Operational and tactical management of qualifications for flexibility optimization of complex manufacturing systems

Antoine Perraudat

▶ **To cite this version:**

Antoine Perraudat. Operational and tactical management of qualifications for flexibility optimization of complex manufacturing systems. Other. Université de Lyon, 2021. English. NNT : 2021LYSEM005 . tel-03470583

HAL Id: tel-03470583
https://theses.hal.science/tel-03470583

Submitted on 8 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N°d'ordre NNT : 2021LYSEM005

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 26/01/2021, par :
Antoine PERRAUDAT

Gestion tactique et opérationnelle des qualifications pour l'optimisation de la flexibilité de lignes de fabrication complexes

Devant le jury composé de :

Roussy, Agnès Professeure à Mines Saint-Etienne	Présidente
Sauer, Nathalie Professeure à Université de Lorraine	Rapporteure
Mönch, Lars Professeur à FernUniversität in Hagen	Rapporteur
Ljubic, Ivana Professeure à ESSEC Business School	Examinatrice
Borodin, Valeria Maître de conférences à Mines Saint-Etienne	Examinatrice
Tollenaere, Michel Professeur à Grenoble INP	Examinateur
Dauzère-Pérès, Stéphane Professeure à Mines Saint-Etienne	Directeur de thèse
Vialletelle, Philippe Ingénieur à STMicroelectronics	Encadrant industriel
Bezal, Olivier, Responsable Performances Industrielles à STMicroelectronics	Invité

N°d'ordre NNT : 2021LYSEM005

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 26/01/2021, par :
Antoine PERRAUDAT

Gestion tactique et opérationnelle des qualifications pour l'optimisation de la flexibilité de lignes de fabrication complexes

Operational and tactical management of qualifications for flexibility optimization of complex manufacturing systems

---

Devant le jury composé de :

Roussy, Agnès Professeure à Mines Saint-Etienne	Présidente
Sauer, Nathalie Professeure à Université de Lorraine	Rapporteure
Mönch, Lars Professeur à FernUniversität in Hagen	Rapporteur
Ljubic, Ivana Professeure à ESSEC Business School	Examinatrice
Borodin, Valeria Maître de conférences à Mines Saint-Etienne	Examinatrice
Tollenaere, Michel Professeur à Grenoble INP	Examinateur
Dauzère-Pérès, Stéphane Professeure à Mines Saint-Etienne	Directeur de thèse
Vialletelle, Philippe Ingénieur à STMicroelectronics	Encadrant industriel
Bezal, Olivier, Responsable Performances Industrielles à STMicroelectronics	Invité

Remerciements

Je souhaite tout d'abord remercier Stéphane et Philippe qui m'ont permis de réaliser ces travaux. Vos commentaires, votre aide, votre expérience, vos contacts au sein de l'entreprise ont grandement participé à la réussite de cette thèse.

Je souhaite également remercier les membres du jury pour avoir accepté d'évaluer mes travaux. Vos questions m'ont également permis de compléter mon manuscrit.

Je souhaite également remercier l'ensemble des membres du laboratoire SFL pour leur accueil et leurs conseils lors des séminaires et réunions. Je souhaite tout particulièrement remercier Pierre Uny et Claude Yugma qui ont pris beaucoup sur leur temps pour venir nous chercher à la gare ou à Rousset lors de nos déplacements.

Je souhaite également remercier Renaud, Vincent et Quentin du service Production & Methods à ST Crolles. Votre aide a été grandement précieuse, notamment pour pouvoir récupérer les données nécessaires aux expérimentations numériques. Je te tiens également à vous remercier de m'avoir sollicité sur d'autres sujets afin que je puisse m'ouvrir à d'autres problématiques liées à l'industrie.

Je tiens à remercier Philippe Grail et Olivier Bezal qui ont pris du temps pour évaluer l'application de mes travaux à travers FlexQual. Je sais que cela vous a pris du temps, et je vous en remercie.

Et enfin, je tiens à remercier l'ensemble des membres de ma famille pour leur soutien pendant toutes mes années d'étude et ces trois années de thèse.

Contents

List of Figures

List of Tables

List of Algorithms

General introduction

The thesis is performed in cooperation with STMicroelectronics Crolles, France through CIFRE cooperation programs and deals with industrial qualification management problems, which are related to process flexibility problems, in the semiconductor manufacturing industry. This thesis follows two previous thesis (Johnzén, 2009; Rowshannahad, 2015).

For semiconductor manufacturing, a qualification is an eligibility or certification for a machine to process one operation of a specific product. A machine cannot process a product without the associated qualification. Qualifications are mandatory to ensure high yield of production lines and products of quality. Qualifications are used to improve the flexibility (ability to respond effectively to changing circumstances) and to configure production capacities of work centers in semiconductor factories as they allow the production volume associated to operations to be processed on different machines. Qualifications are dynamic, *i.e.* time-varying, and new qualifications are frequently planned and developed because of demand changes on the semiconductor market. The larger the number of qualifications in a factory, the better the manufacturing performances, *e.g.* in terms of workload balance, throughput, fabrication time and demand satisfaction. However, developing new qualifications takes times, up to several months, and can be expensive. New qualifications must be anticipated. In addition, qualifications can be lost over time (disqualifications) because of fabrication process problems, then, a machine can lose its qualification for a product part. Then, re-qualifications must be performed to change fabrication process parameters in order to ensure that the machine is again able to process the product part while meeting yield and quality requirements. Re-qualifications can also be expensive.

Because qualifications and re-qualifications are expensive, relevant qualification and re-qualification decisions must be made to optimize the workload balance, throughput, fabrication time and demand satisfaction at the lowest cost. More precisely, we answer the following questions: Given a horizon, a demand forecast by product part, process times, qualification delays and costs and production capacities of machines, how to determine most relevant new qualifications? From disqualifications, how to determine most relevant re-qualifications to improve operational efficiency? Or equivalently, what are the most relevant qualifications and re-qualifications?

Answering these questions is complex. This is because evaluating the quality of qualifications and re-qualifications is complex as machines can share common qualifications and have finite production capacities. Typically, it is difficult to evaluate the utilization rate of a machine after multiple qualification decisions as the workload of a machine can be shared with similarly qualified machines. Evaluating the utilization rates of machines is primordial as they are intrinsically to throughput, fabrication times and demand satisfaction.

Because qualification management is complex, we answer these questions from two standpoints by proposing relevant new optimization models and solution approaches. The first standpoint is an operational standpoint and seeks to answer

the question for re-qualifications. The second standpoint is a tactical standpoint and seeks to answer the question for qualifications. Although both standpoints are closely related, the question is not exactly answered the same way because of the nature of possible qualifications given a horizon.

First, in Chapter 1, an overview of the industrial and scientific context is provided. Following the description of the semiconductor industry and the semiconductor manufacturing process, qualification management, its criticality for manufacturing performances and complexity are detailed. A literature review of process flexibility and qualification management is then provided. Finally, contributions on operational and tactical qualification management are highlighted.

Our contributions on operational qualification management can be found in Chapters 2, 3, 4, 5, and 7. The question "How to determine the most relevant re-qualifications to improve operational efficiency?" is answered from different standpoints. In Chapter 2, re-qualifications are determined to maximize the utilization balance and minimize the total utilization rate of the machines. The problem is equivalent to maximize flexibility measures proposed in the literature. Six solution approaches are compared on industrial data from STMicroelectronics. Solution approaches guided by the dual variables of qualification constraints are shown to outperform other solution approaches when a small computational budget (a few minutes) is given. In Chapter 3, re-qualifications are determined in terms of throughput. Due to some characteristics of work centers in semiconductor production facilities, such as production variability (no constant pattern over time of product quantities and production capacities) and highly automated dispatching decisions at the shop floor, maximizing the utilization balance of the machines might not be always equivalent to maximizing the throughput, in particular for short term horizons (*e.g.* a few hours). To better capture these characteristics, bilevel optimization models are proposed. Bilevel optimization models are shown to be better suited to maximize the throughput on short term horizons. To further better capture production variability, dynamic bilevel optimization models are motivated and proposed in Chapter 4. They are compared to static bilevel optimization models and shown to better capture dynamic product quantities and production capacities on some industrial instances. In Chapter 5, the impact of re-qualifications on cycle times is evaluated. Simple closed-form solutions are shown to exist to describe the mean cycle times for short term horizons. Closed-form solutions are validated by using industrial historical data. Then, we show on industrial data that re-qualifications that maximize the throughput are not necessarily the same that minimize the mean cycle time, mostly due to production variability. Pratical uses of optimization models are discussed at the end of each chapter. Finally, in Chapter 7, we show how optimization models and solution approaches were embedded in a fully functional decision support system used at STMicroelectronics at Crolles, France.

Our contributions on tactical qualification management can be found in Chapter 6. In Chapter 6, the question "How to determine the most relevant new qualifications to satisfy the demand and cover the demand uncertainty while minimizing qualification costs?" is answered with new optimization models. First, a deterministic modeling is proposed to satisfy the nominal demand. Then, because products are most often not equivalent in terms of workload for work centers due to different

processing times and operations, a modeling of the demand, based on robust optimization and product cannibalization, and an optimization model are proposed. Finally, a binary search is proposed to characterize the robustness of a work center against demand uncertainty. We show that (1) A limited number of well-chosen qualifications are required to achieve the same robustness than the one obtained by performing all qualifications, and (2) Implementing the qualifications determined by only considering the nominal demand can lead to capacity constraint violations. We also discuss practical uses of proposed optimization models.

Chapter 1

Industrial and scientific contexts

This chapter provides an overview of the industrial and scientific contexts related to semiconductor manufacturing and qualification management. In particular, the semiconductor manufacturing context is presented, and why qualification management is a critical component of manufacturing operations in the semiconductor industry is discussed. In Section 1.1, the semiconductor manufacturing industry is described from a general standpoint, followed in Section 1.2, by the manufacturing process of products in front end factories. In Section 1.3, challenges and problems related to qualification management are presented. In Section 1.4, an up-to-date literature review on qualification management in the semiconductor industry is proposed. Finally, in Section 1.5, our industrial and scientific contributions are outlined.

1.1 Semiconductor industry

The semiconductor industry deals with the design and fabrication of Integrated Circuits (ICs). The Moore's "law" initially drove the semiconductor industry. Gordon Moore stated in 1965 that the transistor count in an IC would double every 2 years due to shrinking transistor dimensions and other improvements. Moore's law was then a goal for the semiconductor industry. Today, the semiconductor industry is diversified as shrinking transistor dimensions is expensive and is not sustainable for most semiconductor companies (McKinsey & Company, 2011), in particular for medium sized to small sized semiconductor companies. Consequently, for a large number of semiconductor companies, competition is no longer necessarily about minimizing transistor dimensions. Competition is about production costs, selling prices, energy efficiency, customized products and most of all differentiation. The semiconductor industry market represented in 2016 a market of more than $368 billion, and more than $489 billion in 2018 (Deloitte, 2019). The semiconductor market is also characterized by a cyclical market, *i.e.* the market size decreases for a few years before growing up again (McKinsey & Company, 2011).

The semiconductor industry is composed of three different types of companies or business model: IC manufacturer, fabless companies and pure play foundry companies. Fabless companies design and sell their ICs but outsource their production to pure play foundries. IC manufacturers both produce and design their ICs, and can further be distinguished into two groups: Low Mix (LM) and High Mix (HM) manufacturers. LM manufacturers propose a restricted set of products (different types of ICs) to potential clients. They are able to produce each product in large quantities. HM manufacturers propose a large set of products but each product is made in smaller quantities compared to their LM counterparts because more products share the same production capacity. STMicroelectronics is a HM manufacturer, where each factory can produce a portfolio of more than several thousand different products. About several hundred products are being made at any time in each factory. HM manufacturers are characterized by time-varying demand of products, short life times of products, a wide range of products with customization. Figure 1.1 illustrates the historical demand for a given product at STMicroelectronics. The demand was highly variable because a large ramp-up for the product was expected only within a few months. The product was only made for a few months, and then the demand disappeared.

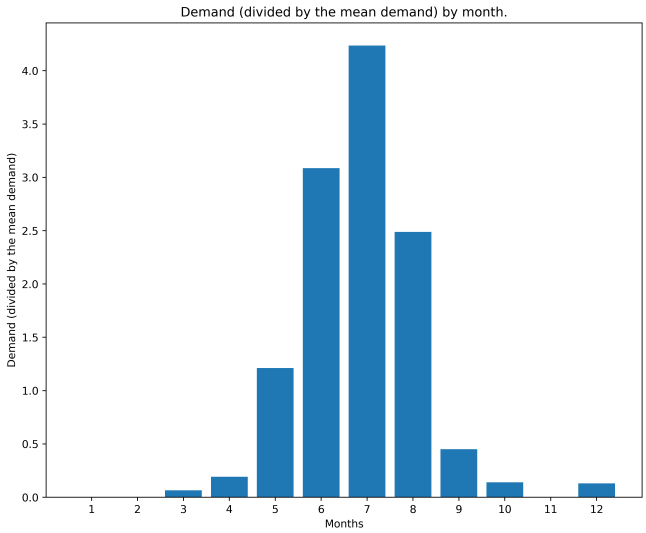


Figure 1.1: An illustration of the time-varying demand for a product.

The manufacturing process of ICs is performed from raw silicon wafers (see Figure 1.2) and is divided into two main fabrication steps: Front end fabrication and back end fabrication. Front end fabrication is associated to wafer fabrication, *i.e.* the fabrication of ICs on wafers, and wafer probing, *i.e.* electrical and functional testing of fabricated ICs to identify potential undetected defective ICs. Back end fabrication is associated to assembly, packaging and final testing of ICs before sending them to final customers. In the remainder of this section, we focus on front end fabrication as the thesis was performed in a front end wafer fab. Wafer fabs are often distinguished by the wafer size that they can produce. In general, wafer sizes used in the semiconductor industry varies between 100 mm and 300 mm. A wafer contains hundreds to thousands of ICs depending on their respective size.

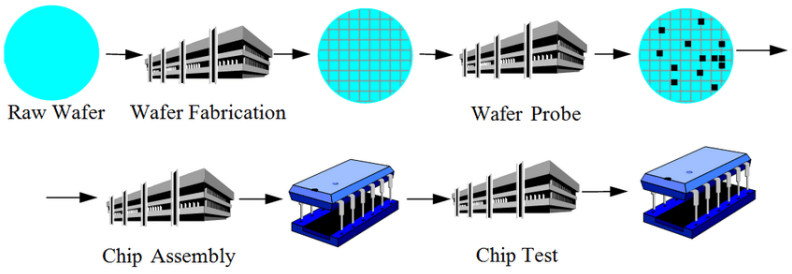


Figure 1.2: Manufacturing process of ICs (Schömig and Fowler, 2000).

Common applications for ICs are: Telecommunications, Internet Of Things (IOT), robotics, home automation, wireless payment, radio frequency identification, smart phones, computers, automotive industry, spatial industry, digital image processing, digital voice recorder, etc. ICs can even be found in shoes and clothing.

ICs are very often made of millions (billions for the largest ICs) of transistors and other electric components such as diodes, capacitors.

Manufacturing ICs is actually expensive. A wafer fab typically costs several billion dollars. More than 80% of the total cost corresponds to machine purchasing cost (Quirk and Serda, 2001; Delp et al., 2006; Mönch et al., 2012; McKinsey & Company, 2011). A single machine usually costs more than a few millions dollars, and wafer fabs typically consist of several hundreds of machines. For instance, most recent lithography machines (see Figure 1.3) can cost up to $100 million. Developing new technologies is also extremely expensive. McKinsey & Company (2011) reports that developing a 45 nm logic process technology costs $600 million. Operating costs are also expensive, notably because (i) machine parts are frequently changed during maintenance operations, (ii) machines can be energy intensive and (iii) because operating a wafer fab implies controlling the air quality, reducing pressure with respect to outside pressure, controlling hydrometry, controlling temperature to avoid introducing, generating or keeping particles, which may ultimately lead to contamination and thus create defective ICs. Reducing production costs is therefore particularly critical due to high capital cost of wafer fabs, and can be done in three main ways (Hahn, 2001):

1. Improving the yield, *i.e.* the ratio between the number of working ICs at the end of the manufacturing process and the total number of ICs produced,
2. Improving the machine utilization,
3. Increasing the wafer size.

Among these three options, improving the yield is not expected to provide major cost reductions as the yield is already large, above 90 to 95% for all mature products. Using a wafer size of 450mm is prohibitive as machines for 450mm wafers are not existing today. The development of 450mm technologies has been abandoned due to their prohibitive costs. Improving machine utilization is therefore the major source for reducing production costs in existing wafer fabs with fixed wafer sizes. For a given product flow in a wafer fab, improving machine utilization should be understood as reducing idle times of machines, and therefore as increasing the throughput and reducing the fabrication time of ICs. Improving machine utilization without improving the overall throughput or reducing the fabrication time does not reduce production costs (Atherton and Atherton, 1995). Increasing the throughput leads to more sales, thus larger revenues. Reducing fabrication times is associated to shorter development cycle, smaller inventory holding costs and greater market responsiveness (Atherton and Atherton, 1995). Reducing fabrication times is also linked to yield reduction. Reducing fabrication times reduces the number of out of specification wafers as wafer probing is performed sooner and reduces the deposit of undesired particles on wafers (Atherton and Atherton, 1995; Leachman and Ding, 2010). Operational efficiency is expected to be a major source of production cost reduction. It is important to mention that the throughput and the cycle time are two intrinsically related metrics. For a given installed capacity in a factory, increasing

the throughput, or similarly the number of started products, also increases fabrication times, which should be minimized in an ideal situation (Hopp and Spearman, 2011).

The work of this thesis is closely related to improving the utilization of machines with relevant qualifications. A qualification is an eligibility or certification for a machine to process one operation of a specific product. A machine cannot process a product without the associated qualification. Qualifications are performed to ensure high yield and products of quality. Qualifications are used to improve the flexibility (ability to respond effectively to changing circumstances) and to configure production capacities of work centers in semiconductor factories. Therefore, they are critical parameters that can be used to better use available machine capacities, reduce idle times of machines, thus reducing fabrication times, improving the throughput and better satisfying the customer demands.

Figure 1.3: An illustration of a lithography machine (ASML, 2011).

1.2 Front end fabrication process

1.2.1 Description of front end fabrication

Front end fabrication arguably includes the most complex processes in semiconductor manufacturing (Mason et al., 2002; Ovacik and Uzsoy, 2012; Mönch et al., 2012). An entity is described as complex if it consists of many different and connected parts.

In front end wafer fabs, wafers are grouped in lots which are handled in a pod, and more precisely with a Front Opening Unified Pod (FOUP) in 300 mm wafer fabs. FOUPs are designed to safely transport lots, from one machine to another, from one machine to a storage location, from a storage location to another storage location, or from a storage location to a machine. An illustration of a FOUP is given in Figure 1.4. In general, 25 wafers constitutes a lot in 300 mm wafer fabs, and each lot corresponds to a specific product.

Figure 1.4: Illustration of a Front Opening Unified Pod (FOUP) (Silicon Connection, 2020).

The fabrication of a wafer is associated to a product fabrication route, which describes the sequence in which operations, *i.e.* elementary fabrication steps, must be performed to fully fabricate ICs associated to the product. Fabrication routes are different from one product to another. The fabrication of a lot includes more than 500 operations, and products can share common operations. Wafers are fabricated layer by layer and contain hundreds to thousands of ICs, depending on the size of the product and the wafer size. The mean fabrication time (cycle time) of a lot in a wafer fab is typically of two to three months. More than 80% of the cycle time of a lot is made of non added value activities, *e.g.* inspection, transportation, and waiting times. This is mostly due to the fact that the utilization of machines in wafer fabs is maximized to reduce production costs, therefore there are a large number of wafers in a wafer fab, thus leading to congestion problems and waiting times.

Each operation on the fabrication route is performed by a machine in a work center. More precisely, at each operation, the machine applies a *recipe* that corresponds to the desired fabrication process. The recipe is a program that defines actions that must be performed by the machine for the fabrication process at the operation. For instance, the recipe defines the pressure, the temperature conditions and the chemicals that must be used. For ion implantation, a recipe determines the dopant (the ion type that will be implanted in the wafer), the energy at which the machine must operate, and the dose. Recipes can be very different from one product to another, leading to different (fabrication) process times. Several main work centers (or operation types) can be distinguished (Atherton and Atherton, 1995; Hutcheson, 2000; Mönch et al., 2012):

Diffusion, oxidation and layer/film deposition. Oxide layer are grown from silicon with oxygen. This operation is used to grow oxide layer of transistor gates. It is done at a very high temperature, $> 1000°C$. During metallization operations, conductive layers are grown on the surface of the wafer. Conductive layers are used to access components, connect transistors and are used as an interface between packaging and the IC. Layers of electrical insulators are also grown on the surface to create capacitors in memory ICs, separate conductive layers, isolate transistors, and protect ICs from external contamination. These operations are often performed with chemical vapor deposition processes.

Photolithography. A photolithography operation delimits where dry etch and ion implantation operations should be performed. Regions are delimited by using a reticle (or mask) and a photosensitive resin (or photoresist) that transfers a pattern

onto the wafer. A photolithography operation takes place in three stages: (1) The photosensitive resin is first spread on the wafer, (2) It is then exposed to ultraviolet light where it is not protected by the reticle, and (3) Exposed regions of the photosensitive resin are removed from the surface of the wafer. The next operations take place in exposed regions. Photolithography machines are expensive and photolithography work centers often constitutes a bottleneck/critical (limiting the overall manufacturing performances) work center for wafer fabs.

Etch. A dry etch operation takes place after a photolithography operation. It is used to remove matter from the wafer in exposed photosensitive resin regions. For instance, it removes some regions of previously grown layers. Dry etch operations can also use polymers to better respect delimited regions by the exposed photosensitive resin. Polymers are then removed during cleaning operations. Etch operations are described as dry when a plasma is used instead of chemical solutions, etch operations are then described as wet. Wet etch operations are uniform on the surface of the wafer, and are used to remove complete layers, *e.g.* for rework purposes, and recycle test wafers.

Ion implantation. Ion implantation operations consists in doping the wafer with ions (*e.g.* phosphorus, boron, arsenic and indium) where dry etch operations have been performed. Ion implantation gives the ability for a transistor to conduct electric current when a voltage is applied on its gate. Ion implantation can only be performed in exposed regions of the photosensitive resins. Annealing can be used after an ion implantation operation to correct defects from ion implantation and activate dopants. Annealing is performed at a very high temperature ($> 1000°C$) to avoid diffusing dopants in the silicon.

Planarization. Planarization (or chemical mechanical planarization) is used to have a planar layer. Ultimately, a planarization operation reduces the size of layer deposit. Planarization reduces lens focusing problems in photolithography.

Cleaning. Cleaning operations are also frequent to remove contaminants. Cleaning operations are often performed after photolithography, dry etching, ion implantation operations for particle, metallic or organic decontamination. For instance, polymers introduced by dry etching are removed with cleaning.

Inspection. Lots are frequently inspected to control the quality of the wafers, maximize the yield and therefore avoid any defect to reach the final client. Inspected lots are sampled, *i.e.* all lots are not inspected, because inspection capacity is limited as inspection machines are also expensive. In addition, in general, all wafers of a lot are not inspected. Inspection operations are performed in metrology work centers, which control physical parameters on wafers, *e.g.* transistor gate sizes, thickness of films, or defectivity work centers, which control contamination, scratches, defect patterns.

Wafers are fabricated layer by layer. Because machines are expensive, the number of operations to perform to fabricate a wafer is much greater than the number of work centers in wafer fabs. Wafer fabs are therefore characterized by *re-entrant product flows* in work centers. More precisely, a wafer visits multiple times the same work center for different operations on its fabrication route. For instance, for some products, the lithography work center can process the same lot up to 80 times. Work centers are therefore typically organized in a job-shop manner.

Re-entrant product flows make every local decision at a work center, *e.g.* dispatching, maintenance operations and qualification decisions, connected to all other work centers. For instance, if a machine of a critical work center fails, then other work centers can starve as lots are blocked at the critical work center. Re-entrant product flows are arguably one the major factors that makes front end fabrication complex (Mason et al., 2002; Ovacik and Uzsoy, 2012; Mönch et al., 2012). Not only interconnected work centers make front end fabrication complex to manage, but they also create a high production variability, *i.e.* a lack of constant pattern in the demand and in production capacities, which contributes to the increase of the mean cycle time (Hopp and Spearman, 2011) and therefore inventory costs.

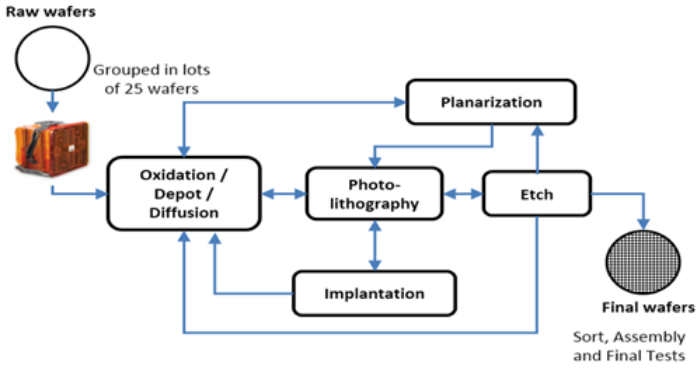


Figure 1.5: An illustration of the wafer fabrication process (Mönch et al., 2011) in a wafer fab.

1.2.2 Qualifications

Equipment qualifications. Once the machine is purchased and being installed in the wafer fab, the machine must undergo a first *qualification* procedure, which qualifies the production environment around and in the machine. More precisely, wafer fabrication is performed in a *clean room*, which is controlled in terms of particle, air flow and pressure to minimize the number of particles that can enter in contact with wafers, and therefore potentially lead to the fabrication of defective ICs. More precisely, norm ISO 14644-1 must be respected. Similarly, particle emission from the machine to the factory is measured. If particle emissions are too large, particle emission sources are investigated. For instance, friction sources are investigated. If necessary, some elements in the machine must be moved or changed. The qualification procedure also ensures that alarms work properly. Different physical parameters can also be measured. For instance, electrostatic fields nearby wafers, glasses, load ports are also measured near ion implantation machines. A qualification procedure requires rigor as every measurement must be properly described and traceable. Each qualification procedure is machine dependent because machines can come from different suppliers and be of different generations.

Recipe-to-machine qualifications. Once their production environment is qualified,

machines must also be qualified for recipes that they will run. These qualifications are *recipe-to-machine* qualifications. Qualifying machines for recipes certifies that machines are able to meet quality and yield requirements when recipes are run. A qualification therefore concerns a couple (recipe, machine). The ultimate goal of qualification procedures is to reach zero defects, avoid infant mortality of wafers and eliminate out of specification products. There exist different qualification levels, which can be distinguished by their delay and their cost.

For new machines, recipes, or products, the qualification level is arguably the most "difficult" one. The qualification procedure for new machines, recipes or products is expensive, energy-consuming, consumes production capacity and takes about two to three months as test lots must complete their production. As previously mentioned in Section 1.1, developing new technologies is also extremely expensive. McKinsey & Company (2011) reports that developing a 45nm logic process technology costs $600 million. These costs are partly due to qualification procedures. During the qualification procedure for a new machine, test runs of recipes are conducted and the machine has an "engineering" status that prevents it from processing standard production lots. For machines that are already installed in the wafer fab, new qualification procedures typically involve the development of new recipes for new products or new technologies. First, the recipe is qualified from a *defectivity* standpoint. It is verified that there is no undesired deposition of particles, corrosion, holes and scratches, which can lead to short-circuits. It is also verified that there are no patterns in defects, *i.e* defects must be uniform on the wafer and not in a single specific region of the wafer. It is also verified that the recipe is in line with specifications, *e.g.* in terms of implanted dose or layer thickness, by inspecting the wafer and measuring physical parameters. If problems are detected, then parameters of the recipes are changed. For instance, different chemicals can be tried, spin speeds of robots can be changed, a larger temperature can be tried, different temperature ramp up and ramp down profiles can be tried, voltage can be changed. Small changes in the parameters of the recipe can increase the yield by 2 to 3%.

Once the qualification procedure is over and fruitful (note that the qualification procedure may *never* be fruitful), it is certified that the machine respects yield and quality requirements and can apply the recipe on wafers. It is then said that the machine is *qualified* for the recipe. Due to technological restrictions on machines, machines are only *qualifiable* for a subset of recipes. Initial technological restrictions for existing machines can be waived with retrofitting, but it induces an additional cost. In addition, due to strong process constraints, several families of recipes may be incompatible and therefore are never qualified at the same time on a given machine. Therefore, some machines can be dedicated to very specific recipes whereas other machines will be qualified for a wide range of recipes. In practice, although a machine could be qualified for all recipes, the number of qualifications of a machine is often limited because qualification costs can be expensive. Maintaining many qualifications on a machine is difficult because of limited human and inspection resources, and because all possible qualifications are probably not necessary to optimize operational efficiency or ensure the satisfaction of the demand. There is a trade-off between the number of qualifications to perform, which must be minimized because qualifications can be expensive, and manufacturing performances,

which must be maximized. Note that, as products can share common operations, then they can also share common recipes. Therefore, once a recipe is qualified on a machine, then the machine can run all the products that require the recipe.

Other qualification levels correspond to *re-qualification* procedures, which are often less 'difficult" than for qualifications of new recipes, machines or products. A machine that is qualified for a recipe does not remain qualified throughout its operation in the wafer fab. Qualifications are time-varying. A machine no longer qualified for a recipe is said to be *disqualified* for the recipe. Depending on the disqualification reason, re-qualification procedures can be immediately done by computer orders, maintenance operations, recipe parameter adjustments or may require the completion of test lots if a detailed analysis of wafers is required. Disqualification reasons condition re-qualification delays and costs, and therefore condition the horizon size on which re-qualification decisions can be considered. In a work center, disqualifications can be frequent and have serious consequences on wafer fab performances, if they are not managed properly or anticipated. To maximize manufacturing performances, in particular in terms of throughput and cycle time, an efficient and effective design and *follow-up* of the qualification configuration of each work center is required (see *e.g.*, Johnzén et al. 2007, 2011; Kabak et al. 2013; Rowshannahad et al. 2015; Kopp et al. 2018). Re-qualifications are therefore also critical.

A *disqualification* can occur for different reasons. A first reason lies in Work-In-Process (WIP) management policies. As two recipes on the same machine can be incompatible, the machine is alternately disqualified for a recipe and qualified for another. Recipes can also be disqualified to orient the WIP on certain machines, which are known to be much faster or give slightly better results in terms of yield. Some recipes are disqualified because they quickly degrade machines, which lead to more maintenance operations. Therefore, in general at any time, only one machine is kept qualified to deal with such cases. Other machines will be re-qualified if initial machines are down, disqualified or the gain on the throughput or cycle time is interesting. These machines are often used in a "back-up" mode. Observe that WIP management policies may vary over time. Recipes can also be disqualified on machines because they have not been run for a long time (qualification time window, Obeid et al. 2014; Kopp et al. 2016). This time window varies from one work center to another. Time windows are practically defined because the quality and yield of a recipe are time-varying and depend on other recipes. Therefore, if a recipe has not been applied for a long time, it is considered as outdated. If the recipe has not been run for more than 6 months, a complete qualification procedure must once again be performed. Recipes can also be disqualified on machines because of unexpected events. For instance, a recipe can be disqualified because of yield, parametric issues or because a consumable (*e.g.*, a bottle of gas) is empty.

Disqualification rates are highly variable from one work center to another. For 50% of the total number of machines in the production facility, disqualifications represent less than 10% of initial qualifications. For 75% of machines, disqualifications represent less than 20% of initial qualifications. Note that it is not unusual to observe that disqualifications on machines in some work centers can represent up to 50% of initial qualifications, which can be due to WIP management policies

and the fact that machines have few qualified operations. Disqualifications can then represent hundreds of operations over all the machines.

Note that qualification procedures are not limited to the semiconductor industry. They may simply not be described with the same term. For instance, being a nurse or a surgeon requires a medical degree, which is a certification ensuring that medical operations will be correctly performed. Air plane pilots must be certified to fly new aircraft. This certifies that the pilot is able to operate the aircraft and know emergency procedures. Spare part suppliers must algo undergo a certification procedure. However, qualifications are very rarely as *dynamic* as in semiconductor manufacturing, in particular in high-mix wafer fabs. This is because qualifications can be frequently lost or because new qualifications are frequently developed as new products and new machines are frequently introduced in wafer fabs. In addition, qualifications are frequently updated due to the time-varying demand. The demand for some products decreases whereas the demand for other products increases. Therefore, if the demand for a product decreases over time, it is possible and even desired to perform new qualifications for other products to keep the required production capacity and satisfy the demand.

In this thesis, we are not interested in optimizing qualification procedures, for instance in terms of costs. We tackle recipe-to-machine qualification management from a capacity planning standpoint and a flexibility standpoint. In other words, we consider features associated to recipe-to-machine qualification procedures such as delays and costs but qualifications are optimized in terms of production capacity to satisfy the demands of products and improve machine utilization.

1.3 Qualification management in manufacturing operations of wafer fabs

The primary goal of a wafer fab is to fabricate wafers and deliver functional ICs to customers. Manufacturing operations correspond to the management, planning and control of production activities ensuring that each client receives its lots on time. Manufacturing operations, which constitute a part of operations management, are generally divided into three decision levels, a strategic decision level, a tactical decision level and an operational decision level. We refer the reader to Stadtler and Kilger (2002); Hopp and Spearman (2011); Mönch et al. (2018) for a description of the different decision levels.

Delivering functional ICs to *all* customers, *i.e.* satisfying the demand for all products and maximizing on-time deliveries is actually complex in HM wafer fabs.

There are several reasons explaining this complexity. Several hundred products compete for the same production machines in HM wafer fabs, which themselves operate with finite production capacities. In addition, the demand by product is time-varying and can be highly uncertain. Products have short lifetimes. There are also manufacturing risks (*e.g.* machine breakdowns, yield losses) that can prevent wafer fabs from satisfying the demand. In addition, wafer fabs are often made of complex information and manufacturing flows, in particular due to re-entrant product flows. When such conditions are met, the need for *flexibility* (the ability to

respond effectively to changing circumstances, see Sethi and Sethi 1990) is imperative (De Toni and Tonchia, 1998). Qualifications are therefore critical *catalysts* for demand satisfaction and on-time deliveries in wafer fabs because:

- Qualification management is closely related to the the notion of *production flexibility*, which is defined as all products a factory is able to produce without requiring additional major capital investment. Production flexibility is the result, among others, of *process flexibility*, which is defined as the ability of processing different products at the same time (Sethi and Sethi, 1990; Jain et al., 2013). Adding new qualifications improves the level of process flexibility of work centers and therefore improves the management of manufacturing risks such as machine downtime.

- Qualifications allow a machine to run recipes, qualifications are therefore used to configure the production capacity of a work center as the production volume associated to operations can be processed to be processed on different machines. The larger the number of qualified machines for an operation, the larger the production capacity for the operation. However, the production capacity is finite. Therefore, qualifications are critical parameters that enable wafer fabs to satisfy the demand. Planning and adding the right new qualifications are critical to anticipate ramp-up product demands.

- New qualifications are determined at a tactical decision level, they have a large influence on the overall performances of wafer fabs as they notably affect production planning and scheduling. If the right qualifications are not carefully determined to satisfy the demand and improve operational efficiency, wafer fabs cannot have high service levels and reduce production costs.

Qualification management is a discipline of manufacturing operations. It refers to the planning and control activities of qualifications. Qualification management can be found at the operational and tactical decision levels of manufacturing operations:

- **Tactical decision level.** There are existing machines in a work center. New machines are being installed. Similarly, new products are being introduced in the factory and the demand for existing products can increase. New qualifications are then necessary to increase the production capacity of new products and increase the production capacity of products already made in the factory with a ramp-up demand. More precisely, given demand forecasts for each product, the fabrication route for each product, an estimate of the available production time by machine, qualification delays, new qualifications must be *planned* to satisfy the demand by product, respect capacity constraints and minimize qualification costs as new qualifications are expensive to develop. Qualification decisions at the tactical decision level are typically made over a horizon varying between 6 and 12 months. Recall that as qualifications are determined at a tactical decision level, they have a large influence on the overall performances of wafer fabs as they notably affect production planning and

scheduling. Therefore, the right qualifications must be carefully determined to anticipate the demand for new products and the increasing demand for currently made products. Very often, only a small number of new qualifications is required among all possible new qualifications, even to cover demand uncertainty. Chapter 6 is dedicated to tactical qualification management.

– **Operational decision level.** At the operational level (production control level), initial qualifications are *already* determined. New qualifications are not performed. This is because qualification delays are much larger than the decision horizon, which is of a few hours to one to two weeks. Instead, a follow-up of qualifications is performed and re-qualifications, from disqualifications of initial qualifications, are optimized. Re-qualifications are determined to ensure that disqualifications do not prevent wafer fabs from satisfying the demand and do not prevent started lots from being shipped on time. As the decision horizon is typically smaller than the fabrication time of a wafer, re-qualifications are optimized in terms of utilization balance of the machines, throughput or cycle time to improve operational efficiency, which can also be seen as alternative operational means to ensure that lots move forward in wafer fabs, and as a result that the demand by product is satisfied on time. Typically, given WIP projections, current disqualifications and estimates of available production times by machine, re-qualifications are determined to maximize the utilization balance and minimize the total utilization rate of the machines, which in turn help to maximize the throughput and minimize the cycle time. Very often, only a small number of re-qualifications is necessary to significantly improve manufacturing performances. Chapters 2, 3, 4 and 5 are dedicated to operational qualification management.

Note that qualification management can also be found at a strategic decision level (Liao et al., 2017), but at a supply chain level and not at a work center level, when it comes to deciding the set of wafer fabs that should be able to produce a specific product or technology node. Qualification management problems at a strategic decision level are not directly addressed in this thesis. Nevertheless, the approaches proposed in Chapter 6 for a work center can be extended to be used at a supply chain decision level (see Section 1.5.3).

An effective tactical qualification management is *essential*. In addition, through multiple applications Johnzén et al. (2007), Johnzén et al. (2011), Kabak et al. (2013), Rowshannahad et al. (2015) and Kopp et al. (2018) show that to maximize manufacturing performances, in particular in terms of throughput and cycle time, an effective *follow-up* of the qualification configuration of each work center is required. If a follow-up of qualifications is not performed, *i.e.* when no re-qualifications are performed, then disqualifications that progressively occur will make operational efficiency plummet because the utilization rates of machines will be poorly balanced and there may be some operations without any qualified machine thus making impossible to satisfy the demand for associated products.

Evaluating the quality of qualification decisions is complex as machines can share common qualifications and have finite production capacities. Typically, it is difficult to evaluate the workload or the utilization rate of a machine due to highly

automated dispatching decisions of jobs and because the workload of a machine can be shared with similarly qualified machines. Evaluating the workload and the utilization rates of machines is primordial as they are intrinsically related to the throughput (see Chapters 3 and 4), the cycle time (see Chapter 5) and demand satisfaction (see Chapter 6). Moreover, process times (or equivalently throughput rates) of recipes on machines can be different from one recipe to another, and from one machine to another. Qualifications are also subject to qualification delays, which can make determining relevant qualifications even more difficult. In addition, determining new qualifications and re-qualifications can be even more complex if some parameters, such as the demand by product, are subject to uncertainty and vary over time (see Chapter 6). Qualification management tends to be increasingly complex as the number of products, recipes and machine increases over time as the portfolio of products grows.

As a result, this raises complex capacity planning and flexibility optimization problems as the right amount capacity must be allocated to each product to satisfy the demand and improve operational efficiency. Therefore, advanced methods, *e.g.* operations research, simulation, statistics, that can identify critical qualifications at the tactical decision level and that can identify critical re-qualifications at the operational decision level are required. Existing methods are reviewed in Section 1.4. Our industrial and scientific contributions are presented in Section 1.5.

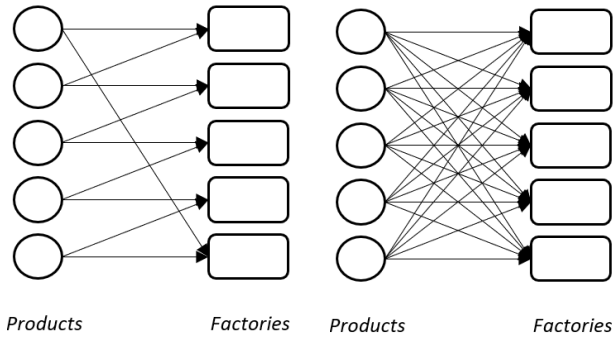
1.4 Literature review

1.4.1 Process flexibility

Qualification management is closely related to the the notion of process flexibility. The scientific literature on process flexibility is mostly interested in measuring the performances of process flexibility designs (which could be called qualification configurations or designs in the context of the thesis) in terms of expected service levels using notably linear programming and max-flow models. The term "link" is preferred to the term qualification. In general, the literature on process flexibility deals with strategic problems at the supply chain level. Links are determined between products and factories. The quality of links (the quality of the process flexibility design) between products and factories is evaluated. Link costs are constrained to a given budget. For instance, if n is the number of factories and products, then 2-chain designs considers at most $2n$ links.

Under balanced (same number of factories and products) and symmetrical assumptions (each unit of product leads to the same amount of workload at any plant), given a set of demand scenarios (demand is assumed to be independent and identically distributed), the seminal work of Jordan and Graves (1995) shows that effective sparse flexibility designs with at most $2n$ links can almost achieve the same benefits as full flexibility designs. In particular, they show that 2-chain designs (also referred as long chain designs in the literature) where each product is exactly linked to two factories (see Figure 1.6a) and where the design forms undirected cycle containing all machines and products, is almost as effective as the full flexibility designs (see Figure 1.6b). with much fewer links. They also show that there can

exist multiple process flexibility designs with similar performances. Chain designs perform better than other sparse designs as they pool more products and factories, thus allowing to better face demand uncertainty. Based on this work, Graves and Tomlin (2003), Chou et al. (2010), Simchi-Levi and Wei (2012), Wang and Zhang (2015), Désir et al. (2016) and Bidkhori et al. (2016) further study, validate and complement the benefits of sparse and chain flexibility designs.



(a) 2-chain design. (b) Full flexibility design.

Figure 1.6: Visual comparison of different flexibility designs.

Nevertheless, the main limits of applying chain flexibility designs to qualification management in semiconductor factories are:

- Most often, only balanced systems (same number of factories and products) are studied, which is unrealistic in semiconductor factories.
- Any plant can be linked to any product. This is impossible in qualification management in semiconductor factories due to continuous investment. Machines belong to different generations, have different software and hardware restrictions and can be of different types. They cannot be qualified for the same fabrication operations. Consequently, chain designs are unlikely.
- Most often, only symmetrical systems (each unit of product leads to the same amount of workload at any plant) are studied. This is not true in semiconductor factories as two products can require different operations with different processing times.
- Link delays are not considered. In semiconductor manufacturing, the qualification process may take several weeks to several months.
- Single period models are considered. However, demands of products are highly dynamic (see *e.g.* Figure 1.1), which cannot be easily captured with single period models.
- Demand is assumed to be independent and identically distributed. In high-mix factories, demands are not independent and identically distributed. Typically, a few products are associated to most of the demand.

Some authors contribute to the process flexibility literature by waiving some of these assumptions. Mak and Shen (2009) propose a two-stage stochastic programming approach to determine process flexibility designs. The studied setting is a balanced system. Process flexibility costs are distinguished by factory. Factories have different production capacities and two products can lead to different workloads. They show that, when the demand by product is heterogeneous, the flexibility design determined with the stochastic programming approach generates a better profit than chain designs. For an unbalanced and unsymmetrical system, Chou et al. (2010) identify underlying conditions such that sparse (not necessarily chained) flexibility designs achieve most of the benefits of the full flexibility design. They also show that adding a restricted number of links is often sufficient to significantly improve the ability of a production system to meet the demand. Deng and Shen (2013) formulate recommendations for process flexibility designs for unbalanced but symmetrical systems. Bidkhori et al. (2016) derive a lower bound for chain designs when systems are unbalanced and factories have different production capacities. Chen et al. (2019) further study unbalanced and unsymmetrical systems by proposing a simple scheme to satisfy the expected demand with high probability in a single-period setting. Shi et al. (2019) study flexibility designs in a multi-period setting. Fiorotto et al. (2018) present a *deterministic* lot-sizing problem motivated by the semiconductor industry. Authors propose two different lot-sizing optimization models to build the best long chain configuration or to find the best links (the total number of links is limited to a given number) while trying to minimize setup, inventory holding and backlogging costs. They analyze different flexibility designs and compare them to different long chain designs (Jordan and Graves, 1995). They show that, when the capacity is tight or when inventory and backlogging costs are very different from one product to another, scenarios that are actually frequently encountered in HM factories, even the best long chain design is not satisfactory. Flexibility links can be misplaced because backlogging costs and setup times are not considered in the long chain principle. The authors show that the optimization obtains better cost effective designs with half the links used by the long chain design.

To improve realism and for a relevant usability in semiconductor manufacturing, most of the assumptions above should be waived, in particular assumptions on the symmetry and balance of production systems. As it is unlikely to determine analytic formulas under such conditions to help determine relevant process flexibility designs, and therefore relevant qualification configurations, solving complex combinatorial optimization problems is required as shown in Mak and Shen (2009) and Fiorotto et al. (2018).

1.4.2 Qualification management in semiconductor manufacturing

Papers that deal with qualification management in semiconductor manufacturing are reviewed, but not papers where qualifications are parameters or constraints of the problem. In total, we found 31 papers that deal with qualification management to improve the manufacturing performances of work centers in semiconductor

manufacturing. The papers were mostly identified by reading two thesis on qualification management (Johnzén, 2009; Rowshannahad, 2015) and reading references therein. Figure 1.7, with the number of cumulative publications by year, shows that the literature is rare and relatively recent. This can be explained by the fact that the semiconductor industry is a complex process industry and, because qualifications are long and expensive, changing qualifications or adding costly qualifications may have not been of great importance in the past. However, with the normalization and development of custom products, products with short life cycles, and because of fierce competition, semiconductor manufacturers are more prone to change or add new qualifications on machines to keep or increase their competitive advantage (Johnzén et al., 2007, 2008). The expertise of semiconductor manufacturers and machine suppliers on machines improved over time, thus allowing more recipes to be qualified at the same time on a given machine without yield losses. In addition, the literature worked first on scheduling problems, which are complex problems, in particular in the semiconductor industry (Tamssaouet, 2019).

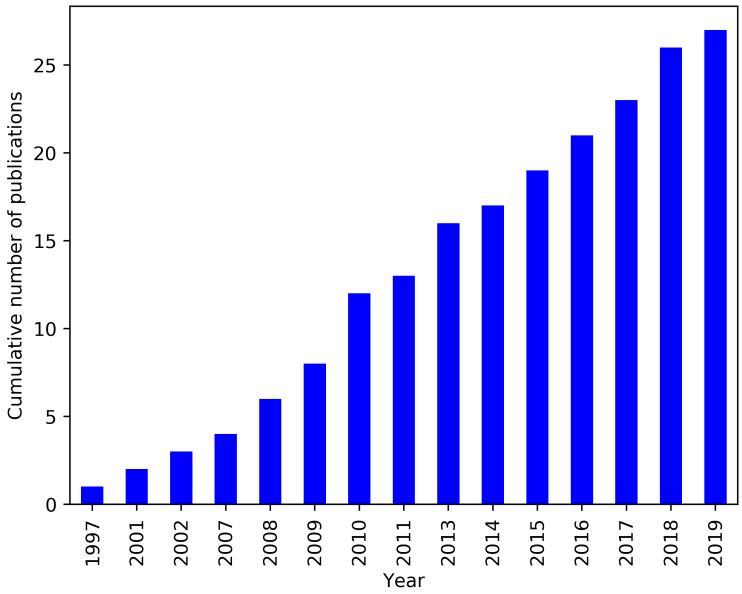


Figure 1.7: Cumulative number of publications on qualification management by year.

Note that the literature is not directly presented in terms of operational and tactical qualification management. This is because most papers do not separate or mention that there exist two separate decision levels for qualification management. A classification of papers is made in the literature review. In addition, note that the literature is presented in a general way. Differences between our contributions and the literature is recalled in each chapter.

1.4.2.1 Extended production planning problems and capacity allocation

Extended production planning problems. Some papers in the literature consider extended production planning problems where qualifications are modeled as addi-

tional "setup" constraints. The setup cost associated to a qualification has to be paid once. Machines can process recipes as long as the qualification constraint is active. The setup cost must be paid once again when the qualification is lost. Extended production planning problems are operational decision problems. Note that authors do not present their work as extended production planning problems although holding inventory and backlogging costs are minimized alongside qualification costs.

Fu et al. (2010) present an extended production planning problem where the objective is to minimize the total production costs, *i.e.*, production, inventory, backlogging and qualification costs. Kopp et al. (2016) consider a qualification management problem for a photolithography work center. The problem is solved by using a Mixed-Integer Linear Programming (MILP) approach. To our knowledge, Kopp et al. (2016) are the first authors to introduce qualification time windows in a qualification management optimization problem. A discrete-event simulation model is used to compare the trade-off between backlogging and qualification costs given different costs and bottleneck utilization scenarios. Kopp et al. (2018) and Kopp et al. (2019) propose a MILP and a simulation model to evaluate different re-qualification strategies for a photolithography work center. Disqualifications occur when qualification time windows expire, with unexpected events on recipes or with machine failures. Re-qualification strategies are assessed in terms of mean cycle time, throughput and tardiness with simulation. Re-qualifications are performed in a rolling horizon to better face uncertainty on the demand and the machine unavailability. Re-qualifications are modeled with a 75-minute delay in the simulation model. In addition, Kopp et al. (2019) propose dispatching strategies to avoid reaching the end of qualification time windows and maintain qualifications determined by the MILP. Kopp and Mönch (2019) propose and compare several heuristics and matheuristics to solve the MILP introduced by Kopp et al. (2016). Finally, in (Fu et al., 2015) consider that the demand in a back end facility is stochastic and described with a probability distribution. The authors use a L-shaped method and develop "qualification cuts" to solve the optimization problem.

Capacity allocation. Ignizio (2009) and Ignizio (2010) deal with qualification management at a tactical decision level. Ignizio (2009) proposes a MILP model to determine the qualification configuration of a work center. The objective consists in maximizing the workload balance and minimizing the number of different photoresists used by machines. A simulation model is used to assess the impact of the qualification configuration on the mean cycle time of the manufacturing facility. Specific constraints of the photolithography work center are also included, *e.g.*, reticle management. Different methods for generating qualification configurations are compared. The mathematical programming approach is shown to give a lower mean cycle time than greedy heuristics or rule-based approaches. Similarly, Ignizio (2010) studies the capacity of different qualification configurations to handle uncertainty in a wafer fab. Ignizio (2010) shows that genetic algorithms are more prone to determine qualification configurations that cope with uncertainty. Liao et al. (2017) propose a two-stage stochastic programming optimization approach to maximize the total profit of a semiconductor company. The first stage problem consists in minimizing qualification costs (production facilities are qualified for products or technology nodes) while the second stage problem consists in

allocating product quantities to production sites to maximize revenue. In the second stage problem, the demand for all products is satisfied and the capacity constraints are respected. Chang and Dong (2017) consider a two-stage capacity allocation stochastic programming problem. The demand is stochastic, and processing times are stochastic. Stochastic capacity losses are also associated to qualification decisions. They propose a Lagrangian relaxation to solve the stochastic programming problem. Uncertainty is described with probabilities.

1.4.2.2 Hierarchical approaches

Klemmt et al. (2010) present a four-stage mathematical programming approach to optimize the performances of a photolithography work center. The authors are motivated by the fact that even best scheduling decisions cannot significantly improve manufacturing performances if qualifications are not correctly prepared with respect to different demand scenarios. A hierarchical approach is then necessary to ensure that all decisions are consistent with each other. The first two stages are strategic stages. They are used to define machine qualifications and photoresist allocations given load balancing and throughput objectives. The two last stages are operational stages. Given qualifications and resist allocations, reticle and lot assignments are optimized. The lower the decision level, the more detailed the process constraints. As qualifications are made in the two first stages, Klemmt et al. (2010) consider the tactical decision level.

Similarly, Kopp and Mönch (2018) introduce a three-level hierarchical approach to better manage machine qualifications in a lithography work center. The top level determines target production quantities. Based on the target production quantities, a MILP problem is solved to recommend re-qualifications. Finally, the base level uses the recommended re-qualifications and simulates the shop-floor dispatching system.

In both papers, information from lower levels is sent back to upper levels and qualification decisions are updated over a rolling horizon to better face uncertainty.

1.4.2.3 Qualification management and production scheduling

Papers in this section mostly deal with qualification management at a tactical decision level. More precisely, papers use simulation to assess different qualification configurations and scheduling policies in terms of mean cycle time.

Fowler et al. (1997) compare different scheduling policies with different qualification configurations. They show that decreasing machine dedication can lead to substantial improvements on the mean cycle time of lots in a work center. Akcalt et al. (2001) use simulation models to assess process control policies and machine dedication policies on the mean cycle time in a lithography work center. Similarly, Kabak et al. (2013) use discrete-event simulation to assess the impact of recipe restrictions and disqualifications on the mean cycle time spent in a lithography work center. Similarly to Fowler et al. (1997), these studies show that adding new qualifications to photolithography machines can lead to substantial reductions in the mean cycle time of lots. However, the cycle time reduction gets smaller as the production

volume increases. Johnzén et al. (2008) study the impact of additional qualifications (or the impact of re-qualifications) on scheduling decisions using simulation models. However, contrary to Fowler et al. (1997), Akcalt et al. (2001), and Kabak et al. (2013), they show that, although additional qualifications increase flexibility, they do not necessarily lead to a decreased cycle time. These different results may be caused by different initial qualification rates and by data sets that are either generated in different manners or coming from different types of factories.

Finally, Aubry et al. (2008) introduce a MILP model to solve a production scheduling problem. The problem consists in finding the minimum number of qualifications so that the production plan is feasible (each recipe is assigned to at least one machine), and machines have the same workload. Preemption is allowed. Furthermore, Rossi (2010) and Aubry et al. (2012) assume that satisfying the demand by product is a key issue to characterize the robustness of a set of qualifications. Rossi (2010) seeks to characterize the robustness of a set of qualifications by determining the largest perturbation from the nominal demand while ensuring that the deadline is respected. Preemption is also allowed. Similarly, Aubry et al. (2012) seek to characterize the robustness of a set of qualifications by determining the largest perturbation from the nominal demand while ensuring that all machines have the same workload and that the qualification cost does not exceed a predefined value. Preemption is also allowed. Dedicated solution approaches are proposed to characterize the robustness of a set of qualifications. In particular, solution approaches are proposed to determine the largest stability radius around the nominal demand.

1.4.2.4 Assessing the qualification setting of a work center

The literature has also studied the definition of Key Performance Indicators (KPIs) to measure the quality of the qualification configuration of a work center and to guide qualification decisions. Most KPIs in the literature relate to flexibility measures, and mathematical models are also introduced to optimize the KPIs. Papers can be both used at operational and tactical decision levels. Flexibility measures can be seen as flexibility guidelines for semiconductor manufacturing similarly to 2-chain designs.

Johnzén et al. (2009) and Johnzén et al. (2011) propose flexibility measures to maximize the workload balance and minimize the total workload. “WIP”, “time” and “toolset” flexibility measures are proposed. The “WIP” (“WIP” stands for Work-In-Process) seeks to evaluate how balanced are recipe quantities, not in terms of workload but in terms of recipe units, between machines of the work center. Similarly, the “time” flexibility measure seeks to evaluate the workload balance between machines of the work center. The “toolset” flexibility measure seeks to evaluate the risk of having too many recipes with a small number of qualified machines. A system flexibility measure is also introduced, which is a weighted sum of the three previous flexibility measures. Flexibility measures are used to identify bottlenecks, the lack of flexibility and to assess the impact of a qualification or disqualification on the performance of a work center. Johnzén (2009) and Johnzén et al. (2011) propose a nonlinear qualification management optimization model to determine a single optimal qualification in terms of workload balance. Johnzén (2009) propose

simple greedy heuristics and local search approaches to solve the multiple qualification version of the optimization model. This work is extended in (Rowshannahad et al., 2015) by considering the finite production capacity, and thus the utilization rate, of each machine, to optimize the utilization balance and total utilization rate of the machines. No solution approach is proposed to solve the multi-qualification version of the problem in (Rowshannahad et al., 2015).

Rowshannahad and Dauzère-Pérès (2013) illustrate how the capacitated "time" flexibility measure can be used to better use the production capacity of machines. Rowshannahad and Dauzère-Pérès (2013) extend the "time" flexibility measure by considering batch size constraint. Rowshannahad et al. (2014) propose another measure to assess the workload variability between machines in a work center. Numerical experiments show that reducing workload variability between machines with additional qualifications significantly improves the workload balance. Finally, Pianne et al. (2016) introduce ideal and potential flexibility measures, and also consider the work center robustness, that is if a machine is sufficiently qualified to mitigate the down times of other machines.

1.4.2.5 Decision support systems

Interestingly, although the literature is rare on qualification management, there exist cooperation projects on qualification management between academics and semiconductor manufacturers (Leachman et al., 2002; Johnzén et al., 2009; Rowshannahad, 2015; Liao et al., 2017). Leachman et al. (2002) present a project and a decision support system (DSS) that enabled a wafer fab to significantly reduce its mean cycle time and make substantial savings. A key element for this success was the preparation of the right qualifications with respect to the production plan. Rowshannahad (2015) and Johnzén et al. (2009) describe qualification management software solutions that implement the "WIP", "time", "toolset" and system flexibility measures to recommend a single qualification decision to production personnel. Finally, Liao et al. (2017) also describe a DSS.

1.5 Industrial and scientific contributions

In this section, the industrial and scientific contributions are presented. The thesis pursues the work of two previous thesis on qualification management (Johnzén, 2009; Rowshannahad, 2015). The purpose of the thesis is to provide effective and efficient methods for qualification management at both the operational and tactical decision levels for high-mix wafer fabs. In particular, the purpose of the thesis is to answer the following questions:

- **Operational decision level.** How to determine the most relevant re-qualifications to improve operational efficiency? In other words, how to determine the most relevant re-qualifications to maximize the utilization balance and minimize the total utilization rate of the machines in order to maximize the throughput and minimize the cycle time?

– **Tactical decision level.** How to determine the most relevant new qualifications to satisfy the demand and cover the demand uncertainty, while respecting production capacities of machines and minimizing qualification costs?

These questions are answered in the industrial context of front-end manufacturing in a 300 mm wafer fab of STMicroelectronics in Crolles, France. The wafer fab is characterized by the following features:

– Several hundred products compete for the same production machines.

– Frequent product changes and demand uncertainty.

– Production variability (time-varying demand, WIP, production capacities, qualifications).

– High automation level in terms of dispatching and transportation decisions.

– Work centers are “unbalanced”. The number of products and recipes is (much) larger than the number of machines.

– Work centers are “unsymmetrical”. The wafer of a given product does not lead to the same workload as the wafer of another product, as both products may not have the same fabrication routes and re-entrant factor flows and may not require the same operations and recipes.

These features have a significant influence on the modeling choices in the thesis.

Note that although the questions are answered in the context of 300 mm front-end manufacturing facility, the same qualification management problems can be found in other front-end wafer fabs and also in back end factories. Consequently, the methods proposed in the thesis are not limited to 300 mm wafer fabs and can be applied to other semiconductor factories, and other industries with similar flexibility problems.

New optimization models, *i.e.* operations research techniques, and solution approaches are proposed in the thesis to answer the questions at the operational and tactical decision levels. The optimization models and solution approaches are included in decision support systems. For each proposed optimization model, a computational study is performed. Not only is the computational study used to validate the fact that the optimization model is relevant, but also used to show the limits of the optimization model. In addition, managerial recommendations are drawn in each chapter, which are not necessarily restricted to qualification management. The remainder of the thesis manuscript is organized as described in Sections 1.5.1 and 1.5.2.

1.5.1 Operational qualification management

In the context of operational qualification management, the question ”How to determine the most relevant re-qualifications to improve operational efficiency?” is answered from different standpoints.

In terms of utilization balance and total utilization rate. In Chapter 2, we pursue the work of Johnzén et al. (2011) and Rowshannahad et al. (2015) on qualification management for non-identical parallel machines. In other words, we seek to answer the question: How to determine the most relevant re-qualifications to minimize the total utilization rate and maximize the utilization balance of the machines?

To answer this question, Johnzén et al. (2011) propose a nonlinear qualification management optimization model to determine a single optimal qualification in terms of workload balance and total workload. This work is extended in (Rowshannahad et al., 2015) by considering the finite production capacity of each machine. New qualifications are therefore evaluated in terms of utilization balance and total utilization rate of the machines. To the best of our knowledge, the qualification management optimization problem with multiple qualifications and finite production capacity has never been considered. We propose and evaluate new efficient optimization approaches to answer the question. Optimization approaches determine in almost real time, *i.e.* in a small computational time, the best re-qualifications of operations in a work center with non-identical parallel machines. More precisely, the number of re-qualifications and the quantities by operation to process are given, and the objective consists in maximizing the utilization balance and minimizing the total utilization rate of the machines. Six new solution approaches, notably inspired by heuristics for discrete location problems and based on the analysis of dual variables, are proposed and compared on industrial data from a 300 mm wafer fab. The use of dual variables leads to heuristics that are effective both in terms of solution quality and computational time. The most relevant approach is now implemented in the decision support system presented in Chapter 7.

In terms of throughput. In Chapter 3, we seek to answer the question: "How to determine the most relevant re-qualifications to maximize the throughput?" The nonlinear optimization model proposed in Chapter 2 is extended to bilevel optimization approaches to answer the question in terms of throughput. In other words, in Chapter 3, we seek to determine the best re-qualifications to maximize the throughput. First, we argue why a bilevel optimization approach is a suitable approach to maximize the throughput at an operational decision level, and why the approach presented in Chapter 2 can be limited in some cases. Furthermore, to the best of our knowledge, there is no contribution in the literature that proposes to model disqualification decisions whereas can be important for operational qualification management. Therefore, a bilevel optimization model is proposed to cover the case where disqualification decisions must be made. A bilevel optimization approach is also proposed to combine qualification and disqualification decisions. Finally, a computational study on industrial data from a 300 mm wafer fab is performed to validate the proposed bilevel optimization models.

The bilevel optimization approach presented in Chapter 3 is extended to a *dynamic* (multi-period) bilevel optimization approach in Chapter 4. More precisely, the demand and production capacities vary over time due to production and demand variability. We show, on industrial data from a 300 mm wafer fab, that the dynamic approach can be more appropriate than the (static) bilevel optimization

approach presented in Chapter 2 to propose relevant qualifications when they are subject to lead times or induce maintenance operations.

In terms of mean cycle time. In Chapter 5, we seek to answer the question: "How to determine the most relevant re-qualifications to minimize the mean cycle time?" More precisely, Chapter 5 is dedicated to the study of the effect of re-qualifications on the mean cycle time. This is motivated by the fact that minimizing the cycle time reduces production costs. It is first argued that simple closed-form solutions describing the mean cycle time are available at an operational level for work centers. Second, the relevance and the limits of closed-form solutions are shown for different work centers on industrial data from a 300 mm wafer fab. The effect of re-qualifications on short-term cycle time is then illustrated. In particular, we show that there can exist two re-qualifications that lead to the same gain on the throughput but different gains on the mean cycle time. In addition, it is shown that most re-qualifications are irrelevant to minimize the cycle time but relevant re-qualifications can significantly minimize the mean cycle time.

Decision support system. A fully functional decision support system used at STMicroelectronics by production personnel is presented in Chapter 7, in particular, the purpose, the functioning, and the content of the decision support system. The decision support system embeds all theoretical developments of Chapters 2, 3, 4 and 5. The decision support system is now included in the decision process of some work centers.

1.5.2 Tactical qualification management

In the context of operational qualification management, the question "How to determine the most relevant new qualifications to satisfy the demand and cover the demand uncertainty, while respecting production capacities of machines and minimizing qualification costs?" is answered.

Chapter 6 is dedicated to tactical qualification management. Optimization approaches are proposed to answer the question. A new mixed integer linear programming mathematical model is proposed for the considered tactical qualification management problem when the demand is deterministic and qualification lead times are considered. Qualification costs must be minimized while the demand by product and production capacities of machines must be satisfied.

As the demand by product is subject to uncertainty, the choice of robust optimization is motivated. An uncertainty set based on the budget of uncertainty (Bertsimas and Sim, 2004) is proposed to cover the uncertainty on the demand by product. A new robust optimization model is introduced. In addition, a new decision-dependent uncertainty mathematical program is proposed to characterize the robustness of a set of qualifications, therefore the robustness of a work center, against demand uncertainty. A binary search is proposed to characterize the robustness of a set of qualifications because the decision-dependent uncertainty program is NP-Complete. In a computational study on industrial data from a 300 mm wafer

fab, we show that: (1) The price of uncertainty is acceptable, often less than a few additional qualifications by machine, (2) It is possible to achieve the same level of robustness as the case where all new qualifications are performed with only a restricted number of relevant qualifications, (3) Depending on the forecast uncertainty and the work center, the robust optimization problem can be difficult to solve, and (4) Using the nominal set of qualifications can lead to significant capacity constraint violations although it can be used for some work centers when the forecast uncertainty is small. Finally, practical applications and implications of the proposed models are discussed.

1.5.3 Remarks

Chapters 2, 3, 4, 5 and 7 present scientific and industrial contributions for operational qualification management (qualification management performed at an operational decision level), while Chapter 6 relates to contributions for tactical qualification management (qualification management performed at a tactical decision level). Finally, conclusions and perspectives are outlined in Chapter 8.

It is worth mentioning that, although some chapters are more dedicated to operational qualification management or tactical qualification management, the proposed mathematical models can be used for qualification management at other decision levels. For instance, the optimization model for determining qualification plans in Chapter 2 can be used to further distinguish performances of qualification configurations that would have the same cost in Chapter 6 (see Appendix F). Similarly, optimization models proposed in Chapter 6 can be adapted for qualification management at a supply chain level. It is sufficient to replace input parameters by, for instance, replacing machines by wafer fabs, and production capacities of machines by production capacities of wafer fabs.

Chapter 2

Managing re-qualifications to optimize utilization balance and total utilization rate of machines

In this chapter, the question "How to determine the most relevant re-qualifications to improve operational efficiency?" is answered from a utilization balancing standpoint. More precisely, we put ourselves in the shoes of a work center manager who must decide on the short term the best re-qualifications to maximize the utilization balance and minimize the total utilization rate of the machines. We show on industrial data that it is possible to efficiently and effectively maximize the utilization balance and minimize the total utilization rate of the machines by using the dual variables of qualification constraints to guide solution approaches*.

*Most of this chapter has been submitted to an international journal. Solution approaches were also presented at the ROADEF 2020 conference, and the author was finalist for the best student paper.

2.1 Introduction

We pursue the work of Johnzén et al. (2011) and Rowshannahad et al. (2015) on qualification management on non-identical parallel machines. Johnzén et al. (2011) propose a nonlinear qualification management optimization model to determine a single optimal qualification in terms of workload balance and total workload. This work is extended in (Rowshannahad et al., 2015) by considering the finite production capacity, and thus the utilization rate, of each machine, to optimize the utilization balance and total utilization rate of the machines. To the best of our knowledge, the qualification management optimization problem with multiple qualifications and finite production capacity has never been considered. Therefore, we propose and evaluate new effective and efficient optimization approaches that determine in real time, *i.e.* in a small computational time, the best re-qualifications of operations in a work center with non-identical parallel machines. The most relevant approach is now implemented in a decision support system presented in Chapter 7. The remainder of the chapter is organized as follows. The problem is formalized as a Mixed Integer NonLinear Program (MINLP) in Section 2.2, and solution approaches are proposed in Section 2.3. In Section 2.4, computational experiments on industrial data are presented and discussed, followed by recommendations based on the numerical results in Section 2.4. Finally, we conclude and give perspectives in Section 2.6.

2.2 Problem definition and analysis

Let us consider a work center of M non-identical parallel machines which must process R different operations with a strictly positive demand. Machines are non-identical, both in terms of qualifications and throughput rates. More precisely, machines are unrelated, *i.e.* there is no machine that is systematically faster than another machine for all operations. Machines performing the same type of operations were most often not acquired together, and thus belong to different generations. In addition, machines do not have the same core competencies, *i.e.* all machines do not process the same types of operations. A machine can only process qualified operations, and a qualifiable operation can be processed on a machine if it is already qualified. The qualification matrix between operations and machines is known, and each operation has a throughput rate on the machines on which it is qualifiable. Each machine has a finite capacity, which can be different from other machines. Among the qualifiable pairs (operation, machine) not already qualified, the objective is to determine a re-qualification plan consisting of k re-qualifications in order to maximize the utilization balance and minimize the total utilization rate of the machines.

2.2.1 Problem modeling

The notations used in the chapter are listed below.

Parameters:

$q_{r,m} \in \{0,1,2\}$: Is equal to 1 if machine m is currently qualified for operation, r, is equal to 2 if machine m is qualifiable for operation r, and is equal to 0 if machine m cannot be operation for product r,
k: Number of re-qualifications,
$tp_{r,m}$: Throughput rate (in number of wafers per second) of operation r on machine m,
c_m: Production availability (in seconds) of machine m,
d_r: Quantity of operation r to produce,
γ: Utilization balancing parameter, which is strictly greater than 1.

Variables:
$OQ_{r,m} \in \{0,1\}$: Is equal to 1 if operation r should be qualified on machine m, and is equal to 0 otherwise,
U_m: Utilization rate of machine m,
$WIP_{r,m}$: Quantity of operation r assigned to machine m.

The problem is formalized below as a Mixed Integer NonLinear Program (MINLP):

$$f_1 = \min \quad \sum_m U_m^\gamma \tag{2.1}$$

$$\text{s.t.} \quad \sum_{r,m} OQ_{r,m} \leq k \tag{2.2}$$

$$U_m = \sum_r \frac{WIP_{r,m}}{tp_{r,m} c_m} \qquad \forall m \tag{2.3}$$

$$\sum_m WIP_{r,m} = d_r \qquad \forall r \tag{2.4}$$

$$WIP_{r,m} \leq d_r \qquad \forall r, \forall m \mid q_{r,m} = 1 \tag{2.5}$$

$$WIP_{r,m} \leq d_r OQ_{r,m} \qquad \forall r, \forall m \mid q_{r,m} = 2 \tag{2.6}$$

$$WIP_{r,m} \leq 0 \qquad \forall r, \forall m \mid q_{r,m} = 0 \tag{2.7}$$

$$WIP_{r,m} \geq 0 \qquad \forall r, \forall m \tag{2.8}$$

$$OQ_{r,m} \in \{0,1\} \qquad \forall r, \forall m \tag{2.9}$$

The objective function (2.1) aims at finding a compromise between the utilization balance of the machines and their total utilization rate. The larger γ, the higher the priority on the utilization balance (see Section 2.2.3). Constraint (2.2) limits the number of re-qualifications, *i.e.* the size of the optimized qualification plan, to at most k. Constraints (2.3) compute the utilization rate for each machine in the work center. The machine utilization rate should be understood as the "implied" machine utilization rate by the operation quantities assigned to the the machine. A machine utilization rate is not necessarily lower than or equal to 1 if the machine cannot process all its assigned operation quantities on the horizon. Constraints (2.4) ensure that the demand of each operation is fully assigned to the machines. Constraints (2.5)-(2.7) ensure that machine m can only process operation r if r is currently qualified on m ($q_{r,m} = 1$) or is both qualifiable and proposed to be qualified ($q_{r,m} = 2$ and $OQ_{r,m} = 1$). Note that the dual prices of Constraints (2.5)-(2.7) indicate the potential gain in terms of utilization balance (Bazaraa et al., 2013), and will be used in

some of the heuristics proposed in Section 2.3. Finally, Constraints (2.8) are the non-negativity constraints, and Constraints (2.9) are the binary constraints.

Let us discuss below some important characteristics of our problem:

- This problem is equivalent to maximizing the capacitated time flexibility measure proposed by Rowshannahad et al. (2015), which evaluates the quality of the balance of the qualification configuration of a work center in terms of utilization rates of machines. Solving this optimization model enables decision makers to compare re-qualifications in terms of utilization balance of the machines, and therefore select the best re-qualifications to reduce productivity losses.
- All re-qualifications require the same cost and time. This assumption comes from production personnel that can hardly differentiate between re-qualifications at the operational level. On a longer horizon of several weeks or months, where new qualifications need to be planned (tactical qualification management), considering different costs and times for qualifications is relevant, although the information might not be easy to obtain. This is studied in Chapter 6, which is dedicated to tactical qualification management.
- We also assume that re-qualifications can be performed very quickly. Even tough it is not always the case depending on disqualification reasons, assuming that re-qualifications can be almost done instantly provides insights on critical qualifications which should have been active to maximize the workload balance and minimize the total workload. This assumption is relaxed in Chapter 4.
- Demands and production capacities varying over time and disqualifications are not considered. This is because the problem is solved regularly, *e.g.* once every shift of 8 hours for the next 24 hours, and the qualifications are frequently updated given the current disqualifications and a new estimate of the quantities to process. Including disqualifications and time varying demands and production capacities in the problem on a longer planning horizon is studied in Chapters 3 and 4.

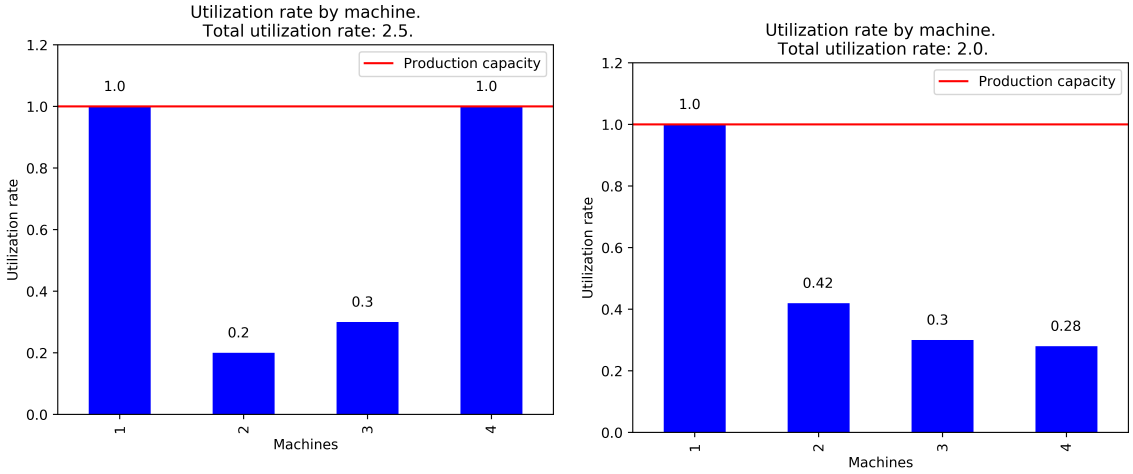
2.2.2 Illustrative example

Consider a work center with four machines and seven operations with the following parameters:

$$q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix}, tp = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.8 & 0.2 & 0 \\ 0 & 0.2 & 0.8 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 1 \end{pmatrix}$$

$$d = \begin{pmatrix} 100 & 200 & 200 & 100 & 100 & 100 & 300 \end{pmatrix}, c = \begin{pmatrix} 300 & 200 & 200 & 300 \end{pmatrix}$$

Machines are initially poorly balanced (see Figure 2.1a and Table 2.1). This can be due to the fact that a critical machine is currently down, or because of the demand mix. Re-qualifications must be performed to maximize the utilization balance and minimize the total utilization rate of the machines.



(a) Initial machine utilization rates.

(b) New machine utilization rates.

Figure 2.1: Comparison of the initial machine utilization rates (a) and the new machine utilization rates after one qualification (b) with $\gamma = 4$.

In Figure 2.1a, the utilization rates of machines 1 and 4 are equal to 1.0. Both machines are then critical, and an option to reduce their utilization rates is to re-assign part of their workload to machines 2 or 3, which are less loaded. This is possible by performing a re-qualification (or new qualification).

If operation 7 is qualified on machine 2, a large part of the workload of machine 4 is reassigned to machine 2. The utilization rate of machine 4 is strongly reduced, from 1.0 to 0.28, and the utilization rate of machine 2 is increased, from 0.2 to 0.42 (see Figure 2.1b). Machines 2, 3 and 4 are now better balanced, and the manufacturing performance of the work center is expected to improve. Concretely, maximizing the utilization balance and minimizing the total utilization rate of the machines improve productivity as more wafers should be produced in less time. A better utilization balance of the machines means a better throughput and less backlog. In addition, machines can better undergo the failure of a critical machine when the utilization rates of the machine are better balanced. The objective function is reduced by 48% (Table 2.1), and a single qualification significantly improves the utilization balance of the machines. Chapter 3 details why the throughput cannot be directly maximized (mostly due to highly automated dispatching decisions). The quality of utilization balance in terms of throughput is further studied in Chapter 3 and in Chapter 4, and the quality of utilization balance in terms of cycle time is studied in Chapter 5.

	Utilization rate by machine				
Configuration	1	2	3	4	$\sum_m U_m^\gamma$
Before	1.00	0.20	0.30	1.00	2.01
After	1.00	0.42	0.30	0.28	1.04

Table 2.1: Comparison of the initial objective function (Figure 2.1a) and after a re-qualification (Figure 2.1b), $\gamma = 4$.

2.2.3 Justification of the nonlinear objective function

The objective of this chapter consists in determining re-qualifications that both minimize the total utilization rate and maximize the utilization balance of the machines. The objective of this chapter is not to solve a biobjective optimization problem because, in the industrial context, only one solution is expected by the decision makers. The main advantage of the nonlinear objective function is to provide flexibility in the trade-off between the total utilization rate and the utilization balance, while avoiding a biobjective problem without explicitly formalizing the utilization balance. Formalizing a criterion that models the utilization balance of the machines is actually complex, and classical indicators fail to do so because qualifications and throughput rates vary from one machine to another (see Section 2.2.3.2). By selecting an appropriate value of γ (see Section 2.2.3.1 and Rowshannahad et al. 2015), the total utilization rate or the utilization balance is more emphasized. However, considering a nonlinear objective function increases the computational burden, which motivates the development of efficient and effective solution approaches, in particular because the solution approaches must be embedded in an operational decision support system.

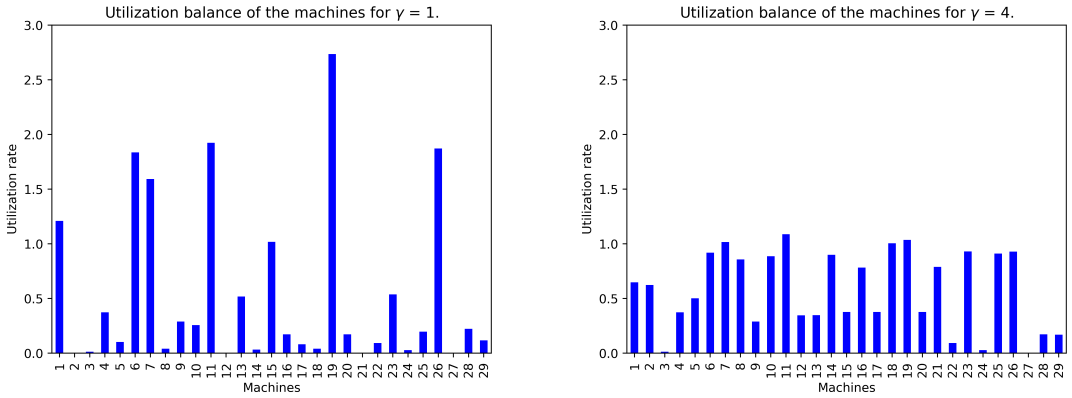
First, the influence of γ on the total utilization rate of the machines is illustrated in Section 2.2.3.1 with illustrative examples. Then, in Section 2.2.3.2, the nonlinear objective function is compared to two classical machine utilization balancing approaches. Finally, in Section 2.2.3.3, interpretations of the nonlinear objective function are proposed.

2.2.3.1 Influence of γ

γ is used to model a compromise between the total machine utilization rate and the machine utilization balance. With $\gamma = 1$, only the total machine utilization rate is minimized, and the larger γ, the more important the maximization of the machine utilization balance compared to the minimization of the total machine utilization rate (see also Rowshannahad et al. 2015). γ is therefore a choice. For instance, when $\gamma = 4$, a solution where two machines have utilization rates of 0.95 and 0.2 ($f_1 = 0.82$) is preferred to solution where both machines have a utilization rate of 0.9 ($f_1 = 1.31$). However, if maximizing the utilization balance of the machines is highly prioritized then, by using $\gamma = 20$ in the nonlinear objective function, the second solution where both machines have a utilization rate of 0.9 ($f_1 = 0.24$) is preferred to the first solution ($f_1 = 0.36$).

In particular for operational decision levels, Rowshannahad et al. (2015) recommend adjusting γ according to the real workload distribution on the shop floor, for instance by using historical data. Due to production variability and the short-sighted aspect of dispatching rules (see Chapters 3 and 4), for short-term horizons, small values of γ are more appropriate because the utilization balance of the machines is not perfect. For larger horizons such as one month, larger γ are appropriate because production variability is absorbed over time, and machines tend to be more naturally balanced. In the considered manufacturing system, for an horizon of 24 hours, $\gamma = 4$ and $\gamma = 6$ are considered appropriate values. For an horizon of one month, larger values of γ, such as 20, are acceptable.

Note that too small values of γ can lead to unrealistic solutions. Let us illustrate the practical use of the nonlinear objective function in the decision support system. Using $\gamma = 1$, where only the total machine utilization rate is minimized, would indicate that a machine is critical, *i.e.* overloaded with respect to other machines in the work center. Consider the illustrative example on industrial data in Figure 2.2. For $\gamma = 1$, the utilization rate of machine 19 is inflated, and some machines are not even used such as machines 2 and 12, contrary to the utilization rates obtained by solving the optimization model for $\gamma = 4$. From practical experience, this is often unrealistic. This would suggest to production personnel that the machine is critical, thus that qualification decisions (even postponing maintenance operations) are required to reduce its utilization rate. However, in practice, the critical machine is not necessarily as critical as initially thought. By increasing γ, the utilization rate of the critical machine can be greatly reduced by balancing it with other machines. Therefore, using $\gamma = 1$ would lead to poor decision making in practice.



(a) Utilization balance of machines for $\gamma = 1$.

(b) Utilization balance of machines for $\gamma = 4$.

Figure 2.2: Comparison of utilization balances of machines for (a) $\gamma = 1$ and (b) $\gamma = 4$.

2.2.3.2 Comparison to other balancing approaches

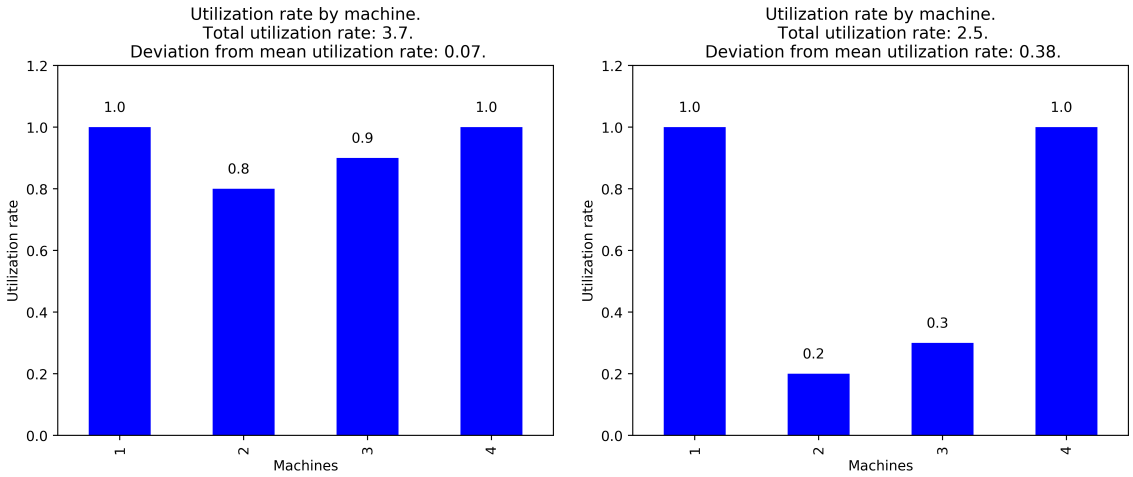
The utilization balance of the machines can be optimized with other objective functions than the nonlinear objective function $\min \sum_m U_m^\gamma$, namely a min-max objec-

tive function, $\min \max_m U_m$, and an objective function that measures the total deviation from the mean utilization rate of the machines, $\min \sum_m |U_m - \bar{U}|$ where $\bar{U} = \frac{1}{M} \sum_m U_m$. However, the utilization rate of some machines can be left unoptimized or inflated.

Deviation from mean utilization rate. An objective function that optimizes the total deviation from the mean utilization rate of the machines (hereafter called mean deviation approach) can suffer from the same problem than the nonlinear function objective if it was used with a very large value of γ. Machines can artificially have a large utilization rate, and thus possibly both a large total utilization rate and a large maximum utilization rate, to optimize the total deviation from the mean utilization rate. This is actually mostly due to the fact that machines have non-identical throughput rates.

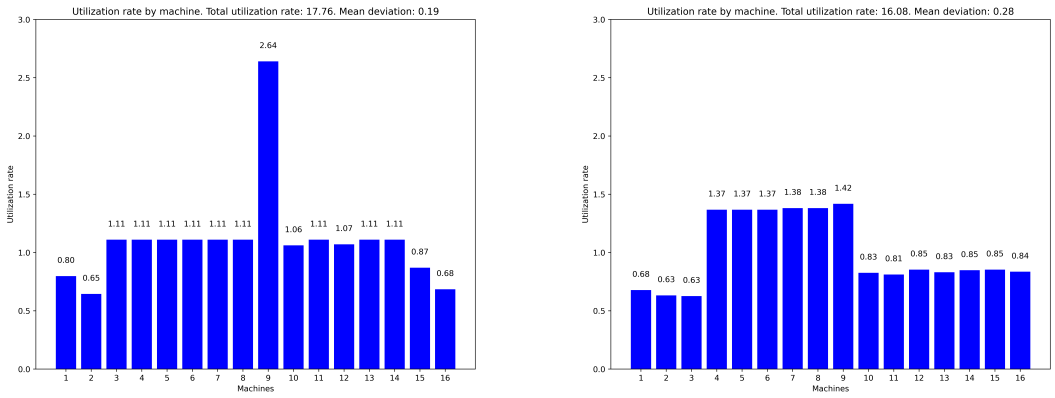
From a practical standpoint, showing the utilization balance of the machines obtained from the mean deviation approach to production personnel will lead them: (1) To incorrectly believe that some machines are extremely loaded or that machines are unbalanced, thus leading to poor decision making, or (2) To discard the proposed solution because it is unrealistic. First, let us consider the illustrative example in Section 2.2.2. Figure 2.3 compares the machine utilization rates when the deviation from the mean utilization rate of the machines is minimized and when the nonlinear function is minimized. The machine utilization rates obtained by minimizing the deviation from the mean utilization rate of the machines are much larger, and actually inflated to minimize the objective function, than the utilization rates of the machines obtained by minimizing the nonlinear function. Similarly, consider the example in Figure 2.4 with industrial data. It does not make sense that machine 9 has such a large utilization rate, whereas it can be balanced with machines 4, 5, 6, 7 and 8. Similarly, it can be observed that the utilization rates of machines 10, 11, 12, 13, 14, 15 and 16 are artificially increased to minimize the total deviation from the mean utilization rate, although the utilization rates of these machines are smaller with the nonlinear objective function.

Min-max approach (makespan minimization). When machines are non-identical in terms of qualifications, *i.e.* they cannot be qualified for the same operations, then min-max approaches may not be suitable to maximize the utilization balance or minimize the total utilization rate of the machines. This is because a min-max approach *only* considers the maximum machine utilization rate. Consider Figure 2.4, where both solutions are optimal for the min-max approach due to the fact that machines 1 and 4 cannot be balanced with other machines. Similarly, consider Figure 2.1, where both solutions are also optimal, even after one qualification, because machine 1 cannot be balanced with other machines. If the maximum machine utilization rate could not be minimized, then it would be possible to conclude that there is no qualification that can lead to a better utilization balance of the machines. Nevertheless, for a work center with a large number of machines, there may exist machines with utilization rates that are not equal to the maximum utilization rate and that can still be better balanced with qualifications.



(a) Mean deviation approach. (b) Nonlinear approach, $\gamma = 4$.

Figure 2.3: Comparison of the machine utilization rates obtained with the mean deviation approach and with the nonlinear objective function for the initial qualification configuration for the illustrative example.



(a) Mean deviation approach. (b) Nonlinear approach, $\gamma = 4$.

Figure 2.4: Comparison of the machine utilization rates obtained with the mean deviation approach and with the nonlinear objective function for the initial qualification configuration of a real work center.

2.2.3.3 Interpretation of the objective function

The nonlinear objective function cannot be directly interpreted as it does not represent something physical, such as the total machine utilization rate, the maximum machine utilization rate, the standard deviation of the machine utilization rates or the mean deviation from the mean utilization rate of the machines. This is why in practice, the nonlinear objective function is never presented or interpreted as is. In fact, the literature actually proposes two alternatives:

- It is possible to interpret the objective function by using other indicators than the objective function. For instance, the nonlinear optimization problem is solved, then the mean machine utilization rate, the maximum machine utilization rate, and the mean deviation from the mean utilization of the machines rate are presented along with the utilization balance. Presenting the solution, rather than indicators, *i.e.* the utilization balance of the machines, allows the objective function to be understood.

- As minimizing the objective function is equivalent to maximizing the capacitated time flexibility measure proposed by Rowshannahad et al. (2015), the objective function can be interpreted in terms of flexibility, which evaluates the quality of the utilization balance for a given qualification configuration.

- Another alternative consists in interpreting the objective function in terms of throughput or cycle time. This is particularly interesting at operational levels as production personnel is challenged to maximize the throughput and minimize cycle times. This option is investigated in Chapters 3, 4 and 5.

2.2.4 Computational complexity

Determining a single relevant re-qualification on the illustrative example presented in Figure 2.1 is straightforward, because the example is simple and the number of machines is limited. In practice, the throughput rates significantly vary from one operation to another and from one machine to another, and the numbers of operations and machines are large. Moreover, the effect of multiple additional re-qualifications on the utilization balance of the machines is difficult to capture as an initially overloaded machine can become less loaded than an initially underloaded machine after several re-qualifications.

Johnzén (2009) shows that optimizing the "WIP" flexibility measure is a strongly NP-Hard problem by reduction from the 3-partition problem (Garey and Johnson, 1979). The proof is based on the proof given in (Aubry et al., 2008) for the Minimum Cost Load Balanced Configuration Problem (MCLBCP). Optimizing the "WIP" flexibility measure is a special case of our problem, even when $tp_{r,m} = tp \ \forall r, \forall m$, and $c_m = 1 \ \forall m$. The proof in (Johnzén, 2009) is recalled in Appendix A, Section A.1 for the sake of completeness.

The studied optimization is NP-Hard. In addition, we want to tackle large scale industrial instances (see Section 2.4.1). Efficient solution approaches must thus be designed to propose effective re-qualification plans that can be used by production personnel in factories.

2.2.5 Outer linearization algorithm for solving the nonlinear program

Solving the continuous relaxation (or when $k = 0$) of the MINLP (2.1)-(2.9) is performed by using an outer linearization algorithm. The outer linearization algorithm is motivated by the fact that the nonlinearity only comes from the objective function. Hence, the objective function is separable on the decision variables U_m, and it is possible to give realistic bounds to U_m.

Consider Figure 2.5 for a given machine m, which illustrates how $f(U_m) = U_m^\gamma$ can be linearized using outer linearization. Outer linearization constraints of $f(U_m) = U_m^\gamma$ are given for $U_m = 0.5$ and $U_m = 1.0$. At u^0, the outer linearization equation is equal to $u_o^\gamma + \gamma u_o^{\gamma-1}(U_m - u_o)$. By adding a sufficient number of outer linearization constraints, the continuous relaxation (or when $k = 0$) of the MINLP (2.1)-(2.9) can be solved. Nevertheless, adding all possible outer linearization constraints is unpractical, as it will lead to adding an infinite number of constraints. Adding the most relevant outer linearization constraints is therefore critical to quickly solve the MINLP.

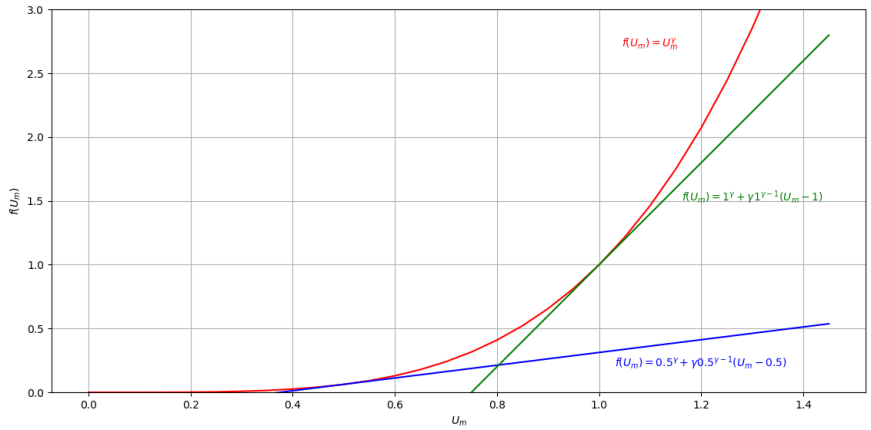


Figure 2.5: Outer linearization example for $f(U_m) = U_m^\gamma$ for machine m.

The outer linearization is performed for all machines separately. Consider that $O_m = \max_{o \in \mathcal{O}_m}(u_o^\gamma + \gamma u_o^{\gamma-1}(U_m - u_o))$, where $\mathcal{O}_m$ is the set of outer linearization points for machine m. Intuitively, O_m represents the value of U_m^γ when it is linearized by outer linearization. The objective function (2.1) then becomes min $\sum_m O_m$, where $O_m \geq u_o^\gamma + \gamma u_o^{\gamma-1}(U_m - u_o)$ $\forall m$, $\forall o \in \mathcal{O}_m$. The Linear Program (2.10)-(2.12) below provides a lower bound on the objective function:

$$\min \quad \sum_m O_m \tag{2.10}$$

$$\text{s.t.} \quad O_m \geq u_o^\gamma + \gamma u_o^{\gamma-1}(U_m - u_o) \qquad \forall m, \forall o \in \mathcal{O}_m \tag{2.11}$$

$$(2.2) - (2.9) \tag{2.12}$$

Equation (2.10) is the objective function. Constraints (2.11) are the outer linearization constraints. Constraints (2.12) are the qualification constraints, the utilization rate computation constraints, and the constraints ensuring that the total demand of operations must be assigned to qualified machines.

First, each set $\mathcal{O}_m$ is initialized with $0 \leq u \leq 8$. This is because, in industrial data and by experience, it is very unlikely for U_m to be larger than 8, even in a factory subject to high production variability. Once the linear program (2.10)-(2.12) is solved, $\boldsymbol{U}$ can be extracted from the incumbent solution to compute an upper bound on the objective function $\sum_m U_m^\gamma$. Then, additional outer linearization constraints are added to the sets $\mathcal{O}_m \; \forall m$ until the stopping condition, *i.e.* a small relative gap ϵ between the lower and upper bounds, is met. The outer linearization is detailed in Algorithm 1.

Algorithm 1 Outer linearization algorithm

procedure OUTER LINEARIZATION ALGORITHM
  $u_{min} \leftarrow 0$
  $u_{max} \leftarrow 8$
  $u_{step} \leftarrow 0.1$
  for $m = 1$ to M **do**
    $u_o \leftarrow u_{min}$
    while $u_o \leq u_{max}$ **do**
      $\mathcal{O}_m \leftarrow \mathcal{O}_m \cup u_o$
      $u_o \leftarrow u_o + u_{step}$
    end while
  end for
  gap $\leftarrow \infty$
  while gap $> \epsilon$ **do**
    Solve Linear Program (2.10)-(2.12) and compute $LB \leftarrow \sum_m O_m$
    $\boldsymbol{SU} \leftarrow \boldsymbol{U} = (U_1, ..., U_M)$
    $UB \leftarrow \sum_m SU_m^\gamma$
    gap $\leftarrow \frac{UB-LB}{LB}$
    for $m = 1$ to M **do**
      $\mathcal{O}_m \leftarrow \mathcal{O}_m \cup SU_m$
    end for
  end while
end procedure

For $\gamma = 4$, a gap of 0.0001 and the values of u_{min}, u_{max} and u_{step} in Algorithm 1, empirical observations on the industrial instances of Section 2.4 show that the algorithm converges in less than ten iterations. Comparing solution approaches to solve nonlinear programs could be valuable, but is beyond the scope of this study and is left for future research.

2.3 Solution approaches

In this section, new solution approaches are proposed to solve the optimization problem with multiple qualifications formalized in Section 2.2.

2.3.1 Constructive greedy heuristic

The first proposed algorithm is a greedy heuristic, which is inspired by the "ADD" heuristics for discrete location problems (Daskin (2011)). The greedy heuristic is a constructive heuristic that, at each iteration, selects the single best re-qualification that optimizes the nonlinear objective function and updates the qualification matrix. The procedure is repeated until a re-qualification plan of k re-qualifications is determined. The pseudo code of the algorithm can be found in Algorithm 2.

Algorithm 2 Greedy heuristic

Input data: q

procedure GREEDY HEURISTIC
  Best Plan $\leftarrow \emptyset$
  $f^* \leftarrow \infty$
  for $i = 1$ to k **do**
    $(r*,m*) \leftarrow \emptyset$
    for $r = 1$ to R **do**
      for $m = 1$ to M **do**
        if $(r,m) \notin$ Best Plan **then**
          $\boldsymbol{q}^{temp} \leftarrow \boldsymbol{q}$
          $q^{temp}_{r,m} \leftarrow 1$
          $f^{temp} \leftarrow f_1(\arg\min(2.1) - (2.9)$ with $\boldsymbol{q}^{temp}, k = 0)$
          if $f^{temp} < f*$ **then**
            $(r*, m*) \leftarrow (r, m)$
            $f* \leftarrow f^{temp}$
          end if
        end if
      end for
    end for
    Update qualification matrix, $q_{r*,m*} \leftarrow 1$
    Best Plan $\leftarrow$ Best Plan $\cup$ $(r*, m*)$
  end for
  return Best Plan
end procedure

2.3.2 Local search

The local search is a best improvement local search approach and is inspired by the "ADD-REMOVE" heuristics for discrete location problems (Daskin (2011)). The first

step consists in determining a feasible re-qualification plan with the greedy heuristic. Once a feasible re-qualification plan is determined, the local search removes one re-qualification at a time and tries to determine a better re-qualification. The local search terminates when it is no longer possible to determine a better re-qualification in the best re-qualification plan. The pseudo code of the local search can be found in Algorithm 3.

Algorithm 3 Local search

Input data: q

procedure LOCAL SEARCH
  Best Plan $\leftarrow$ Call Greedy Heuristic, Algorithm 2 (initialization step)
  $j \leftarrow 0$
  $i \leftarrow 0$
  while $j \neq k-1$ **do**
    Current Plan $\leftarrow$ Best Plan
    Current Plan $\leftarrow$ Remove the re-qualification at the i-th index
    $i \leftarrow i+1$
    $f^* \leftarrow \infty$
    $(r*, m*) \leftarrow \emptyset$
    for $r = 1$ to R **do**
      for $m = 1$ to M **do**
        if $(r,m) \notin$ Current Plan **then**
          $\boldsymbol{q^{temp}} \leftarrow \boldsymbol{q} \cup$ Current Plan
          $q^{temp}_{r,m} \leftarrow 1$
          $f^{temp} \leftarrow f_1(\arg\min(2.1)-(2.9)$ with $\boldsymbol{q^{temp}}, k=0)$
          if $f^{temp} < f*$ **then**
            $(r*, m*) \leftarrow (r, m)$
            $f* \leftarrow f^{temp}$
          end if
        end if
      end for
    end for
    Current Plan $\leftarrow$ Current Plan $\cup$ $(r*, m*)$
    if f_1(Best Plan) $<$ f_1(Current Plan) **then**
      Best Plan $\leftarrow$ Current Plan
      $j \leftarrow 0$
    else
      $j \leftarrow j+1$
    end if
    if $i = k$ **then**
      $i \leftarrow 0$
    end if
  end while
  return Best Plan
end procedure

2.3.3 Dual prices

Although heuristics presented in Sections 2.3.1 and 2.3.2 are starting points to determine good re-qualification plans, the number of re-qualifications to evaluate from one iteration to another can be substantial when the number of operations and machines are large. On industrial instances, a few thousand re-qualifications have to be evaluated, which is not acceptable when short computational times are required. Given the problem structure and the nature of the data, we know from practical (industrial) experience that only a restricted set of qualifiable pairs (operation, machine) can lead to valuable re-qualification plans in terms of utilization balance and total utilization rate of the machines.

For instance, let us consider the illustrative example in Section 2.2.2. The initial utilization balance is presented in Figure 2.1. Machines 1 and 4 are critical (*i.e.* $U_1 = 1.0$ and $U_4 = 1.0$) while machines 2 and 3 are underloaded (*i.e.* $U_2 < 1.0$ and $U_3 < 1.0$). Adding re-qualifications to machines 1 and 4 is irrelevant in terms of utilization balance because the machines would be even more loaded. Therefore, in this example, the search of the optimal re-qualifications can potentially be restricted to machines 2 and 3. All possiblere- qualifications could be tested for the example presented in Figure 2.1 as the number of operations and the number of machines are small. However, because many operations could be qualified on many machines in industrial data, evaluating all the possible re-qualifications is most often too time-consuming when short computational times are required.

To identify the most promising operations and machines, and therefore to reduce the number of re-qualifications from one iteration to another, the dual prices of the relevant constraints of the following reformulation (when $k = 0$) of the optimization model (2.1)-(2.9) can be used:

$$f_1 = \min \quad \sum_m U_m^\gamma \tag{2.13}$$

$$\text{s.t.} \quad U_m = \sum_r \frac{WIP_{r,m} d_r}{tp_{r,m} c_m} \qquad \forall m \tag{2.14}$$

$$\sum_m WIP_{r,m} = 1 \qquad \forall r \tag{2.15}$$

$$WIP_{r,m} \leq 1 \qquad \forall r, \forall m \mid q_{r,m} = 1 \tag{2.16}$$

$$WIP_{r,m} \leq 0 \qquad \forall r, \forall m \mid q_{r,m} \neq 1 \tag{2.17}$$

$$WIP_{r,m} \geq 0 \qquad \forall r, \forall m \tag{2.18}$$

The objective function (2.13) aims at finding a compromise between the utilization balance and the total utilization rate of the machines. Constraints (2.14) compute the utilization rate of each machine in the work center. Constraints (2.15) ensure that the demand of each operation is fully assigned to the machines. Constraints (2.16) and (2.17) ensure that machine m can only process operation r if it is qualified on m. Finally, Constraints (2.8) are the non-negativity constraints for variables $WIP_{r,m}$.

The optimization model is close to the initial model (2.1)-(2.9), but has some

significant differences. First, $WIP_{r,m}$ is redefined as the ratio of the quantity of operation r which is assigned to machine m. Second, the constraints imposing that the current qualifications are satisfied are differentiated. With these modifications, before any re-qualification decision, this optimization model can be solved and the dual variable of each Constraint (2.17) can be analyzed. The dual variable can then be interpreted as an approximation of the gain on the nonlinear objective function f_1 if operation r is qualified on machine m, as dual variables can be interpreted as "the marginal rate of change in the objective function with respect to perturbations in the right-hand side of a constraint" (Bazaraa et al., 2013). f_1 would become $\sum_m U_m^{\gamma} + \lambda_{r,m}$, where $\lambda_{r,m}$ is the dual variable for the pair (r,m) of Constraint (2.17). Analyzing the value of $\lambda_{r,m}$ for each pair (r,m), when $q_{r,m} = 2$, allows the most promising re-qualification decisions *to be ranked*. Note that the dual variables associated to Constraint (2.17) cannot be strictly positive because re-qualifications cannot decrease f_1.

By embedding the use of the dual variables in the greedy heuristic, instead of testing every possible re-qualification at each iteration, the search space can be greatly reduced to the N_{dual} most promising re-qualifications. For instance, at each iteration of the greedy heuristic, instead of testing 800 re-qualifications, only $N_{dual} = 10$ are tested. If the re-qualifications are tested in parallel, N_{dual} can be limited to the number of cores of the CPU. If, at a given iteration of the greedy heuristic, more than N_{dual} dual variables have the same value, the first ones in the list are arbitrarily selected. The pseudo code of the greedy heuristic with dual variables is provided in Algorithm 4. The same principle can be applied to the local search.

Another "Instantaneous" Greedy Heuristic (IGH) can be designed by using dual variables in a more straightforward way. IGH builds a feasible re-qualification plan with the k re-qualifications associated to the k smallest dual variables. Contrary to the greedy heuristic in Algorithm 4, IGH is not an iterative procedure since the k re-qualifications are taken just after the dual variables are computed. The pseudo code of the instantaneous greedy heuristic can be found in Algorithm 5.

2.3.4 Branch and bound

Similarly to the motivations that led to the use of dual variables, the design of the branch and bound (B&B) approach is motivated by the practical experience:

- Since there are only a limited number of preexisting qualifications, the qualification matrix is sparse and the overall number of possible re-qualifications is small.
- Industrial instances are considered, hence *non-randomly generated*, and only a restricted set of re-qualifications can lead to valuable re-qualification plans in terms of machine utilization balance.
- We are only interested in taking a limited number of re-qualification decisions.
- The continuous relaxation of optimization problem (2.1)-(2.9) is then "strong" in the sense that only a few re-qualification decision variables are not binary.

Algorithm 4 Greedy heuristic with dual prices

Input data: q

procedure GREEDY HEURISTIC WITH DUAL PRICES
  Best Plan $\leftarrow \emptyset$
  $f^* \leftarrow \infty$
  for $i = 1$ to k **do**
    Solve optimization model (2.13)-(2.18) with Algorithm 1
    L $\leftarrow$ Dual variables of Constraints (2.17)
    L' $\leftarrow$ Qualifications corresponding to the N_{dual} smallest dual variables in L
    for each $(r, m) \in$ L' **do**
      $\boldsymbol{q^{temp}} \leftarrow \boldsymbol{q}$
      $q^{temp}_{r,m} \leftarrow 1$
      $f^{temp} \leftarrow f_1(\arg\min(2.13) - (2.18)$ with $\boldsymbol{q^{temp}}, k = 0)$
      if $f^{temp} < f*$ **then**
        $(r*, m*) \leftarrow (r, m)$
        $f* \leftarrow f^{temp}$
      end if
    end for
    Update qualification matrix, $q_{r*,m*} \leftarrow 1$
    Best Plan $\leftarrow$ Best Plan $\cup$ $(r*, m*)$
  end for
  return Best Plan
end procedure

Algorithm 5 Instantaneous Greedy Heuristic (IGH)

procedure GREEDY HEURISTIC WITH DUAL PRICES
  Best Plan $\leftarrow \emptyset$
  Solve optimization problem (2.13)-(2.18) with Algorithm 1
  L $\leftarrow$ Dual variables of Constraints (2.17)
  L' $\leftarrow$ Qualifications corresponding to the N_{dual} smallest dual variables in L
  for each $(r, m) \in$ L' **do**
    Best Plan $\leftarrow$ Best Plan $\cup$ (r, m)
  end for
  return Best Plan
end procedure

– An initial feasible solution can be quickly determined by using IGH that is presented in Section 2.3.3 and Algorithm 5.

Based on these observations, a branch and bound approach can potentially be an efficient solution approach. If the continuous relaxation is strong, the number of re-qualifications is small and the initial solution is good, then we can prune nodes and find an optimal solution very quickly, even if the optimization problem is nonlinear. In the branch and bound solution approach, a best first approach is explored. Branching is performed on the re-qualification decision variable $OQ_{r,m}$ that is the closest to one. Bounding is performed by solving the continuous relaxation of the optimization model (2.1)-(2.9). A priority queue on the smallest lower bound is implemented to explore the tree. Finally, as explained in the hypothesis, a feasible solution can be quickly generated by using the dual variables. The optimization model (2.13)-(2.18) is solved and the k smallest dual variables are used to identify the most promising re-qualification decisions, leading to an initial feasible re-qualification plan. The pseudo code of the branch and bound algorithm is provided in Algorithm 6.

2.4 Computational study

In this section, the solution approaches presented in Section 2.3 are compared on industrial instances. The objective is to determine the most suited solution approaches by work center given the required small computational time (a few minutes at most).

2.4.1 Instance characterization

The computational study is performed by using historical data extracted from a manufacturing facility with a large variety of operations and which is located in Crolles, France. The factory is characterized by shifting bottleneck work centers, frequent product mix changes, high production variability, frequent disqualifications and large machine utilization rates.

In Tables 2.2, 2.3, and 2.4, the industrial instances used for the computational study are described. To preserve confidentiality, industrial instances are not given as is because they may contain critical information of the factory. Instead, coefficients of variability are used to present the industrial data. However, authors interested in the data can contact one of the authors to obtain more information.

Table 2.2 shows the number of operations R and machines M for the two work centers used in the computational study. The Coefficient of Variability (CV), defined as the standard deviation over the mean of a data set of the throughput rate can be found in Table 2.2. The “Operation TH coefficient of variability” corresponds to the coefficient of variability of the throughput rate for a given operation over all initially qualified and qualifiable machines. The machine “TH coefficient of variability” corresponds to the variability of the throughput rate for a given machine over all initially qualified and qualifiable operations. The minimum coefficient of

Algorithm 6 Branch and bound algorithm

procedure BRANCH AND BOUND ALGORITHM
  Best Plan ← Call IGH, Algorithm 5 (initialization step)
  $UB \leftarrow f_1$(Best Plan)
  $\mathcal{Q} \leftarrow \emptyset$
  $\boldsymbol{OQ} \leftarrow \arg\min$ (2.1)-(2.9) when relaxing binary constraints
  $LB \leftarrow f_1(\boldsymbol{OQ})$
  $\mathcal{Q} \leftarrow \mathcal{Q} \cup (\boldsymbol{OQ}, LB)$
  while $\mathcal{Q} \neq \emptyset$ or $\frac{UB-LB}{LB} > \epsilon$ **do**
    Take a node N $(\boldsymbol{OQ}*, f*)$ off $\mathcal{Q}$
    if $\boldsymbol{OQ}*$ binary and $f* \leq UB$ **then**
      Best Plan ← Re-qualifications from $\boldsymbol{OQ}*$
      $UB \leftarrow f*$
    end if
    if $\boldsymbol{OQ}*$ non binary **then**
      Let $(r*, m*)$ be the largest non binary variable in $\boldsymbol{OQ}*$
      $\boldsymbol{OQ}_0 \leftarrow \arg\min$(2.1) – (2.9) when relaxing binary constraints and $OQ_{r*,m*} = 0$
      $\boldsymbol{OQ}_1 \leftarrow \arg\min$(2.1) – (2.9) when relaxing binary constraints and $OQ_{r*,m*} = 1$
      if $f_1(\boldsymbol{OQ}_0) \geq UB$ **then**
        Prune node
      else
        $\mathcal{Q} \leftarrow \mathcal{Q} \cup (\boldsymbol{OQ}_0, f_1(\boldsymbol{OQ}_0))$
      end if
      if $f_1(\boldsymbol{OQ}_1) \geq UB$ **then**
        Prune node
      else
        $\mathcal{Q} \leftarrow \mathcal{Q} \cup (\boldsymbol{OQ}_1, f_1(\boldsymbol{OQ}_1))$
      end if
    end if
  end while
  return Best Plan
end procedure

variability is not presented because it is always equal to zero because there is always an operation that is qualified on only one specific machine.

Table 2.3 presents the data on the matrix density. The qualifiable density is presented, it is the number of entries that are equal to two in the qualification matrix over $R \times M$. Similarly, the qualified density is also presented, it is the number of entries that are equal to 1 in the qualification matrix over $R \times M$. The overall density is the sum of the qualifiable and qualified densities. Finally, "Operation CV density" and "Machine CV density" are described. The Operation CV density is the coefficient of variability of the number of qualified and qualifiable machines by operation. Similarly, the Machine CV density is the coefficient of variability of the number of qualified and qualifiable operations by machine.

In total, 24 instances are used by work center to compare solution approaches, and the production quantities for one day in each work center are used. Both work centers are characterized by a very large number of operations. Work center A has a limited number of machines, and work center B has a large number of machines. However, the number of operations that can be run by a machine in work center A is significantly larger than a machine in work center B. For each work center, the qualification matrix is sparse, in particular very sparse for work center B. For work center A, the qualifiable density in work center A varies between 1.8% and 3.6%, and the qualified density varies between 18.6% and 21.6%. The operation CV density is approximately equal to 0.45 and the machine CV density is approximately equal to 0.6. For work center B, the qualifiable density in work center B varies between 0.720% and 0.825%, and the qualified density varies between 2.3% and 2.5%. The operation CV density is approximately equal to 0.63 and the machine CV density is approximately equal to 3. This shows a large variability in the number of qualified and qualifiable machines by operation, and a large variability in the number of qualified and qualifiable operations by machine for both work centers.

Note that, although the qualifiable density in work center B is smaller than the qualifiable density in work center A, the number of qualifiable operations can be larger because the numbers of operations and machines are larger. Finally, in both work centers, machines are unrelated. The operation TH mean coefficient of variability is twice larger than the operation TH mean coefficient of variability in work center B. However, note that, the operation TH maximum coefficient of variability in work center B can be larger than the operation TH maximum coefficient of variability in work center A. The machine TH mean coefficient of variablity is approximately 33% larger in work center A than in work center B. In work center A, the mean machine TH coefficient of variability is approximately equal to 0.5. In work center B, it is approximately equal to 0.3. This shows that the throughput rates, given a machine and from one operation to another, or given a operation and from one machine to another, are very variable.

Finally, the coefficient of variability of the demand by operation is relatively constant from one instance to another for each work center. The coefficient of variability of the demand is approximately equal to 1.2 for work center A and equal to 1.4 for work center B.

Table 2.4 shows the coefficient of variability of the production capacity and the maximum production capacity divided by the minimum production capacity

by work center. The coefficient of variability of the production capacity is equal to 0.05 for work center A and to 0.12 for work center B. The fraction between the maximum production capacity and the minimum production capacity is equal to 1.40 for work center A and to 1.39 for work center B. Machines have different production capacities because they do not necessarily process the same operations. In addition, machines are of different ages since they were acquired at different times in the life of the factory.

The values of R, M, coefficients of variability and densities, and in particular the difference between the values from one instance to another, highlight the characteristics of the manufacturing facility: Frequent product mix changes, high production variability, frequent disqualifications.

Table 2.2: Description of industrial instances (1/3).

	Work center A						Work center B					
Instance	R	M	Operation TH Mean CV	Operation TH Max CV	Machine TH Mean CV	Machine TH Max CV	R	M	Operation TH Mean CV	Operation TH Max CV	Machine TH Mean CV	Machine TH Max CV
1	668	18	0.11	0.86	0.46	0.69	807	191	0.06	0.99	0.28	0.65
2	658	18	0.11	0.78	0.46	0.70	781	191	0.06	0.99	0.29	0.65
3	674	18	0.11	0.86	0.45	0.57	830	191	0.06	3.00	0.31	3.25
4	647	18	0.10	0.64	0.46	0.65	812	191	0.06	1.59	0.30	3.09
5	550	18	0.10	1.19	0.46	0.93	795	193	0.05	1.97	0.35	3.87
6	539	18	0.10	1.37	0.48	1.09	794	193	0.07	1.97	0.35	3.85
7	533	18	0.10	1.16	0.45	0.90	807	192	0.07	1.97	0.34	3.78
8	543	18	0.10	0.64	0.43	0.54	837	191	0.04	1.33	0.33	3.85
9	570	18	0.10	0.64	0.43	0.57	854	191	0.05	1.33	0.34	3.95
10	566	18	0.10	0.64	0.46	1.03	844	191	0.04	1.33	0.34	3.78
11	565	18	0.11	0.64	0.43	0.57	809	192	0.05	1.33	0.33	3.57
12	586	18	0.11	0.64	0.44	0.60	822	192	0.06	1.33	0.34	3.58
13	579	18	0.10	0.72	0.43	0.52	830	192	0.05	0.99	0.29	0.67
14	604	18	0.11	0.64	0.43	0.54	813	192	0.05	0.99	0.39	5.24
15	592	18	0.10	0.64	0.44	0.55	825	193	0.05	0.99	0.29	0.66
16	586	18	0.11	1.08	0.48	0.95	816	193	0.06	0.99	0.28	0.66
17	636	18	0.11	1.08	0.47	0.90	822	193	0.05	0.99	0.28	0.65
18	633	18	0.11	1.08	0.47	0.89	822	193	0.06	0.99	0.28	0.66
19	606	18	0.10	0.64	0.46	0.65	818	193	0.06	0.99	0.28	0.61
20	583	18	0.10	0.66	0.46	0.67	794	191	0.06	0.99	0.28	0.66
21	561	18	0.10	0.64	0.46	0.63	776	193	0.05	0.99	0.29	0.63
22	567	18	0.10	0.64	0.45	0.59	773	193	0.06	1.97	0.31	3.75
23	589	18	0.11	0.64	0.45	0.63	793	193	0.06	0.99	0.28	0.61
24	602	18	0.10	0.66	0.44	0.57	780	193	0.06	0.99	0.29	0.60

Table 2.3: Description of industrial instances (2/3).

	Work center A					Work center B				
Instance	Qualifiable density	Qualified density	Overall density	Operation CV density	Machine CV density	Qualifiable density	Qualified density	Overall density	Operation CV density	Machine CV density
1	2.54	18.73	21.27	0.43	0.70	0.72	2.53	3.25	0.63	1.26
2	2.52	18.64	21.16	0.43	0.70	0.73	2.54	3.26	0.64	1.23
3	1.79	19.35	21.14	0.43	0.63	0.76	2.56	3.32	0.65	1.26
4	1.57	19.24	20.81	0.43	0.66	0.75	2.54	3.29	0.64	1.28
5	3.38	20.06	23.44	0.47	0.56	0.77	2.39	3.15	0.65	1.18
6	3.33	20.23	23.56	0.46	0.55	0.75	2.42	3.16	0.65	1.19
7	3.24	20.34	23.58	0.44	0.56	0.78	2.40	3.19	0.64	1.17
8	2.40	21.56	23.96	0.46	0.55	0.81	2.38	3.19	0.66	1.19
9	3.26	19.69	22.94	0.46	0.56	0.79	2.40	3.18	0.65	1.20
10	3.04	20.48	23.53	0.44	0.58	0.87	2.31	3.17	0.65	1.19
11	3.09	20.54	23.63	0.46	0.57	0.82	2.40	3.22	0.62	1.16
12	3.16	20.39	23.55	0.46	0.55	0.82	2.37	3.20	0.63	1.16
13	2.50	21.07	23.58	0.46	0.53	0.88	2.29	3.17	0.65	1.22
14	2.01	21.47	23.48	0.46	0.58	0.82	2.35	3.17	0.65	1.27
15	2.06	21.10	23.16	0.46	0.58	0.80	2.35	3.15	0.65	1.21
16	2.28	20.76	23.04	0.45	0.60	0.78	2.38	3.17	0.65	1.22
17	2.38	20.78	23.17	0.47	0.58	0.79	2.37	3.17	0.64	1.23
18	2.55	20.95	23.50	0.47	0.57	0.81	2.33	3.15	0.65	1.22
19	3.11	20.32	23.42	0.47	0.55	0.80	2.33	3.12	0.64	1.27
20	3.15	19.90	23.05	0.48	0.56	0.80	2.39	3.19	0.64	1.29
21	3.45	20.12	23.57	0.47	0.54	0.83	2.39	3.22	0.62	1.18
22	3.47	20.07	23.54	0.47	0.55	0.82	2.40	3.22	0.61	1.25
23	3.04	19.85	22.88	0.48	0.56	0.86	2.39	3.25	0.64	1.16
24	2.83	19.98	22.81	0.48	0.58	0.81	2.37	3.18	0.62	1.23

Table 2.4: Description of industrial instances (3/3).

Work center	CV	Max / min
A	0.05	1.40
B	0.12	1.39

2.4.2 Design of experiments

In the computational study, the horizon is 24 hours. Solution approaches are compared for a number of re-qualifications $k \in \{1, 2, 3, 4, 5, 6, 7, 8, 40, 100\}$. We study all values between 1 and 8 because, in most cases, it is unnecessary to make a larger number of re-qualifications to significantly improve the utilization balance of the machines. In other words, the three best re-qualifications lead to better increase on the utilization balance of the machines than the following three best re-qualifications, even if the utilization balance of the machines still improves. In addition, in practice, only a limited number of re-qualifications is usually allowed on 24 hours. Larger values of k, *i.e.* 40 and 100, are studied to evaluate the performances of solution approaches in a limited computational time.

Solution approaches for the two different work centers presented in Tables 2.2 and 2.3. Two maximum computational times are considered: 30 seconds and 180 seconds (3 minutes). In addition, two initial qualification configurations are studied:

> *First qualification configuration.* It consists in taking the industrial qualification matrix as is to test our approaches for real-life qualification configurations.
>
> *Second qualification configuration.* We are also interested in testing our approaches for more extreme cases. This configuration consists in making qualifiable the qualifications that are not initially qualifiable (*i.e.* when $Q_{r,m} = 0$). For each machine, the associated throughput for these cases is set to the mean throughput over other initially qualified and qualifiable machines. The density of the qualification matrix is then close to 100%. This case can be useful to identify whether or not qualifying an operation that has never been qualified is actually beneficial for the utilization balance of the machines in the work center. In addition, this case can also be used to identify the limit of each algorithm.

The tested solution approaches are presented in Table 2.5. They are summarized by their name and whether dual prices are used. In total, six different solution approaches are compared to generate a re-qualification plan for short-term qualification management. For the sake of presentation, short names are given to the solution approaches (see Table 2.5) to present the numerical results in Section 2.4.3.

2.4.3 Numerical results

Two metrics are presented by instance to compare solution approaches: The relative gain (%) on the utilization balance of the machines with respect to the initial

Table 2.5: Solution approaches tested in the computational study.

Algorithm	Dual prices	Short name	Reference section
Greedy heuristic	Off	GH	2.3.1
Local search	Off	LS	2.3.2
Greedy heuristic	On	GHDP	2.3.3
Local search	On	LSDP	2.3.3
Instantaneous Greedy heuristic (branch and bound)	On	IGH	2.3.3
Branch and bound	-	B&B	2.3.4

qualification configuration and the computational time (in seconds). Numerical results are not detailed instance by instance to limit the length of the chapter. More precisely, the relative gain (%) is equal to $\frac{f1^{before}-f1^{after}}{f1^{before}} * 100$.

In the numerical experiments, γ is set to four. The outer linearization algorithm is stopped when a relative gap lower than 0.0001 is reached. N_{dual} is set to 8 for all algorithms. Each iteration of the outer linearization algorithm is solved by CLP, which is an open source solver (Lougee-Heimer, 2003; Löhndorf, 2016). Dual variables are then computed with CLP when the outer linearization algorithm is stopped. All solution approaches are implemented in Java 8 on a computer with an Intel(R) Xeon(R) CPU E3-1240 v5 @3.50GHz with 4 cores and 32 Go of RAM. Note that all solution approaches are parallelized, including the branch and bound algorithm. The maximum number of re-qualification plans that are simultaneously evaluated is equal to the number of logical threads, *e.g.* 8 logical threads on the used computer. For instance, 8 re-qualification plans are tested in parallel in the greedy heuristic of Section 2.3.1. When a solution approach is running, the current computational time is compared to the maximum computational time every second. If a solution approach is running when the maximum computational time is reached, running and waiting threads are terminated and the solution approach is stopped. Only finished threads that are not interrupted are considered to improve the objective function. Finally, in B&B, we set an optimality gap, *i.e.* $\frac{UB-LB}{LB}$, of 0.0001. If B&B is runnning but the gap is lower than 0.0001, then B&B is stopped and the best solution found so far is considered as numerically optimal.

2.4.3.1 Work center A

Numerical results for work center A are details in Appendix A.2, Section A for space limitations. Results are summarized here.

2.4.3.1.1 First qualification configuration

Numerical results show that B&B outperforms all other solution approaches both in terms of solution quality and computational time, and should be run for this work center and the first qualification configuration. This also shows that using empirical observations and dual variables, which are part of the B&B solution approach, is relevant for this work center. Numerical results show that including dual variables to guide solution approaches is relevant, otherwise the search space at each iteration

of GH and LS is too large for short computational time limits. Finally, we can observe that GHDP is often close to the optimal solution and can challenge B&B when the computational time limit is 180 seconds.

2.4.3.1.2 Second qualification configuration

Numerical results show that LSDP is the best option for work center A because it outperforms all other heuristics, even GHDP. Another interesting conclusion that can be drawn from these numerical experiments is that the gain between the first and second qualification configurations are very different. Consider $k = 1$ where the optimal solution is found for all instances by B&B. For the first qualification configuration, the mean gain is equal to 2.7% whereas it is equal to 15.4% for the second qualification configuration. The difference is significant. This shows that machines that cannot be qualified for some operations, *i.e.* such that $q_{r,m} = 0$ in the first configuration, could potentially lead to substantial improvements for the work center in terms of utilization balance of the machines. This may be worth to investigate, and to check if these forbidden qualifications could actually be made, *i.e.* whether the associated $q_{r,m} = 0$ in the first configuration could be changed to $q_{r,m} = 2$.

2.4.3.2 Work center B

The numerical results for work center B can be found in Tables 2.6 through 2.11. A first general observation is that the numerical results for the first qualification configuration for work center B behave similarly than the numerical results for the second qualification configuration for work center A.

2.4.3.2.1 First qualification configuration

Only GHDP and LSDP determine satisfactory re-qualification plans that scale with the number of re-qualifications. For $k = 1$, the mean gain with GHDP is equal to 15.8% and increases to 27.7% for $k = 8$. For larger values of k and a computational time limit of 30 seconds, GHDP does not determine better re-qualification plans because it reaches the computational time limit. Similarly to work center A for the first qualification configuration, LSDP leads to a modest improvement of the utilization balance of the machines. For instance, for $k = 3$ and 4, LSDP improves the utilization balance of the machines by 0.1% compared to GHDP although the computational time is twice as large.

For $k = 1$, the mean gain with IGH is equal to 15.1%, which is close to the mean gain with GHDP of 15.8%. However, as k increases, the difference between the mean gains of both solution approaches increases. Similarly to work center A, the difference can be explained by the fact that the dual prices only indicate a potential decrease of the utilization balance of the machines. Note that the optimization model (2.13)-(2.18) is also more computationally expensive to solve than for work center A. The mean computational time of the model is approximately 2.3 seconds, more than ten times longer than for work center A. This is mainly due to the fact that work center B has approximately ten times more machines than work center A.

Table 2.6: Mean gain (%) and CPU (s) over all instances for work center B for the first qualification configuration and a run time of 30 seconds by solution approach.

	GH		GHDP		LS		LSDP		IGH		B&B	
k	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)
1	15.4	33.5	15.8	2.6	15.4	33.3	15.8	5.0	15.1	2.3	15.9	7.6
2	15.4	33.1	20.8	5.0	15.4	33.1	20.8	10.0	17.6	2.3	20.9	19.2
3	15.4	33.5	23.0	7.3	15.4	33.5	23.1	15.8	18.9	2.3	23.2	26.7
4	15.4	33.2	24.6	9.7	15.4	33.5	24.7	21.2	19.9	2.4	24.3	30.3
5	15.4	33.2	25.6	12.2	15.4	33.3	25.8	26.8	20.4	2.3	24.1	30.8
6	15.4	33.2	26.5	14.6	15.4	33.4	26.7	30.1	20.9	2.5	21.9	30.8
7	15.4	33.3	27.2	17.4	15.4	33.3	27.3	31.2	21.6	2.3	21.6	30.8
8	15.4	33.3	27.7	19.6	15.4	33.1	27.7	31.4	21.9	2.4	21.9	30.8
40	15.4	33.1	28.7	31.3	15.4	33.4	28.7	31.3	25.9	2.5	25.9	30.8
100	15.4	33.2	28.7	31.4	15.4	33.7	28.7	31.3	28.3	2.6	28.3	30.8

Table 2.7: Mean gain (%) and CPU (s) over all instances for work center B for the first qualification configuration and a run time of 180 seconds by solution approach.

	GH		GHDP		LS		LSDP		IGH		B&B	
k	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)
1	15.9	182.0	15.8	2.7	15.9	187.3	15.8	4.9	15.1	2.3	15.9	7.5
2	16.5	186.7	20.8	5.0	16.5	189.7	20.8	10.1	17.6	2.3	20.9	35.5
3	16.5	186.4	23.0	7.5	16.6	188.3	23.1	15.5	18.9	2.4	23.2	81.1
4	16.5	185.9	24.6	9.8	16.6	188.4	24.7	20.9	19.9	2.3	24.8	128.8
5	16.6	185.8	25.6	12.1	16.6	188.5	25.8	27.2	20.4	2.3	25.6	164.7
6	16.6	187.3	26.5	14.5	16.6	186.5	26.7	32.5	20.9	2.4	25.4	170.0
7	16.6	187.4	27.2	17.3	16.6	186.9	27.3	39.8	21.6	2.3	26.0	179.3
8	16.6	185.8	27.7	19.6	16.6	188.9	27.8	47.6	21.9	2.5	25.8	180.9
40	16.6	188.9	29.5	97.8	16.6	188.0	29.5	181.2	25.9	2.5	25.9	180.8
100	16.6	187.0	29.6	181.4	16.6	186.2	29.6	181.3	28.3	2.7	28.3	180.8

In contrast with work center A for the first qualification configuration, GH and LS perform poorly. When the computational time limit is 30 seconds, GH cannot complete its first iteration before reaching 30 seconds. When the computational time limit is 180 seconds, GH can complete its first iteration for some instances but never completes its second iteration. LS slightly improves the utilization balance of the machines for some instances, at most by 0.1% on average. Similarly to work center A for the second qualification configuration, for $k = 1$, GH is able to determine qualification plans that are close in terms of quality to the re-qualification plans determined by GHDP. However, it is mostly by "chance" because good solutions are among the first ones tested.

Finally, similarly to work center A for the second qualification configuration, on average, B&B determines poor re-qualification plans. For $k = 1$, B&B is efficient and determines optimal solutions for all instances (see Table 2.8). Note that the continuous relaxation for work center B is weaker than for work center A. An optimal solution has been found for all instances whereas the mean final gap is of 4.86%. This means that the optimal solution has been found by pruning nodes with bounds. These numerical results suggest that work center B is less suitable for the branch and bound approach. Finally, note that, when the computational time limit is 30

Table 2.8: Details of the branch and bound solution approach for work center B and the first qualification configuration.

	30 seconds				180 seconds			
k	Initial Gap	Final Gap	Number Explored Nodes	Number optimal instances	Initial Gap	Final Gap	Number Explored Nodes	Number optimal instances
1	5.75%	4.86%	3.7	24	5.75%	4.86%	3.4	24
2	8.89%	3.43%	20.8	15	8.89%	3.43%	43.8	22
3	10.33%	3.44%	35.6	9	10.33%	3.40%	116.0	18
4	11.30%	3.63%	49.2	1	11.30%	3.06%	214.4	12
5	12.01%	4.90%	55.8	0	12.01%	3.07%	316.7	5
6	12.16%	10.17%	61.1	0	12.16%	3.88%	369.9	3
7	11.75%	11.75%	61.8	0	11.75%	3.80%	409.4	1
8	11.85%	11.85%	61.4	0	11.85%	4.31%	421.5	0
40	7.05%	7.05%	60.4	0	7.05%	7.05%	459.8	0
100	2.59%	2.59%	61.5	0	2.59%	2.59%	454.5	0

seconds, B&B does not find optimal solutions from $k = 5$, and the mean final gap is large, above 3%. When the computational time limit is 180 seconds, B&B does not find optimal solutions from $k = 8$.

Generally, the numerical results show that B&B is not suitable for work center B, in particular because of the very large number of machines. As for work center A, GHDP determines satisfactory re-qualification plans and is the most appropriate approach for work center B and the first qualification configuration. LSDP is at least as good as GHDP, but does not significantly improve the utilization balance of the machines and requires larger computational times.

2.4.3.2.2 Second qualification configuration

For $k = 1$ and contrary to the first qualification configuration, GH often determines *unsatisfactory* re-qualification plans for the second qualification configuration, whether the computational time limit is 30 or 180 seconds. The mean gain is equal to 0.8% for a computational time limit of 30 seconds, and only increases to 2.8% for a computational time limit of 180 seconds. The mean gain with GHDP is equal to 35.4% with a computational time of approximately 6 seconds. Such a difference is due to the significant combinatorial explosion associated to work center B. For instance, consider instance 1 of work center B. There are 807 operations and 191 machines (see Table 2.2). The initial qualifiable density is approximately equal to 2.53% (see Table 2.3). For the second qualification configuration, this means that the total number of qualifiable pairs (operation, machine) is equal to $\frac{100-2.53}{100} \times (807 \times 191) = 150,238$.

As mentioned for work center A, GHDP and LSDP are "immunized" against the combinatorial explosion because the number of re-qualifications that are tested from one iteration to another is constant and equal to N_{dual}. Moreover, these re-qualifications are relevant to improve the utilization balance of the machines.

Similarly to GH, B&B performs poorly. When the computational time limit is 30

seconds, the mean number of explored nodes is approximately equal to 1. Moreover, only one optimal solution is determined for $k = 1$. In contrast with the first qualification configuration, the number of explored nodes is very small but only one optimal solution is determined. Several reasons explain this. First, the continuous relaxation of the optimization model (2.1)-(2.9) is actually more computationally difficult to solve than the optimization model (2.13)-(2.18) that is used to evaluate a qualification plan, for instance in GH. In the continuous relaxation of the optimization model (2.1)-(2.9), there are *twice* as many decision variables as in the optimization model (2.13)-(2.18). There are also $2\times R\times M$ more constraints (due to the bound constraints on $OQ_{r,m}$). Similar observations can be made for work center A, but the practical impact of the resolution of the relaxed programs is lesser.

Contrary to the first qualification configuration, LSDP is able to improve the initial re-qualification plan determined by GHDP. For instance, when $k = 3$, the mean gain with GHDP is equal to 50.8%, whereas the mean gain with LSDP is equal to 51.8%. When $k = 4$, the mean gain with GHDP is equal to 55.7%, whereas the mean gain with LSDP is equal to 56.5%. However, the increase of the utilization balance of the machines impacts the computational times, as the mean computational time of LSDP is approximately equal to three times the mean computational time of GHDP.

As for the second qualification configuration for work center A and the first qualification configuration for work center B, GHDP, and possibly LSDP, seems to be most relevant approach to tackle the studied optimization problem on very large scale industrial instances, even for a small computational budget.

Table 2.9: Mean gain (%) and CPU (s) over all instances for work center B for the second qualification configuration and a run time of 30 seconds by solution approach.

	GH		GHDP		LSDP		IGH		B&B	
k	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)
1	0.8	33.2	35.3	3.3	35.3	6.2	32.3	2.7	35.7	32.7
2	-	-	44.5	6.1	44.8	12.7	34.6	2.7	38.3	34.6
3	-	-	50.8	9.2	51.8	20.5	35.2	2.7	35.2	34.4
4	-	-	55.7	12.2	56.5	27.1	35.3	2.7	35.3	33.6
5	-	-	59.5	15.1	61.2	31.3	35.3	2.7	35.3	33.3
6	-	-	63.4	18.3	64.3	31.5	35.3	2.7	35.3	32.8
7	-	-	65.8	21.1	66.7	31.5	35.3	2.7	35.3	33.2
8	-	-	68.3	24.1	69.0	31.7	35.3	2.7	35.3	33.5
40	-	-	73.0	31.7	73.3	31.8	35.9	2.7	35.9	35.3
100	-	-	73.5	31.9	73.5	31.5	37.1	2.9	37.1	35.1

2.5 Recommendations from the computational study

Numerical results in Sections 2.4.3.1 and 2.4.3.2 show that all algorithms do not perform equally. Generally, GH and LS are irrelevant because GHDP and LSDP determine re-qualification plans of similar quality in smaller computational times. However, depending on the work center, the qualification configuration and the

Table 2.10: Mean gain (%) and CPU (s) over all instances for work center B for the second qualification configuration and a run time of 180 seconds by solution approach.

	GH		GHDP		LSDP		IGH		B&B	
k	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)
1	2.8	189.8	35.3	3.4	35.3	6.1	32.3	2.7	36.0	69.2
2	-	-	44.5	6.5	44.8	12.4	34.6	2.7	46.5	188.2
3	-	-	50.8	9.5	51.8	20.4	35.2	2.7	53.4	190.4
4	-	-	55.7	12.0	56.5	28.0	35.3	2.7	56.3	192.6
5	-	-	59.5	15.3	61.5	39.3	35.3	2.7	35.3	191.1
6	-	-	63.4	18.1	64.5	47.3	35.3	2.7	35.3	193.4
7	-	-	65.8	20.9	67.0	61.5	35.3	2.7	35.3	196.9
8	-	-	68.3	23.9	69.3	64.4	35.3	2.7	35.3	198.6
40	-	-	88.1	120.9	88.7	182.0	35.9	2.7	35.9	208.2
100	-	-	90.2	181.2	90.2	181.6	37.1	2.9	37.1	198.1

Table 2.11: Details of the branch and bound solution approach for work center B and the second qualification configuration.

	30 seconds				180 seconds			
k	Initial Gap	Final Gap	Number Explored Nodes	Number optimal instances	Initial Gap	Final Gap	Number Explored Nodes	Number optimal instances
1	10.95%	3.91%	1.8	1	10.95%	3.44%	4.8	24
2	30.71%	21.73%	1.9	0	30.71%	5.37%	24.1	0
3	51.04%	51.04%	1.6	0	51.04%	6.21%	22.0	0
4	71.36%	71.36%	1.4	0	71.36%	11.53%	20.6	0
5	91.01%	91.01%	1.2	0	91.01%	91.01%	18.2	0
6	111.19%	111.19%	1.0	0	111.19%	111.19%	17.0	0
7	131.72%	131.72%	1.0	0	131.72%	131.72%	16.3	0
8	152.43%	152.43%	1.0	0	152.43%	152.43%	15.5	0
40	696.11%	696.11%	1.0	0	696.11%	696.11%	13.0	0
100	890.51%	890.51%	1.5	0	890.51%	890.51%	17.1	0

computational budget, the other solution approaches are valuable to a certain extent.

For a very small computational budget, instantaneous or of a few seconds, allowed in the Decision Support System, IGH is the most suitable approach, in particular for $k > 1$, because the computational time is independent of k, no matter the work center and the qualification configuration. However, a re-qualification plan determined by IGH may be of poor quality compared to GHDP, because one machine could inappropriately be overqualified at the expense of other machines. Therefore, a re-qualification plan may need manual rework by production personnel in the Decision Support System.

For work center A, and more generally, for work centers with a small number of machines, *e.g.* $M = 20$, and for the first qualification configuration, B&B is a good approach. However, although B&B performs slightly better than GHDP on average on our test instances, it is possible that B&B fails on other instances in terms of

worst-case performances (*e.g.*, see Table A.6) and determines poor re-qualification plans.

Generally, as GHDP determines satisfactory re-qualification plans that are very close to the optimal solutions for the first qualification configuration and very good qualification plans for the second qualification configuration, using GHDP is the best policy for the optimization model for any work center. GHDP surpasses IGH in terms of solution quality because it iteratively reconsiders better re-qualifications. LDSP can be selected if production personnel accepts a larger computational time, which can be conceivable for large work centers as work center B, or for the second qualification configuration where the difference in terms of relative gain with GHDP can be appreciable.

Note that if many dual variables have the same value, or are very close, as in the second qualification configuration, the solution approaches that are based on dual variables lose quality if a restricted number of qualifications is tested at each iteration. However, numerical results on industrial data show that this loss is not substantial and does not seem to depend on the number of operations R and machines M. If the loss was significant, the number of re-qualifications tested at each iteration in GHDP and LSDP could be increased to overcome the loss of quality.

Finally, this study shows that, although an optimization problem can be NP-Hard, studying the nature of the data is primordial to design efficient solution approaches. For manufacturing facilities with a large operation variety, using dual variables to guide the solution approach is shown to be effective and efficient for two different types of work centers and qualification configurations.

2.6 Conclusions and perspectives

In this chapter, we propose new solution approaches to determine optimized re-qualification plans in work centers with non-identical parallel machines to maximize the utilization balance and minimize the total utilization rate of the machines. In particular, dual prices are used to derive heuristics that are quickly guided towards good solutions. The proposed approaches are compared on industrial data on two different work centers and two different qualification configurations. Recommendations are finally provided. The approaches are now embedded in a decision support system that determines and proposes effective re-qualification plans to production personnel twenty minutes before every shift (every 8 hours). The decision support is used to enhance their decision process and better manage work centers (see Chapter 7 for more details).

An extension of the utilization balancing optimization approach proposed in Section 2.2.1 is proposed in Chapter 3 to maximize the throughput. We also study the effect of re-qualifications and disqualifications on the throughput. In addition, in Chapter 4, we further study how re-qualification delays and time varying demands and production capacities affect the throughput.

We believe the following perspectives are worth investigating in the future (out of the scope of the thesis):

1. Some parameters might be subject to uncertainty, such as the operation quantities and the machines capacities, and designing robust qualification plans is an interesting research avenue. In Chapter 6, a robust optimization approach is proposed for tactical qualification management, which could be extended for operational qualification management.

2. Workload variables are continuous but, in practice, some machines run operation quantities by batches. Hence, the consideration of batching constraints could be explored as in (Rowshannahad and Dauzère-Pérès, 2013).

3. An outer linearization algorithm is used to solve nonlinear programs. Other algorithms, such as active-set methods or sequential quadratic methods (Rowshannahad et al., 2015), could be compared to the outer linearization algorithm to further reduce computational times.

4. Solution approaches could be compared on data from other factories to further validate the relevance of dual variable solution approaches.

5. It would be relevant to study the robustness of solution approaches, *e.g.* under which conditions using dual prices do not provide good solutions.

6. Additional branching and exploring strategies could be explored for the B&B solution approach.

7. It would be interesting to better understand the impact of different γ settings on solution quality and computing time.

Chapter 3

A single period bilevel optimization approach for throughput maximization

In Chapter 2, the question "How to determine the most relevant re-qualifications to improve operational efficiency?" is answered from a utilization balancing standpoint. Nevertheless, a utilization balancing approach may actually be limited to maximize the throughput on short term horizons for factories subject to high production variability. In this case, modeling approaches that are more suitable may be required. In addition, disqualifications are not considered in Chapter 2, whereas disqualifications can be a critical component in operational qualification management*.

*Large parts of this chapter has been submitted to an international journal.

3.1 Introduction

In general, when a lot arrives in a work center, if a machine is both idle and qualified for the operation required by the lot, the lot might be assigned to this machine even if it is very slow to process the operation compared to other qualified machines. This might be the right decision to temporarily and locally maximize the throughput with local perception. However, this assignment can be a poor decision for the overall throughput over a day when a wave (peak) of lots with a faster operation is expected, because the production capacity of each machine is *finite*. Such assignments are *frequently* observed in manufacturing systems with many operations because dispatching engines work only with the current lots in a work center, or with a limited vision of the lots arriving in the work center. Dispatching decisions are *short-sighted* and can be a source of capacity loss. This means that two re-qualifications that better balance the utilization rates of the machines may not lead to the same gain on the throughput.

Assume that a small quantity of a very slow operation is expected to consume the production capacity of a machine that also should process a large quantity of a much faster different operation. The slow operation is thus disqualified on the machine, so that only the faster operation is processed and the slow operation is assigned to other machines. Disqualifications are also made when the throughput rate of an operation depends on the machine health (*i.e.*, the process quality of a machine). This is the case for ion implantation machines because the throughput rate of an operation depends on the wear of the ion source. The more significant the wear, the lower the throughput rate. If there exist other qualified machines for the operation, it can be preferable to disqualify the operation on a machine to process it on other qualified machines. Nevertheless, it is difficult to evaluate the timing of such decisions without models. More generally, disqualifications are frequently made when the throughput rates between two machines are very different for a given operation. In addition, disqualifications are used by production personnel to disqualify operation on machines that do not have an optimal yield. Therefore, disqualifications are used to better manage the production capacity of each machine and maximize the throughput.

Nevertheless, too many disqualifications can lead to poorly balanced machines in terms of utilization rates. This can also be a source of throughput loss even though initially disqualified machines are much slower than other initially qualified machines. In this case, production personnel must determine *re-qualifications* to rebalance the utilization rates of the machines in the work center. Re-qualifications are then also used to maximize throughput, anticipate future bottleneck (overloaded) machines and better use the production capacity of each machine. Production personnel must therefore find a balance between re-qualification and disqualification decisions to maximize the manufacturing performances of their work center, and in particular to maximize the throughput.

Determining relevant re-qualifications and disqualifications is crucial to maximize the throughput. However, this is complex because the effect of qualifications and disqualifications on the throughput depends, among other reasons, on the number of operations that must be processed, the quantity for each operation, the initial

set of qualifications and disqualifications, the machine states and *dispatching decisions* (Johnzén et al., 2008; Gurumurthi and Benjaafar, 2004; Kopp et al., 2019).

The remainder of Chapter 3 is organized as follows. A bilevel optimization approach for operational management of re-qualifications and disqualifications is motivated in Section 3.2. In particular, it is motivated with respect to the existing literature and to the utilization balancing approach presented in Chapter 2. Bilevel optimization models are proposed in Section 3.3. In Section 3.4, we show the benefit of using a bilevel optimization approach to optimize the throughput with re-qualifications *and* disqualifications. The bilevel optimization approach is notably compared to the utilization balancing optimization approach studied in Chapter 2. In Section 3.5, we formulate recommendations for production personnel. Finally, in Section 3.6, we conclude and give perspectives.

3.2 Motivation

3.2.1 Optimizing the throughput with a utilization balancing approach

The utilization balancing approach presented in Chapter 2 can be seen a surrogate mean to improve the throughput. For instance, the utilization balancing optimization model can be used to identify short-term bottleneck (overloaded) machines and therefore recommend qualifications to improve the utilization balance between the machines. If machine utilization rates are better balanced, then it is reasonable to assume that the throughput will improve, in particular if we are able to reassign some of the workload of a bottleneck machine to a under loaded machine.

However, the utilization balancing approach does not *necessarily* ensure that the throughput is optimized over a finite planning horizon, a major performance indicator on which production personnel are challenged and that is ultimately used to complete orders on time. Consider the illustrative example in Tables 3.1 and 3.2 of a work center of four machines. The utilization balancing approach is used to determine the utilization rate of each machine when no new re-qualification decision is performed. Machine 1 has a utilization rate of 3.0, machine 2 of 1.2, machine 3 of 1.2 and machine 4 of 0.8. When the utilization rate of a machine is larger than 1, then there is a backlog at the end of the horizon and the machine is bottleneck (critical).

First, assume that it is possible to make a re-qualification decision to optimize the utilization balancing objective function $\sum_m U_m^\gamma$ of Chapter 2 by moving some workload of machine 1 to machine 2 (see Table 3.1), $\gamma > 1$. After one re-qualification decision, the utilization balancing objective function decreases from 83.1 to 38.9, which is significant. Nevertheless, because the total utilization rate remains *unchanged*, it is unclear to see why the throughput should increase, in particular because both machines have a utilization rate larger than 1 before and after the re-qualification.

Then, assume that it is possible to make a re-qualification decision to optimize the utilization balancing objective function $\sum_m U_m^\gamma$ by transferring some workload

of machine 3 to machine 4 (see Table 3.2). After one re-qualification, the utilization balancing objective function decreases from 2.5 to 2, which is a significantly smaller reduction with respect to the re-qualification that balances the utilization rates of machines 1 and 2.

Therefore, if we had to choose the best re-qualification in terms of utilization balance and total utilization rate, the first re-qualification, *i.e.* the re-qualification that balances the utilization rates of machines 1 and 2, would be chosen. Nevertheless, this is probably not the relevant re-qualification in terms of throughput. This is because the second re-qualification, *i.e.* the re-qualification that balances the utilization rates of machines 3 and 4, allows all backlog quantities by machine 3 to be processed by machine 4 before the end of the horizon even tough the total utilization rate remains unchanged. This is because the initial utilization rate of machine 4 is equal to 0.8. As a result, the throughput is expected to increase. In other words, the best re-qualification in terms of utilization balance is not necessarily the best re-qualification in terms of throughput. For a given horizon, balancing the utilization rate of two overloaded machines may not lead to a throughput gain, in particular if the total utilization rate of both machines is not reduced.

Note that, if the throughput cannot be improved by optimizing the utilization balancing objective function, then it is possible to focus on some re-qualification decisions to still keep optimizing the utilization balance while ensuring that the throughput is improved. For instance, re-qualification decisions could be restricted underloaded machines, *i.e.* the machines that have still some production capacity left by the end of the horizon, or to overloaded machines that are faster for some operations than currently qualified overloaded machines to minimize the total utilization rate. However, this is no longer a mathematical model but a solution approach. The underlying objective that consists in optimizing the throughput is not modeled. A more relevant mathematical model can help put the problem faced by production personnel into perspective to better support decision making in terms of throughput.

Table 3.1: An illustrative example why optimizing the utilization balance may not be equivalent to throughput maximization ($\gamma = 4$) (1/2).

	Current utilization rate	Utilization rate after one re-qualification
Machine 1	3	2.1
Machine 2	1.2	2.1
Total utilization rate $\sum_m U_m$	4.2	4.2
$\sum_m U_m^\gamma$	83.1	38.9

In addition, a utilization balancing approach does not consider dispatching rules whereas they strongly affect the relevance of re-qualification decisions (see Section 3.2.2 and Johnzén et al. 2008; Gurumurthi and Benjaafar 2004; Kopp et al. 2019).

Table 3.2: An illustrative example why optimizing the utilization balance may not be to throughput maximization ($\gamma = 4$) (2/2).

	Current utilization rate	Utilization rate after one re-qualification
Machine 3	1.2	1
Machine 4	0.8	1
Total utilization rate $\sum_m U_m$	2	2
$\sum_m U_m^\gamma$	2.5	2

3.2.2 Short-sighted aspect of dispatching engines

To maximize the throughput with qualification management, a simple optimization model could be used and would consist in directly maximizing *the number of wafers* processed by the end of the planning horizon. This model is similar to the model proposed by Chang and Dong (2017):

$$\max \quad \sum_{r,m} WIP_{r,m} \tag{3.1}$$

$$\text{s.t.} \quad \sum_{r} \frac{WIP_{r,m}}{tp_{r,m}} \leq c_m \qquad \forall m \tag{3.2}$$

$$\sum_{m} WIP_{r,m} \leq d_r \qquad \forall r \tag{3.3}$$

$$\sum_{r,m} OQ_{r,m} \leq k \tag{3.4}$$

$$WIP_{r,m} \leq d_r OQ_{r,m} \qquad \forall r, \forall m \mid q_{r,m} = 2 \tag{3.5}$$

$$WIP_{r,m} \leq d_r \qquad \forall r, \forall m \mid q_{r,m} = 1 \tag{3.6}$$

$$WIP_{r,m} \leq 0 \qquad \forall r, \forall m \mid q_{r,m} = 0 \tag{3.7}$$

$$WIP_{r,m} \geq 0 \tag{3.8}$$

$$OQ_{r,m} \in \{0,1\} \qquad \forall r, \forall m \tag{3.9}$$

The objective function (3.1) that consists in maximizing *the number of wafers* processed by the end of the planning horizon. Constraint (3.2) ensures that the production capacity of each machine in the work center is respected. Constraint (3.4) limits the size of the re-qualification plans to k re-qualifications. Constraint (3.3) ensures that the number of wafers processed by operation cannot exceed the total quantity. Constraint (3.5)-(3.7) ensures that wafers of operation r can only be assigned to machine m if the machine m is qualified for operation r. Finally, Constraint (3.8) is the non-negativity constraint and Constraint (3.9) is the binary constraint.

The optimization model (3.1)-(3.9) is very optimistic and always assigns wafers to the fastest qualified machine while satisfying capacity constraints to maximize the throughput. Presenting capacity allocation deduced from the optimization model to

production personnel may be unrealistic, because very often the same operations will be backlogged. This is similar to the justification for the use of a nonlinear function for utilization balancing. The objective function seems appropriate but solutions may not be acceptable in practice. From practical experience, given a machine, although an operation is much slower than the fastest operation, the slow operation will likely be run before faster operations, and this is even desirable because the slow operation may be linked to operations with strong due date commitments. Similarly, given an operation, the workload is not necessarily associated to the fastest machine. This is due to the short-sighted aspect of dispatching engines. The way operations *compete* at the expense of each other for machines with finite production capacity is not captured. High priority operations are likely to be run before low priority operations. Considering operation queues is probably necessary to make better re-qualification decisions. Johnzén et al. (2008) and Gurumurthi and Benjaafar (2004) actually show that re-qualifications do not necessarily improve the throughput because of the short-sighted aspect of dispatching engines. Similarly, Kopp et al. (2019) show that dispatching decisions strongly influence the relevance of additional qualifications.

A way to overcome the problem using optimization models such as (3.1)-(3.9) is to add weights to decision variables in the objective function. However, relevant weights, which could be seen as inventory backlog or holding costs, to make capacity allocation more realistic are complex to define, in particular on short horizons (a few hours to a few days), for several reasons. First, the fabrication process takes several months, therefore defining inventory holding costs on short horizons is less relevant. Inventory holding and backlog costs are often optimized at tactical or strategic decision levels (master planning) when product quantities to release in the factory are decided. In addition, the decision level for re-qualifications is actually close to the scheduling level, where the concepts of inventory and backlog do not really exist and are implicitly coded using priorities of lots.

3.2.3 Evaluating disqualification decisions

To the best of our knowledge, although disqualification management is a critical component of short-term decision making for production personnel, there is no contribution that deals with the use or modeling of disqualification decisions in semiconductor manufacturing. However, some approaches can be adapted to support disqualification decisions.

First, the linear optimization model proposed by Chang and Dong (2017), or similarly the linear optimization model proposed by Rowshannahad et al. (2015), cannot be be used to recommend *disqualifications* that *improve* the throughput. If there is at least one disqualification (the number of "1" in the qualification matrix decreases), the new value of the objective function cannot be larger than the value of the objective function with the initial set of qualifications. This is because, if the number of qualifications (number of entries that are equal to "1") is reduced with respect to the initial set of qualifications, then the solution set of the optimization models with the new set of qualifications is included in the solution set of the initial set of qualifications. Therefore, after performing disqualifications, the objective

function cannot strictly increase.

From a more general standpoint, it seems difficult or even impossible to define a *single-level* optimization approach to optimize the throughput with *disqualifications*, because disqualifications can only decrease the throughput. However, this is not what can be observed in practice by production personnel. This means, that in order to simultaneously consider the impact of qualifications *and* disqualifications, whether it is with a user-defined scenario in a DSS or in an optimization algorithm, qualification decisions and dispatching decisions must be separated, as in practice, in particular if we want to thoroughly model the impact of disqualification on the throughput. Note that separating qualification and dispatching decisions is relevant both for disqualification and qualification decisions.

A first option consists in using *simulation* approaches, which can determine disqualifications that improve the throughput. For instance, a discrete-event simulation can be used to simulate dispatching decisions where qualifications are input parameters, which can be defined by an optimization model. Simulation models are used by Fowler et al. (1997); Akcalt et al. (2001); Kabak et al. (2013); Ignizio (2009, 2010); Kopp et al. (2016, 2018, 2019). Simulation models are used to evaluate the effect of re-qualifications on indicators such as in terms of mean cycle time, throughput, mean tardiness and the number of performed re-qualification activities. Note that the short-sighted aspect of dispatching decisions is still present in simulation models because dispatching decisions do not consider upcoming peaks of the wafer quantities.

These contributions could be adapted to embed disqualification decisions, for instance by designing an iterative optimization-simulation approach. Such approach is popular for instance for production planning in semiconductor manufacturing (*e.g.* see Hung and Leachman 1996). Discrete-event simulation models are particularly interesting if we want to include special features of a work center, such as the management of masks in a lithography work center. However, if short computational times are required, running a discrete-event simulation can be undesirable for purposes other than defining strategies. It is particularly undesirable if a discrete-event simulation must be run for each scenario defined by production personnel in a DSS. Moreover, developing and *maintaining* an up-to-date discrete-event simulation model is time-consuming (Shanthikumar et al., 2007), in particular if we want to apply the approach to any work center and not just to a particular work center, as the lithography work center.

Another option consists in using a *bilevel* optimization approach. It is a suitable approach for production personnel because there is no need to run discrete-event simulations, and bilevel optimization models can be used to model a hierarchical decision making, and therefore separate the re-qualification and dispatching decisions (see *e.g.* Stackelberg 1952; Bracken and McGill 1973; Sinha et al. 2017). Note that bilevel optimization problems are more challenging to solve than "classical" optimization problems (Fischetti et al., 2017).

3.2.4 Concluding remarks and contributions

For short-term optimization, the utilization balancing problem without making any qualification decision presented in Chapter 2 can actually be seen as a scheduling problem on parallel machines with preemption and no release dates where the objective consists in maximizing the utilization balance and minimizing the total utilization rates of the machine. Although the utilization balancing problem can take into account the fact that a machine is idle or that an operation is assigned to a machine, it cannot fully *simulate* dispatching rules and the way operations *compete* for machines.

A bilevel optimization approach is able to consider the short-sighted aspect of dispatching engines. Based on a utilization balancing problem, a bilevel optimization approach is suitable to simulate dispatching rules and their impact on the throughput. A bilevel optimization approach can also be seen as a method to evaluate the quality of the utilization rates of the machines in terms of throughput. Finally, a bilevel optimization approach covers a broader set of decisions than the utilization balancing approaching to improve the throughput (see Figure 3.1) and the same set of decisions that are available to production personnel. Note that, the optimistic optimization model (3.10)-(3.20) cannot be used as the lower-level optimization problem in the bilevel optimization approach. This is because the optimistic optimization model already computes the throughput, and as the upper-level computes the throughput, both decision levels would cooperate and be unable to determine disqualifications that improve the throughput.

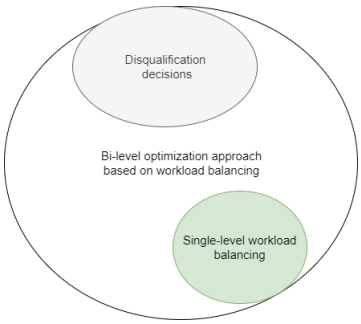


Figure 3.1: Set of decisions covered by the different optimization approaches.

Our contributions are summarized below:

- A bilevel optimization model is proposed to maximize the throughput with re-qualification decisions.
- A bilevel optimization model is proposed to cover the case where disqualification decisions must be made.
- A bilevel optimization model is proposed to combine re-qualification and disqualification decisions.
- A computational study is performed on industrial data from a 300 mm wafer fab located in France to validate bilevel optimization models.

3.3 Bilevel optimization models

To support short-term re-qualification and disqualification decisions that maximize the throughput, a bilevel optimization approach is proposed because it is suitable to model hierarchical decision making (Stackelberg 1952; Bracken and McGill 1973; Sinha et al. 2017). A bilevel optimization problem is a classical optimization problem, but in which at least one constraint is another optimization problem. The nested optimization problem is referred as the lower-level optimization problem. The outer optimization problem is referred as the upper optimization problem. The lower-level (also known as follower) optimization problem has its own objective function and constraints, that depend on decision variables of the upper-level (also known as leader) optimization problem (Sinha et al., 2017).

In this chapter, the bilevel optimization approach follows the separation of the production control and dispatching decision levels. The upper-level optimization problem is used to model the production control level. Re-qualification and disqualification decisions are therefore made in the upper-level optimization problem, where the objective criterion is the throughput, and used as arguments in the lower-level optimization problem. The lower-level optimization problem is modeled as a utilization balancing optimization problem as in Chapter 2, subject to re-qualification and disqualification decisions of the upper-level optimization problem, to build realistic queues of operations in front of machines by using empirical observations of dispatching engines. Such an approach is motivated by the fact that, from a general perspective, dispatching engines maximize the utilization of machines in order to maximize the throughput. Once queues of operations are defined by the lower-level optimization problem, the throughput is computed by the upper-level optimization problem.

In bilevel optimization, there exist two types of positions for the lower-level optimization problem: An optimistic position and a pessimistic position (Sinha et al., 2017). If multiple optimal solutions are available for the lower-level optimization problem, an optimistic solution is a solution that maximizes the upper-level objective function. A pessimistic position is a solution that minimizes the upper-level objective function. Our bilevel optimization formulation is neither pessimistic nor optimistic. In practice, dispatching or scheduling engines, even they are a source of production capacity loss, try to maximize the throughput by balancing the utilization rates of the machines. Extending the current formulation to optimistic (or even pessimistic) formulations is worth investigating but left for future research. It is also worth mentioning that there may exist multiple optimal solutions associated to the lower-level optimization problem. Some solutions may lead to a better throughput than others. Studying this is left for future research, *e.g.* by giving bounds on the throughput to production personnel.

A bilevel optimization problem where both the upper and lower optimization problems are linear is NP-Hard (see *e.g.*, Ben-Ayed and Blair 1990; Bard 1991). Bilevel optimization problems presented in Sections 3.3.3, 3.3.4 and 3.3.5 are therefore NP-Hard.

3.3.1 Problem statement

Consider a work center consisting of M non-identical, both in terms of throughput rates and re-qualifications, parallel machines. On a given time horizon, R different operations, each with a strictly positive quantity, must be processed by the machines. The throughput rate of each operation on each machine is known. However, each machine has a finite production capacity and can only process qualified operations. Similarly, an operation can only be processed by qualified machines. The qualification matrix, *i.e.* the operations that are initially qualified and that can be qualified for each machine, is known. We assume that, when a re-qualification decision is made, then the re-qualification is immediately active at the beginning of the horizon. We also assume that, when a disqualification decision is made, then it it made at the beginning of the horizon.

The objective is to determine k qualifiable pairs (operation r, machine m) to maximize the throughput.

3.3.2 Notations

In this section, the notations used in the models are presented.
Indices and sets:
m: Index for machines, $\in \{1,..,M\}$,
r: Index for operations, $\in \{1,..,R\}$.
Parameters:
k: Number of re-qualification (or disqualification) decisions to be made at the beginning of the planning horizon,
$q_{r,m} \in \{0,1,2\}$: Is equal to 1 if machine m is qualified for operation r, is equal to 2 if machine m is qualifiable for operation r, and is equal to 0 if machine m cannot be qualified for operation r,
$tp_{r,m}$: Throughput rate (in number of wafers by second) of operation r on machine m,
c_m: Initial availability time (in seconds) of machine m over the planning horizon,
d_r: Quantity of operation r to process,
γ: Utilization balancing parameter, strictly greater than 1.
Decision variables:
$OQ_{r,m} \in \{0,1\}$: Is equal to 1 if a re-qualification procedure is proposed for operation r on machine m at the beginning of the planning horizon, and 0 otherwise,
U_m: Utilization rate of machine m,
$WIP_{r,m}$: Quantity of operation r processed by machine m,
$\mathcal{R}^q(\mathbf{q},\mathbf{OQ}) = \{r \mid \sum_{m=1}^{M}(\mathbb{1}(q_{r,m}) + OQ_{r,m}) > 0\}$: Set of operations with at least one qualified machine with some capacity on the planning horizon, where $\mathbb{1}(q_{r,m}) = 1$ if $q_{r,m} = 1$, and 0 otherwise.

3.3.3 Bilevel optimization model for re-qualifications

Upper-level optimization model:

$$\max \qquad TH = f(\mathbf{U},\mathbf{WIP}) \tag{3.10}$$

$$\text{s.t.} \quad \sum_{r,m} OQ_{r,m} \leq k \tag{3.11}$$

$$\mathbf{U}, \mathbf{WIP} \in \arg\min LBP(\mathbf{OQ}) \tag{3.12}$$

$$OQ_{r,m} \in \{0,1\} \qquad \forall r, \forall m \tag{3.13}$$

Lower-level optimization model:

$$LBP(\mathbf{OQ}) = \min \quad \sum_{m} U_m^{\gamma} \tag{3.14}$$

$$\text{s.t.} \quad \sum_{m} WIP_{r,m} = d_r \qquad \forall r \in \mathcal{R}^q(\mathbf{q}, \mathbf{OQ}) \tag{3.15}$$

$$U_m = \sum_{r} \frac{WIP_{r,m}}{tp_{r,m} c_m} \qquad \forall m \tag{3.16}$$

$$WIP_{r,m} \leq d_r \qquad \forall r, \forall m \mid q_{r,m} = 1 \tag{3.17}$$

$$WIP_{r,m} \leq d_r OQ_{r,m} \qquad \forall r, \forall m \mid q_{r,m} = 2 \tag{3.18}$$

$$WIP_{r,m} \leq 0 \qquad \forall r, \forall m \mid q_{r,m} = 0 \tag{3.19}$$

$$WIP_{r,m} \geq 0 \qquad \forall r, \forall m \tag{3.20}$$

Upper-level optimization model. The objective function (3.10) aims at maximizing the throughput that is computed from the utilization rates of the machines in the work center. Constraints (3.11) set to k the maximum number of re-qualifications to be performed at the beginning of the planning horizon. Constraints (3.12) link the upper-level and lower-level problems. Constraints (3.13) are the binary constraints for the re-qualification decisions.

Lower-level optimization model. The lower-level optimization model corresponds to a utilization balancing optimization problem. Equation (3.14) is the objective function that consists in maximizing the utilization balance and minimizing the total utilization rates of the machine. Constraints (3.15) define the flow conservation on the planning horizon. Constraints (3.16) compute the utilization rate of each machine in the work center. Constraints (3.17)-(3.18) ensure that wafers of operation r can only be assigned to machine m if r is qualified on machine m. Constraints (3.19) ensure that, if operation r is not qualified and cannot be qualified on machine m, then operation quantities of operation r is never assigned to machine m. Constraints (3.20) are the non-negativity constraints for $WIP_{r,m}$.

As previously mentioned, there may exist multiple solutions to the lower-level optimization problem. In the thesis, we consider only the initial solution obtained when solving the lower-level optimization problem.

3.3.4 Bilevel optimization model for disqualifications

The proposed optimization model to make disqualification decisions differs from the optimization model to make re-qualification decisions presented in Section 3.3.3. A new binary decision variable, $DOQ_{r,m}$, is introduced, is equal to 1 if a disqualification procedure is proposed for operation r on machine m, and 0 otherwise. In addition, a new set $\mathcal{R}^d$ describes the set of operation that still have at least one

qualified machine with some capacity over the planning horizon after a disqualification. $\mathcal{R}^d$ is defined as follows: $\mathcal{R}^d(\mathbf{q}, \mathbf{DOQ}) = \{r \mid \sum_{m=1}^{M}(\mathbb{1}(q_{r,m}) - DOQ_{r,m}) > 0\}$. The bilevel optimization model is formalized below:

Upper-level optimization model:

$$\max \quad TH = f(\mathbf{U}, \mathbf{WIP}) \tag{3.21}$$

$$\text{s.t.} \quad \sum_{r,m} DOQ_{r,m} \leq k' \tag{3.22}$$

$$\mathbf{U}, \mathbf{WIP} \in \arg\min LBP(\mathbf{DOQ}) \tag{3.23}$$

$$DOQ_{r,m} \in \{0, 1\} \qquad \forall r, \forall m \tag{3.24}$$

Lower-level optimization model:

$$LBP(\mathbf{DOQ}) = \min \quad \sum_{m} U_m^{\gamma} \tag{3.25}$$

$$\text{s.t.} \quad \sum_{m} WIP_{r,m} = d_r \qquad \forall r \in \mathcal{R}^d(\mathbf{q}, \mathbf{DOQ}) \tag{3.26}$$

$$U_m = \sum_{r} \frac{WIP_{r,m}}{tp_{r,m} c_m} \qquad \forall m \tag{3.27}$$

$$WIP_{r,m} \leq d_r - d_r DOQ_{r,m} \qquad \forall r, \forall m \mid q_{r,m} = 1 \tag{3.28}$$

$$WIP_{r,m} \leq 0 \qquad \forall r, \forall m \mid q_{r,m} = 2 \tag{3.29}$$

$$WIP_{r,m} \leq 0 \qquad \forall r, \forall m \mid q_{r,m} = 0 \tag{3.30}$$

$$WIP_{r,m} \geq 0 \qquad \forall r, \forall m \tag{3.31}$$

Upper-level optimization model. The upper-level optimization model is similar to the one presented in Section 3.3.3. Only Constraint (3.22) that limits the number of disqualifications to k' changes.

Lower-level optimization model. Similarly, the lower-level optimization model is close to the one presented in Section 3.3.3. Only constraint (3.28) changes. It ensures that operation quantities of operation r cannot be assigned to machine m if m is disqualified at the beginning of the planning horizon. All other constraints are identical.

By setting $k = 0$ and by modifying the initial qualification matrix $q_{r,m}$ following a custom scenario defined by production personnel, it is possible to quickly simulate the impact of multiple disqualification decisions on the throughput, and thus improve manufacturing performances.

If the throughput rate $tp_{r,m}$ is identical for all operations and machines, then disqualifying operations cannot lead to gain on the throughput. This is because fast operations cannot be processed first because all operations are processed at the same rate. Similarly, to a certain extent, if $tp_{r,m}$ is not varying much from one machine to another and from one operation to another, then the gain on throughput with disqualifications may be limited.

Finally, when making one disqualification decision, it is meaningless to disqualify a pair (r,m) for which $WIP_{r,m}$ is initially equal to 0, *i.e.* when the bilevel optimization problem is solved for $k' = 0$. This is because the qualification pair (r,m) is not

used by the lower-level optimization problem and therefore has no effect on the utilization balance.

3.3.5 Combining re-qualifications and disqualifications

The bilevel optimization models for re-qualification and disqualification management can be combined to determine both re-qualifications and disqualifications. The resulting model is detailed in Appendix B, Section B.0.1.

3.3.6 Computation of the throughput

Although all wafers are assigned to machines in the work center with Equations (3.15), (3.26) and (B.8), this does not mean that the throughput is equal to the demand. This is because each machine has a finite production capacity that is shared between wafers. For machines that have a utilization rate lower than 1, *i.e.* for machines such that $U_m < 1$, all wafers assigned are processed by the end of the horizon. However, when machines have a utilization rate larger than 1, some wafers will not be processed by the end of the horizon and will be backlogged. Therefore, a choice must be made to differentiate processed wafers from backlogged wafers. In the thesis, processed wafers are distinguished from backlogged wafers by simulating dispatching rules. Let us consider the illustrative example in Table 3.3. Consider a work center with two machines. A total demand of 110 wafers is expected on the horizon, *e.g.* on the next 8 hours. Consider two configuration of qualifications. With the first configuration, the solution of the lower-level optimization model is a solution where machine 1 has a utilization rate of 1.2 and machine 2 has a utilization rate of 0.9. 60 wafers are allocated to machine 1 and 50 wafers are allocated to machine 2. Because machine 2 has a utilization rate lower than 1, it can process all allocated wafers by the end of the horizon. However, because machine 1 has a utilization rate greater than 1, it cannot process all allocated wafers. A cross-product can give an estimation of the throughput: $\frac{1}{1.2} \times 60 + 50 = 100$. With the second configuration, the solution of the lower-level optimization model is a solution where both machines have a utilization rate of 0.95. Therefore, both machines can process all allocated wafers by the end of the horizon. In this case, the throughput is equal to the demand and is therefore equal to 110.

Table 3.3: Illustrative example on the computation of the throughput in bilevel optimization models.

	First configuration		Second configuration	
	Machine 1	Machine 2	Machine 1	Machine 2
Utilization rate, U_m	1.2	0.9	0.95	0.95
Allocated wafers, $\sum_r WIP_{r,m}$	60	50	55	55

The computation of the throughput can be more complex because operations often have priorities. Two modes for computing the throughput $TH = f(\mathbf{U}, \mathbf{WIP})$ are used to simulate First-In First-Out (FIFO) and priority-based dispatching rules:

1. The first mode simulates a FIFO dispatching rule and is defined as "an average product mix": $f(\mathbf{U}, \mathbf{WIP}) = \sum_m \frac{1}{\max\{1, U_m\}} \sum_r WIP_{r,m}$. From an aggregate point of view, as each operation allocated to a machine is equally processed in proportion, it can be seen as reproducing a FIFO dispatching rule.

2. The second mode simulates a priority-based dispatching rule, where priority operations are processed first on machines. If two operations have the same priority, the one with the largest $WIP_{r,m}$ is processed first to avoid setups. Finally, if $WIP_{r,m} = WIP_{r',m}$ for two operations r and r', then the fastest operation is processed first.

3.3.7 Single-level reductions

It is actually possible to perform a *single-level* reduction of the bilevel optimization models. Lower-level optimization models (3.14)-(3.20), (3.25)-(3.31) and (B.7)-(B.13) are convex optimization models with affine contraints, and thus can be replaced by their Karush–Kuhn–Tucker (KKT) conditions, which are necessary and sufficient (Boyd and Vandenberghe, 2004). The single-level reduction of a bilevel optimization problem is complex to solve, even in the convex case, because using KKT conditions conditions implies using constraints with bilinear terms. The single-level reduction is therefore relevant to define a single-level optimization model for disqualifications, which, however, is less intuitive.

3.4 Computational study

The objective of the computational study is to evaluate if the bilevel optimization models lead to a better decision-making in terms of throughput. If this is the case, then solution approaches may be worth investigating. Consequently, in this computational study, we compare the relevance of extending the utilization balancing approach in Chapter 2 to a bilevel optimization approach that maximizes the throughput. We also show that disqualification decisions can improve the throughput.

3.4.1 Design of experiments

Numerical experiments are carried out on industrial data from a manufacturing facility located in France, which is characterized by frequent product mix changes, shifting bottleneck work centers and high production variability. Two work centers, called work center A and work center B, are used for the computational study. For each work center, 19 industrial instances are considered. Both work centers have completely different machines and operations. In the computational study, optimization approaches are compared for a number of re-qualifications $k = 1$. This allows us to compare the optimal solution of optimization approaches in terms of throughput. We compare the optimization approaches on two different simulated dispatching rules: a FIFO dispatching rule, and a priority-based dispatching rule. Two different horizons are considered: 8 hours and 24 hours.

The industrial data of the computational study in Chapter 2 are again used. Note that the considered work centers are not necessarily identical to those studied in Chapter 2. Industrial data are summarized in Appendix B, Section B.0.2.

3.4.2 Numerical results

We cannot directly compare the value of the objective function of both the utilization balancing optimization approach and the bilevel optimization approach as they do not have the same objective function. This is because the utilization balancing optimization approach presented in Chapter 2 is not a direct measure of the throughput. In addition, both optimization approaches are compared in terms of gains on the throughput rather than in terms of decisions. For instance, a re-qualification proposed by the utilization balancing optimization approach could be different from the re-qualification proposed by the bilevel optimization approach, although both could lead to the same throughput as there may exist multiple optimal re-qualifications.

To make the comparison as fair as possible, once the utilization balancing optimization problem is solved, the optimal re-qualification is identified and evaluated with the bilevel optimization problem. When the bilevel optimization approach is used, the throughput can directly be computed with the objective function. After this transformation, both optimization approaches can be compared in terms of throughput.

It is worth mentioning that there is no bias toward or against the bilevel optimization model. Consider any re-qualification plan. Both the utilization balancing and the bilevel optimization models have the same solution in terms of utilization rate of the machines and capacity allocation, and therefore the same throughput. The only difference between the optimization models is that the final selection of the best re-qualification plan is either based on the utilization balancing or the throughput.

For each instance, we compare the throughput in terms of relative gain(%) with respect to the initial throughput when no re-qualification (or disqualification) is made. The initial throughput is computed with the optimization model (3.10)-(3.20) when $k = 0$. If the relative gains of optimization approaches are close on large number of instances and for each work center, then the bilevel optimization approach may not be as relevant as motivated in Section 3.2. If relative gains are very different, in particular on some instances, then the bilevel optimization approach brings information that is not captured by the utilization balancing optimization approach. Note that we do not report computational times as the purpose of the computational study is to evaluate the difference between optimization approaches on the throughput. Nevertheless, the bilevel optimization approach is expected to be more difficult to solve than the utilization balancing approach, especially since the utilization balancing can be solved efficiently (see Chapter 2).

In the numerical experiments, $\gamma = 4$ and $k = 1$. To solve the bilevel optimization models, we evaluate every possible re-qualification decision (or disqualification decision), we solve the lower-level optimization problem, then compute the throughput. The best qualification decision is kept. This procedure corresponds to the constructive greedy heuristic presented in Chapter 2. The lower-level optimization

problem is a nonlinear optimization problem. To solve it, we linearize the objective function and use a cutting-plane algorithm (see also Chapter 2). The cutting-plane algorithm is stopped when a relative gap lower than 0.0001 is reached. Each iteration of the cutting-plane algorithm is solved by CLP, which is an open source solver (Lougee-Heimer, 2003; Löhndorf, 2016). The cutting-plane algorithm is implemented in Java 8 on a computer with an Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz with 4 cores and 8 GB of RAM on Windows 10.

Note that, when performing disqualifications, we limited ourselves to disqualifications that respect some industrial constraints, more precisely, disqualification decisions are restricted to low priority operations. Disqualification decisions are also restricted to operations that are not subject to "time constraints". A time (soft) constraint is a practical maximum cycle time defined between two operations to ensure the yield and quality of products (*e.g.*, see Lima et al. 2019, 2020). In addition, disqualification decisions may be forbidden if they create line stops or there is a single qualified machine for a operation.

3.4.2.1 Numerical results with one qualification

Table 3.4, respectively Table 3.5, compares the gain on the throughput of the utilization balancing approach and bilevel optimization approaches for work center A, respectively work center B. Numerical results for the bilevel optimization model are obtained by solving the bilevel optimization model presented in Section 3.3.3.

From a general standpoint, contrary to the bilevel optimization model, the utilization balancing approach may "fail" to determine one re-qualification that improves the throughput. However, this strongly depends on the simulated dispatching rules, the duration of the horizon, the instance and the work center.

For work center A, the number of instances where the relative gain determined by the utilization balancing is negative, is equal to 19 for a horizon of 8 hours, and 5 for a horizon of 24 hours (see Table 3.4). In general, there are two times more instances where the relative gain is negative when operations are processed with respect to their priority. Similarly, for work center B, the number of instances where the relative gain determined by the utilization balancing is negative, is equal to 17 for a horizon of 8 hours, and 6 for a horizon of 24 hours (see Table 3.5).

This is mainly because of *production variability*, the lack of constant pattern in the demand. As motivated in Section 3.2.1, optimizing the throughput is not strictly equivalent to optimizing the utilization balance. In practice if two machines are overloaded, then adding a re-qualification means trying to load the machines even more, which can be irrelevant if the throughput must be optimized. The bilevel optimization approach is able to detect the best re-qualification among two possible re-qualifications that would lead to similar utilization balances but to different throughputs. The bilevel optimization approach can thus better determine re-qualification plans that optimize the throughput.

We observe that the difference between the relative gains between both optimization approaches is largest when operation priorities are considered. For instance, for instance 14 and work center A, the relative gain is equal to -2.66% when the utilization balancing optimization approach is used whereas the relative gain is

equal to 1.17% when the bilevel optimization approach is used. The difference is much smaller when the simulated dispatching rule is FIFO. Therefore, the choice of the operation, and not only the quantity to re-balance, with respect to its priority, seems also crucial to optimize throughput. As a matter of fact, qualifying operations with low priority may be inefficient to optimize the throughput. This is because they can still be backlogged as they are processed last although the initially disqualified machine is faster and underloaded. Qualifying operations with low priority is efficient if a large workload can be transferred from one machine to another.

Note that there also some instances where the relative gains of both optimization approaches are almost identical. Consider work center A, when operations priorities are considered, the relative gains are very close for instances 11, 19 and a horizon of 8 hours. For a horizon of 24 hours, instances 7, 8, 9 and 10 are very close. The difference between relative gains is at most 0.2%. When a FIFO rule is simulated, the relative gains are very close for instances 7, 8, 18 and 19. Similar results can be observed for work center B. For instance, when operations priorities are considered, the relative gains are very close for instances 13, 14 for a horizon of 8 hours. For a horizon of 24 hours, the relative gains are very close for instances 1, 9, 10, 11, 12, 17 and 19. Although the utilization balancing optimization approach "fails" to determine one satisfactory re-qualification that improves the throughput, on a large number of instances, the utilization balancing optimization approach leads to very similar gains. This happens when the throughput can be optimized, no matter the operation queues. It would be interesting to automatically analyze instances when this happens to solve the utilization balancing optimization problem first, as the bilevel optimization approach is expected to be more computationally expensive.

Interestingly, we can also observe that the difference between the quality of re-qualifications determined by optimization approaches is large when the horizon is 8 hours, but the difference is much smaller when the horizon is 24 hours. The difference is smaller both in terms of relative gain and the number of instances with a negative relative gain. This is counter-intuitive but the difference can also be explained by the production variability. On short horizons, re-qualifications are made to tackle demand peaks, therefore the variability of the demand profile. Operation queues will have a more critical role in the determination of satisfactory re-qualifications than on larger horizons, even of only 24 hours. Considering the way operations compete, *i.e.* operation queues, for the same production resources is relevant and can lead to better re-qualification plans to optimize the throughput. This confirms existing results in the literature that shows that dispatching decisions strongly influence the relevance of additional qualifications (Johnzén et al., 2008; Gurumurthi and Benjaafar, 2004; Kopp et al., 2019). For short horizons, production variability has a significant impact of the quality of the proposed re-qualifications. To a certain extend, this also means that defining re-qualification strategies for factories with high production variability may be inefficient to optimize the throughput.

Furthermore, we observe in the numerical results that there are frequently only a limited number of re-qualifications that lead to an interesting increase of the throughput. Most re-qualifications do not improve the throughput.

Finally, we also observe that the mean relative gain determined by the bilevel optimization approach is larger for a horizon of 8 hours than for a horizon of 24

hours. Consider the case when operation priorities are considered. For work center A, the mean relative gain is equal to 1.27% for a horizon of 8 hours whereas the mean relative gain is equal to 0.84% for a horizon of 24 hours. For work center B, the mean relative gain is equal to 1.40% for a horizon of 8 hours whereas the mean relative gain is equal to 0.89% for a horizon of 24 hours. This may suggest that it is more relevant to frequently make re-qualification decisions, for instance one re-qualification decision every 8 hours, than one re-qualification every 24 hours because the real gain on the throughput may be larger.

Table 3.4: Comparison of the relative gain (%) on the throughput between the utilization balancing optimization approach and the bilevel optimization approaches for work center A (Bold values are negative gains).

	8 hours				24 hours			
	Utilization Balancing		Bilevel Approach		Utilization Balancing		Bilevel Approach	
Instance	FIFO	Priority	FIFO	Priority	FIFO	Priority	FIFO	Priority
1	**-0.02**	**-0.18**	1.13	1.03	0.08	0.26	0.32	0.52
2	0.11	1.03	1.10	1.80	0.26	0.03	0.43	0.63
3	0.12	0.64	0.19	1.89	0.11	0.44	1.13	0.96
4	0.12	**-0.04**	0.12	0.20	0.10	**-0.13**	0.10	0.19
5	0.00	**-1.46**	0.54	0.54	0.03	**-0.14**	0.04	0.30
6	**-0.38**	**-0.79**	0.04	0.15	0.04	0.24	0.04	0.70
7	**-1.31**	**-0.89**	1.01	1.02	0.60	1.50	0.60	1.61
8	**-1.43**	**-2.37**	1.40	1.47	0.51	1.28	0.51	1.28
9	0.01	**-0.33**	1.37	1.40	0.01	**-0.38**	0.78	0.77
10	0.03	0.23	0.03	1.01	0.02	**-0.50**	0.02	0.19
11	0.00	**-0.19**	0.22	1.41	0.05	0.95	0.24	0.95
12	0.17	1.87	1.17	1.87	0.13	1.10	0.48	1.10
13	**-0.68**	**-0.82**	0.73	0.72	0.09	0.47	0.68	0.94
14	**-0.57**	**-2.66**	0.49	1.17	0.01	0.19	1.02	0.98
15	0.03	**-0.45**	2.28	2.32	0.09	0.12	1.10	1.09
16	0.02	**-0.01**	1.50	2.10	0.03	0.25	0.03	0.82
17	0.03	**-1.19**	0.25	1.45	0.04	**-0.19**	0.17	1.19
18	0.06	0.26	0.14	1.34	0.10	0.23	0.15	0.90
19	0.04	1.01	0.23	1.29	0.15	0.07	0.15	0.29
Mean	***-0.20***	***-0.41***	*0.76*	*1.27*	*0.13*	*0.32*	*0.44*	*0.84*
Number of Instances With Negative Gains	6	13	0	0	0	5	0	0

Table 3.5: Comparison of the relative gain (%) on the throughput between the utilization balancing and the bilevel optimization approaches for work center B (Bold values are negative gains).

	8 hours				24 hours			
	Utilization Balancing		Bilevel Approach		Utilization Balancing		Bilevel Approach	
Instance	FIFO	Priority	FIFO	Priority	FIFO	Priority	FIFO	Priority
1	**-0.30**	**-0.47**	1.32	1.29	2.10	2.25	2.10	2.25
2	0.15	0.20	1.30	1.74	0.48	0.49	1.07	1.16
3	0.15	**0.00**	1.21	1.67	0.24	**0.00**	1.14	1.22
4	**-1.84**	**-2.69**	0.51	0.51	0.32	0.25	0.55	0.46
5	**-1.02**	**-1.14**	0.59	0.55	**-0.65**	**-0.85**	0.58	0.51
6	**-0.59**	**-0.88**	1.44	1.44	0.30	0.24	0.86	0.88
7	0.23	0.09	1.65	1.64	0.19	0.01	0.74	0.68
8	0.23	0.00	0.41	0.38	0.19	0.00	0.88	0.88
9	**-0.01**	0.00	0.46	0.62	0.56	0.57	0.56	0.57
10	0.48	0.00	0.48	0.54	0.41	**-0.03**	0.52	0.52
11	0.04	0.03	1.19	1.03	0.74	0.75	0.74	0.75
12	0.15	0.55	1.61	1.46	0.73	0.76	0.73	0.76
13	2.79	2.77	2.79	2.77	0.00	**0.00**	0.61	0.60
14	3.29	3.25	3.29	3.25	0.00	0.00	0.81	0.82
15	**-0.09**	0.01	0.56	0.58	0.06	**-0.02**	0.35	0.32
16	**-0.03**	**-0.08**	2.67	2.63	0.41	0.41	0.84	0.89
17	**-0.30**	**-1.18**	1.00	1.05	1.38	1.47	1.38	1.47
18	**-1.23**	**-0.72**	1.39	1.15	1.03	0.93	1.48	1.29
19	0.49	0.41	2.29	2.32	0.98	0.87	0.98	0.87
Mean	*0.14*	*0.01*	*1.38*	*1.40*	*0.50*	*0.43*	*0.89*	*0.89*
Number of Instances With Negative Gains	9	8	0	0	1	5	0	0

3.4.2.2 Numerical results with one disqualification

This section shows on industrial data that is possible to find disqualifications that improve the throughput. Table 3.6 illustrates the relative gain on the throughput by instance, by work center and by simulated dispatching rule for one disqualification decision. Numerical results are obtained by solving the bilevel optimization model presented in Section 3.3.4. Note that, for experiments on disqualification decisions, additional industrial constraints were added to the disqualification optimization approach. Disqualifications are not evaluated (1) for operations that are run for priority lots, (2) if they create an operation without any qualified machine, and (3) for operations that must be run for lots in time constraints.

Similarly to one re-qualification decision, one disqualification decision can improve the throughput. However, the relative gain strongly depends on the instance, the work center, the simulated dispatching rule, and the horizon. Furthermore, there are only a limited number of disqualifications that lead to a significant increase of the throughput. Most disqualifications do not improve the throughput.

Similarly to observations made in Section 3.4.2.1, the relative gain strongly depends on the simulated dispatching rules. For a horizon of 8 hours and a FIFO simulated dispatching rule, the mean relative gain is equal to 0.29% for work center

A whereas the mean relative gain is equal to 0.23% for work center B. When operation priorities are considered, the mean relative gain for work center A is equal to 1.56% whereas the mean relative gain is equal to 0.24%. Therefore, considering the way dispatching decisions are designed in a work center can lead to better disqualification decisions in terms of throughput.

Table 3.6 shows that some work centers can be relatively insensible to one disqualification decision. For instance, for a horizon of 8 hours, the mean relative gain is approximately equal to 0.23% for work center B. However, the relative gain also depends on the instance. On instances 12, 13 and 14, the relative gain is larger than 0.43% when a FIFO dispatching rule is simulated. However, dispatching decisions are particularly relevant for work center A. Consider a horizon of 8 hours and the priority based simulated dispatching rules, for which the mean relative gain is equal to 1.56%. The relative gain strongly varies from one instance to another. However, the minimum relative gain is equal to 0.47%. This shows a significant difference between the impact of disqualification decisions between two work centers. Therefore, generalizing disqualification decisions to all work centers may be irrelevant. To a certain extent, we may find work centers for which disqualification decisions are always irrelevant.

Table 3.6: Comparison of the relative gain (%) on the throughput for one disqualification decision by instance, by work center and by simulated dispatching rule.

	Work center A				Work center B			
	8 hours		24 hours		8 hours		24 hours	
Instance	FIFO	Priority	FIFO	Priority	FIFO	Priority	FIFO	Priority
1	0.18	1.51	0.05	0.39	0.10	0.17	0.00	0.01
2	0.45	1.99	0.11	0.87	0.02	0.04	0.00	0.02
3	0.02	2.05	0.05	0.78	0.05	0.10	0.00	0.02
4	0.24	1.23	0.02	0.38	0.13	0.53	0.09	0.09
5	1.09	0.59	0.01	0.66	0.06	0.15	0.07	0.10
6	0.34	0.74	0.08	0.56	0.01	0.06	0.04	0.04
7	0.28	2.97	0.03	1.83	0.27	0.29	0.14	0.02
8	0.55	1.21	0.01	0.87	0.17	0.17	0.15	0.06
9	0.38	3.12	0.01	0.37	0.12	0.15	0.00	0.02
10	0.07	2.36	0.00	0.32	0.12	0.42	0.23	0.24
11	0.04	1.09	0.00	0.11	0.16	0.03	0.10	0.01
12	0.25	2.72	0.07	0.78	0.43	0.22	0.06	0.09
13	0.19	0.91	0.04	0.51	0.86	0.11	0.14	0.35
14	0.13	0.47	0.13	0.62	0.76	0.38	0.13	0.31
15	0.33	0.71	0.23	0.57	0.12	0.25	0.09	0.10
16	0.25	1.73	0.01	0.50	0.13	0.36	0.16	0.19
17	0.42	1.60	0.01	1.34	0.14	0.13	0.13	0.25
18	0.14	1.48	0.00	0.65	0.10	0.24	0.01	0.00
19	0.08	1.15	0.00	0.78	0.63	0.66	0.00	0.00
Mean	*0.29*	*1.56*	*0.05*	*0.68*	*0.23*	*0.24*	*0.08*	*0.10*

3.4.2.3 Numerical results with one re-qualification and one disqualification

Finally, in the computational study, we are interested in illustrating the potential of combining one re-qualification decision (OQ) and one disqualification decision

(DOQ) to improve the throughput with the bilevel optimization model introduced in Section 3.3.5. This numerical study is relevant to validate the bilevel optimization model. As a matter of fact, we could intuitively conclude that disqualifications and re-qualifications are incompatible. This is because disqualification decisions use the differences of throughput rates between machines and therefore lead to a decrease in the utilization balance whereas re-qualification decisions use the fact that machines are poorly balanced to improve the throughput. Note that, contrary to the numerical experiments when only one re-qualification decision or one disqualification decision is made, optimal solutions are not presented. To solve the bilevel optimization model presented in Section 3.3.5, a greedy heuristic was used. It is similar to the constructive greedy heuristic presented in Chapter 2. An action plan is iterately built. First, every possible disqualification decision is evaluated. The disqualification that best optimizes the throughput is kept in the action plan. Then, every possible re-qualification decision is evaluated. The re-qualification that optimizes the throughput is kept in the action plan. Another greedy heuristic consists in selecting first the best re-qualification decision instead of the best disqualification decision. Both greedy heuristics provide similar numerical results. The difference on the relative gain by instance is different by at most 0.2%. Therefore, we only show the numerical results for the greedy heuristic when the disqualification decision is made first. Moreover, the horizon here is 8 hours. Table 3.7 presents the numerical results by instance, by work center and by simulated dispatching rule.

The numerical results show that, if the re-qualification and disqualification decisions are carefully determined, it is possible to combine these decisions to improve the throughput (see Table 3.7). In general, the mean relative gain obtained after one re-qualification and one disqualification decision is slightly less than the sum of the mean relative gains obtained after one re-qualification and one disqualification decision.

Consider work center A, when operation priorities are considered, the mean relative gain after combining one re-qualification decision and one disqualification decision is equal to 2.45%. The mean relative gain after one qualification decision is only 1.56%. When a FIFO dispatching rule is simulated, the mean relative gain is equal to 1.00%, and the mean relative gain after one qualification decision is equal to 1.00%. As in Sections 3.4.2.1 and 3.4.2.2, the mean gain is strongly dependent on the instance.

For work center B, as partly mentioned in Section 3.4.2.2, the mean relative gain after combining one re-qualification decision and one disqualification decision is almost equal to the mean relative gain after making only one re-qualification decision.

Therefore, it is possible to combine disqualification and qualification decisions. As in general, the mean relative gain obtained after one re-qualification and one disqualification decision is slightly less than the sum of the mean relative gains obtained after one re-qualification and one disqualification decision, we expect that only a limited number of disqualification and re-qualification decisions in the same action plan can actually reach the best possible improvement on the throughput.

Table 3.7: Comparison of the relative gain (%) on the throughput after making one disqualification decision (DOQ) and one qualification decision (OQ) by instance, by work center and by simulated dispatching rules.

		FIFO			Priority		
Work center	Instance	DOQ	OQ	DOQ then OQ	DOQ	OQ	DOQ then OQ
A	1	0.18	1.13	1.29	1.51	1.03	2.27
	2	0.45	1.10	1.54	1.99	1.80	3.09
	3	0.02	0.19	0.18	2.05	1.89	3.66
	4	0.24	0.12	0.34	1.23	0.20	1.38
	6	1.09	0.54	1.41	0.59	0.54	1.13
	7	0.34	0.04	0.38	0.74	0.15	1.02
	8	0.28	1.01	1.34	2.97	1.02	3.97
	9	0.55	1.40	1.95	1.21	1.47	2.14
	10	0.38	1.37	1.75	3.12	1.40	4.47
	11	0.07	0.03	0.08	2.36	1.01	2.59
	12	0.04	0.22	0.26	1.09	1.41	1.95
	13	0.25	1.17	1.42	2.72	1.87	3.89
	14	0.19	0.73	0.92	0.91	0.72	1.48
	15	0.13	0.49	0.62	0.47	1.17	0.97
	16	0.33	2.28	2.61	0.71	2.32	2.68
	17	0.25	1.50	1.75	1.73	2.10	3.65
	18	0.42	0.25	0.68	1.60	1.45	2.81
	19	0.14	0.14	0.28	1.48	1.34	1.98
	20	0.00	0.23	0.22	1.15	1.29	1.33
Mean		*0.28*	*0.73*	*1.00*	*1.56*	*1.27*	*2.45*
B	1	0.10	1.32	1.41	0.17	1.29	1.46
	2	0.00	1.30	1.30	0.02	1.74	1.76
	3	0.00	1.21	1.22	0.04	1.67	1.71
	4	0.01	0.51	0.52	0.01	0.51	0.53
	6	0.06	0.59	0.59	0.12	0.55	0.66
	7	0.00	1.44	1.45	0.06	1.44	1.51
	8	0.18	1.65	1.83	0.07	1.64	1.72
	9	0.16	0.41	0.56	0.09	0.38	0.47
	10	0.06	0.46	0.52	0.14	0.62	0.77
	11	0.03	0.48	0.51	0.10	0.54	0.64
	12	0.16	1.19	1.19	0.03	1.03	1.06
	13	0.43	1.61	2.13	0.22	1.46	1.61
	14	0.80	2.79	3.58	0.08	2.77	2.87
	15	0.73	3.29	4.02	0.24	3.25	3.50
	16	0.12	0.56	0.68	0.25	0.58	0.83
	17	0.07	2.67	2.74	0.08	2.63	2.72
	18	0.05	1.00	1.04	0.01	1.05	1.05
	19	0.01	1.39	1.40	0.04	1.15	1.19
	20	0.01	2.29	2.30	0.03	2.32	2.35
Mean		*0.16*	*1.38*	*1.53*	*0.09*	*1.40*	*1.49*

3.5 Recommendations

Production personnel should pay attention to dispatching decisions when making qualification and disqualification decisions. Dispatching decisions strongly affect the quality of decisions. Our computational study on industrial instances shows that using the bilevel optimization models is more appropriate than using a utilization balancing optimization approach to determine qualification and disqualification decisions that optimize the throughput. When small computational times are required, such as in a Decision Support System (DSS), we recommend to use bilevel optimization models because they are able to capture the effect of dispatching decisions. However, when small computational times are not required, for instance for factories with low production variability where qualification strategies are acceptable instead of tailored decisions for each instance, then an optimization-simulation approach can be suitable.

Nevertheless, numerical results show that as the length of the horizon increases, the best re-qualifications determined between the utilization balancing and the bilevel optimization approaches are the same or, in other words, lead to the same gain on the throughput. This is because, as the length of the horizon increases, demand peaks of (priority) operations are averaged over the planning horizon, therefore the utilization rates of the machines tend to be naturally more balanced. Therefore, for large horizons, longer than a few days, the utilization balancing approach to determine re-qualifications is relevant, and appealing as it can solved efficiently (see Chapter 2). Nevertheless, the bilevel optimization approach is still relevant to consider the throughput instead of utilization balance, which cannot be directly performed with the utilization balancing approach. Finally, note that, because operation priorities are frequently updated at the production control level, it is probably irrelevant to use the priority based computation mode for the throughput for horizons ranging from a few hours to a few days depending on the work center and the update frequency. In other words, on the long term, no operation is prioritized and, therefore, the computation mode “average product mix” may be better suited for large horizons.

In some cases, disqualification decisions are particularly relevant to optimize manufacturing performances. For instance, when a qualified machine for a operation is known to have a significant lower yield than other qualified machines, production personnel disqualifies the machine if the impact on the throughput is limited. Disqualifications decisions also make sense if the throughput rate of the operation of the machine has significantly decreased due to its wear. In this case, it is preferable for the operation to be processed by other qualified machines.

However, managing the production capacity of a work center for *throughput optimization purposes* with disqualification decisions may not be a *satisfactory industrial practice*. For instance, disqualifying operations on machines that do not have yield issues can be dangerous, for instance, in terms of interruption of the production flow, increase of production variability and unexpected future *quality* problems on operations that could have been measured or managed if the operations were qualified. Disqualifications can also be forgotten. This is why it is preferable to temporarily disqualify operations, notably to avoid reaching the end of qualification

time windows (Obeid et al., 2014; Nattaf et al., 2019; Kopp et al., 2019) and frequently reevaluating the disqualification decisions as production personnel tends to do nowadays. In some work centers, disqualifications can be relevant. Sometimes, some operations are known to wear machines more than other operations. For these operations, at most one machine is kept qualified and others are disqualified. In this case, bilevel optimization models can help deciding the machine that should be kept qualified in terms of throughput. In addition, if process times depend on machine wear, bilevel optimization models can help deciding if it is worth in terms of throughput to keep the qualification given the current machine wear.

Moreover, although managing the production capacity of machines with disqualifications without yield problems can unsuitable to a certain extent, it is "difficult" for production personnel not use disqualification decisions given they are challenged to optimize manufacturing performance as dispatching engines are shortsighted and a source of capacity loss, and that this way of working did not lead to major production problems.

We argue that disqualification decisions could be avoided if dispatching engines had a better vision of upcoming demand peak from upstream work centers. For instance, this could be translated by giving penalties to operation quantities allocated to a machine beyond a given threshold that can be determined after solving the bilevel optimization problem for $k = 0$ (or equivalently the utilization optimization problem). In other words, the utilization rates of the machines computed with the utilization balancing approach could be given to dispatching decisions as guidelines. However, including these penalties and guidelines in a dispatching engine may be complex. Using optimization solutions such as solving multi-objective parallel-machine or complex job-shop scheduling problems may be valuable (Knopp et al., 2017). Solving scheduling problems is also computationally complex and very challenging. In the near future, disqualification decisions will *probably* remain relevant to better manage the production capacity of a work center.

3.6 Conclusions and perspectives

In this chapter, we propose bilevel optimization models for short-term qualification and disqualification management in semiconductor manufacturing. To the best of our knowledge, there is no paper that copes with disqualification management in semiconductor manufacturing. Then, we validate the proposed bilevel optimization models on industrial data. Contrary to utilization balancing optimization approach on some instances, the bilevel optimization approach considers dispatching decisions, which is in some cases critical to determine relevant re-qualifications in terms of throughput. On horizons larger than 24 hours, the bilevel optimization approach and the utilization balancing optimization approach propose re-qualifications that have similar gains. If only re-qualifications are desired, the utilization balancing optimization approach is preferable for long horizons as it can be solved efficiently and effectively (see Chapter 2). The numerical results show that the bilevel optimization approach can also be used to determine relevant disqualifications in terms of throughput. We argue that disqualifications may not always be a satisfactory

industrial practice even though they can be used to improve the throughput.

In Chapter 4, we extend the single-period bilevel optimization model to a multi-period bilevel optimization model. We show that the multi-period bilevel optimization model is particularly relevant when re-qualifications are subject to delays or induce capacity losses. In Chapter 5, we also show that for cycle time forecasts, the multi-period bilevel optimization model is more appropriate than the single-period bilevel optimization model.

We believe the following perspectives are worth investigating in the future (out of the scope of the thesis):

1. Some parameters used in the bilevel optimization models might be subject to uncertainty, *e.g.* the demand by operation and the production capacity by machine. From the production personnel's point of view, dispatching decisions can also be subject to uncertainty. Determining robust re-qualification and disqualification plans can be profitable.

2. In the computational study, we limited ourselves to the case where $k = 1$. However, designing efficient and effective solution approaches, that quickly determine re-qualification and disqualification plans for $k > 1$, is a relevant research avenue. Note that, although no thorough computational study was performed to evaluate efficient and effective solution approaches, in the decision support system presented in Chapter 7, constructive greedy heuristics are used to solve the bilevel optimization models presented in this chapter. More precisely, to avoid evaluating every re-qualification and disqualification at each iteration of the greedy heuristic, pre-processing rules are used. For instance, pre-processing rules are based on the values of the dual variables of qualification constraints for re-qualification. Pre-processing rules are also based on the fact that potential effective re-qualifications to perform (at least for some work centers) are the re-qualifications that consider re-qualifying operations for underloaded machines, *i.e.* for machines such that $U_m < 1$, or re-qualifying operations that can be processed much faster on disqualified machines than on currently qualified machines.

3. Similarly to Chapter 2, it would be interesting to better understand the impact of different γ settings on re-qualification decisions and computing time.

4. Considering multiple objectives in short-term qualification and disqualification management is a relevant research avenue.

5. In some work centers where wafers are processed by batch (Rowshannahad and Dauzère-Pérès, 2013), it is desirable to consider batch size constraints in the bilevel optimization models.

6. As already discussed, there may be multiple solutions to the lower-level optimization problem, and some solutions may lead to a better throughput than others, because the proposed bilevel formulations are neither optimistic or pessimistic. Studying this is left for future research, *e.g.* by giving bounds on the throughput to production personnel. For instance Fischetti et al. (2018)

propose a refinement procedure that can be used to obtain an optimistic solution from the lower-level.

Chapter 4

A multi-period bilevel optimization approach for throughput maximization

In this chapter, the bilevel optimization model presented in Chapter 3 is extended to better consider re-qualification delays and maintenance operations potentially required for re-qualification activities. We show that if re-qualifications are subject to lead times or induce maintenance operations, then considering a multi-period (dynamic) approach can lead to making re-qualifications that will have a better result on the throughput than re-qualifications made with a single-period (static) approach*.

*Large parts of this chapter can be found in (Perraudat et al., 2019).

4.1 Introduction

To the best of our knowledge, the fact that re-qualifications can be subject to delays or can require maintenance operations is rarely considered in the literature. Chang and Dong (2017) consider a single-period stochastic optimization approach where the number of processed wafers must be maximized. The demand is considered as uncertain and qualifications induce stochastic capacity losses, for instance due to maintenance operations. Kopp et al. (2018) and Kopp et al. (2019) propose a mixed integer linear program and the effect re-qualifications of performances are assessed with a simulation model. Re-qualification lead times are not included in the mixed integer linear program but are included in the simulation model, which are modeled as a 75-minute delay to consider measurement activities. Re-qualification delays and maintenance operations can strongly influence the relevance of re-qualification decisions.

Figure 4.1 illustrates three common different demand profiles that can be observed for operations in wafer fabs due to production variability. Operation A is expected to have a constant profile with about 500 wafers arriving in the work center at every period, *e.g* every hour, every shift (8 hours) or every day. A large peak of demand (Work-in-Process bubble) is expected for operation B. Finally, there is no demand for operation C in the first four periods but a large number of wafers is expected afterwards.

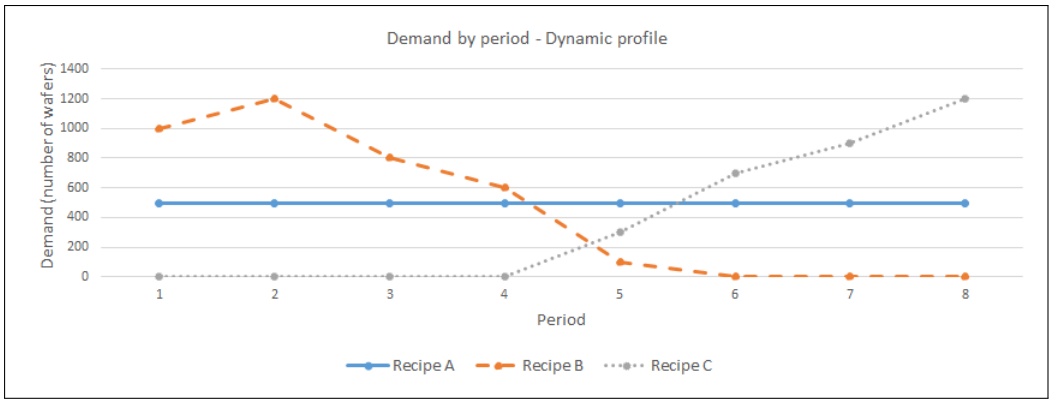


Figure 4.1: Illustrative example of the dynamic profile of the demand for three operations (recipes).

Suppose that only one re-qualification is allowed, and that the single-period bilevel optimization approach proposes to re-qualify operation B on a disqualified machine. However, a re-qualification lead time or a maintenance operation of 4 periods is expected. Then, re-qualifying operation B on a disqualified machine is actually irrelevant. This is because short-sighted dispatching decisions do not wait for the end of the re-qualification procedure. A large number of wafers requiring operation B and arriving in the work center in the first periods has already been processed by other qualified machines than the re-qualified machine. In other words, the single-period bilevel optimization approach proposes a re-qualification, which could be optimal in terms of throughput, that has a impact in practice on

the throughput because of re-qualification lead times. The effect on the throughput might even be null if operation B is highly prioritized. The real demand profile for operations seen by the single-period bilevel optimization approach is illustrated in Figure 4.2.

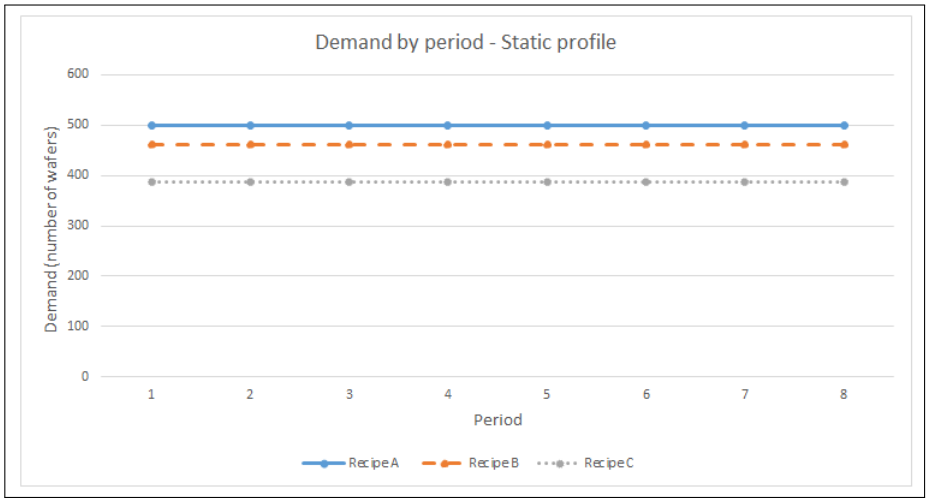


Figure 4.2: Operation (recipe) demand profiles seen by the single-period bilevel optimization approach.

If we had known that the re-qualification was ineffective to improve the throughput, we would have chosen to re-qualify a machine for operation A, with a constant demand profile, or even for operation C with a expected demand peak later in the horizon even if the re-qualifications initially looked sub-optimal.

Because the single-period optimization approach is unable to differentiate re-qualifications for operations that have different demand profiles, a multi-period optimization approach seems necessary. This is the case for work centers which frequently undergo large fluctuations in operation demand profiles, mostly due to production variability and the short-sighted aspect of dispatching rules.

This effect is not limited to delays or maintenance operations induced by re-qualifications. It can also be observed if a work center manager wants to simulate a maintenance operation, even a short maintenance operation, on a machine in a decision support system and to determine a set of re-qualifications. It can also be observed if a machine is currently down. For instance, the single-period optimization approach can recommend re-qualifying a operation on the affected machine by the maintenance operation because the machine has still some capacity left at the end of the horizon. However, most wafers of the corresponding operation will be processed by other qualified machines due to the short-sighted aspect of dispatching rules.

The bilevel optimization model presented in Chapter 3 is extended in two ways: (1) An extended single-period bilevel optimization model is proposed to better consider re-qualifications that lead to maintenance operations, and (2) A multi-period bilevel optimization model is proposed to better consider both re-qualification delays and re-qualifications that lead to maintenance operations. In addition, the

effect of the multi-period bilevel optimization approach on the throughput is compared to the effect of the single-period bilevel optimization approach. In particular, we show that, if re-qualifications are subject to delays or induce maintenance operations, then considering a multi-period (dynamic) approach can lead to proposing re-qualifications that will have a better result on the throughput than re-qualifications proposed with a single-period (static) approach. In this chapter, a dynamic approach should be understood as a multi-period approach. A static approach should be understood as a single-period approach.

This remainder of this chapter is organized as follows[†]. In Section 4.1, the dynamic approach is motivated with a practical illustrative example. In Section 4.2, the extended single-period and the multi-period bilevel optimization approaches are presented. In Section 4.3, the extended single-period and the multi-period bilevel optimization approaches are compared on industrial data. Practical insights are also proposed. Finally, in Section 4.5, conclusions and perspectives are outlined.

4.2 Problem statement

Consider a work center consisting of M non-identical parallel machines. On a given time horizon consisting of T periods, R different operations, each with a strictly positive quantity, must be processed by the machines. The throughput rate of each operation on each machine is known. However, each machine has a finite production capacity by period and can only process qualified operation. Similarly, a operation can only be processed by qualified machines. The qualification matrix, *i.e.* the operation that are initially qualified and that can be qualified, is known. We assume that, when a re-qualification decision of operation r is made on machine m, either a re-qualification lead time $l_{r,m}$, or a maintenance operation of duration c^{loss}. Re-qualification decisions are assumed to be made at $t = 0$. The objective is to determine k qualifiable pairs (operation r, machine m) to maximize the throughput.

4.2.1 Notations

Indices and sets:
m: Index for machines, $\in \{1,..,M\}$,
r: Index for operations, $\in \{1,..,R\}$,
t: Index for periods, $\in \{1,..,T\}$.
Parameters:
k: Number of re-qualification decisions to be made at the beginning of the planning horizon,
$q_{r,m}$: Is equal to 1 if machine m is initially qualified for operation r, is equal to 2 if machine m is initially qualifiable for operation r, is equal to 0 if machine m cannot be qualified for operation r,
$tp_{r,m}$: Throughput rate (in number of wafers by seconds) of operation r on machine m,

[†]Parts of this Chapter have been presented at the EURO 2019 conference and Winter Simulation Conference 2019 (see Perraudat et al. 2019).

$c_{t,m}$: Initial availability time (in seconds) of machine m over the planning horizon,
$c^{loss}_{t,r,m}$: Capacity loss generated (in seconds) by re-qualifying operation r on machine m at period t,
$l_{r,m}$: Re-qualification lead time required for re-qualifying operation r on machine m,
$d_{t,r}$: New arriving wafers in the work center requiring operation r at period t,
$I_{0,r}$: Initial operation quantity in number of wafers r at $t = 0$ in the work center,
γ: Utilization balancing parameter strictly greater than one.

Decision variables:
$OQ_{r,m} \in \{0,1\}$: Is equal to 1 if a re-qualification procedure is proposed for operation r on machine m at the beginning of the planning horizon, and 0 otherwise,
$U_{t,m}$: Utilization rate of machine m,
$C^{eff}_{t,m}$: Effective availability time (in seconds) of machine m at period t,
$I_{t,r}$: Number of wafers of operation r in the work center at the end of period t,
$WIP_{t,r,m}$: Quantity of operation r allocated to machine m at period t, $TH_{t,r}$: Number of wafers of operation r processed at period t,
$D_{t,r}$: Total demand in number of wafers for operation r processed at period t,
$Q_{t,r,m}$: New state of the qualification at period t for operation r, machine m processed at period t,
$C^{neg}_{t,m}$: Indicator variable that is strictly negative if there is still capacity loss to impute at $t+1$ on machine m,
$\mathcal{R}^{q}_{t}(\mathbf{Q}, \mathbf{C^{eff}}) = \{r \mid \sum_{m=1|C^{eff}_{t,m}>0}^{M} \mathbb{1}(Q_{t,r,m}) > 0\}$: Set of operations with at least one qualified machine with some capacity at period t, where $\mathbb{1}(x) = 1$ if $x = 1$, and 0 otherwise.

4.2.2 Extended single-period bilevel optimization model

In this section, the period index t is omitted as only one period is considered. The extended single-period bilevel optimization model can be found below.

Upper-level optimization problem:

$$\max \quad TH = f(\mathbf{U}, \mathbf{WIP}) \tag{4.1}$$

$$\text{s.t.} \quad \sum_{r,m} OQ_{r,m} = k \tag{4.2}$$

$$Q_{r,m} = OQ_{r,m} \qquad \forall r, \forall m \mid q_{r,m} = 2 \tag{4.3}$$

$$Q_{r,m} = q_{r,m} \qquad \forall r, \forall m \mid q_{r,m} \neq 2 \tag{4.4}$$

$$C^{eff}_{m} = \max(c_m - \sum_{r} c^{loss}_{r,m} OQ_{r,m}, 0) \qquad \forall m \tag{4.5}$$

$$U_m, WIP_{r,m} \in \arg\min LBP(\mathbf{Q}, \mathbf{C^{eff}}) \tag{4.6}$$

$$OQ_{r,m} \in \{0,1\} \qquad \forall r, \forall m \tag{4.7}$$

Lower-level optimization problem:

$$LBP(\mathbf{Q}, \mathbf{C^{eff}}) = \min \quad \sum_m U_m^\gamma \tag{4.8}$$

$$\text{s.t.} \quad \sum_m WIP_{r,m} = d_r + I_{0,r} \qquad \forall r \in \mathcal{R}^q(\mathbf{Q}, \mathbf{C^{eff}}) \tag{4.9}$$

$$U_m = \sum_r \frac{WIP_{r,m}}{tp_{r,m} C_m^{eff}} \qquad \forall m \mid C_m^{eff} > 0 \tag{4.10}$$

$$U_m = 0 \qquad \forall m \mid C_m^{eff} = 0 \tag{4.11}$$

$$WIP_{r,m} \leq d_r + I_{0,r} \qquad \forall r, \forall m \mid q_{r,m} = 1 \tag{4.12}$$

$$WIP_{r,m} \leq (d_r + I_{0,r}) OQ_{r,m} \qquad \forall r, \forall m \mid q_{r,m} = 2 \tag{4.13}$$

$$WIP_{r,m} \leq 0 \qquad \forall r, \forall m \mid q_{r,m} = 0 \tag{4.14}$$

$$WIP_{r,m} \geq 0 \qquad \forall r, \forall m \tag{4.15}$$

Upper-level optimization problem. The objective function (4.1) maximizes the throughput. Constraint (4.2) sets to k the number of re-qualifications that must be performed at the beginning of the horizon. Constraints (4.3)-(4.4) determine the new state of each qualification after re-qualification decisions. Constraints (4.5) define the effective availability time of machine m if there are capacity losses due to re-qualification procedures, *e.g.* due to maintenance operations. Constraints (4.6) link the upper-level and lower-level optimization problems. Constraints (4.7) are the binary constraints for the re-qualification decisions.

Lower-level optimization problem. The objective function (4.8) of the lower-level optimization problem consists in maximizing the utilization balance and minimizing the total utilization rate of the machines. Constraints (4.9) ensure that all quantities must be assigned to qualified and available machines. Operations with a production capacity interruption, i.e. such that all qualified machines are down or if there is no qualified machine, are not assigned to machines. Constraints (4.10) and (4.11) defines the utilization rate of each machine in the work center. Constraints (4.12) and (4.13) ensure that the workload can only be assigned to machine m if operation r is qualified on machine m. Constraints (4.14) ensure that if operation r is not qualified and cannot be qualified on machine m, then the workload for operation r will never be assigned to machine m.

Note that $\mathbf{C_{eff}}$ and $\mathbf{OQ}$ are not decision variables of the lower-level optimization problem (4.8)-(4.15). $\mathbf{C_{eff}}$ and $\mathbf{OQ}$ are parameters of the lower-level optimization problem as they are decided in the upper-level optimization problem.

4.2.3 Multi-period bilevel optimization model

The multi-period bilevel optimization model is proposed below.

Upper-level optimization problem:

$$\max \quad \sum_{r,t} TH_{t,r} \tag{4.16}$$

$$\text{s.t.} \quad \sum_{r,m} OQ_{r,m} = k \tag{4.17}$$

$$I_{t,r} = I_{t-1,r} + d_{t,r} - TH_{t,r} \qquad \forall t > 1, \forall r \tag{4.18}$$

$$I_{1,r} = I_{0,r} + d_{1,r} - TH_{t,1} \qquad \forall r \tag{4.19}$$

$$D_{t,r} = I_{t-1,r} + d_{t,r} \qquad \forall t > 1, \forall r \tag{4.20}$$

$$D_{1,r} = I_{0,r} + d_{t,r} \qquad \forall r \tag{4.21}$$

$$TH_{t,r} = f(\mathbf{U}, \mathbf{WIP}) \qquad \forall t, \forall r \tag{4.22}$$

$$Q_{1,r,m} = OQ_{r,m} \qquad \forall r, \forall m \mid l_{r,m} = 0, q_{r,m} = 2 \tag{4.23}$$

$$Q_{1,r,m} = 0 \qquad \forall r, \forall m \mid l_{r,m} > 0, q_{r,m} = 2 \tag{4.24}$$

$$Q_{t,r,m} = Q_{t-1,r,m} + OQ_{r,m} \qquad \forall t > 1, \forall r, \forall m \mid 1 + l_{r,m} = t, q_{r,m} = 2 \tag{4.25}$$

$$Q_{t,r,m} = Q_{t-1,r,m} \qquad \forall t > 1, \forall r, \forall m \mid 1 + l_{r,m} \neq t, q_{r,m} = 2 \tag{4.26}$$

$$Q_{t,r,m} = q_{r,m} \qquad \forall t, \forall r, \forall m \mid q_{r,m} \neq 2 \tag{4.27}$$

$$C^{eff}_{1,m} = \max(c_{1,m} - \sum_{r} c^{loss}_{r,m} OQ_{r,m}, 0) \qquad \forall r, \forall m \tag{4.28}$$

$$C^{neg}_{1,m} = \min(c_{1,m} - \sum_{r} c^{loss}_{r,m} OQ_{r,m}, 0) \qquad \forall r, \forall m \tag{4.29}$$

$$C^{eff}_{t,m} = \max(c_{1,m} + C^{neg}_{t-1,m}, 0) \qquad \forall t > 1, \forall r, \forall m \tag{4.30}$$

$$C^{neg}_{t,m} = \min(c_{2,m} + C^{neg}_{t-1,m}, 0) \qquad \forall t > 1, \forall r, \forall m \tag{4.31}$$

$$U_{t,m}, WIP_{t,r,m} \in \arg\min LBP_t(\mathbf{D_t}, \mathbf{Q_t}, \mathbf{C^{eff}}) \qquad \forall t \tag{4.32}$$

$$OQ_{r,m} \in \{0, 1\} \qquad \forall r, \forall m \tag{4.33}$$

Lower-level optimization problem:

$$LBP_t(\mathbf{D_t}, \mathbf{Q_t}, \mathbf{C^{eff}}) = \min \quad \sum_{m} U^{\gamma}_{t,m} \tag{4.34}$$

$$\text{s.t.} \quad \sum_{m} WIP_{t,r,m} = D_{r,t} \qquad \forall r \in \mathcal{R}^q_t(\mathbf{Q}, \mathbf{C^{eff}}) \tag{4.35}$$

$$U_{t,m} = \sum_{r} \frac{WIP_{t,r,m}}{tp_{r,m} C^{eff}_{t,m}} \qquad \forall m \mid C^{eff}_{t,m} > 0 \tag{4.36}$$

$$U_{t,m} = 0 \qquad \forall m \mid C^{eff}_{t,m} = 0 \tag{4.37}$$

$$WIP_{t,r,m} \leq D_r \qquad \forall r, \forall m \mid Q_{t,r,m} = 1 \tag{4.38}$$

$$WIP_{t,r,m} \leq 0 \qquad \forall r, \forall m \mid Q_{r,m} \neq 1 \tag{4.39}$$

$$WIP_{t,r,m} \geq 0 \qquad \forall r, \forall m \tag{4.40}$$

Upper-level optimization problem. The objective function (4.16) consists in maximizing the throughput over the horizon. The throughput is computed from the uti-

lization rates of machines determined by the lower-level optimization problem (see Chapters 2 and 3). Constraint (4.17) sets to k the number of qualifications that must be performed at the beginning of the horizon. Constraints (4.18)-(4.19) are flow conservation constraints. Constraints (4.20)-(4.21) compute the demand for all operations and all periods from the current number of products in the work center and new arriving products. Constraints (4.22) compute the throughput from the utilization rates of the machines determined by the lower-level optimization problem (see Chapter 2). Constraints (4.23)-(4.27) determine the new state of each qualification at each period for all operations and machines from re-qualification decisions made at $t = 0$ and re-qualification delays. Constraints (4.23)-(4.26) concern the qualifiable pairs (operation r, machine m), *i.e.* such that $q_{r,m} = 2$, while Constraints (4.27) guarantee that the qualification status of the other pairs (operation r, machine m), *i.e.* such that $q_{r,m} = 0$ or $q_{r,m} = 1$, remains the same throughout the planning horizon. Constraints (4.23) and (4.25) ensure that machine m becomes qualified for operation r as soon as the re-qualification lead time $l_{r,m}$ is reached. Constraints (4.24) and (4.26) ensure both that (1) machine m is not qualified for operation r before its lead time and that (2) machine m remains qualified for operation r in the planning horizon once it has been qualified. Constraints (4.28)-(4.31) ensure the effective capacity of machine m at period t if a re-qualification requires a maintenance operation. Constraints (4.32) link the upper-level and lower-level problems. Finally, Constraints (4.33) are the binary constraints for the qualification decisions.

Lower-level optimization problem: The objective function (4.34) of the lower-level optimization problem consists in maximizing the utilization balance and minimizing the total utilization rate of the machines. Constraints (4.35) ensure that all quantities must be assigned to qualified and available machines. Line stop operations are not assigned to machines. Operations with no available machines are not assigned. For these operations, $TH_{t,r}$ is equal to zero, and $I_{t,r}$ necessarily increases. Constraints (4.36) and (4.37) compute the utilization rate of each machine. Finally, Constraints (4.40) are the non-negativity constraints.

Note that $\mathbf{C_{eff}}$, $\mathbf{Q}$, $\mathbf{OQ}$, and $\mathbf{D}$ are not decision variables of the lower-level optimization problem (4.34)-(4.40). $\mathbf{C_{eff}}$, $\mathbf{Q}$, $\mathbf{OQ}$, and $\mathbf{D}$ are parameters of the lower-level optimization problem as they are decided in the upper-level optimization problem.

In the multi-period optimization approach, if the entire capacity loss due to the qualification cannot only be attributed to the first period, then the remaining capacity loss is attributed to the next period, until there is no capacity loss left. For instance, this can happen if the maintenance operations lasts 12 hours whereas a period lasts 8 hours. For delays, we proceed in a similar way. As re-qualifications are performed at the beginning of the planning horizon, *i.e* at $t = 0$, the qualification matrix is updated at period t if $1 + L = t$. From period t and for the rest of the planning horizon, the machine is re-qualified for the operation. Finally, for the multi-period optimization model, the lower-level optimization problem is solved for each period of the planning horizon. All wafers that cannot be processed at period t are backlogged at period $t + 1$.

Furthermore, the difference between the single-period and multi-period approaches is that, in the multi-period approach, each operation queue in front of

machines is re-evaluated in each period to better consider operation priorities and backlogging. In addition, the single-period bilevel optimization model assumes that all wafers are ready to be processed at the beginning of the period, which can be unrealistic. The multi-period bilevel optimization model alleviates this assumption to a certain extent.

For the sake of brevity, only the multi-period bilevel optimization approach with re-qualification decisions is presented in this section. Nevertheless, for the sake of completeness, the multi-period bilevel optimization approach with re-qualification and disqualification decisions is presented in Appendix C, Section C.1.

4.3 Computational study

The computational study is performed to study if considering dynamic WIP quantities, *i.e.* a demand that varies with time, and production capacities affect the choice of re-qualifications. And if it does, to what extent. In Section 4.3.1, we briefly describe the industrial instances used to perform the computational study. For more details on industrial instances, we refer the reader to Chapter 2. In Section 4.3.2, the design of experiments of the computational study is described. Finally, in Section 4.3.3, numerical results associated to the computational study are presented.

4.3.1 Instance generation

The single-period and multi-period optimization models are compared on industrial data extracted from a 300 mm High Mix (HMLV) wafer fab located in Crolles, France. The wafer fab is characterized by shifting bottleneck work centers, short product life cycles, frequent product mix changes, a high production variability with frequent disqualifications, very high utilization rates of machines and strong machine dedication constraints.

Data were extracted on 15 different weeks in 2018 and 2019. 60 industrial instances are used to compare both optimization models on four different work centers.

Work center A is characterized by a large number of different operations. An operation in work center A can have very different throughput rates from one machine to another. Machines in work centers A and B are cluster machines. In work center C, some machines process batches of lots, while the other machines process lots wafer by wafer. In work center D, machines process lots wafer by wafer. In general, most machines allow several lots to be processed at the same time.

4.3.2 Design of experiments

Table 4.1 presents the design of experiments. We did not run experiments where qualification procedures simultaneously require maintenance operations and are subject to delays. This is left for future research. We limit ourselves to $k = 1$ so that we can study and compare the optimal solution of both optimization models. Both optimization models are studied on a 24-hour planning horizon. To solve the lower-level utilization balancing problem, a cutting plane algorithm (see Chapter

2) is used. γ is set to 6. The algorithm stops when a relative gap of 0.00001 is reached. All experiments are run using Java 8 and CLP Java (Lougee-Heimer, 2003) and Löhndorf (2016) as the linear solver for solving the cutting plane algorithm.

To search for the best re-qualification, all re-qualifications are tested separately. More precisely, for each possible re-qualification, the input qualification matrix $q_{r,m}$ is first modified. Then for each period, the actual demand for each operation is computed, then the lower-level utilization balancing problem is solved, and then the upper-level problem is solved to compute the throughput.

Table 4.1: Design of experiments.

Parameters	Values
Lead time (in 8-hour shifts)	0, 1, 2
Capacity loss $c_{loss}^{r,m}$ (in hours)	0, 4, 8, 12
T (in 8-hour shifts)	3
Number of qualifications k	1
Simulated dispatching	FIFO, Priority
Work center	A, B, C, D
Optimization model	Single-period, Multi-period

4.3.3 Numerical results

As the single-period model does not model delays and as the objective function of the single-period and multi-period optimization models is not computed the same way, to have a fair comparison between both optimization models, when the single-period model is solved, the best re-qualification is retrieved and is used to compute the qualification matrix of the multi-period optimization model, which is used to compute the throughput. Both approaches can then be compared in a fair way. In the rest of this section, the single-period optimization approach is denoted SP, and the multi-period optimization approach is denoted MP. In addition, "base case" refers to the case where a re-qualification does not require maintenance operation or is not subject to a lead time.

4.3.3.1 Capturing dynamic WIP quantities and production capacities

Table 4.2 compares the mean gap(%) $= 100 \times \frac{MP-SP}{MP}$, in terms of throughput between the single-period (SP) and multi-period (MP) approaches. Table 4.2 enables us to assess if the single-period optimization approach is able to capture the dynamic WIP and capacity. Numerical results show that the single-period optimization model can lead to less relevant re-qualification decisions. The largest mean and maximum gaps are observed for the work center A when operation priorities are considered. Even without any lead time or capacity loss, the mean gap is of 1.50%. For the work center A, this indicates that the single-period optimization model does not always capture dynamic WIP quantities. It proposes to qualify a operation with higher demand on average whereas higher gains can be achieved by focusing on operations with high peaks of demand on certain shifts. For the work center D, the mean gap

is of 0.42%. For other work centers, mean gaps are closer. Nevertheless, maximum gaps are always greater than 0.6%.

When there are capacity losses and operation priorities are considered, mean and maximum gaps are significant. The highest mean gap, 2.36%, is observed for the work center A. The highest maximum gap, 5.02%, is also observed for the work center A. The maximum gap for the work center B is about 4.36% when an 8-hour maintenance operation is required. The maximum gap for the work center D is equal to 1.62%. The maximum gap for the work center C is equal to 1.39%. Overall, mean gaps are always greater than 0.46%. Mean and maximum gaps are smaller when re-qualifications are subject to lead time than when they require maintenance operations. This can be surprising because the single-period optimization model does not consider delays. However, this can be explained by the fact that delays do not interrupt production contrary to maintenance operations.

Table 4.2: Mean and maximum gaps (%) $= 100 \times \frac{MP-SP}{MP}$, in terms of throughput between the single-period (SP) and the multi-period (MP) optimization models.

				Lead time (in shifts)				Capacity loss (in hours)					
		Base case		1		2		4		8		12	
Throughput	Work center	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max
Priority	A	1.50	4.65	0.77	1.97	0.79	2.38	2.32	4.96	2.29	5.58	2.36	5.02
	B	0.16	0.60	0.24	2.32	0.31	1.70	0.46	2.15	0.72	4.36	0.51	4.30
	C	0.23	0.68	0.08	0.41	0.11	0.27	0.53	1.67	0.46	1.35	0.48	1.39
	D	0.42	1.58	0.23	0.76	0.16	0.67	0.66	1.95	0.82	1.78	0.63	1.62
FIFO	A	0.05	0.65	0.15	1.22	0.20	1.43	0.10	0.82	0.10	0.29	0.05	0.42
	B	0.01	0.15	0.01	0.15	0.06	0.56	0.00	0.00	0.00	0.00	0.01	0.08
	C	0.00	0.04	0.02	0.22	0.01	0.10	0.00	0.04	0.00	0.18	0.00	0.06
	D	0.00	0.01	0.05	0.37	0.09	0.40	0.01	0.09	0.01	0.19	0.01	0.15

In addition, we can observe that gaps are also smaller when a FIFO queue is considered. This can be due to the fact that the backlogged quantities are less variable contrary to when operation priorities are considered. Table 4.3 reinforces this idea. When considering a FIFO, both optimization models propose much more frequently the same re-qualification than when operation priorities are considered. For instance, for an 8-hour capacity loss, both optimization models propose eight times the same re-qualification plan when a FIFO queue is considered, and only twice when operation priorities are considered. Operation priorities are then a source of production variability for qualification management but must be considered.

Table 4.2 also shows that mean gaps are often far from maximum gaps. For example, for the work center B, when there is a 4-hour capacity loss, the mean gap is equal to 0.46% whereas the mean gap is equal to 2.15%. This is something that we can observe for all work centers, in particular when maintenance operations are required. Moreover, as on a non-negligible amount of instances, both optimization models propose the same re-qualification (see Table 4.3), this indicates that there exists, even among the same work center, a large disparity between instances. There are instances where the gap between both optimization models is equal to zero or very small whereas other instances where the gap is very large.

Finally, Table 4.4 shows the mean gain (%) on the throughput after performing a re-qualification. As mean gaps between both optimization models can be large (Ta-

Table 4.3: Number of identical re-qualification plans (out of 15) recommended by both optimization models.

			Lead time (in shifts)		Capacity loss (in hours)		
Throughput	Work center	Base case	1	2	4	8	12
Priority	A	4	3	1	1	2	0
	B	5	11	7	7	10	12
	C	8	8	3	5	2	1
	D	7	4	2	3	2	0
FIFO	A	12	10	8	11	8	8
	B	14	13	9	11	11	11
	C	11	10	8	11	8	9
	D	10	8	6	10	7	8

ble 4.2), in general, the multi-period optimization model more frequently proposes re-qualifications that capture dynamic WIP quantities. When operation priorities are considered, the mean gain can be quite different between both optimization models. For instance, in the work center A, when there is a 4-hour capacity loss, the single-period optimization model proposes a re-qualification plan that leads to a diminution of the throughput by -1.60%. Instead, the multi-period optimization model proposes a re-qualification that leads to an increase of the throughput by 0.74%! This situation is observed for most work centers when there is capacity loss. There is only in the work center C where the single-period optimization model with a 12-hour maintenance operation does not induce a negative mean gain. However, the mean gain is equal to 0.07%, which is very small, compared to the mean gain of the multi period optimization model that is equal to 0.56%. We also observe that for a 12-hour maintenance operation, the mean gain of the multi-period optimization model is negative for the work center A. However, the mean gain in almost ten times worse with the single-period optimization model (-0.29% versus -2.65%). Overall, the multi-period optimization model proposes re-qualification plans that achieve better mean gain than the single-period optimization model. This means that the single-period optimization model can propose *wrong* re-qualification decisions. When re-qualifications are only subject to delays, mean gains are closer. However, they remain significant for the work center A with a difference greater than 0.7%. When a FIFO queue is considered, mean gains are *very* close. In other words, if a FIFO queue is the method used to simulate dispatching decisions and if re-qualification decisions are assumed to have no lead time and induce no capacity loss, then a multi-period bilevel optimization model is not relevant in most cases.

4.3.3.2 Influence of/on dispatching rules

Table 4.2 shows that although the capacity loss/lead time increases, the mean gap does not necessarily increases. Table 4.5 shows that for some instances, performing a re-qualification with a lead time greater than zero better maximizes the throughput than performing a re-qualification with no lead time. These results seems

Table 4.4: Comparison of the mean gain(%) on the throughput after performing a re-qualification between the single-period (SP) and multi-period (MP) optimization models. Bold values are negative mean gain.

				Lead time (in shifts)				Capacity loss (in hours)					
		Base case		1		2		4		8		12	
Throughput	Work center	MP	SP	MP	SP	MP	SP	MP	SP	MP	SP	MP	SP
Priority	A	2.04	0.51	1.30	0.52	1.14	0.33	0.74	**-1.60**	0.21	**-2.09**	**-0.29**	**-2.65**
	B	0.88	0.72	1.02	0.78	0.89	0.57	0.55	0.09	0.50	**-0.22**	0.42	**-0.09**
	C	0.79	0.56	0.46	0.38	0.31	0.21	0.68	0.15	0.66	0.20	0.56	0.07
	D	1.53	1.11	0.78	0.54	0.45	0.29	1.08	0.42	0.49	**-0.34**	0.39	**-0.24**
FIFO	A	1.57	1.52	1.38	1.23	1.01	0.81	1.32	1.22	0.94	0.91	0.55	0.50
	B	0.63	0.62	0.63	0.62	0.52	0.46	0.61	0.61	0.53	0.53	0.47	0.47
	C	0.58	0.58	0.55	0.54	0.37	0.36	0.52	0.51	0.42	0.39	0.31	0.31
	D	1.16	1.16	0.93	0.88	0.61	0.52	0.97	0.96	0.76	0.74	0.52	0.50

counter-intuitive. Actually, this effect is due to the way the throughput is computed, and more generally, how the production system works with dispatching rules. When lots arrive in an work center, a dispatching engine assigns lots on machines. The dispatching engine is shortsighted. It does not consider lots that arrive one or two shifts later. In addition, it does not necessarily challenge previous dispatching decisions made when a new lot arrives. This means that, if an operation is qualified on a machine, the dispatching engine will take advantage of the re-qualification and assign lots to the machine. Thus, if this re-qualification decision is taken right now for an operation with longer process times than those already qualified on the same machine, the throughput per shift slightly decreases due to the fact the average throughput rate on that machine decreases. The magnitude of this effect varies with WIP variability over time and if priorities are considered. This effect is also observed in (Gurumurthi and Benjaafar, 2004; Johnzén et al., 2008) where numerical experiments are run to assess the effect of a re-qualification on the mean cycle time. After qualifying machines, the mean cycle time did not necessarily decrease. A similar explanation is also detailed in (Johnzén et al., 2008). Therefore, re-qualifications influence dispatching rules decisions, and dispatching rules also influence re-qualification decisions.

4.4 Practical insights

Numerical results highlight the fact that proposing the best re-qualifications is a complex procedure, and that performing the re-qualifications at the right time is critical to improve the throughput. Re-qualification decisions are influenced by WIP and capacity variability over time but also by decision maker preferences or dispatching rules and priorities. In addition, numerical experiments show that performing a re-qualification may lead to uncompensated capacity losses, *e.g.* due to required maintenance operations. Thus, after performing a re-qualification, the throughput can be lower than in the case where no re-qualification is performed. Instead of only considering the throughput, other indicators might be interesting to assess the quality of a re-qualification by for example prioritizing lots with large pri-

Table 4.5: Number of instances by work center where performing a re-qualification with lead time gives a larger throughput than performing qualification with no lead time.

		Lead time (in shifts)			
		1		2	
Throughput	Work center	MP	SP	MP	SP
Priority	A	5	7	5	6
	B	5	6	5	4
	C	0	4	0	3
	D	2	6	1	6
Average product mix	A	0	1	1	2
	B	1	1	0	0
	C	5	4	1	2
	D	4	2	3	3

orities. For instance, although the overall throughput decreases, if the mean cycle time of priority lots also decreases, then a re-qualification can be acceptable. Since maximizing the throughput is not always the best option, qualification management could also therefore be modeled and solved as a multi-objective problem.

Numerical results highlight the fact that the single-period and multi-period optimization models can propose different re-qualifications respect to the demand profile of the operations. Depending on the demand profiles, a model is more appropriate than the other. In general, in work centers where lots come by wave, the multi-period optimization model should be more suited because it better captures demand peaks. This model is then useful to identify and fix short-term bottlenecks with cross qualifications. It is also more robust again highly variable demand and capacity profiles.

Another significant of the multi-period optimization model is that, it considers several periods over the horizon, it also leads to better estimate of the mean cycle time spent by lots in the work center (see Chapter 5). This means that, even if a FIFO queue is the method used to simulate dispatching decisions and if re-qualification decisions are assumed to have no lead time and induce no capacity loss, then a multi-period bilevel optimization model may still in fact be relevant because it better estimates the mean cycle time spent by lots in the work center although there is no large difference with the single period optimization model to estimate the throughput

Numerical results also show that dispatching rules significantly affect the quality of a re-qualification plan (Gurumurthi and Benjaafar, 2004; Johnzén et al., 2008; Kopp et al., 2019). Therefore, how the lots are scheduled should be considered in qualification management, in particular in operational qualification management.

As we study a high mix production facility subject to high production variability, the demand and capacity can be uncertain. Therefore, it can be preferable to perform a re-qualification that requires no lead time or maintenance operation and looks sub-optimal, rather than perform an “optimal” re-qualification with a larger

lead time or longer maintenance operation. If all re-qualification decisions are subject to delays or maintenance operations, shortest ones should be preferred. In addition, uncertainty can be managed by using a rolling horizon approach (Clark and Clark, 2000; Curcio et al., 2018). A re-qualification plan is determined at the beginning of the first shift of the planning horizon. At the beginning of the next shift, new information is revealed, the optimization model is solved and a new re-qualification plan is determined.

4.5 Conclusions and perspectives

In this chapter, a single-period optimization model and a multi-period optimization model are compared to maximize the throughput with re-qualification plans. Dispatching rules are included and simulated in optimization models. The dynamic qualification optimization model is used to better consider qualification delays and maintenance operations. Numerical experiments on industrial data show the relevance of the multi-period qualification optimization model. In particular, numerical experiments show that the choice of the model can have a significant effect on the re-qualification plan, and therefore on the mean gain in terms of throughput. The mean gain is particularly affected when operation priorities are considered and a maintenance operation is required to re-qualify operations on machines. However, the single-period optimization model remains relevant for some instances. In addition, in Chapter 5, we further demonstrate that a multi-period optimization model is better suited to propose more relevant re-qualifications in terms of mean cycle time.

We believe the following perspectives are worth investigating in the future (out of the scope of the study):

1. We limit ourselves to $k = 1$. Efficient and effective solution approaches can be designed to propose re-qualification plans for large values of k for both optimization models. The solution approaches presented in Chapter 2 could be extended to solve the single-period and multi-period bilevel optimization problems. For instance, preliminary results showed that combining the use of dual variables and other preprocessing rules (only re-qualifications on machines with utilization rates lower than 1 or only re-qualifications of operations that are faster on currently disqualified machines than currently qualified machines) allows computational times to be significantly reduced while only slightly impacting the quality of the re-qualification plan, in particular when FIFO queues are simulated.

2. Similarly to Chapters 2 and 3, it would be interesting to better understand the impact of different γ settings on re-qualification decisions and computing time.

3. New methods to simulate the throughput or dispatching and scheduling decisions could be proposed to be closer to the real behavior of the dispatching and scheduling engines, *e.g.* by including batching constraints when relevant

(Rowshannahad and Dauzère-Pérès, 2013) or using machine learning techniques.

4. Our numerical experiments show that the single-period optimization model often proposes the same re-qualifications as the multi-period optimization model on industrial data. It would be interesting to automatically identify when the single-period optimization model is likely to suggest the same re-qualifications. Doing this would save a lot of time when searching for re-qualifications.

Chapter 5

Evaluating the impact of re-qualifications on cycle times

In Chapters 2, 3 and 4, the question "How to determine the most relevant re-qualifications to improve operational efficiency?" is answered from a utilization balancing standpoint or from a throughput standpoint. In this chapter, we are interested in evaluating the effect of re-qualification decisions in terms of cycle time in a work center. In particular, we show that it is possible to have closed-formed solutions for cycle time forecasts based on simple assumptions. In addition, we show that re-qualifications that maximize the throughput are not necessarily the re-qualifications that minimize the mean cycle time due to production variability.

5.1 Introduction

In this chapter, we evaluate the effect of a re-qualification on the mean cycle time of lots in a work center. First, we show that it is possible to have closed-formed solutions for cycle time forecasts based on simple assumptions, in particular on the fact that work centers are unlikely to be empty, *i.e.* without any lot to process, and on deterministic laws for lot arrivals and departures from a work center. Depending on the work center and the length of the horizon, we show that cycle time forecast errors can be smaller than 5%. Second, by using the closed-formed solutions, we show that, due to production variability, re-qualifications that maximize the throughput are not necessarily the re-qualifications that minimize the mean cycle time.

The remainder of this chapter is organized as follows. In Section 5.1.1, we review related work on qualification management in semiconductor manufacturing to control or minimize the mean cycle time. After reviewing the literature, our contributions are explained. In Section 5.2, we argue that closed-form solutions are available at a production control level, *i.e.* at an operational decision level, for a work center manager. This is because closed-form solutions can be derived by assuming deterministic arrivals and departures and are realistic. Two closed-form solutions are then derived. One closed-form solution is close the one proposed by Leachman (2015). In Section 5.3, we show the limits and the relevance of closed-form solutions for short-term cycle time forecasts for different work centers on industrial data. In Section 5.4, we show the effect of re-qualification decisions on the short-term mean cycle time with derived closed-forms on industrial data. Note that the effect of disqualification decisions could also be illustrated but is not performed in this chapter for space constraints. In Section 5.5, we provide recommendations for production personnel for the management of re-qualifications with respect to the mean cycle time. We also argue that closed-form solutions can be used for different goals than qualification management. Examples are provided. Finally, in Section 5.6, we conclude and give perspectives.

5.1.1 Related work

Most papers that deal with qualification management and cycle time use optimization models to, first, determine qualifications in terms of production costs, throughput or workload balancing, and then, use simulation models to evaluate the mean cycle time (Akcalt et al., 2001; Ignizio, 2009, 2010; Kabak et al., 2013; Kopp et al., 2018, 2019). In other words, qualifications are often not directly optimized in terms of mean cycle time. Another computationally cheap alternative consists in using queuing theory (Shanthikumar et al., 2007) to evaluate the mean cycle time.

An alternative to simulation to evaluate the mean cycle time is to use queuing theory, which has the benefit of being computationally inexpensive compared to simulation models. Nevertheless, queuing theory has given unsatisfactory results (Shanthikumar et al., 2007), in particular in a predictive use. In addition, queuing theory can be intractable if embedded in optimization models as it involves highly non linear equations. However, using queuing theory to control, reduce or estimate decisions on the mean cycle time is a well developed practice in the semiconductor

industry. Aurand and Miller (1997) and Brown et al. (2010) show how queuing theory is used at IBM to manage the mean cycle time. Similarly, Kalir and Bouhnik (2006) and Li et al. (2007) show practices at Intel Corporation to manage the mean cycle time with queuing theory. Sattler (1996), Potti and Whitaker (2003) and Schelasin (2011) show practices related to queuing theory at Texas Instrument. Although queuing theory has given unsatisfactory results in terms of predicting the mean cycle time, applying queuing theory principles, *i.e.* improving workload balancing, reducing production variability, and increasing the number of qualifications, have shown to decrease the mean cycle time. In addition, recent works (see *e.g.* Delp et al. 2006; Morrison and Martin 2006, 2007) have shown on industrial data that if queuing theory had given unsatisfactory results in a predictive use, this may be because many factors contributing to increasing the mean cycle time of lots in semiconductor manufacturing have been not considered in classical G/G/m queues. In particular, Morrison and Martin (2006) and Morrison and Martin (2007) show that including the mean cycle time offsets, *e.g.* transportation times, hold times, idle with WIP, defections of lots from a failed machine, and a better modeling of parallel processing offered by some machines can lead to an accurate estimation of the mean cycle time in a work center.

From a qualification management standpoint, queuing theory is particularly convenient because it provides a closed-form solution between the utilization rates of machines, the number of machines in the work center, and the mean cycle time of lots in a work center. Therefore, the mean cycle time of lots can be decreased simply by assessing decisions using spreadsheets. Fowler et al. (1997) show that adding new qualifications to machines can lead to substantial reduction of the mean cycle time from queuing theory. Fowler et al. (1997) compare different dispatching policies with different qualification rates. They validate with simulation that adding new qualifications strongly reduce the mean cycle time. Similarly, Leachman (2012) illustrate how the mean cycle time can be reduced when new qualifications are added by using queuing theory.

Although G/G/m queues have been applied with success on some work centers (Morrison and Martin, 2006, 2007), they may not be applied with success for all work centers. This is because a fundamental assumption of G/G/m is that there are m identical machines in parallel able to process a lot, which is not always the case. Machines can be of different generations, thus not identical in terms of throughput rates and qualifications. G/G/m queues also imply that machines must be qualified for all operations. However, the variability on the number of qualified machines can actually be large from one operation to another in a work center. This is because old machines do not necessarily process the same layers or the same products as newly purchased and installed machines (Miltenburg et al., 2002). In addition, qualifications can be lost due to maintenance operations, recipe validation problems, yield losses, parametric issues (see *e.g.* Kopp et al. 2018) or WIP management policies. Similarly to the variability in arrival rates and process times, the variability in machine qualifications could negatively affect the mean cycle time (Sattler, 1996; Hopp and Spearman, 2011). The variability in the number of qualified machines by operation could be one major reason why Shanthikumar et al. (2007) state that queuing theory has unsatisfactory results in a predictive use.

Work centers are actually flexible queuing systems where lot types, *e.g.* operations or recipes, do not queue the same machines (see Figure 5.1). To better consider flexible queuing systems, a possibility is to artificially decompose a large work center into smaller work centers that do not share qualifications with other work centers so that G/G/m holds for smaller work centers. However, this is not always possible, in particular for high mix low volume factories. Instead, Sattler (1996) and Juang and Huang (2000) propose to redefine the number of machines m as an effective number of machines $m*$. Veeger et al. (2010) generate queuing models with simulation based on the notion of effective process times. In their model, the number of machines m is a parameter that can be fitted to better approximate the mean cycle time.

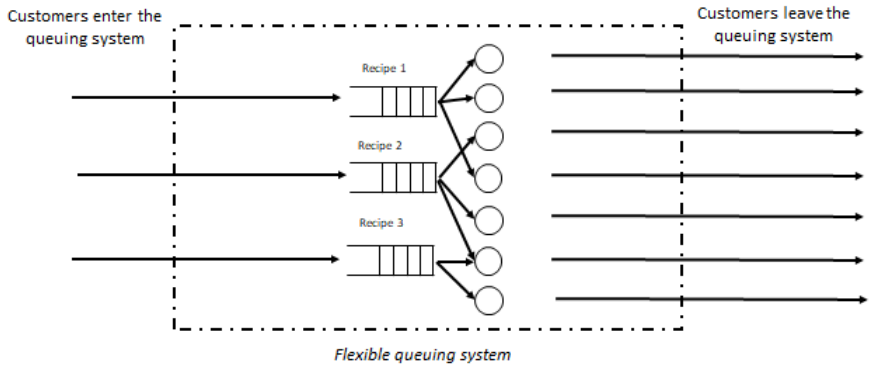


Figure 5.1: Flexible queuing system.

5.1.2 Contributions

In this chapter, we are interested in evaluating the effect of re-qualifications on the mean cycle time in a work center. Simulation models are prohibitive because they require expensive computational times. To the best of our knowledge, the assessment of re-qualification decisions on short term cycle times does not exist in the literature. In addition, evaluating the effect of re-qualification decisions on the mean cycle time, a relevant cycle time measurement, must be derived at an operational level for a work center in order to determine relevant re-qualifications. Therefore, our contributions are three-fold:

- Although there are papers on short-term cycle time forecasts using machine learning techniques (see *e.g.* Can and Heavey 2016; Wang et al. 2017), we show that for short-term horizons, relevant simple closed-form solutions are available to model the mean cycle time.
- The proposed closed-form solutions are studied in a computational study on industrial data. In particular, the limits of the closed-formed solutions are shown.
- We show the closed-form solutions can be used by production personnel to assess the relevance of short-term re-qualification decisions in terms of mean cycle time.

5.2 Closed-form solutions for cycle time modeling

In this section, relevant simple closed-form solution are derived to model short-term cycle times. In particular, closed-form solutions are derived by assuming deterministic laws for the arrivals and departures of products in work centers, which is motivated by practical experience.

5.2.1 Motivations

Assume that the mean cycle time of lots in front a machine or a work center for a finite time horizon, *e.g.* a week, must be studied. The system (the machine or the work center) is empty at $t = 0$. The system is assumed to be immediately in steady state. Lot and wafer arrivals follow a deterministic law. The arrival rate is equal to $\frac{1}{\lambda}$. The process time $\frac{1}{\mu}$ follows en exponential distribution and is equal to 300 seconds. Assume that the number of lots and wafers that will arrive over the finite time horizon is known exactly. Then, the utilization rate of the system u is know because $u = \frac{\lambda}{\mu} = \frac{d}{T\mu}$. As the queuing system is described by a D/M/1 queue and in steady state, the mean queuing time (QT) is given by the following equation:

$$QT = \frac{1}{2\mu}\frac{u}{1-u} \tag{5.1}$$

Table 5.1 gives mean queuing times for utilization rates close to one. For $u = 0.9999$, the mean cycle time is equal to 1,499,999,851 seconds, more than the time horizon of a week!

Table 5.1: QT (M/D/1) for utilization rates close to one

Utilization rate (u)	QT(M/D/1) (s)
0.99	14,850
0.999	149,850
0.999	1,499,850
0.9999	1,499,999,851

Such values for the mean queuing time is impossible in practice. As the system is assumed to be empty at the beginning of the time horizon, the number of customers and the process time are known, and the utilization rate is less than one, the mean queuing time cannot exceed T seconds, where T is the length of the time horizon.

This example may be twisted but illustrates what production personnel daily faces at the operational decision level. For a given work center, the demand for the next shift or 24 hours is known and *finite*, and very often, the utilization rate of several machines is very close to one, frequently exceeding one because of machine down times and production variability. As the utilization rate is greater than one, a part of upcoming wafers accumulate, the mean queuing time increases (and it is always defined because finite), but never skyrockets as it is the case in D/M/1

queues. This leads to a time-dependent queuing system that is not modeled in classical queuing models.

The problem consists in applying queuing theory, *e.g.* D/M/1 or G/G/m queues, to a (short-term) finite horizon. Consider a work center with M parallel machines. G/G/m queues make four fundamental assumptions: The work center is in a steady state, the utilization of each machine is less than 1, the population of lots is infinite and there is flow conservation, *i.e.* the arrival rate is equal to the departure rate. At an operation level, these four assumptions do *not* hold because of production variability and finite time horizons. In G/G/m queues, the notion of time horizon does not exist and, at an operational level the population of lots is always finite because the time horizon is finite. For instance, machines are frequently, but temporarily, overloaded, therefore their utilization rate can exceed 1. This does not mean that the classical queuing theory is wrong or cannot be useful to derive relevant decisions to improve the cycle time (see Section 5.1.1). It means that it cannot be used to propose accurate values of the cycle time on the short term, and thus to offer the best ranking of potentially relevant re-qualification decisions.

5.2.2 Deterministic arrivals and departures

Consider a lithography work center. As lithography machines are expensive, up to $100M dollars by unit for the most recent machines, lithography work centers are often bottleneck work centers and it is *extremely* unlikely that all machines are simultaneously idle because the capacity utilization rates of machines are critical indicators for semiconductor factories. In other words, assuming that there are always wafers in lithography work centers is realistic.

As all machines in the semiconductor industry tend to be expensive, it can be reasonable to assume that there is *always* some WIP (Work-In-Process) in any work center in a semiconductor factory. In other words, work centers are never totally empty. Nevertheless, this is not always the same machines that have qualified WIP in front of them, otherwise, the long-term mean cycle time would significantly increase. By assuming that both arrivals and departures follow a *deterministic* law, it is possible to compute and determine a closed-from solution for the mean WIP in a work center over a finite time horizon by using the integral version of Little's law (Little and Graves, 2008; Little, 2011).

We assume that deterministic assumptions for arrivals and departures are reasonable for small horizons, *i.e.* horizons of a few fours to at most 24 hours. This is because the state of each work center can be assumed to be relatively stable over small horizons. For instance, on small horizons, the number of processed products, the number of down machines, the congestion in the transportation system, the qualifications can be assumed to be relatively stable.

5.2.3 First approach

Assume that the mean WIP must be computed on the horizon $[0, T]$, with $N(t)$ the cumulative number of arrivals at time t, $D(t)$ the cumulative number of departures at time t. $N(0)$ is the WIP already in the system at $t = 0$ and $N(T)$ is the total

number of arrivals at the end of the horizon. $D(T)$ is the total number of departures (throughput) by the end of the horizon. These notations will be used throughout the chapter. The mean WIP, $\overline{WIP}$, over $[0, T]$ in each work center can be computed as follows if the work center is never empty:

$$\overline{WIP} = \frac{1}{T}\int_0^T (N(t) - D(t))dt \tag{5.2}$$

Note that $N(t) > D(t)\ , \forall t$ as work centers are never empty. As $N(t)$ and $D(t)$ follow a deterministic law, computing $\int_0^T (N(t) - D(t))dt$ can be done as shown below:

$$\int_0^T (N(t) - D(t))dt = \int_0^T (N(0) + N'(t) - D(t))dt \tag{5.3}$$

where $N(0)$ is the number of wafers in the work center at $t = 0$, and $N'(t)$ the new arrivals by time t. Finally, if λ^{in} is the deterministic (independent of t) arrival rate and λ^{out} is the deterministic departure rate, Equation (5.3) can be rewritten:

$$\begin{aligned}
&\int_0^T (N(0) + N'(t) - D(t))dt = \int_0^T (N(0) + \lambda^{in}t - \lambda^{out}t)dt \\
&\int_0^T (N(0) + \lambda^{in}t - \lambda^{out}t)dt = N(0) + \frac{\lambda^{in}}{2}T^2 - \frac{\lambda^{out}}{2}T^2 \\
&\overline{WIP} = \frac{N(0)T + \frac{\lambda^{in}}{2}T^2 - \frac{\lambda^{out}}{2}T^2}{T} \\
&\overline{WIP} = N(0) + \frac{\lambda^{in}}{2}T - \frac{\lambda^{out}}{2}T \\
&\overline{WIP} = N(0) + \frac{N'(T)}{2} - \frac{D(T)}{2}
\end{aligned} \tag{5.4}$$

Equation (5.4) shows that, under a deterministic assumption for arrivals and departures, $\overline{WIP}$ is actually the arithmetic mean between WIP at $t = 0$ and $t = T$. From $\overline{WIP}$, it is possible to compute the mean cycle time (CT) experienced by lots in the work center with Little's law (Little and Graves, 2008; Little, 2011):

$$CT^{add} = \frac{\overline{WIP}}{\frac{N(T)}{T}} \tag{5.5}$$

Note that as $N(0) > 0$, CT^{add} is not strictly equal to the mean cycle time experienced by lots in the work center. It is actually equal to the *additional* mean cycle time experienced by lots in the work center (Little and Graves, 2008; Little, 2011). Note that as CT^{add} is the *additional* mean cycle time, $CT^{add} \leq T$. In addition, as there exist re-entrant product flows in semiconductor manufacturing, $N(T)$ corresponds to the total number of couples (operation, lot) that will arrive in the work center by the end of the horizon. In other words, Equation (5.5) evaluates the additional mean cycle by lot *and* by operation.

As there are lots in the system at $t = 0$, lots have been waiting and the associated waiting times contribute to the overall mean cycle time:

$$
\begin{aligned}
CT^{overall} &= \frac{1}{N(T)} \sum_{k=1}^{N(T)} (CT_k^{init} + CT_k^{add}) \\
CT^{overall} &= \frac{1}{N(T)} \sum_{k=1}^{N(0)} (CT_k^{init}) + \frac{1}{N(T)} \sum_{k=1}^{N(T)} (CT_k^{add}) \\
CT^{overall} &= \frac{CT^{init}}{N(T)} + \frac{N(T) CT^{add}}{N(T)} \\
CT^{overall} &= \frac{CT^{init}}{N(T)} + CT^{add} \qquad (5.6)
\end{aligned}
$$

where CT_k^{init} is the initial cycle time already spent by lot k in the work center at $t = 0$, CT_k^{add} is the additional cycle time of lot k in the work center, $CT^{init} = \sum_{k=1}^{N(0)} CT_k^{init}$, and $CT^{add} = \frac{1}{N(T)} \sum_{k=1}^{N(T)} CT_k^{add}$.

Equation (5.4) corresponds to the mean WIP for a work center over all its operations with WIP on the horizon $[0, T]$. Equation (5.4) can be used at different aggregation levels. For instance, it can be used to compute the mean WIP for a work center over all its operations with a WIP that belongs to a specific fabrication layer. The additional mean cycle time by fabrication layer on the horizon $[0, T]$ can then be estimated with Equation (5.6). Other aggregation levels are possible, for instance, by product or product family.

Equation (5.6) naturally considers the effect of re-qualifications because the throughput, *i.e.* the number of departures $D(T)$, can be improved with re-qualification decisions (see Chapters 3 and 4). Note that minimizing (5.5) or (5.6) is equivalent because both equations differ by a constant term.

Leachman (2015) proposes in his lecture (page 7) a similar formula, $CT = \frac{\overline{WIP}}{\frac{D(T)}{T}}$, to estimate the mean cycle time, named "throughput time", for a single operation. This is only valid when there is flow conservation, *i.e.* when $\lambda^{in} = \lambda^{out}$, which may hold for large or infinite horizons but is not acceptable for *finite* short-term horizons due to production variability, *e.g.* machine breakdowns. Moreover, in the context of high mix manufacturing, we show that estimating the cycle time by operation is largely inaccurate and statistically irrelevant. This is mainly because the assumption on the non-emptiness of the system, which consists of a set of operations for a given work center, is often violated when only a small number of operations is considered. Numerical results show that estimating the mean cycle time is only possible over a large set of operations at a given work center. It is worth mentioning that Equation (5.4) is used in production planning optimization problems to define clearing functions (see *e.g.* Kacar et al. 2011). Clearing functions are used to model production capacities of work centers. They define a relationship between the number of departures (throughput) from a work center and a workload estimation of the work center, *e.g.* based an evaluation of the mean WIP. Clearing functions are typically fitted from simulation data where the horizon is typically of one week.

We could not find in the literature contributions that show the limits of Equation (5.4) to estimate the mean WIP on industrial data. In particular, we numerically show that Equation (5.4) is not representative of the mean WIP in a work center when the horizon exceeds one day. Equation (5.4) cannot be applied for large planning horizons but can be refined for large horizons by dividing them in smaller time periods (see Section 5.2.4).

5.2.4 Second approach

For "large" horizons, the deterministic assumption may no longer be realistic due to production variability. This is because machines fail, which leads to time-varying arrivals and departures. Moreover, new products are introduced, new machines are installed, and qualifications are updated over time, which also have the effect of creating time-varying arrivals, processing times and departures. Nevertheless, it is possible to relax the deterministic assumption by assuming that large horizons can be divided into smaller periods of time where the deterministic assumption still holds in each period. In other words, each period has its own deterministic arrival and departure rates that can be different from other periods. For instance, consider the illustrative example in Figure 5.2. A horizon of one day can be divided into three periods of eight hours each. The second approach therefore is used to better capture the dynamic of work centers in terms of mean cycle times, which can result in better decision making.

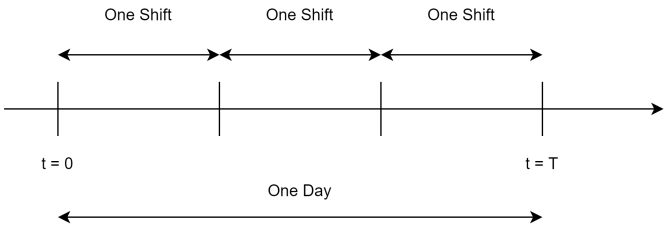


Figure 5.2: Dividing the horizon.

By dividing the horizon into smaller periods, it is still possible to use equation (5.4) to compute the mean WIP by period and finally compute the overall mean WIP, $\overline{WIP}$, over the horizon. First, for sake of simplicity, consider that the horizon, as in Figure 5.2, is divided into three periods. Let us define $t1$ as the end date of the first time period and the start date of the second time period, $t2$ as the end date of the second time period and the start date of the third time period. The mean WIP of each period can be computed as follows:

$$\overline{WIP_1} = \frac{1}{(t1-0)} \int_0^{t1} (N(t) - D(t))dt$$

$$\overline{WIP_2} = \frac{1}{(t2-t1)} \int_{t1}^{t2} (N(t) - D(t))dt$$

$$\overline{WIP_3} = \frac{1}{(T-t2)} \int_{t2}^{T} (N(t) - D(t))dt$$

(5.7)

By additivity of integrals, WIP can be computed as follows:

$$\overline{WIP} = \frac{1}{T}\int_0^T (N(t) - D(t))dt$$

$$\overline{WIP} = \frac{1}{T}[\int_0^{t1} (N(t) - D(t))dt + \int_{t1}^{t2} (N(t) - D(t))dt + \int_{t2}^{T} (N(t) - D(t))dt]$$

$$\overline{WIP} = \frac{1}{T}[(t1 - 0)\overline{WIP_1} + (t2 - t1)\overline{WIP_2} + (T - t2)\overline{WIP_3}] \quad (5.8)$$

WIP_1, WIP_2 and WIP_3 can each be computed with Equation (5.4). Note that, if the horizon is divided into periods of equal duration, the overall mean WIP is simply equal to the sum of the mean WIP over all periods divided by the number of periods. Equation (5.8) is expected to better capture dynamic WIP quantities and production capacities than Equation (5.4) as the horizon is divided in smaller periods of time.

5.2.5 Illustrative example

Consider the illustrative example shown in Tables 5.2 and 5.3. The horizon is made of twelve periods of equal duration. The initial WIP in the work center is equal to 5,000 wafers. The production capacity of the work center is constant and is equal to 2,000 wafers per period. The total number of arrivals, including the initial WIP, is equal to 24,000 wafers. The total number of departures is also equal to 24,000 wafers. In the illustrative example, two different cases are compared in terms of arrival profile (see Table 5.2). In case 1, the arrival profile is constant over time, *i.e.* at each period 2,000 wafers arrive. In case 2, the arrival profile strongly varies over time, a large peak of arrivals is expected between periods 3 and 5. For each period, we compute the mean WIP, $\overline{WIP}$, the WIP at Beginning of Period (WIP BOP) and the WIP at End of Period (WIP EOP).

In Table 5.3, we compute the mean WIP and additional CT using the first approach, *i.e.* with Equation (5.4), and the second approach, *i.e.* with Equation (5.8). Using the first approach, we show that both arrival profiles have strictly the same effect on the additional mean CT. However, when using the second approach, the additional mean CT is actually larger by 49% than in case 1. A larger cycle time should be expected in case of large variability in arrivals (Shanthikumar et al., 2007; Hopp and Spearman, 2011).

Table 5.3: Mean WIP and additional mean CT (ACT) in seconds.

	First approach (Equation (5.4))		Second approach (Equation (5.8))	
	Case 1	Case 2	Case 1	Case 2
WIP	5,000	5,000	5,000	7,450
ACT	14,897	14,897	14,897	22,196

Table 5.2: Illustrative example on the importance of the demand profile. Production capacity = 2,000 wafers per period.

	First approach (Equation (5.4))				Second approach (Equation (5.8))			
Period	*WIP BOP*	*WIP EOP*	N	$\overline{WIP}$	*WIP BOP*	*WIP EOP*	N	$\overline{WIP}$
1	5,000	5,000	2,000	5,000	5,000	5,000	2,000	5,000
2	5,000	5,000	2,000	5,000	5,000	5,300	2,300	5,150
3	5,000	5,000	2,000	5,000	5,300	6,600	3,300	5,950
4	5,000	5,000	2,000	5,000	6,600	8,700	4,100	7,650
5	5,000	5,000	2,000	5,000	8,700	10,000	3,300	9,350
6	5,000	5,000	2,000	5,000	10,000	9,900	1,900	9,950
7	5,000	5,000	2,000	5,000	9,900	9,400	1,500	9,650
8	5,000	5,000	2,000	5,000	9,400	8,600	1,200	9,000
9	5,000	5,000	2,000	5,000	8,600	7,800	1,200	8,200
10	5,000	5,000	2,000	5,000	7,800	7,000	1,200	7,400
11	5,000	5,000	2,000	5,000	7,000	6,100	1,100	6,550
12	5,000	5,000	2,000	5,000	6,100	5,000	900	5,550

Tables 5.2 and 5.3 show that neglecting production variability leads to inaccurate estimates of the mean cycle time, which can result in poor decision making in terms of re-qualifications. Therefore, considering a dynamic approach, *i.e.* where arrivals and production capacities are time-varying, is critical. Re-qualifications by using Equation (5.8) are therefore expected to be relevant for the mean cycle time in practice for real manufacturing systems than re-qualifications relying on Equation (5.4). This example also shows that optimizing the throughput is not strictly equivalent to optimizing the mean cycle time contrary to what Equation (5.4) suggests, although increasing the throughput decreases the mean cycle time. Again, this is due to production variability.

From a long term perspective, effective re-qualifications can help to minimize the mean cycle time (see Figure 5.3), which helps reduce inventory costs. Note that minimizing the cycle time with the help of re-qualifications and Equation (5.4) or Equation (5.8) reduces the effect of the variability component on the cycle time as the (long term) utilization rate of machines has already been decided by tactical and strategic decision levels. In addition, for instance, production plans in semiconductor manufacturing can be determined from historical mean cycle times (see *e.g.* Christ 2020). Minimizing the cycle time with re-qualifications is a catalyst to improve responsiveness because of shorter cycle times and to reduce backlog penalties if the same cycle times are proposed to clients.

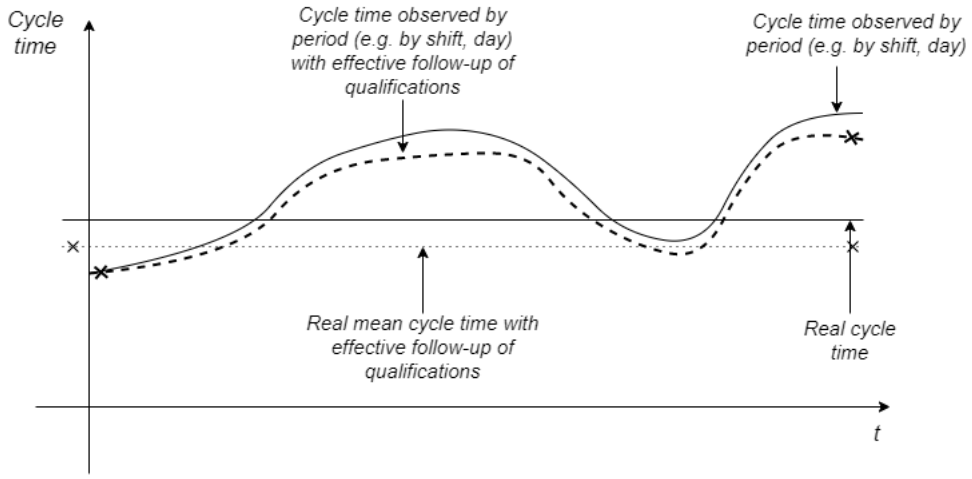


Figure 5.3: Illustration of the reduction of the mean cycle with re-qualifications at a work center.

5.3 Computational experiments

Numerical experiments are used to validate Equations (5.4) and (5.8) with industrial data from a 300mm manufacturing facility located in Crolles, France. In other words, numerical experiments are used to verify if deterministic assumptions for arrivals at and departures from a work center hold, and, if they hold, to what extent.

5.3.1 Design of experiments

Effect size. The effect size (value) of each coefficient before $N(0)$, $N'(T)$ and $D(T)$ is evaluated with an ordinary least square linear regression. The effect size is compared to what is derived in Equation (5.4). The ordinary least square regression is performed by using the *statmodels* library in Python. Both the fitted value and the 95-% confidence interval around the fitted value are reported.

Statistical test. Two-tailed t-tests are used to evaluate if the coefficients before $N(0)$, $N'(T)$ and $D(T)$ are statistically different from zero. t-test results are also provided by the *statmodels* library in Python. Recall that, in linear regressions, the null hypothesis states that the coefficient before a predictor is equal to zero. To each predictor, including the intercept, a p-value is associated. A p-value less than a given threshold, usually 0.05, indicates that the null hypothesis can be rejected because, on the long run and in less than 5% of time, a mistake will be made on the relevance of the predictor. A p-value is a relevant information only if the statistical test has enough statistical power.

Sample size. To perform the numerical experiments, industrial data are used from a 300mm manufacturing facility located in France. However, it is impossible to extract all possible data. The appropriate sample size is computed to achieve a desirable statistical power, which corresponds in this study, to the probability of

a observing a coefficient before $N(0)$, $N'(T)$ and $D(T)$ different from zero if it is actually different from zero in practice. To compute the appropriate sample size, the open source statistical software *G-Power* is used (Faul et al., 2007). For a strong effect size of $f^2 = \frac{R^2}{1-R^2} = 0.35$ (Cohen, 2013), an α level of 0.05, a statistical power of 0.95 and three predictors ($N(0)$, $N'(T)$ and $D(T)$), 40 samples are required (two-tailed t-tests, "Linear Multiple Regression: Fixed Model, single regression coefficient" in G-Power). Note that R^2 scores are partial R^2 scores in the sample size computation.

Predictability of future WIP. To evaluate the predictability of future WIP, the Mean Absolute Percentage Error (MAPE) and the Out of Sample (OOS) R^2 scores are reported. The OSS R^2 score is computed with the theoretical equation and not the equation fitted by the linear regression. The OSS R^2 score then mainly measures the correlation between the outputs of Equation (5.4) and historical data. These two indicators are then computed for the entire data set. It is important to report both R^2 scores and MAPE because although R^2 scores can be large, MAPE can be also large (for instance see Table 5.5). All historical data are used to compute MAPE and OOS R^2 scores as there is no training step: The theoretical equation is compared to the real situation.

Data collection. All lot transactions, *i.e.* all lots that arrived and exited, are retrieved for 11 months between 2018 and 2019. This is more than sufficient to meet the sample size requirements. For each lot and operation, the arrival time and departure time are known and reliable. Note that a lot appears multiple times in the data due to re-entrant product flows. However, only the couple (lot, operation) is unique in the data. No outliers were removed. As data are not formatted in a simple way to compute the historical mean WIP on a given horizon, we first divide the 11 months of data in custom horizons to evaluate Equations (5.8) and (5.4). We study six different horizons: One half-shift (4 hours), one shift (8 hours), one half-day (12 hours), one day (24 hours), one week (7 days or 168 hours), one month (35 days or 840 hours). Note that a month is equal to 35 days so that one month is an integral multiple of one week. Note that a horizon of one month will not be used for statistical tests as there are not enough months in the data to meet the sample size requirement. However, it can still be used to compare Equations (5.8) and (5.4) in terms of predictability (see Section 5.3.2.3).

Studied work centers. Five different work centers are studied. Work centers completely perform different operations. Machines in work centers are non-identical. Numerical experiments are conducted for a large variety of work centers in terms of machine types (see Table 5.4). This allows us to evaluate the limit of Equations (5.8) and (5.4), in particular for complex machines such as lithography machines or machines with long process times such as diffusion machines. The studied work centers are subject to medium to high variability based on the Coefficient of Variability (CV) of arrivals and the CV of process times according to Hopp and Spearman (2011).

Table 5.4: Different studied work centers.

Work centers	Machine type
Dielectric	Parallel and serial process chambers
Diffusion	Single wafer and batching machines
Etch	Parallel and serial process chambers
Implantation	Single wafer machines
Lithography	Job cascading machines

5.3.2 Numerical results

In Sections 5.3.2.1 and 5.3.2.2, the relevance of Equation (5.4) to statistically describe the mean WIP is studied. t-tests are performed to identify whether $N(0)$, $N'(T)$ and $D(T)$ are statistically significant. The predictability of the mean WIP is shown. The limits of Equation (5.4) on large horizons and product families are also shown. Then, it is shown in Section 5.3.2.3 that Equation (5.8) can better predict the future mean WIP than Equation (5.4), even for large horizons.

5.3.2.1 Numerical results using first approach by work center

The numerical results in Table 5.5 show that Equation (5.4) is relevant and can accurately predict the mean WIP over a horizon, even when assuming deterministic laws for arrivals and departures. For a horizon of 4 hours, MAPE is below 2.4%. However, the larger the horizon, the larger MAPE. MAPE is multiplied by two between a horizon of 4 hours and a horizon of 12 hours. MAPE is still below 5.1% for all work centers. These results can be surprising because a much larger MAPE is expected for the work center with batching machines because many lots often leave at the same time the work center at the end of batches, which could make the deterministic assumption on departures unrealistic. For horizons larger than 12 hours, MAPE increases significantly. For a horizon of 24 hours, MAPE is between 5% and 8%. For horizons of 168 hours, Equation (5.4) is inaccurate because MAPE exceeds 15%.

For a horizon of 168 hours, $N'(T)$, $N(0)$, and $D(T)$ are very often statistically significant but can be quite far from the expected values using Equation (5.4). For instance, for the etch work center, the coefficient before $N(0)$ is 0.743 whereas the theoretical one is equal to 1.0. Note that, for the diffusion work center, only $N(0)$ is a statistically significant predictor. However, its coefficient is equal to 0.59, which is far from the expected coefficient, *i.e.* 1.0, computed from Equation (5.4). Moreover, MAPE varies between 12.0% and 16.2%. Therefore, although $N'(T)$, $N(0)$, and $D(T)$ are statistically significant and R^2 scores (for most work center greater than 0.7) indicate that there is a strong correlation between historical data and Equation (5.4), Equation (5.4) should not be used to estimate the mean WIP because MAPE is too large. For horizons smaller or equal to 24 hours, the coefficients determined by the linear regression are close to the coefficients derived by assuming deterministic laws. Predictors, $N'(T)$, $N(0)$, and $D(T)$, are all significant, even for

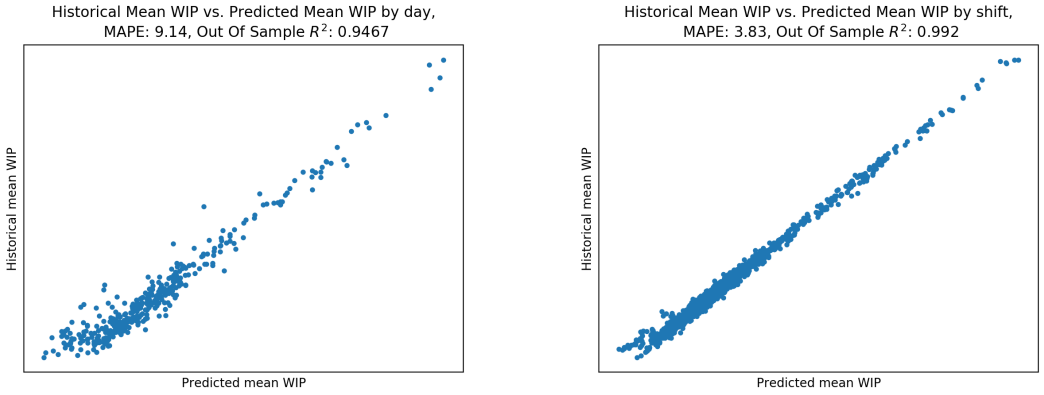
the diffusion work center with process time of several hours. However, it is interesting to note that confidence intervals almost never contain derived values of $N'(T)$, $N(0)$, and $D(T)$. This probably means that the true coefficients before $N'(T)$, $N(0)$, and $D(T)$ are not exactly those derived by assuming deterministic laws for arrivals and departures. However, from a practical standpoint, Equation (5.4) is appropriate because MAPE is reasonable and gets smaller as the horizon gets smaller. Similarly, R^2 scores (greater than 0.9) indicate a good predictability. Using Equation (5.4) is thus acceptable and relevant for horizons smaller than or equal to 24 hours. It is also interesting to observe that, in many cases although the null hypothesis is rejected for the intercept, the associated t-test lacks of statistical power (real partial R^2 scores associated do not lead to a statistical power of 0.95). In general, the intercept accounts for bias that is not captured by other terms in the linear regression. Nevertheless, it is difficult to interpret what the intercept accounts for in practice in a work center, and if it is actually accounts for something given its low statistical power.

Table 5.5: Statistical results for Equation (5.4). The expected coefficients before $N(0)$ is equal to 1, before $N'(T)$ to 0.5, and before $D(T)$ to -0.5. Bold values indicate p-values strictly larger than α. Italic values indicate a low achieved statistical power (< 0.95).

		Coefficient before						
	Horizon (h)	$N(0)$	$N'(T)$	$D(T)$	Intercept	MAPE	R^2	OOS R^2
Diel	4	1, [1, 1]	0.52, [0.51, 0.53]	-0.51, [-0.52, -0.5]	*-34.88, [-60.04, -9.73]*	2.03	0.997	0.997
	8	0.99, [0.98, 0.99]	0.51, [0.5, 0.52]	-0.51, [-0.52, -0.49]	***17.56, [-41.01, 76.13]***	2.98	0.992	0.992
	12	0.97, [0.96, 0.98]	0.52, [0.5, 0.53]	-0.51, [-0.53, -0.5]	***92.61, [-9.37, 194.6]***	3.91	0.985	0.984
	24	0.9, [0.88, 0.92]	0.46, [0.44, 0.48]	-0.45, [-0.48, -0.43]	*453.79, [183.54, 724.03]*	7.07	0.951	0.951
	168	0.73, [0.59, 0.86]	0.47, [0.36, 0.59]	-0.47, [-0.58, -0.35]	***-232.17, [-1938.28, 1473.94]***	15.76	0.752	0.715
Diffusion	4	1, [0.99, 1]	0.52, [0.51, 0.53]	-0.51, [-0.52, -0.5]	***-6.71, [-52.52, 39.1]***	1.72	0.995	0.995
	8	0.97, [0.97, 0.98]	0.52, [0.51, 0.53]	-0.51, [-0.52, -0.49]	183.48, [86.12, 280.84]	2.35	0.989	0.989
	12	0.95, [0.94, 0.96]	0.5, [0.49, 0.51]	-0.5, [-0.51, -0.48]	546.47, [393.04, 699.9]	2.90	0.983	0.983
	24	0.82, [0.8, 0.84]	0.44, [0.42, 0.46]	-0.43, [-0.45, -0.41]	1555.6, [1167.44, 1943.77]	5.65	0.948	0.946
	168	0.59, [0.33, 0.84]	**0.21, [-0.01, 0.43]**	**-0.19, [-0.41, 0.04]**	***-507.32, [-4674.78, 3660.14]***	11.99	0.369	0.290
Etch	4	1, [0.99, 1]	0.53, [0.52, 0.54]	-0.51, [-0.52, -0.5]	*-52.45, [-102.09, -2.81]*	1.57	0.997	0.997
	8	0.98, [0.98, 0.99]	0.52, [0.51, 0.53]	-0.51, [-0.52, -0.5]	***113.34, [2.11, 224.57]***	2.21	0.994	0.994
	12	0.97, [0.96, 0.98]	0.51, [0.5, 0.52]	-0.5, [-0.51, -0.49]	*258.44, [74.7, 442.19]*	2.77	0.989	0.989
	24	0.88, [0.86, 0.9]	0.46, [0.45, 0.48]	-0.45, [-0.47, -0.43]	928.49, [446.54, 1410.43]	5.42	0.966	0.965
	168	0.73, [0.52, 0.95]	0.34, [0.17, 0.51]	-0.31, [-0.48, -0.14]	***-4480.33, [-9214.87, 254.2]***	14.36	0.569	0.478
Implantation	4	1, [1, 1]	0.52, [0.51, 0.53]	-0.52, [-0.53, -0.51]	***-29.88, [-67.86, 8.1]***	2.21	0.998	0.998
	8	1, [0.99, 1]	0.53, [0.51, 0.54]	-0.53, [-0.54, -0.52]	*103.35, [9.83, 196.88]*	3.28	0.994	0.993
	12	0.98, [0.98, 0.99]	0.52, [0.5, 0.54]	-0.52, [-0.54, -0.51]	*245.32, [76.47, 414.18]*	4.39	0.987	0.987
	24	0.95, [0.93, 0.97]	0.49, [0.47, 0.51]	-0.49, [-0.52, -0.46]	*685.74, [237.88, 1133.6]*	7.62	0.963	0.963
	168	0.84, [0.73, 0.95]	0.52, [0.42, 0.63]	-0.51, [-0.61, -0.4]	***-3079.74, [-6137.04, -22.44]***	12.34	0.880	0.861
Lithography	4	1, [1, 1]	0.54, [0.53, 0.55]	-0.53, [-0.54, -0.52]	***-50.37, [-101.01, 0.27]***	2.48	0.997	0.997
	8	0.99, [0.98, 0.99]	0.51, [0.5, 0.52]	-0.52, [-0.53, -0.5]	*169.76, [43.79, 295.73]*	3.83	0.992	0.992
	12	0.99, [0.98, 1]	0.51, [0.5, 0.53]	-0.52, [-0.54, -0.5]	*315.93, [87.59, 544.26]*	5.21	0.984	0.984
	24	0.95, [0.93, 0.97]	0.48, [0.45, 0.51]	-0.49, [-0.52, -0.46]	*886.9, [262.33, 1511.48]*	9.14	0.947	0.947
	168	0.76, [0.66, 0.86]	0.47, [0.36, 0.58]	-0.46, [-0.57, -0.35]	***371.6, [-3142.97, 3886.17]***	16.21	0.829	0.817

Finally, a visual comparison between the predicted mean WIP and the historical mean WIP cound be found in Figures5.4 and 5.5. Figure 5.4 presents the predicted mean WIP against the historical mean WIP for the lithography work center. Figure 5.5 presents the predicted mean WIP and the historical mean WIP for the lithography work center over time. Note that only the results for the lithography work center is presented because other work centers have similar results. Note that scales are hidden for confidentiality purposes.

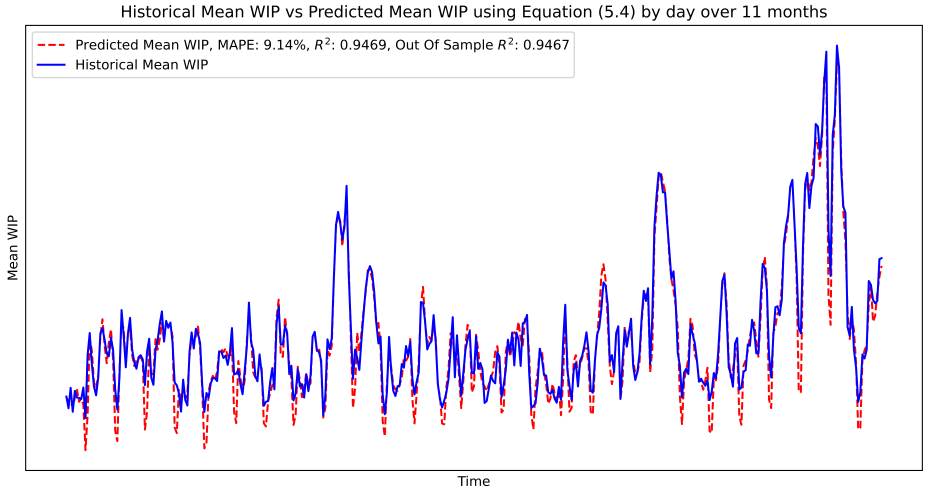
Consider a horizon of one day. Figures 5.4 and 5.5 show that, although the out of sample R^2 is greater than 0.9, Equation (5.4) can be slightly biased, in particular when the WIP is low. The predicted mean WIP is always smaller than the real mean WIP when the real mean WIP is low, which indicates that there is still room for improvements. This is where Equation (5.8) is relevant. Note also that Figure 5.5 also highlights the fact that the work center is subject to a great production variability since the historical mean WIP *strongly* varies over time.



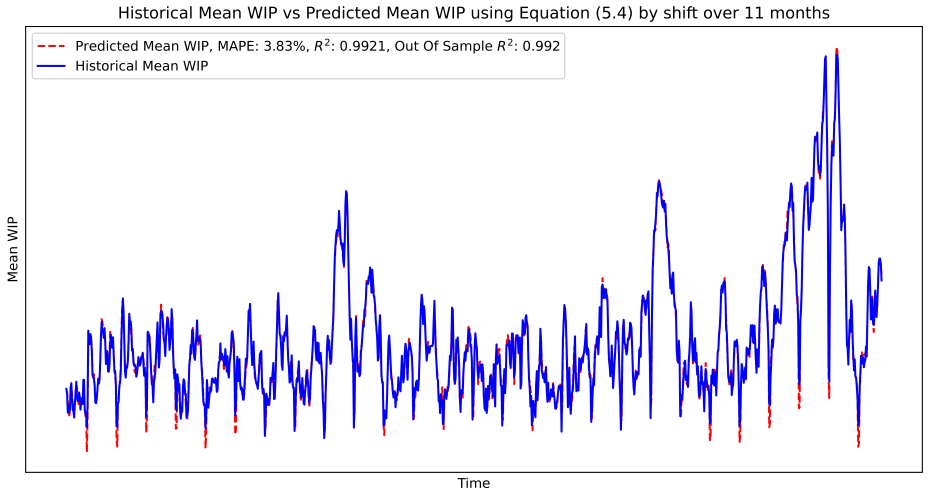
(a) Horizon of one day.

(b) Horizon of one shift day.

Figure 5.4: Predicted mean WIP against historical mean WIP for lithography work center. Scales are hidden for confidentiality purposes.



(a) Horizon of one day.



(b) Horizon of one shift.

Figure 5.5: Visual comparison between predicted mean WIP and historical mean WIP over time for lithography work center over time. Scales are hidden for confidentiality purposes.

5.3.2.2 Numerical results using first approach by product family by work center

Table 5.6, respectively Table 5.7, presents the numerical results as in Table 5.5 but for one low volume product family, respectively for one large volume product family. In particular, Tables 5.6 and 5.7 show that it is still possible to use Equation (5.4) to predict the mean WIP for aggregation levels of operations, such as product families, with large WIP, but its use becomes questionable for aggregation levels of operations with low WIP.

Consider the low volume product family. Table 5.6 shows that, even for a horizon of 4 hours, Equation (5.4) is unable to provide accurate forecasts of the historical mean WIP. For instance, MAPE is equal to 38.7% for the dielectric work center although $N(0)$, $N'(T)$, and $D(T)$ are statistically significant. MAPE is always greater than 15% although all predictors are significant except for the intercept. Similar results were observed for most low volume product families.

Table 5.6: Small volume product family. Statistical results for Equation (5.4). The expected coefficients before $N(0)$ is equal to 1, before $N'(T)$ to 0.5, and before $D(T)$ to -0.5. Bold values indicate p-values strictly larger than α. Italic values indicate a low achieved statistical power (< 0.95).

		Coefficient before						
	Horizon (h)	$N(0)$	$N'(T)$	$D(T)$	Intercept	MAPE	R^2	OOS R^2
Diel	4	0.94, [0.93, 0.95]	0.51, [0.5, 0.53]	-0.44, [-0.45, -0.42]	***-0.35, [-1.27, 0.57]***	38.67	0.941	0.937
	8	0.87, [0.85, 0.9]	0.46, [0.44, 0.48]	-0.39, [-0.41, -0.37]	***-0.46, [-2.21, 1.28]***	50.38	0.891	0.881
	12	0.77, [0.74, 0.8]	0.43, [0.4, 0.45]	-0.33, [-0.36, -0.3]	***-0.62, [-3.16, 1.92]***	40.71	0.845	0.813
	24	0.62, [0.57, 0.67]	0.37, [0.33, 0.4]	-0.29, [-0.33, -0.25]	***0.61, [-3.75, 4.97]***	49.10	0.771	0.698
	168	0.22, [0.13, 0.31]	0.12, [0.06, 0.18]	**-0.1, [-0.16, -0.04]**	***3.24, [-7.3, 13.77]***	48.49	0.742	0.386
Diffusion	4	0.96, [0.96, 0.97]	0.52, [0.51, 0.54]	-0.45, [-0.46, -0.43]	***-0.78, [-1.98, 0.43]***	16.18	0.962	0.960
	8	0.92, [0.9, 0.93]	0.49, [0.47, 0.51]	-0.41, [-0.43, -0.39]	***-0.77, [-3.01, 1.48]***	21.50	0.938	0.933
	12	0.93, [0.9, 0.95]	0.49, [0.47, 0.51]	-0.44, [-0.47, -0.41]	***-2.43, [-5.89, 1.02]***	22.01	0.909	0.906
	24	0.83, [0.79, 0.88]	0.44, [0.4, 0.48]	-0.39, [-0.43, -0.35]	***3.25, [-3.5, 10.01]***	25.87	0.840	0.832
	168	0.5, [0.35, 0.65]	0.33, [0.22, 0.45]	-0.32, [-0.43, -0.2]	***4.45, [-20.53, 29.44]***	33.24	0.696	0.625
Etch	4	0.96, [0.95, 0.97]	0.53, [0.52, 0.55]	-0.44, [-0.45, -0.42]	**-2.56, [-4.37, -0.74]**	15.75	0.958	0.956
	8	0.92, [0.9, 0.94]	0.49, [0.47, 0.51]	-0.41, [-0.44, -0.39]	***-1.46, [-5.09, 2.17]***	17.44	0.927	0.922
	12	0.91, [0.88, 0.93]	0.5, [0.47, 0.52]	-0.42, [-0.45, -0.39]	**-5.31, [-10.49, -0.13]**	19.51	0.907	0.899
	24	0.86, [0.81, 0.9]	0.45, [0.41, 0.49]	-0.41, [-0.45, -0.36]	***-2.48, [-12.18, 7.22]***	22.61	0.852	0.841
	168	0.48, [0.35, 0.62]	0.22, [0.12, 0.33]	**-0.21, [-0.32, -0.09]**	***-2.86, [-33.89, 28.16]***	29.50	0.805	0.753
Implantation	4	0.98, [0.98, 0.99]	0.51, [0.5, 0.52]	-0.47, [-0.48, -0.45]	***-1.1, [-2.38, 0.18]***	42.52	0.964	0.963
	8	0.95, [0.93, 0.96]	0.49, [0.46, 0.51]	-0.43, [-0.45, -0.41]	***-0.26, [-3, 2.48]***	25.29	0.919	0.914
	12	0.9, [0.87, 0.92]	0.5, [0.47, 0.53]	-0.46, [-0.49, -0.43]	**5.53, [1.65, 9.42]**	27.25	0.889	0.883
	24	0.79, [0.75, 0.83]	0.44, [0.4, 0.47]	-0.4, [-0.44, -0.36]	**11.37, [4.96, 17.77]**	25.54	0.843	0.828
	168	0.56, [0.41, 0.71]	0.27, [0.15, 0.39]	-0.26, [-0.38, -0.14]	***11.45, [-13.12, 36.03]***	33.12	0.690	0.634
Lithography	4	0.94, [0.93, 0.95]	0.52, [0.51, 0.54]	-0.42, [-0.44, -0.4]	***-0.85, [-2.66, 0.95]***	23.43	0.924	0.919
	8	0.88, [0.86, 0.9]	0.49, [0.47, 0.52]	-0.4, [-0.43, -0.37]	***0.38, [-3.29, 4.05]***	26.15	0.859	0.848
	12	0.86, [0.83, 0.89]	0.43, [0.4, 0.46]	-0.39, [-0.42, -0.36]	**7.94, [2.52, 13.36]**	27.66	0.801	0.796
	24	0.83, [0.78, 0.88]	0.4, [0.36, 0.45]	-0.38, [-0.42, -0.33]	**10.78, [1.81, 19.76]**	28.85	0.742	0.732
	168	0.56, [0.35, 0.78]	**0.23, [0.08, 0.38]**	**-0.22, [-0.37, -0.07]**	***16.86, [-17.55, 51.28]***	34.92	0.431	0.350

Now consider the large volume product family. Contrary to Table 5.6, Table 5.7 shows similar results than Table 5.5. However, MAPE is larger. MAPE is less than 10% for horizons smaller than 24 hours. MAPE is approximately equal to 5% for a horizon of 4 hours.

Table 5.7: Large volume product family. Statistical results for Equation (5.4). The expected coefficients before $N(0)$ is equal to 1, before $N'(T)$ to 0.5, and before $D(T)$ to -0.5. Bold values indicate p-values $> \alpha$. Italic values indicate a low achieved statistical power (< 0.95).

		Coefficient before						
	Horizon (h)	$N(0)$	$N'(T)$	$D(T)$	Intercept	MAPE	R^2	OOS R^2
Diel	4	1, [0.99, 1]	0.51, [0.49, 0.52]	-0.48, [-0.49, -0.47]	-8.63, [-13.03, -4.23]	5.01	0.993	0.993
	8	0.99, [0.98, 1]	0.51, [0.49, 0.52]	-0.48, [-0.5, -0.47]	**-13.21, [-23.16, -3.26]**	6.90	0.985	0.985
	12	0.98, [0.97, 0.99]	0.52, [0.5, 0.54]	-0.49, [-0.51, -0.47]	**-19.62, [-35.47, -3.76]**	8.46	0.976	0.976
	24	0.93, [0.9, 0.95]	0.48, [0.45, 0.51]	-0.45, [-0.48, -0.42]	***-20.45, [-58.62, 17.71]***	13.17	0.940	0.938
	168	0.63, [0.51, 0.76]	0.39, [0.29, 0.48]	-0.37, [-0.47, -0.27]	***-35.61, [-234.37, 163.15]***	24.33	0.777	0.742
Diffusion	4	0.99, [0.98, 0.99]	0.51, [0.5, 0.53]	-0.49, [-0.5, -0.47]	***-4.54, [-18.22, 9.14]***	3.66	0.981	0.981
	8	0.95, [0.94, 0.96]	0.5, [0.48, 0.51]	-0.46, [-0.48, -0.45]	***23.14, [-3.4, 49.67]***	4.73	0.966	0.965
	12	0.92, [0.9, 0.93]	0.49, [0.47, 0.51]	-0.45, [-0.47, -0.42]	***38.81, [-4.31, 81.92]***	6.03	0.942	0.939
	24	0.74, [0.71, 0.78]	0.4, [0.38, 0.43]	-0.35, [-0.39, -0.32]	**212.11, [123.82, 300.41]**	9.47	0.871	0.856
	168	0.49, [0.28, 0.7]	0.33, [0.19, 0.48]	-0.31, [-0.46, -0.16]	***-78.6, [-486.45, 329.25]***	13.83	0.694	0.504
Etch	4	0.99, [0.98, 0.99]	0.51, [0.5, 0.52]	-0.48, [-0.5, -0.47]	***-12.65, [-25.33, 0.04]***	3.07	0.991	0.991
	8	0.97, [0.96, 0.98]	0.51, [0.5, 0.52]	-0.49, [-0.51, -0.47]	***15.82, [-10.87, 42.5]***	4.19	0.982	0.982
	12	0.95, [0.93, 0.96]	0.5, [0.48, 0.52]	-0.47, [-0.49, -0.45]	***23.02, [-17.47, 63.51]***	5.13	0.973	0.972
	24	0.86, [0.83, 0.88]	0.45, [0.42, 0.47]	-0.42, [-0.44, -0.39]	**128.88, [39.03, 218.72]**	8.25	0.934	0.930
	168	0.59, [0.39, 0.79]	0.34, [0.21, 0.47]	-0.32, [-0.46, -0.18]	*-147, [-593.85, 299.86]*	16.74	0.726	0.654
Implantation	4	1, [1, 1]	0.53, [0.52, 0.54]	-0.5, [-0.52, -0.49]	**-23.04, [-32.32, -13.76]**	4.38	0.993	0.993
	8	1, [0.99, 1]	0.54, [0.52, 0.55]	-0.52, [-0.54, -0.5]	**-22.43, [-43.56, -1.29]**	6.04	0.983	0.983
	12	0.97, [0.96, 0.98]	0.51, [0.49, 0.53]	-0.49, [-0.52, -0.47]	***8.37, [-26.58, 43.32]***	7.39	0.971	0.971
	24	0.93, [0.91, 0.96]	0.49, [0.47, 0.52]	-0.47, [-0.5, -0.44]	***13.28, [-59.87, 86.44]***	10.55	0.942	0.940
	168	0.74, [0.63, 0.86]	0.46, [0.37, 0.55]	-0.45, [-0.54, -0.35]	*-222.21, [-567.43, 123.01]*	18.61	0.833	0.779
Lithography	4	0.99, [0.99, 0.99]	0.52, [0.51, 0.54]	-0.48, [-0.5, -0.47]	-36.33, [-49.87, -22.79]	6.47	0.989	0.988
	8	0.98, [0.97, 0.99]	0.51, [0.49, 0.53]	-0.49, [-0.51, -0.47]	***-28.28, [-56.17, -0.38]***	8.43	0.979	0.979
	12	0.98, [0.97, 0.99]	0.51, [0.49, 0.53]	-0.51, [-0.53, -0.49]	***35.08, [-7.6, 77.76]***	9.71	0.969	0.969
	24	0.95, [0.92, 0.97]	0.49, [0.46, 0.52]	-0.5, [-0.53, -0.46]	**138.35, [42.06, 234.65]**	14.25	0.929	0.929
	168	0.69, [0.6, 0.78]	0.41, [0.32, 0.5]	-0.41, [-0.5, -0.31]	***98.26, [-280.03, 476.55]***	22.98	0.825	0.812

A potential explanation is that such differences between low volume and big volume product families is that the variability of the arrival rate of low volume product families is much greater than the variability of big volume product families. This is because there is no continuous flow of products arriving in the work center for small volume product families. Equation (5.4) does not capture well discontinuous product flows.

Although there are many more low volume product families than large volume product families, low volume products do not prevent Equation (5.4) from accurately estimating the overall mean WIP in the work center. This is because their total volume is much more smaller than the total volume of big volume products for which Equation (5.4) is relevant. In addition, to a certain extent, it is also possible that the error on the estimated mean WIP for low volumes products compensate, thus increasing the overall accuracy of Equation (5.4).

However, a problem with Equation (5.4), in particular for "large" horizons, is that it can underestimate the mean WIP when the real mean WIP is low, which can lead to conclude that re-qualifications are unnecessary. Decreasing the horizon or dividing the horizon in multiple periods decrease this problem (see Section 5.3.2.3). However, in practice, this is not necessarily a problem. This is because, if the real mean WIP at the work center is low, the mean cycle time at the work center is also low. Therefore, re-qualifications are not mandatory.

5.3.2.3 Numerical results using second approach by work center

Using MAPE, Table 5.8 compares MAPE the first approach using Equation (5.4) and the second approach using Equation (5.8) to estimate the historical mean WIP. Note that only MAPE is reported in this section because Equation (5.4) has been validated in Section 5.3.2.1, in particular for horizons smaller than or equal to 24 hours.

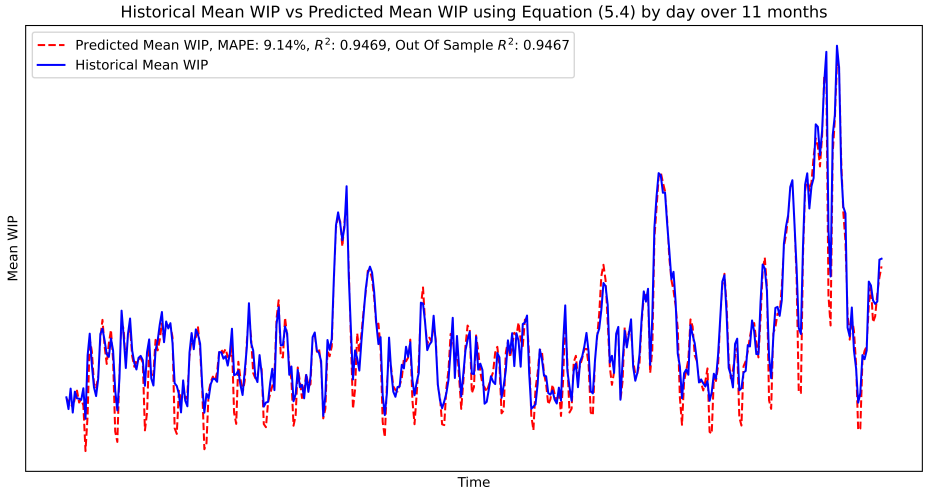
Table 5.8 shows that using Equation (5.8) is particularly relevant to better estimate the historical mean WIP because MAPE is strongly reduced. For instance, consider the case where the horizon of one day (24 hours) is divided into six half shifts of four hours. MAPE is four times smaller than when only one period of 24 hours is considered. Similar observations can be made when the overall horizon is of one week or one month. Dividing the week into days makes MAPE two to four times smaller than considering a single period. Dividing any horizon into periods of a few hours enables to make great estimates of the historical mean WIP because MAPE is lower than 3% in most cases and often very close to 1%. Therefore, Equation (5.8) is more relevant than using Equation (5.4) to estimate the historical mean WIP. The relevance of Equation (5.8) over Equation (5.4) is emphasized in Figure 5.6. Note that similar improvements were observed when the aggregation level is by product family. Consequently, this also emphasizes that the multi-period bilevel optimization approach presented in Chapter 4, Section 4.2.3, should be used to better estimate the mean cycle time.

Table 5.8: Comparison on MAPE between using Equations (5.4) and (5.8) to estimate historical mean WIP.

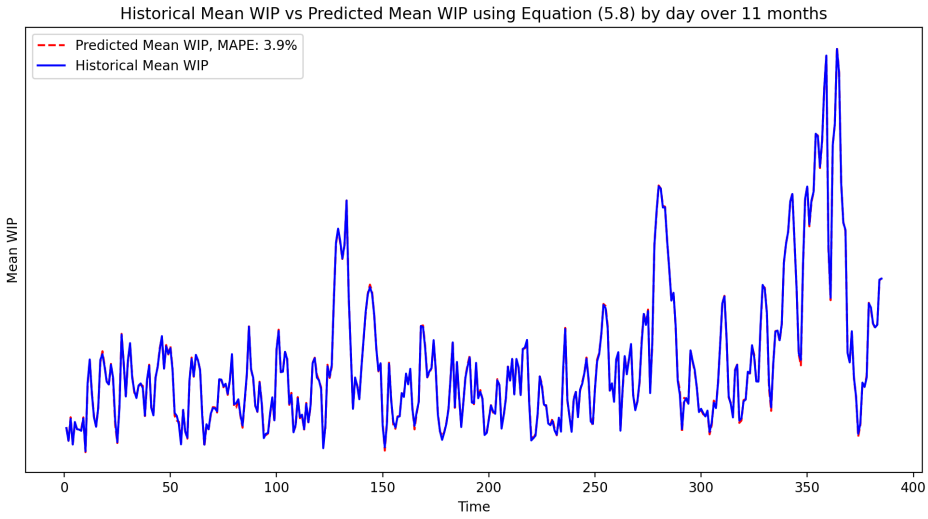
		First approach	Second approach with Equation (5.8)				
Horizon	Work center	with Equation (5.4)	Half Shift	Shift	Half Day	Day	Week
Day	Dielectric	7.1	1.6	2.0	2.9	-	-
	Diffusion	5.6	1.0	1.5	2.1	-	-
	Etch	5.4	1.5	1.8	2.4	-	-
	Implant	7.6	1.4	2.1	3.3	-	-
	Lithography	9.1	3.9	4.5	5.4	-	-
Week	Dielectric	15.8	1.5	1.4	1.3	3.0	-
	Diffusion	12.0	0.7	0.6	0.6	2.7	-
	Etch	14.4	1.3	1.4	1.7	3.5	-
	Implant	12.3	1.3	1.7	2.2	4.6	-
	Lithography	16.2	4.1	4.4	5.0	7.2	-
Month	Dielectric	22.0	1.3	1.2	1.0	1.4	12.2
	Diffusion	12.1	0.4	0.3	0.3	2.0	6.6
	Etch	15.4	1.1	1.2	1.5	3.1	8.3
	Implant	15.5	1.2	1.6	2.0	4.0	7.1
	Lithography	13.4	4.0	4.2	4.9	6.9	8.2

5.4 The effect of one re-qualification on mean cycle times

In this section, the effect of one re-qualification on the additional mean cycle time spent by lots in a work center is illustrated. In particular, it is shown that re-



(a) Historical Mean WIP vs Predicted Mean WIP using Equation (5.4).



(b) Historical Mean WIP vs Predicted Mean WIP using Equation (5.8).

Figure 5.6: Visual comparison of Equations (5.8) and (5.4) for the lithography work center. Scales are hidden for confidentiality purposes.

qualifications that maximize the throughput are not necessarily the ones that minimize the mean cycle time. The mean WIP is computed by using Equation (5.8), and the mean additional cycle time is then computed by using Equation (5.5). Note that minimizing (5.5) or (5.6) is equivalent because both equations differ by a constant term.

5.4.1 Industrial data

To estimate the mean WIP with Equation (5.8), the number of departures from the work center, *i.e.* the throughput at each time period, must be computed. It is computed by using the multi-period bilevel optimization proposed presented in Chapter 4, Section 4.2.3. We are interested in the case where one re-qualification is made at a time.

The computational study is performed by using industrial data from a 300mm wafer fab located in Crolles, France. Two different work centers are studied, work center A and work center B. Both work centers have different types of machines, and run different types of operations. Work center B is subject to strong dedication constraints, *i.e.* a large number of operations are qualified on a single machine, while work center A is subject to mild dedication constraints. The horizon is of 12 hours, and it is divided into three periods of 4 hours. For each work center, 19 different instances are considered. Instances are extracted from different weeks in the year 2019. For both work centers, the number of operations R vary between 400 and 500. The number of machines M is approximately equal to 20. Table 5.9 details the characteristics of the instances. Parameters of the multi-period bilevel optimization approach are similar to those presented in Chapter 2.

Table 5.9: Characteristics of industrial instances.

	Work center A		Work center B	
Instance	R	M	R	M
1	496	18	433	22
2	480	18	456	21
3	464	18	462	21
4	431	18	456	22
5	442	18	487	22
6	436	18	473	22
7	423	18	468	22
8	438	18	465	22
9	431	18	493	22
10	420	18	486	22
11	429	18	481	22
12	419	18	483	22
13	453	18	477	22
14	454	18	466	22
15	441	18	465	22
16	436	18	474	22
17	402	18	455	24
18	420	18	454	23
19	455	18	463	24

5.4.2 Numerical results

It is possible to solve to optimality the multi-period bilevel optimization problem presented in Chapter 4, because we are interested in the case where a single re-qualification is selected at a time.

For each instance and each work center, the relative mean cycle time gain, a negative value indicating a reduction of the mean cycle time, associated to the re-qualification that minimizes the mean cycle time is reported. Similarly, the relative throughput gain, a positive value indicating an increase of the throughput, associated to the re-qualification that maximizes the throughput is reported. In addition, we report if the re-qualification that maximizes the overall throughput is also the one that minimizes the mean additional cycle time. Similarly, we report if the re-qualification that best minimizes the mean additional cycle time is also the one that maximizes the overall throughput. The initial situation is computed when no re-qualification decision is made. Table 5.10 presents numerical results for the considered work centers and instances. Table 5.11 presents the mean relative gaps to the smallest mean cycle time for re-qualifications that maximize the throughput but do not minimize the mean cycle time. Conversely, Table 5.11 presents the mean relative gaps to the largest throughput for re-qualifications that minimize the mean cycle time but do not maximize the throughput.

Numerical results show that relevant re-qualification decisions can significantly

improve manufacturing performances of a work center, both in terms of mean cycle time and throughput. Nevertheless, in most instances, the re-qualification that minimizes the mean cycle time is not necessarily the one that maximizes the throughput. Conversely, the re-qualification that maximizes the throughput is not necessarily the one that minimizes the mean cycle time. This can be explained by production variability, in particular the variability in arrivals. If the cycle time must be minimized, then it is likely that it is better to consider Equation (5.8) than Equation (5.4) as the dynamic behavior of a work center, both in terms of arrivals and throughput, is better captured. In addition, numerical results show that there are often only a few re-qualifications that lead to significant gains. Most re-qualifications do not improve the mean cycle time or throughput at all, in particular because they are unable to better balance the utilization rates of the machines. Finally, numerical results show that qualifying an operation on a machine may have a limited effect on the throughput but a large effect on the mean cycle time if the operation was subject to a production capacity interruption. Correctly managing and anticipating operations subject to production capacity interruptions, *e.g.* by qualifying least loaded machines, is critical to improve the mean cycle time.

Consider work center A. The relative gains strongly vary from one instance to another (Table 5.10). For instance, the mean cycle time can be decreased by -2.23% in instance 8, by -3.32% in instance 5 and by -0.40% in instance 13. In only four out of 19 instances, the re-qualification that minimizes the mean cycle time is the same as the re-qualification that maximizes the throughput. Moreover, Table 5.11 show that, when the re-qualification that minimizes the mean cycle time is not the one that maximizes the throughput, the gap can be significant. For instance, consider instance 11, where the gap of the re-qualification that maximizes the throughput to the smallest mean cycle time is equal to 1.62%. The best re-qualification in terms of throughput is therefore quite far from the best re-qualification in terms of mean cycle time. For other instances, the difference may be smaller. The gap is equal to 0.11% for instance 6 and equal to 0.05% for instance 12.

Consider work center B. The relative gains also strongly vary from one instance to another (Table 5.10). Relative gains in terms of mean cycle times are greater than for work center A. Work center B is subject to strong dedication constraints, *i.e.* two machines have few qualifiable operations in common. Machines in work center B are also used in a "back-up" mode, some machines are qualified if the main machine is down. This means that as soon as the main machine is down, wafer quantities cannot move and are blocked at the work center. This is a production capacity interruption, or simpler line stop. Equation (5.8) helps us to understand that even limited wafer quantities subject to line stops can contribute to significantly increase the mean cycle time, in particular if most wafer quantities subject to line stops are already at the work center. For instance, consider instance 10 in Table 5.11. The re-qualification that minimizes the mean cycle time is about 2.21% away from the re-qualification that maximizes the throughput. Therefore, for work center B, it is often preferable to minimize the mean cycle time by qualifying back-up machines even for a relatively small amount of wafer quantities subject to line stops. More precisely, it can be preferable, in terms of mean cycle time, to qualify an operation on a machine to avoid having 400 wafers subject to line stops for twelve hours (and

possibly accumulating more line stop quantities later), than to qualify an operation on a machine to improve the throughput by 500 wafers (because the 500 wafers will arrive latter in the work center). Note that qualifying additional machines for wafer quantities that are in the work center at $t = 0$ but are never processed, has the same effect on the mean cycle time as fixing line stop operations.

For some instances, the numerical experiments show that two re-qualifications can be equivalent in terms of throughput but not necessarily equivalent in terms of mean cycle time. This can be observed for instance 10 and work center B. There is one re-qualification that simultaneously maximizes the throughput and minimizes the mean cycle time (see Table 5.10) but there are also other re-qualifications that maximize the throughput but fails to minimize the mean cycle time (see Table 5.11).

Table 5.10: Relative gain (%) on the mean Additional Cycle Time (ACT) and the throughput (TH) by work center and by instance. Bold values indicate when at least one re-qualification simultaneously minimizes the cycle time and maximizes the throughput.

	Work center A		Work center B	
Instance	Gain ACT (%)	Gain TH (%)	Gain ACT (%)	Gain TH (%)
1	-1.13	1.54	-1.36	1.28
2	**-1.75**	**1.35**	-0.77	0.46
3	-0.83	0.20	**-0.66**	**0.46**
4	-1.55	0.60	-1.25	0.94
5	**-3.32**	**1.58**	**-1.91**	**0.95**
6	-1.24	1.10	**-5.42**	**2.81**
7	**-1.36**	**1.57**	**-6.18**	**3.14**
8	-2.23	1.36	-0.87	1.38
9	-1.21	0.58	**-3.12**	**2.63**
10	-0.61	1.03	-0.88	1.75
11	-1.51	0.78	-1.18	1.04
12	-1.99	1.07	**-2.33**	**1.19**
13	-0.40	0.52	-1.89	0.40
14	-0.57	0.78	**-2.35**	**1.14**
15	-0.80	0.55	-1.06	1.08
16	**-2.43**	**1.20**	-0.63	0.54
17	-0.56	0.59	**-1.75**	**1.00**
18	-2.00	1.01	**-3.05**	**1.12**
19	-3.32	1.57	-0.68	0.57
Mean	-1.52	1.00	-1.96	1.26

Table 5.11: Gap (%) to the smallest mean Additional Cycle Time (ACT) for the re-qualifications that maximize the throughput. Gap (%) to the largest throughput (TH) for the re-qualifications that minimize the cycle time.

	Work center A		Work center B	
Instance	Gap to smallest ACT(%)	Gap to largest TH(%)	Gap to smallest ACT(%)	Gap to largest TH(%)
1	0.52	0.50	0.13	0.95
2	-	-	0.05	0.09
3	0.31	0.72	-	-
4	0.16	0.75	0.24	0.80
5	-	-	-	-
6	0.11	0.89	-	-
7	-	-	-	-
8	0.82	0.62	0.99	0.59
9	0.49	1.31	-	-
10	0.14	0.81	1.34	2.12
11	1.62	0.82	0.28	0.26
12	0.05	1.64	-	-
13	0.20	0.24	0.34	1.73
14	0.39	0.27	-	-
15	0.56	0.52	0.85	0.16
16	-	-	0.19	0.01
17	0.30	1.10	-	-
18	0.56	0.16	-	1.34
19	0.89	0.74	0.21	0.82

It is worth mentioning that the overall mean cycle time may decrease at the expense of some operations or products. In other words, the overall mean cycle time decreases but the mean cycle time of specific operations may slightly increase. This is because qualification decisions strongly affect dispatching decisions (Gurumurthi and Benjaafar, 2004; Johnzén et al., 2008; Kopp et al., 2019), and because products compete for the same production resources that operate at finite production capacity. A multi-objective approach may be appropriate to propose re-qualifications that balance the mean cycle time between, for instance, operation types or products.

5.5 Practical use and recommendations

At the operational level, excellent predictions on the short term of cycle time are not necessarily mandatory. Work center managers are more interested in evaluating and *comparing* the impact of different decisions on cycle times to help them choose the most appropriate decisions without resorting to detailed simulations models which are known to be expensive to maintain and run (Shanthikumar et al., 2007). Equations (5.4) and (5.8) give this support to production personnel while being

relatively accurate. There is no need to run expensive simulation models to evaluate the effect of re-qualifications on the cycle time.

From a qualification management standpoint, Equation (5.8) can be used to determine the best re-qualifications to minimize the cycle time. Beyond its purely predictive use, Equations (5.4) and (5.8) are particularly interesting to compare re-qualification plans in terms of cycle time. For instance, Equations (5.4) and (5.8) can be used to identify two re-qualification plans that would have the same effect on the number of departures, *i.e.* the throughput, by the end of the horizon but that would have different effects on the cycle time. Finally, a better management and follow-up of qualifications will have a beneficial effect on cycle times. For instance, two re-qualifications can be compared in terms of potential gains for the cycle time of critical fabrication layers (recall that wafers are fabricated layer by layer, and a layer is thus a set of operations) or products in the factory. Equations (5.4) and (5.8) can therefore be used to follow qualifications that are critical to optimize cycle times. The numerical results in Section 5.4 show that an efficient and proactive management of operations subject to line stops is also imperative to minimize and control cycle time, in particular for work centers subject to strong dedication constraints. This also calls for a better planning of maintenance operations.

From a practical standpoint, Equation (5.8) is not limited to re-qualification decisions. Any decision that improves the number of departures, *e.g.* minimizing idle times or minimizing the total utilization rate, in a work center will have a beneficial effect on the cycle time in the work center. However, this not necessarily means that all decisions are equivalent for the overall cycle time over all work centers because decisions at an up-stream work center can create variability at a downstream work center. Similarly, any decision that reduces variability in arrivals and departures will have a beneficial effect on the cycle time at the work center. For instance, Meidan et al. (2011), Wribhu (2018) and Wang et al. (2018) provide a list of factors that strongly affect the cycle time. Relevant decisions can be proposed from these factors. For instance, Equation (5.8) can be used to better plan maintenance operations by considering upstream and downstream work centers.

The application of Equation (5.8) is not limited to short-term horizons. Equation (5.8) can also be used to estimate the cycle time, *e.g.* the cycle time by layer, after determining a production plan along with a product start plan (see *e.g.* Hung and Leachman 1996) without resorting to discrete-event simulation approaches. Equations (5.4) and (5.8) can be used on any horizon, even of several months, as long as the horizon is divided into small periods, *e.g.* in periods of 24 hours, and as long as it is possible to compute, or at least estimate, the number of arrivals and departures for all periods within the considered horizon. Moreover, as Equation (5.8) is linear in terms of arrivals and departures, Equation (5.8) could be used as the objective function in linear programs that define production plans and product start plans without increasing the computational burden. Nevertheless, it is possible that using Equation (5.8) for more than a few days can lead to poor estimates of cycle times. This is because, contrary to historical data, departures and arrivals in a work center are estimates, and thus subject to uncertainty. The uncertainty is also probably larger for late periods than early periods in the horizon. In addition, the process times of some machines can be large, *e.g.* of a few hours for diffusion

machines, which can lead to difficulties in estimating future number of departures.

Finally, we recommend using Equations (5.4) and (5.8) only for high levels of aggregation of operations. Typically, the aggregation must be done over all operations or over a large set of operations, *e.g.* by layer, by product family or operation family, so that there is always enough WIP for the considered aggregation level at the work center. Thus, the deterministic assumption on arrivals and departures holds, and it is relevant to use Equations (5.4) and (5.8). If the aggregation level does not have enough WIP, then Equations (5.4) and (5.8) may no longer be relevant. Therefore, for smaller aggregation levels, *e.g.* for a specific operation, unless the operation is common to many products and that there is always WIP at this operation, a discrete-event simulation is required.

5.6 Conclusions and perspectives

In this chapter, first, we motivated the use of closed-form solutions for the mean cycle time in a work center at the operational level. Two closed-form solutions were derived by assuming deterministic arrivals and departures in a work center. One closed-form solution is similar to the one proposed by Leachman (2015). Classical G/G/m queues cannot be used at an operational level because they require the work center to be in a steady state with utilization rates lower than one for all machines, which cannot be ensured because of production variability. Then, the closed-form solutions were validated on industrial data by comparing the historical mean WIP and the predicted mean WIP. The limits of the closed-form solutions were also shown, as they could be inaccurate and irrelevant for decision making when there is not enough WIP in the system. Finally, a computational study was performed to show how closed-form solutions could be used to propose re-qualifications that minimize the mean cycle time in a work center.

We believe the following perspectives are worth investigating in the future (out of the scope of the study):

1. It would be extremely relevant to study the effect of machine qualifications, *e.g.* the effect of variability of machine qualifications, in G/G/m queues. For instance, this could lead to better decisions at a tactical level when new qualifications are determined. Nevertheless, considering machine qualifications in queuing theory is complex. This is because not only the number of qualifications matters but also the qualifications themselves are critical. Two sets of qualifications with the same number of qualified operations by machine and the same number of qualified machines by operation will often lead to different cycle times as shown in the numerical results. Instead of trying to directly determine the cycle time with qualifications, a solution could consist in determining and exploiting upper bounds on the cycle time. For instance, it is known that the mean cycle time of G/G/m queues is bounded by the mean cycle time of m G/G/1 queues working in parallel. Exploiting this bound could provide valuable insights to capacity planners to decide the qualifications that should be added to a machine to reduce the mean cycle time. Knowing the exact value of the mean cycle time is not mandatory to make relevant deci-

sions (see Section 5.1.1). Other bounds may be available. Recently, bounds on the mean cycle time were proposed based on robust optimization (Bandi et al., 2015; Bertsimas et al., 2018; Whitt and You, 2019). New bounds could also be derived using machine learning techniques and historical data.

2. Multi-objective considerations should also be studied. A re-qualification may decrease the mean cycle time of a particular layer at the work center. Nevertheless, the same re-qualification may also increase the mean cycle time of another layer.

3. Consistent qualification decisions between work centers should also be studied. A local decision at a work center may not be ultimately relevant if all work centers were simultaneously considered.

Chapter 6

Robust tactical qualification management to cover demand uncertainty

This chapter deals with tactical qualification management*. More precisely, we want to answer the question "How to determine the most relevant new qualifications to satisfy the demand and cover the demand uncertainty while minimizing qualification costs?" We show on industrial data that: (1) A limited number of well-chosen qualifications are required to achieve the same robustness than the one obtained by performing all qualifications, and (2) Implementing the qualifications determined by only considering the nominal demand can lead to capacity constraint violations.

*This chapter has been submitted to an international journal.

6.1 Introduction

Satisfying the demand associated to each product is difficult in semiconductor manufacturing. Several hundred products compete for the same production machines in high mix manufacturing facilities. In addition, the demand by product is time-varying, often significantly from one month to another, and can be highly uncertain. There are also manufacturing risks (*e.g.* machine breakdowns, yield losses) that can prevent manufacturing facilities from satisfying the demand. When such conditions are met, the need for *flexibility* (the ability to respond effectively to changing circumstances, see Sethi and Sethi 1990) is imperative (De Toni and Tonchia, 1998). Qualification management is closely related to the the notion of *production flexibility*, which is defined as all products a factory is able to produce without requiring additional major capital investment. Production flexibility is the result, among others, of *process flexibility*, which is defined as the ability of processing different products at the same time (Sethi and Sethi, 1990; Jain et al., 2013). Adding new qualifications improves the level of process flexibility of work centers and therefore improves the capacity of a factory to satisfy the demand.

In this chapter, we are interested in the qualification optimization problem that typically arises at a tactical decision level where the planning horizon is between six and twelve months. The considered qualification optimization problem is a tactical capacity planning problem: The production capacity of a work center must be configured to satisfy the demand. There are existing machines in the work center, and new machines might be installed. Similarly, new products are being introduced in the factory, and new qualifications are necessary to increase the production capacity of new products and increase the production capacity of already existing products already made by factory with a ramp-up demand. This is because new qualifications enable operations associated to the product to be processed on more machines. More precisely, a set of new qualifications, *i.e.* new couples (operation, machine) to qualify, must be determined so that the demand for all products is satisfied while respecting production capacity constraints. The couple (operation, machine) must be either determined as to be qualified or not to be qualified.

Because new qualifications can be expensive and time-consuming, between one week and several months mainly in the form of delay as test lots must complete their production to validate the operation on the machine, the number of new qualifications to perform must be minimized and anticipated. Moreover, the demand by product, which is an external parameter to the company, is affected by uncertainty. In factories with a high product mix, *i.e.* many products, the uncertainty on the demand by product is particularly strong, as factories face frequent product mix changes with products that have short lifetimes. In other words, the set of qualifications determined to satisfy a nominal demand by product may be inappropriate if the realized demand by product is too different from the nominal demand by product. A significant change in the demand can significantly decrease the manufacturing performances. This is because the wafer of a product does not lead to the same workload of a wafer of another product due to different re-entrant flow factors and throughput rates (Kotcher and Chance, 1999). Determining a “robust” set of new qualifications, which covers the uncertainty on the demand, is therefore

also critical.

Let us recall that the literature related to tactical qualification management on process flexibility cannot most often be directly applied to semiconductor manufacturing (see the literature review, in Chapter 1, Section 1.4.1).

In addition, the literature is scarce on the design of qualification configurations in semiconductor manufacturing, in particular when the demand is uncertain. Stochastic programming has been a method of choice so far to deal with the uncertainty on the demand. Klemmt et al. (2010) propose to design qualification configurations for a specific work center by covering a few scenarios on the demand by product, which is a common practice in the semiconductor industry. Nevertheless, the approach is not entirely detailed. Chang and Dong (2017) propose a stochastic programming optimization approach to maximize the weighted expected number of processed product quantities. The demand and the production capacities are subject to uncertainty. In addition, they consider that new qualifications lead to a stochastic capacity loss that can be described with a distribution probability. However, the approach proposed by Chang and Dong (2017) cannot be used at a tactical level. This is because their stochastic model does not ensure that the demand by operation has to be satisfied. Then, only a fraction of the operations corresponding to a product could be qualified, and the product could potentially never be delivered. Fu et al. (2015) also consider that the demand is uncertain in a qualification management optimization problem. Nevertheless, the problem is treated from an extended production planning standpoint and not from a capacity planning standpoint. Consequently, similarly to Chang and Dong (2017), the work of Fu et al. (2015) cannot be used at a tactical level. Liao et al. (2017) propose a two-stage stochastic programming optimization approach to maximize the total profit of a semiconductor company. The first stage problem consists in minimizing qualification costs while second stage problem consists in allocating product quantities to production sites to maximize revenue.

However, stochastic programming implies characterizing demand scenarios and associated probabilities. This is difficult as products tend to have dependent demands due to product cannibalization, which is not mentioned in the literature. Product cannibalization is particularly critical for manufacturers with a high product mix. Determining nominal demands and plausibility limits is a promising alternative: It is as natural as defining demand scenarios without requiring probabilities and can consider product cannibalization.

Our contributions to the qualification management and robust optimization literature are as follows:

- We propose a new mixed integer linear programming mathematical model for the tactical qualification management problem when the demand is deterministic and the qualification lead times are considered.
- As the demand by product can be subject to uncertainty, we motivate the choice of robust optimization for the considered problem.
- We propose an uncertainty set based on the budget of uncertainty (Bertsimas and Sim 2004), to cover the demand uncertainty and product cannibalization.

- We propose a new robust reformulation of the deterministic model when the demand is considered as uncertain but can be described by $\mathcal{D}_t$.
- We propose a new decision-dependent uncertainty linear program to characterize the robustness of a set of qualifications. As the problem is NP-complete, a binary search solution approach is proposed when the uncertainty on the demand is symmetrical.
- In the computational study, we show on industrial data that the price of uncertainty is small, often less than a few qualifications, whereas the qualifications determined for the nominal demand often lead to capacity constraint violations.

The remainder of this chapter is organized as follows. In Section 6.2, we describe and motivate the type of demand uncertainty faced in semiconductor manufacturing. We motive the use of robust optimization to cover demand uncertainty. In Section 6.3, the deterministic mathematical model is presented. Then, a mathematical robust optimization approach is proposed to cover demand uncertainty. In Section 6.4, we propose a mathematical model and discuss several approaches to determine the robustness of a given set of qualifications (*e.g.* the set of initial qualifications). In Section 6.5, a computational study on industrial data is conducted to evaluate the price of uncertainty (Gorissen et al., 2015), the practical tractability of the proposed optimization models, and possible the capacity constraint violations and consequences if the set of qualifications obtained by solving the deterministic optimization problem is used. In Section 6.6, we discuss how the proposed optimization models can be used for a practical use by capacity planners in a decision support system. Finally, in Section 6.7, we conclude and give some perspectives.

6.2 Uncertainty on the demand

6.2.1 Demand uncertainty and product cannibalization

Processing times, production capacities, qualification lead times and the demand by product can be subject to uncertainty. In this chapter, only the demand uncertainty is considered, which is critical to a manufacturing company. The uncertainty on the demand is an external uncertainty, which is difficult, if not impossible, to control with discount prices and incentives even if the product is innovative. Considering the uncertainty on other parameters is left for future research.

Note that the uncertainty on the demand by operation is a consequence of the uncertainty on the demand by product. In the semiconductor industry, operations need to be run to process a product. However, all products do not share the same operations. Moreover, although two products share common operations, operations will not have the same processing times. This is due to differences in the re-entrant product flows. We are therefore interested in characterizing and modeling the demand uncertainty and linking it to the uncertainty on the demand by operation.

Although it is possible to accurately predict the total quantity of products that a manufacturing facility must complete in the future, it is often impossible to exactly

know the quantity of each product. One important reason why is that high-tech companies such as semiconductor manufacturers with a large portfolio of products often face *product cannibalization* (Moorthy and Png 1992; Kim and Chhajed 2000). Product cannibalization occurs when a company manufactures different products that compete with each other on the same market. Consider the following example. A client that seeks to design an electronic system has the choice between several micro-controller among those that the company sells. A micro-controller is integrated circuit with essentially the same features as modern computers, *i.e.* computing unit, memory, input and output interfaces, but are dedicated to specific applications and require little energy. Several micro-controllers are suitable for a given application, and the final choice will be made based on cost, energy consumption and memory among other characteristics. The client will probably never buy *all* suitable micro-controllers. Therefore, selling one unit of a product may mean selling fewer units of other products. Nevertheless, a product cannot be replaced by any other product because all products are not used for the same application. Some products will be used in the automotive industry, whereas others will be used for industrial applications in factories, or telecommunication applications. Products are distinguished by their *family*. A product family is then a set of products that have similar characteristics, can be used for similar applications, and therefore compete on the same market segment.

6.2.2 Managing the demand uncertainty

To cover the demand uncertainty, two main methods exist: Stochastic optimization and robust optimization. Stochastic optimization assumes that the probability distribution of the demand uncertainty is known. Then, in general, the expected value of the objective function is optimized. In this chapter, the objective would consist in minimizing the expected number of qualifications after generating, possible many, scenarios from the estimated probability distribution (Birge and Louveaux, 2011). Robust optimization is different because the probability distribution of the uncertainty is not required. In robust optimization, the objective function must be minimized while ensuring that the constraints are *never* violated (Ben-Tal and Nemirovski, 2002; Ben-Tal et al., 2009; Bertsimas et al., 2011; Gorissen et al., 2015).

Robust optimization is more relevant when determining a set of qualifications at the tactical decision level. First, estimating the probability distribution of the demand of a product when it is correlated to the demands of other products is difficult. Furthermore, estimating the probability distribution of the demand for new products is difficult. This is because semiconductor manufacturers may not have enough data on the demands to derive *relevant* distribution probabilities as they experience frequent product mix changes (Bertsimas and Thiele, 2006a), *i.e.* the demand for a product strongly varies from one month to another. Figure 6.1 provides an illustrative example using historical industrial data on the changes of the demand for 5 different products over 12 months. Note that the charts have different Y-scales. For confidentially purposes, product names are not mentioned. In addition, the monthly demand is divided by the mean demand over the 12 months. For all products, the mean demand is of several hundreds. The demand for Product

A (see Figure 6.1a) is very different from one month to another. The demand for the first month is about 2.2 while the demand for the seventh month is equal to 0. The demand for the eighth month is about 1.7. The demand for Product B also changes a lot from one month to another. A large demand peak is observed for the sixth and seventh months. Then, the demand is decreasing. The demand for product C is particularly interesting. There is a large, quick and intense ramp-up demand for product C. Nevertheless, the demand for Product C quickly fades away. Product D is a more stable product with a larger demand even though it also suffers from large fluctuations, in particular between the ninth and tenth months when the demand increases by about 2.3. Finally, product E also suffers from large fluctuations. The demand is multiplied by two between the first and the seventh months, then it divided by two between the seventh and the twelve months, with a demand almost equal to zero in the ninth month. These demand fluctuations are critical, especially since their intensities are extremely difficult to predict in advance. This is when a robust optimization based approach is relevant to cover the demand uncertainty.

Second, it is critical to anticipate relevant qualifications to cover the demand uncertainty. This is because, in general, it is important to perform the right qualifications and not all qualifications to respect capacity constraints and satisfy the demand (Jordan and Graves, 1995; Benjaafar et al., 1995; Graves and Tomlin, 2003; Chou et al., 2010; Johnzén et al., 2011; Fiorotto et al., 2018; Chen et al., 2019).

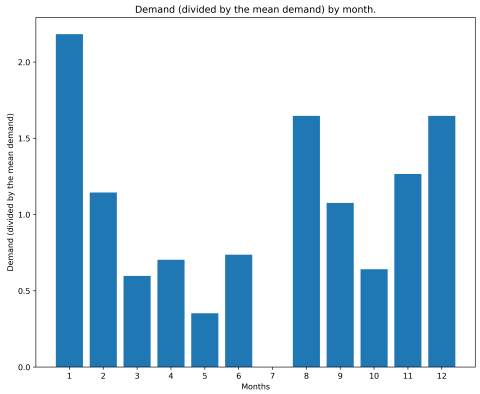
Furthermore, as qualification decisions are made at a tactical decision level, they have a major impact of all production planning and control management issues (Hopp and Spearman, 2011). For instance, if new qualifications are not properly determined, then effective robust production plans may not be found to satisfy the demand. Determining the right set of new qualifications is thus critical for manufacturing and financial performances.

Third, in practice, a way to deal with uncertainty is to frequently adjust the current set of qualifications by performing new qualifications when the nominal demand is updated. However, this is not always possible because the qualification process may sometimes take several weeks or months to validate the quality and the yield of the operation. Therefore, if the demand is updated late, it may be impossible to perform additional qualifications to satisfy the demand. Then, anticipating the right qualifications to cover the demand uncertainty is critical. Also, determining a set of robust qualifications could save critical time for capacity planners. This is because the set of qualifications would be determined in a less reactive manner but in a more proactive manner against demand changes. Capacity planners could therefore be assigned to other tasks. Note that the set of qualifications would still need to be adjusted when completely new products are introduced or old products are reintroduced because of unnoticed disqualifications.

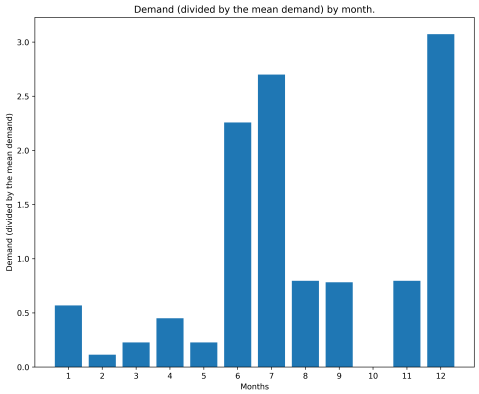
6.3 Problem modeling

6.3.1 Problem description

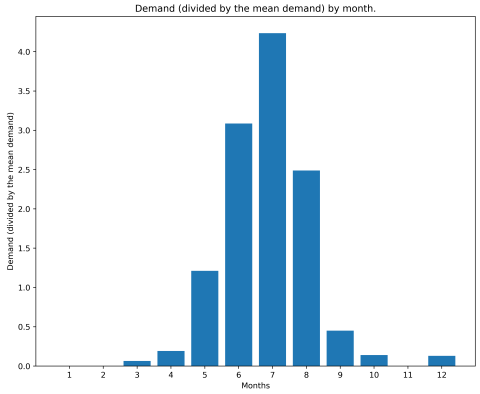
Let us consider a work center of M unrelated parallel machines, both in terms of qualifications and throughput rates, which must process R different operations. Ma-



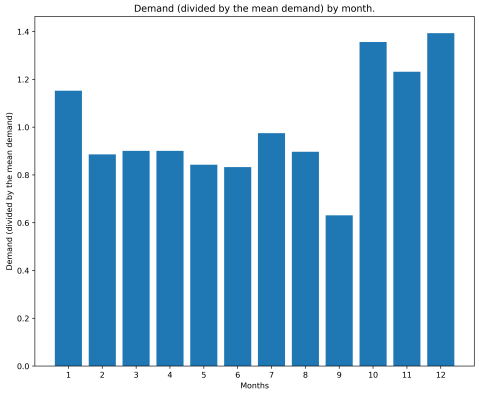
(a) Product A.



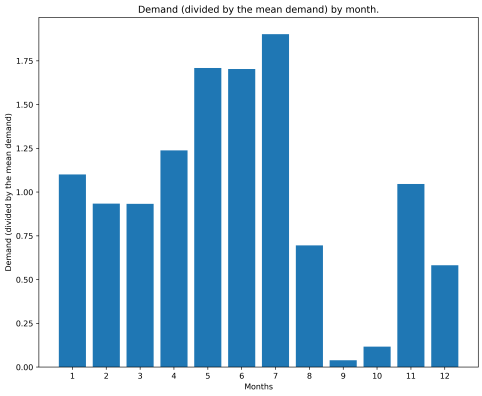
(b) Product B.



(c) Product C.



(d) Product D.



(e) Product E.

Figure 6.1: Illustrative example of demand (in number of wafers) fluctuation over time.

chines are unrelated because they are of different generations. A demand is associated to each operation on the considered horizon. The horizon consists of T periods. The work center is asymmetrical and unbalanced, *i.e.* the demand varies from one operation to another and the number of operations is much greater than the number of machines. A machine can only process qualified operations, and a qualifiable operation can only be processed on a machine if it is qualified. Qualifying an operation on a machine induces a qualification cost and is subject to a qualification lead time. The qualification matrix defines the initial set of active qualifications. A qualification is therefore a pair (operation, machine). The initial set of active qualifications is known and deterministic. Each machine has a finite production capacity that must be respected at each period on the considered horizon.

The objective is to minimize the cost of performed qualifications to perform, among the qualifiable pairs (operation, machine) not already qualified, while satisfying the demand and respecting capacity constraints.

This problem will be referred as the Minimum Cost Qualification Configuration Problem (MCQCP) in the remainder of the chapter.

6.3.2 Deterministic modeling

Parameters:
M: Number of machines,
R: Number of operations,
P: Number of products,
T: Number of periods,
$q_{r,m}$: Is equal to 1 if machine m is initially qualified for operation r, to 2 if machine m is qualifiable for operation r, to 0 if machine m cannot be qualified for operation r,
$tp_{r,m}$: Throughput rate (per hour) of operation r on machine m,
$c_{t,m}$: Production availability (in hours) of machine m at period t,
$u_{t,m}^{max}$: Maximum utilization rate allowed for machine m at period t,
$rf_{p,r}$: How many times (re-entrant flow factor) operation r needs to be run to produce one unit of product p,
$d_{t,p}$: Demand for product p at period t,
$l_{r,m}$: Lead time (in number of periods) for qualifying operation r on machine m,
δ_t: Discount factor at period t,
$cq_{r,m}$: Cost of qualifying operation r on machine m.
Decision variables:
$OQ_{t,r,m} \in \{0,1\}$: Is equal to 1 if there is qualification procedure to start for operation r at period t on machine m, and 0 otherwise,
$WIP_{t,r,m} \in [0,1]$: Ratio of the demand for operation r processed by machine m at period t.

$$\min \quad \sum_{t,r,m} \delta_t cq_{r,m} OQ_{t,r,m} \tag{6.1}$$

$$\text{s.t.} \quad \sum_{r} \frac{(\sum_p rf_{p,r}d_{t,p})WIP_{t,r,m}}{tp_{r,m}} \leq c_{t,m}u_{t,m}^{max} \qquad \forall t, \forall m \tag{6.2}$$

$$\sum_{m} WIP_{t,r,m} = 1 \qquad \forall t, \forall r \mid \sum_{p} rf_{p,r}d_{t,p} > 0 \tag{6.3}$$

$$WIP_{t,r,m} \leq q_{r,m} \qquad \forall t, \forall r, \forall m \mid q_{r,m} \neq 2 \tag{6.4}$$

$$WIP_{t,r,m} \leq \sum_{t'=1 \mid t-t' \geq l_{r,m}}^{t} OQ_{t',r,m} \qquad \forall t, \forall r, \forall m \mid q_{r,m} = 2 \tag{6.5}$$

$$WIP_{t,r,m} \geq 0 \qquad \forall t, \forall r, \forall m \tag{6.6}$$

$$OQ_{t,r,m} \in \{0,1\} \qquad \forall t, \forall r, \forall m \tag{6.7}$$

The objective function (6.1) minimizes the cost of performing qualifications on the planning horizon. The discount factor is used to decide if qualifications must be made as soon as possible or as late as possible. For instance, qualification procedures must be started as late as possible. This is possible by ensuring that $\delta_t \geq \delta_{t+1} \ \ \forall t \in \{1, ..., T-1\}$. Constraints (6.2) ensure that the capacity constraint for each machine m and each period t are respected. Constraints (6.2) also limit the utilization rate of machine m at period t to a maximum of $u_{t,m}^{max}$. This controls the mean cycle time (fabrication time) in the work center as the mean cycle time increases exponentially with the utilization rate, and improves the responsiveness of the work center (Hopp and Spearman, 2011). Constraint (6.3) are the flow constraints. They ensure that the demand by operation must be satisfied. Constraints (6.3) are active only if there is demand for operation r at period t, $\forall t, \forall r \mid \sum_p rf_{p,r}d_{t,p} > 0$. This avoids qualifying operations on machines in the early periods if the demand is only expected in the late periods. Constraints (6.4)-(6.5) are the qualification constraints. They ensure that machine m is qualified for operation r at period t, if it has been newly qualified or was initially qualified while considering qualification lead times. Finally, Constraints (6.6) are the non-negativity constraints and Constraints (6.7) are the binary constraints.

Let us discuss below some important characteristics of our problem:

- The deterministic optimization model is relevant, although it does not consider demand uncertainty, because it considers essential features of qualifications which are qualification costs and delays, and models unbalanced and unsymmetrical systems. In the computational study on industrial data, we found that the deterministic model is easy to solve (see Section 6.5) for the considered work centers.

- MCQCP can also be solved factory-wide, *i.e.* by considering all work centers simultaneously. However, as two different work centers do not share operations, optimality is preserved when breaking down the problem by work center to reduce the size of the problem in terms of machines and operations.

- It is important to mention that MCQCP can be infeasible if the production capacities of machines are too small and if too few qualifiable pairs (operation,

machine) exist to better balance the workload between the machines. Note that in the numerical experiments performed in Section 6.5, MCQCP is always feasible contrary to its robust counterpart.

- The deterministic model can still be used to determine a set of qualifications even if lead times are not modeled, *i.e.* if $l_{r,m} = 0 \ \forall r, \forall m$. In this case, decision variables $OQ_{t,r,m}$ should be interpreted as the period at which operation r must be qualified on machine m if $OQ_{t,r,m} = 1$.

- In order to correctly consider new machines, it is sufficient to set $c_{t,m}$ to appropriate values until machine m is actually started-up in the factory. Start-up periods are notably used for qualification purposes. For instance, if the horizon is of 3 months with 3 periods of one month and the start-up period lasts one month, then $c_{t,m}$ must be equal to zero for the first two months.

- New qualifications can lead to capacity losses in the considered work center as it is required to run quality tasks on machines by using test products. Chang and Dong (2017) model this aspect by using a probability distribution. As it is complex to define relevant probability distributions, capacity losses due to new qualifications are modeled with available historical data as exogenous factors in the production capacity of each machine. Note that quality tasks are also frequently run even for existing qualifications, which is also considered in the production capacity of each machine.

- Contrary to operational qualification management, the problem is not modeled as a bilevel optimization model. This is because the goal is to satisfy all demands, and thus fast operations are processed as often as slow operations.

6.3.3 Robust modeling

6.3.3.1 Polyhedral uncertainty with budget of uncertainty

To consider demand uncertainty and product cannibalization, a polyhedral uncertainty set, based on budget uncertainty proposed by Bertsimas and Sim (2004), is used. Let us introduce the new notations below:

New parameters:
F : Number of product families,
$\overline{d}_{t,p}$: Nominal demand for product p at period t,
$\widehat{d}_{t,p} \leq \overline{d}_{t,p}$: Maximum deviation from nominal demand for product p at period t,
$\alpha_{p,f}$: Is equal to 1 if product p belongs to product family f, and 0 otherwise,
$\Gamma_{t,f}$: Budget of uncertainty for product family f at period t.

The demand $d_{t,p}$ is assumed to be a random variable that takes values as follows: $d_{t,p} \in [\overline{d}_{t,p} - \widehat{d}_{t,p}, \overline{d}_{t,p} + \widehat{d}_{t,p}] \ \forall t, \forall p$. $\overline{d}_{t,p} - \widehat{d}_{t,p}$ and $\overline{d}_{t,p} + \widehat{d}_{t,p}$ are the plausibility limits for product p at period t. The uncertainty set $\mathcal{D}_t$ that models the effect of product cannibalization by product family at period t is described below:

$$\mathcal{D}_t = \{d_{t,p} \mid d_{t,p} \geq \overline{d}_{t,p} - \widehat{d}_{t,p} \quad \forall p, d_{t,p} \leq \overline{d}_{t,p} + \widehat{d}_{t,p} \quad \forall p, \sum_{p|\alpha_{p,f}=1} d_{t,p} \leq \Gamma_{t,f} \quad \forall f\} \tag{6.8}$$

In $\mathcal{D}_t$, the total demand by product family f at period t is limited to the budget of uncertainty $\Gamma_{t,f}$, which is the maximum demand to cover for product family f at period t. Therefore, if the demand for a product in the product family increases above its nominal value, then the increase is made at the expense of another product in the product family, whose demand must decrease. In addition, if $\Gamma_{t,f} = \sum_{p|\alpha_{p,f}=1} \overline{d}_{t,p}$, then, for each product family f, the maximum overall quantity to produce is equal to the overall quantity in the nominal case, but the distribution of the demand between the products in the product family is unknown. Setting $\Gamma_{t,f} = \sum_{p|\alpha_{p,f}=1} \overline{d}_{t,p}$ is a practical assumption. This ensures that qualifications are not determined to cover extreme cases where the quantity by product family would actually be much larger than the nominal quantity by product family, which is often unrealistic. Instead, qualifications are optimized to cover any demand realization given an overall quantity by product family. Note that, although the uncertainty set $\mathcal{D}_t$ ensures that the total demand of all products in a family is not large, the total demand over all operations arriving in work centers can significantly increase as re-entrant flow factors significantly vary from one product to another.

Parameters $\overline{d}_{t,p}$ and $\widehat{d}_{t,p}$ do not necessarily reflect the real uncertainty on the demand of product p at period t. They can be defined in a such way that they correspond to the uncertainty capacity planners *want* to manage if the real uncertainty is too expensive to cover (Bertsimas and Sim, 2004).

6.3.3.2 Static reformulation

We investigate a static reformulation of the deterministic optimization problem. We follow Ben-Tal and Nemirovski (2002), Gorissen et al. (2015) and Yanıkoğlu et al. (2019) to write the robust formulation of MCQCP. First, constraints with uncertain parameters, the demand, need to be identified, then the robust counterpart can be derived.

There are two constraints with uncertain parameters: The flow constraints (6.3), and the capacity constraints (6.2).

Flow constraints (6.3). The demand is used to control when the flow constraint must be active. To make sure the flow constraints hold for any demand realization within $\mathcal{D}_t$, it is sufficient to replace the condition

$$\sum_m WIP_{t,r,m} = 1 \quad \forall t, \forall r \mid \sum_p rf_{p,r} d_{t,p} > 0 \qquad \text{by}$$

$$\sum_m WIP_{t,r,m} = 1 \quad \forall t, \forall r \mid \sum_p rf_{p,r} (\overline{d}_{t,p} + \widehat{d}_{t,p}) > 0.$$

Capacity constraints (6.2). If the demand uncertainty is row-wise and the uncertainty is compact, then an optimal solution for the static reformulation problem

is also an optimal solution for the adjustable robust reformulation problem (Ben-Tal et al., 2004; Yanıkoğlu et al., 2019). In this chapter, the uncertainty set $\mathcal{D}_t$ is compact: The uncertainty set $\mathcal{D}_t$ is bounded, because $0 \leq d_{t,p} \leq \overline{d}_{t,p} + \widehat{d}_{t,p} \ \ \forall t, \forall p$, and is closed because $\mathcal{D}_t$ consists of a set of closed half spaces described by linear inequalities. However, the uncertainty is not row-wise because the uncertain parameter for period t, *i.e.* $d_{t,p}$, is found in the capacity constraint of each machine. The uncertainty would be row-wise if the demand for a product also depended on the machine, which is impossible.

Note that the problem also has two stages by nature: (i) Here-and-now: we decide about new qualifications before the realization of the demand, (ii) Wait-and-See: we decide about dispatching, modeled by variables $WIP_{t,r,m}$, after the observation of the demand. Investigating adjustable robust reformulations could therefore be relevant but is left for future research.

By considering the uncertainty set $\mathcal{D}_t$ to model the demand uncertainty, capacity constraints become in a static reformulation:

$$\sum_r \frac{(\sum_p rf_{p,r} d_{t,p}) WIP_{t,r,m}}{tp_{r,m}} \leq c_{t,m} u_{t,m}^{max} \qquad \forall t, \forall m, \forall \mathbf{d} \in \mathcal{D}_t$$

Robust counterpart: The next step consists in determining the robust counterpart of the capacity constraints. The robust counterpart is independently determined from one capacity constraint to another. Consider *one* capacity constraint for a given machine m at period t:

Step 1 (worst-case reformulation):

$$\max_{\mathbf{d} \in \mathcal{D}_t} \sum_r \frac{(\sum_p rf_{p,r} d_{t,p}) WIP_{t,r,m}}{tp_{r,m}} \leq c_{t,m} u_{t,m}^{max}$$

$$\max_{\mathbf{d} \in \mathcal{D}_t} \sum_p d_{t,p} (\sum_r \frac{rf_{p,r} WIP_{t,r,m}}{tp_{r,m}}) \leq c_{t,m} u_{t,m}^{max}$$

Intuitively, covering the worst-case realization in the uncertainty set $\mathcal{D}_t$ will conduct to add qualifications to machines for operations common to many products, or for operations associated to products with large re-entrant flow factors, as they are the operations that will impact the most the utilization rate of machines. In particular, operations with large demands and currently a single qualified machine are particularly constraining for a work center.

Step 2 (duality):

The next step consists in taking the dual of the inner maximization problem. The inner maximization problem and its dual, which is a minimization problem, have the same objective value because the inner maximization problem is linear. For a given period t, the following optimization problem must be solved:

$$\begin{aligned} \max \quad & \sum_p d_{t,p} (\sum_r \frac{rf_{p,r} WIP_{t,r,m}}{tp_{r,m}}) & & \\ \text{s.t.} \quad & d_{t,p} \geq \overline{d}_{t,p} - \widehat{d}_{t,p} & \forall p & \qquad (6.9) \end{aligned}$$

$$d_{t,p} \leq \overline{d}_{t,p} + \widehat{d}_{t,p} \qquad \forall p \qquad (6.10)$$

$$\sum_{p|\alpha_{p,f}=1} d_{t,p} \leq \Gamma_{t,f} \qquad \forall f \qquad (6.11)$$

The dual variables associated to Constraints (6.9)-(6.11) are listed in Table 6.1.

Table 6.1: Dual variables associated to constraints in the uncertainty set $\mathcal{D}_t$ for a capacity constraint (6.2).

Constraints	Dual variables
(6.9)	y_p^{min}
(6.10)	y_p^{max}
(6.11)	y_f^{gamma}

The dual of the inner maximization problem is a minimization problem. The minimization problem for a given capacity constraint for machine m at period t is modeled below:

$$\begin{aligned}
\min \quad & \sum_p (-(\overline{d}_{t,p} - \widehat{d}_{t,p}) y_p^{min}) + \sum_f (\Gamma_{t,f} y_f^{gamma}) \\
& + \sum_p ((\overline{d}_{t,p} + \widehat{d}_{t,p}) y_p^{max}) \\
\text{s.t.} \quad & -y_p^{min} + y_p^{max} + \sum_{f|\alpha_{p,f}=1} y_f^{gamma} \geq \sum_r \frac{rf_{p,r} WIP_{t,r,m}}{tp_{r,m}} && \forall p \\
& y_p^{min}, y_p^{max} \geq 0 && \forall p \\
& y_f^{gamma} \geq 0 && \forall f
\end{aligned}$$

Step 3 (Robust Counterpart): The final step consists in omitting the minimization term to obtain the robust counterpart. Therefore, the robust counterpart of the capacity constraint for a given machine m and a given period t can be found below:

$$\begin{aligned}
& \sum_p (-(\overline{d}_{t,p} - \widehat{d}_{t,p}) y_p^{min}) + \sum_f (\Gamma_f y_f^{gamma}) \\
& + \sum_p ((\overline{d}_{t,p} + \widehat{d}_{t,p}) y_p^{max}) \leq c_{t,m} u_{t,m}^{max} \\
& -y_p^{min} + y_p^{max} + \sum_{f|\alpha_{p,f}=1} y_f^{gamma} \geq \sum_r \frac{rf_{p,r} WIP_{t,r,m}}{tp_{r,m}} && \forall p \\
& y_p^{min}, y_p^{max} \geq 0 && \forall p \\
& y_f^{gamma} \geq 0 && \forall f
\end{aligned}$$

6.3.3.3 Robust Optimization Model

By deriving the robust counterpart for each capacity constraint and each time period and indexing the dual variables by period t and machine m, the overall robust optimization problem is:

$$\min \quad \sum_{t,r,m} \delta_t cq_{r,m} OQ_{t,r,m} \tag{6.12}$$

$$\text{s.t.} \quad (6.4)-(6.7)$$

$$\sum_{p}(-(\overline{d}_{t,p}-\widehat{d}_{t,p})y^{min}_{t,m,p}) + \sum_{f}(\Gamma_{t,f}y^{gamma}_{t,m,f}) + \sum_{p}((\overline{d}_{t,p}+\widehat{d}_{t,p})y^{max}_{t,m,p}) \leq c_{t,m}u^{max}_{t,m} \qquad \forall t, \forall m \tag{6.13}$$

$$-y^{min}_{t,m,p} + y^{max}_{t,m,p} + \sum_{f|\alpha_{p,f}=1} y^{gamma}_{t,m,f} \geq \sum_{r} \frac{rf_{p,r}WIP_{t,r,m}}{tp_{r,m}} \qquad \forall t, \forall m, \forall p \tag{6.14}$$

$$\sum_{m} WIP_{t,r,m} = 1 \qquad \forall t, \forall r \mid \sum_{p} rf_{p,r}(\overline{d}_{t,p}+\widehat{d}_{t,p}) > 0 \tag{6.15}$$

$$y^{min}_{t,m,p}, y^{max}_{t,m,p} \geq 0 \qquad \forall t, \forall m, \forall p \tag{6.16}$$

$$y^{gamma}_{t,m,f} \geq 0 \qquad \forall t, \forall m, \forall f \tag{6.17}$$

The objective function (6.12) minimizes the cost of performing qualifications, while Constraints (6.13)-(6.14) are the "robustification" constraints. Constraints (6.15) ensure that the demand by operation must be satisfied if there is demand. Note that Constraints (6.15) are slightly different from Constraint (6.3) as it must be active when $\sum_p rf_{p,r}(\overline{d}_{t,p} + \widehat{d}_{t,p}) > 0$ for operation r at period t instead of $\sum_p rf_{p,r}(\overline{d}_{t,p}) > 0$. Constraints (6.16)-(6.17) correspond to the non-negativity constraints introduced by the "robustification" procedure.

Note that the robust optimization model (6.12)-(6.15) can still be used when a product belongs to several product families. Only input parameters must be changed.

The robust optimization problem will referred as the Minimum Cost Robust Qualification Configuration Problem (MCRQCP) in the remainder of the chapter.

Similarly to MCQCP, it is important to mention that MCRQCP can be infeasible if the production capacities of machines are too small and if too few qualifiable pairs (operation, machine) exist to better balance the workload between the machines. Note that in the numerical experiments performed in Section 6.5, MCRQCP is infeasible for some values of $\overline{d}_{t,p}$ and $\widehat{d}_{t,p}$.

6.3.4 Illustrative example on tractability

MCQCP and MCRQCP can be both modeled with mixed integer linear programs, and thus can be solved by standard solvers. Although no new binary variables are

required in the robust reformulation of MCQCP, reformulating capacity constraints can modify the problem structure and introduce many more variables and constraints. The reformulation can also worsen the quality of linear relaxations, thus increasing the computational time required to reach an optimal solution in a branch and cut algorithm. It is then expected that MCRQCP requires more computational time to be solved than MCQCP. MCQCP has $T \times M + T \times R + 4 \times T \times R \times M$ constraints, $T \times R \times M$ continuous variables and $T \times R \times M$ binary variables. MCRQCP has $2 \times T \times M \times P + T \times M \times F$ more continuous variables and $3 \times T \times M \times P + T \times M \times F$ more constraints than MCQCP.

Table 6.2 illustrates the additional computational effort required to solve MCRQCP by reformulating capacity constraints with respect to MCQCP in terms of number of decision variables and constraints. The number of decision variables and constraints of MCQCP and MCRQCP are given for $P = 238$, $R = 1208$, $F = 3$, $M = 20$, $T = 7$. These values come from one work center (work center A) in the computational study. Assuming that the demand and worst-case demand are greater than 0 for all products and all periods, the increase of the number of continuous variables is equal to 16.6% and the increase of the number of constraints is equal to 12.8%. This makes the robust optimization problem more difficult to solve than the deterministic optimization problem as the robust optimization problem also tighten the capacity constraints. In practice, the robust optimization problem is much more difficult to solve than the deterministic optimization problem although most optimal solutions can be found in one hour (see Section 6.5.3.2).

Table 6.2: Comparison of the number of variables and constraints between MCQCP and MCRQCP. $P = 238$, $R = 1208$, $F = 3$, $M = 20$, $T = 7$.

	Optimization problem		
Number of	MCQCP	MCRQCP	Increase(%)
Continuous variables	169,120	202,860	16.6
Binary variables	169,120	169,120	0.0
Constraints	685,076	785,456	12.8

6.4 Characterizing the robustness of a set of qualifications

6.4.1 Motivation

Determining intuitively relevant values for $\widehat{d}_{t,p}$ can be difficult. The first option consists in using values estimated by decision-makers in charge of defining and predicting future demands. However, determining relevant values can be difficult for some products, in particular for new products because data can be insufficient.

If it is too difficult to provide relevant $\widehat{d}_{t,p}$ for each product, another option is to propose initial values for $\widehat{d}_{t,p}$. $\widehat{d}_{t,p}$ can first roughly initialized, *e.g.* initialized to

$\overline{d}_{t,p}$, and then refined by characterizing the robustness of a set of qualifications (typically the set of initial qualifications) with respect to the demand uncertainty. More precisely, characterizing the robustness of a set of qualifications means determining to what extent a work center is able to correctly absorb the demand uncertainty. Characterizing a set of qualifications is similar to determining the largest $\widehat{d}_{t,p}$ for each product p at period t. Therefore, determining the robustness of a set of qualifications provides capacity planners with the tolerated changes on the demand by a work center. Then, if possible, demand changes should be made in the bounds defined by $\overline{d}_{t,p}$ and $\widehat{d}_{t,p}$ to limit additional costs with outsourcing or new machines.

In the context of qualification management, Rossi (2010) and Aubry et al. (2012) assume that satisfying the demand by product is a key issue to characterize the robustness of a set of qualifications. Rossi (2010) seeks to characterize the robustness of a set of qualifications by determining the minimum additional quantity of products from the nominal demand that can be absorbed without the makespan exceeding a specified value. Robustness is defined as a distance in (Rossi, 2010). Similarly, Aubry et al. (2012) seeks to characterize the robustness of a set of qualifications by determining the largest perturbation from the nominal demand while ensuring that all machines have the same workload and that qualification costs do not exceed a predefined value. The L-1 norm is used. Similarly to Rossi (2010) and Aubry et al. (2012), we assume that satisfying the demand by product is a key issue when characterizing the robustness of a set of qualifications. The major differences with Rossi (2010) and Aubry et al. (2012) are that: (1) We do not assume that machines are uniform or related; (2) We consider large scale production systems with hundreds of products and thousands of operations; (3) Product cannibalization and correlated demands are considered. To characterize the robustness of a set of qualifications, we resort to robust optimization and the uncertainty set $\mathcal{D}_t$. More precisely, we seek to determine to what extent a set of qualifications is able to absorb the demand uncertainty when it is described by the uncertainty set $\mathcal{D}_t$. Assessing the robustness of a set of qualifications depends on the *utility function* used to evaluate it. First, we propose a generic mathematical model to model the robustness of a set of qualifications with respect to the demand uncertainty. Second, we propose a solution approach, based on a binary search approach, to determine the robustness of a set of qualifications.

6.4.2 Problem statement

The problem is mostly identical to the problem introduced in Section 6.3.1. The only difference is that the objective is to characterize the robustness of a set of qualifications. This problem will be referred as the Maximum Robustness Budgeted Qualification Problem (MRBQP) in the remainder of the chapter.

6.4.2.1 Problem modeling

Let us introduce a new decision variable $\theta_{t,p} \geq 0 \ \ \forall t, \forall p$ that is used to evaluate the robustness of a set of qualifications. Let us assume that $d_{t,p}$ is a random variable that depends on $\theta_{t,p}$: $d_{t,p} \in [\overline{d}_{t,p} - \widehat{d}_{t,p}\theta_{t,p}, \overline{d}_{t,p} + \widehat{d}_{t,p}\theta_{t,p}] \ \ \forall t, \forall p$. Formally, the problem can

be modeled as follows:

$$\max \quad f(\boldsymbol{\theta}) \tag{6.18}$$

$$\text{s.t.} \quad \sum_{r} \frac{(\sum_{p} rf_{p,r} d_{t,p}) WIP_{t,r,m}}{tp_{r,m}} \leq c_{t,m} u_{t,m}^{max} \qquad \forall t, \forall m, \forall \mathbf{d} \in \mathcal{D}_t(\boldsymbol{\theta}) \tag{6.19}$$

$$(6.3) - (6.6)$$

$$\theta_{t,p} \leq \frac{\overline{d}_{t,p}}{\widehat{d}_{t,p}} \qquad \forall t, \forall p \tag{6.20}$$

$$\theta_{t,p} \geq 0 \qquad \forall t, \forall p \tag{6.21}$$

The objective function (6.18) maximizes a utility function of $\boldsymbol{\theta}$. The capacity constraints (6.19) depend on $\boldsymbol{\theta}$, which are used to control the demand uncertainty. Constraints (6.20) ensure that the demand by product cannot be negative. Finally, Constraints (6.21) are the non-negativity constraints.

Solving MRBQP is equivalent to determining the robustness of the initial set of qualifications, or any set of qualifications as input parameter.

6.4.2.2 Robust counterpart of capacity constraints

MRBQP is an optimization problem under *decision-dependent uncertainty* because the capacity constraints depend on the matrix $\boldsymbol{\theta}$ used to control the demand uncertainty to cover. Optimization problems under decision-dependent uncertainty are known to be difficult to solve. When the uncertainty set is polyhedral, the objective function is linear and the constraints are linear, the optimization problem is NP-Complete (Nohadani and Sharma, 2018; Lappas and Gounaris, 2018).

Nevertheless, as in the classical robust optimization paradigm, it is possible to reformulate decision-dependent uncertainty constraints with duality. This is because $\boldsymbol{\theta}$ is not a decision variable of the inner robust maximization problem. Let us consider the same uncertainty set in Equation (6.8). The only difference stems from the fact that the plausibility limits of $d_{t,p}$ are now dependent on $\theta_{t,p}$. Similarly to Section 6.3.3.2, it is possible to "robustify" the capacity constraints (6.19).

Consider the capacity constraints of a given machine m at period t. We follow the same procedure as the one in Section 6.3.3.2, and the same notations for dual variables are used. Steps 1 and 2 are similar to Section 6.3.3.2. The robust counterpart of the capacity constraint for machine m at period t is written below:

$$\sum_{p}(-(\overline{d}_{t,p} - \widehat{d}_{t,p}\theta_{t,p})y_p^{min}) + \sum_{f}(\Gamma_{t,f} y_f^{gamma}) + \sum_{p}((\overline{d}_{t,p} + \widehat{d}_{t,p}\theta_{t,p})y_p^{max}) \leq c_{t,m} u_{t,m}^{max} \tag{6.22}$$

$$-y_p^{min} + y_p^{max} + \sum_{f|\alpha_{p,f}=1} y_f^{gamma} \geq \sum_{r} \frac{rf_{p,r} WIP_{t,r,m}}{tp_{r,m}} \qquad \forall p \tag{6.23}$$

$$y_p^{min}, y_p^{max} \geq 0 \qquad \forall p \tag{6.24}$$

$$y_f^{gamma} \geq 0 \qquad \forall f \qquad (6.25)$$

6.4.2.3 Robust reformulation

By deriving the robust counterpart for each capacity constraint, it is possible to write the robust reformulation of MRBQP below:

$$\max \quad f(\boldsymbol{\theta}) \qquad (6.26)$$

$$\text{s.t.} \quad (6.4)-(6.7)$$

$$\sum_p (-(\overline{d}_{t,p} - \widehat{d}_{t,p}\theta_{t,p}) y_{t,m,p}^{min}) + \sum_f (\Gamma_{t,f} y_{t,m,f}^{gamma}) + \sum_p ((\overline{d}_{t,p} + \widehat{d}_{t,p}\theta_{t,p}) y_{t,m,p}^{max}) \leq c_{t,m} u_{t,m}^{max} \quad \forall t, \forall m \qquad (6.27)$$

$$\sum_m WIP_{t,r,m} = 1 \qquad \forall t, \forall r \mid \sum_p rf_{p,r}(\overline{d}_{t,p} + \widehat{d}_{t,p}) > 0 \qquad (6.28)$$

$$-y_{t,m,p}^{min} + y_{t,m,p}^{max} + \sum_{f|\alpha_{p,f}=1} y_{t,m,f}^{gamma} \geq \sum_r \frac{rf_{p,r} WIP_{t,r,m}}{tp_{r,m}} \qquad \forall t, \forall m, \forall p \qquad (6.29)$$

$$y_{t,m,p}^{min}, y_{t,m,p}^{max} \geq 0 \qquad \forall t, \forall m, \forall p \qquad (6.30)$$

$$y_{t,m,f}^{gamma} \geq 0 \qquad \forall t, \forall m, \forall f \qquad (6.31)$$

$$\theta_{t,p} \leq \frac{\overline{d}_{t,p}}{\widehat{d}_{t,p}} \qquad \forall t, \forall p \qquad (6.32)$$

$$\theta_{t,p} \geq 0 \qquad \forall t, \forall p \qquad (6.33)$$

The objective function (6.26) maximizes the robustness of a set of qualifications. Constraints (6.27)-(6.31) correspond to the "robustification" constraints. They ensure that the capacity constraints must be respected for any realization in the uncertainty set $\mathcal{D}_t$. Constraints (6.32) are the constraints for the upper bound $\theta_{t,p}^{min}$. Finally, Constraints (6.33) are the non-negativity constraints.

Solving MRBQP leads to determining the largest $\theta_{t,p}$ for product p at period t, and consequently to characterize the robustness of a set of qualifications. The main drawback of MRBQP is that it is computationally challenging to solve. This is because MRBQP contains products of variables, $\theta_{t,p}$ and $y_{t,m,p}^{max}$, and $\theta_{t,p}$ and $y_{t,m,p}^{min}$, which are introduced by the "robustification" procedure for capacity constraints. Existing MILP reformulations are possible when one the variables is a binary variable (Nohadani and Sharma 2018; Lappas and Gounaris 2018). To determine an estimate of the robustness of a set of qualifications, a binary search solution approach is presented in Section 6.4.3.

6.4.3 Binary search approach

To characterize the robustness of a set of qualifications, it is possible, for each period, to maximize $\theta_{t,p}$ assuming that $\theta_{t,p} = \theta_t \ \forall p$. For this objective, Algorithm 7, which a binary search like algorithm, can be used. The initial value for $\widehat{d}_{t,p}$ is $\widehat{d}^0_{t,p}$.

Algorithm 7 Binary search

Input data: $\widehat{\boldsymbol{d}}^{\boldsymbol{0}}$

procedure BINARY SEARCH
  $\widehat{d}_{t,p} \leftarrow \widehat{d}^0_{t,p} \ \forall t, \forall p$
  $\theta^{max}_t \leftarrow \min_p \frac{\overline{d}_{t,p}}{\widehat{d}_{t,p}} \ \forall t$
  $\theta^{min}_t \leftarrow 0 \ \forall t$
  $\theta_t \leftarrow 0 \ \forall t$
  for $i = 1$ to T **do**
    $\theta_i \leftarrow \frac{\theta^{max}_i + \theta^{in}_i}{2}$
    while $\theta^{max}_i > \epsilon$ and $\frac{\theta^{max}_i - \theta^{min}_i}{\theta^{max}_i} > \epsilon$ **do**
      Verify that MRBQP is feasible for $\boldsymbol{\theta}$ at period t (no capacity constraint violation at period t)
      if feasible **then**
        $\boldsymbol{\theta^{min}} \leftarrow \boldsymbol{\theta}$
      else
        $\boldsymbol{\theta^{max}} \leftarrow \boldsymbol{\theta}$
      end if
      $\theta_i \leftarrow \frac{\theta^{max}_i + \theta^{in}_i}{2}$
    end while
  end for
  return $\boldsymbol{\theta^{min}}$
end procedure

The computational difficulty in Algorithm 7 comes from solving multiple large-scale linear programs. The computational burden can be lowered by warm-starts as only the coefficients $y^{min}_{t,m,p}$, $y^{max}_{t,p,m}$, and $y^{gamma}_{t,m,f}$ variables in the "robustification" constraints must be changed.

Note that if θ is assumed to be identical for all periods and products, Algorithm 7 returns the smallest θ^{min} over all periods. From a practical standpoint, some products can be filtered out of Algorithm 7 if there is no uncertainty on the product, or if the uncertainty on the product does not need to be covered.

If Algorithm 7 is run when all new qualifications are started at $t = 0$, then an ideal value of $\boldsymbol{\theta}$ is computed. This is the largest value of θ for which the demand uncertainty can be covered in the work center. Reporting this value is interesting for capacity planners to assess the robustness of the work center against an ideal situation.

6.5 Computational study

The computational study is performed to answer the following questions: What is the price of uncertainty? Is it risky to use the set of qualifications determined by considering only the nominal demand? Is the robust optimization problem difficult to solve?

In Section 6.5.1, the instances used for the computational study and generated from industrial data are described. For confidentially purposes, raw values by product, by operation, by product family and by machine of parameters are not provided. Instead, means, minimums, maximums and standard deviations are presented. In Section 6.5.2, the design of experiments is presented, and the numerical results in Section 6.5.3. We show that the price of uncertainty, defined by comparing the number of qualifications determined for the robust optimization problem and for the deterministic optimization problem, when the demand is fully known (perfect handsight), is actually very small. Moreover, in a large number of experiments, the robustness of the set of qualifications determined by solving the deterministic optimization problem with the nominal demand is far from the robustness of the set of qualifications determined by solving the robust optimization problem whereas both qualification matrices have about the same number of qualifications. In addition, we show that only considering the nominal demand can lead to a large number of capacity constraint violations. The computational study highlights that selecting the right qualifications is more important for robustness than the number of qualifications.

6.5.1 Instance generation

Work center: The computational study is performed by using industrial data from a semiconductor factory located at Crolles, France. Two critical work centers, work center A and work center B, of the factory are considered. Work center A has $M = 20$ machines. Work center B has $M = 30$ machines.

Demand: A horizon of 7 periods, *i.e.* $T = 7$, is considered. Each period corresponds to one month. The nominal demand by product is given by internal forecasts for each period of the horizon. Exact demand values are not provided for confidentiality reasons. Instead, Table 6.3 illustrates the number of products with a strictly positive demand by period and the Coefficient of Variability (CV) of the demand by period. On the horizon, there are in total 238 products, *i.e.* $P = 238$. For work center A, these 238 products lead to 1,208 operations, *i.e.* $R = 1,208$. For work center B, these 238 products lead to 401 operations. There is no uncertainty on the demand for the first month.

Production capacities: For work center A, $u_{t,m}^{max} = 0.95 \ \forall t, \forall m$ in the industrial data. Consider a given period t. For work center B, the mean of $u_{t,m}^{max}$ is equal to 0.80, the minimum of $u_{t,m}^{max}$ to 0.63, the maximum of $u_{t,m}^{max}$ to 0.87, and the standard deviation of $u_{t,m}^{max}$ to 0.079. Both work centers do not have the same values for $u_{t,m}^{max}$ because machine types are completely different. Note that $u_{t,m}^{max}$ is constant from one period to another. Similarly, the production capacity by machine $c_{t,m}$ is constant from one period to another, but is different from one machine to another.

This is mainly because machines are non-identical and are of different ages and generations. Values for $c_{t,m}$ are given based on the length of period t. For work center A, the mean of $c_{t,m}$ is equal to 59% of the length of the period, the minimum of $c_{t,m}$ to 44%, the maximum to 66%, and the standard deviation to 6%. For work center B, the mean of $c_{t,m}$ is set to 75%, the minimum to 36%, the maximum to 85%, and the standard deviation to 9%. $c_{t,m}$ is not equal to 100% because machines have capacity losses, *e.g.* due to maintenance operations, engineering operations, setup times.

Re-entrant flow factors: For work center A, the re-entrant flow factors vary between 14 and 72, with a mean of 41.2 and a standard deviation of 11.0. For work center B, the re-entrant flow factors vary between 1 and 28, with a mean of 16.0 and a standard deviation of 4.3.

Product families: There are three product families, *i.e.* $F = 3$. Each product belongs to exactly one product family. The first product family contains 120 products. The second product family contains 64 products. The third product family contains 54 products.

Qualification matrix: The initial set of qualifications is partially initialized, in particular because some machines are already qualified for existing operations. Consider work center A. The mean number of qualified machines by operation is equal to 4.2, and the standard deviation to 2.0. The minimum, respectively maximum, number of qualified machines for an operation is equal to 1, respectively 13. The mean number of qualified operations by machine is equal to 251.3, and the standard deviation to 188.0. The minimum, respectively maximum, number of qualified operations for a machine is equal to 25, respectively 645. Note that some operations cannot be qualified on some machines due to technological restrictions. In total, 2,843 new qualifications are possible in work center A. Consider work center B. The mean number of qualified machines by operation is equal to 3.5, and the standard deviation to 1.6. The minimum, respectively maximum, number of qualified machines for an operation is equal to 1, respectively 6. The mean number of qualified operations by machine is equal to 48.0, and the standard deviation to 43.8. The minimum, respectively maximum, number of qualified operations for a machine is equal to 0, respectively 130. Note that some operations cannot be qualified on some machines due to technological restrictions. In total, 1,266 new qualifications are possible in work center B. Some machines have no qualified operations because they are being started up.

Qualification costs: We could not access to the qualification costs. Therefore, we assume that all qualification costs are identical and equal to one. This is a common assumption made by capacity planners in practice. Hence, in the computational study, the number of qualifications to perform must be minimized.

Qualification lead times: Qualification lead times are rough estimates of the lead times to perform the qualification procedures. Qualification lead times vary between several days and two months. For qualification lead times that are smaller than 2 weeks, $l_{r,m}$ is set to 0 because the considered period in the computational study is one month. Consider work center A. The minimum lead time for all operations and machines is equal to 0 period, the mean to 1.6, the standard deviation to 0.8, and the maximum to 2. Consider work center B. The minimum

lead time for all operations and machines is equal to 0 period, the mean to 1.1, the standard deviation to 0.4, and the maximum to 2.
Throughput rates: Throughput rates strongly vary from one machine to another and from one operation to another. Consider work center A. The minimum throughput rate for all operations and machines is equal to 11.4 wafers per hour, the mean to 221.6, the standard deviation to 126.7, and the maximum to 527.8. Consider work center B. The minimum throughput rate for all operations and machines is equal to 6.8 wafers per hour, the mean to 48.0, the standard deviation to 13.9, and the maximum to 83.3.

Table 6.3: Nominal demand by month.

	Month						
	1	2	3	4	5	6	7
CV	2.78	1.97	2.88	2.29	2.06	3.58	3.09

6.5.2 Design of experiments

We consider that $\widehat{d}_{t,p} = \theta \overline{d}_{t,p} \ \ \forall t, \forall p$, *i.e.* θ is assumed to be identical for all periods and products. Different values of θ are studied: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7. Values of 0.2 and 0.3 for θ is not unusual even for early periods of the horizon for high mix factories. Larger values of θ are not considered because the robust optimization problem becomes infeasible from $\theta = 0.770$ for work center A (see Section 6.5.3). In addition, the robust optimization problem becomes infeasible from $\theta = 0.294$ for work center B. The budget of uncertainty $\Gamma_{t,f}$ is set to $\sum_{p|\alpha_{p,f}=1} \overline{d}_{t,p}, \ \ \forall t, \forall f$. The discount factor δ_t is set to 1 $\ \forall t$ in numerical experiments. This means that there are no incentives on performing qualifications as soon as possible or as late as possible. In Algorithm 7, $\widehat{d}^0_{t,p} = \overline{d}_{t,p}, \ \ \forall t, \forall p$.

In the experiments, MCQCP is solved once. The robustness of the optimized set of qualifications is evaluated with Algorithm 7. Then, for each possible value of θ, MCRQCP is solved. For each value of θ, 3,600 demand scenarios are generated to evaluate the capacity constraint violations if the nominal set of qualifications was considered, and the price of uncertainty. Because the true distribution of the demand is unknown and the demand between products is correlated, scenarios are randomly generated by using a linear program. The linear program, described in Appendix D.1, . The linear program generates for a given θ a scenario on the demand by product and by period for a given demand level $\eta_{t,f}$ by product family and by period. In the experiments, it is assumed that $\eta_{t,f}$ is equal to the nominal demand by product family.

For the sake of presentation, in the remainder of the computational study, the set of qualifications determined by solving MCQCP for the nominal demand are called nominal qualifications, the set of qualifications determined by solving MCQCP for

the perfect handsight demand scenario, perfect handsight qualifications, and the set of qualifications determined by solving MCRQCP, robust qualifications.

Note that the robust and nominal qualifications are not compared in a rolling horizon in the computational study, *i.e.* where qualifications could be updated at each period after demand realizations for the following reasons: (1) It is difficult to known the final practical decision when capacity constraint violations occur; (2) Qualification decisions must be anticipated due to long qualification processes; (3) Comparing the robust and nominal set of qualifications is possible and fair because both are computed from "static" optimization models.

6.5.3 Numerical results

Mathematical models and Algorithm 7 are implemented in Java 8 on a computer with an Intel Xeon CPU W3530 running at 2.80GHz with 8 threads and 12GB of RAM. Mathematical models are solved by using the solver IBM ILOG CPLEX 12.9 with default parameters. A computational time limit of one hour is given to the solver, ϵ is set to 0.0001 in Algorithm 7.

Section 6.5.3.1 answers the question "What is the price of uncertainty?", Section 6.5.3.2 the question "It the robust optimization problem difficult to solve?", and Section 6.5.3.3, the question "Is it risky to use the set of qualifications determined by considering only the nominal demand?"

6.5.3.1 What is the price of uncertainty?

The Price of Uncertainty (PoU) is computed by comparing the number of robust qualifications and the number of perfect handsight qualifications. Gorissen et al. (2015) argue that a low mean PoU and standard deviation indicate a good robust solution. Table 6.4 shows the mean PoU, its standard deviation (std.) and its maximum afor each value of θ. Note that as for $\theta > 0.294$, MCRQCP is infeasible for work center B, PoU is not presented.

Consider work center A. The mean PoU varies between 1.08 qualifications on average for $\theta = 0.1$, with a standard deviation of 0.30, and 31.99 qualifications on average for $\theta = 0.7$ with a standard deviation of 2.03. Note that the increase of PoU when θ increases is mainly due to the fact that the number of robust qualifications increases (see Figure 6.2). The standard deviation of PoU is small with respect to the mean PoU. To better put into perspective, the meaning of about 30 qualifications, consider $\theta = 0.7$. In the worst case, PoU is equal to 35. Recall that the number of machines in work center A is equal to 20. In other words, to cover the demand uncertainty, it is required to add $\frac{35}{20} = 1.75$ qualifications on average to each machine, each having a few hundred qualifications on average, which seems acceptable in practice. Therefore, robust qualifications for work center A appear to be good solutions. In addition, a small number of additional qualifications, in the worst case 35, is required to cover the demand uncertainty. This is small compared to the 2,843 possible new qualifications. This suggests that it is possible to be robust by performing the right qualifications. Robust qualifications are also relevant because they can avoid capacity constraint violations contrary to nominal qualifications (see Section 6.5.3.3).

Similar observations can be observed for work center B (see Table 6.4 and Figure 6.3). The maximum PoU varies between 5 and 19 qualifications. Similarly to work center A, to better put into perspective the meaning of a PoU of 19 qualifications, recall that the number of machines in work center B is equal to 30. With respect to the perfect handsight qualifications, to cover the demand uncertainty, it is required to add $\frac{19}{30} = 0.63$ qualifications on average to each machine, each having a few tens of qualifications on average. This is also small compared to the 1,266 possible new qualifications. This again suggests that it is possible to be robust by performing the right qualifications.

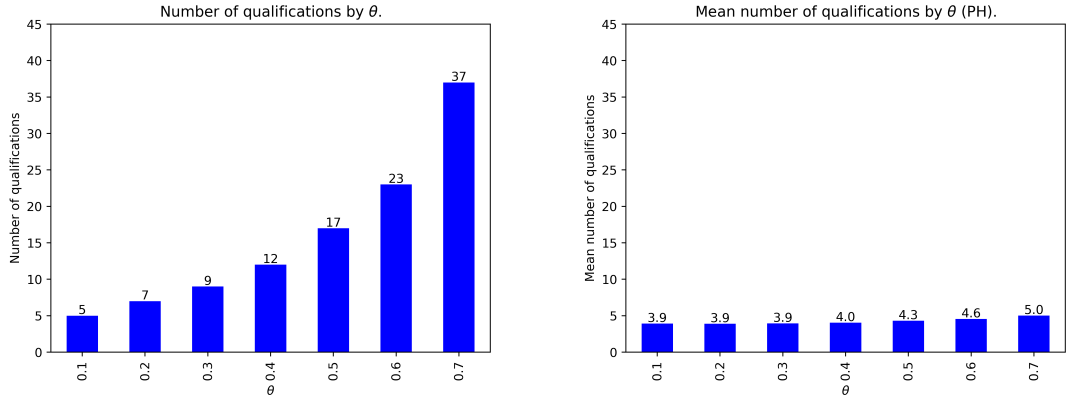
Implementing perfect handsight qualifications is impossible because it is impossible to know in advance the demand realizations. A more practical price of uncertainty can be computed by comparing the number of robust qualifications and nominal qualifications. We found that the actual price of uncertainty is close to the PoU presented in Table 6.4. This is because, for both work centers, the number of nominal qualifications is equal to 4 and the mean number of perfect handsight scenarios is also close to 4 (see Figures 6.2 and 6.3).

Now consider the case where $\theta = \theta^{max}$, where θ^{max} is the largest possible value pf θ for the considered work center. It can be computed by running Algorithm 7 when all new qualifications are started at $t = 0$. For work center A, this gives $\theta^{max} = 0.77$. When MCRQCP is solved for $\theta = \theta^{max}$, 96 new qualifications are required (the set of new qualifications is optimal). $\frac{96}{2{,}843} \times 100 = 3.37\%$ of all new possible qualifications are required to reach the same robustness than the one when all qualifications are performed. For work center B, $\theta^{max} = 0.294$. When MCRQCP is solved for $\theta = \theta^{max}$, 135 new qualifications are required (optimality gap of 25.0% after 3,600 seconds). $\frac{135}{1{,}266} \times 100 = 10.6\%$ of all new possible qualifications are required to reach the same robustness than the one when all qualifications are performed. In the best case, $\lceil 135 - 0.25 \times 135 \rceil = 102$ new qualifications are required, which corresponds to 8.05% of all possible new qualifications. This further suggests that it is possible to be robust by performing a limited number of qualifications. In other words, it can be ineffective to add many qualifications, if they are irrelevant. Similar observations can be found in other contributions on flexibility, *e.g.* on the long-chain and closed-chain principles (Jordan and Graves, 1995; Chou et al., 2010). Thus, relevant qualifications must be carefully optimized and planned to immunize a work center against demand uncertainty.

One of the reasons why PoU is small is that qualification costs are assumed identical in the computational study, which is in fact a common assumption in practice. PoU could potentially be larger if qualification cost profiles are different from one machine to another and from one operation to another. Nevertheless, PoU is not necessarily expected to be significantly larger since new qualifications must be paid in the nominal, perfect handsight and robust cases for the following reasons: (1) Qualifications are made for *new* operations or *new* machines, or existing operations that have never been qualified on existing machines and (2) A ramp-up demand for a product, even uncertain, implies adding new qualifications to machines to increase product capacity and balance the workload between the machines. If new qualifications are not performed, then it is impossible to satisfy the demand, and both MCRQCP and MCQCP are infeasible. If qualification cost profiles are very dif-

Table 6.4: Price of Uncertainty (PoU).

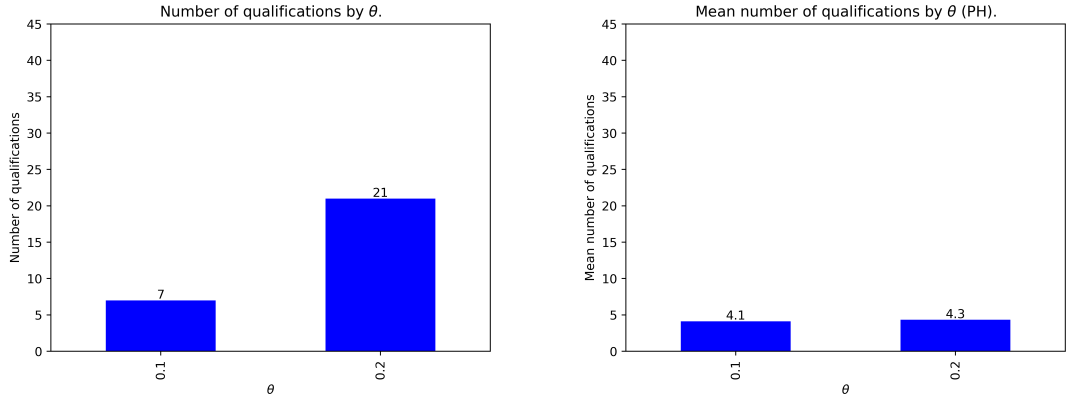
	Work center A			Work center B		
θ	Mean	Std.	Max.	Mean	Std.	Max.
0.1	1.08	0.30	2	2.88	0.41	5
0.2	3.10	0.63	4	16.66	0.94	19
0.3	5.05	0.85	7	-	-	-
0.4	7.96	1.11	10	-	-	-
0.5	12.70	1.41	15	-	-	-
0.6	18.44	1.68	21	-	-	-
0.7	31.99	2.03	35	-	-	-



(a) Number of robust qualifications by θ.

(b) Mean number of qualifications by θ (PH).

Figure 6.2: Work center A. Number of qualifications by θ.



(a) Number of robust qualifications by θ.

(b) Mean number of qualifications by θ (PH).

Figure 6.3: Work center B. Number of qualifications by θ.

ferent, it may be possible to keep a small PoU by performing a lot of inexpensive qualifications and avoid performing expensive qualifications whenever possible. Finally, PoU is also small because of product cannibalization that limits the overall demand of products.

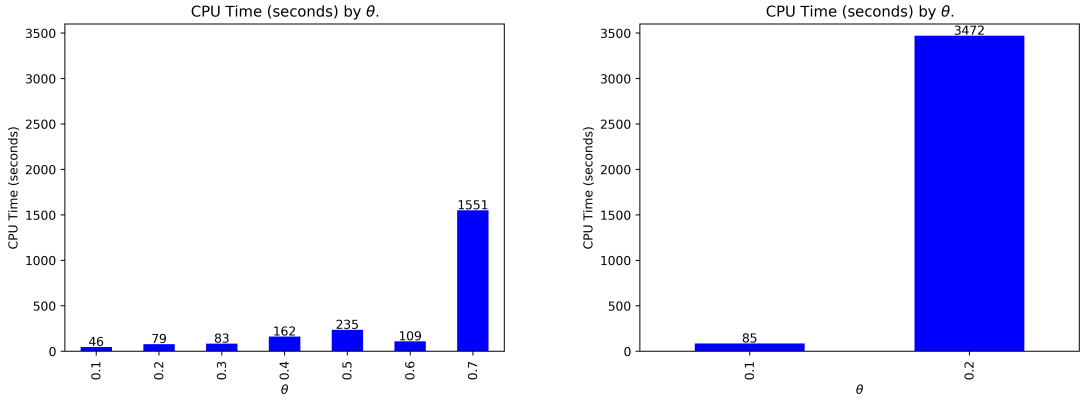
Now assume that the manufacturer faces an extreme case where too many qualifications must be performed to cover the demand uncertainty with respect to the number of nominal qualifications. This information is still valuable for capacity planners because they will have to refine plausibility limits to limit additional outsourcing and machine purchasing costs. In this situation, MRBQP is relevant to help refining plausibility limits.

Finally, from a practical standpoint, as both work centers are located in the same factory, covering the demand uncertainty for θ larger than 0.3 in work center A is probably unnecessary as θ^{max} is equal to 0.294 for work center B.

6.5.3.2 Is the robust optimization problem difficult to solve?

Consider work center A. For all values of θ, a set of optimal robust qualifications is determined. However, determining optimal robust qualifications is much more time consuming than determining optimal nominal qualifications (about 3 seconds). Similarly, determining optimal perfect handsight qualifications requires between 2 and 6 seconds. Determining optimal robust qualifications requires between 46 seconds for $\theta = 0.1$ and 1,551 seconds for $\theta = 0.7$ (see Figure 6.4). For θ^{max}, the optimal set of robust qualifications is determined in 656 seconds. It is also worth mentioning that all optimal nominal qualifications are determined at the root node by IBM ILOG CPLEX. Except for $\theta = 0.4, 0.5, 0.7$ and $\theta = \theta^{max}$, all robust qualifications are also determined at the root node by IBM ILOG CPLEX. This can be explained by the fact that modern solvers such as IBM ILOG CPLEX embed advanced preprocessing, probing, heuristic and cutting plane routines that are used to strengthen the linear relaxation of mixed integer linear problems (see *e.g.* Savelsbergh, 1994; Atamtürk et al., 2000) and quickly to determine good solutions. It can also be observed that it is faster to get the optimal robust qualifications for $\theta = 0.6$ than for $\theta = 0.5$.

For work center B, determining nominal qualifications takes about 1 second, while, similarly to work center A, determining optimal robust qualifications is more difficult. For $\theta = 0.1$, optimal robust qualifications are determined in 85 seconds (see Figure 6.4), and in 3,472 seconds for $\theta = 0.2$. Branching in IBM ILOG CPLEX is required for both $\theta = 0.1$ and $\theta = 0.2$.



(a) Work center A. (b) Work center B.

Figure 6.4: Computational time (in seconds) required to determine the set of robust qualifications by θ.

6.5.3.3 Is it risky to use the set of qualifications determined by only considering the nominal demand?

The numerical experiments show that it can be risky to implement nominal qualifications because it can lead to capacity constraint violations, which are computed with the following procedure:

1. A demand scenario is generated with the linear program in Appendix D.1.
2. Then, for the set of nominal qualifications and the generated demand, the Total Overtime (OT) is minimized with the linear program (D.5)-(D.10) in Appendix D.2.
3. If $OT > 0$, then there is at least one capacity constraint violation for the considered scenario. In this case, to put into perspective what a positive overtime means, in particular in terms of machine utilization rates, we solve the nonlinear utilization balancing optimization problem of Chapter 2. This avoids the problem where the total overtime for a period is set to a specific machine whereas, in practice, it would be balanced with similarly qualified machines. The utilization balancing exponent γ is set to 20 in this chapter.

For scenario i, the procedure enables us to determine what would be the utilization rate $U^i_{t,m}$ of machine m at period t for a given demand by product and a given set of qualifications (here the nominal qualifications) if there is a capacity constraint violation. If $U^i_{t,m} > u^{max}_{t,m}$, then there is a capacity constraint violation for scenario i. Repeating this procedure for the 3,600 scenarios enables us to estimate the capacity constraint violations if only nominal qualifications were implemented.

Table 6.5 shows capacity constraint violations. Column "A" corresponds to the percentage of scenarios where there is at least one capacity constraint violation, Column "B" to the number of capacity constraint violations. Mathematically, the mean number of capacity constraint violations is computed as follows: $\frac{1}{3{,}600\times T\times M\times I}\sum_{i=1}^{3{,}600}\mathbb{1}(U^i_{t,m}-u^{max}_{t,m})$, where $\mathbb{1}(x)=1$ if $x>0$, and 0 otherwise and

$I = 3,600$. The maximum (max.) number of capacity constraint violations is computed as follows: $\max_i(\sum_{t,m} \mathbb{1}(U^i_{t,m} - u^{max}_{t,m}))$. Column "C" quantifies capacity constraint violations when there is at least one capacity constraint violation. Mathematically, the mean capacity constraint violation is computed as follows: $\frac{1}{3,600 \times T \times M \times I} \sum_{i=1}^{3,600} (\sum_{t,m} \max(0, U^i_{t,m} - u^{max}_{t,m}))$, and the maximum capacity constraint violation is computed as follows $\max_{i,t,m} \max(0, U^i_{t,m} - u^{max}_{t,m})$. Columns "B" and "C" are computed only if there is at least one capacity constraint violation.

Table 6.5: Capacity constraint violations.

	Work center A					Work center B				
	A	B		C		A	B		C	
θ		Mean	Max.	Mean	Max.		Mean	Max.	Mean	Max.
0.1	0.72%	0.058	8	0.004	0.010	15.56%	7.096	11	0.010	0.029
0.2	15.64%	1.241	8	0.010	0.029	44.28%	8.947	22	0.014	0.048
0.3	26.19%	2.080	12	0.015	0.046	-	-	-	-	-
0.4	30.78%	2.670	13	0.020	0.066	-	-	-	-	-
0.5	38.39%	3.693	24	0.024	0.086	-	-	-	-	-
0.6	45.31%	5.058	36	0.027	0.106	-	-	-	-	-
0.7	57.83%	7.704	51	0.030	0.126	-	-	-	-	-

Consider work center A, $\theta = 0.1$, 0.72% of the scenarios have a capacity constraint violation (see Table 6.5), *i.e.* a relatively small number of scenarios. In addition, the number of capacity constraint violations is relatively small. In the worst case, 8 out of 140 ($T = 7$ and $M = 20$) capacity constraints are violated, capacity constraint violations are not very large, on average 0.004 and at most 0.010. This means that if the maximum utilization rate of a machine was set to 0.95, then on average, its real utilization rate would be equal to 0.9504, at most 0.96. Therefore, for $\theta = 0.1$, using the nominal qualifications is probably acceptable. For $\theta = 0.2$, 15.64% of the scenarios have a capacity constraint violation, which is significantly larger than for $\theta = 0.1$. On average, the capacity constraint violation is equal to 0.010, and in the worst case to 0.029, which starts to be appreciable. For larger values of θ, using nominal qualifications is more risky. For instance, consider $\theta = 0.4$ where 30.78% of the scenarios have at least one capacity constraint violation (see Table 6.5). In the worst case, 13 out of 140 capacity constraints are violated. In addition, the largest capacity constraint violation is equal to 0.066. This means, that if the maximum utilization rate of a machine was set to 0.95, then its real utilization rate would be equal to 1.003. The same observations can be made for larger values of θ. Utilization rates near 1.0 are not sustainable in terms of service levels. This is due to the fact that the cycle time increases almost exponentially with the utilization rate (queuing theory) and due to production variability (Hopp and Spearman, 2011). In other words, even small capacity constraint violations should be avoided.

Capacity constraint violations are more critical for work center B than for work center A. For $\theta = 0.1$, 15.56% of the scenarios lead to at least one capacity constraint violation, and for $\theta = 0.2$, 44.28% of the scenarios. In the worst case, there are 11

capacity constraint violations for $\theta = 0.1$ and 22 capacity constraint violations for $\theta = 0.2$. For $\theta = 0.1$, the mean capacity constraint violation is equal to 0.010. $u_{t,m}^{max}$ is set to low values (compared to work center A) because it is known that, in the industrial context, small increases of utilization rates can lead to much larger cycle times due to production variability.

Using nominal qualifications can lead to capacity constraint violations because nominal qualifications are not robust against demand uncertainty, and are in fact much less robust than robust qualifications. For work center A, Algorithm 7 for the nominal qualifications gives $\theta = 0.043$, and $\theta = 0.024$ for work center B. With a limited number of additional qualifications, robust qualifications lead to a much better robustness (see Section 6.5.3.1). Consider work center A and $\theta = 0.2$, 7 robust qualifications are required instead of 4 nominal qualifications to avoid capacity constraint violations in 15.64% of the scenarios. For $\theta = 0.3$, 9 robust qualifications are sufficient to avoid capacity constraint violations in 26.19% of the scenarios. Similar observations can be made for work center B. For instance, for $\theta = 0.1$, 7 robust qualifications are required instead 4 of nominal qualifications to avoid capacity constraint violations in 15.56% of the scenarios. It is worth mentioning that robust qualifications are more robust against demand uncertainty because more qualifications are performed. Nevertheless, even by adding a large number of qualifications, the nominal qualifications are still outperformed by the robust qualifications in terms of demand uncertainty coverage. Let us consider the case where α-flexibility designs are enforced when nominal qualifications are determined by the optimization model (6.1)-(6.7). An α-flexibility design enforces that at least α machines must be qualified by operation. For operations where it is not possible to have α qualified machines, the number of largest number of qualified machines is enforced. An α-flexibility design is enforced by adding the two following constraints: (1') $\sum_{t'=1|t'+l_{r,m}\leq t}^{t} OQ_{t',r,m} \leq 1 \ \forall t, \forall r, \forall m$, (2') $\sum_{m} \sum_{t'=1|t'+l_{r,m}\leq t}^{t} OQ_{t',r,m} \geq \min(\alpha, \alpha') - \alpha'' \ \forall t, \forall r$. α' is the the number of qualifiable and qualified couples (operation, machine) for a given period, and α'' is the number of qualified couples (operation, machine) for a given period. Constraints (1') are required, otherwise Constraints (2') could be satisfied by performing the new qualification at different periods. Table 6.6 shows that enforcing α-flexibility designs for the nominal qualifications does not lead to a better robustness against demand uncertainty than the robust qualifications even though many qualifications are performed. This is because there are many different ways to enforce an α-flexibility design. This reinforces the idea that if qualifications are not optimized, then even many qualifications may not be effective to tackle demand uncertainty.

Practical consequences of capacity constraint violations are lower service levels, larger cycle time and larger inventory holding costs. Due to to capacity constraint violations, the number of products in the factory would have to be decreased so that the real utilization rates of machines violating their capacity constraint in the factory is at least lower than 1.0, and ideally lower than $\boldsymbol{u}^{\boldsymbol{max}}$ to control the cycle times. This can severely affect deliveries and the production objectives of the factory.

Products can then be backlogged, which can be expensive. Outsourcing is an alternative to ensure on-time deliveries but can also be expensive or impossible. Nevertheless, it is difficult to predict what would the final decision, as it depends

Table 6.6: Number of qualifications (NQ) and robustness (θ) of nominal qualifications when an α-flexibility design is enforced.

	Work center A		Work center B	
α	NQ	θ	NQ	θ
1	4	0.043	4	0.024
2	84	0.043	14	0.021
3	251	0.0606	77	0.012
4	611	0.071	224	0.087
5	1119	0.152	394	0.140

on the context in which optimization problems are solved and the utility function to describe backlogs, which is not as simple as minimizing the total backlog cost or maximizing the total number of product units made. Ultimately, the choice of using nominal qualifications instead of robust qualifications should be left to capacity planners based on estimated trade-offs between capacity constraint violations and the number of robust qualifications. It is likely that, for non critical work centers with low production variability and short qualification lead times, implementing nominal qualifications can be acceptable because capacity constraint violations may have little impact in practice. Nevertheless, for critical work centers, anticipating critical qualifications and implementing robust qualifications is primordial to ensure a high service level. Nevertheless and generally speaking, implementing robust qualifications is interesting as it avoids capacity violation constraints for a small number of additional qualifications.

In practice, a method to deal with uncertainty is to continuously updating nominal qualifications each time the demand is updated. This should be avoided. This is because, as mentioned in Section 6.2.1, this does not guarantee to find feasible nominal qualifications because the qualification process may sometimes take several weeks or months to validate the quality and the yield of the operation. As the demand by product for the early months on the horizon is also subject to uncertainty, determining and planning robust qualifications is preferable for the whole horizon.

It is worth observing that, if θ is not adequately selected, there may also exist multiple sets of robust qualifications with the same number of qualifications. However, some sets of robust qualifications may actually be better to cover a larger demand uncertainty than other sets of robust qualifications, which is not captured by the robust optimization model because it only seeks to immunize the work center against the specified uncertainty. This is why Algorithm 7 is relevant to identify the most robust set of qualifications among all robust sets of qualifications. These observations are consistent with other observations in the literature: There may exist multiple robust solutions to an optimization problem. Although these robust solutions have the same worst-case objective value, they can have different performances for the nominal scenario (Iancu and Trichakis, 2014; Gorissen et al., 2015; de Ruiter et al., 2016, 2017; Yanıkoğlu et al., 2019).

6.6 Practical use of optimization models

6.6.1 Determining qualification decisions

A straightforward use of the robust optimization model (6.12)-(6.14) is to determine new qualifications to perform to satisfy the demand while respecting capacity constraints and covering the demand uncertainty.

6.6.2 Further improving manufacturing performances

As illustrated on the industrial data in Section 6.5, a small number of qualifications among several hundreds of new qualifications is sufficient to cover the demand uncertainty. Consequently, it is likely that there are two different sets of robust qualifications that cover the demand uncertainty but lead to different performances, for instance in terms of utilization balance of the machines or production variability. It is necessary to distinguish them to further improve manufacturing performances. Differentiating identical sets of robust qualifications in terms of number of qualifications can be done by populating the solution pool after determining the minimum number of qualifications to perform. Modern solvers such as IBM ILOG CPLEX provide this functionality:

1. Two sets of robust qualifications may not be identical in terms of robustness. Algorithm 7 can be used to identify the most robust set of qualifications.

2. Two sets of robust qualifications may also be different in terms of real utilization rates although they all satisfy capacity constraints. Johnzén et al. (2011) and Rowshannahad et al. (2015) propose a "time flexibility measure" to evaluate sets of new qualifications in terms of total utilization rate and utilization balance of the machines. This flexibility measure is interesting as maximizing the utilization balance contributes to further control and reduce cycle times. However, Johnzén et al. (2011) and Rowshannahad et al. (2015) do not consider that demand uncertainty. Their model need to be robustifieds.

3. Robust qualifications can be differentiated in terms of production variability as a large production variability contributes to significantly increase cycle times (Hopp and Spearman, 2011). In semiconductor factories, partly due to reentrant flow, it is unlikely that products arrive continuously in work centers. Work centers are often subject to large Work-In-Process (WIP) peaks leading to congestion. To better capture this phenomenon, Johnzén et al. (2011) propose "a toolset" flexibility measure that captures the fact that operations with large demands must be more qualified than operations with low demands. Pianne et al. (2016) argue that qualified process times should be balanced between machines in the work center. A machine should not be overqualified at the expense of other machines. This is because machines with few qualifications must process almost all their qualified products every qualified to meet the optimized utilization balance, which is difficult due to production variability. Associated flexibility measures are proposed in (Pianne et al., 2016). They

can be seen as ways to measure the quality of the balancing of the qualified process times, and not the quality of the utilization balance of the machines.

4. The principle of large closed chains or long chains can also be used to differentiate sets of qualifications. If one set of qualifications creates more closed chains or larger closed chains between machines and operations than other sets of qualifications, it is very likely that the former will deal better with WIP peaks than the latter (Jordan and Graves, 1995; Graves and Tomlin, 2003).

5. Another straightforward way of differentiating sets of qualifications consists in enforcing α-flexibility designs. However, note that enforcing α-flexibility designs without optimizing a criterion that helps to tackle WIP peaks, such as flexibility measures, may not necessarily lead to better performances (see Section 6.5.3.3).

6.6.3 Exploiting dual variables of robust reformulation

Bertsimas and Thiele (2006b) report that dual variables correspond to the sensitivity of the objective function to changes in parameters of the budget uncertainty set for an inventory management problem. Similarly to Bertsimas and Thiele (2006b) and what was done in Chapter 2, dual variables of the robust optimization model, namely $y^{min}_{t,m,p}$, $y^{max}_{t,m,p}$, $y^{gamma}_{t,m,f}$, can also be exploited:

- $y^{min}_{t,m,p}$ is the sensitivity of the number of qualifications to perform to changes in the parameter $\overline{d}_{t,p} - \widehat{d}_{t,p}$. In other words, if $\overline{d}_{t,p} - \widehat{d}_{t,p}$ increases, $y^{min}_{t,m,p}$ indicates the potential reduction of the number of qualifications.
- $y^{max}_{t,m,p}$ is the sensitivity of the number of qualifications to perform to changes in the parameter $\overline{d}_{t,p} + \widehat{d}_{t,p}$. In other words, if $\overline{d}_{t,p} + \widehat{d}_{t,p}$ decreases, $y^{max}_{t,m,p}$ indicates the potential reduction of the number of qualifications.
- $y^{gamma}_{t,m,f}$ is the sensitivity of the number of qualifications to perform to changes in the parameter $\Gamma_{t,f}$. In other words, if $y^{gamma}_{t,m,f}$ decreases, $\Gamma_{t,f}$ indicates the potential reduction of the number of qualifications.

Exploiting the values of dual variables is particularly relevant from an industrial standpoint to identify if the demand uncertainty on some products or product families is very expensive in terms of number of qualifications. Reporting the values of dual variables can be used by capacity planners to refine the uncertainty set, *i.e.* by defining a smaller uncertainty set, and initiate a discussion with the departments in charge of defining future demands in the case where the number of qualifications to perform is overwhelming. Capacity planners can also initiate a discussion with the departments in charge of defining future demands that the demand uncertainty on some products or product families is not constraining for the production system. The departments can therefore consider new future potential product mixes, *i.e.* by defining a larger uncertainty set, that would have never been initially considered.

6.6.4 On infeasibilities

The optimization problems can be infeasible (see Section 6.5.3). For instance, this can be caused by large qualification lead times and too small production capacities to cover the demand uncertainty. Determining that optimization problems are infeasible is also valuable in practice.

If the nominal optimization problem is infeasible, it indicates capacity planners that the demand must be changed. It is difficult to anticipate how would be the new demand as it depends on different stakeholders (*e.g.* capacity planning, demand planning) within a company. For instance, if the nominal optimization problem is infeasible, the demand for products that generate a large workload at the work center can be decreased while the demand for products that generate a lesser workload can be increased. In this case, the total number of product units made may not decrease, backlogging costs may be acceptable, but lost sales may be incurred on critical products.

If both MCQCPLT or MCRQCPLT cannot be solved because capacity constraints cannot be respected, it is also possible to solve a utilization balancing problem where the demand is described by the uncertainty $\mathcal{D}_t$ to highlight critical machines, *i.e.* machines for which $U_{t,m} > u_{t,m}^{max}$. For instance, the utilization balancing approach in Chapter 2 can be robustified (see Appendix F). In a decision support system, systematically solving a robust utilization balancing problem is relevant to either identify infeasibilities or most loaded and critical machines.

6.7 Conclusions and perspectives

In this chapter, we first proposed a new mixed-integer linear programming mathematical model for a tactical qualification management problem when the demand is deterministic. Second, we motivated the choice of robust optimization when the demand is uncertain, in particular for high mix factories. We proposed an uncertainty set based on the budget of uncertainty to describe product cannibalization and cover the demand uncertainty. Third, we proposed a new robust reformulation of the deterministic model when the demand is described by product cannibalization. Four, we proposed a linear program and a binary search approach to characterize the robustness of a set of qualifications when the demand is uncertain. Fifth, we performed a computational study by using industrial data from a high mix semiconductor manufacturer. In particular, we showed that, (1) The price of uncertainty is acceptable, often less than a few additional qualifications for each machine, (2) It is possible to achieve the same level of robustness as the case where all new qualifications are performed by only performing a restricted number of relevant qualifications, (3) Depending on the forecast uncertainty and the work center, the robust optimization problem can be difficult to solve, and (4) Using the nominal set of qualifications can lead to significant capacity constraint violations, although it can be used for some work centers when the forecast uncertainty is small. Finally, practical applications and implications of the developed models are discussed.

We believe the following perspectives are worth investigating in the future:

1. Other parameters can also be subject to uncertainty, *e.g.* production capacities, throughput rates of operations on machines, qualification costs and lead times. Studying the relevance and effect of the uncertainty of these parameters can be valuable.

2. Considering qualification available times due to disqualifications could be interesting to further distinguish best new qualifications.

3. A large number of qualifications can be difficult to maintain at an operational level. Including disqualification decisions, *e.g.* constraining the number of qualifications by machines, or constraining the total number of qualifications in each period, could be relevant.

4. Extending the static robust reformulation to adjustable robust reformulations may be valuable to further reduce qualification costs.

5. For work centers where the numbers of operations and machines are large, more efficient solution approaches might be necessary. An option consists in using a cutting-plane solution approach with lazy constraints as proposed by Bertsimas et al. (2016). This might be a viable approach as the computational time required to solve MCQCP is small.

6. As there may exist several sets of robust qualifications in terms of number of qualifications given an immunization level, it would be interesting to use additional objective functions, and thus multi-objective optimization, to select the best set of robust qualifications.

7. Other solution approaches for MRBQP can be considered. Iterated max-min approaches are probably relevant not to restrict to the same value of θ for all products and periods.

8. Studying the effect of different qualification cost profiles by machine or by machine and time dependent qualification decisions on the price of uncertainty can be interesting.

9. As the ability of qualifications to cover the uncertainty on the demand strongly depends on the machines in the work center (Hopp and Spearman, 2011), considering the investment decisions in terms of machines could also be investigated to cover the uncertainty on the demand.

10. Capacity allocation is optimistic. For a given operation r, the workload is allocated to the fastest machines in priority to satisfy capacity constraints. Nevertheless, because dispatching and scheduling decisions are highly automated and due to production variability in semiconductor factories, it is unlikely that operational decisions will exactly match capacity allocation decisions. Machines can ultimately have larger workload and utilization rates than expected, because all operation quantities will not necessarily be allocated to fastest machines. This difference should be considered as it may lead capacity constraint violations (see Appendix G for more details).

Chapter 7

Industrial applications and decision support system

A fully functional decision support system, called “FlexQual”, for operational qualification management used at STMicroelectronics by production personnel is presented in this chapter. In particular, the purpose, the functioning, the content of FlexQual are presented. FlexQual embeds all theoretical developments made in Chapters 2, 3, 4 and 5. FlexQual is now included in the decision process of production personnel.

7.1 Introduction

Theoretical developments in Chapter 2, 3, 4 and 5 are included in FlexQual, a fully functional decision support system, that is currently used at STMicroelectronics for operational qualification management.

In the given industrial context, re-qualification decisions are not automated, *i.e.* re-qualifications are not automatically updated in the Manufacturing Execution System (MES) after running the solution approaches. All mathematical models and solution approaches developed for the operational decision level presented in the thesis are included in the decision support system called FlexQual. A decision support system (DSS) is a system that aims at supporting decision-making by presenting relevant information and recommendations on decisions that should be made. Making a decision support system is relatively easy, cheap, and still effective to improve manufacturing performances although decision support systems can at first slow down production personnel in their decision process (Sharda et al., 1988). Moreover, production personnel can evaluate scenarios (what-if analysis), which would be impossible if re-qualification decisions were automated. In addition, it is actually difficult to automate a decision process in a complex and already automated manufacturing system. Plus even the best automated system cannot take into account the whole context motivating a decision. The availability of technicians, their skills hence the knowledge of coming maintenance operations or recent history of the equipment may not be known to the system.

The remainder of the chapter is organized as follows. In Section 7.2, we present questions that are answered by production personnel at a production control level to manage their work center and ensure best possible manufacturing performances. In Sections 7.4 and 7.3, we present the fully functional decision support system called FlexQual used at STMicroelectronics to support production personnel in their decision process, notably by ensuring that questions presented in Section 7.2 can be answered. More precisely, Section 7.3 presents the content of FlexQual. Section 7.4 presents how FlexQual works. Section 7.5 presents some use cases and feedback of production personnel on FlexQual. Finally, in Section 7.6, we conclude and give some perspectives.

7.2 Decision-making by production personnel

At an operational decision level, production personnel usually want to answer the following questions:

- Are there line stop operations, *i.e.* operations for which there is currently no machine validated or qualified?
- Are there priority lots subject to line stops?
- Are machines unbalanced?
- What are the critical re-qualifications that should be made or should have been active to improve manufacturing performances?

- Will the work center be able to meet daily production objectives?
- Should a maintenance operation be postponed or scheduled sooner?
- Is there budget, *i.e.* capacity margins, for maintenance or engineering operations on a particular machine or machine set?

These questions are all linked together. For instance, if the work center is able to meet its daily production objectives, then there is probably budget for maintenance operations. In addition, identifying if a maintenance operation can be postponed or scheduled sooner can be done by identifying if machines have production capacity margins to satisfy daily production objectives. FlexQual aims at effectively answering these questions. Therefore, FlexQual must present information so that answering the questions is eased. Three major axes are considered for FlexQual:

- Present relevant performances indicators, *e.g.* in terms of utilization rates of the machines, cycle time, throughput, and line stop operations.
- Proposing re-qualification decisions to improve performance indicators and thus meet production objectives.
- Standardize decision making across team shifts of a work center, and across all work centers.

7.3 Content of FlexQual

FlexQual is made of two parts:

1. An e-mail body summarizing critical information for quick and relevant decision-making.
2. An Excel^{TM} file containing additional information on the utilization rate by machine, the mean utilization rate by machine set, on OEE(%) by machine, on OEE(%) by machine set, on the throughput by product family, on disqualifications, on priority WIP, on line stop operations.

In this section, the content of FlexQual will be presented. Note that chart legends and column headers are in French because FlexQual is currently used at a factory located in France. In addition, note that most of the data is blurred for confidentiality purposes. Similarly, X and Y scales of charts are omitted for confidentiality purposes.

7.3.1 E-mail content

The e-mail contains a body summarizing critical information for quick and relevant decision-making, *i.e.* line stop operations, priority WIP (Work-In-Process) with no more than one qualified machine, and a re-qualification plan to improve the throughput (see Figure 7.1). All indicators are computed by using the multi-period bilevel optimization approach proposed in Chapter 4, and the considered horizon is

of 24 hours (divided into three periods of 8 hours). As priorities of operations are continuously modified by the manufacturing system, the "FIFO" computation mode for the throughput is considered. As team shifts work for 8 hours (in general), the e-mail body only contains line stop operations for the next 8 hours. Similarly, the e-mail body only shows the priority WIP with no more than one qualified machine for the next 8 hours based on forecasted lot arrivals. Oppositely, the proposed re-qualification plan is determined for a horizon of 24 hours. Consequently, the e-mail body can answer the questions "Are there line stop operations?", "Is there priority WIP subject to line stops?", and "What are the critical re-qualifications that should be made or should have been active to improve manufacturing performances?".

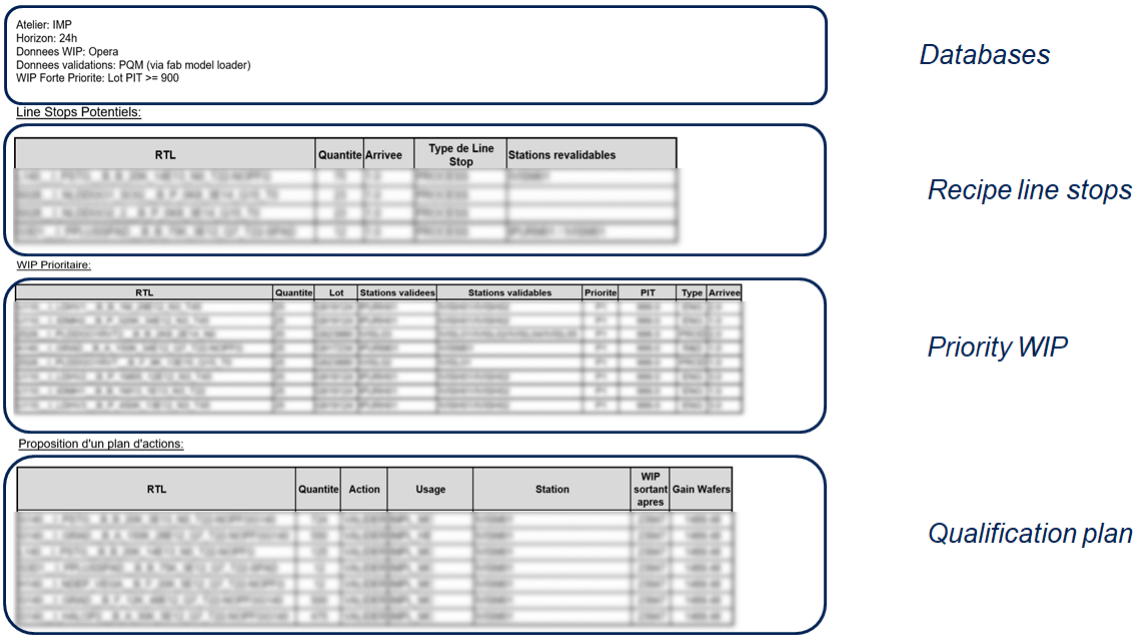


Figure 7.1: E-mail content.

The e-mail body also contains three charts:

1. The throughput, *i.e.* the number of wafers produced, by machine set for which production objectives are defined. This answers the question "Will the work center be able to meet daily production objectives?".

2. The utilization rate of machines for machine sets that do no meet their production objectives. This partially answers the questions "Are machines unbalanced?" because only the utilization rate of machines for machine sets that do not meet their production objectives is presented in the e-mail body. This is done to present only concise information. Note that if all machine sets meet their production objectives, the utilization rate of all machines is presented.

3. The qualification rates of machines for machines sets that do no meet their production objectives. Including qualification rates enables production personnel to better estimate when a maintenance operation for a machine is

mandatory. For instance, in the ion implantation work center, operations are disqualified if processing conditions for the considered operation become too unstable due to ion source wear. Qualification rates are practical estimates to perform ion source changes. Similarly, to the utilization rates of machines, if all machine sets meet their production objective, the utilization rates of all machines is presented.

These three charts give production personnel visibility on significant metrics for the next 24 hours. At a glance, production personnel can thus identify and anticipate potential problems and fix them with appropriate decisions, *e.g.* by requalifying some machines for some operations or by postponing non-critical maintenance operations.

7.3.2 Excel™ file

Although most questions can be answered with the e-mail body, additional information is provided with an Excel™ File. The Excel™ file contains the same information as the e-mail body, as the e-mail body is built upon information contained in the Excel™ file. Nevertheless, the Excel™ file contains additional information to support decision-making and answer some questions that could not be answered in the e-mail body, or were only partially answered in the e-mail body, and the possibility to evaluate what-if scenarios. More precisely, the Excel™ file has seven tabs:

1. “Board” evaluates and presents the manufacturing performance of the work center for the next 24 hours. This tab then gives visibility on the utilization rates of machines, on OEE(%), on the throughput by machine set, on the throughput by product family, on qualification rates, and therefore on the capacity of the work center to meet daily production objectives.

2. “Projections” presents WIP quantities by operation that should arrive in the work center by the next 24 hours. In the tab, each entry is associated to an operation. For each operation, the mean throughput rate and the standard deviation of throughput rates is presented. In addition, qualified and qualifiable machines are presented. If there is only one qualified machine, then the operation is highlighted in orange. If their is no qualified machine, the operation is in bold and also highlighted in orange.

3. “WIPHighPrio” presents priority WIP that should arrive in the work center by the next 24 hours. In the tab, only highly priority lots are presented. They have a priority greater than a specified threshold. For each priority lot, the associated operation and qualified machines are presented. If there is no more than one qualified machine, the entry is also highlighted in orange.

4. “LineStops” presents operation quantities that are subject to production capacity interruptions, which is either due to the fact that operations have currently no qualified machines or that all qualified machines are down. “Line Stops” must be anticipated and managed because: (1) they can severely increase the mean cycle time if unnoticed; (2) at the work center level, production is

usually managed with aggregated indicators (unless for highly priority WIP), *e.g.* in terms of throughput by machine set or operation family. If a line stop operation is not fixed, it may be still possible to attain a large throughput. Nevertheless, some lots could be backlogged or behind schedule because they are stopped at an operation as there is no associated production capacity.

5. "Devalidations" presents all disqualifications in the work center. More precisely, for each disqualified pair (operation, machine), the tab presents disqualification reason codes, disqualification costs (categorical costs, *e.g.* small, medium, large), the latest qualification date. Disqualification reason codes are critical as they describe if a disqualification can be easily performed. They are used by the solution approaches to discard qualifications that cannot be done in 24 hours. Similarly, the latest qualification date is reported because it is critical to respect internal quality standards due to process obsolescence.

6. "Scenarios" enables production personnel to evaluate a scenario (what-if analysis), which can be composed of maintenance operations, disqualifications and re-qualifications.

7. "Actions" proposes a re-qualification plan to improve the manufacturing performances. The objective function used in the optimization process is either the utilization balance and the total utilization rate of the machines and throughput depending on the horizon. The mean cycle time is reported as an additional indicator to better support decision-making but is not currently optimized.

For space limitations, more details on the content and the use of the Excel^{TM} file are provided in Appendix E Section E.1.

7.4 How does FlexQual work?

Early versions of FlexQual were running on the computer of production personnel. This caused problems as personal computers have a limited amount of RAM and can be slow to solve optimization problems. Latest versions run on a dedicated virtual machine hereafter called *server*. The server has scheduled runs that consists in sending e-mails containing FlexQual to production personnel.

When production personnel wants to evaluate a scenario, the servers receives a request, interprets it, and then sends a new e-mail containing an updated e-mail body and Excel^{TM} file with respect to the evaluated scenario. The use of e-mails, even to receive the results of a scenario, is interesting for production personnel because only the server runs the optimization engine, which can be a computationally expensive task, in particular on computers used by production personnel that are not necessarily properly suited. E-mail bodies contain only the most relevant information for decision-making and is easily accessible, even remotely from smart-phones by production personnel. Nevertheless, the use of e-mails can lead to an excessive amount of e-mails. In addition, production personnel share the server. This means that if a scenario or a scheduled run takes too long to complete,

production personnel will wait a great amount of time before receiving the results of their scenario. Note that this has not caused any problem so far. An alternative to the use of e-mails would consists in developing a dedicated web application, or something similar.

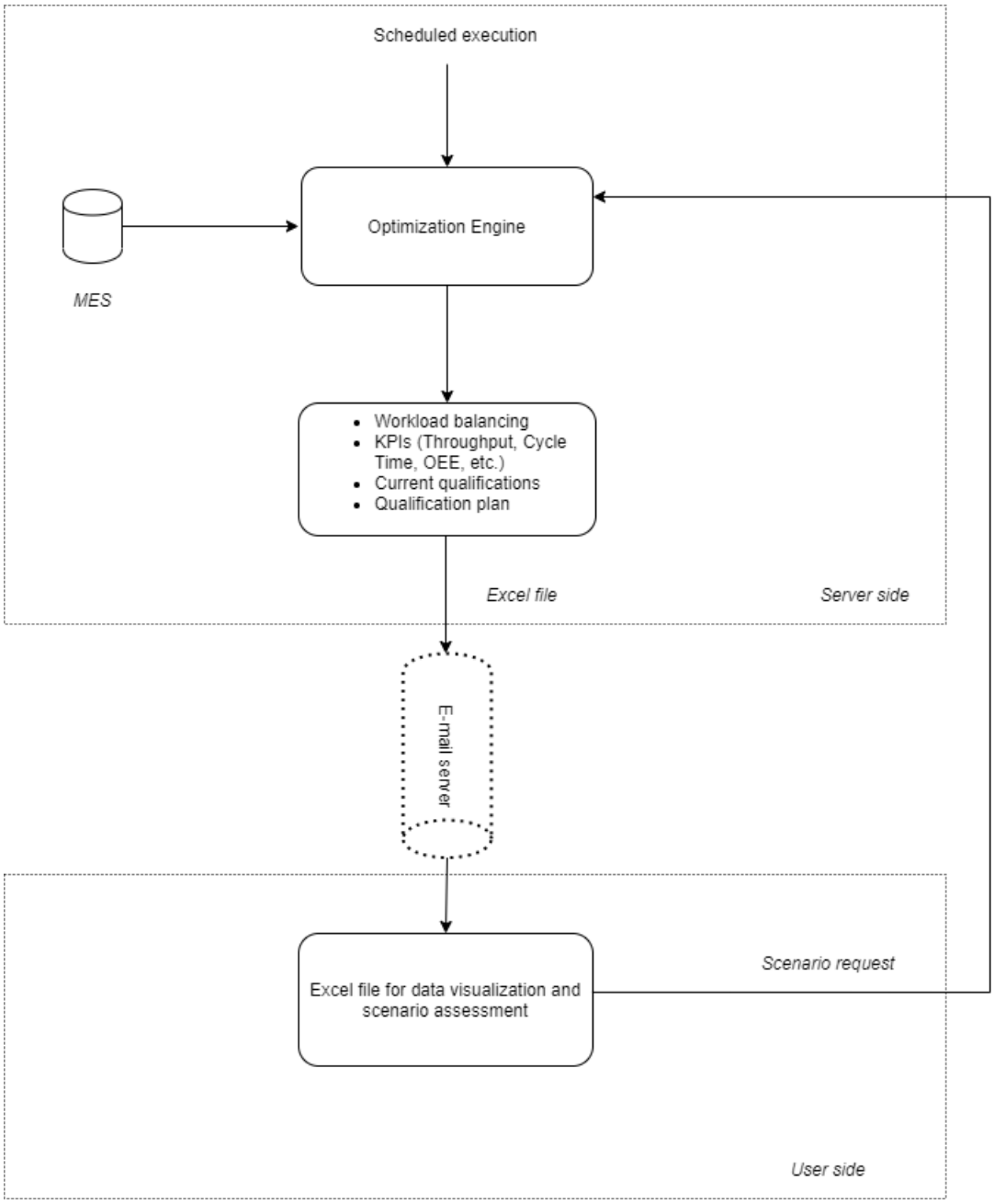


Figure 7.2: How does FlexQual work?

More details are provided in Appendix E Section E.2.

7.5 Use cases and experience feedback

7.5.1 Short-term use cases

Main use case. The server has scheduled runs consisting in sending to production personnel FlexQual and results associated to the optimization engine 20 minutes before shift changes (see Figure 7.2), approximately every eight hours. In this case, the length of the horizon is equal to 24 hours.

Ideal re-qualifications. Another potential use case is not to show only possible re-qualifications that can be done given the considered horizon and re-qualification lead times and costs, but to show re-qualifications with respect to an ideal situation. If this qualification was active, what would the throughput and cycle time? By evaluating every re-qualification separately, no matter its cost or lead time, it is possible to evaluate the potential impact of the qualification on the throughput and cycle time. By frequently repeating this operation, it is possible to create a database that highlights the most critical re-qualifications to perform, and consequently, highlight the most critical re-qualifications that should not be lost. This information is still relevant for production personnel but differs from a classical use case.

Short-term bottlenecks. Even without determining re-qualifications, the utilization balancing approach developed in Chapter 2 can be used to detect future short-term bottlenecks machines or machine groups if their utilization rate is expected to exceed 100% for instance on a 24-hour horizon.

Prioritizing maintenance operations. The bilevel optimization model is also used to identify machines that are down but critical to maximize the throughput. To identify the machines, we use what-if scenarios and compute with the bilevel optimization model what would be the throughput if the machine that is currently down is assumed to be up. This computation is ideal as maintenance operations cannot be done instantaneously. Nevertheless, after finishing all what-if scenarios, it is possible to rank machines by their gain on the throughput and prioritize some maintenance operations over others.

Better dispatching decisions. The utilization balancing approach developed in Chapter 2 can be used to provide guidelines, for instance the maximum number of wafers for a operation quantity to assign to a machine, to dispatching engines as it also optimizes the total utilization rate of the machines.

Further benchmarking work centers. All indicators are also computed for two ideal cases. The first ideal case is when it is assumed that there is no disqualification. The second ideal case is when it is assumed that there is no disqualification and there is no machine down time. Both ideal cases allow work center mangers to further benchmark work centers and evaluate if there is a large gap between between current performances and ideal performances. If yes, then some critical decisions should be made in terms of qualifications or maintenance operations.

Experience feedback. Developed mathematical models and solution approaches can be applied for all work centers in the factory. Although FlexQual can be used for *any* work center, standardizing its use across different work centers is a difficult task. This is because the use of FlexQual is different from a work center to another because work centers are different and thus may have different needs.

In work centers where there are many possible re-qualifications, *e.g.* where there are many machines (process chambers) such as in Dry etch, developed models and solution approaches are particularly relevant to identify critical qualifications that should be performed to optimize performances. This is because the number of operations and machines can be overwhelming. In addition, in work centers where disqualification decisions are frequent, *e.g.* in Ion implantation for WIP management policies, proposing re-qualifications plans is particularly relevant to evaluate if machines are too unbalanced, which may affect the throughput and the cycle time. This is because Ion Implantation tools may have very different throughputs depending on the operation processed, on the wear of their ion source and usually process hundreds of different operations each. In addition, Ion implantation is a work center with a high degree of re-entrancy, making it a critical work center for the whole production facility.

In work centers where the overall number of qualifications, the degree of re-entrancy or the number of machines is relatively small, production personnel already know most critical qualifications that should not be lost. Therefore, *optimizing* re-qualification plans might be less relevant. The decision support is still relevant to better follow-up qualifications, *i.e.* anticipate disqualifications faster and more often, as production personnel are interested in identifying unnoticed disqualifications and in controlling qualification rates. Developed models and solution approaches remain relevant, for instance, to identify the bottleneck machines and therefore anticipate WIP bubbles or line stop operations. For these work centers, in most cases, this is sufficient to ensure a good throughput. If it is insufficient, production personnel can still use the re-qualification plan that is automatically proposed by FlexQual, as it will probably show to production personnel critical points to address for the next 24 hours. Furthermore, for work centers where qualification delays can be long, in particular in Diffusion where re-qualifications may require the use of test wafers and last a few hours, considering a horizon of a few hours to 24 hours can be limited because potential gains may never be significant.

Developed approaches are particularly interesting for bottleneck or near-bottleneck work centers. Determining if there exist critical qualifications to improve the throughput or the cycle time, is particularly interesting as bottleneck work centers constrain the whole production facility. For non-bottleneck work centers, identifying and anticipating line stop operations or operations with only one qualified machine or priority WIP, and identifying recent disqualifications is often good enough to maintain a good throughput and a low cycle time. Nevertheless, an unnoticed disqualification can still occur and strongly affects the throughput or the cycle time of non-bottleneck work centers. Improving the throughput of non-bottleneck work centers can also be useful to avoid starving bottleneck work centers, reducing production variability and thus minimizing the total cycle time, which can be performed with developed approaches.

Future steps. Next steps consists in computing more frequently than every 8 hours information proposed in FlexQual to better follow operational conditions (*e.g.* machine downtime, WIP peaks, disqualifications), and therefore to work in a rolling-horizon manner. Information is computed every 3 hours for an horizon of 24 hours. Each re-qualification plan contains a single re-qualification. Instead of sending an e-mail, information is printed with the help of a business intelligence software (TIBCO Spotfire) on a single page that is directly visible in the clean room by shift teams. The Excel^{TM} file is kept if what-if scenarios are required.

7.5.2 Medium-term use cases

Main use case. Developed mathematical models and solution approaches are also used for medium-term decision support. Similarly, the server has also scheduled runs where it computes optimized re-qualification plans for a horizon of one to two weeks. The server automatically sends an e-mail each week before the review and update of the production plan. In this case, the considered horizon is of one week. Production personnel receive the same information as they would receive if the considered horizon were of 24 hours. A one-week report gives production personnel visibility on machines that may be overloaded, and therefore that are critical. They can also use this report to exchange with production planners on the feasibility on the new production plan and critical operations.

Next steps. Similarly to short-term use cases, next steps also consists in computing more frequently information proposed in FlexQual for a horizon between one and two weeks.

7.5.3 Feedback and decision process

FlexQual received positive feedbacks from Diffusion and Ion implantation production personnel. At the time where the development of FlexQual started with Diffusion and Ion implantation because they were bottleneck. Production personnel of other work centers also received automatic e-mails. Nevertheless, production personnel of other work centers were less involved in the development, refinement and adjustment of FlexQual for operational needs.

Beyond qualification management, FlexQual changes the management philosophy: More anticipation and preparation. Production personnel no longer work with an instantaneous vision of the WIP. They work on previsions, *i.e.* WIP projection, for the next eight to twenty-four hours, even if the projection may not be extremely accurate. This is not necessarily a problem as work centers tend to work with a rolling horizon approach where decisions, including re-qualification decisions, are frequently reassessed. This gives them visibility, time to plan decisions, simulate maintenance operations, qualification and disqualification decisions to manage the work centers. Shift teams are also more autonomous in their decision-making with

respect to priority WIP and line stop operations. All necessary information for decision-making is contained in a single file, which is a considerable time saver. In addition, FlexQual is now fully integrated in their decision process between shift changes and in the daily activity reviews. At 5 AM, a first analysis is made by production personnel to identify critical machines, line stop operations and potential levers by shift teams. At 9 AM, production personnel make a second analysis. An estimate of the number of wafers processed by product families with respect to daily production objectives is communicated to other work centers (see Figure 7.3).

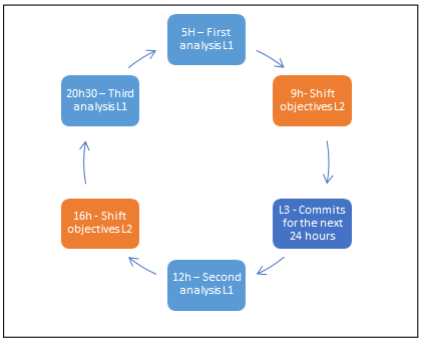


Figure 7.3: Daily decision process.

A mistake, when designing and developing decision support systems, is to present an overwhelming number of indicators. First, users probably do not need all of them. Indicators for aggregated entity levels, *e.g.* the cycle time by product family or operation family, is probably enough. Indicators for smaller aggregated levels is probably irrelevant due to production variability. It is enough to have a proactive behavior and proposing a few indicators even if some must be removed or others added in later development stages. In addition, optimization only represents a fraction of a decision support system.

7.6 Conclusions and perspectives

In this chapter, we presented questions that answered by production personnel several times a day to ensure manufacturing performances, in particular to ensure a high throughput, low cycle time and to meet production objectives. We presented a fully functional decision support system that is used by production personnel at STMicroelectronics. Driving work centers are Diffusion and Ion implantation. The decision support system, FlexQual, is now included in their daily decision process and at production shift start. Early involvement of work centers for the development of FlexQual was crucial. In the context of this thesis, early involvement of different work centers (Diffusion and Ion implantation) enabled us to develop FlexQual by following guidelines provided by production personnel. The first meeting consisted in us presenting and proposing an initial and simple version of FlexQual. Then, regular meetings were scheduled with production personnel. Production personnel frequently requested to add new information or remove information, which were no longer judged relevant for decision-making. Similarly, we frequently pro-

posed modifications and new features for decision-making. FlexQual was kept as simple as possible and included in Excel^{TM}. Excel^{TM} was naturally selected as the graphical user interface as most employees in the company frequently use it. Note that FlexQual was developed for operational qualification management. Improving FlexQual with the work presented in Chapter 6 on tactical qualification management is in progress.

FlexQual can be improved:

1. With small changes, it could be used by other wafer fabs.

2. FlexQual could also benefit from new developments by modeling specific features for work centers.

3. Other reporting could be imagined. Reporting critical qualifications could be relevant. Critical qualifications could correspond to the qualifications that lead to a large decrease of the utilization balance if they are lost. Critical qualifications could correspond to the smallest set of qualifications that needs to be maintained to keep the same utilization balance and total utilization rate of the machines as the utilization balance and total utilization rates of the machine with the initial set of qualifications. This could encourage work centers to quickly re-qualify critical qualifications if they are lost.

4. Further studying the effect of currently down machines on manufacturing performances could be extremely relevant for production personnel to better determine priority maintenance operations.

5. Another significant aspect would be to be able consider all work centers at the same time in the solution approaches. This is because although local qualification decisions are relevant to optimize some local aspects of manufacturing performances, in particular the throughput or cycle time of the work center, local qualification decisions may not consider other crucial aspects. Considering all work centers at the same time could be done by considering a multiple objective optimization approach. For instance, the objective function would consist in maximizing the satisfaction of production objectives, which can be defined in terms of machine set, of layer, of product family, etc. As all work centers have the same production objectives as they are defined from a production plan optimized for a horizon of one week (see Mhiri 2016; Christ 2020), considering a multiple objective optimization approach would be a first step to considering all work centers at the same time. Production objectives could also be weighted to better consider the capacity of downstream work centers. Finally, note that a simple multiple objective optimization approach is implemented in FlexQual based on the work of Tamssaouet (2019) who define in his thesis an objective function called Target Satisfaction Indicator (TSI), based on the concept of generalized mean, to maximize the satisfaction of production objectives. More precisely, given appropriate weights, we are able to determine a re-qualification (and disqualification) plan to optimize TSI with the constructive greedy heuristic and the bi-level optimization approach. Nevertheless, a complete study on the most suitable way to consider

different objective functions, *e.g.* with *a priori* approach or *a posteriori* approach, and on the trade-off between the satisfaction of different production objectives after qualification and disqualification decisions would be valuable. Another option could consists in using production objectives as the demand. In this case, re-qualification decisions optimized in terms of utilization balance and total utilization rate simultaneously maximize the throughput and optimize the satisfaction of production objectives. This is simpler than solving multiobjective optimization problems but this does not consider the "real" WIP in work centers, which could be a problem for short horizons. The cycle time computed by using Chapter 5 may also no longer be representative of the real cycle time. In the current version of FlexQual, considering production objectives as the demand is possible through the configuration file.

6. Another interesting study would be to integrate a notion of feasibility of the qualification with respect to equipment performance. In Ion Implantation for example, throughput and stability of the process decrease with time and operations may get disqualified because too unstable. Re-qualifying these operations is not a good option.

Chapter 8

Conclusions and perspectives

This thesis dealt with qualification management in manufacturing operations, both at tactical and operational decision levels and in particular for HMLV wafer fabs, and followed two previous theses on qualification management (Johnzén, 2009; Rowshannahad, 2015). The thesis aimed at providing new optimization models and efficient and effective solution approaches to answer the following two questions: (1) "How to determine the most relevant re-qualifications to improve operational efficiency?" and (2) "How to determine the most relevant new qualifications to satisfy the demand and cover the demand uncertainty while minimizing qualification costs?". Our contributions and main findings are highlighted in Section 8.1, and relevant perspectives on our work are outlined in Section 8.2.

8.1 Conclusions

8.1.1 How to determine the most relevant re-qualifications to improve operational efficiency?

In Chapters 2, 3, 4, and 5, we answered the question: "How to determine the most relevant re-qualifications to improve operational efficiency?"

In terms of utilization balance and total utilization of the machines. In Chapter 2, most relevant re-qualifications are determined to maximize the utilization balance and minimize the total utilization rate of the machines. New effective and efficient solution approaches were proposed. In particular, dual prices are used to derive heuristics that are quickly guided towards good solutions, *i.e* an effective set of re-qualifications. Numerical results on industrial data of two different work centers and two different qualification configurations from a 300 mm wafer fab showed that using dual prices is particularly relevant. From a general standpoint, we recommend studying the nature of the data, as it is primordial to design efficient and effective solution approaches. Solution approaches are embedded in the fully functional decision support system presented in Chapter 7. Most relevant re-qualifications are proposed to work center managers twenty minutes before shift changes.

In terms of throughput. In Chapter 3, the utilization balancing approach of Chapter 2 is extended to a bilevel optimization approach to better consider dispatching rules, which have a significant effect on the relevance of re-qualifications, to better model disqualification decisions, and to better consider the throughput as optimization criterion. To the best of our knowledge, there is no contribution in the literature that proposes to model disqualification decisions, whereas it is a fundamental aspect of operational qualification management. Numerical results showed that the bilevel optimization approach can be used to determine relevant disqualification decisions in terms of throughput. On short horizons, the bilevel optimization approach leads to better solutions in terms of throughput than the utilization balancing approach. On larger horizons of several days to a few weeks, the utilization balancing approach and the bilevel optimization approach lead to similar re-qualification decisions. Bilevel optimization approaches are embedded in the fully functional decision support system presented in Chapter 7.

In Chapter 4, the bilevel optimization approach is extended to a dynamic bilevel optimization approach, where dynamic means that demand and production capacities vary over time due to production variability. Numerical results on industrial data from a 300 mm wafer fab showed that, when qualification decisions are subject to lead times or induce maintenance operations, the dynamic approach is more relevant than the (static) bilevel optimization approach presented in Chapter 3 to determine relevant qualification decisions. In other words, re-qualification decisions determined with a dynamic bilevel optimization approach is likely to lead to a larger throughput in practice than re-qualification decisions determined with a static bilevel optimization approach. The dynamic bilevel optimization approach is

embedded in the fully functional decision support system presented in Chapter 7.

In terms of mean cycle time. In Chapters 2, 3, and 4, the optimization criterion is either the utilization balance and the total utilization rate or the throughput. In Chapter 5, the effect of re-qualification decisions on the mean cycle time on the short term is studied. It is first argued that closed-form solutions describing the mean cycle time are available at an operational level for work centers. Second, the relevance and the limits of closed-form solutions are shown for different work centers on industrial data from a 300 mm wafer fab. Then, the impact of re-qualification decisions on short-term cycle times is highlighted. In particular, we show that it is possible for two qualification decisions to lead to the same gain on the throughput but different gains on the mean cycle time. In addition, it is shown that most qualification decisions are irrelevant to minimize the cycle time, but that relevant qualification decisions can significantly minimize the mean cycle time. Closed-form solutions for the short-term mean cycle time are embedded in the fully functional decision support system presented in Chapter 7.

Decision support system. Chapter 7 presented a fully functional decision support system developed for operational qualification management in a 300 mm wafer fab of STMicroelectronics. The decision support system embeds all theoretical development performed throughout the thesis. A configuration file manages the use of the most relevant mathematical models and solution approaches depending on the horizon and the considered work center. The decision support system usually proposes re-qualification plans optimized in terms of utilization balance and total utilization rate of the machines or throughput by using the solution approaches developed in Chapter 2. Although re-qualification plans are determined for a specific criterion, the decision support system reports many relevant indicators to work center managers.

Take away message. In Chapters 2, 3, 4 and 5, the computational study highlights that there are frequently only a *handful* of re-qualifications and disqualifications that can significantly improve a given criterion. Because semiconductor production facilities are complex, unbalanced, unsymmetrical and time-varying systems, advanced methods are required to identify those relevant re-qualifications and disqualifications. This is what optimization models and solution approaches proposed and validated in Chapters 2, 3, 4 and 5 can be used for and are currently being used at STMicroelectronics in the decision support system presented in Chapter 7.

8.1.2 How to determine the most relevant new qualifications to satisfy the demand and cover the demand uncertainty while minimizing qualification costs?

In Chapter 6, we answered the question: “How to determine the most relevant new qualifications to satisfy the demand and cover the demand uncertainty while minimizing qualification costs?”

Optimizing the set of new qualifications to perform is critical because the qualification configuration of a work center is the underlying structure enabling a work center to have a great operational efficiency in terms of utilization balance, throughput and cycle time. To determine a cost effective set of qualifications that satisfy the demand and respect production capacities of machines, a new mixed integer linear programming mathematical model is proposed when the demand is deterministic and the qualification delays are considered. As the demand by product is subject to uncertainty, a robust reformulation is proposed to cover the uncertainty based on the concept of budget of uncertainty. In addition, a new decision-dependent uncertainty model is proposed to characterize the robustness of a set of qualifications, therefore the robustness of a work center, against demand uncertainty. A binary search is proposed to compute the robustness of a set of qualifications because the decision-dependent uncertainty program is NP-Complete. In a computational study on industrial data from a 300 mm wafer fab, the price of uncertainty is shown to be small. The set of qualifications determined by only considering the nominal demand is shown to lead to capacity constraint violations. Covering the uncertainty on the demand uncertainty is critical as being robust is actually affordable.

Take away message: To satisfy the demand by product and cover its uncertainty, there is no need to conduct *many* qualifications. Similarly to re-qualifications, only relevant new qualifications are mandatory. As the qualification process may take several weeks or months, it is also critical to anticipate the right new qualifications. As it is complex to identify because semiconductor production facilities are complex, unbalanced, unsymmetrical and time-varying systems, the optimization models proposed and validated in Chapter 6 can be used.

8.2 Perspectives

Section 8.2.1, resp. Section 8.2.2, outlines relevant perspectives to further improve operational qualification management, resp. tactical qualification management. Finally in Section 8.2.3, perspectives on the decision support system are outlined.

8.2.1 Further improving operational qualification management

8.2.1.1 Improving the utilization balancing approach

A utilization balancing approach is proposed to determine re-qualifications in Chapter 2. However, the utilization balancing approach can be extended, and several research avenues are proposed below:

1. Some parameters might be subject to uncertainty, such as operation quantities and machines capacities, and designing robust qualification plans should be interesting. A first idea to investigate is how the robust optimization approach proposed in Chapter 6 for tactical qualification management can be extended for operational qualification management.

2. Workload variables are continuous but, in practice, some machines run operation quantities by batches. Hence, the consideration of batching constraints could be explored as in Rowshannahad and Dauzère-Pérès (2013).

3. An outer linearization algorithm is used to solve nonlinear programs. Other algorithms, such as active-set methods or sequential quadratic methods (Rowshannahad et al., 2015) could be compared to the outer linearization algorithm to further reduce computational times.

4. Solution approaches could be compared on data from other factories to further validate the relevance of the dual variable solution approaches.

5. It would be relevant to study the robustness of the solution approaches, *e.g.* under which conditions using dual prices does not provide good solutions.

6. Additional branching and exploring strategies could be explored for the Branch-and-Bound solution approach.

7. It would be interesting to better understand the impact of different γ settings on solution quality and computing time.

8.2.1.2 Improving the bilevel optimization approaches

Several research avenues to extend the bilevel optimization approaches are discussed below:

1. Some parameters used in the bilevel optimization models might be subject to uncertainty, *e.g.* the demand by operation and the production capacity by machine. From the work center managers' point of view, dispatching decisions can also be subject to uncertainty. Determining robust re-qualification and disqualification plans can be relevant.

2. In the computational studies, we limited ourselves to the case where $k = 1$. Hence, it seems relevant to design efficient and effective solution approaches that quickly determine re-qualification and disqualification plans for $k > 1$. Note that, in the decision support system presented in Chapter 7 and although no thorough computational study was performed to evaluate their efficiency, constructive greedy heuristics are used to solve the bilevel optimization models. More precisely, to avoid evaluating every re-qualification and disqualification at each iteration of the greedy heuristic, pre-processing rules are used. For instance, pre-processing rules are based on the values of the dual variables of the qualification constraints for re-qualification. Pre-processing rules are also based on the fact that potential effective re-qualifications to perform (at least for some work centers) are the re-qualifications that consider re-qualifying operations for underloaded machines, *i.e.* for machines such that $U_m < 1$, or re-qualifying operations that can be processed much faster on disqualified machines than on currently qualified machines.

3. It would be also interesting to better understand the impact of different γ settings on re-qualification decisions and computing time.

4. Studying the multi-objective aspect of short-term qualification and disqualification management would be interesting to explore.

5. It is relevant to consider batch size constraints in the bilevel optimization models when wafers are processed by batch (Rowshannahad and Dauzère-Pérès, 2013).

6. As already discussed, there may be multiple solutions to the lower-level optimization problem, and some solutions may lead to a better throughput than others, because the proposed bilevel formulations are neither optimistic or pessimistic. Studying this is left for future research, *e.g.* by giving bounds on the throughput to production personnel. For instance Fischetti et al. (2018) propose a refinement procedure that can be used to obtain an optimistic solution from the lower-level.

7. New methods to simulate the throughput or dispatching decisions could be proposed to be closer to the behavior of the actual dispatching decisions, *e.g.* by including batching constraints when relevant (Rowshannahad and Dauzère-Pérès, 2013) or machine learning techniques.

8. Numerical results showed that the single-period optimization model often proposes the same re-qualifications as the multi-period optimization model on industrial data. It would be interesting to automatically identify when this is the case, to save significant computational times when searching for re-qualifications.

8.2.1.3 Improving cycle time management

At the operational decision level for a work center, cycle time management is possible because we showed in Chapter 5 that it is possible to evaluate re-qualification decisions with a simple and relevant formula for the mean WIP.

It would be relevant to study the effect of machine qualifications, *e.g.* the effect of variability of machine qualifications, in G/G/m queues. This would, for instance, help to support better decisions at tactical decision levels when new qualifications are determined. Nevertheless, considering machine qualifications in queuing theory is complex. This is because not only the number of qualifications matters but the qualifications themselves are also critical. Two qualification matrices could have the same number of qualifications by machines and the same number of qualifications by operation, but they would probably lead to different cycle times as illustrated in the numerical results. Instead of trying to directly estimate the new cycle time with qualifications, a solution could consists in determining and exploiting upper bounds on the cycle time. For instance, it is known that the mean cycle time of G/G/m queues is bounded by the mean cycle time of m-G/G/1 queues working in parallel. Exploiting this bound could provide valuable insights to capacity planners to decide the qualifications that should be added to a machine to reduce the mean cycle time. Knowing the exact value of the mean cycle time is not mandatory to make relevant decisions (see Section 5.1.1). Other bounds may be available. Recently, bounds on the mean cycle time were proposed based on robust optimization (Bandi

et al., 2015; Bertsimas et al., 2018; Whitt and You, 2019). New bounds could also be derived using machine learning techniques and historical data. Consistent qualification decisions between work centers should also be studied. A local decision at a work center may not be ultimately relevant if all work centers were simultaneously considered. For instance, if a re-qualification decision is made to optimize the throughput at an upstream work center such that a downstream work center is unable to absorb the additional workload. In this case, another re-qualification decision, which may improve less the throughput, would be more relevant. Multiple objectives could also be considered. A re-qualification decision may decrease the mean cycle time of a particular layer at the work center. Nevertheless, the same re-qualification decision may also increase the mean cycle time of another layer.

8.2.1.4 Considering all work centers simultaneously

Another interesting perspective would be to simultaneously consider all work centers in the solution approaches. Although local qualification decisions are relevant to optimize some local aspects of manufacturing performances, in particular the throughput or cycle time of the work center, local qualification decisions may not consider other crucial aspects. For instance, improving the throughput of non-bottleneck work centers can be useful to better feed bottlenecks work centers but, from a flow management perspective, one re-qualification might be better than another. For instance, machines that mainly process a specific layer might be overloaded, whereas machines that mainly process another layer might be underloaded. Therefore, a re-qualification decision that improves the throughput of the underloaded machines is probably better.

8.2.1.5 Short-sighted aspect of dispatching decisions

In this thesis, we considered one of the short-sighted aspects of dispatching decisions. Namely, we were interested in the way operations compete for the capacity of a given machine. This was modeled as bilevel optimization models. Nevertheless, capacity allocation might be optimistic for the bilevel optimization models and optimization models presented in Chapter 6. Capacity might be allocated to the fastest machine in priority. In other words, the utilization rates of the machines might be underestimated. This other short-sighted aspect of dispatching decisions could be also be considered. Proposals are provided in Appendix G.

8.2.1.6 Endogenous demand

In the thesis, the demand by operation is assumed to be exogenous in the chapters dedicated to operational qualification management. In practice, due to the re-entrant product flows, the demand by operation is endogenous. In other words, qualification and disqualification decisions have a effect on the demand by operation. Modeling the fact that the demand by operation is endogenous could lead to more effective qualification decisions on manufacturing performances. Two options could be considered. A first option consists in an iterative solution procedure between projections and qualification decisions. More precisely, a projection of the

demand by operation is first determined, then qualification decisions are optimized. Then, a new projection of the demand by operation is determined given the new qualification decisions, and so on. A second option to model an endogenous demand would be to assume that the demand by operation is uncertain.

8.2.2 Further improving tactical qualification management

In Section 8.2.2.1, perspectives related to the tactical decision level are discussed. In particular, we believe it would be relevant to better model production variability, which may induce an (too) optimistic satisfaction of production capacities. In Section 8.2.2.2, it is argued that investigating qualification management problems at the strategic level, at the supply chain level, by following the work of Liao et al. (2017) is a prominent research avenue.

8.2.2.1 Improving tactical qualification management

We believe the following perspectives are worth investigating in the future for tactical qualification management:

1. Other parameters than demand can also be subject to uncertainty, *e.g.* production capacities, throughput rates of operations on machines, qualification costs and lead times. Studying the relevance and effect of uncertainty on these parameters can be valuable.

2. Considering qualification available times due to disqualifications could be interesting to further distinguish best new qualifications.

3. A large number of qualifications can be difficult to maintain at an operational level. Including disqualification decisions, *e.g.* constraining the number of qualifications by machines or constraining the total number of qualifications in each period, could be relevant.

4. Extending the static robust reformulation to adjustable robust reformulations may be valuable to further reduce qualification costs.

5. For work centers with a large number of operations and machines, efficient solution approaches can be valuable. An option consists in using a cutting-plane solution approach with lazy constraints as proposed by Bertsimas et al. (2016). This might be a viable approach as the computational time required to solve MCQCP is small.

6. As there may exist several sets of robust qualifications in terms of number of qualifications given an immunization level, it would be interesting to use additional objective functions to select the most robust set of qualifications. This would lead to considering a multi-objective optimization approach.

7. Other solution approaches for MRBQP can be considered. Iterated max-min approaches are probably relevant not to restrict to the same value of θ for all products and periods.

8. Studying the effect of different qualification cost profiles by machine or of time dependent qualification decisions on the price of uncertainty can be interesting.

9. As the ability of qualifications to cover demand uncertainty strongly depends on the machines in the work center (Hopp and Spearman, 2011), considering the investment decisions in terms of machines could also be investigated.

10. Capacity allocation tends to be optimistic. For a given operation, the workload is allocated to the fastest machines in priority to satisfy capacity constraints. Nevertheless, because dispatching decisions are highly automated and due to production variability in semiconductor factories, it is unlikely that operational decisions will exactly match capacity allocation decisions. Machines can ultimately have larger workload and utilization rates than expected, because all operation quantities will not necessarily be allocated to fastest machines. This difference should be considered as it may lead to capacity constraint violations (see Appendix G for more details).

8.2.2.2 From tactical to strategic qualification management

Strategic qualification management is also a prominent research avenue to reduce production costs. Strategic qualification management seeks to determine the most relevant set of wafer fabs to qualify for each product or technology. The number of wafer fabs to qualify should be minimized, in particular for new products, as developing qualifications is expensive. To the best of our knowledge, Liao et al. (2017) are the only authors who deal with strategic qualification management in semiconductor manufacturing. We believe that our work on tactical qualification management in Chapter 6 can be extended for strategic qualification management. For instance, machines can be replaced by wafer fabs, and production capacities of machines by production capacities of wafer fabs. Therefore, the deterministic and robust optimization models proposed in Chapter 6 should still be valid for strategic qualification management. The optimization model characterizing the robustness of a qualification configuration should also be valid, and could be used to characterize the robustness of the company to cover demand uncertainty. Nevertheless, to better suit strategic decisions, our work on tactical qualification management probably requires to be extended by including additional features such as outsourcing costs, production costs, profits and yields.

This work is not limited to the semiconductor industry, and can be extended to any industry that has to deal with similar process flexibility issues.

8.2.3 Industrial perspectives

The decision support system can be improved. With small changes, it could be used by other wafer fabs. The decision support system could also benefit from new developments by modeling specific features of specific work centers. Other relevant reports could be imagined, such as the reporting of critical qualifications. Critical qualifications could correspond to the qualifications that lead to a large decrease

of the utilization balance if they are lost. Critical qualifications could also be the smallest set of qualifications that need to be maintained to keep the same utilization balance and total utilization rate of the machines than with the initial set of qualifications. This could encourage work centers to quickly re-qualify critical qualifications if they are lost. Further studying the effect of currently down machines on manufacturing performances could be extremely relevant for work center managers to better prioritize and plan maintenance operations.

Bibliography

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. 23, 24, 67, 104

ASML (December 2011). Lithography machine. `https://spectrum.ieee.org/semiconductors/devices/euv-faces-its-most-critical-test`. accessed in 2020. vii, 9

Atamtürk, A., Nemhauser, G. L., and Savelsbergh, M. W. (2000). Conflict graphs in solving integer programming problems. *European Journal of Operational Research*, 121(1):40–55. 158

Atherton, L. F. and Atherton, R. W. (1995). *Wafer fabrication: Factory performance and analysis*, volume 339. Springer Science & Business Media. 8, 10

Aubry, A., Jacomino, M., Rossi, A., and Espinouse, M.-L. (2012). Maximizing the configuration robustness for parallel multi-purpose machines under setup cost constraints. *Journal of Scheduling*, 15(4):457–471. 24, 148

Aubry, A., Rossi, A., Espinouse, M.-L., and Jacomino, M. (2008). Minimizing setup costs for parallel multi-purpose machines under load-balancing constraint. *European Journal of Operational Research*, 187(3):1115–1125. 24, 40, 205

Aurand, S. S. and Miller, P. J. (1997). The operating curve: A method to measure and benchmark manufacturing line productivity. In *1997 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop ASMC 97 Proceedings*, pages 391–397. IEEE. 105

Bandi, C., Bertsimas, D., and Youssef, N. (2015). Robust queueing theory. *Operations Research*, 63(3):676–700. 131, 186

Bard, J. F. (1991). Some properties of the bilevel programming problem. *Journal of optimization theory and applications*, 68(2):371–378. 69

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2013). *Nonlinear programming: theory and algorithms*. John Wiley & Sons. 33, 46

Ben-Ayed, O. and Blair, C. E. (1990). Computational difficulties of bilevel linear programming. *Operations Research*, 38(3):556–560. 69

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*, volume 28. Princeton University Press. 137

Ben-Tal, A., Goryashko, A., Guslitzer, E., and Nemirovski, A. (2004). Adjustable robust solutions of uncertain linear programs. *Mathematical programming*, 99(2):351–376. 144

Ben-Tal, A. and Nemirovski, A. (2002). Robust optimization–methodology and applications. *Mathematical Programming*, 92(3):453–480. 137, 143

Benjaafar, S., Talavage, J., and Ramakrishnan, R. (1995). The effect of routeing and machine flexibility on the performance of manufacturing systems. *International Journal of Computer Integrated Manufacturing*, 8(4):265–276. 138

Bertsimas, D., Brown, D. B., and Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM review*, 53(3):464–501. 137

Bertsimas, D., Dunning, I., and Lubin, M. (2016). Reformulation versus cutting-planes for robust optimization. *Computational Management Science*, 13(2):195–217. 166, 188

Bertsimas, D., Gupta, V., and Kallus, N. (2018). Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292. 131, 187

Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations research*, 52(1):35–53. 28, 135, 142, 143

Bertsimas, D. and Thiele, A. (2006a). Robust and data-driven optimization: modern decision making under uncertainty. In *Models, methods, and applications for innovative decision making*, pages 95–122. INFORMS. 137

Bertsimas, D. and Thiele, A. (2006b). A robust optimization approach to inventory theory. *Operations research*, 54(1):150–168. 164

Bidkhori, H., Simchi-Levi, D., and Wei, Y. (2016). Analyzing process flexibility: A distribution-free approach with partial expectations. *Operations Research Letters*, 44(3):291–296. 19, 20

Birge, J. R. and Louveaux, F. (2011). *Introduction to stochastic programming*. Springer Science & Business Media. 137

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press. 74

Bracken, J. and McGill, J. T. (1973). Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44. 67, 69

Brown, S. M., Hanschke, T., Meents, I., Wheeler, B. R., and Zisgen, H. (2010). Queueing model improves ibm's semiconductor capacity and lead-time management. *Interfaces*, 40(5):397–407. 105

Can, B. and Heavey, C. (2016). A demonstration of machine learning for explicit functions for cycle time prediction using mes data. In *2016 Winter Simulation Conference (WSC)*, pages 2500–2511. IEEE. 106

Chang, X. and Dong, M. (2017). Stochastic programming for qualification management of parallel machines in semiconductor manufacturing. *Computers & Operations Research*, 79:49–59. 23, 65, 66, 88, 135, 142

Chen, X., Ma, T., Zhang, J., and Zhou, Y. (2019). Optimal design of process flexibility for general production systems. *Operations Research*, 67(2):516–531. 20, 138

Chou, M. C., Chua, G. A., Teo, C.-P., and Zheng, H. (2010). Design for process flexibility: Efficiency of the long chain and sparse structure. *Operations research*, 58(1):43–58. 19, 20, 138, 156

Christ, Q. (2020). *Optimisation et aide à la décision pour la planification de production opérationnelle en fabrication de semiconducteurs*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne. 113, 178

Clark, A. R. and Clark, S. J. (2000). Rolling-horizon lot-sizing when set-up times are sequence-dependent. *International Journal of Production Research*, 38(10):2287–2307. 101

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press. 115

Curcio, E., Amorim, P., Zhang, Q., and Almada-Lobo, B. (2018). Adaptation and approximate strategies for solving the lot-sizing and scheduling problem under multistage demand uncertainty. *International Journal of Production Economics*, 202:81–96. 101

Daskin, M. S. (2011). *Network and discrete location: models, algorithms, and applications*. John Wiley & Sons. 43

de Ruiter, F. J., Ben-Tal, A., Brekelmans, R. C., and den Hertog, D. (2017). Robust optimization of uncertain multistage inventory systems with inexact data in decision rules. *Computational Management Science*, 14(1):45–66. 162

de Ruiter, F. J., Brekelmans, R. C., and den Hertog, D. (2016). The impact of the existence of multiple adjustable robust solutions. *Mathematical Programming*, 160(1-2):531–545. 162

De Toni, A. and Tonchia, S. (1998). Manufacturing flexibility: a literature review. *International journal of production research*, 36(6):1587–1617. 16, 134

Deloitte (April 2019). Semiconductor - the next wave. `https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/technology-media-telecommunications/deloitte-cn-tmt-semiconductors-the-next-wave-en-190422.pdf`. accessed in 2020. 6

Delp, D., Si, J., and Fowler, J. W. (2006). The development of the complete x-factor contribution measurement for improving cycle time and cycle time variability. *IEEE Transactions on Semiconductor Manufacturing*, 19(3):352–362. 8, 105

Deng, T. and Shen, Z.-J. M. (2013). Process flexibility design in unbalanced networks. *Manufacturing & Service Operations Management*, 15(1):24–32. 20

Désir, A., Goyal, V., Wei, Y., and Zhang, J. (2016). Sparse process flexibility designs: is the long chain really optimal? *Operations Research*, 64(2):416–431. 19

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191. 115

Fiorotto, D. J., Jans, R., and de Araujo, S. A. (2018). Process flexibility and the chaining principle in lot sizing problems. *International Journal of Production Economics*, 204:244–263. 20, 138

Fischetti, M., Ljubić, I., Monaci, M., and Sinnl, M. (2017). A new general-purpose algorithm for mixed-integer bilevel linear programs. *Operations Research*, 65(6):1615–1637. 67

Fischetti, M., Ljubić, I., Monaci, M., and Sinnl, M. (2018). On the use of intersection cuts for bilevel optimization. *Mathematical Programming*, 172(1):77–103. 85, 186

Fowler, J., Brown, S., Gold, H., and Schoeming, A. (1997). Measurable improvements in cycle-time-constrained capacity. In *1997 IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings (Cat. No. 97CH36023)*, pages A21–A24. IEEE. 23, 24, 67, 105

Fu, M., Askin, R., Fowler, J., and Zhang, M. (2015). Stochastic optimization of product–machine qualification in a semiconductor back-end facility. *IIE Transactions*, 47(7):739–750. 22, 135

Fu, M., Haghnevis, M., Askin, R., Fowler, J., and Zhang, M. (2010). Machine qualification management for a semiconductor back-end facility. In *Proceedings of the 2010 Winter Simulation Conference*, pages 2486–2492. IEEE. 22

Garey, M. R. and Johnson, D. S. (1979). *Computers and intractability*, volume 174. freeman San Francisco. 40, 205

Gorissen, B. L., Yanıkoğlu, İ., and den Hertog, D. (2015). A practical guide to robust optimization. *Omega*, 53:124–137. 136, 137, 143, 155, 162

Graves, S. C. and Tomlin, B. T. (2003). Process flexibility in supply chains. *Management Science*, 49(7):907–919. 19, 138, 164

Gurumurthi, S. and Benjaafar, S. (2004). Modeling and analysis of flexible queueing systems. *Naval Research Logistics (NRL)*, 51(5):755–782. 63, 64, 66, 77, 99, 100, 128

Hahn, P. O. (2001). The 300 mm silicon wafer—a cost and technology challenge. *Microelectronic Engineering*, 56(1-2):3–13. 8

Hopp, W. J. and Spearman, M. L. (2011). *Factory physics*. Waveland Press. 9, 12, 15, 105, 112, 115, 138, 141, 160, 163, 166, 189

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. 67, 129

Hutcheson, J. D. (2000). Introduction to semiconductor equipment. *Handbook of Semiconductor Manufacturing Technology*, pages 23–33. 10

Iancu, D. A. and Trichakis, N. (2014). Pareto efficiency in robust optimization. *Management Science*, 60(1):130–147. 162

Ignizio, J. P. (2009). Cycle time reduction via machine-to-operation qualification. *International Journal of Production Research*, 47(24):6899–6906. 22, 67, 104

Ignizio, J. P. (2010). The impact of operation-to-tool dedications on factory stability. In *Proceedings of the Winter Simulation Conference*, pages 2606–2613. Winter Simulation Conference. 22, 67, 104

Jain, A., Jain, P., Chan, F. T., and Singh, S. (2013). A review on manufacturing flexibility. *International Journal of Production Research*, 51(19):5946–5970. 16, 134

Johnzén, C. (2009). *Modeling and optimizing flexible capacity allocation in semiconductor manufacturing*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne. 1, 21, 24, 25, 40, 181, 205

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. 14, 17, 24, 27, 32, 138, 163

Johnzén, C., Dauzère-Pérès, S., Vialletelle, P., and Yugma, C. (2007). Importance of qualification management for wafer fabs. In *2007 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 166–169. IEEE. 14, 17, 21

Johnzén, C., Dauzère-Pérès, S., Vialletelle, P., and Yugma, C. (2009). Optimizing flexibility and equipment utilization through qualification management. In *2009 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pages 137–140. IEEE. 24, 25

Johnzén, C., Vialletelle, P., Dauzère-Pérès, S., Yugma, C., and Derreumaux, A. (2008). Impact of qualification management on scheduling in semiconductor manufacturing. In *Proceedings of the 40th Conference on Winter Simulation*, pages 2059–2066. Winter Simulation Conference. 21, 24, 63, 64, 66, 77, 99, 100, 128

Jordan, W. C. and Graves, S. C. (1995). Principles on the benefits of manufacturing process flexibility. *Management Science*, 41(4):577–594. 18, 20, 138, 156, 164

Juang, J.-Y. and Huang, H.-P. (2000). Queueing network analysis for an ic foundry. In *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on*, volume 4, pages 3389–3394. IEEE. 106

Kabak, K. E., Heavey, C., Corbett, V., and Byrne, P. (2013). Impact of recipe restrictions on photolithography toolsets in an asic fabrication environment. *IEEE Transactions on Semiconductor Manufacturing*, 26(1):53–68. 14, 17, 23, 24, 67, 104

Kacar, N. B., Irdem, D. F., and Uzsoy, R. (2011). An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms. *IEEE Transactions on Semiconductor Manufacturing*, 25(1):104–117. 110

Kalir, A. and Bouhnik, S. (2006). Achieving reduced cycle times in semiconductor manufacturing via effective control of the pk equation factors. *IFAC Proceedings Volumes*, 39(3):65–69. 105

Kim, K. and Chhajed, D. (2000). Commonality in product design: Cost saving, valuation change and cannibalization. *European Journal of Operational Research*, 125(3):602–621. 137

Klemmt, A., Lange, J., Weigert, G., Lehmann, F., and Seyfert, J. (2010). A multistage mathematical programming based scheduling approach for the photolithography area in semiconductor manufacturing. In *Proceedings of the Winter Simulation Conference*, pages 2474–2485. Winter Simulation Conference. 23, 135

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. 84

Kopp, D. and Mönch, L. (2018). A hierarchical approach to qualification management in wafer fabs. In *2018 Winter Simulation Conference (WSC)*, pages 3514–3525. IEEE. 23

Kopp, D. and Mönch, L. (2019). Fast heuristics for making qualification management decisions in wafer fabs. In *2019 Winter Simulation Conference (WSC)*, pages 2348–2359. IEEE. 22

Kopp, D., Mönch, L., Pabst, D., and Stehli, M. (2016). An optimization model for qualification management in wafer fabs. In *Proceedings of the 2016 Winter Simulation Conference*, pages 2610–2620. IEEE Press. 14, 22, 67

Kopp, D., Mönch, L., Pabst, D., and Stehli, M. (2018). Simulation-based performance assessment of tool requalification strategies in wafer fabs. In *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, pages 638–641. IEEE. 14, 17, 22, 67, 88, 104, 105

Kopp, D., Mönch, L., Pabst, D., and Stehli, M. (2019). Qualification management in wafer fabs: Optimization approach and simulation-based performance assessment. *IEEE Transactions on Automation Science and Engineering*. 22, 63, 64, 66, 67, 77, 84, 88, 100, 104, 128

Kotcher, R. and Chance, F. (1999). Capacity planning in the face of product-mix uncertainty. In *1999 IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings (Cat No. 99CH36314)*, pages 73–76. IEEE. 134

Lappas, N. H. and Gounaris, C. E. (2018). Robust optimization for decision-making under endogenous uncertainty. *Computers & Chemical Engineering*, 111:252–266. 149, 150

Leachman, R. (2015). Cycle time management. `http://courses.ieor.berkeley.edu/ieor130/CT%20Management%20Nov%202015.pdf`. accessed in 2020. 104, 110, 130

Leachman, R. C. (2012). The engineering management of speed. In *Proceedings of the 2012 Industry Studies Association Annual Conference*. 105

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. 8

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. 25

Li, N., Zhang, M. T., Deng, S., Lee, Z.-H., Zhang, L., and Zheng, L. (2007). Single-station performance evaluation and improvement in semiconductor manufacturing: A graphical approach. *International Journal of Production Economics*, 107(2):397–403. 105

Liao, B., Yano, C. A., and Esturi, S. (2017). Optimizing site qualification across the supply network at western digital. *Interfaces*, 47(4):305–319. 17, 22, 25, 135, 188, 189

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. 76

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*, pages 1–25. 76

Little, J. D. (2011). Or forum—little's law as viewed on its 50th anniversary. *Operations research*, 59(3):536–549. 108, 109

Little, J. D. and Graves, S. C. (2008). Little's law. In *Building intuition*, pages 81–100. Springer. 108, 109

Löhndorf, N. (2016). Java interface for the clp solver. `https://github.com/quantego/clp-java`. accessed in 2017. 53, 76, 96, 229

Lougee-Heimer, R. (2003). The common optimization interface for operations research: Promoting open-source software in the operations research community. *IBM Journal of Research and Development*, 47(1):57–66. 53, 76, 96, 229

Mak, H.-Y. and Shen, Z.-J. M. (2009). Stochastic programming approach to process flexibility design. *Flexible services and manufacturing journal*, 21(3-4):75–91. 20

Mason, S. J., Fowler, J. W., and Matthew Carlyle, W. (2002). A modified shifting bottleneck heuristic for minimizing total weighted tardiness in complex job shops. *Journal of Scheduling*, 5(3):247–262. 9, 12

McKinsey & Company (Autumn 2011). Mckinsey on semiconductors. accessed in 2020. 6, 8, 13

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. 129

Mhiri, E. (2016). *Planification de la production à capacité finie dans un contexte à forte variabilité, application à l'industrie des semi-conducteurs*. PhD thesis. 178

Miltenburg, J., Cheng, C. H., and Yan, H. (2002). Analysis of wafer fabrication facilities using four variations of the open queueing network decomposition model. *Iie Transactions*, 34(3):263–272. 105

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. vii, 12

Mönch, L., Fowler, J. W., and Mason, S. J. (2012). *Production planning and control for semiconductor wafer fabrication facilities: modeling, analysis, and systems*, volume 52. Springer Science & Business Media. 8, 9, 10, 12

Mönch, L., Uzsoy, R., and Fowler, J. W. (2018). A survey of semiconductor supply chain models part i: semiconductor supply chains, strategic network design, and supply chain simulation. *International Journal of Production Research*, 56(13):4524–4545. 15

Moorthy, K. S. and Png, I. P. (1992). Market segmentation, cannibalization, and the timing of product introductions. *Management science*, 38(3):345–359. 137

Morrison, J. R. and Martin, D. P. (2006). Cycle time approximations for the g/g/m queue subject to server failures and cycle time offsets with applications. In *The 17th Annual SEMI/IEEE ASMC 2006 Conference*, pages 322–326. IEEE. 105

Morrison, J. R. and Martin, D. P. (2007). Practical extensions to cycle time approximations for the $g/g/m$-queue with applications. *IEEE Transactions on Automation Science and Engineering*, 4(4):523–532. 105

Nattaf, M., Dauzère-Pérès, S., Yugma, C., and Wu, C.-H. (2019). Parallel machine scheduling with time constraints on machine qualifications. *Computers & Operations Research*, 107:61–76. 84

Nohadani, O. and Sharma, K. (2018). Optimization under decision-dependent uncertainty. *SIAM Journal on Optimization*, 28(2):1773–1795. 149, 150

Obeid, A., Dauzère-Pérès, S., and Yugma, C. (2014). Scheduling job families on non-identical parallel machines with time constraints. *Annals of Operations Research*, 213(1):221–234. 14, 84

Ovacik, I. M. and Uzsoy, R. (2012). *Decomposition methods for complex factory scheduling problems*. Springer Science & Business Media. 9, 12

Perraudat, A., Dauzère-Pérès, S., and Vialletelle, P. (2019). Evaluating the impact of dynamic qualification management in semiconductor manufacturing. In *Proceedings of the 2019 Winter Simulation Conference (2019)*, page 12 pages. IEEE Press. 87, 90

Pianne, A., Rivero, L., Dauzère-Pérès, S., and Vialletelle, P. (2016). Ideal and potential flexibility measures for qualification management in semiconductor manufacturing. In *Winter Simulation Conference (WSC), 2016*, pages 2621–2632. IEEE. 25, 163

Potti, K. and Whitaker, M. (2003). Cycle time reduction at a major texas instruments wafer fab. In *Advanced Semiconductor Manufacturing Conference and Workshop, 2003 IEEEI/SEMI*, pages 106–110. IEEE. 105

Quirk, M. and Serda, J. (2001). *Semiconductor manufacturing technology*, volume 1. Prentice Hall Upper Saddle River, NJ. 8

Rossi, A. (2010). A robustness measure of the configuration of multi-purpose machines. *International journal of production research*, 48(4):1013–1033. 24, 148

Rowshannahad, M. (2015). *Qualification Management and Closed-Loop Production Planning in Semiconductor Manufacturing*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne. 1, 21, 25, 181

Rowshannahad, M. and Dauzère-Pérès, S. (2013). Qualification management with batch size constraint. In *2013 Winter Simulations Conference (WSC)*, pages 3707–3718. IEEE. 25, 60, 85, 102, 185, 186

Rowshannahad, M., Dauzère-Pérès, S., and Cassini, B. (2014). Qualification management to reduce workload variability in semiconductor manufacturing. In *Proceedings of the 2014 Winter Simulation Conference*, pages 2434–2443. IEEE Press. 25

Rowshannahad, M., Dauzère-Pérès, S., and Cassini, B. (2015). Capacitated qualification management in semiconductor manufacturing. *Omega*, 54:50–59. 14, 17, 25, 27, 32, 34, 36, 37, 40, 60, 66, 163, 185, 234

Sattler, L. (1996). Using queueing curve approximations in a fab to determine productivity improvements. In *IEEE/SEMI 1996 Advanced Semiconductor Manufacturing Conference and Workshop. Theme-Innovative Approaches to Growth in the Semiconductor Industry. ASMC 96 Proceedings*, pages 140–145. IEEE. 105, 106

Savelsbergh, M. W. (1994). Preprocessing and probing techniques for mixed integer programming problems. *ORSA Journal on Computing*, 6(4):445–454. 158

Schelasin, R. (2011). Using static capacity modeling and queuing theory equations to predict factory cycle time performance in semiconductor manufacturing. In *Proceedings of the Winter Simulation Conference*, pages 2045–2054. Winter Simulation Conference. 105

Schömig, A. and Fowler, J. (2000). Modeling semiconductor manufacturing operations. In *Proceedings of the 9th ASIM dedicated conference simulation in production and logistics*, pages 55–64. vii, 7

Sethi, A. K. and Sethi, S. P. (1990). Flexibility in manufacturing: a survey. *International journal of flexible manufacturing systems*, 2(4):289–328. 16, 134

Shanthikumar, J. G., Ding, S., and Zhang, M. T. (2007). Queueing theory for semiconductor manufacturing systems: a survey and open problems. *IEEE Transactions on Automation Science and Engineering*, 4(4):513–522. 67, 104, 105, 112, 128

Sharda, R., Barr, S. H., and MCDonnell, J. C. (1988). Decision support system effectiveness: a review and an empirical test. *Management science*, 34(2):139–159. 168

Shi, C., Wei, Y., and Zhong, Y. (2019). Process flexibility for multiperiod production systems. *Operations Research*, 67(5):1300–1320. 20

Silicon Connection (May 2020). Front opening unified pod. `https://www.si-cnx.com/products/microelectronics-packaging/front-opening-unified-pod-foup/`. accessed in 2020. vii, 10

Simchi-Levi, D. and Wei, Y. (2012). Understanding the performance of the long chain and sparse designs in process flexibility. *Operations research*, 60(5):1125–1141. 19

Sinha, A., Malo, P., and Deb, K. (2017). A review on bilevel optimization: From classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2):276–295. 67, 69

Stackelberg, H. v. (1952). Theory of the market economy. 67, 69

Stadtler, H. and Kilger, C. (2002). *Supply chain management and advanced planning*, volume 4. Springer. 15

Tamssaouet, K. (2019). *Multiobjective Complex Job-Shop Scheduling: Application to Semiconductor Manufacturing*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne. 21, 178

Veeger, C., Etman, L., Van Herk, J., and Rooda, J. (2010). Generating cycle time-throughput curves using effective process time based aggregate modeling. *IEEE Transactions on Semiconductor Manufacturing*, 23(4):517–526. 106

Wang, J., Zhang, J., and Wang, X. (2017). Bilateral lstm: A two-dimensional long short-term memory model with multiply memory units for short-term cycle time forecasting in re-entrant manufacturing systems. *IEEE Transactions on Industrial Informatics*, 14(2):748–758. 106

Wang, J., Zhang, J., and Wang, X. (2018). A data driven cycle time prediction with feature selection in a semiconductor wafer fabrication system. *IEEE Transactions on Semiconductor Manufacturing*, 31(1):173–182. 129

Wang, X. and Zhang, J. (2015). Process flexibility: A distribution-free bound on the performance of k-chain. *Operations Research*, 63(3):555–571. 19

Whitt, W. and You, W. (2019). Time-varying robust queueing. *Operations Research*, 67(6):1766–1782. 131, 187

Wribhu, A. (2018). Identifying cycle time factors and its relative impact on tools in semi-conductor fab using statistical inferences. In *2018 29th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, pages 170–173. IEEE. 129

Yanıkoğlu, İ., Gorissen, B. L., and den Hertog, D. (2019). A survey of adjustable robust optimization. *European Journal of Operational Research*, 277(3):799–813. 143, 144, 162

Appendix A

Appendix of Chapter 2

A.1 NP-Hardness of the multi-qualification problem

Johnzén (2009) shows that optimizing the "WIP" flexibility measure is a strongly NP-Hard problem by reduction from the 3-partition problem (Garey and Johnson, 1979). The proof is based on the proof in Aubry et al. (2008) for the Minimum Cost Load Balanced Configuration Problem (MCLBCP). Optimizing the "WIP" flexibility measure is a special case of the problem considered in this paper, even when $tp_{r,m} = tp\ \forall r, \forall m$, and $c_m = 1\ \forall m$. Let us recall the proof in Johnzén (2009) for the sake of completeness.

The "WIP" flexibility measure optimization problem, referred as the flexibility problem hereafter, is shown to have the same complexity as the 3-partition problem, which is NP-complete in the strong sense (Garey and Johnson, 1979). The 3-partition problem can be stated as follows: Consider a set S of $3M$ integers $s_1, s_2, ..., s_{3M}$ larger than or equal to one, and a positive integer B such that $\frac{B}{4} < s_r < \frac{B}{2}$ $\forall r \in [1, 3M]$. The question is "Is it possible to partition S into M disjoints sets $SS_1, SS_2, ..., SS_M$ such that $\sum_{s|s_r \in SS_m} s_r = B\ \forall m \in [1, M]$?

The 3-partition problem can be transformed into a special case of the flexibility problem where $tp_{r,m} = tp\ \forall r, \forall m$, and $c_m = 1\ \forall m$. The $3M$ elements s_r are defined as $3M$ products where the demand d_r of operation r is equal to s_r, $\forall r \in [1, 3M]$. The M disjoint sets corresponds to the M different machines. As the M sets are disjoint, the M different machines are disjoint, i.e. demand d_r of operation r can also be assigned to only one machine. With M machines and $3M$ operations, the number of necessary re-qualifications k is equal to $3M$. The "WIP" flexibility measure, F^{WIP}, is equal to one if the machines process the same quantities (see Johnzén 2009). The question for the flexibility problem can be stated as follows: "It is possible to determine a qualification matrix with $3M$ re-qualifications such that machines have the same workload B, i.e. such that $F^{WIP} = 1$?"

Identification of the flexibility problem with the 3-partition problem.

The qualification matrix is built from the 3-partition problem as follows: $q_{r,m} = 1$ if $s_r \in SS_m\ \forall (r, m) \in [1, 3M] \times [1, M]$. Because $\frac{B}{4} < s_r < \frac{B}{2}\ \forall r \in [1, 3M]$, there are three elements in a disjoint set. Each machine $m \in [1, M]$ then processes the following products quantities:

$$U_m = \sum_{r=1}^{3M} d_r q_{r,m} = \sum_{r|s_r \in SS_m} d_r = \sum_{r|s_r \in SS_m} s_r = B$$

Therefore, if the 3-partition problem has an affirmative answer, then the flexibility problem has an affirmative answer.

Identification of the 3-partition problem with the flexibility problem.

It is assumed that the qualification matrix has $3M$ qualifications. Furthermore, exactly one machine is qualified for each operation. This ensures that the demand for an operation is entirely allocated to a single machine. It is also assumed that the quantities by operation can be perfectly balanced on the machines, then we have $U_m = B \; \forall m \in [1, M]$.

The disjoint sets $SS_m \; \forall m \in [1, M]$ contain the elements s_r such that $q_{r,m} = 1$. Each set has three elements. For each disjoint set SS_m, it follows that:

$$\sum_{r|s_r \in SS_m} s_r = \sum_{r|s_r \in SS_m} d_r = \sum_{r=1}^{3M} d_r q_{r,m} = U_m = B \tag{A.1}$$

To conclude, if it is possible to balance the quantities by operation on M machines such that $F^{WIP} = 1$, then it is possible to partition the set S into M disjoint sets such that $\sum_{r|s_r \in S_m} d_r = B$.

A.2 Work center A

Table A.1, resp. A.4, shows the numerical results for work center A and the first, resp. second, qualification configuration for a run time of 30 seconds, while Tables A.2 and A.5 show the numerical results for a run time of 180 seconds. Table A.3, resp. A.6, provides details on the Branch and Bound algorithm for the first, resp. second, qualification configuration, such as the initial relaxation gap at the root node, the final relaxation gap when the algorithm stops, the total number of explored nodes and the number of instances where the optimal solution is found.

A.2.1 First qualification configuration

First, note that all solution approaches determine satisfactory re-qualification plans. However, some solution approaches are more efficient than others. B&B is very efficient at determining excellent re-qualification plans. In less than 30 seconds, B&B determines an optimal solution for 87.5% of the instances. In less than 180 seconds, B&B determines an optimal solution for 93.75% of the instances (see Table A.3). For a computational time limit of 30 seconds, it outperforms all other solution approaches. Even for $k = 40$ or $k = 100$, B&B does not reach the computational time limit on average. The dominance of B&B is confirmed for a computational time limit of 180 seconds. For $k = 40$, the mean gain of 10.6% for a computational time limit of 30 seconds only increases to 10.7% for a computational time limit of 180 seconds. Similarly, for $k = 100$, the mean gain is not significantly improved but B&B is able to prove that the best solution found is numerically optimal by reducing the mean final relaxation gap and by pruning more nodes (see Table A.3). On average, a limited number of nodes is explored before finding an optimal solution. For $k = 1$, the mean number of explored nodes is equal to 0.3 for both computational time limits. For $k = 2$, the mean number of explored nodes is equal to 0.6. In other words, branching is not required for some studied instances because the continuous relaxation at

the root node determines binary values for the re-qualification variables $OQ_{r,m}$, or because IGH provides an optimal solution. For $k = 100$, fewer nodes are explored by B&B than when $k = 40$, which can be counter intuitive because more combinations should be tested. However, as the number of re-qualifications increases, almost all relevant re-qualification decisions are already binary in the continuous relaxation at the root node (due to the nature of data), and thus considered in the initial feasible re-qualification plan determined by IGH. Hence, the required branching effort is reduced because the resulting number of "choices" is smaller. Similarly, almost all relevant re-qualifications are determined by using the k largest dual variables. Therefore, on industrial data, as soon as k exceeds a few re-qualifications, even if the optimization problem is NP-Hard, the theoretical combinatorial aspect of the problem fades.

	GH		GHDP		LS		LSDP		IGH		B&B	
k	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)
1	2.7	4.2	2.7	1.7	2.7	7.9	2.7	3.0	2.1	0.2	2.7	0.5
2	4.1	8.1	4.1	3.0	4.1	18.5	4.1	5.6	3.4	0.2	4.1	1.0
3	5.1	12.5	5.1	4.3	5.1	25.5	5.1	8.4	4.5	0.2	5.1	1.9
4	5.9	17.9	5.9	5.7	5.9	29.5	5.9	11.5	5.1	0.3	5.9	2.2
5	6.5	21.4	6.5	7.0	6.5	31.5	6.6	15.0	5.7	0.2	6.6	3.8
6	6.9	25.4	7.0	8.7	6.9	31.9	7.1	18.5	6.0	0.2	7.1	4.1
7	7.1	27.9	7.4	9.9	7.2	31.6	7.5	20.9	6.3	0.2	7.5	5.7
8	7.3	30.0	7.8	11.2	7.4	31.6	7.8	24.6	6.7	0.2	7.8	5.6
40	7.5	31.7	9.7	30.4	7.6	31.7	9.7	30.3	9.6	0.2	9.9	24.3
100	7.6	31.8	9.7	30.4	7.7	31.9	9.7	30.3	10.6	0.2	10.6	17.5

Table A.1: Mean gain (%) and CPU (s) over all instances for work center A for the first qualification configuration and a run time of 30 seconds by solution approach.

	GH		GHDP		LS		LSDP		IGH		B&B	
k	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)
1	2.7	4.1	2.7	1.7	2.7	7.9	2.7	3.1	2.1	0.2	2.7	0.5
2	4.1	7.9	4.1	2.8	4.1	17.7	4.1	5.6	3.4	0.2	4.1	1.0
3	5.1	12.2	5.1	4.2	5.1	30.0	5.1	8.4	4.5	0.2	5.1	1.8
4	5.9	19.0	5.9	5.6	5.9	45.1	5.9	11.5	5.1	0.2	5.9	2.2
5	6.5	22.8	6.5	7.0	6.6	62.2	6.6	14.9	5.7	0.2	6.6	3.8
6	7.1	29.6	7.0	8.5	7.1	76.9	7.1	18.5	6.0	0.3	7.1	4.2
7	7.5	35.6	7.4	9.9	7.5	92.7	7.5	21.0	6.3	0.3	7.5	5.6
8	7.8	43.8	7.8	11.2	7.8	105.8	7.8	25.0	6.7	0.2	7.8	7.5
40	9.8	173.6	10.4	55.1	10.0	181.7	10.4	145.3	9.6	0.2	10.0	107.0
100	10.0	181.7	10.8	126.8	10.1	181.7	10.8	168.8	10.6	0.2	10.7	56.4

Table A.2: Mean gain (%) and CPU (s) over all instances for work center A for the first qualification configuration and a run time of 180 seconds by solution approach.

IGH determines satisfactory re-qualification plans for a nearly instantaneous result. The mean computational time of IGH does not exceed 0.3 seconds. However, it should be noted that IGH does not necessarily determine the optimal re-qualification plan, in particular for $k = 1$. This may be due to the fact that dual variables are only indicative. They do not necessarily guarantee that the marginal rate can be fully

	30 seconds				180 seconds			
k	Initial Gap	Final Gap	Number Explored Nodes	Number optimal instances	Initial Gap	Final Gap	Number Explored Nodes	Number optimal instances
1	0.60%	0.02%	0.3	24	0.60%	0.02%	0.3	24
2	0.89%	0.06%	0.9	24	0.89%	0.06%	0.9	24
3	0.76%	0.05%	2.7	24	0.76%	0.05%	2.7	24
4	0.92%	0.04%	3.2	24	0.92%	0.04%	3.2	24
5	0.99%	0.04%	5.8	24	0.99%	0.04%	5.8	24
6	1.22%	0.02%	6.6	24	1.22%	0.02%	6.6	24
7	1.30%	0.01%	10.7	24	1.30%	0.01%	10.7	24
8	1.26%	0.03%	12.3	23	1.26%	0.01%	13.5	24
40	0.96%	0.64%	78.7	7	0.96%	0.48%	349.7	13
100	0.20%	0.15%	56.6	12	0.20%	0.10%	184.2	20

Table A.3: Details of the Branch and Bound solution approach for work center A and the first qualification configuration.

reached. In this problem, it is also possible that several dual variables have the same value but, in practice, does not lead to the same gain on the utilization balance of the machines.

GHDP and GH also determine satisfactory re-qualification plans. For a computational time limit of 30 seconds, the mean GHDP performs better on average than GH from $k = 6$. The mean gain with GH is equal to 6.9% whereas the mean gain with GHDP is equal to 7.0%. The larger k, the larger the difference between GHDP and GH. This is due to the fact that, although the mean run time of GH is equal to 26.3 seconds, on several instances GH cannot find a complete re-qualification plan because it reaches the computational time limit. This is confirmed by experiments for $k = 7$ and $k = 8$. This shows that, for a small computational time limit, using dual variables is valuable. For a computational time limit of 180 seconds, GH actually performs slightly better on average than GHDP for $k = 6$ and $k = 7$. For instance, for $k = 7$, the mean gain of the re-qualification plan determined b GH is 7.5% whereas the gain of the re-qualification plan determined by GHDP is 7.4%. This is because the dual variables are only indicative of the marginal increase in the objective function. However, when $k = 40$ or $k = 100$, GHDP determines better re-qualification plans than GH because GH reaches the computational time limit.

When the computational time limit is 180 seconds, GHDP determines re-qualification plans that are at least as good as B&B (see Table A.2). For $k = 40$ and $k = 100$, GHDP performs better than B&B because B&B only finds optimal solutions for a few instances. For $k = 40$, the mean gain with GHDP is equal to 10.4%, whereas the mean gain with B&B is equal to 10.0%. When the computational time limit is 30 seconds, for $k = 40$ and $k = 100$, GHDP performs poorly compared to B&B because the computational time limit of 30 seconds is reached for all instances and the algorithm does not have enough time to find a complete re-qualification plan. For instance, when $k = 40$ or $k = 100$, the mean gain with GHDP is equal to 9.7% whereas the mean gain with B&B is equal to 9.9%. The benefit of the local search is very limited for a substantial increase in the computational time. On

average, LSDP only slightly improves (at most by 0.1%) the mean gain determined by GHDP.

Numerical results show that B&B outperforms all other solution approaches both in terms of solution quality and computational time, and should be run for this work center and the first qualification configuration. This also shows that using empirical observations and dual variables, which are part of the B&B solution approach, is relevant for this work center. Numerical results show that including dual variables to guide solution approaches is relevant, otherwise the search space at each iteration of GH and LS is too large for short computational time limits. Finally, we can observe that GHDP is often close to the optimal solution and can challenge B&B when the computational time limit is 180 seconds.

A.2.1.0.1 Second qualification configuration

From a general perspective, only GHDP and LSDP determine satisfactory re-qualification plans. The use of dual variables is relevant. Contrary to all other solution approaches (except IGH), restricting the search space to the N_{dual} best dual variables at each iteration of GHDP and LSDP "immunizes" both solution approaches against the increase in the number of qualifiable operations on each machine. GH and LS are very inefficient and never complete the first iteration of GH, because there are thousands of re-qualifications to evaluate. Therefore, the search of re-qualification plans for the second qualification configuration shows that the use of dual variables is particularly relevant to determine re-qualification plans. Contrary to the first qualification configuration, LSDP is relevant here and leads to an interesting increase in the utilization balance of the machines. For instance, for $k = 2$, LSDP increases the utilization balance of the machines by 1.2% with respect to GHDP. For $k = 3$, the increase is 1%.

	GH		GHDP		LSDP		IGH		B&B	
k	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)
1	12.0	32.3	15.4	1.7	15.4	3.2	13.7	0.2	15.4	1.8
2	-	-	23.5	2.9	24.7	6.1	15.3	0.3	25.1	6.8
3	-	-	30.7	4.3	31.7	9.4	16.1	0.2	31.3	17.1
4	-	-	35.3	5.8	36.3	14.4	16.4	0.2	29.1	26.9
5	-	-	38.9	7.2	40.2	17.7	17.4	0.3	17.4	30.4
6	-	-	42.2	8.8	43.2	21.4	18.8	0.3	21.6	28.3
7	-	-	44.5	10.0	46.1	24.8	19.9	0.2	19.9	30.4
8	-	-	46.7	11.4	48.6	27.8	20.2	0.2	20.2	30.4
40	-	-	57.7	30.6	57.7	30.7	43.1	0.3	43.1	30.4
100	-	-	57.7	30.6	57.7	30.7	53.0	0.4	53.0	30.4

Table A.4: Mean gain (%) and CPU (s) over all instances for work center A for the second qualification configuration and a run time of 30 seconds by solution approach.

When $k = 1$, GH determines a re-qualification plan that is close in terms of quality to the qualification plan determined by GHDP. However, such a quality in the qualification plans is almost by "chance" because the computational time limit of 30 or 180 seconds is always reached (see Tables A.4 and A.5) and good solution

	GH		GHDP		LSDP		IGH		B&B	
k	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)	Gain (%)	CPU (s)
1	15.0	182.7	15.4	1.7	15.4	3.0	13.7	0.2	15.4	1.8
2	-	-	23.5	2.9	24.7	6.1	15.3	0.3	25.1	6.9
3	-	-	30.7	4.4	31.7	9.4	16.1	0.2	32.1	19.8
4	-	-	35.3	5.9	36.3	14.1	16.4	0.2	35.2	67.9
5	-	-	38.9	7.1	40.2	17.6	17.4	0.2	31.7	154.8
6	-	-	42.2	8.5	43.2	21.4	18.8	0.3	24.7	158.9
7	-	-	44.5	10.0	46.1	25.3	19.9	0.2	20.9	174.6
8	-	-	46.7	11.4	48.6	30.8	20.2	0.3	20.2	180.5
40	-	-	60.4	55.8	61.0	178.8	43.1	0.3	43.1	180.4
100	-	-	62.4	138.5	62.4	180.8	53.0	0.3	53.3	176.2

Table A.5: Mean gain (%) and CPU (s) over all instances for work center A for the second qualification configuration and a run time of 180 seconds by solution approach.

	30 seconds				180 seconds			
k	Initial Gap	Final Gap	Number Explored Nodes	Number optimal instances	Initial Gap	Final Gap	Number Explored Nodes	Number optimal instances
1	2.04%	0.19%	1.8	24	2.04%	0.19%	1.8	24
2	13.78%	0.20%	10.0	24	13.78%	0.20%	10.0	24
3	25.15%	1.63%	32.1	20	25.15%	0.22%	32.5	24
4	35.62%	14.76%	72.5	6	35.62%	4.20%	153.3	21
5	44.27%	44.27%	98.7	0	44.27%	17.62%	422.8	7
6	51.13%	46.28%	91.7	3	51.13%	38.51%	517.3	4
7	57.49%	57.49%	98.6	0	57.49%	55.79%	577.6	1
8	64.39%	64.39%	98.9	0	64.39%	64.39%	597.0	0
40	60.79%	60.79%	98.6	0	60.79%	60.79%	596.8	0
100	32.32%	32.32%	98.7	0	32.32%	31.77%	582.2	1

Table A.6: Details of the branch and bound solution approach for work center A and the second qualification configuration.

are among the first ones tested. Note that the computational time limit is always reached as soon as $k = 1$, only numerical results for $k = 1$ and GH is presented because numerical results would be identical for larger values of k and LS.

Contrary to the first qualification configuration, B&B does not outperform all other solution approaches and perform poorly on a large number of experiments. For a computational time limit of 30 seconds, B&B is only relevant when $k \leq 2$ and quickly determines optimal solutions for all instances (see Table A.6). For a computational time limit of 180 seconds, B&B provides optimal solutions for all instances when $k \leq 3$ (see Table A.6). However, as the number of re-qualifications to make increases, B&B quickly becomes irrelevant. For instance, when the computational time limit is 30 seconds, B&B cannot determine any optimal solution for $k = 7$, $k = 8$, $k = 40$ and $k = 100$. Similar observations can be made when the computational time limit is 180 seconds. Moreover, re-qualification plans determined by the Branch and Bound approach are not close to sub-optimal solutions, *e.g.* when the

final relaxation gap would be lower than 0.5%. When the computational time limit is 30 seconds, from $k = 4$, the mean final gap is larger than 14.76% and reaches 64.36% for $k = 8$. By increasing the computational time limit to 180 seconds, the mean final gap slightly decreases (*e.g.* for $k = 4$, 5, 6 or 7) or does not decrease at all (*e.g.* for $k = 8$) although the number of explored nodes greatly increases compared to the first qualification configuration. For $k > 7$, the number of explored nodes is approximately equal to 100 when the computational time limit is equal to 30 seconds, and is approximately equal to 600 when the computational time limit is equal to 180 seconds.

Contrary to the first qualification configuration, the poor performance of B&B can be explained by the fact that empirical observations that motivate B&B do not longer hold and cause a combinatorial explosion. For instance, many re-qualification decisions are relevant and the continuous relaxation may no longer be strong. Many re-qualifications can be relevant to improve the utilization balance of the machines and the qualification matrix is now dense.

Finally, although the mean run time is still very small (< 0.5 seconds), IGH is less relevant to determine re-qualification plans in the second qualification configuration than in the first qualification configuration. For instance, for $k = 2$, GHDP determines a re-qualification plan that improves the objective function by 23.5%, whereas IGH improves the objective function by 15.3% on average (see Tables A.4 and A.5). For $k = 3$, GHDP determines a re-qualification plan that improves the objective function by 30.7%, whereas IGH improves the objective function by 16.4% (see Tables A.4 and A.5). IGH is far from the best solution found by other solution approaches because many dual variables that rank among the best ones when assessing the initial situation often correspond to the same operation, or the same machine. In practice, qualifying the same operation, or the same machine, many times is irrelevant to efficiently improve the utilization balance of the machines.

From a general perspective, the numerical results show that LSDP is the best option for work center A because it outperforms all other heuristics, even GHDP. Another interesting conclusion that can be drawn from these numerical experiments is that the gain between the first and second qualification configurations are very different. Consider $k = 1$ where the optimal solution is found for all instances by B&B. For the first qualification configuration, the mean gain is equal to 2.7% whereas it is equal to 15.4% for the second qualification configuration. The difference is significant. This shows that machines that cannot be qualified for some operations, *i.e.* such that $q_{r,m} = 0$ in the first configuration, could potentially lead to substantial improvements for the work center in terms of utilization balance of the machines. This may be worth to investigate, and to check if these forbidden qualifications could actually be made, *i.e.* whether the associated $q_{r,m} = 0$ in the first configuration could be changed to $q_{r,m} = 2$.

Appendix B

Appendix of Chapter 3

B.0.1 Combining re-qualifications and disqualifications

Let us introduce another set $\mathcal{R}^{ov}(\mathbf{OQ},\mathbf{DOQ}) = \mathcal{R}^{q}(\mathbf{q},\mathbf{OQ}) \cap \mathcal{R}^{d}(\mathbf{q},\mathbf{DOQ})$. The combined optimization can be defined as follows:

Upper-level optimization model:

$$\max \quad TH = f(\mathbf{U},\mathbf{WIP}) \tag{B.1}$$

$$\text{s.t.} \quad \sum_{r,m} DOQ_{r,m} \leq k' \tag{B.2}$$

$$\sum_{r,m} OQ_{r,m} \leq k \tag{B.3}$$

$$\mathbf{U},\mathbf{WIP} \in \arg\min LBP(\mathbf{OQ},\mathbf{DOQ}) \tag{B.4}$$

$$DOQ_{r,m} \in \{0,1\} \qquad \forall r, \forall m \tag{B.5}$$

$$OQ_{r,m} \in \{0,1\} \qquad \forall r, \forall m \tag{B.6}$$

Lower-level optimization model:

$$LBP(\mathbf{OQ},\mathbf{DOQ}) = \min \quad \sum_{m} U_m^{\gamma} \tag{B.7}$$

$$\text{s.t.} \quad \sum_{m} WIP_{r,m} = d_r \qquad \forall r \in \mathcal{R}^{ov}(\mathbf{OQ},\mathbf{DOQ}) \tag{B.8}$$

$$U_m = \sum_{r} \frac{WIP_{r,m}}{tp_{r,m} c_m} \qquad \forall m \tag{B.9}$$

$$WIP_{r,m} \leq d_r - d_r DOQ_{r,m} \qquad \forall r, \forall m \mid q_{r,m} = 1 \tag{B.10}$$

$$WIP_{r,m} \leq d_r OQ_{r,m} \qquad \forall r, \forall m \mid q_{r,m} = 2 \tag{B.11}$$

$$WIP_{r,m} \leq 0 \qquad \forall r, \forall m \mid q_{r,m} = 0 \tag{B.12}$$

$$WIP_{r,m} \geq 0 \qquad \forall r, \forall m \tag{B.13}$$

Upper-level optimization model: The upper-level optimization model is similar to the one presented in Sections 3.3.3 and 3.3.4. Equation (B.1) is the objective function. Constraints (B.2) and (B.3) ensure that at most k re-qualifications and k' disqualifications. An identical number of re-qualifications and disqualifications, *i.e.* $k = k'$ enables the work center managers to limit the loss of flexibility. Another option consists in replacing constraints (B.2) and (B.3) by a constraint ensuring that the overall number of qualifications does not change with respect to the initial number of qualifications. The surrogate constraint would be defined as $\sum_{r,m} DOQ_{r,m} = \sum_{r,m} OQ_{r,m}$. Constraint (B.4) links the upper and lower decision levels. Constraints (B.5) and (B.6) are binary constraints for disqualification and re-qualification decisions.

Lower-level optimization model: Similarly, the lower-level optimization model is close to the one presented in Sections 3.3.3 and 3.3.4. Equation (B.7) is the objective function. Constraints (B.8) define the flow conservation for all operations that have at least one qualified machine with some capacity over the planning horizon. Constraints (B.9) compute the capacity utilization rate for each machine in the work center. Finally, constraints (B.10)-(B.13) ensure that wafer quantities of operation r to process can only be assigned to machine m if machine m is qualified.

B.0.2 Design of experiments

Tables B.1 and B.2 describe the industrial data. Table B.1 shows the number of operations R and machines M by instance, and the Coefficient of Variability (CV) for the throughput rates. The CV is defined as the standard deviation over the mean of a data set. The CV is used to represent the industrial data for confidentiality reasons. For a given operation $r*$, the "Operation TH CV" is the CV of the throughput rate $tp_{r*,m}$ over all initially qualified and qualifiable machines. For a given machine $m*$, the "Machine TH CV" is the throughput rate $tp_{r,m*}$ over all initially qualified and qualifiable operations. We report the mean and maximum of "Machine TH CV" over all operations and the mean and maximum of "Operation TH CV" over all machines for each instance. The minimum coefficient of variability is not presented because it is always equal to zero. This is because there is always an operation that has only one qualified machine. Table B.2 shows the industrial data on the qualification matrix by instance and by work center. Table B.2 first presents the qualifiable density, *i.e.* the total number of "2" in the qualification matrix over $R \times M$, and the qualified density, *i.e.* the total number of "1" in the qualification matrix over $R \times M$. Table B.2 also presents the "Operation density CV", *i.e.* the mean over all operations of the CV representing the number of initially qualified and qualifiable machines, and the "Machine density CV", *i.e.* the mean over all operations of the CV representing the number of initially qualified and qualifiable operations. Again, the CV is used to represent the industrial data for confidentiality reasons.

The number of operations R varies between 500 and 700, and the number of machines M is approximately equal to 20 in both work centers. In each work center, machines are in practice unrelated. The mean of "Operation TH CV" is approximately equal to 0.10 for both work centers. For some operations, the "Operation TH CV" can be much larger. The maximum of "Operation TH CV" varies between 0.64 and 0.91 for work center A and varies between 0.51 and 1.32 for work center B. In terms of qualifications, work center A and work center B are very different. Machines in work center A are in general qualified for a large number of operations whereas machines in work center B are more dedicated to specific operations. The number of qualifications is therefore much smaller in work center B than in work center A. Moreover, the qualification matrix of work center B is more complex than the qualification matrix of work center A because the "Operation density CV" density and the "Machine density CV" are larger than for work center A. Finally, the coefficient of variability of the demand by operation is high but relatively constant from one instance to another. It is approximately equal to 1.2 for work center A and 1.4 for work center B.

Tables B.1 and B.2 show real industrial data and highlight variability that can be found in complex manufacturing systems. There is not a constant pattern in throughput rates and qualification matrices are complex.

Table B.1: Description of industrial instances (1/2).

	Work center A						Work center B					
			Operation TH CV		Machine TH CV				Operation TH CV		Machine TH CV	
Instance	R	M	Mean	Max	Mean	Max	R	M	Mean	Max	Mean	Max
1	658	18	0.11	0.91	0.59	1.03	549	22	0.11	0.51	0.33	2.24
2	656	18	0.11	0.91	0.59	1.09	562	21	0.11	0.94	1.82	3.33
3	645	18	0.11	0.91	0.59	1.09	561	21	0.12	0.94	1.81	3.31
4	575	18	0.10	0.64	0.44	0.60	554	22	0.11	1.32	0.43	2.72
5	604	18	0.10	0.64	1.45	2.94	591	22	0.10	0.61	1.74	3.31
6	607	18	0.10	0.64	0.47	1.06	570	22	0.11	0.51	1.01	2.48
7	623	18	0.11	0.64	0.44	0.60	571	22	0.12	0.94	1.58	3.30
8	616	18	0.11	0.64	0.44	0.61	575	22	0.12	0.67	0.58	2.83
9	628	18	0.11	0.64	0.44	0.54	583	22	0.11	0.51	0.43	2.82
10	619	18	0.11	0.64	0.44	0.55	575	22	0.11	0.98	0.41	2.92
11	603	18	0.10	0.64	0.45	0.59	577	22	0.11	0.97	0.30	2.56
12	615	18	0.10	0.64	0.45	0.76	576	22	0.11	0.97	0.29	2.45
13	630	18	0.11	0.64	0.45	0.57	586	22	0.11	0.97	0.30	2.54
14	630	18	0.11	0.64	0.45	0.57	575	22	0.11	0.97	0.30	2.54
15	618	18	0.10	0.64	0.46	0.66	581	22	0.12	0.97	0.30	2.45
16	608	18	0.10	0.66	0.45	0.64	585	22	0.12	0.97	2.00	3.00
17	568	18	0.10	0.64	0.46	0.61	558	24	0.12	0.99	0.47	3.02
18	581	18	0.10	0.64	0.44	0.58	563	23	0.12	0.99	0.52	3.03
19	601	18	0.11	0.66	0.46	0.64	583	24	0.15	1.31	1.44	3.24

Table B.2: Description of industrial instances (2/2).

	Work center A				Work center B			
Instance	Qualifiable density	Qualified density	Operation CV density	Machine CV density	Qualifiable density	Qualified density	Operation CV density	Machine CV density
1	2.52	18.66	0.43	0.70	5.09	12.40	0.53	0.89
2	1.74	19.44	0.44	0.65	5.46	12.98	0.54	0.86
3	1.59	19.57	0.44	0.65	5.49	13.08	0.54	0.86
4	3.61	20.23	0.47	0.55	5.17	12.32	0.51	0.90
5	2.45	21.32	0.46	0.56	5.92	11.85	0.52	0.90
6	3.08	20.21	0.45	0.57	5.34	12.40	0.53	0.90
7	3.26	20.11	0.46	0.56	5.25	11.75	0.49	0.89
8	3.31	19.88	0.47	0.56	5.10	11.87	0.50	0.89
9	2.60	20.92	0.47	0.55	4.81	12.86	0.52	0.91
10	2.06	21.15	0.46	0.58	4.81	12.96	0.54	0.92
11	2.27	21.06	0.46	0.58	4.63	13.01	0.50	0.92
12	2.44	20.31	0.44	0.60	4.50	12.80	0.50	0.92
13	2.54	20.62	0.45	0.57	4.45	12.91	0.51	0.91
14	2.54	20.76	0.45	0.57	4.44	12.87	0.52	0.91
15	3.09	20.18	0.47	0.55	4.63	11.90	0.49	0.88
16	3.10	19.65	0.48	0.55	4.56	11.99	0.50	0.88
17	3.45	19.93	0.47	0.55	4.05	11.03	0.54	0.95
18	3.37	20.04	0.47	0.55	4.52	11.45	0.54	0.94
19	3.36	19.44	0.49	0.55	4.65	10.49	0.53	0.98

Appendix C

Appendix of Chapter 4

C.1 Multi-period bi-level optimization approach with re-qualifications and disqualifications

Let us introduce a new binary decision variable, $DOQ_{r,m}$ that is equal to one if there is disqualification decisions for operation r on machine m at $t = 0$. In addition, let us introduce k^{oq} as the number of re-qualifications to perform and k^{doq} the number of disqualifications to perform. Recall that $\mathcal{R}_t^q(\mathbf{Q}, \mathbf{C^{eff}}) = \{r \mid \sum_{m=1|C_{t,m}^{eff}>0}^{M} \mathbb{1}(Q_{t,r,m}) > 0\}$ is the set of operations with at least one qualified machine with some capacity at period t, where $\mathbb{1}(x) = 1$ if $x = 1$, and 0 otherwise (see Section 4.2). Considering disqualification decisions, the multi-period bi-level optimization approach with re-qualifications presented in Section 4.2.3 becomes:

Upper-level optimization problem:

$$\max \quad \sum_{r,t} TH_{t,r} \tag{C.1}$$

$$\text{s.t.} \quad \sum_{r,m} OQ_{r,m} = k^{oq} \tag{C.2}$$

$$\sum_{r,m} DOQ_{r,m} = k^{doq} \tag{C.3}$$

$$I_{t,r} = I_{t-1,r} + d_{t,r} - TH_{t,r} \qquad \forall t > 1, \forall r \tag{C.4}$$

$$I_{1,r} = I_{0,r} + d_{1,r} - TH_{t,1} \qquad \forall r \tag{C.5}$$

$$D_{t,r} = I_{t-1,r} + d_{t,r} \qquad \forall t > 1, \forall r \tag{C.6}$$

$$D_{1,r} = I_{0,r} + d_{t,r} \qquad \forall r \tag{C.7}$$

$$TH_{t,r} = f(\mathbf{U}, \mathbf{WIP}) \qquad \forall t, \forall r \tag{C.8}$$

$$Q_{1,r,m} = OQ_{r,m} \qquad \forall r, \forall m \mid l_{r,m} = 0, q_{r,m} = 2 \tag{C.9}$$

$$Q_{1,r,m} = 0 \qquad \forall r, \forall m \mid l_{r,m} > 0, q_{r,m} = 2 \tag{C.10}$$

$$Q_{t,r,m} = Q_{t-1,r,m} + OQ_{r,m} \qquad \forall t > 1, \forall r, \forall m \mid 1 + l_{r,m} = t, q_{r,m} = 2 \tag{C.11}$$

$$Q_{t,r,m} = Q_{t-1,r,m} \qquad \forall t > 1, \forall r, \forall m \mid 1 + l_{r,m} \neq t, q_{r,m} = 2 \tag{C.12}$$

$$Q_{t,r,m} = q_{r,m} - DOQ_{r,m} \qquad \forall t, \forall r, \forall m \mid q_{r,m} \neq 2 \tag{C.13}$$

$$C_{1,m}^{eff} = \max(c_{1,m} - \sum_{r} c_{r,m}^{loss} OQ_{r,m}, 0) \qquad \forall r, \forall m \tag{C.14}$$

$$C_{1,m}^{neg} = \min(c_{1,m} - \sum_{r} c_{r,m}^{loss} OQ_{r,m}, 0) \qquad \forall r, \forall m \tag{C.15}$$

$$C^{eff}_{t,m} = \max(c_{1,m} + C^{neg}_{t-1,m}, 0) \qquad \forall t > 1, \forall r, \forall m \qquad \text{(C.16)}$$

$$C^{neg}_{t,m} = \min(c_{2,m} + C^{neg}_{t-1,m}, 0) \qquad \forall t > 1, \forall r, \forall m \qquad \text{(C.17)}$$

$$U_{t,m}, WIP_{t,r,m} \in \arg\min LBP_t(\mathbf{D_t}, \mathbf{Q_t}, \mathbf{C^{eff}}) \qquad \forall t \qquad \text{(C.18)}$$

$$OQ_{r,m} \in \{0,1\} \qquad \forall r, \forall m \qquad \text{(C.19)}$$

$$DOQ_{r,m} \in \{0,1\} \qquad \forall r, \forall m \qquad \text{(C.20)}$$

Lower-level optimization problem:

$$LBP_t(\mathbf{D_t}, \mathbf{Q_t}, \mathbf{C^{eff}}) = \min \quad \sum_m U^{\gamma}_{t,m} \qquad \text{(C.21)}$$

$$\text{s.t.} \quad \sum_m WIP_{t,r,m} = D_{r,t} \qquad \forall r \in \mathcal{R}^q_t(\mathbf{Q}, \mathbf{C^{eff}}) \qquad \text{(C.22)}$$

$$U_{t,m} = \sum_r \frac{WIP_{t,r,m}}{tp_{r,m} C^{eff}_{t,m}} \qquad \forall m \mid C^{eff}_{t,m} > 0 \qquad \text{(C.23)}$$

$$U_{t,m} = 0 \qquad \forall m \mid C^{eff}_{t,m} = 0 \qquad \text{(C.24)}$$

$$WIP_{r,m} \leq D_r \qquad \forall r, \forall m \mid Q_{t,r,m} = 1 \qquad \text{(C.25)}$$

$$WIP_{r,m} \leq D_r Q_{t,r,m} \qquad \forall r, \forall m \mid Q_{t,r,m} = 2 \qquad \text{(C.26)}$$

$$WIP_{r,m} \leq 0 \qquad \forall r, \forall m \mid Q_{r,m} = 0 \qquad \text{(C.27)}$$

$$WIP_{r,m} \geq 0 \qquad \forall r, \forall m \qquad \text{(C.28)}$$

$$WIP_{t,r,m} \geq 0 \qquad \forall r, \forall m \qquad \text{(C.29)}$$

Upper-level optimization problem. Equation (C.1) is the objective function that consist in maximizing the throughput over the horizon. The throughput is computed from the workload balancing determined by the lower-level optimization problem (see Chapter 2). Constraint (C.2) sets to k^{oq} the number of qualifications that must be performed at the beginning of the horizon. Constraint (C.3) sets to k^{doq} the number of disqualifications that must be performed at the beginning of the horizon. Constraints (C.4)-(C.5) are flow conservation constraints. Constraints (C.6)-(C.7) compute the demand for all operations and all periods from the current number of wafers in the work center and new arriving wafers. Constraints (C.8) compute the throughput from the utilization balance determined by the lower-level optimization problem (see Chapters 2 and 3). Constraints (C.9)-(C.12) determine the new state of each qualification from re-qualification decisions made at $t = 0$, re-qualification lead times and disqualification decisions. Constraints (C.9)-(C.12) concern the qualifiable pairs (operation r, machine m), *i.e.* such that $q_{r,m} = 2$, while Constraints (4.27) guarantee that the qualification status of the other pairs (operation r, machine m), *i.e.* such that $q_{r,m} = 0$ or $q_{r,m} = 1$, remains the same throughout the planning horizon. Constraints (4.23) and (4.25) ensure that machine m becomes qualified for operation r as soon as the re-qualification lead time $l_{r,m}$ is reached. Constraints (4.24) and (4.26) ensure both that (1) machine m is not qualified for operation r before its lead time and that (2) machine m remains qualified for operation r in the planning horizon once it has been qualified. Constraints (C.13) ensure that machine m is disqualified for operation r throughout the horizon if a disqualification is made for machine m and operation r. Constraints (C.14)-(C.17) ensure

that the effective capacity of machine m at period t if a re-qualification requires a maintenance operation. Constraints (C.18) link the upper-level and lower-level problems. Finally, constraints (C.19) and (C.20) are the binary constraints for the re-qualification and disqualification decisions.

Lower-level optimization problem. Equation (C.21) defines the objective of the lower-level problem, *i.e.* that consists in maximizing the utilization balance and minimizing the total utilization rate of the machines. Constraints (C.22) ensure that all operation quantities must be assigned to qualified and available machines. Operations that have no qualified and available machines are not assigned to machines. For these operations, $TH_{t,r}$ is equal to zero, and $I_{t,r}$ necessarily increases. Constraints (C.23) and (C.24) compute the utilization rate of each machine. Finally, Constraints (C.29) are the non-negativity constraints.

Appendix D

Appendix of Chapter 6

D.1 Linear programming for scenario generation

The linear program (D.1)-(D.4) consists in simulating a (perfect handsight) scenario on the demand from a nominal demand and the uncertainty parameters defined in the uncertainty set $\mathcal{D}_t$. The $\mathbf{w}$ parameters are weights and can be randomly drawn to generate a scenario on the demand. Note that $d_{t,p}$ is a decision variable in the linear program (D.1)-(D.4) as a scenario on the demand must be generated.

$$\min \quad \sum_{t,r} w_{t,r} \sum_{p} r f_{p,r} d_{t,p} \tag{D.1}$$

$$\text{s.t.} \quad d_{t,p} \geq \overline{d}_{t,p} - \widehat{d}_{t,p} \qquad \forall t, \forall p \tag{D.2}$$

$$d_{t,p} \leq \overline{d}_{t,p} + \widehat{d}_{t,p} \qquad \forall t, \forall p \tag{D.3}$$

$$\sum_{p|\alpha_{p,f}=1} d_{t,p} = \eta_{t,f} \qquad \forall t, \forall f \tag{D.4}$$

Equation (D.1) is the objective function that is used to simulate a scenario on the demand from the nominal demand. If weights $\mathbf{w}$ are randomly generated, *e.g.* between -1 and 1, the objective function can be used to generate random scenarios. Constraints (D.2)-(D.4) are the constraints that correspond to the uncertainty set $\mathcal{D}_t$.

D.2 Total overtime minimization for evaluating capacity constraint violations

Let us introduce the new decision variable $O_{t,m}$ for machine m at period t. $O_{t,m}$ is greater than 0 if there is an overtime on machine m at period t. The linear program (D.5)-(D.10) minimizes the total overtime over the planning horizon:

$$\min \quad \sum_{t,m} O_{t,m} \tag{D.5}$$

$$\text{s.t.} \quad \sum_{r} \frac{(\sum_p r f_{p,r} d_{t,p}) WIP_{t,r,m}}{tp_{r,m}} \leq c_{t,m} u_{t,m}^{max} + O_{t,m} \qquad \forall t, \forall m \tag{D.6}$$

$$\sum_{m} WIP_{t,r,m} = 1 \qquad \forall t, \forall r \mid \sum_{p} r f_{p,r} d_{t,p} > 0 \tag{D.7}$$

$$WIP_{t,r,m} \leq q_{r,m} \qquad \forall t, \forall r, \forall m \mid q_{r,m} \neq 2 \tag{D.8}$$

$$WIP_{t,r,m} \leq 0 \qquad \forall t, \forall r, \forall m \mid q_{r,m} = 2 \tag{D.9}$$

$$WIP_{t,r,m} \geq 0 \qquad \forall t, \forall r, \forall m \tag{D.10}$$

Here, q is the initial set of qualifications with new (nominal) qualifications.

Appendix E

Appendix of Chapter 7

E.1 Excel™ file

Are machine unbalanced? In the e-mail body, production personnel have a partial vision of the utilization rates because only the utilization rates of machines that belong to machine sets that do not meet their production objectives is presented. In the Excel™ file, all machines are presented. Figure E.1 provides an example of the chart in FlexQual. Blue bars correspond to the utilization rates. Hatched orange bars correspond to WIP quantities by operation that can only be assigned to the current machine (named station in Figure E.1) because other qualified machines are either down or disqualified. Hatched green bars correspond to the single machine workload that can be assigned to multiple machine sets thanks to cross-qualifications. Hatched green bars indicate to which machines these workload should in fact be allocated in priority to maximize the utilization balance and to maximize the throughput of the work center. Note that hatched bars are also expressed in terms of utilization rate and represent a subtotal of the utilization rate of each machine.

This charts enables production personnel to identify critical machines, *i.e.* machines that are more loaded than the rest of the work center and machines with a large portion of operation quantities that can only be assigned to only one machine (single machine workload). For instance, if a machine has a large portion of single machine workload, work centers may want to postpone maintenance operations or qualify another machines for corresponding operations. Note that solution approaches, in particular solution approaches that seek to maximize the utilization balance and minimize the total utilization rate of the machines, will automatically propose re-qualification decisions, if they exist, to reduce the utilization rates of critical machines.

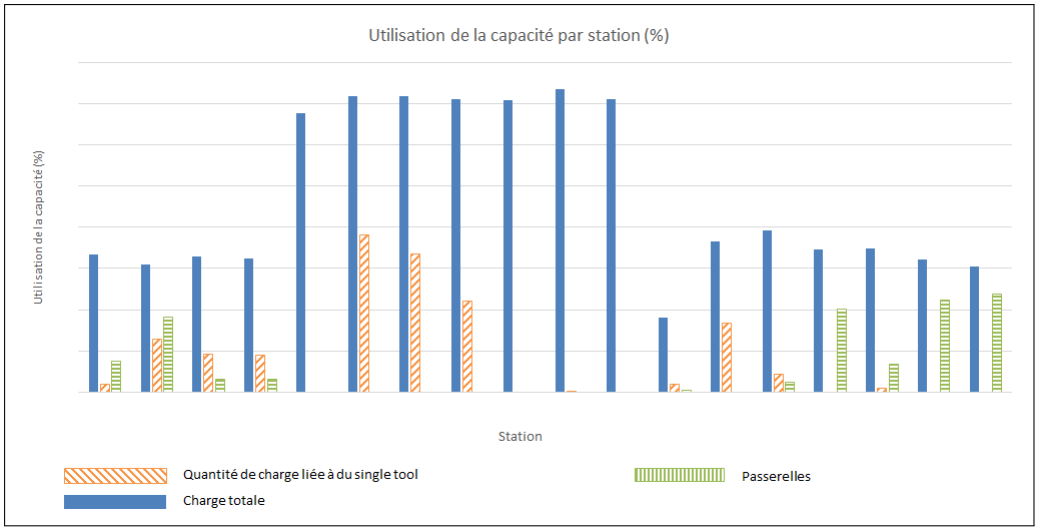


Figure E.1: The utilization rate by machine.

What are the critical re-qualifications that should be made or critical qualifications that should have been active to improve manufacturing performances? Figure E.2 illustrates the presentation of the re-qualification plan in FlexQual. For each re-qualification plan, a collection of re-qualification decisions (*operation*, *station*) is proposed. Some indicators associated to the re-qualification plan are presented. For instance, the throughput after and before the re-qualification is presented. In addition, the minimum and maximum satisfaction of production objectives after and before the re-qualification are presented. Note that in general, only one re-qualification plan is proposed in FlexQual. This is not restrictive as production personnel can define their own re-qualification plan and use FlexQual to evaluate it. For instance, production personnel can select a few re-qualifications among those proposed by double clicking on the qualifications, which will automatically include them in the scenario. Work center managers can also add a few qualifications, for instance, by adding re-qualifications associated to either line stop operations or single machine workload and evaluate the scenario with FlexQual to ensure that, for instance, the throughput is not (or only slightly) decreased.

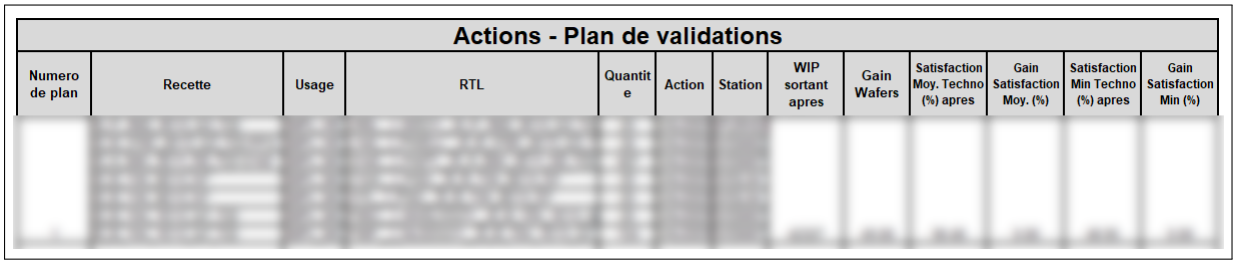


Figure E.2: Proposed re-qualification plan.

Should a maintenance operation be postponed or scheduled sooner? Is there budget, *i.e.* capacity margins, for maintenance or engineering operations on a particular machine or machine set? These are two questions that production

personnel frequently ask themselves. On the one hand, some parts of machines must be replaced, *e.g.* consumables that are empty, but this often requires a maintenance operation. On the other hand, the throughput must be maximized, the cycle time must be minimized and the production objectives must be satisfied. These are two conflicting decisions as machine downtime is one major source of capacity loss, which therefore affects manufacturing performances. To support decision-making, production personnel can use the chart summarizing the satisfaction of production objectives by machine set (see Figure E.3), which gives information on capacity margins with respect to the production objective.

Consider Figure E.3. The production objective (in number of wafers over the considered period) of the machine set is in green. The estimated wafer quantity produced by the machine set corresponds to the blue bar. The estimated wafer quantity that cannot be produced by the machine set, and therefore should remain in the work center by the end of the horizon, corresponds to hatched red bars. Consider the second machine set, starting from right side. The machine set largely satisfies its production objective. Therefore, there is capacity margin for maintenance operations. Work center can therefore schedule maintenance operations that were postponed, or in a more proactive manner, schedule sooner future maintenance operations. Note that production personnel will schedule maintenance operations if machines in the considered machine set do not have a large single machine workload.

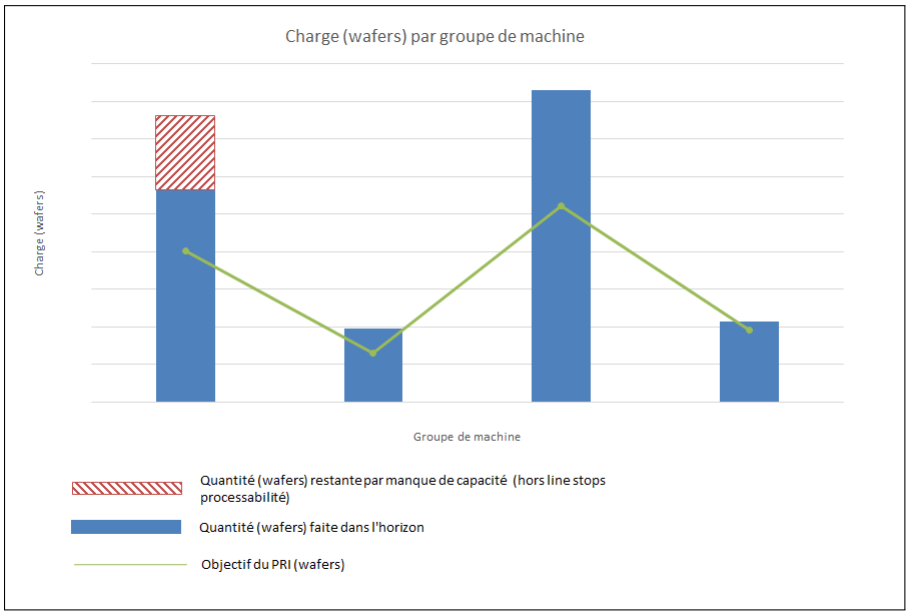


Figure E.3: Throughput by machine group.

Will the work center be able to meet daily production objectives? The chart presented in Figure E.3 is actually primarily used to evaluate if daily production objectives can be met. It provides information on machine sets that largely satisfy their production objective, *e.g.* the second machine set starting from the right side, or on machine set that may not satisfy their production objective, . It also can be

analyzed to have an idea of the problem causing the machine set not to satisfy its production objective. Consider the first machine set starting from the right side. The machine set may be unable to satisfy its production objective because the total number of wafers arriving to the work center is barely greater than the production objective. Such information can lead to better WIP management decisions, *e.g.* by prioritizing the processing of wafers that must be run by the machine set on upstream work centers. Note that production personnel in a given work center are not responsible for making WIP management decisions across all work centers. However, they can communicate this information to other work centers managers, Production Control and WIP management teams.

Scenario management in FlexQual Another critical feature of FlexQual is its capacity to evaluate custom scenarios (what-if analysis) made by production personnel. The scenario interface in FlexQual is kept as simple as possible. It is illustrated in Figure E.4. The scenario interface has the machine list of the work center. For each machine, there are five possibilities. The "Maintenance/Eng" column is used to simulated maintenance or engineering operations on the associated machine. Work center managers can either indicate a number, *e.g.* 24, correspond to the duration of maintenance operations in hours or two dates determining the start and end of maintenance operations. The four other possibilities are associated to qualification and disqualification decisions. In addition, production personnel has the option to include the proposed re-qualification plan by clicking the "Importer plan actions" button, which includes the initially proposed re-qualification decisions into the scenario. In addition, in the "Projection" tab, when double clicking on an entry, a pop menu appears and proposes either to qualify one qualifiable machine or to disqualify one qualified machine. The re-qualification is then automatically included in the "Scenario" tab. After defining a scenario, production personnel has two options: He/she can either ask the server to compute indicators associated to this scenario by clicking the "Jouer scenario rapide" button or by he/she can ask the server the compute indicators to this scenario and propose a new re-qualification plan by assuming that modifications defined in the scenario will be implemented. The scenario management feature is particularly critical for maintenance operations. This is because maintenance operations are currently poorly anticipated and often postponed.

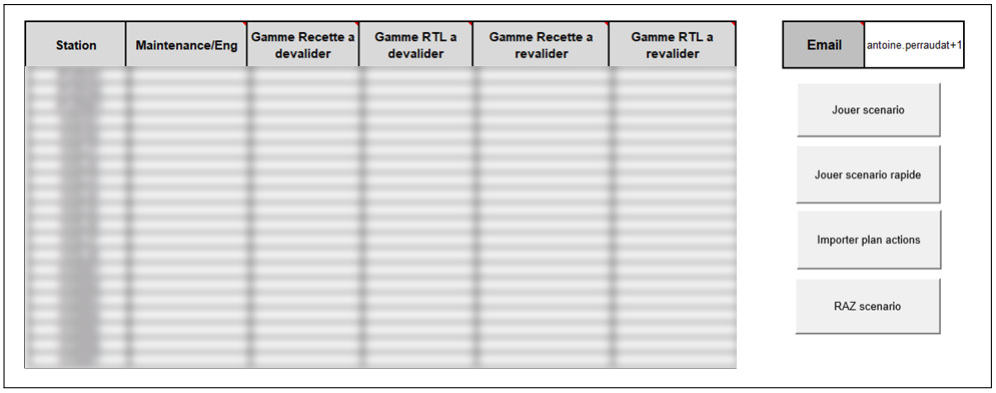


Figure E.4: Scenario tab in FlexQual.

E.2 How does FlexQual work?

WIP projections. In general, the demand by operation is estimated by performing a projection of lots with the help of historical data on the planned cycle time by operation. In other words, each lot is moved forward from the current operation to the next after waiting the planned historical cycle time at the current operation. This procedure is repeated until the sum of cycle times over operations exceeds the duration of the horizon. The throughput rate by operation and by machine is estimated by using statistical models developed and maintained by the Industrial engineering service. The production capacity by machine is also estimated by statistical models developed and maintained by the Industrial engineering service. The production capacity by machine can be refined when production personnel enters scenarios on maintenance operations or with scheduled maintenance operations. It is also possible to define the demand by operation by considering only the WIP which is currently in the considered work center.

The demand by operation could be estimated in a different way. For instance, the demand by operation could be equal to the production objective of the operation (in number of wafers) over the considered horizon. This avoids that the demand is subject to projection uncertainty because production objectives by operation are totally deterministic. Nevertheless, this does not consider the real state of the WIP, *e.g.* WIP peaks (WIP bubbles) at certain operations, which can be critical to improve manufacturing performances. In other words, we could determine optimal re-qualifications no matter the objective function when the demand by operation is equal to the production objective of the operation that has no real effect on manufacturing performances. Therefore, the second option is activated only on large horizons, in general of at least one week.

Machine status. Depending on the horizon, machine status (machine availability on the horizon) is estimated in four different ways: (1) The instantaneous status of the machine is assumed to remain the same over the entire horizon, (2) machines are assumed to be fully available and production personnel uses what-if scenarios to refine machine availability based on the duration of current scheduled and unscheduled down times, (3) scheduled maintenance operations are used to

estimate machine availability, (4) an internal static model is used to estimate, on average, what is the machine availability. (1) or (2) are used for horizons smaller than 24 hours. (2) is used for horizons larger than or equal to 24 hours. (4) is used for larger horizons than a few days. (3) is less frequently used as maintenance operations are frequently postponed due to operational conditions.

Disqualifications. The instantaneous status of a qualification is assumed to remain the same over the entire horizon, unless it is proposed in a re-qualification plan. In this case, it is assumed that no lead time is required and there is no capacity loss due to the re-qualification (no industrial data is available).

Type of qualifications assessed. Depending on the disqualification reasons, given a horizon, some re-qualifications can or cannot be done. Typically, for horizons smaller than a hours days, only disqualifications made for WIP management policies, *e.g.* disqualifying a slow operation on a machine to favour its dispatching on faster machines, can be re-qualified. As the horizon gets larger, the number of possible re-qualifications increases. For horizons larger than a few days/one week, disqualifications for yield losses can be re-qualified.

Mathematical models used. FlexQual includes all mathematical models and solution approaches presented in the thesis. A configuration file manages the use of most relevant mathematical models and solution approaches depending on the horizon and the considered work center.

FlexQual usually propose re-qualification plans optimized by solving the multiperiod bilevel optimization problem presented in Chapter 5. Re-qualification plans are then optimized in terms of throughput. To solve the multiperiod bilevel optimization problem, we combine the use of dual variables and other preprocessing rules (only re-qualifications on machines with utilization rates lower than 1 or only re-qualifications of operations that are faster on currently disqualified machines than currently qualified machines). Nevertheless, the FlexQual is also able to propose to production personnel re-qualification plans that are constructed to optimize the throughput or the mean cycle time. Computational time given to solution approaches are relatively small, *e.g.* a few minutes. This has two main benefits. First, FlexQual is responsive, which is what should be expected for decision-making at a production control level where scenarios are expected to be evaluated. Second, small computation times can generate creativity and the development of innovative solution approaches. Computational times can be larger when decision makers do not need to assess scenarios.

Although re-qualification plans can be determined by only maximizing the utilization balance and minimizing the total utilization rate of the machines, results are always shown by using the bilevel optimization model. Not only does the bilevel optimization models provide insights on the overall throughput or cycle time by work center but it also provides information on the utilization rate of each machine in the work center. This information can then be used and aggregated to standardized levels, *e.g.* by operation family, by product family, by layers, by machine set, in the factory to better support decision making.

Technology used for FlexQual. On the server, linear optimization problems are solved by busing an open source solver (Lougee-Heimer, 2003; Löhndorf, 2016) with Java 8. Python code is used for the server side to detect scenario requests and detect when new automatic e-mail must be sent to production personnel. The Excel^{TM} file also includes Visual Basic Application (VBA) macros to better navigate between different charts in the "Board" tab. It also includes VBA macros to ease the design of a scenario with custom made pop menus.

Appendix F

Robust utilization balancing optimization model

Let $U_{t,m}$ be the utilization rate of machine m at period t, k the given budget for performing qualifications and $\gamma > 1$. The same notations as in Section 6.3.3.2 are used for the model. The workload balancing problem presented in Chapter 2 when the demand is uncertain can be formulated with Equations (F.1)-(F.11):

$$\min \quad \sum_{t,m} U_{t,m}^{\gamma} \tag{F.1}$$

$$\text{s.t.} \quad \sum_{p}(-(\overline{d}_{t,p} - \widehat{d}_{t,p})y_{t,m,p}^{min}) + \sum_{f}(\Gamma_{t,f} y_{t,m,f}^{gamma})$$

$$+ \sum_{p}((\overline{d}_{t,p} + \widehat{d}_{t,p})y_{t,m,p}^{max}) = U_{t,m} \qquad \forall t, \forall m \tag{F.2}$$

$$- y_{t,m,p}^{min} + y_{t,m,p}^{max}$$

$$+ \sum_{f|\alpha_{p,f}=1} y_{t,m,f}^{gamma} \geq \sum_{r} \frac{rf_{p,r} WIP_{t,r,m}}{tp_{r,m}} \qquad \forall t, \forall m, \forall p \tag{F.3}$$

$$\sum_{m} WIP_{t,r,m} = 1 \qquad \forall t, \forall r \mid \sum_{p} rf_{p,r}(\overline{d}_{t,p} + \widehat{d}_{t,p}) > 0 \tag{F.4}$$

$$WIP_{t,r,m} \leq q_{r,m} \qquad \forall t, \forall r, \forall m \mid q_{r,m} \neq 2 \tag{F.5}$$

$$WIP_{t,r,m} \leq \sum_{t'=1|t-t' \geq l_{r,m}}^{t} OQ_{t',r,m} \qquad \forall t, \forall r, \forall m \mid q_{r,m} = 2 \tag{F.6}$$

$$\sum_{t,r,m} cq_{r,m} OQ_{t,r,m} \leq k \tag{F.7}$$

$$y_{t,m,p}^{min}, y_{t,m,p}^{max} \geq 0 \qquad \forall t, \forall m, \forall p \tag{F.8}$$

$$y_{t,m,f}^{gamma} \geq 0 \qquad \forall t, \forall m, \forall f \tag{F.9}$$

$$WIP_{t,r,m} \geq 0 \qquad \forall t, \forall r, \forall m \tag{F.10}$$

$$OQ_{t,r,m} \in \{0,1\} \qquad \forall t, \forall r, \forall m \tag{F.11}$$

The objective function (F.1) maximizing the utilization balance and minimizing the total utilization rate of the machine. Constraints (F.2) and (F.3) are the "robustification" constraints. Constraints (F.2) computes the worst-case utilization rate of machines. Constraints (F.3) are the new constraints introduced by the "robustification" procedure. Constraints (F.4) ensure that the demand is satisfied. Constraints

(F.5)-(F.6) are the qualification constraints. Constraint (F.7) corresponds to the qualification budget. Constraints (F.8)-(F.10) are the non-negativity constraints. Constraints (F.11) are the binary constraints on qualification decisions.

If qualification costs are assumed to be identical for all operations and machines, then algorithms presented in Chapter 2 can be naturally used to solve this problem.

Appendix G

Short-sighted dispatching rules and utilization rate estimation

Capacity allocation in optimization models, including utilization balancing and bilevel optimization models, might be optimistic. Capacity might be allocated to the fastest machine in priority to satisfy capacity constraint. In practice, this is unlikely that the real capacity allocation can be this way. This is because the real capacity allocation is affected by short-sighted dispatching rules and production variability, in particular, time-varying arrival rates, down time, disqualifications. In other words, once a machine is qualified for an operation r, it is likely that the machine runs operation r even though it is not the fastest machine for operation r (see also Chapters 3 and 4). Consequently, if this short-sighted aspect is not considered in the computation of utilization rates of machines, the utilization rates of machines may be underestimated. This is true for both operational and tactical qualification management.

To address this issue, short-sighted dispatching rules should be also be modeled when utilization rates are computed, both at the operational and tactical decision levels. It can be reasonable to assume that capacity allocation variables are subject to some limits that depend on qualifications. Mathematically, this can be modeled with the following constraints:

$$\max_m WIP_{t,r,m} - \min_m WIP_{t,r,m} \leq \theta_r^{disp} \quad \forall t, \forall r \tag{G.1}$$

where $\theta_r^{disp} \in [0,1]$ represents the effect of variability and short-sighted aspect of dispatching rules on capacity allocation. If $\theta_r^{disp} = 0$, then each qualified machine runs the same quantity of operation r. If $\theta_r^{disp} = 1$, then dispatching rule constraints are not considered since $WIP_{t,r,m} \leq 1 \quad \forall t, \forall r, \forall m$. Constraint (G.1) can be linearized with Constraints (G.2)-(G.3):

$$WIP_{t,r,m'} - WIP_{t,r,m} \leq \theta_r^{disp} + (2 - Y_{t,r,m} - Y_{t,r,m'}) \quad \forall t, \forall r, \forall m, \forall m' \mid m \neq m' \mid q_{r,m} > 0 \tag{G.2}$$

$$WIP_{t,r,m'} - WIP_{t,r,m} \geq -\theta_r^{disp} - (2 - Y_{t,r,m} - Y_{t,r,m'}) \quad \forall t, \forall r, \forall m, \forall m' \mid m \neq m' \mid q_{r,m} > 0 \tag{G.3}$$

where $Y_{t,r,m}$ is a state variable indicating if operation r is qualified on machine m at period t. Recall that $q_{r,m} > 0$ when operation r is or can be qualified on machine m. The advantage of this method is that θ_r^{disp} can be computed (extracted) from historical data. However, dispatching rule constraints (G.2)-(G.3) lead to numerous additional constraints and worsen the linear relaxation. Consequently, if dispatching rule constraints are included in optimization models presented throughout the manuscript, they should be computationally more difficult to solve.

Another approach could consist in changing the balancing parameter γ to better represent the short-sighted aspect of dispatching engines as suggested by Rowshannahad et al. (2015).

For tactical qualification management, an approach could consist in reducing the production capacity $c_{t,m}$ for a machine m at period t to model short-sighted aspect of dispatching rules. The main advantage of this approach is that the structures of the optimization problems remain unchanged and no additional constraints are introduced contrary to Constraints (G.2)-(G.3). However, it may be more difficult to define a practical value for $c_{t,m}$.

Combinations of these approaches could be also adopted.

Mines Saint-Étienne

NNT : 2021LYSEM005

Antoine Perraudat

Operational and tactical management of qualifications for flexibility optimization of complex manufacturing systems

Speciality: Industrial Engineering

Keywords: Semiconductor manufacturing industry, Process flexibility, Qualification management, Capacity optimization, Nonlinear programming, Duality, Bilevel optimization, Robust optimization

Abstract:

For semiconductor manufacturing, a qualification is a certification for a machine to process one operation of a specific product. A machine cannot process a product without the associated qualification. Qualifications are mandatory to ensure high yield of production lines and products of quality. Qualifications are used to improve the flexibility (ability to respond effectively to changing circumstances) and to configure production capacities of work centers in semiconductor factories. Because qualifications take time, up to several months, and can be expensive, only relevant qualifications must be planned and determined to optimize the utilization balance of the machines, throughput, fabrication time and demand satisfaction at the lowest cost. More precisely, the following question is answered: Given a horizon, a demand forecast by product and operation, processing times, qualification delays and costs and production capacities of machines, what are the most relevant qualifications?

Answering these questions is actually complex because it is difficult to evaluate the utilization rate of a machine after multiple qualification decisions. Evaluating the utilization rates of the machines is yet primordial as the utilization rates of the machines is related to the throughput, the fabrication time and the demand satisfaction. Because qualification management is complex, we answer these questions from two standpoints by proposing relevant new optimization models and solution approaches. The first standpoint is an operational standpoint where most the follow up of qualifications is optimized. The second standpoint is a tactical standpoint where new qualifications must be planned and anticipated to satisfy the demand. We show throughout the thesis that a small number of relevant qualifications is often sufficient to optimize a given criterion.

Mines Saint-Étienne

NNT : 2021LYSEM005

Antoine Perraudat

Gestion tactique et opérationnelle des qualifications pour l'optimisation de la flexibilité de lignes de fabrication complexes

Spécialité : Génie Industriel

Mots clefs : Industrie de fabrication de composants semiconducteurs, Flexibilité des procédés, Gestion des qualifications, Optimisation de la capacité, Programmation non linéaire, Dualité, Optimisation biniveau, Optimisation robuste

Résumé :

Pour la fabrication de composants semiconducteurs, une qualification est une certification pour une machine permettant de traiter une opération d'un produit spécifique. Une machine ne peut pas traiter un produit sans la qualification associée. Les qualifications sont impératives pour garantir un rendement élevé des lignes de production et des produits de qualité. Les qualifications sont utilisées pour améliorer la flexibilité (capacité à répondre efficacement à des circonstances changeantes) et servent à configurer la capacité de production des ateliers de fabrication dans les usines de semiconducteurs. Comme les qualifications peuvent prendre du temps et sont coûteuses, seules les décisions pertinentes de qualifications et de requalifications doivent être déterminées et planifiées pour optimiser l'équilibre de la charge de travail des machines, la capacité de production, le temps de fabrication et la satisfaction de la demande. Plus précisément, nous voulons répondre à la question suivante : Compte tenu d'un horizon, d'une prévision de la demande par produit et par opération, du temps opératoire par opération, des délais et des coûts de qualification et des capacités de production des machines, quelles les qualifications les plus pertinentes ?

Répondre à cette question est en fait complexe parce qu'il est en général d'évaluer les taux d'utilisation des machines, qui sont liés à la capacité de production, le temps de fabrication et la satisfaction de la demande, après plusieurs décisions de qualifications. Parce que la gestion des qualifications est complexe, nous répondons à cette question sous deux angles en proposant des nouveaux modèles d'optimisation et des nouvelles approches de résolutions pertinentes. Le premier angle est un angle opérationnel où le suivi des qualifications est optimisé, le second angle est un angle tactique où les nouvelles qualifications doivent être planifiées et anticipées pour répondre à la demande. Nous montrons tout au long de la thèse qu'en général, un petit nombre de qualifications, du moment qu'elles sont pertinentes, est souvent suffisant pour optimiser un critère donné.