control, and are even all improved compared to the simulation model without the global scheduling approach.

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77.0
Average Cycle Times (days)	**34.5**	**60.2**	37.9	**34.7**	73.7
Cycle Time Gaps	**11.3%**	**2.0%**	-17.6%	**8.4%**	-4.3%
Release Quantities	563	563	541	541	541
Completed Quantities	405	314	412	396	306

Table 4.19: Simulation-based method, simulation with global scheduling approach, instance 2 with 12 blocks of operations

Instance 2

Table 4.19, resp. Table 4.20, presents the results for instance 2 obtained with the global scheduling approach with 12 blocks of operations, resp. 50 blocks of operations. The results are worse than in instance 1 for 12 blocks of operations, with three products that have a positive cycle time gap and a maximum cycle time gap of 11.3%. Again, the improvements when using 50 blocks of operations are significant, as all products have an average cycle time which is lower than the corresponding cycle time target as shown in Table 4.20.

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77.0
Average Cycle Times (days)	25.0	57.3	40.0	30.1	77.0
Cycle Time Gaps	-19.4%	-2.9%	-13.0%	-5.9%	0.0%
Release Quantities	563	563	541	541	541
Completed Quantities	471	337	387	425	281

Table 4.20: Simulation-based method, simulation with global scheduling approach, instance 2 with 50 blocks of operations

Reducing the cycle time targets of products

This section aims at illustrating that our global scheduling approach helps to control cycle times by reducing the cycle time targets of different products in instances 1 and 2. Because of the quality of the results obtained in sections 4.3.6.3, 50 blocks of operations are considered in the remaining experiments.

First, the cycle time target of product 3 in instance 1, whose average cycle time is equal to 47.3 days in Table 4.18, is decreased from 51 days to 42 days. The results in Table 4.21 show that the average cycle time decreases from 47.3 to 43.3 days, only 3% above the target cycle time, and the average cycle times of the other products remain under control since the largest cycle time gap is equal to 3.7%.

The cycle time target of product 4 is decreased from 40 to 35 days in Table 4.22, and its average cycle time decreases from 40.0 days in Table 4.18 to 32.3 days, again with a limited impact on the satisfaction of other cycle time targets, since the largest cycle time gap is equal to 2.8% for product 5.

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	**42.0**	40.0	39.0
Average Cycle Times (days)	42.0	42.9	**43.3**	41.5	38.3
Cycle Time Gaps	-7.9%	-18.2%	3.0%	3.7%	-1.8%
Release Quantities	926	721	541	541	541
Completed Quantities	676	524	383	391	403

Table 4.21: Simulation-based method, simulation with global scheduling approach, instance 1, reducing the cycle time target of product 3

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	**35.0**	39.0
Average Cycle Times (days)	41.0	43.8	50.6	**32.3**	40.1
Cycle Time Gaps	-14.5%	-12.40%	-0.8%	-7.7%	2.8%
Release Quantities	926	721	541	541	541
Completed Quantities	679	526	368	416	390

Table 4.22: Simulation-based method, simulation with global scheduling approach, instance 1, reducing the cycle time target of product 4

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	**27.0**	77.0
Average Cycle Times (days)	25.3	57.0	42.6	**26.8**	78.1
Cycle Time Gaps	-18.4%	-3.4%	-7.4%	-0.7%	1.4%
Release Quantities	563	563	541	541	541
Completed Quantities	472	359	378	423	273

Table 4.23: Simulation-based method, simulation with global scheduling approach, instance 2, reducing the cycle time target of product 9

Considering now instance 2, the cycle time target of product 9 is decreased from 32 to 27 days in Table 4.23. The resulting average cycle time of product 9 decreases from 30.1 days in Table 4.20 to 26.8 days, with a very small maximum cycle time gap of 1.4% for product 10.

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	**70.0**
Average Cycle Times (days)	27.1	61.3	46.2	31.6	**70.0**
Cycle Time Gaps	-12.6%	3.9%	0.4%	-1.3%	0.0%
Release Quantities	563	563	541	541	541
Completed Quantities	466	304	360	421	290

Table 4.24: Simulation-based method, simulation with global scheduling approach, instance 2, reducing the cycle time target of product 10

In the last experiment, the cycle time target of product 10 is decreased from 77 to 70 days. Table 4.24 shows that the average cycle time of product 10 exactly reaches its cycle time target and, as importantly, the other products remain under control with a cycle time gap always smaller than 3.9%.

Quadratic Tardiness in the Objective Function

In this experiment instead of using a linear cost on the tardiness in the objective function, a quadratic cost is used in the objective function (4.35).

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{N} \max(0, p + (r)A - T_{gl})^2 Z^P_{glpr} + \sum_{r=1}^{p} \max(0, p - r - T_{gl})^2 Z^F_{glpr} \right) \quad (4.35)$$

Compared with the results in Table 4.18, the maximum negative deviation is reduced by 0.6% for the first instance in Table 4.25. In Table 4.26, the maximum negative deviation for the second instance is reduced by 6.4% compared to Table 4.20.

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	41.0	46.7	46.2	**40.1**	37.1
Cycle Time Gaps	-14.6%	-6.6%	-9.4%	**0.2**%	-5.1%
Release Quantities	926	721	541	541	541
Completed Quantities	697	528	376	387	412

Table 4.25: Simulation-based method, simulation with global scheduling approach, instance 1, quadratic cost on tardiness

Compared with the results in Tables 4.18 and 4.20 where a linear cost on the tardiness is used in the objective function, the negative deviations decrease for products 1, 2 and 5 (Table 4.25) and for products 6 and 7 (Table 4.26) while they increase for product 3

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77
Average Cycle Times (days)	30.5	57.5	40.0	28.0	**78.4**
Cycle Time Gaps	-1.6%	-2.5%	-13.0%	-12.5%	**1.8**%
Release Quantities	563	563	541	541	541
Completed Quantities	450	365	387	449	267

Table 4.26: Simulation-based method, simulation with global scheduling approach, instance 2, quadratic cost on tardiness

(Table 4.25) and product 9 (Table 4.26). In addition, the positive deviations are observed for product 4 (Table 4.25) and product 10 (Table 4.26). The problem of reducing positive and negative deviations on cycle times are discussed in Chapter 5.

4.4 Conclusions and Perspectives

This chapter has proposed and studied different global scheduling strategies to minimize and control cycle times. Two different strategies were compared for minimization of cycle times, the push strategy and the time at operation strategy. The time at operation strategy provides very good results in terms of cycle times and throughput compared with the results of the push strategy. This is because the time at operation strategy manages better the Work-In-Process. Different quantities of Work-In-Process that arrived at different times in an operation are penalized differently in order to prioritize , the processing of the Work-In-Process that has spent the most time in the operation.

Controlling cycle times is very challenging in complex manufacturing systems such as semiconductor manufacturing. The aim of the strategy proposed in the second part of this chapter is to ensure that the cycle time targets are met. The proposed global scheduling model minimizes the gap between the observed cycle times and the cycle time targets of products throughout their production. This is done by using the release dates of products and the temporal tracing of the Work-In-Process . Two methods to determine cycle time targets in blocks (subsequences of operations) of product routes are presented and compared. Numerical results on industrial data show that the global scheduling strategy is effective in steering the manufacturing factory to control the cycle times. This strategy opens a new way of explicitly controlling cycle times in complex manufacturing systems that we hope other researchers will exploit.

As future agenda, various directions of research may be explored. First, the investigation on how to mix the cycle time control strategies with other strategies, such as those in Chapter 3 in a multi-objective approach by combining, for example, the objective of satisfying cycle times and that of minimizing the variability of cycle times. Other objectives can be combined with the control of cycle times such as the level of throughput satisfaction. In practice, for example, the demand for certain products increases while the demand of other products decreases. Therefore, managers may need to increase and decrease the throughput of products while maintaining the cycle time targets. In semiconductor manufacturing, it is always difficult to achieve these two objectives since the increase in throughput of certain products influences the cycle times of other products. The problem can be managed with

multi-objective approaches by defining compromises on the level of throughput satisfaction for certain products and the level of cycle time degradation for other products.

Chapter 5

Multi-objective Optimization for Cycle Time Control

5.1 Introduction

This chapter studies multi-objective strategies for cycle time control. The global scheduling model of Chapter 4 only accelerates products that are behind the cycle time targets of their blocks of operations. The aim is to minimize the positive gap between the observed cycle times and the cycle time targets. However, significant negative gaps on cycle time targets are observed for certain products. In this chapter, in addition to the tardiness, earliness is minimized in the global scheduling model. The objective is to minimize the positive and negative gaps between the observed cycle times and the cycle time targets. Different objective functions are presented. Two multi-objective approaches are used, the first one is using a weighted sum and the second one a lexicographic approach.

5.2 Weighted Sum Approach

5.2.1 Modeling

The weighted sum approach to solve multi-objective optimization problems was introduced by Zadeh (1963), and is the most straightforward multi-objective approach. The idea is to transform a multi-objective optimization problem into a single-objective optimization problem, for which there are many solution methods. This is made possible by assigning a weight w_i to each objective function $f_i(x)$ with a weighted coefficient in order to minimize the weighted sum of the objective functions:

$$Min \sum_{i=1}^{k} w_i f_i(x) \tag{5.1}$$

where $w_i \geq 0$ and should be strictly positive for at least one objective function, such that $\sum_{i=1}^{k} w_i = 1$ (Jaimes et al. (2009)). Although this approach is easy to implement, it nevertheless presents many disadvantages such as the lack of efficiency in the search for solutions enclosed in a concavity, the choice of weights, etc.

In this section, various objective functions are presented. These objective functions combine the speed up of products that are behind (tardiness) their cycle time targets and

the slow down of products that are ahead (earliness) their cycle time targets. These objectives are presented below:

– In the first objective function (5.2), linear costs on both tardiness and earliness are used. The goal is to speed up and slow down linearly quantities of products behind and ahead their cycle time targets.

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{R} (f_1^P - f_2^P) Z_{glpr}^P + \sum_{r=1}^{p} (f_1^F - f_2^F) Z_{glpr}^F \right) \tag{5.2}$$

– In the second objective function (5.3), quadratic costs on tardiness and linear costs on earliness are used. This aims to speed up quadratically quantities of products behind their cycle time targets and to slow down linearly quantities of products ahead their cycle time targets. As the crucial goal is to meet the cycle time targets of products, with quadratic costs on tardiness, the objective function strongly penalizes the large delays on cycle time targets while minimizing at the same time earliness.

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{R} ((f_1^P)^2 - f_2^P) Z_{glpr}^P + \sum_{r=1}^{p} ((f_1^F)^2 - f_2^F) Z_{glpr}^F \right) \tag{5.3}$$

– The third objective function (5.4) strongly penalizes both tardiness and earliness, so that the cycle time targets of products are met with minimum positive and negative deviations. Quadratic costs are used for both tardiness and earliness to speed up quantities of products behind their cycle time targets and to slow down quantities of products ahead their cycle time targets.

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{R} ((f_1^P)^2 + (f_2^P)^2) Z_{glpr}^P + \sum_{r=1}^{p} ((f_1^F)^2 + (f_2^F)^2) Z_{glpr}^F) \right) \tag{5.4}$$

Note that $f_1^P = \max(0, p + (r)A - T_{gl})$ corresponds to the tardiness on the cycle time target of blocks of product g in operation l released in period $-r$ in the past, and $f_1^F = \max(0, p - r - T_{gl})$ corresponds to the tardiness on the cycle time target of blocks of product g in operation l released in period r in the future. In the same way, $f_2^P = \min(0, p + (r)A - T_{gl})$ corresponds to the earliness on the cycle time target of blocks of product g in operation l released in period $-r$ in the past, and $f_2^F = \min(0, p - r - T_{gl})$ corresponds to the earliness of product g in operation l released in period r in the future.

Note that the earliness is penalized positively in objective functions (5.2), (5.3) and (5.4). Indeed, the larger the Work-In-Process Z_{glpr} of product g at operation l at the end of period p released in the past (period $-r$) or released in the future (period r), the larger the earliness. If the earliness is negatively penalized, the optimization model will prioritize the processing of products that are ahead and with short cycle times to increase Z_{glpr} at downstream operations. As some products are faster than others (because of their short cycle times or their small number of operations), the processing of faster products will be prioritized. However, the positive penalization of the earliness prevents the optimization

model to keep large quantities of products ahead their cycle time targets. In addition, the objective function is optimized when the quantities of products are on time.

Because of the quality of the results obtained in sections 4.3.6.3, 50 blocks of operations are considered throughout the remaining experiments of this chapter. Objective function (5.2), (5.3) and (5.4) are used with constraints (4.19 - 4.33) presented in section 4.3.4.

5.2.2 Computational Experiments and Analysis

This section uses the same instances and the experiment configuration as those used in Chapter 4. Section 5.2.2.1 presents the numerical results on the first instance, while Section 5.2.2.2 presents the numerical results on the second instance. Finally, a general discussion on the results is provided in Section 5.2.2.3.

5.2.2.1 Numerical results on instance 1

Tables 5.1 and 5.3 show that the cycle times are not under control when the tardiness and earliness are penalized identically, either linearly or quadratically. In Table 5.1, the maximum positive and negative gaps on cycle time targets are very high, respectively 38.7% and -54.1%. The same observation can be made when looking at Table 5.3, where the maximum positive and negative gaps on cycle time targets are, respectively 49.7% and -60.6%.

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	47.1	46.2	23.4	45.9	54.1
Cycle Time Gaps	-1.8%	-7.6%	**-54.1%**	15.0%	**38.7%**

Table 5.1: Results based on objective function (5.2), instance 1

Table 5.2 presents results when the tardiness is penalized quadratically and the earliness is penalized linearly. The cycle times are better managed compared with the results of Tables 5.1 and 5.3. Compared with results when only the tardiness is linearly penalized (objective function (4.18)), almost all the negative gaps on cycle time targets have increased, except for product 5 where the negative gap is decreased by 3.4%. In addition, not all cycle time targets are met, product 4 has small positive gap of 3%.

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	39.9	44.2	44.5	41.3	37.0
Cycle Time Gaps	-16.8%	-11.6%	-12.7%	**3%**	-5.1%

Table 5.2: Results based on objective function (5.3), instance 1

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	52.2	46.9	20.1	55.4	58.4
Cycle Time Gaps	8.7%	-6.2%	**-60.6%**	39.0%	**49.7%**

Table 5.3: Results based on Objective function (5.4), instance 1

5.2.2.2 Numerical results on instance 2

As in the first instance, Tables 5.4 and 5.6 show that the cycle times are not under control when the tardiness and earliness are penalized identically, either linearly or quadratically. Compared with results when only the tardiness is linearly penalized (objective function (4.18)), in Table 5.4, the maximum negative gap is decreased by 5.2%. However, products 6 and 7 have positive gaps of respectively 1.2% and 5.7%. In Table 5.6, the maximum negative gap is decreased by 3.8%, but there are positive gaps of 10.6% and 3.1% respectively for products 6 and 9.

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77
Average Cycle Times (days)	31.4	62.4	43.1	29.8	66.1
Cycle Time Gaps	**1.2%**	**5.7%**	-6.3%	-6.9%	-14.2%

Table 5.4: Results based on objective function (5.2), instance 2

Table 5.5 presents results when the tardiness is penalized quadratically and the earliness is penalized linearly. The cycle times are better managed in comparison to the results in Tables 5.4 and 5.6. In addition, the results in Table 5.5 are better than the results when only the tardiness is linearly penalized (objective function (4.18)). The maximum negative gap is decreased by 5% and all the cycle time targets are satisfied.

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77
Average Cycle Times (days)	31.0	58.0	44.1	27.4	74.0
Cycle Time Gaps	0.0%	-1.7%	-4.1%	**-14.4%**	-3.9%

Table 5.5: Results based on objective function (5.3), instance 2

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77
Average Cycle Times (days)	34.3	59.0	43.5	33.0	65.0
Cycle Time Gaps	**10.6%**	0.0%	-5.4%	**3.1%**	-15.6%

Table 5.6: Results based on objective function (5.4), instance 2

5.2.2.3 Discussion

In summary, the results of this section show that the cycle times are not better managed when the tardiness and earliness are penalized in the same way in a single objective function, either linearly or quadratically (objective functions (5.2) and (5.4)). This is due to the structure of semiconductor manufacturing systems which include re-entrant flows and the competition on the same resources between products at different stages of their routes . The speed up of some products implies the slow down of other products. However, if a product can be ahead of its cycle time target in block b but behind its cycle time target in block $b+1$, delaying the product in block b can increase its delay in block $b+1$ which can cause a situation where its late acceleration will not be sufficient to meet its cycle time target. The same goes for products behind their cycle time targets in block b but ahead of their cycle time targets in the next block $b+1$. Speeding up the products in block b can lead to the situation where it is difficult to delay them so that they are close to their cycle time targets. Thus, some products are earliers and other are later when the products are sped up and slowed down in the same way in the single objective function.

Another reason is that, in the simulation, if all the production targets are met in period p, the simulation model continues to produce according to the FIFO dispatching rules. This is normal in practice if capacity is available. It can therefore happen, even if the global scheduling model does not send production targets for a product that the dispatching rule prioritizes this product in the queue line of the work-center if all production targets are met and machines in the work-centers are available. This problem can be solved using more sophisticated dispatching rules than FIFO at the local level.

The goal being to remain under or equal to the cycle time targets of the products, the results Section 5.2.2.1 show also that, when the tardiness and the earliness are optimized in the same objective function and when the former is more penalized than the latter, the cycle time targets are better managed, see Tables 5.2 and 5.5 where the tardiness is quadratically penalized and the earliness is linearly penalized (objective function (5.3)).

5.3 Lexicographic Approach

5.3.1 Modeling

The lexicographic approach is introduced in Chapter 3. In this section, the tardiness on cycle time targets is penalized in the first stage of the lexicographic approach, while the earliness on cycle time targets is penalized in the second stage. As the goal is to have the cycle times of all the products under or equal to their cycle time targets, the objective function which penalizes tardiness is considered as the most important objective. In this work, the lexicographic approach aims to reduce the positive and negative gaps on cycle time targets.

Objective function (5.5) corresponds to the first stage of the lexicographic approach when linear penalty costs on tardiness are used, while objective function (5.6) corresponds to the second stage when linear penalty costs on the earliness are used.

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{R} f_1^P Z_{glpr}^P + \sum_{r=1}^{p} f_1^F Z_{glpr}^F \right) \tag{5.5}$$

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{R} (-f_2^P) Z_{glpr}^P + \sum_{r=1}^{p} (-f_2^F) Z_{glpr}^F \right) \tag{5.6}$$

Objective function (5.7) corresponds to the first stage when quadratic penalty costs on tardiness are used, while objective function (5.8) corresponds to the second stage when quadratic penalty costs on earliness are used.

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{R} (f_1^P)^2 Z_{glpr}^P + \sum_{r=1}^{p} (f_1^F)^2 Z_{glpr}^F \right) \tag{5.7}$$

$$Min \sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{R} (f_2^P)^2 Z_{glpr}^P + \sum_{r=1}^{p} (f_2^F)^2 Z_{glpr}^F \right) \tag{5.8}$$

Three different combinations of tardiness and earliness penalizations are used:

- **Scenario 1**. Tardiness is penalized linearly in the first stage and earliness is penalized linearly in the second stage,
- **Scenario 2**. Tardiness is penalized quadratically in the first stage and earliness is penalized linearly in the second stage,
- **Scenario 3**. Tardiness is penalized quadratically in the first stage and earliness is penalized quadratically in the second stage

The two stages are solved using Constraints (4.19 - 4.33) introduced in chapter 4, Section 4.3.4. However, the second stage has additional constraints. Let S^* be the optimal solution of the first stage. Constraints (5.9) are added when the tardiness is penalized linearly in the first stage, while Constraints (5.10) are added when the tardiness is penalized quadratically in the first stage.

$$\sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{R} f_1^P Z_{glpr}^P + \sum_{r=1}^{p} f_1^F Z_{glpr}^F \right) \leq S^* \tag{5.9}$$

$$\sum_{g\in\mathcal{G}} \sum_{l\in\mathcal{L}_g} \sum_{p=1}^{H} \left(\sum_{r=1}^{R} (f_1^P)^2 Z_{glpr}^P + \sum_{r=1}^{p} (f_1^F)^2 Z_{glpr}^F \right) \leq S^* \tag{5.10}$$

5.3.2 Computational Experiments and Analysis

This section uses the same instances and experiment configuration than those used in Chapter 4. The numerical results based on different combinations of tardiness and earliness penalization scenarios as presented in Section 5.3.1 are analyzed. Section 5.3.2.1 presents numerical results based on the first instance, while Section 5.3.2.2 presents numerical results based on the second instance.

5.3.2.1 Numerical Results on Instance 1

Table 5.7 presents the results obtained when tardiness and earliness are penalized linearly, respectively in the first stage and second stage. Almost all cycle times are under their cycle time targets except product 4 with a small positive gap of 3%. However, the maximum negative gap on cycle times is still very high (-16.0%).

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	39.4	43.7	44.5	41.3	37.1
Cycle Time Gaps	**-16.8%**	-11.6%	-12.7%	**3%**	-5.1%

Table 5.7: Instance 1, linear penalty costs on tardiness in stage 1, linear penalty costs on earliness in stage 2

Table 5.8 shows results when the tardiness is penalized quadratically in the first stage while the earliness is penalized linearly in the second stage. All product cycle times are under their cycle time targets. Compared to the results in Table 5.7, the positive gap of product 4 reaches 0.0%. In addition, the maximum negative gap decreases from -16.8% in the Table 5.7 to -15.2% in Table 5.8.

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	40.7	44.5	44.3	40.0	38.5
Cycle Time Gaps	**-15.2%**	-11.0%	-13.1%	0.0%	-1.3%

Table 5.8: Instance 1, quadratic penalty costs on tardiness in stage 1, linear penalty costs on earliness in stage 2

Table 5.9 shows results when both the tardiness and the earliness are penalized quadratically in the first and second stages. Almost all product cycle times are under their cycle time targets. The maximum negative gap decreases from -15.2% in Table 5.8 to -14.0% in Table 5.9.

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	48.0	50.0	51.0	40.0	39.0
Average Cycle Times (days)	41.1	46.0	44.4	40.0	37.2
Cycle Time Gaps	**-14.4%**	-8.0%	-12.9%	0.0%	-4.6%

Table 5.9: Instance 1, quadratic penalty costs on tardiness in stage 1, quadratic penalty costs on earliness in stage 2

Changing cycle time targets, instance 1

In the following experiments, we modify the product cycle time targets. The goal is to illustrate that our multi-objective global scheduling approach helps to control cycle times. Table 5.10 presents results when the tardiness and earliness are penalized linearly in the first and second stages. All product cycle time targets are satisfied with small negative deviations, except for product 4 where a positive gap of 5.0% is observed. Compared to Table 5.10, the results in Table 5.11 show that the maximum positive and negative gaps on cycle time targets are decreased, respectively from 5.0% to 2.9% and from -4.2% to -3.1% when the tardiness is penalized quadratically in the first stage and the earliness is penalized linearly in the second stage .

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	38.0	44.0	43.0	42.0	38.0
Average Cycle Times (days)	36.4	42.9	41.6	41.2	39.9
Cycle Time Gaps	**-4.2%**	-2.5%	-3.2%	-2.0%	**5.0**%

Table 5.10: Instance 1 with linear penalty costs on tardiness in stage 1 and linear penalty costs on earliness in stage 2, changing cycle time targets

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	38.0	44.0	43.0	42.0	38.0
Average Cycle Times (days)	36.8	43.1	43.0	41.0	39.1
Cycle Time Gaps	**-3.1%**	-2.0%	-0.0%	-2.4%	**2.9**%

Table 5.11: Instance 1 with quadratic penalty costs on tardiness in stage 1 and linear penalty costs on earliness in stage 2, changing cycle time targets

Compared to Table 5.11, the results in Table 5.12 show that the maximum positive and negative gaps of cycle times are decreased, respectively from 2.9% to 0.0% and from -3.1% to -2.3% when the tardiness and earliness are penalized quadratically, respectively in the first and second stages. The cycle times of all products are satisfied.

Let us compare the results in Table 5.12 with the results when the tardiness is penalized quadratically in the global scheduling model using a single objective function. The results of

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	38.0	44.0	43.0	42.0	38.0
Average Cycle Times (days)	37.1	43.8	43.0	41.4	38.0
Cycle Time Gaps	**-2.3%**	-0.4%	-0.0%	-1.4%	**0.0%**

Table 5.12: Instance 1 with quadratic penalty costs on tardiness on stage 1 and quadratic penalty costs on earliness, changing cycle time targets

the multi-objective global scheduling model in Table 5.12 outperform the results of the single objective global scheduling model because the cycle time targets of all products are satisfied. In addition, in Table 5.12 the maximum negative gap on cycle time targets is -2.3%, which is smaller than -3.6% in Table 5.13.

	Products				
	1	**2**	**3**	**4**	**5**
Cycle Time Targets (days)	38.0	44.0	43.0	42.0	38.0
Average Cycle Times (days)	36.6	42.1	43.9	41.7	37.9
Cycle Time Gaps	-3.6%	-4.3%	**2.1%**	-0.7%	-0.3%

Table 5.13: Instance 1, quadratic penalty costs on tardiness in single objective function, changing cycle time targets

5.3.2.2 Numerical Results on Instance 2

Table 5.14 presents results when the tardiness and earliness are penalized linearly, respectively in the first and second stages. Almost all product cycle times are under their cycle time targets except product 6 with a positive gap of 5.8%. However, the maximum negative gap on cycle times is still very high (-11.8%).

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77.0
Average Cycle Times (days)	32.8	58.6	42.7	28.2	77.0
Cycle Time Gaps	**5.8%**	-0.7%	-7.2%	**-11.8%**	0.0%

Table 5.14: Instance 2, linear penalty costs on tardiness in stage 1, linear penalty costs on earliness in stage 2

Table 5.15 shows results when the tardiness is penalized quadratically in the first stage while the earliness is penalized linearly in the second stage. Almost all product cycle times are under their cycle time targets except products 6 and 10 with positives gaps of 3.2% and 1.9% respectively. Compared with the results in Table 5.14, the maximum positive and negative gaps are smaller, respectively decreasing from 5.8% to 3.2% and from -11.8% to -10.3%.

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77.0
Average Cycle Times (days)	32.0	58.3	43.2	28.7	78.5
Cycle Time Gaps	**3.2%**	-1.2%	-6.1%	**-10.3%**	1.9%

Table 5.15: Instance 2, quadratic penalty costs on tardiness in stage 1, linear penalty costs on earliness in stage 2

Table 5.16 shows results when the tardiness and earliness are penalized quadratically, respectively in the first and second stages. All product cycle times are under their cycle time targets. Compared with results in Table 5.15, the maximum positive and negative gaps change, respectively, from 3.2% to 0.0% and from -10.3% to -9.0%.

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	31.0	59.0	46.0	32.0	77.0
Average Cycle Times (days)	31.0	58.3	43.7	29.1	77.0
Cycle Time Gaps	**0.0%**	-1.2%	-5.0%	**-9.0%**	0.0%

Table 5.16: Instance 2, quadratic penalty costs on tardiness in stage 1, quadratic penalty costs on earliness in stage 2

Changing cycle time targets on instance 2

As in Section 5.3.2.1, in the following experiments, we modify the product cycle time targets. Table 5.17 presents results when the tardiness and earliness are penalized linearly, respectively in the first and second stages. The negative and positive gaps of cycle times are -16.1% and 0.7% respectively.

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	35.0	62.0	48.0	28.0	75.0
Average Cycle Times (days)	32.7	61.7	43.9	23.5	75.5
Cycle Time Gaps	-6.6%	-0.5%	-8.0%	**-16.1%**	**0.7**%

Table 5.17: Instance 2 with linear penalty costs on tardiness in stage 1 and linear penalty costs on earliness in stage 2, changing cycle time targets

Compared with the results in Table 5.17, Table 5.18 show that the maximum positive and negative gaps of cycle times decrease, respectively from 0.7% to 0.5% and from -16.1% to -6.3%.

Compared with the results in Table 5.18, the maximum positive and negative gaps of cycle times in Table 5.19 decrease respectively, from 0.5% to 0.0% and from -6.3% to -5.4%. The cycle times of all products are under their cycle time targets.

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	35.0	62.0	48.0	28.0	75.0
Average Cycle Times (days)	32.8	59.5	45.6	26.9	75.4
Cycle Time Gaps	**-6.3%**	-4.0%	-5.0%	-3.9%	**0.5**%

Table 5.18: First stage: Quadratic penalty costs on tardiness, second stage: Linear penalty costs on earliness, changing cycle time targets, instance 2

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	35.0	62.0	48.0	28.0	75.0
Average Cycle Times (days)	33.1	58.9	45.8	27.1	75.0
Cycle Time Gaps	-5.4%	**-5.0%**	-4.6%	-3.2%	**0.0%**

Table 5.19: Instance 2 with quadratic penalty costs on tardiness in stage 1 and quadratic penalty costs on earliness in stage 2, changing cycle time targets

Let us compare the results in Table 5.19 with the results when the tardiness is penalized quadratically in the global scheduling model using a single objective function. The results of the multi-objective global scheduling model in Table 5.19 outperform the results of the single objective global scheduling model in terms of positive and negative gaps on cycle times. There are no positive gaps in Table 5.19 and the maximum negative gap is equal to -5.4%, while, in Table 5.20, the cycle time targets of all products are not satisfied and the maximum negative gaps is equal to -14.4%.

	Products				
	6	**7**	**8**	**9**	**10**
Cycle Time Targets (days)	35.0	62.0	48.0	28.0	75.0
Average Cycle Times (days)	31.5	57.7	41.1	24.0	75.8
Cycle Time Gaps	-10.0%	-6.9%	**-14.4%**	-14.3%	1.1%

Table 5.20: Instance 2, quadratic penalty costs on tardiness in single objective function, changing cycle time targets

5.4 Conclusions

This chapter has discussed multi-objective optimization strategies for cycle time control. Penalty costs on tardiness and earliness are used in two different approaches. In the weighted sum approach, the objective functions that penalize tardiness and earliness are combined in the same objective function. A lexicographic approach is also proposed, where the objective function that penalizes tardiness is considered as the most important objective, thus used in the first stage, while the objective function that penalizes earliness is used in the second stage. The results show that the lexicographic approach provides better results than weighted sum

approach. The lexicographic approach allows not only the satisfaction of all product cycle time targets, but it also reduces the negative and positive gaps on the cycle time targets.

Chapter 6

General Conclusions and Perspectives

6.1 General Conclusions

Due to the complexity of semiconductor manufacturing, scheduling decisions are generally managed by either simple dispatching rules or dedicated scheduling algorithms at each work-center. However, this local management, which only considers local information at the level of each work-center, does not take into account the interactions between the different work-centers. This may lead to unbalanced flows in the factory. Some work-centers are more congested than others, which implies a deterioration of some critical KPIs.

To effectively manage scheduling decisions, this thesis proposes a global scheduling approach based on a two-level structure of the operational decision level:

- The global level (factory level), on the basis of global information from the factory, optimizes the Work-In-Process and provides production targets, i.e., production quantities to complete for each product, at each operation and at each period on a scheduling horizon and,
- The local level (work-center level), which manages the scheduling decisions using simple dispatching rules (First-In-First-Out rule) and a *Production Target Dispatching Rule* (PTDR), i.e., a mechanism to ensure that production targets received from the global level are followed at the local level.

Production targets are the main mechanism for steering scheduling decisions at work-center level. The main contributions of the thesis are based on the proposed approach, i.e., the global scheduling approach and on its evaluation.

Global scheduling approach

The global scheduling approach includes the principle of the approach, i.e., the determination of the production targets to be followed at work-center level and to be updated regularly, as well as the strategies to follow (global scheduling strategies). These strategies are based on Work-In-Process management techniques and are implemented through Linear Programming models (global scheduling models). They aim to optimize different objectives such as:

- Output variability on cycle times and throughput. This objective ensures that all the products move forward properly in their production stages. It prevents some products from slowing down others when no particular product is prioritized. This objective is optimized using a new Work-In-Process management strategy based on balancing

coefficients, i.e., a defined percentage of the Work-In-Process on the total Work-In-Process of the system that should remain in the system at the end of each period for each product. The balancing coefficients define the flow rate with which the production of each product should be carried out.

- Throughput (number of wafers produced). This objective is optimized together with the output variability on cycle times and throughput in a multi-objective strategy. The strategy implemented for the maximization of throughput is based on a Work-In-Process management technique called pull technique. This strategy ensures that the more products advance in their production route, the more important they become to exit the system as quickly as possible and increase the number of wafers produced.

- Cycle time (total time product spends in the production system). First, the cycle time is minimized with a strategy which consists in processing quantities of products differently according to their arrival time in an operation. A product which arrives at time $t-1$ in an operation does not have the same holding cost as the product which arrives at time t in the same operation. Second, cycle times are controlled to satisfy pre-defined cycle time targets given as input parameters. Cycle time targets can be determined by historical data or by simulation or defined by production managers. A global scheduling strategy is implemented on the basis of three key variables: The release dates, the temporal tracing of the Work-In-Process, and the cycle time targets. To control the cycle time in the production line, the operations in a route are subdivided into subsequences of operations (blocks of operations). Cycle time targets are assigned to each block. The control of cycle times is carried out using single objective and multi-objective optimization. In the single-objective optimization, the tardiness on the cycle time targets of blocks is minimized. In the multi-objective optimization, we first minimize the tardiness on the cycle time targets of blocks, and then we minimize the earliness on the cycle time target of blocks.

The global scheduling approach also includes parameters (global scheduling parameters) such as the scheduling horizon, the length of each period in the scheduling horizon and the time when the global scheduling model is solved again (triggering horizon).

Evaluation of the global scheduling approach

The evaluation of the global scheduling approach is made possible by using a simulation model representing the factory. The global scheduling strategy is called at each end of the triggering horizon during the simulation and provides production targets to be completed. A control variable mechanism for monitoring these production targets is implemented in the simulation model. It ensures that a product g remains in the state of production as long as its product target has not yet been reached. Once it production target is reached, the production of product g is stopped, and will only resume if the production targets for all other products are reached or if product g is alone in the queue of a resource.

The computational experiments carried out in this thesis were performed on industrial data. The results obtained show that the production targets provided by the different global scheduling strategies and sent to the local level optimize different objectives even when a simple dispatching rule is used at the local level. The proposed approach and the results obtained are promising and consistent. However, there are some features that need to be further developed and explored, and which are discussed in Section 6.2.

6.2 Perspectives

We classify the perspectives in two parts, the global scheduling approach in Section 6.2.1 and the evaluation of the approach in Section 6.2.2.

6.2.1 Global Scheduling Approach

In this section, the main perspectives are based on the global scheduling strategies used to optimize various factory objectives and the parameters used to evaluate the approach. Since the global scheduling model is called regularly, various parameters are required for its evaluation. These parameters include the scheduling horizon, the triggering horizon (the moment when the global scheduling strategy is regularly called) and the length of the period in the scheduling horizon. It would be interesting to perform an in-depth study of the impact of these parameters. The objective will be to determine if a single configuration of these parameters is sufficient for all strategies or if a particular configuration is required for each strategy. This task can be accomplished by performing additional experiments for each global scheduling strategy.

Chapter 4 explained the major disadvantage of using static balancing coefficients throughout the production horizon to control the Work-In-Process in the system. These balancing coefficients may or may not give expected results if the system is in a disruption state, which can be caused by machine breakdowns, discontinuous and non-uniform releases of products in the system,etc. This may cause some products to slow down while others are accelerated. The consequence is that some products will not respect the flow given by the balancing coefficients. To overcome this drawback, a dynamic determination of the balancing coefficients is necessary. The balancing coefficients will be adjusted according to the release scheme and the progress of each product in the system. This will allow a better control of the Work-In-Process, especially when there are disruptions. Optimization or specific algorithms can be used to dynamically determine balancing coefficients before each call of the global scheduling model.

In chapters 5 and 6, strategies for controlling cycle times have been proposed. In the short-term, it would be interesting to study new multi-objective optimization model with other objective functions such as the minimization of the variability of cycle times, the maximization of the number of moves and the use of production capacities, etc. Next, implementing new multi-objective approaches such as goal programming or desirability functions will helps to compare the numerical results obtained with those of the lexicographic approach. Beyond parameters such as the release dates and the temporal tracing of the Work-In-Process, the way the cycle time targets of products are distributed on the different blocks of operations in the route influences the results of global scheduling strategies. In this thesis, the cycle time targets of blocks of operations are static. Additional insights in the control of cycle times can be obtained by using dynamic cycle time targets for blocks of operations. This can be achieved by using dedicated algorithms. Another important aspect to be addressed for cycle time control is the study of a particular allocation of the number of blocks of operations for each product such that products with different number of operations in their routes have different numbers of blocks of operations. In addition, in the same spirit of cycle time control, release dates are among the important parameters used when modeling the global scheduling strategy for cycle time control. In the present work, release dates are aggregated into release classes on the basis of 21 periods (one week). It would be interesting to extend the study of the aggregation of release dates in release classes by reducing the

number of periods in each release class.

The fact that products share the same resources in semiconductor manufacturing influences the flow of each product and especially the cycle times. It may be interesting to make a preliminary study on the product mix, like the one in Chang (2016). The result of this study can be used as an additional parameter in the global scheduling model for cycle time control. There are also other factors impacting the cycle times in semiconductor manufacturing and which could be taken into account when modeling. Among these factors we have for example, the time constraints which can be important to take into account when allocating cycle time targets to the blocks of operations. Let us recall that a time constraint corresponds to the maximum time allowed between two operations, which are consecutive or not, to ensure yield and quality (Lima et al. (2019) and Lima et al. (2020)).

6.2.2 Evaluation of the Global Scheduling Approach

Different perspectives are possible for the evaluation of the global scheduling approach with the simulation model. The short-term and long-term perspectives are based on how the production targets provided by the global scheduling strategies are followed at work-center level, the environment and the configuration of the evaluation of the approach in the simulation model. In the short term, we could study:

- The monitoring of production targets in a dependent scheme period after period. In the experiments carried out in this thesis, the production targets are monitored independently in each period, i.e., if there are still production targets at the end of period p, these production targets are not taken into account in $p + 1$. The simulation only tracks the production targets dedicated to a current period. It would be interesting to experiment a monitoring of production targets, correlating multiple periods.

- The evaluation of the performance of the global scheduling approach using new dispatching rules or dedicated scheduling algorithms at each work-center such as the one proposed in Tamssaouet (2019).

In the long term two perspectives are discussed below:

- The evaluation of the approach in a transient environment. For the computational experiments carried out in this thesis, the simulation model was in a steady state. It would be interesting to add disturbances such as machines breakdowns to create more variability in the simulation without the global scheduling strategy. The objective would be to observe if this variability is better controlled when the global scheduling approach is used. In the same spirit, it would also be interesting to test the effectiveness of the global scheduling approach by introducing engineering lots during the simulation. The goal is to observe the impact on cycle times and variability of the global scheduling approach coupled to the simulation model compared with the behavior of the simulation model without the global scheduling approach. Note that, in semiconductor manufacturing, engineering lots are usually given higher priority in the manufacturing process in order to improve the manufacturing process and/or to facilitate the development of new products (Chang (2016)).

- The evaluation of the approach on several product families. In this thesis, the global scheduling strategies are evaluated on a set of five product families. Adding more product families can add complexity to solve the global scheduling models as this

will require more memory space and computational time. Proposing new solution approaches can help handling this complexity. These approaches can be a step-by-step resolution on a limited number of periods, then set the values of the variables on these periods and then solve again on the remaining periods. The complexity can also be reduced by carrying out experiments to select representative operations in the route of each product instead of optimizing the production on its entire route.

A study on product release policies would also be interesting to investigate. One of the difficult parts of production planning is the release mechanism of products into the factory (see Rezaie et al. (2009)). In our case, the product release mechanism is used as input parameters. Studying different release policies will allow to understand their effects on global scheduling strategies. A review on product release policies can be found in Li et al. (2011).

In a global vision of the approach, the outputs of the global scheduling models can be thought differently, instead of sending only production targets to be completed, the global scheduling approach can reinforce the interactions between the different work-centers by sending additional information. To do this, the global scheduling approach could share with each work-center the upstream and downstream Work-In-Process information. This information could then be used to build new dispatching rules at the local level. It would also be interesting to extend the simulation model by adding, for example, cluster machines, batching machines and the transportation system. A library dedicated to Automated Material Handling Systems (AMHS) is available in the latest version of the AnyLogic software.

Manufacturers are now talking about artificial intelligence, Internet of Things, cloud computing and other sophisticated high-tech tools for better decision-making in industries. It would be interesting to get real-time factory information in the simulation model, and thus get a digital twin, when evaluating global scheduling strategies.

List of Figures

List of Tables

Bibliography

Ab Rahim, S. R., Ahmad, I. and Chik, M. A. (2012). Technique to improve visibility for cycle time improvement in semiconductor manufacturing, *2012 10th IEEE International Conference on Semiconductor Electronics (ICSE)*, IEEE, pp. 627–630.

Akçali, E., Uzsoy, R., Hiscock, D. G., Moser, A. L. and Teyner, T. J. (2000). Alternative loading and dispatching policies for furnace operations in semiconductor manufacturing: a comparison by simulation, *Proceedings of the 32nd conference on Winter simulation*, Society for Computer Simulation International, pp. 1428–1435.

Akcalt, E., Nemoto, K. and Uzsoy, R. (2001). Cycle-time improvements for photolithography process in semiconductor manufacturing, *IEEE Transactions on Semiconductor Manufacturing* **14**(1): 48–56.

Akhavan-Tabatabaei, R., Ding, S. and Shanthikumar, J. G. (2009). A method for cycle time estimation of semiconductor manufacturing toolsets with correlations, *Winter Simulation Conference*, pp. 1719–1729.

Arisha, A. and Young, P. (2005). Simulation in semiconductor manufacturing facilities.

Arisha, A., Young, P. and El Baradie, M. (2004). A simulation model to characterize the photolithography process of a semiconductor wafer fabrication, *Journal of materials processing technology* **155**: 2071–2079.

Babbs, D. and Gaskins, R. (2007). Effectiveness of small batch size on cycle time reduction in a conventional 300mm factory, *2007 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, IEEE, pp. 105–110.

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.

Barbosa, C. and Azevedo, A. (2017). Hybrid simulation for complex manufacturing value-chain environments, *Procedia Manufacturing* **11**: 1404–1412.

Barhebwa-Mushamuka, F., Dauzère-Pérès, S. and Yugma, C. (2019a). Multi-objective optimization for work-in-process balancing and throughput maximization in global fab scheduling, *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, IEEE, pp. 697–702.

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

Ben-Salem, A., Yugma, C., Troncet, E. and Pinaton, J. (2016). Amhs design for reticles in photolithography area of an existing wafer fab: Ie: Industrial engineering, *Advanced Semiconductor Manufacturing Conference (ASMC), 2016 27th Annual SEMI*, IEEE, pp. 110–115.

Bitar, A., Dauzère-Pérès, S., Yugma, C. and Roussel, R. (2016). A memetic algorithm to solve an unrelated parallel machine scheduling problem with auxiliary resources in semiconductor manufacturing, *Journal of Scheduling* **19**(4): 367–376.

Boardman, J. and Sauser, B. (2006). System of systems-the meaning of of, *2006 IEEE/SMC International Conference on System of Systems Engineering*, IEEE, pp. 6–pp.

Bonal, J., Fernandez, M., Maire-Richard, O., Aparicio, S., Oliva, R., Gonzalez, S. G. B., Rodriguez, L., Rosendo, M., Villacieros, J. and Becerro, J. (2001). A statistical approach to cycle time management, *2001 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, IEEE, pp. 11–15.

Borshchev, A. (2013). *The big book of simulation modeling: multimethod modeling with AnyLogic 6*, AnyLogic North America.

Brown, S., Domaschke, J. and Leibl, F. (1998). Cycle time reductions for test area bottleneck equipment, *Proceedings of the Second Annual SEMI Test, Assembly, and Packaging Automation and Integration Conference*, pp. B1–B5.

Brown, S., Domaschke, J. and Leibl, F. (1999). No cost applications for assembly cycle time reduction, *International Conference on Semiconductor Manufacturing Operational Modeling and Simulation*, pp. 159–163.

Bureau, M., Dauzère-Pérès, S., Yugma, C. and Vermarien, L. (2007). An approach for simulating consistent global and local scheduling, *2007 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, IEEE, pp. 96–99.

Bureau, M., Dauzère-Pérès, S., Yugma, C., Vermariën, L. and Maria, J.-B. (2007). Simulation results and formalism for global-local scheduling in semiconductor manufacturing facilities, *Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come*, IEEE Press, pp. 1768–1773.

Butterbaugh, J. W. (2004). Strategies for cycle time reduction in batch cleaning, *2004 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, IEEE, pp. 52–56.

Byrne, M. and Bakir, M. A. (1999). Production planning using a hybrid simulation–analytical approach, *International Journal of Production Economics* **59**(1-3): 305–311.

Cardarelli, E. and Pelagagge, P. M. (1995). Simulation tool for design and management optimization of automated interbay material handling and storage systems for large wafer fab, *IEEE Transactions on Semiconductor Manufacturing* **8**(1): 44–49.

Chance, F., Robinson, J. and Fowler, J. W. (1996). Supporting manufacturing with simulation: model design, development, and deployment, *Proceedings of the 28th conference on Winter simulation*, IEEE Computer Society, pp. 114–121.

Chang, K.-H. (2016). Risk-controlled product mix planning in semiconductor manufacturing using simulation optimization, *IEEE Transactions on Semiconductor Manufacturing* **29**(4): 411–418.

Chang, S.-C. (1999). Demand-driven, iterative capacity allocation and cycle time estimation for re-entrant lines, *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No. 99CH36304)*, Vol. 3, IEEE, pp. 2270–2275.

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

Chen, T. (2013). A systematic cycle time reduction procedure for enhancing the competitiveness and sustainability of a semiconductor manufacturer, *Sustainability* **5**(11): 4637–4652.

Chien, C.-F., Hsiao, C.-W., Meng, C., Hong, K.-T. and Wang, S.-T. (2005). Cycle time prediction and control based on production line status and manufacturing data mining, *ISSM 2005, IEEE International Symposium on Semiconductor Manufacturing, 2005.*, IEEE, pp. 327–330.

Chien, C.-F., Hsu, C.-Y. and Hsiao, C.-W. (2012). Manufacturing intelligence to forecast and reduce semiconductor cycle time, *Journal of Intelligent Manufacturing* **23**(6): 2281–2294.

Chien, C.-F. and Hu, C.-H. (2006). Segmented wip control for cycle time reduction, *2006 IEEE International Symposium on Semiconductor Manufacturing*, IEEE, pp. 265–268.

Christopher, J., Kuhl, M. E. and Hirschman, K. (2005). Simulation analysis of dispatching rules for automated material handling systems and processing tools in semiconductor fabs, *Semiconductor Manufacturing, 2005. ISSM 2005, IEEE International Symposium on*, IEEE, pp. 84–87.

Chung, J. and Jang, J. (2009). A wip balancing procedure for throughput maximization in semiconductor fabrication, *IEEE Transactions on Semiconductor Manufacturing* **22**(3): 381–390.

Chung, S.-H. and Huang, H.-W. (2002). Cycle time estimation for wafer fab with engineering lots, *Iie Transactions* **34**(2): 105–118.

Collins, D. W., Lakshman, V. and Collins, L. (2001). Dynamic simulator for wip analysis in semiconductor manufacturing, *Semiconductor Manufacturing Symposium, 2001 IEEE International*, IEEE, pp. 71–74.

Dabbas, R. M. and Fowler, J. W. (2003). A new scheduling approach using combined dispatching criteria in wafer fabs, *IEEE Transactions on Semiconductor Manufacturing* **16**(3): 501–510.

Dauzère-Pérès, S. and Lasserre, J.-B. (2002). On the importance of sequencing decisions in production planning and scheduling, *International transactions in operational research* **9**(6): 779–793.

Dauzère-Péres, S. and Lasserre, J.-B. (2012). *An integrated approach in production planning and scheduling*, Vol. 411, Springer Science & Business Media.

Deb, K. (2001). *Multi-objective optimization using evolutionary algorithms*, Vol. 16, John Wiley & Sons.

Dequeant, K., Vialletelle, P., Lemaire, P. and Espinouse, M.-L. (2016). A literature review on variability in semiconductor manufacturing: the next forward leap to industry 4.0, *Proceedings of the 2016 Winter Simulation Conference*, IEEE Press, pp. 2598–2609.

Domaschke, J., Brown, S., Robinson, J. and Leibl, F. (1998). Effective implementation of cycle time reduction strategies for semiconductor back-end manufacturing, *1998 Winter Simulation Conference. Proceedings*, Vol. 2, IEEE, pp. 985–992.

Eberts, D., Keil, S., Peipp, F. and Lasch, R. (2015). Shortening of cycle time in semiconductor manufacturing via meaningful lot sizes, *2015 26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, IEEE, pp. 34–41.

Ehm, H., McGinnis, L. and Rose, O. (2009). Are simulation standards in our future?, *Winter Simulation Conference*, Winter Simulation Conference, pp. 1695–1702.

Ehrgott, M. (2005). *Multicriteria optimization*, Vol. 491, Springer Science & Business Media.

Ehteshami, B., Petrakian, R. G. and Shabe, P. M. (1992). Trade-offs in cycle time management: hot lots, *IEEE Transactions on Semiconductor Manufacturing* **5**(2): 101–106.

El Adl, M., Rodriguez, A. A. and Tsakalis, K. S. (1996). Hierarchical modeling and control of re-entrant semiconductor manufacturing facilities, *Decision and Control, 1996., Proceedings of the 35th IEEE Conference on*, Vol. 2, IEEE, pp. 1736–1742.

El-Khouly, I. A., El-Kilany, K. S. and El-Sayed, A. E. (2009). Modelling and simulation of re-entrant flow shop scheduling: An application in semiconductor manufacturing, *Computers & Industrial Engineering, 2009. CIE 2009. International Conference on*, IEEE, pp. 211–216.

Figueira, G. and Almada-Lobo, B. (2014). Hybrid simulation–optimization methods: A taxonomy and discussion, *Simulation Modelling Practice and Theory* **46**: 118–134.

Fordyce, K., Dalton, D., Gerard, B., Jesse, R. R., Sell, R. and Sullivan, G. G. (1992). Daily output planning: Integrating operations research, artificial intelligence, and real-time decision support with apl2, *Expert Systems with Applications* **5**(3–4): 245–256.

Fowler, J. W. and Rose, O. (2004). Grand challenges in modeling and simulation of complex manufacturing systems, *Simulation* **80**(9): 469–476.

Freitag, M. and Hildebrandt, T. (2016). Automatic design of scheduling rules for complex manufacturing systems by multi-objective simulation-based optimization, *CIRP Annals* **65**(1): 433–436.

Fronckowiak, D., Peikert, A. and Nishinohara, K. (1996). Using discrete event simulation to analyze the impact of job priorities on cycle time in semiconductor manufacturing, *Advanced Semiconductor Manufacturing Conference and Workshop, 1996. ASMC 96 Proceedings. IEEE/SEMI 1996*, IEEE, pp. 151–155.

Ghasemi, A., Heavey, C. and Laipple, G. (2018). A review of simulation-optimization methods with applications to semiconductor operational problems, *2018 Winter Simulation Conference (WSC)*, IEEE, pp. 3672–3683.

Gorod, A., Sauser, B. and Boardman, J. (2008). System-of-systems engineering management: A review of modern history and a path forward, *IEEE Systems Journal* **2**(4): 484–499.

Govind, N., Bullock, E. W., He, L., Iyer, B., Krishna, M. and Lockwood, C. S. (2008). Operations management in automated semiconductor manufacturing with integrated targeting, near real-time scheduling, and dispatching, *IEEE Transactions on Semiconductor Manufacturing* **21**(3): 363–370.

Govind, N. and Fronckowiak, D. (2003). Setting performance targets in a 300 mm wafer fabrication facility, *Advanced Semiconductor Manufacturing Conference and Workshop, 2003 IEEEI/SEMI*, IEEE, pp. 75–79.

Haimes, Y. (1971). On a bicriterion formulation of the problems of integrated system identification and system optimization, *IEEE transactions on systems, man, and cybernetics* **1**(3): 296–297.

Hassoun, M. (2013). On improving the predictability of cycle time in an nvm fab by correct segmentation of the process, *IEEE Transactions on Semiconductor Manufacturing* **26**(4): 613–618.

Hsieh, S.-J. T. (2002). Hybrid analytic and simulation models for assembly line design and production planning, *Simulation Modelling Practice and Theory* **10**(1-2): 87–108.

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 programmingÂ—simulation approaches for production planning, *IEEE Transactions on Semiconductor Manufacturing* **23**(3): 442–455.

Jaimes, A. L., Martınez, S. Z. and Coello, C. A. C. (2009). An introduction to multiobjective optimization techniques, *Optimization in Polymer Processing* pp. 29–57.

Jeong, S. J., Lim, S. J. and Kim, K. S. (2006). Hybrid approach to production scheduling using genetic algorithm and simulation, *The International Journal of Advanced Manufacturing Technology* **28**(1-2): 129–136.

Jimenez, J., Kim, B., Fowler, J., Mackulak, G., Choung, Y. I. and Kim, D.-J. (2002). Material handling: operational modeling and simulation of an inter-bay amhs in semiconductor wafer fabrication, *Proceedings of the 34th conference on Winter simulation: exploring new frontiers*, Winter Simulation Conference, pp. 1377–1382.

Johnzén, C., Dauzère-Pérès, S. and Vialletelle, P. (2011). Flexibility measures for qualification management in wafer fabs, *Production Planning and Control* **22**(1): 81–90.

Jung, C., Pabst, D., Ham, M., Stehli, M. and Rothe, M. (2014). An effective problem decomposition method for scheduling of diffusion processes based on mixed integer linear programming, *IEEE Transactions on Semiconductor Manufacturing* **27**(3): 357–363.

Kalisch, S., Ringel, R. and Weigang, J. (2008). Managing wip and cycle time with the help of loop control, *2008 Winter Simulation Conference*, IEEE, pp. 2298–2304.

Kao, Y.-T., Chang, C.-M. and Chang, S.-C. (2014). Do we still need daily production target setting in fully automated fabs?, *2014 e-Manufacturing & Design Collaboration Symposium (eMDC)*, IEEE, pp. 1–4.

Kao, Y.-T. and Chang, S.-C. (2018). Setting daily production targets with novel approximation of target tracking operations for semiconductor manufacturing, *Journal of Manufacturing Systems* **49**: 107–120.

Kiba, J.-E., Lamiable, G., Dauzère-Pérès, S. and Yugma, C. (2009). Simulation of a full 300mm semiconductor manufacturing plant with material handling constraints, *Simulation Conference (WSC), Proceedings of the 2009 Winter*, IEEE, pp. 1601–1609.

Kim, B.-I., Jeong, S., Shin, J., Koo, J., Chae, J. and Lee, S. (2009). A layout-and data-driven generic simulation model for semiconductor fabs, *IEEE Transactions on Semiconductor Manufacturing* **22**(2): 225–231.

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

Kim, Y.-D., Dong-Ho, L., Jung-Ug, K. and Roh, H.-K. (1998). A simulation study on lot release control, mask scheduling, and batch scheduling in semiconductor wafer fabrication facilities, *Journal of Manufacturing Systems* **17**(2): 107.

Kim, Y.-D., Kim, J.-G., Choi, B. and Kim, H.-U. (2001). Production scheduling in a semiconductor wafer fabrication facility producing multiple product types with distinct due dates, *IEEE Transactions on Robotics and Automation* **17**(5): 589–598.

Knopp, S., Dauzère-Pérès, S. and Yugma, C. (2017). A batch-oblivious approach for complex job-shop scheduling problems, *European Journal of Operational Research* **263**(1): 50–61.

Kohn, R., Noack, D., Mosinski, M., Zhou, Z. and Rose, O. (2009). Evaluation of modeling, simulation and optimization approaches for work flow management in semiconductor manufacturing, *Simulation Conference (WSC), Proceedings of the 2009 Winter*, IEEE, pp. 1592–1600.

Kong, S. H. (2007). Two-step simulation method for automatic material handling system of semiconductor fab, *Robotics and Computer-Integrated Manufacturing* **23**(4): 409–420.

Koo, P.-H., Park, M.-J. and Koh, S.-G. (2016). Simulation analysis of operational control decisions in semiconductor wafer fabrication, *Proc. ICAOR*, p. 102.

Kramer, S. S. (1989). Total cycle time management by operational elements, *IEEE/SEMI International Semiconductor Manufacturing Science Symposium*, IEEE, pp. 17–20.

Kuhl, M. E. and Laubisch, G. R. (2004). A simulation study of dispatching rules and rework strategies in semiconductor manufacturing, *IEEE/SEMI advanced semiconductor manufacturing conference*, pp. 4–6.

Leachman, R. C. and Ding, S. (2010). Excursion yield loss and cycle time reduction in semiconductor manufacturing, *IEEE Transactions on Automation science and engineering* **8**(1): 112–117.

Leachman, R. C., Kang, J. and Lin, V. (2002). Slim: Short cycle time and low inventory in manufacturing at samsung electronics, *Interfaces* **32**(1): 61–77.

LeBaron, H. T. and Pool, M. (1994). The simulation of cluster tools: a new semiconductor manufacturing technology, *Simulation Conference Proceedings, 1994. Winter*, IEEE, pp. 907–912.

Lee, B., Lee, Y., Yang, T. and Ignisio, J. (2008). A due-date based production control policy using wip balance for implementation in semiconductor fabrications, *International Journal of Production Research* **46**(20): 5515–5529.

Lee, Y. H. and Lee, B. (2003). Push-pull production planning of the re-entrant process, *The International Journal of Advanced Manufacturing Technology* **22**(11-12): 922–931.

Lee, Y. H., Park, J. and Kim, S. (2002). Experimental study on input and bottleneck scheduling for a semiconductor fabrication line, *IIE transactions* **34**(2): 179–190.

Li, S., Tang, T. and Collins, D. W. (1996). Minimum inventory variability schedule with applications in semiconductor fabrication, *IEEE Transactions on Semiconductor Manufacturing* **9**(1): 145–149.

Li, Y., Jiang, Z., Li, N. and Li, C. (2011). A review on release policies in semiconductor wafer fabrication system, *Industrial Engineering and Management* **16**(6): 108–114.

Liao, D.-Y., Chang, S.-C., Pei, K.-W. and Chang, C.-M. (1996). Daily scheduling for r&d semiconductor fabrication, *IEEE transactions on Semiconductor Manufacturing* **9**(4): 550–561.

Lima, A., Borodin, V., Dauzère-Pérès, S. and Vialletelle, P. (2019). Sampling-based release control of multiple lots in time constraint tunnels, *Computers in Industry* **110**: 3–11.

Lima, A., Borodin, V., Dauzère-Pérès, S. and Vialletelle, P. (2020). A sampling-based approach for managing lot release in time constraint tunnels in semiconductor manufacturing, *International Journal of Production Research* pp. 1–25.

Lin, J. and Long, Q. (2011). Development of a multi-agent-based distributed simulation platform for semiconductor manufacturing, *Expert systems with applications* **38**(5): 5231–5239.

Lin, J. T. and Chen, C.-M. (2015). Simulation optimization approach for hybrid flow shop scheduling problem in semiconductor back-end manufacturing, *Simulation Modelling Practice and Theory* **51**: 100–114.

Lin, Y.-H. and Lee, C.-E. (2001). A total standard wip estimation method for wafer fabrication, *European Journal of Operational Research* **131**(1): 78–94.

Lin, Y.-H., Shie, J.-R. and Tsai, C.-H. (2009). Using an artificial neural network prediction model to optimize work-in-process inventory level for wafer fabrication, *Expert Systems with Applications* **36**(2): 3421–3427.

Liu, C.-M., Kuo, C.-J. and Chi, C.-Y. (2006). A dynamic method for optimal wip allocation and control in a semiconductor manufacturing system, *2006 IEEE International Symposium on Semiconductor Manufacturing*, IEEE, pp. 61–65.

Liu, J., Li, C., Yang, F., Wan, H. and Uzsoy, R. (2011). Production planning for semiconductor manufacturing via simulation optimization, *Proceedings of the winter simulation conference*, Winter Simulation Conference, pp. 3617–3627.

Lu, S. C., Ramaswamy, D. and Kumar, P. (1994). Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants, *IEEE Transactions on Semiconductor Manufacturing* **7**(3): 374–388.

Lu, S., Ramaswamy, D. and Kumar, P. (1993). Scheduling semiconductor manufacturing plants to reduce mean and variance of cycle-time, *Proceedings. IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, IEEE, pp. 83–85.

Mack, C. A. (2005). Lithography simulation in semiconductor manufacturing, *Advanced Microlithography Technologies*, Vol. 5645, International Society for Optics and Photonics, pp. 63–84.

Mackulak, G. T., Lawrence, F. P. and Colvin, T. (1998). Effective simulation model reuse: a case study for amhs modeling, *Proceedings of the 30th conference on Winter simulation*, IEEE Computer Society Press, pp. 979–984.

Majorana, A. and Iuliano, G. (1997). Improving cycle time through managing variability in a dram production line, *1997 IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings*, IEEE, pp. A29–A32.

Marler, R. T. and Arora, J. S. (2004). Survey of multi-objective optimization methods for engineering, *Structural and multidisciplinary optimization* **26**(6): 369–395.

Mati, Y., Dauzère-Pérès, S. and Lahlou, C. (2011). A general approach for optimizing regular criteria in the job-shop scheduling problem, *European Journal of Operational Research* **212**(1): 33–42.

May, G. S. and Spanos, C. J. (2006). *Fundamentals of semiconductor manufacturing and process control*, John Wiley & Sons.

Meidan, Y., Lerner, B., Rabinowitz, G. and Hassoun, M. (2011). Cycle-time key factor identification and prediction in semiconductor manufacturing using machine learning and data mining, *IEEE Transactions on Semiconductor Manufacturing* **24**(2): 237–248.

Meyersdorf, D. and Yang, T. (1997). Cycle time reduction for semiconductor wafer fabrication facilities, *1997 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop ASMC 97 Proceedings*, IEEE, pp. 418–423.

Miettinen, K. (1999). Nonlinear multiobjective optimization, volume 12 of international series in operations research and management science.

Mittler, M., Schoemig, A. and Gerlich, N. (1995). Reducing the variance of cycle times in semiconductor manufacturing systems, *In International Conference on Improving Manufacturing Performance in a Distributed Enterprise: Advanced Systems and Tools*.

Mittler, M. and Schoemig, A. K. (1999). Comparison of dispatching rules for semiconductor manufacturing using large facility models, *WSC'99. 1999 Winter Simulation Conference Proceedings.'Simulation-A Bridge to the Future'*, Vol. 1, IEEE, pp. 709–713.

Miyashita, K., Okazaki, T. and Matsuo, H. (2004). Simulation-based advanced wip management and control in semiconductor manufacturing, *Proceedings of the 2004 Winter Simulation Conference, 2004.*, Vol. 2, IEEE, pp. 1943–1950.

Mönch, L., Fowler, J. W., Dauzère-Pérès, S., Mason, S. J. and Rose, O. (2011). A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations, *Journal of scheduling* **14**(6): 583–599.

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., Rose, O. and Sturm, R. (2003). A simulation framework for the performance assessment of shop-floor control systems, *Simulation* **79**(3): 163–170.

Mueller, R., Alexopoulos, C. and McGinnis, L. F. (2007). Automatic generation of simulation models for semiconductor manufacturing, *Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come*, IEEE Press, pp. 648–657.

Nayani, N. and Mollaghasemi, M. (1998). Validation and verification of the simulation model of a photolithography process in semiconductor manufacturing, *Proceedings of the 30th conference on Winter simulation*, IEEE Computer Society Press, pp. 1017–1022.

Ndiaye, M. A., Dauzère-Pérès, S., Yugma, C., Rullière, L. and Lamiable, G. (2016a). Automated transportation of auxiliary resources in a semiconductor manufacturing facility, *Proceedings of the 2016 Winter Simulation Conference*, IEEE Press, pp. 2587–2597.

Ndiaye, M. A., Dauzère-Pérès, S., Yugma, C., Rullière, L. and Lamiable, G. (2016b). Management of crisis situations in a large unified amhs of a semiconductor manufacturing facility: Ie: Industrial engineering, *Advanced Semiconductor Manufacturing Conference (ASMC), 2016 27th Annual SEMI*, IEEE, pp. 106–109.

Negahban, A. and Smith, J. S. (2014). Simulation for manufacturing system design and operation: Literature review and analysis, *Journal of Manufacturing Systems* **33**(2): 241–261.

Nemoto, K., Akcali, E. and Uzsoy, R. M. (2000). Quantifying the benefits of cycle time reduction in semiconductor wafer fabrication, *IEEE Transactions on Electronics Packaging Manufacturing* **23**(1): 39–47.

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.

Pierce, N. G. and Yost, A. (1995). Cycle time metrics for r&d semiconductor wafer fabrication, *Proceedings of SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, IEEE, pp. 105–110.

Potti, K., Bunch, T., Clark, C. and Wallers, K. (1994). Using simulation modeling to calculate wip levels in semiconductor manufacturing, *Proceedings of 1994 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop (ASMC)*, IEEE, p. 193.

Qi, C., Tang, T. K. and Sivakumar, A. L. (2002). Modeling methodology: simulation based cause and effect analysis of cycle time and wip in semiconductor wafer fabrication, *Proceedings of the 34th conference on Winter simulation: exploring new frontiers*, Winter Simulation Conference, pp. 1423–1430.

Rentmeesters, M. J., Tsai, W. K. and Lin, K.-J. (1996). A theory of lexicographic multi-criteria optimization, *Proceedings of ICECCS'96: 2nd IEEE International Conference on Engineering of Complex Computer Systems (held jointly with 6th CSESAW and 4th IEEE RTAW)*, IEEE, pp. 76–79.

Rezaie, K., Eivazy, H. and Nazari-Shirkouhi, S. (2009). A novel release policy for hybrid make-to-stock/make-to-order semiconductor manufacturing systems, *2009 Second international conference on developments in esystems engineering*, IEEE, pp. 443–447.

Robinson, J. and Chance, F. (2000). Wafer fab cycle time management using mes data, *Proceedings of the 2000 Modeling and Analysis for Semiconductor Manufacturing Conference (MASM 2000), Tempe, AZ*.

Robinson, J. K. et al. (2002). Understanding and improving wafer fab cycle times, *Semiconductor FabTech* **17**(April).

Rozen, K. and Byrne, N. M. (2016). Using simulation to improve semiconductor factory cycle time by segregation of preventive maintenance activities, *Proceedings of the 2016 Winter Simulation Conference*, IEEE Press, pp. 2676–2684.

Sada, T., Yuen, R. A., Ichikawa, M., Yamada, M. and Kabata, K. (2001). Simple tool of analysis for cycle time reduction, *2001 IEEE International Symposium on Semiconductor Manufacturing. ISSM 2001. Conference Proceedings*, IEEE, pp. 79–82.

Sadeghi, R. (2017). *Consistency of global and local scheduling decisions in semiconductor manufacturing*, PhD thesis, Ecole des Mines de Saint-Etienne.

Sadeghi, R., Dauzere-Pérès, S. and Yugma, C. (2016). A multi-method simulation modelling for semiconductor manufacturing, *IFAC-PapersOnLine* **49**(12): 727–732.

Sargent, R. G. (2013). Verification and validation of simulation models, *Journal of simulation* **7**(1): 12–24.

Sarin, S. C., Varadarajan, A. and Wang, L. (2011). A survey of dispatching rules for operational control in wafer fabrication, *Production Planning and Control* **22**(1): 4–24.

Schmidt, K. (2007). Improving priority lot cycle times, *2007 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, IEEE, pp. 117–121.

Schoemig, A. K. (1999). On the corrupting influence of variability in semiconductor manufacturing, *Proceedings of the 31st conference on Winter simulation: Simulation—a bridge to the future-Volume 1*, ACM, pp. 837–842.

Shannon, R. E. (1998). Introduction to the art and science of simulation, *Proceedings of the 30th conference on Winter simulation*, IEEE Computer Society Press, pp. 7–14.

Shanthikumar, J. G. and Sargent, R. G. (1983). A unifying view of hybrid simulation/analytic models and modeling, *Operations research* **31**(6): 1030–1052.

Shikalgar, S. T., Fronckowiak, D. and MacNair, E. A. (2002). 300 mm wafer fabrication line simulation model, *Simulation Conference, 2002. Proceedings of the Winter*, Vol. 2, IEEE, pp. 1365–1368.

Sivakumar, A. I. (2000). Simulation based cause and effect analysis of cycle time distribution in semiconductor backend, *Proceedings of the 32nd conference on Winter simulation*, Society for Computer Simulation International, pp. 1464–1471.

Spearman, M. L., Woodruff, D. L. and Hopp, W. J. (1990). Conwip: a pull alternative to kanban, *The International Journal of Production Research* **28**(5): 879–894.

Sturm, R., Frauenhoffer, F., Dorner, J., Kirschenhofer, O. and Reisinger, T. (1999). Advanced wip control for make-to-order wafer fabrication, *10th Annual IEEE/SEMI. Advanced Semiconductor Manufacturing Conference and Workshop. ASMC 99 Proceedings (Cat. No. 99CH36295)*, IEEE, pp. 31–36.

Swe, A. N., Gupta, A. K., Sivakumar, A. I. and Lendermann, P. (2006). Cycle time reduction at cluster tool in semiconductor wafer fabrication, *2006 8th Electronics Packaging Technology Conference*, IEEE, pp. 671–677.

Tai, Y., Pearn, W. and Lee, J. (2012). Cycle time estimation for semiconductor final testing processes with weibull-distributed waiting time, *International Journal of Production Research* **50**(2): 581–592.

Tamssaouet, K. (2019). *Ordonnancement multi-objectif d'ateliers complexes de type job-shop : application à la fabrication de semiconducteurs*, Theses, Université de Lyon.
URL: *https://tel.archives-ouvertes.fr/tel-02884923*

Tirkel, I. (2011). Cycle time prediction in wafer fabrication line by applying data mining methods, *2011 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, IEEE, pp. 1–5.

Tirkel, I., Reshef, N. and Rabinowitz, G. (2009). In-line inspection impact on cycle time and yield, *IEEE Transactions on Semiconductor Manufacturing* **22**(4): 491–498.

T'kindt, V. and Billaut, J.-C. (2006). *Multicriteria scheduling: theory, models and algorithms*, Springer Science & Business Media.

Tsakalis, K. S., Flores-Godoy, J.-J. and Rodriguez, A. A. (1997). Hierarchical modeling and control for re-entrant semiconductor fabrication lines: a mini-fab benchmark, *Emerging Technologies and Factory Automation Proceedings, 1997. ETFA'97., 1997 6th International Conference on*, IEEE, pp. 508–513.

van der Eerden, J., Walbrick, W., Niesing, H., Saenger, T. and Schuurhuis, R. (2006). Litho area cycle time reduction in an advanced 300mm semiconductor manufacturing line, *The 17th Annual SEMI/IEEE ASMC 2006 Conference*, IEEE, pp. 114–119.

Varadarajan, A. and Sarin, S. C. (2006). A survey of dispatching rules for operational control in wafer fabrication, *IFAC Proceedings Volumes* **39**(3): 715–726.

Vargas-Villamil, F. D., Rivera, D. E. and Kempf, K. G. (2003). A hierarchical approach to production control of reentrant semiconductor manufacturing lines, *IEEE Transactions on control systems technology* **11**(4): 578–587.

Vialletelle, P. and France, G. (2006). An overview of an original wip management framework at a high volume/high mix facility, *IFAC Proceedings Volumes* **39**(3): 89–92.

Wang, C.-N. and Wang, C.-H. (2007). A simulated model for cycle time reduction by acquiring optimal lot size in semiconductor manufacturing, *The International Journal of Advanced Manufacturing Technology* **34**(9-10): 1008–1015.

Wang, J. and Zhang, J. (2016). Big data analytics for forecasting cycle time in semiconductor wafer fabrication system, *International Journal of Production Research* **54**(23): 7231–7244.

Wang, Y.-C., Chen, T.-C. T. and Wang, L.-C. (2017). Simulating a semiconductor packaging facility: Sustainable strategies and short-time evidences, *Procedia Manufacturing* **11**: 787–795.

Wang, Z., Wu, Q. and Qiao, F. (2007). A lot dispatching strategy integrating wip management and wafer start control, *IEEE Transactions on Automation Science and Engineering* **4**(4): 579–583.

Wein, L. M. (1992). On the relationship between yield and cycle time in semiconductor wafer fabrication, *IEEE transactions on semiconductor manufacturing* **5**(2): 156–158.

Werner, S., Horn, S., Weigert, G. and Jahnig, T. (2006). Simulation based scheduling system in a semiconductor backend facility, *Simulation Conference, 2006. WSC 06. Proceedings of the Winter*, IEEE, pp. 1741–1748.

Wu, G.-L., Wei, K., Tsai, C.-Y., Chang, S.-C., Wang, N.-J., Tsai, R.-L. and Liu, H.-P. (1998). Tss: a daily production target setting system for fabs, *1998 Semiconductor Manufacturing Technology Workshop (Cat. No. 98EX133)*, IEEE, pp. 86–98.

Yoon, H. J. and Lee, D. Y. (2000). A control method to reduce the standard deviation of flow time in wafer fabrication, *IEEE transactions on Semiconductor Manufacturing* **13**(3): 389–392.

Yugma, C., Dauzère-Pérès, S., Artigues, C., Derreumaux, A. and Sibille, O. (2012). A batching and scheduling algorithm for the diffusion area in semiconductor manufacturing, *International Journal of Production Research* **50**(8): 2118–2132.

Zadeh, L. (1963). Optimality and non-scalar-valued performance criteria, *IEEE transactions on Automatic Control* **8**(1): 59–60.

Zarifoglu, E., Hasenbein, J. J. and Kutanoglu, E. (2012). Lot size management in the semiconductor industry: Queueing analysis for cycle time optimization, *IEEE Transactions on Semiconductor Manufacturing* **26**(1): 92–99.

Zhou, Z. and Rose, O. (2010). A pull/push concept for toolgroup workload balance in wafer fab, *in* B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan and E. Yucesan (eds), *Proceedings of the 2010 Winter Simulation Conference*, Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, pp. 2516–2522.

Zhou, Z. and Rose, O. (2011). A composite rule combining due date control and wip balance in a wafer fab, *in* S. Jain, R. Creasey, J. Himmelspach, K. White and M. Fu (eds), *Proceedings of the 20011 Winter Simulation Conference*, Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, pp. 2085–2092.

Zhou, Z. and Rose, O. (2012). Wip control and calibration in a wafer fab, *in* C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose and A. M. Uhrmacher (eds), *Proceedings of the 2012 Winter Simulation Conference*, Institute of Electrical and Electronics Engineers, Inc., Piscataway, New Jersey, pp. 2007–2018.

Zhou, Z. and Rose, O. (2019). A global wip oriented dispatching scheme: Work-center workload balance without relying on target wip, *2019 Winter Simulation Conference (WSC)*, IEEE, pp. 2212–2223.

École Nationale Supérieure des Mines de Saint-Étienne

NNT : **2020LYSEM020**

Félicien BARHEBWA - MUSHAMUKA

Novel Optimization Approaches for Global fab Scheduling in Semiconductor Manufacturing

Speciality: Industrial Engineering

Keywords: Global scheduling, linear programming, Work-In-Process control, cycle time control, semiconductor manufacturing

Abstract:

In semiconductor manufacturing, microelectronic components require several hundred operations on several hundred machines grouped into different work centers. Each work center specializes in handling one type of operation, and these are usually very different from one work center to another. These characteristics, along with reentrant flows and long cycle times (6 to 8 weeks) greatly complicate scheduling decisions. It is therefore very difficult to manage in detail all of the scheduling decisions in all the work centers of a factory. Thus, this thesis proposes a global scheduling approach based on a structure in two levels of the operational level (global level and local level). This approach aims at steering scheduling decisions at the work center level using production targets. These production targets are expressed as the quantities of components to be achieved for each operation and for each period over a scheduling horizon. Different mathematical models called global scheduling models (linear programs) proposed in the thesis determine these quantities. These global scheduling models correspond to different global scheduling strategies of the factory such as minimizing variability, controlling cycle times, etc. The local scheduling level aims to achieve the objectives set by the global scheduling models, while optimizing its own criteria and respecting its constraints. The approach is validated by experiments based on a simulation model and industrial data.

École Nationale Supérieure des Mines de Saint-Étienne

NNT : **2020LYSEM020**

Félicien BRAHEBWA - MUSHAMUKA

Nouvelles Approches d'Optimisation pour l'Ordonnancement Global en Fabrication de Semi-conducteurs

Spécialité: Génie industriel

Mots clefs : Ordonnancement global, optimisation linéaire, control des encours (WIP), control du temps de cycle, fabrication de semi-conducteurs

Résumé :

Dans les usines de fabrication de semi-conducteurs, les composants micro-électroniques nécessitent plusieurs centaines d'opérations sur plusieurs centaines des machines regroupées en différents ateliers. Chaque atelier est spécialisé dans le traitement d'un type d'opérations, et ces dernières sont en général très différentes d'un atelier à l'autre. Ces caractéristiques, ainsi que les flux réentrants et les temps de cycle longs (de 6 à 8 semaines) complexifient grandement les décisions d'ordonnancement. Il est par conséquent très difficile de gérer de manière détaillée l'ensemble des décisions d'ordonnancement dans tous les ateliers d'une usine. Ainsi, cette thèse propose une approche d'ordonnancement global reposant sur une structure en deux niveaux du niveau opérationnel (niveau global et niveau local). Cette approche permet de piloter les décisions d'ordonnancement au niveau des ateliers en utilisant des objectifs de production. Ces objectifs de production sont exprimés comme des quantités de composants à réaliser à chaque opération et à chaque période sur un horizon d'ordonnancement. Ces quantités sont déterminées par différents modèles mathématiques d'ordonnancement global (programmes linéaires) proposés dans la thèse, qui correspondent à différentes stratégies d'ordonnancement global de l'usine comme la minimisation de la variabilité, le contrôle des temps de cycle, etc. Le niveau d'ordonnancement local vise à atteindre les objectifs fixés par l'ordonnancement global, tout en optimisant ses propres critères et en respectant ses contraintes. L'approche est validée par des expérimentations reposant sur un modèle de simulation et des données industrielles.