Optimisation et aide à la décision pour la planification de production opérationnelle en fabrication de semi-conducteurs

Quentin Christ

N°d'ordre NNT : 2020LYSEM002

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'École des Mines de Saint-Etienne

École Doctorale N° 488
Sciences, Ingénierie, Santé

Spécialité de doctorat : Génie Industriel

Soutenue publiquement le 16/01/2020, par :
Quentin Christ

Optimisation et aide à la décision pour la planification de production opérationnelle en fabrication de semi-conducteurs

---

Devant le jury composé de :

Amodeo, Lionel, Professeur, Université de Technologie de Troyes	Président
Kedad-Sidhoum, Safia, Professeur, CNAM Paris	Rapporteuse
Hadj-Hamou, Khaled, Professeur, INSA Lyon	Rapporteur
Amodeo, Lionel, Professeur, Université de Technologie de Troyes	Examinateur
Gicquel, Céline, Professeur assistant, Université Paris Sud	Examinatrice
Dauzère-Pérès, Stéphane, Professeur, EMSE, Gardanne	Directeur de thèse
Absi, Nabil, Professeur, EMSE, Gardanne	Co-directeur
Lepelletier, Guillaume, Ingénieur, STMicroelectronics, Crolles	Encadrant Industriel
Vialletelle, Philippe, Ingénieur, STMicroelectronics, Crolles	Encadrant Industriel
Najji, Jean-Paul, Ingénieur, STMicroelectronics, Rousset	Invité

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

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

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

Mise à jour : 30/08/2018

Remerciements

Parce qu'un projet comme une thèse n'est pas que le travail d'un individu sur une sujet donné, mais l'aboutissement de collaborations et de rencontres d'un ensemble de personnes, je prends ces quelques lignes pour remercier ceux qui m'ont accompagné dans cette aventure.

Tout d'abord, je tiens à remercier mon directeur Stéphane Dauzère-Pérès, avec lequel j'ai pris un immense plaisir à collaborer. Merci pour le temps que tu m'as accordé, malgré une emploie du temps de ministre qui en rien ne su entacher ta disponibilité, que ce soit pour suivre l'avancée de la thèse, ou pour toutes ces discutions stimulantes et enrichissantes qui ont fait que j'ai beaucoup appris à tes cotés. Ensuite, merci à Guillaume Lepelletier, mon encadrant de thèse à ST Mircoelectronics. Que ce soit pour ta disponibilité à toutes mes questions, ton ouverture d'esprit et ta confiance au travers de l'autonomie que tu m'as laissé, mais également ton humour et ta bonne humeur, je ne garde que de bons souvenirs de ces trois années (et plus en fait) à travailler ensemble.

Merci également à mon co-directeur Nabil Absi. Malgré le caractère très opérationnel de la thèse, tu as été à l'écoute et su m'aider lorsque je venais à toi avec mes interrogations. Merci à Philippe Vialletelle, mon co-encandrant industriel, qui a toujours gardé un oeil sur la thèse et a été un support dans la gestion du projet.

Merci à l'ensemble du laboratoire SFL, permanents et non permanents, grâce à qui ce fût toujours un plaisir de venir en mission à Gardanne.

La thèse CIFRE amène également sont lot de rencontres professionnelles. Ce fut notamment le cas des services Industrial Engineering et Wip Management avec qui j'ai eu l'occasion de beaucoup collaborer et que je remercie pour le temps qu'ils m'ont accordé. Merci aussi à toute l'équipe AMM/PM, à Guillaume L, Renaud, Cedric, Soidri, Guilaume S, Claire, Emannuelle, Vincent, Philippe et Emmanuel pour leur accueil et la super ambiance dans laquelle j'ai pu évoluer tout au long de cette thèse.

Merci à tous mes collègues et amis, avec qui j'ai passé plus de trois années, en commençant par le "box de l'ambiance" jusqu'à la tables des "Cool Kids". Parce qu'il y aurait beaucoup trop de mondes à citer, et que tenter d'être exhaustif amène toujours son risque d'oublier quelqu'un, je tenais simplement à remercier toute cette team, de stagiaires, alternants, ingénieur ou docteurs en devenir, qui plus qu'un lieu de travail, ont fait d'ST un lieu où il faisait bon retrouver ses amis.

Je voudrais également remercier ma famille, dont leur support et leur fierté ont été l'un des principaux moteurs durant ces trois années, et bien avant.

Enfin, merci à toi Karelle, ma meilleure amie, ma coéquipière de vie, et maintenant ma femme. Merci pour ton soutiens, tout au long de cette thèse, et notamment dans les derniers moments où, malgré tes propres difficultés, tu as toujours tout fait pour rendre les miennes plus simples à surmonter.

Table des matières

Introduction générale

Le marché de la micro-électronique a connu un formidable essor depuis ces 50 dernières années. Dans les sociétés connectées d'aujourd'hui, ordinateurs, capteurs, centres de données, électronique automobile ou dispositifs portables, les systèmes électroniques sont devenus omniprésents. Au cœur de ces produits, se trouvent les semi-conducteurs, dont les systèmes de production actuels font partie des plus complexes au monde. Compte tenu d'un marché à la concurrence féroce, obligeant à réduire les prix tout en maintenant un haut niveau de service client, l'optimisation des flux de production est devenue un élément critique dans le maintien de la compétitivité des entreprises. Dans cette thèse, nous traitons du problème de planification de production opérationnelle, dont l'objectif est de faire le lien entre la planification de production tactique, définissant les plans de lancement des produits dans l'usine, et l'ordonnancement détaillé de la production. Du fait de la complexité des systèmes de production en fabrication de semi-conducteurs, considérer ce problème frontière est indispensable, mais a fait l'objet de peu de travaux de recherche. Cette thèse a été réalisée en entreprise, l'objectif était donc à la fois d'avancer sur les aspects scientifiques de développement de nouvelles méthodes, mais également de réfléchir à l'intégration de ces méthodes dans le système industriel.

Dans le chapitre 1, nous présentons le contexte économique et industriel dans lequel s'inscrit cette thèse. Nous décrivons le processus de fabrication des semi-conducteurs, lequel fait partie des plus complexes au monde, et nous introduisons les grandes fonctions de planification de production qui y sont rencontrées. Le chapitre se conclut sur la présentation de l'entreprise et du site de production dans lequel s'est déroulée cette thèse, ainsi que sur une présentation de la problématique rencontrée à laquelle nous souhaitons répondre.

Une revue de la littérature autour des problématiques de planification est présentée dans le chapitre 2. Après un aperçu des fonctions de planification au sein de la chaîne logistique, l'accent est mis sur la planification de production et les principaux travaux et concepts qui y sont associés. Nous nous concentrons ensuite sur la planification de production en fabrication de semi-conducteurs, à travers notamment une revue détaillée de la littérature autour des problématiques de planification de production tactique et opérationnelle. Enfin, nous introduisons le problème de planification de production opérationnelle, et positionnons notre travail par rapport à la littérature existante.

Notre problème frontière de planification de production est modélisé dans le chapitre 3. Après avoir montré que le problème est NP-Difficile, et suite à une analyse expérimentale montrant l'impossibilité de résoudre le problème de manière exacte pour des instances industrielles, nous présentons une approche heuristique composée de trois modules. Cette approche est intégrée dans un outil d'aide à la décision, initialement en place dans l'usine, permettant de définir des plans de production opérationnels en quelques minutes. À la fin du chapitre, une étude expérimentale sur de très petites instances souligne les bonnes performances de l'approche heuristique en termes de temps et de qualité des solutions.

Les plans de production fournis par l'outil d'aide à la décision tiennent comptent de la capacité de production de l'usine. Un module important de l'approche heuristique consiste à évaluer l'impact du plan de production sur la charge des machines, c'est le problème d'*équilibrage des charges* qui est étudié dans la chapitre 4. Ce problème est difficile du fait des caractéristiques complexes du système de production en fabrication de semi-conducteurs. Ce problème d'équilibrage des charges revêt de plus une grande importance dans la détection des machines limitantes et pour orienter les décisions de management de la capacité. La version initiale de l'approche heuristique montrait des difficultés dans l'accomplissement de ces tâches, et une nouvelle méthode, inspirée des problèmes de partage équitable des ressources, est introduite, permettant de fournir des solutions d'équilibrage de qualité et riches de sens pour les utilisateurs.

Le chapitre 5 s'intéresse au troisième module de l'approche heuristique, dont le but est d'assurer la faisabilité des plans de production au regard des contraintes de capacité des machines. Pour cela, le module transforme un plan de production initialement non faisable via une procédure de lissage des charges. Dans ce chapitre, nous présentons trois travaux menés en vue d'améliorer l'approche initiale. Nous introduisons et évaluons différentes règles de lissage permettant d'optimiser la qualité des plans de production selon plusieurs critères de performances. Nous avons également développé une nouvelle approche d'évaluation du taux de charge tenant compte des relations entre machines équilibrées. Enfin, face à ces approches ne pouvant transformer la solution courante qu'en la dégradant, une approche complémentaire de lissage par anticipation est introduite afin de tirer profit de l'excès de capacité de certaines machines.

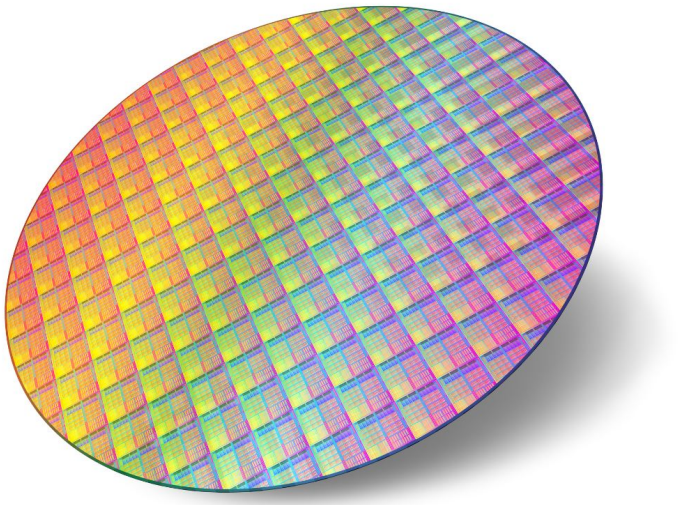
L'un des principaux intérêts d'un projet de thèse en entreprise est de pouvoir à la fois avancer sur les aspects scientifiques de développement de nouvelles méthodes, mais également de réfléchir à l'intégration de ces méthodes dans le système industriel. Dans le chapitre 6, deux réalisations significatives sont présentées. La première concerne la création d'un cadre facilitant le développement, la validation et l'intégration de nouvelles versions de l'approche heuristique dans l'outil de planification de production. La deuxième réalisation est le développement d'une interface permettant aux utilisateurs de facilement disposer des informations pertinentes et nécessaires à la prise de décision.

Chapitre 1

Contexte industriel

.

La production d'une puce électronique peut nécessiter plus d'un millier d'étapes de traitement et prendre jusqu'à 3 mois. Ces étapes sont produites dans les usines parmi les plus chères que l'on puisse trouver dans l'industrie. Dans un marché extrêmement compétitif et à l'intensité capitalistique élevée, prendre les décisions les plus judicieuses aux différents niveaux de planification est primordial. L'objectif de cette thèse est d'améliorer la planification à l'échelle de l'usine afin de mieux piloter les flux de production. Ces travaux de recherche s'inscrivent dans un projet industriel au sein d'un site de production et de R&D, dont les applications peuvent être étendues à d'autres sites, mais aussi à d'autres industries.



1.1 Introduction

Le marché de l'électronique et du numérique a connu une importante croissance au cours des 50 dernières années. Depuis Gordon Moore et l'annonce de sa loi en 1965, selon laquelle la densité d'intégration des dispositifs devait doubler environ tous les deux ans, les applications ont fini par devenir omniprésentes : ordinateurs, capteurs, centres de données, électronique automobile ou dispositifs portables. Aujourd'hui, les appareils électroniques sont partout et la fabrication de semi-conducteurs reste le processus central à la base de toutes ces applications.

Un semi-conducteur est un matériau dont la conductivité est intermédiaire entre celle d'un métal et celle d'un isolant. Son principe de fonctionnement est à la base du fonctionnement des composants de l'électronique moderne : diodes, transistors, etc. C'est ensuite la combinaison de différents types de composants électroniques qui permet la création de Circuits Intégrés (CI), aussi appelés puces électroniques, et dont le but est de reproduire plusieurs fonctions électroniques plus ou moins complexes. Avant d'être vendus, ces CI sont généralement encapsulés dans des boîtiers afin de donner des CI complets, prêts à être montés sur une carte.

Cette position centrale du marché des semi-conducteurs à la base de beaucoup d'autres, explique que la production et la vente de circuits intégrés représentaient en 2015 un chiffre d'affaire dans le monde de plus de $330 milliards, comme le montre la Figure 1.1 issue des données de la Semiconductor Industry Association (SIA (2015)).

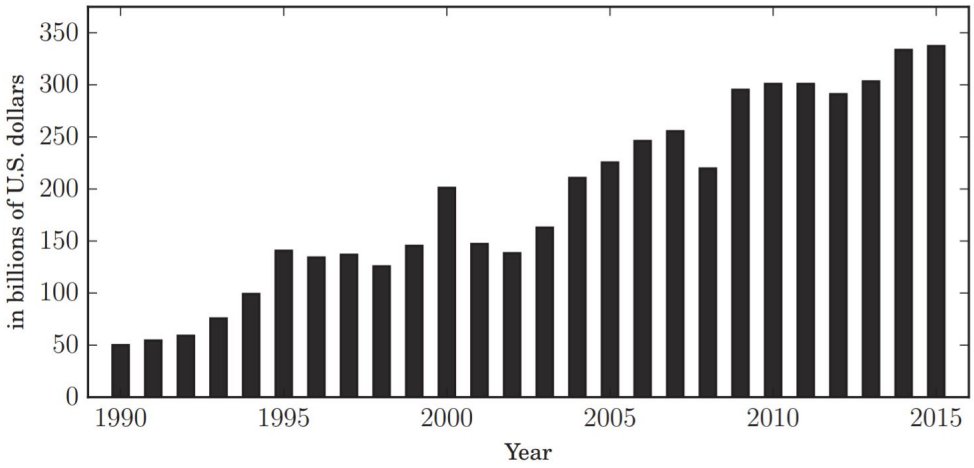


FIGURE 1.1 – Évolution du revenu annuel du marché des semi-conducteurs (SIA (2015))

Cette augmentation importante du marché s'est également accompagnée d'une concurrence de plus en plus forte, forçant les acteurs à être toujours plus compétitifs tant dans le catalogue des produits proposés que leurs prix.

Or, le coût élevé de fabrication des circuits intégrés est un facteur clé de cette industrie. Les salles blanches (environnements où les conditions hygrométriques, de température ou de qualité de l'air sont contrôlées) coûtent très cher avec un prix au mètre carré avoisinant les $10 000 . Les machines représentent un coût important, pouvant aller de quelques centaines de milliers de dollars à plus de $100 millions. On trouve généralement plusieurs centaines de machines dans une seule unité de fabrication de semi-conducteurs (généralement nommée "fab" ou "wafer fab"). Quirk and Serda (2001) rapportent d'ailleurs qu'environ 75% de l'investissement d'une usine est consacré aux machines. Il n'est alors pas étonnant qu'aujourd'hui une seule usine puisse coûter plusieurs milliards de dollards. L'usine fab15

de l'entreprise TSMC, construite en 2010, avait par exemple coûté prêt de $10 milliards, et l'usine fab18 prévue pour 2020 devrait avoir un coût dépassant les $17 milliards (Shilov (2018)). Face à un tel coût, une utilisation efficace des ressources en vue de la maximisation de la rentabilité est primordiale.

Afin d'augmenter la rentabilité, de nombreuses options de réduction des coûts, telles que l'augmentation de la taille des plaquettes et/ou la miniaturisation continue des puces (en lien avec la loi de Moore), ont déjà été largement exploitées, au point qu'il devient difficile de puiser davantage dans ces options. Aujourd'hui, la réduction des coûts d'exploitation à travers de meilleurs systèmes décisionnels est considérée comme une orientation importante. D'un point de vue académique, ce domaine présente également d'importants challenges car l'industrie de la micro-électronique, et notamment la fabrication des circuits intégrés, est considérée comme l'un des systèmes les plus complexes à ce jour (Uzsoy et al. (1992), Mönch et al. (2012)).

1.2 Processus de fabrication des circuits intégrés

1.2.1 Description générique du processus de fabrication

Le processus de fabrication d'un circuit intégré est résumé par la Figure 1.2.

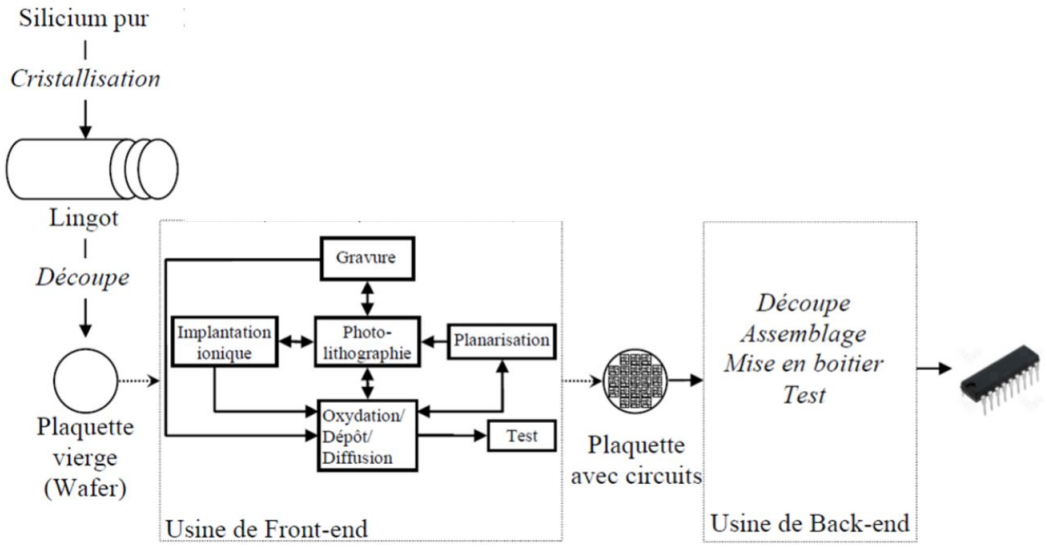


FIGURE 1.2 – Processus de fabrication de circuits intégrés (adapté de Bettayeb (2012))

Les plaquettes brutes sont la matière première utilisée pour produire les circuits intégrés (aussi appelés puces). Une plaquette est une fine tranche de matériau (le plus souvent de silicium) obtenue à partir d'un lingot monocristallin qui a été découpé. À partir d'une seule plaquette, des centaines ou des milliers de puces peuvent être produites, en fonction de la taille des puces et du diamètre de la plaquette. Les diamètres habituels des plaquettes ont augmenté avec le temps pour atteindre des diamètres de 200 mm ou 300 mm, qui sont désormais les tailles standards dans la plupart des usines de fabrication. Un circuit intégré se compose d'une structure tridimensionnelle de résistances, de transistors et d'isolants. Cette structure est créée par ajout successif de couches sur la plaquette. Cette phase de création des différentes couches est faîte dans les usines de *front end*, appelées "*wafer fab*" ou simplement

"*fab*". Une fois les différentes couches déposées sur la plaquette, cette dernière passe une phase de tests afin de détecter les zones non fonctionnelles. Les plaquettes sont généralement transportées dans l'usine par lots de 25, dans des boîtes appelées FOUP (pour *Front Opening Unified Pod*), et dont une illustration est présentée en Figure 1.3.

Une fois les différentes couches réalisées et les tests fonctionnels effectués, les parties fonctionnelles sont gardées, découpées et intégrées dans des boîtiers avant de passer une dernière phase de test. Cette phase est appelée le *back end* et est généralement exécutée dans une usine différente de celle du *front end*.

Notre travail de thèse se concentre sur la partie *front end*, qui est la partie la plus complexe du processus de fabrication. La suite de cette section décrit donc plus en détail le processus de fabrication en jeu dans une usine *front end*.

La fabrication de plaquettes s'effectue dans de grands bâtiments où des conditions strictes de salle blanche doivent être assurées afin d'éviter la contamination de celles-ci, le dépôt d'une particule étrangère de l'épaisseur d'un cheveux pouvant compromettre le fonctionnement final de centaines de puces.

Dans cette première phase, la structure de chaque plaquette est construite couche par couche. Les différentes étapes de traitement nécessaires à la production d'une seule couche sont effectuées dans des zones spécialisées composées de machines ayant des caractéristiques similaires appelées *ateliers*. Chaque couche étant réalisée l'une après l'autre et du fait du nombre important d'étapes de process requis, ceci conduit à des flux réentrants où chaque plaquette visite chaque atelier plusieurs fois. Afin d'assurer la qualité des plaquettes produites, des procédures d'inspection et de mesure sont effectuées entre les étapes de production sur des lots échantillonnés. Si une plaquette est endommagée, cette plaquette peut dans certains cas être traitée de nouveau, sinon elle sera mise au rebut.

L'interaction entre les différents ateliers de fabrication est illustrée dans la phase *front end* de la Figure 1.2. Mönch et al. (2012) définissent 5 grands ateliers, dont une description est proposée ci-dessous:

Diffusion/Oxydation Le processus de diffusion disperse un matériau sur la surface de la plaquette. Le processus d'oxydation fait croître une couche d'oxyde sur la surface d'une plaquette nettoyée. Ces couches sont modifiées par des étapes de traitement ultérieures afin de développer des dispositifs semi-conducteurs connectés (par exemple des transistors, des résistances ou des diodes) qui constituent le circuit intégré. Les étapes de diffusion et d'oxydation peuvent exiger des durées de traitement très élevées, de 12 heures ou plus. La diffusion et l'oxydation sont des procédés à haute température réalisés sur des fours horizontaux ou verticaux. Ces fours sont généralement des machines à *batch*, c'est à dire qu'ils peuvent traiter plusieurs lots de plaquettes en même temps.

Photolithographie Le procédé de photolithographie consiste à altérer une couche de résine photosensible préalablement déposée sur la plaque, via une source lumineuse passant au travers d'un masque. Ce masque représente un motif, dépendant de la puce que l'on souhaite produire et de la couche en cours de réalisation. Les motifs temporaires qui en résultent sont ensuite rendus permanents au cours des étapes ultérieures de gravure ou d'implantation ionique. Pour effectuer une opération de photolithographie, en plus de la machine de photolithographie, un masque (ou réticule) est nécessaire comme ressource auxiliaire. Cette zone de travail contient les machines les plus chères d'une usine qui peuvent coûter jusqu'à $50 millions et constitue souvent, du fait de leur coût, un goulot d'étranglement dans l'ensemble de l'usine.

Gravure Le procédé de gravure permet d'éliminer la matière superflue à la surface de la plaquette. Elle peut-être à motifs ou non. La gravure à motifs enlève, comme son nom l'indique, un motif qui a été déposé sur la plaquette durant la photolithographie. La gravure sans motif réduit l'épaisseur de toute la surface de la plaquette. Il existe deux types de gravure : la gravure sèche expose la surface de la plaquette à un plasma tandis que la gravure humide élimine la matière en utilisant des solutions chimiques.

Implantation Le processus d'implantation introduit des dopants (des ions, des impuretés souhaitées) dans la structure cristalline du matériau semi-conducteur (généralement le silicium) afin de modifier sa conductivité.

Planarisation Le procédé de planarisation utilise la combinaison d'un procédé mécanique et chimique afin d'aplanir la surface de la plaquette et ainsi réduire les écarts d'épaisseur. Ce procédé est réalisé dans l'atelier CMP (pour Chemical and Mechanical Planarization) et est effectué avant chaque ajout d'une nouvelle couche. Cette technique permet d'éviter l'accumulation d'une topologie inégale sur plusieurs couches et évite ainsi de multiples problèmes liés à la non planéité, comme les problèmes de mise au point des lentilles en photolithographie.

Ainsi, on trouve une grande diversité parmi les machines présentes dans ces différents ateliers. Certaines d'entre elles sont capables de traiter plusieurs lots à la fois (fonctionnement par *batchs*) tandis que d'autres traitent chaque plaquette l'une après l'autre. D'autres encore impliquent des temps de préparation en fonction de la séquence de produits réalisés (temps de setup), alors que certaines machines permettent de démarrer une plaquette alors qu'une autre est toujours en cours de process (chevauchement de plusieurs opérations de process).

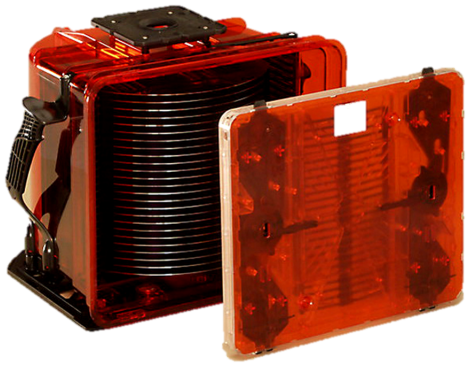


FIGURE 1.3 – Image d'un FOUP contenant 25 plaquettes de silicium [Source: *A300 Carrier Wafer*, http://www.rosefinchtech.com/tw/goods.php?id=72]

1.2.2 Principaux éléments en fabrication de semi-conducteurs

L'industrie des semi-conducteurs, comme beaucoup d'autres industries, apporte son vocabulaire spécifique. Celui-ci peut parfois être en conflit avec le vocabulaire généralement rencontré dans la littérature scientifique, il nous apparaît donc important d'introduire et de définir un certain nombre de termes qui seront régulièrement rencontrés dans cette thèse.

Plaque, plaquette ou "*wafer*" Disque fin d'un matériau semi-conducteur, généralement du silicium, servant de support à la fabrication des circuits intégrés. Quand il est question des plaquettes dans l'usine, ces dernières ne sont généralement plus vierges et sont déjà passées par certaines de leurs étapes de fabrication. Une partie des couches a donc déjà était déposée. Ces plaquettes sont cependant incomplètes et doivent suivre le reste de leur gamme de fabrication avant de pouvoir sortir de l'usine *front end*. Un exemple de plaquette est illustré en Figure 1.4.

"Step", étape Un step caractérise une étape élémentaire de fabrication et correspond au passage d'une plaquette sur une machine. On distingue généralement les steps à valeur ajoutée (steps de process) des steps sans valeur ajoutée (tels que de nettoyage ou de mesure des plaques).

Opération Une opération est un regroupement de plusieurs steps, généralement un step de process et des steps de nettoyage et/ou de métrologie.

Route Une route désigne la gamme de fabrication d'un lot, c'est à dire l'ensemble des étapes de fabrication (aussi appelées *steps*) par lesquelles le lot doit passer avant de pouvoir sortir de l'usine. Il existe une multitude de routes différentes selon le type de produit considéré. Une route se compose généralement de plusieurs centaines d'étapes, et il n'est pas rare aujourd'hui de trouver des routes contenant plus d'un milliers d'étapes de fabrication.

Produit Un produit spécifie le circuit intégré fabriqué. Il requiert un enchaînement spécifique de motifs à réaliser pour chaque couche, et donc un enchaînement spécifique de steps (donc une route spécifique).

Lot Un lot est un ensemble de plaques. Il contient généralement 25 plaquettes au démarrage de sa fabrication, mais peut en avoir moins, notamment si certaines plaquettes doivent être mises au rebut. Les plaquettes constituant un lot sont physiquement disposées dans un FOUP (Figure 1.3). Le lot est l'objet de base de la logistique du système de production, c'est lui qui transite d'équipement en équipement afin que les plaquettes qu'il contient passent sur les différentes étapes.

Recette Ensemble de paramètres servant à piloter un équipement lors de la réalisation d'un step de process donné tels que le type de fluide, la pression, la durée ou la température. La recette est donc dépendante du step considéré, mais aussi du type de produit.

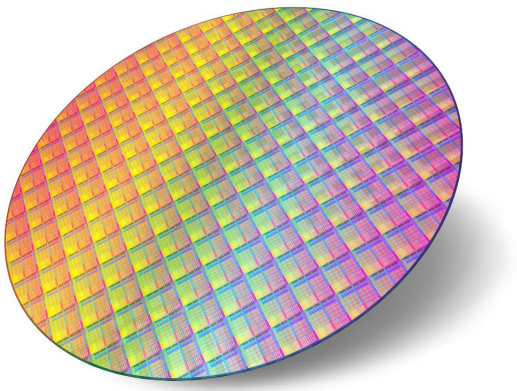


FIGURE 1.4 – Image d'une plaquette ou "*wafer*" (source: Flickr, Rob Bulmahn, http://www.flickr.com/photos/rbulmahn/ (CC License))

Afin d'illustrer la relation entre les différents termes cités précédemment, le tableau 1.1 (inspiré de celui présenté dans Mhiri (2016)) résume l'avancement d'un lot fictif dans sa route et illustre le lien entre les éléments Route/Opération/Step/Machine/Recette. On remarque notamment que la route est constituée de plusieurs opérations, constituées chacune de plusieurs steps. Chacun de ces steps est réalisé sur une machine, et il est possible de rencontrer plusieurs fois la même machine sur différentes opérations. On remarquera également que parfois plusieurs machines sont possibles pour réaliser une même recette, c'est par exemple le cas des machines A et C, toutes les deux capables de réaliser la recette W (étapes 1.1 et 2.1). Ce choix de la machine à attribuer à l'exécution de l'étape de fabrication n'est pas fait à l'avance et dépend du contexte de production.

TABLE 1.1 – Lot passant sur différents éléments de process de sa route (i.e. gamme opératoire)

Route	Opération	Step	Recette	Machine
Lancement du lot				
Route 1	Opération 1	Step 1.1	Recette W	Machine A
Route 1	Opération 1	Step 1.2	Recette X	Machine B
Route 1	Opération 1	Step 1.3	Recette Y	Machine C
Route 1	Opération 1	Step 1.4	Recette Z	Machine D
Route 1	Opération 2	Step 2.1	Recette W	Machine C
Route 1	Opération 2	Step 2.2	Recette V	Machine B
Route 1	Opération 3	Step 3.1	Recette U	Machine E
Route 1	Opération 3	Step 3.2	Recette Y	Machine C
Route 1	Opération 3	Step 3.3	Recette Z	Machine B
...	...	...	...	...
Route 1	Opération n	Step n.1	Recette T	Machine F
Route 1	Opération n	Step n.2	Recette Y	Machine C
Sortie du lot				

1.3 Planification de la production en fabrication de semi-conducteurs

La planification et la prise de décision ne concernent pas la production seulement mais couvrent le large spectre de la chaîne logistique, allant de la planification de la demande au management des stocks de produits, jusqu'à la coordination avec les systèmes de transport.

Concernant l'industrie des semi-conducteurs, un matrice très complète des différents problèmes de planification, disponible en Figure 1.5, a été récemment proposée par Mönch et al. (2018a). On remarque que les auteurs décomposent la chaîne logistique selon quatre grands groupes, à savoir l'approvisionnement, la production, la distribution et enfin la vente. Dans cette section, nous nous concentrerons seulement sur la partie production, en balayant les différents problèmes traités selon l'échelle de temps considérée.

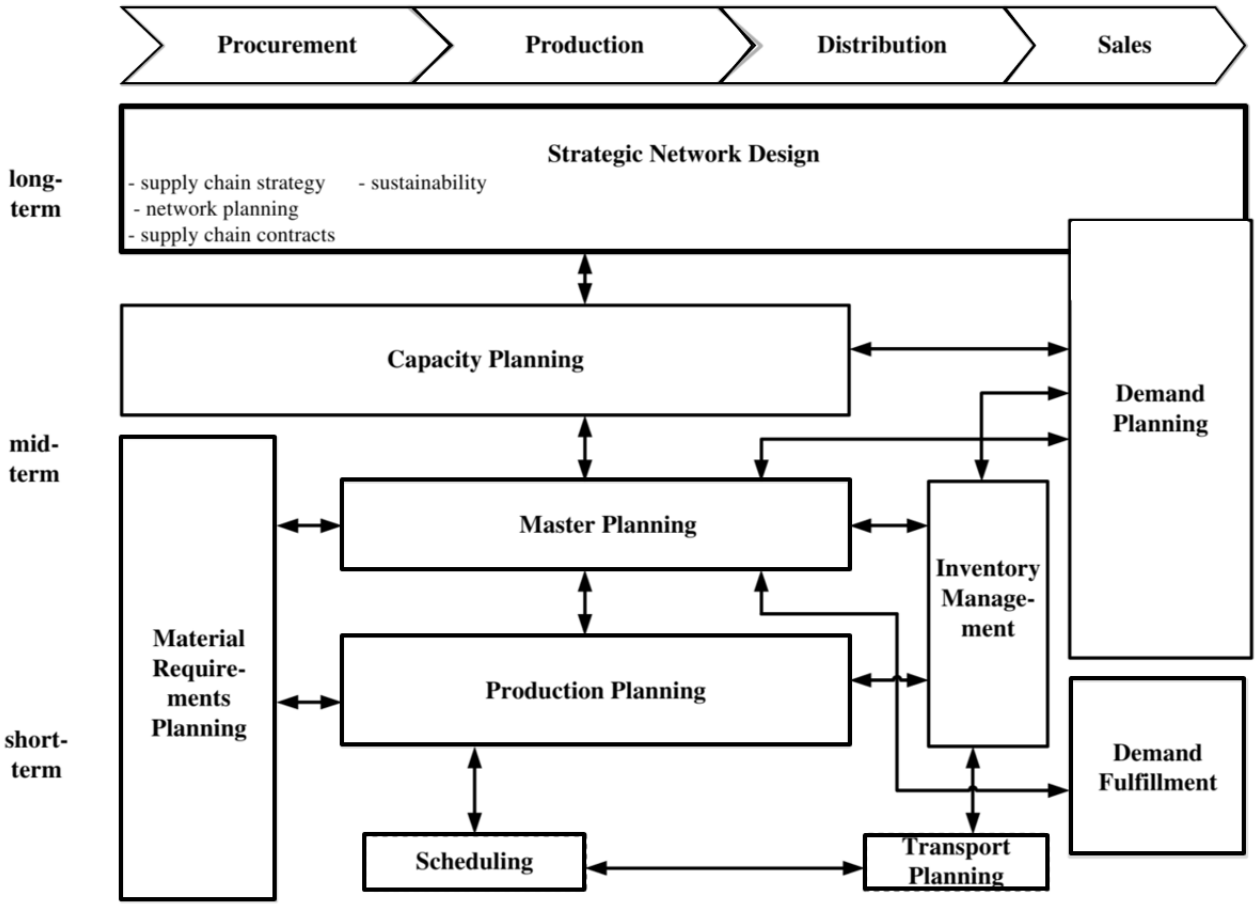


FIGURE 1.5 – Matrice de planification de la chaîne logistique dans l'industrie des semi-conducteurs (adaptée de Mönch et al. (2018a))

1.3.1 Plusieurs échelles de temps en planification de production

Généralement, la planification peut être décomposée selon trois grande échelles, à savoir les échelles *stratégique*, *tactique*, et *opérationnelle*. Cette décomposition était déjà proposée par Anthony (1965) et peut varier d'une industrie à une autre (Vollmann (2005)).

L'horizon temporel, le niveau de détail modélisé et la granularité des décisions à prendre varient selon les différents niveaux de décision. Les livres de Silver et al. (1998) ou Stadtler and Kilger (2015) donnent un aperçu général des niveaux de décision et de leur interrelation. Mönch et al. (2012) distinguent les niveaux de décision au niveau de l'entreprise (long terme), de l'usine (moyen terme) et du poste de travail (court terme). Plus récemment, Mönch et al. (2018a) décrivent les différents types de planification, toujours en fonction du trio d'échelles de temps: long, moyen et court terme. Dans ce qui suit, nous donnons un aperçu des décisions généralement prises à chacun de ces niveaux, dont un résumé est présenté dans le tableau 1.2.

TABLE 1.2 – Problèmes de planification de la production selon l'échelle considérée

Échelle de planification	Problèmes considérés	Horizon de temps
Stratégique	Recherche de nouveaux partenaires industriels Sélection des fournisseurs et sous-traitants Ouverture ou fermeture de sites de production Développement d'un nouveau produit Configuration de l'usine	3 à 5 ans
Tactique	Programme directeur de production, équilibrage de la charge des ressources	3 mois à 1 an
Opérationnelle	Ordonnancement, suivi des ateliers, règles de répartition	De quelques heures à plusieurs semaines

Planification stratégique

Le niveau stratégique comprend la planification à long terme pour un horizon temporel généralement de plusieurs années. À ce niveau, la planification prend généralement en compte la totalité de la chaîne logistique, de la prise en compte des fournisseurs jusqu'à la livraison de commandes aux clients. Ce problème général de planification est présenté sous le nom de *Strategic Network Design* dans Mönch et al. (2018a). Les décisions prises sont de haut niveau et concernent par exemple la construction d'une nouvelle usine, le choix de nouveaux fournisseurs ou clients, généralement dans le but de maximiser les profits. Il peut aussi être question d'anticiper les besoin des différents centres de production en termes de capacité en fonction de la prévision de la demande, amenant généralement à l'achat de nouvelles machines. On parle dans ce cas souvent de Planification de Capacité (*Capacity Planning*). Là encore, les plans sont déterminés sur des horizons de temps assez longs, allant de un à trois ans, du fait des délais important requis pour obtenir et installer de nouveaux équipements (Cakanyildirim and Roundy (1999)). L'horizon de planification est cependant plus court que pour le *Strategic Network Planning*, et c'est pour cette raison que l'on peut considérer la planification de la capacité comme étant à la frontière avec la planification tactique.

Planification tactique

Le niveau tactique comprend la planification à moyen terme pour un horizon temporel de quelques semaines à plusieurs trimestres. Dans ce niveau de planification de la production, on parle très souvent de *Master Planning* et de *Production Planning*. Ces deux problèmes de planification sont assez proches et diffèrent principalement par l'étendue du problème considéré et donc le niveau d'agrégation utilisé. De ce fait, la frontière entre ces deux problèmes de planification est souvent considérée comme assez floue (Fordyce et al. (2012).

Le *Master Planning* considère habituellement plusieurs usines de production, qu'elles soient de *front end* ou de *back end*, et l'horizon de planification peut aller d'un trimestre à un an. La modélisation du problème est assez agrégée, avec des flux physiques répartis selon les clients ou des familles de produit, la capacité étant modélisée à l'échelle des usines ou des ateliers de production qui les composent. Le but est généralement d'optimiser le profit en maximisant les livraisons clients et les principales décisions portent pour chaque usine sur les quantités de plaquettes à produire pour chaque produit pour chaque période de temps (habituellement de la taille d'une semaine ou d'un mois).

Le *Production Planning*, ou planification de production, peut à la fois faire référence à la catégorie générale de planification dédiée aux décisions de production (incluant le *Capacity Planning*, *Master Planning*, etc), mais aussi au problème spécifique de planification généralement situé entre le *Master Planning* et la planification opérationnelle. Afin de différencier ces deux notions, nous utiliserons dans la suite le terme *Production Planning* pour parler du problème de planification tactique, et le terme *planification de la production* pour faire référence à la catégorie générale de planification. Dans l'industrie des semi-conducteurs, le *Production Planning* considère le plus souvent une seule usine de fabrication et est utilisé sur des horizons de temps allant de quelques semaines à quelques mois. L'objectif principal du *Production Planning* est de définir les quantités de lots, pour chaque produit, à lancer en production durant chaque période de temps (le plus souvent une semaine), afin d'assurer un flux constant et uniforme permettant la livraison à temps des lots. Ces dates de livraison sont issues des objectifs de production donnés en amont par le *Master Planning*. La modélisation se veut encore plus précise que pour le *Master Planning*, mais reste encore agrégée (famille de produits, regroupement de machines) par rapport à des problèmes de planification à court terme.

Planification opérationnelle

La planification opérationnelle quant à elle est généralement associée aux problématiques d'ordonnancement. Ici, l'horizon de planification est généralement de quelques heures et les décisions à ce niveau sont prises de manière continue. Les systèmes à ce niveau doivent souvent réagir rapidement afin de prendre des décisions à très court terme. Les décisions d'ordonnancement portent généralement sur quelle machine doit passer tel lot (et avec quels autres lots si la machine fonctionne par *batch* par exemple) et dans quel ordre doivent être traités les différents lots affectés à une même machine ? Dans ce cas, les critères d'optimisation sont par exemple la minimisation du temps d'attente moyen des lots, du temps de cycle, le respect de fenêtres de temps ou la taille moyenne des *batchs*.

Relation entre les échelles de planification

La décomposition de l'ensemble de la planification de la production en sous-problèmes, tels que ceux décrits plus haut, donne des tâches de planification d'une complexité suffi-

samment réduite pour être résolus en des temps acceptables. Car il est difficilement envisageable d'utiliser un modèle de planification intégré, tenant compte des différentes échelles et des différents types de planification, qui fournirait des solutions satisfaisantes en des temps raisonnables. En effet, la complexité des modèles augmentant rapidement avec le nombre d'aspects/échelles de planification considérés en même temps, ceux-ci sont généralement soit très agrégés (donc peu précis), soit très lents (donc inadaptés à des problèmes de grande taille). Cette séparation dans la gestion de ces problèmes de planification nécessite des interfaces verticales et horizontales bien pensées entre les systèmes concernés. Car le risque est d'avoir des objectifs définis à un certain niveau qui ne sont pas réalisables au niveau inférieur. Ceci est principalement dû au fait que plus l'optimisation est à haut niveau, et plus les données sont agrégées pour des soucis de complexité. C'est très souvent cette agrégation qui est source d'erreur et de risques d'inconsistance entre les objectifs fournis et leur réalisabilité (Dauzère-Pérès and Lasserre (2002)). Une revue des différentes approches intégrées et séquentielles pour la planification de production est présentée dans le chapitre 2.

1.3.2 Facteurs complexifiant la planification de production

Nous venons de voir que la planification de production dans l'industrie des semi-conducteurs est généralement séparée en plusieurs blocs selon l'échelle de temps considérée, et que ces blocs sont souvent résolus de façon séquentielle. Mais pour quelle raison la planification de production est-elle si difficile dans l'industrie micro-électronique au point que la planification à chaque échelle prise individuellement est déjà un challenge ? Cette section a pour but de mettre en lumière les principaux facteurs de complexité inhérents à l'industrie de fabrication des semi-conducteurs.

— **Flux réentrants.** Nous avons vu dans la section 1.2.1 que la fabrication d'un circuit intégré se faisait par l'ajout successif de couches sur une plaquette de silicium. Cette multiplicité des couches implique de devoir répéter certaines opérations et donc de repasser plusieurs fois sur le même groupe de machines. Ainsi, l'atelier de photolithographie pourra voir repasser un même lot plusieurs dizaines de fois avant que ce dernier puisse être livré. La conséquence principale de ce type de flux est que l'on retrouve devant une même machine des lots (du même type de produit ou non) qui attendent mais qui n'en sont pas au même point sur leur route. Un lot peut être à quelques opérations de la sortie, alors que son voisin vient seulement d'entrer en production, ce qui complexifie par exemple des fonctions telles que l'affectation ou l'ordonnancement des lots sur les machines.

— **Taille des problèmes.** La planification de production dans une usine *front end* est un problème d'ordonnancement de type job shop complexe, du fait notamment des flux ré-entrants et de la présence de nombreuses machines en parallèle. Mais ce type de problèmes, bien que complexe, pourrait en théorie être résolu si la taille des instances étaient suffisamment petite. Malheureusement, les problèmes liés à la fabrication de semi-conducteurs amènent généralement à considérer des usines de plusieurs centaines de machines et plusieurs milliers de lots. Ces lots sont répartis parmi des centaines de produits différents, et chacun d'entre eux doit passer sur plusieurs centaines d'étapes de process, parfois plus d'un millier. Cette volumétrie fait du développement d'approches de planification de production, à la fois rapides et fournissant des solutions de qualité, un véritable défi.

— **Variété des processus de fabrication.** Un autre facteur de complexité vient du fait

que la nature et la durée des diverses opérations d'un procédé de fabrication de semi-conducteurs varient considérablement. Certaines opérations nécessitent 15 minutes ou moins pour traiter un groupe ou un lot de plaquettes de silicium, tandis que d'autres peuvent prendre plus de 12 heures. Il n'est pas rare qu'un tiers des opérations d'une usine soient des opérations par *batch*. Les machines de traitement par *batch* traitent simultanément plusieurs lots (généralement un à six), puis envoient les lots terminés à d'autres machines qui elles traitent un seul lot à la fois. Ce processus de fabrication par *batch* engendre souvent de longues files d'attente devant ces machines et donc un flux non linéaire de produits dans l'usine.

D'autres machines, telles que les implanteurs, nécessitent des temps de setup importants (changement de configuration des machines) lors du changement de séquences de produits. Si le flux de produit dans l'usine n'est pas contrôlé correctement, ces machines peuvent devenir des goulots d'étranglement dans la production.

— **Grande variété de produits.** Les usines de fabrication de semi-conducteurs peuvent généralement être réparties en deux catégories : *Low-Mix High-Volume* (LMHV) et *High-Mix Low-Volume* (HMLV). Dans le premier cas, l'usine produit une faible variété de produits en grande quantité, profitant ainsi d'une économie d'échelle mais aussi de flux de production moins variables, donc plus simples à piloter. Les entreprises concernées par ce cas sont souvent américaines et asiatiques, profitant d'un important marché permettant la production de gros volumes. A l'opposé de ces entreprises se trouvent celles, en particulier européennes, proposant une grande diversité de produits aux volumes variables, parfois très faibles. Ces entreprises se sont généralement adaptées à un marché européen très diversifié mais héritent alors d'une complexité accrue. On se retrouve avec plusieurs centaines de produits (et donc de routes) différents, partageant les mêmes *steps* élémentaires mais avec des séquences différentes. On se retrouve alors à devoir gérer des centaines de lignes de production, toutes en même temps, toutes en concurrence pour les mêmes ressources (Dequeant (2017)). D'où la complexité accrue des usines de type HMLV.

— **Qualification des recettes sur les machines.** La grande variété des produits, pouvant changer rapidement au fil des semaines, implique que d'une semaine à une autre les ressources nécessaires pour réaliser un volume de produits donné peut changer. Ainsi, si une machine pouvait être suffisante pour gérer le flux d'un certain produit à une étape donnée, il est possible que cette même machine devienne limitante la semaine d'après si le volume du produit en question augmente soudainement. Il est alors nécessaire de rendre une machine supplémentaire capable de traiter le produit pour l'étape en question, on parle de *qualifier la machine* pour une certaine *recette* (Johnzén et al. (2011)). Ce processus peut être long, mais est très fréquent (plusieurs occurrences par semaine) dans les usines HMLV et génère de la complexité pour la gestion des flux de production.

— **Multiplicité des indicateurs considérés.** En fonction de l'échelle considérée, différents indicateurs de performance peuvent être optimisés. Du point de vue de l'usine *front end*, on trouve une variété d'indicateurs. Dans une enquête sectorielle présentée par Pfund et al. (2006), les entreprises considéraient le débit global de l'usine comme l'objectif le plus important. Il est suivi du respect des livraisons clients dans les temps, puis du temps de cycle, des quantités de plaquettes lancées, du rendement des machines et enfin de la capacité de production. Piloter l'usine selon différents critères à optimiser, parfois contradictoires, est une difficulté supplémentaire couramment rencontrée

dans l'industrie des semi-conducteurs.

L'industrie des semi-conducteurs, et notamment la phase *front end*, est donc un système de fabrication parmi les plus complexes mais aussi les plus riches qu'il est possible de rencontrer. Cette difficulté est stimulante pour les chercheurs, et les travaux sur la planification dans ce secteur peuvent très souvent s'étendre par la suite vers d'autres industries avec des processus plus simples.

1.3.3 Outils de planification de la production dans l'industrie des semi-conducteurs

Dans cette section sont présentés les principaux outils et méthodes utilisés dans le management de la production au sens large, ainsi que leurs avantages et limites dans le contexte de la fabrication de semi-conducteurs. Le but est de dresser un portrait des méthodes que l'on peut rencontrer dans l'industrie des semi-conducteurs. Une étude plus large des approches d'optimisation pour la planification de la production est présentée dans le chapitre 2.

Parmi les méthodes industrielles classiques de planification de la production, l'une des plus connues et les plus utilisées est la méthode *Material Requirements Planning* (MRP) développée par Orlicki (1975) à partir de 1965. Le principe de cette méthode pour planifier les besoins de production est de partir de la demande finale et, connaissant les délais de production ou d'approvisionnement (le *Lead Time*), de déterminer à quelle date lancer la fabrication ou la commande des produits. Cette méthode a l'avantage d'être facilement implémentable via des feuilles Excel par exemple. Elle possède cependant certaines limites, dont la principale est qu'elle ne tient pas compte de la capacité de production des machines (Billington et al. (1983), Taal and Wortmann (1997)). Cette première version de la méthode MRP fournit donc un plan de production permettant de suivre le plan de livraison, mais ce plan de production n'est potentiellement pas faisable (ne respectant pas la capacité des machines) et donc impossible à suivre en pratique. Des travaux dans les années 70 vont permettre d'intégrer peu à peu la capacité, et la méthode va ensuite évoluer vers un outil plus global tenant compte notamment des aspects financiers, c'est le *Manufacturing Ressource Planning* (ou MRPII) développé par Wight dans les années 80 (Wight (1995)). Ces méthodes se sont largement répandues dans la plupart des secteurs de production, mais elles ont toujours peiné à s'intégrer dans l'industrie des semi-conducteurs, notamment du fait des flux ré-entrants, des fortes contraintes de capacité et de la grande variété de machines parallèles, amenant souvent à des solutions de faible qualité.

Le développement d'outils de planification a continué via notamment l'intégration toujours plus grande des différentes composantes de l'entreprise telles que la production, les finances ou bien les ressources humaines. Ces évolutions ont abouti à des outils de gestion plus complets tels que les *Enterprise Resource Planning* (ERP). Les ERP sont généralement des progiciels orientés transactions. Ils offrent des fonctionnalités liées à la finance, aux ressources humaines, à la fabrication et à la logistique, et enfin à la vente et à la distribution (Hopp and Spearman (2011)). Le module relatif à la fabrication et à la logistique offre généralement des fonctionnalités de type MRP et reste donc limité pour une utilisation dans le contexte des semi-conducteurs. De ce fait, la fonctionnalité de planification de la production des systèmes ERP n'est utilisée que dans une faible mesure dans les usines *front end*. La fonctionnalité de gestion des commandes offerte par les systèmes ERP est souvent la fonctionnalité la plus importante dans ce contexte, suivie par la fonctionnalité de planification de la demande. Toutefois, au niveau de l'ensemble de l'entreprise, les systèmes ERP sont souvent complétés par des APS (*Advanced Planning System*) (Mönch et al. (2012)).

Les APS sont des compléments aux ERP apportant une composante davantage axée sur l'optimisation et la décision, alors que les outils ERP ont quant à eux pour objectif principal de centraliser les informations issues des différentes composantes d'une entreprise. L'optimisation derrière les APS est assez poussée avec généralement l'utilisation de méthodes de Recherche Opérationnelle (RO) et/ou d'Intelligence Artificielle (IA) (Mönch et al. (2012)). Les APS sont généralement proposés sous forme de progiciels en modules complémentaires inclus dans les ERP. Ils offrent des fonctionnalités liées à la conception stratégique du réseau de fabrication, à la planification de la demande, à la planification du réseau d'approvisionnement, à l'approvisionnement externe, à la planification, à l'ordonnancement de la production, à la planification du transport et à l'ordonnancement des véhicules et à l'exécution des commandes. Ces différentes options sont plus ou moins utilisées dans l'industrie des semi-conducteurs. La planification de production stratégique/tactique (*Master Planning*) est généralement fournie par des APS (voir par exemple Kallrath and Maindl (2006)). Mais plus la planification de production se rapproche du court terme, ou tente simplement de rendre plus finement compte des flux de production et de la capacité, plus les APS laissent place à des méthodes développées localement. C'est notamment le cas pour la planification de production tactique/opérationnelle (Mönch et al. (2012)). Cette difficulté pour les APS à largement se répandre dans l'industrie des semi-conducteurs (du moins pour les fonctions de *Production Planning* et ceux à plus court terme), est que la complexité des systèmes implique souvent de développer des méthodes sur-mesure empêchant de créer des méthodes génériques aisément extensibles à d'autres types d'industries, voire tout simplement à d'autres usines. Les APS sont donc parfois utilisés dans le cadre de l'industrie des semi-conducteurs, mais présentent des faiblesse sur certains aspects et notamment autour des fonctions de *Production Planning*. En effet, des preuves empiriques des défaillances des APS dans la fabrication de semi-conducteurs peuvent être trouvées dans Lin et al. (2006). Il est démontré que les APS dans le contexte de fabrication des semi-conducteurs n'amenait pas à de meilleures performances par rapport à un utilisateur prenant des décisions via des méthodes "manuelles" avec l'ordinateur comme support d'information. Ces résultats restent toutefois à nuancer, car des avancées ont probablement été réalisées depuis plus d'une décennie.

En plus des méthodes MRP et des logiciels intégrés présentés plus haut, on trouve d'autres méthodes classiques de gestion de la production, surtout pour le moyen et court terme. Parmi ces techniques, on peut par exemple citer les méthodes du Juste à Temps (JIT pour *Just In Time* en anglais) (Golhar and Stamm (1991)) ou la théorie des contraintes (TOC pour *Theory Of Constraints* en anglais) (Goldratt (1990)).

Le principe de la méthode JIT est que la production est « tirée » par la demande et non par l'offre : il faut produire puis livrer (dans un temps très court) ce qui est demandé « instantanément » par le client. Cette production en flux tendu est donc opposée aux méthodes MRP qui sont quant à elles en flux poussé. Elle a l'avantage de permettre généralement de réduire le niveau de stock moyen (le WIP dans le cadre des semi-conducteurs), ou bien de réduire le temps de cycle moyen des produits. Cette approche demande cependant une bonne organisation afin de pouvoir rapidement répondre à la demande client. De plus, il a été montré que la méthode JIT éprouve des difficultés dans les systèmes à forte variabilité (Carlson and Yao (1992)), ce qui est souvent le cas dans l'industrie des semi-conducteurs.

La théorie des contraintes, quant à elle, a pour principe fondamental que le flux généré par une organisation est limité par au moins un processus, c'est-à-dire un goulot d'étranglement (*bottleneck* en anglais). La production de valeur ne peut donc être augmentée qu'en augmentant la capacité de production au niveau du goulot d'étranglement. On trouve des intégrations de cette méthode dans l'industrie des semi-conducteurs (Rippenhagen and Kri-

shnaswamy (1998)). Cependant son application est plus difficile et ses performances limitées dans des contextes à forte variabilité avec des changements réguliers des machines bottlenecks. De plus, il a était montré que pour des problèmes de panification à moyen terme, la méthode TOC pouvait fournir des résultats équivalents, voire moins bons que ceux obtenus via des méthodes de programmation linéaire en nombres entiers (PLNE) (Lee and Plenert (1993)).

1.4 Planification de production au sein du site de Crolles

1.4.1 STMicroelectronics et le site de fabrication de Crolles

STMicroelectronics (STM) est une société mondiale qui conçoit, développe, fabrique et commercialise des puces électroniques. Dans le monde, le groupe comptait en 2018 environ 46000 employés, 11 principaux sites de fabrication, des centres de Recherche & Développement avancés dans 10 pays et des bureaux de vente à travers le monde. Pour fournir à ses clients un outil de production indépendant, sécurisé et à un coût efficient, STM s'appuie sur un réseau mondial d'usines *front end* et *back end*. Ses principales usines de fabrication de plaquettes sont situées à Agrate Brianza et Catane (Italie), Crolles, Rousset et Tours (France) et à Singapour. Elles sont complétées par des sites d'assemblage et de tests implantés en Chine, en Malaisie, à Malte, au Maroc, aux Philippines et à Singapour. Créé au début des années 1990, le site de Crolles abrite deux usines *front end*: l'usine Crolles 200 et l'usine Crolles 300 inaugurée en 2003.

L'usine Crolles 200 fabrique des plaquettes de 200mm de diamètre. Elle est la plus ancienne des deux usines, sa construction ayant démarré en 1993. Cette usine est modérément automatisée, avec beaucoup de transports de lots effectués par des opérateurs, bien que certains soient effectués à l'aide de robots situés au sol (AGV). Bien que fournissant des technologies plus anciennes que l'usine Crolles 300, l'usine Crolles 200 reçoit encore un volume important de commandes de plus de 6000 plaquettes par semaine, se répartissant actuellement sur un catalogue de près de 1200 produits différents (dont en moyenne 300 produits actifs pour un instant donné de l'usine), la plaçant dans la catégorie des usines *High-Mix Low-Volume*.

L'usine Crolles 300 fabrique des plaquettes de 300mm de diamètre et a été inaugurée en 2003. Contrairement à l'usine Crolles 200, l'usine Crolles 300, plus moderne, est fortement automatisée avec une très grande majorité des transports et du stockage des lots effectués par un système de robots (AMHS) sous plafond. Depuis quelques années l'entreprise a gagné de nouveaux marchés et a connu une période de montée en volume de près de 40%. L'usine Crolles 300 est également dans la catégorie des usines *High-Mix Low-Volume*, avec des volumes de livraison hebdomadaires dépassant généralement les 5000 plaquettes, réparties parmi un catalogue de plus de 700 produits différents.

Les deux usines ont des différences entre elles, mais gardent les caractéristiques des usines de fabrication *front end*, à savoir les flux ré-entrants, les temps de cycle longs ou bien la grande hétérogénéité des machines.

1.4.2 La planification au sein du site

La planification de production sur le site de production de Crolles est décomposée hiérarchiquement selon plusieurs étapes. Ces dernières sont schématisées sur la figure 6.6.

1. L'usine prend comme principale entrée un plan de livraison, c'est à dire les quantités de chaque produit devant sortir de l'usine chaque semaine. Ces produits ne sont pas directement livrés aux clients, mais aux usines de *back end* en charge notamment de la découpe et de la mise sous boîtier des circuits intégrés. Ce plan de livraison de niveau tactique, aussi nommé plan directeur ou *Master Plan*, est fourni par le service central qui remplit donc la fonction de *Master Planning*.
2. De ce plan de livraison, le service Planning est tenu de fournir un plan de lancement des lots en production, c'est à dire un plan définissant pour chaque type de produit les quantités devant démarrer leur fabrication chaque semaine. Nous nous situons encore à un niveau plutôt tactique de la planification de la production, et c'est généralement cette phase qui prend le nom de *Production Planning*. La définition de ce plan est faîte en partie à l'aide d'un outil d'aide à la décision, tenant compte à la fois des quantités à livrer chaque semaine ainsi que des encours de production (le WIP).
3. L'étape suivante est une étape charnière entre la définition du plan de lancement des lots en production et la planification en temps réel sur les machines. Cette étape, que nous appelons *Operational Production Planning* et qui est à la frontière entre le niveau tactique et opérationnel, a pour but d'aider à manager les flux de production dans leur ensemble en aidant au pilotage des ateliers afin d'éviter un simple cumul d'optima locaux (pour chaque atelier de production). C'est autour de cette étape que s'articule cette thèse.

   Ce pilotage est réalisé à l'aide d'un outil d'aide à la décision tenant notamment compte des plans de livraison, de mise en production, des encours et des caractéristiques détaillées de l'usine (machines, produits, ...) et qui produit un plan de production détaillé (mais qui n'est pas un ordonnancement!). De ce plan de production, plusieurs actions sont mises en oeuvre, que ce soit la mise à jour de la priorité des lots, la création de consignes de productions pour les outils d'ordonnancement ou bien des indications pour les managers pour comprendre les principales lignes directrices à suivre pour piloter au mieux (selon une variété de critères) la production de l'usine. Nous verrons dans la section 1.4.3 que cet outil d'aide à la décision possède un certain nombre de limites sur lesquelles nous avons travaillé durant cette thèse.
4. La dernière étape est telle que décrite dans la section 1.3.1. Le but ici est d'optimiser localement, en temps réel ou sur un horizon de quelques heures, des critères tels que le temps d'attente moyen des lots ou bien le taux d'utilisation des machines. Pour cela, certains ateliers critiques seront aidés par des outils complexes d'ordonnancement et d'affectation des lots sur les machines, tandis que d'autres ateliers utiliseront des outils plus simples de sélection des lots (FIFO, règles classiques de dispatching, ...)

Ainsi, au travers de ces différentes étapes de planification, les consignes stratégiques et agrégées sont peu à peu transformées via des considérations de plus en plus précises et transmises au niveau inférieur jusqu'à aboutir sur des décisions en temps réel sur le choix des lots à lancer sur chaque machine.

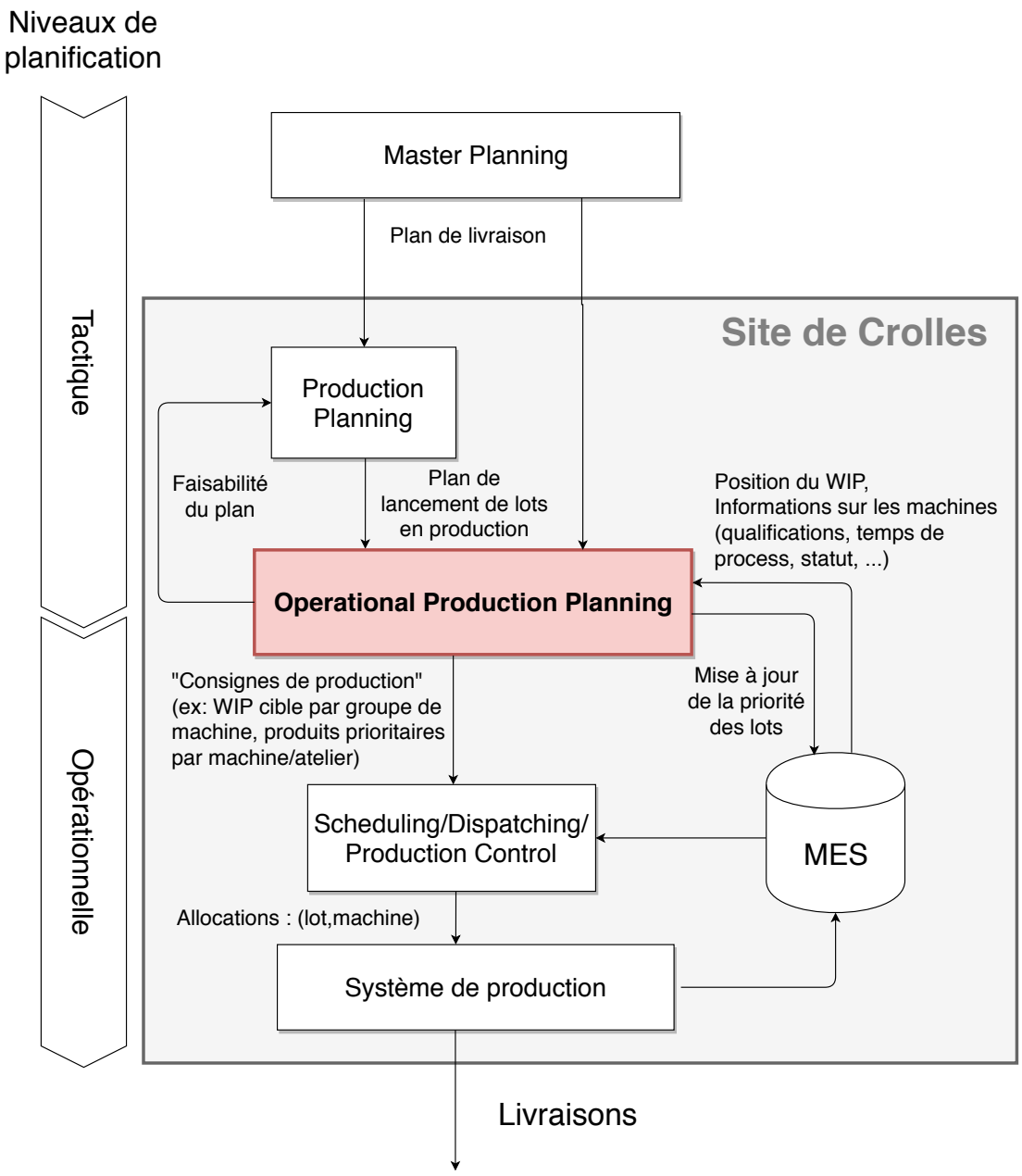


FIGURE 1.6 – Différentes étapes de planification de la production à ST Crolles

1.4.3 Problèmes en planification de production et besoin exprimé

Parmi les différentes étapes de planification de la production présentées dans la figure 6.6, celle liée à la jonction de la planification de production tactique et opérationnelle avait, au commencement de la thèse, un certain nombre des limites.

— Tout d'abord, l'outil d'aide à la fonction d'*Operational Production Planning* (frontière entre le niveau tactique et opérationnel) ne considérait la capacité des machines que de façon agrégée et avait donc des difficultés à proposer des plans de production réalisables par les outils de planification opérationnelle.

— De plus, cet outil dans son mécanisme est amené à faire des choix de répartition de la charge à réaliser sur des groupes de machines. Or, ces répartitions de la charge

semblaient parfois incohérentes, voire clairement sous-optimales pour les utilisateurs, et amenaient ces derniers à douter des consignes issues de l'outil.

— Les utilisateurs étaient également limités dans le choix des critères d'optimisation, l'outil ne cherchant en principe qu'à respecter les plans de livraison donnés en entrée.

— Enfin, les plans de production fournis étant très détaillés, il était parfois difficile pour les utilisateurs de saisir facilement et rapidement les informations principales nécessaires à la prise de décision.

Face à ces multiples limites, l'entreprise a désiré lancer un projet de thèse afin de développer un outil (à partir de celui existant) plus à même d'aider au pilotage de la planification de la production. Cet outil devait fournir à la fois des plans de production tenant mieux compte de la capacité de l'usine, mais aussi plus personalisables dans les critères d'optimisation choisis, et étant plus clairs pour les utilisateurs dans les informations fournies.

1.5 Conclusion

Dans ce chapitre, nous avons présenté le contexte de cette thèse, à savoir l'industrie de la micro-électronique dont le marché revêt aujourd'hui une importance considérable, et plus spécifiquement la fabrication de circuits intégrés. Nous avons montré les principaux aspects du processus de fabrication et notamment de la phase *front end*, chargée de la fabrication des différentes couches de semi-conducteurs nécessaires à la création des circuits.

Les usines en charge de la phase *front end* (communément appelées *wafer fabs*), requièrent des investissements importants du fait de l'environnement contrôlé (salle blanche) et du coût des machines de fabrication. Ajouté à cela une concurrence toujours plus forte, il en résulte un besoin accru de mieux planifier les différents aspects de la chaîne logistique en vue d'optimiser à la fois la rentabilité des investissement faits et la satisfaction des clients.

De plus, les usines ont des flux de production parmi les plus complexes, toutes industries confondues, rendant difficile l'utilisation des outils commerciaux classiques ERP ou APS. Ainsi, plus l'échelle de planification se raccourcit et les modélisations se font plus précises, plus les usines tendent à se doter d'outils développés en interne pour les aider à manager les flux de production.

La thèse a été effectuée en partenariat avec l'entreprise STMicroélectronics, sur le site de Crolles en France, constitué de deux usines. Ces deux usines diffèrent sur plusieurs aspects tels que le type de produits fabriqués ou le niveau d'automatisation, mais possèdent les caractéristiques complexes propres à la fabrication des semi-conducteurs. Cette complexité est encore accrue pour ces usines dont la grande diversité des produits proposés, fabriqués régulièrement en faible volume, les place dans la catégorie des usines *High-Mix Low-Volume* (HMLV).

Les usines de Crolles disposent de plusieurs outils d'aide à la planification de production, allant de la planification tactique du lancement des produits en production, jusqu'à l'ordonnancement de lots sur les machines. Parmi ces outils, celui de planification tactique/opérationnelle, en charge d'un pilotage global des flux de production au sein de chaque usine avait des limites. Ces dernières amenaient à des problèmes de faisabilité, de modularité et de compréhension des solutions, et le but de cette thèse fut de développer un nouvel outil d'aide à la planification tactique/opérationnelle plus performant, plus complet et plus facile d'utilisation.

Dans le prochain chapitre, nous effectuons une revue de la littérature autour des problématiques de planification de la production au sein de l'industrie des semi-conducteurs et positionnons cette thèse au regard des travaux existants.

Chapitre 2

Contexte scientifique

.

La planification est un vaste domaine ayant fait l'objet de nombreux travaux, se décomposant sous différents aspects selon la fonction (allant de la gestion de l'approvisionnement jusqu'à celle des ventes) et l'échelle de temps considérée. La planification de la production couvre un large éventail de décisions, pour lesquelles l'industrie des semi-conducteurs revêt certaines spécificités. Dans cette industrie, l'écart entre planification de production à moyen terme et planification de production à court terme rend difficile leur coordination. Une littérature limitée a traité ce problème frontière auquel ce travail de thèse est dédié.

	Modélisation des étapes	Modélisation des machines	Ensemble des ressources	Équilibrage avec qualifications	Modélisation des lots	Différents Objectifs	Instances réelles	Application industrielle
Fordyce *et al.* (1992)						✓	✓	
Leachman *et al.* (2002)	✓	✓		✓		✓	✓	
Lee et Lee (2003)	✓	✓		✓		✓		
Bureau *et al.* (2007)	✓	✓		✓		✓		
Lee *et al.* (2008)	✓	✓		✓		✓		
Chung et Jang (2009)	✓	✓				✓		
Bard *et al.* (2010)	✓	✓	✓				✓	
Barhebwa-Mushamuka *et al.* (2019)	✓	✓	✓			✓		
Horiguchi *et al.* (2001)		✓			✓			
Habenicht et Mönch (2002)		✓			✓			
Habla *et al.* (2007)		✓			✓			
Mhiri *et al.* (2018)	✓	✓	✓	✓	✓		✓	
Nos travaux	✓	✓	✓	✓	✓	✓	✓	✓

2.1 Introduction

Nous avons vu dans la chapitre 1 que la thèse s'inscrit dans le cadre de l'industrie des semi-conducteurs, et plus précisément sur l'étude d'un problème de planification de production à la frontière entre le tactique *Production Planning*, et le très opérationnel *Production Scheduling*. Cette frontière, que nous appellerons dans ce manuscrit *Operational Production Planning*, se distingue à la fois du *Production Planning* et du *Production Scheduling*, mais partage des similarités avec l'un et l'autre.

Dans ce chapitre, nous abordons dans un premier temps les grands travaux autour de la planification au sens large (tous secteurs confondus), puis nous présentons une brève histoire des méthodes développées afin de traiter les problèmes de planification de production, avant de nous concentrer sur ceux rencontrés en fabrication de semi-conducteurs. Nous rappellerons le cas particulier de notre problème, à la frontière entre deux échelles courantes de planification de production. Une revue de la littérature sur la planification de production moyen terme et sur les problématiques d'ordonnancement dans l'industrie des semi-conducteurs sera ensuite présentée. Enfin, nous positionnerons notre problème par rapport aux travaux les plus proches des nôtres.

2.2 La planification, c'est quoi?

Dans une chaîne logistique, des milliers de décisions doivent être prises chaque minute afin de garantir son bon fonctionnement. Ces décisions sont de types et d'importances variés. Elles vont de décisions très "simples" telles que le choix du prochain lot à traiter par une machine, au choix autrement plus important de la fermeture ou de la création d'une usine. Plus une décision est importante, plus elle a besoin d'être préparée. Cette préparation, c'est le rôle de la *Planification* (Stadtler and Kilger (2015)). Son but est d'identifier un ensemble d'alternatives possibles et de sélectionner parmi elles la (ou les) plus préférable(s).

Les chaînes logistiques étant généralement complexes, tous les détails de la réalité ne peuvent pas et ne doivent pas être considérés lors d'un processus de planification. Il est toujours nécessaire de s'abstraire en partie de la réalité et d'utiliser une copie simplifiée de celle-ci, le modèle, comme base pour établir un plan. Le but étant de représenter la réalité de façon aussi simple que possible mais aussi détaillée que nécessaire, c'est-à-dire sans ignorer les contraintes sérieuses du monde réel.

La planification est généralement classifiée selon trois niveaux, selon la durée de l'horizon sur lequel elle est faîte ainsi que le niveau d'abstraction considéré (Anthony (1965). Ces échelles sont la planification stratégique (long terme), tactique (moyen terme) et opérationnelle (court terme) et ont déjà été présentées dans la section 1.3.1 du chapitre 1, dans le cadre de l'industrie des semi-conducteurs.

Par ailleurs, la planification est généralement catégorisée selon quatre groupes, aux tâches sensiblement différentes mais inter-connectées : l'approvisionnement, la production, la distribution et les ventes. L'approvisionnement comprend tous les sous-processus chargés de gérer les ressources (matériel, personnel, etc.) nécessaires à la production. Planifier la capacité limitée des ressources est la principale considération de la partie production (pouvant elle-même être décomposée en plusieurs modules notamment selon l'échelle de temps considérée). La distribution fait le pont entre les sites de production et les clients (qu'ils soient finaux ou bien d'autres entreprises liées au processus de transformation du produit). Tous les processus logistiques ci-dessus sont guidés par les prévisions de la demande gérées par les

ventes.

La matrice de planification de la chaîne logistique (ou *Supply Chain Planning Matrix*, ou *SCP-Matrix*, voir Stadtler and Kilger (2015)), permet de décomposer ces différents modules de planification selon l'échelle de temps et le champ de la chaîne logistique considéré. En pratique aujourd'hui, ces différents modules de planification de la chaîne logistique sont gérés par des *Enterprise Ressouce Planning* (ERP), eux même souvent complétés par des Systèmes de Planification Avancée (APS pour *Advanced Planning Systems* en anglais). Il existe un grand nombre d'ERP et d'APS intégrés sous forme de logiciels commerciaux avec chacun leurs spécificités. Cependant, Meyr et al. (2015) ont pu synthétiser de façon générique les différents modules généralement utilisés pour répondre à chaque bloc de planification. Cette synthèse est présentée dans la figure 2.1.

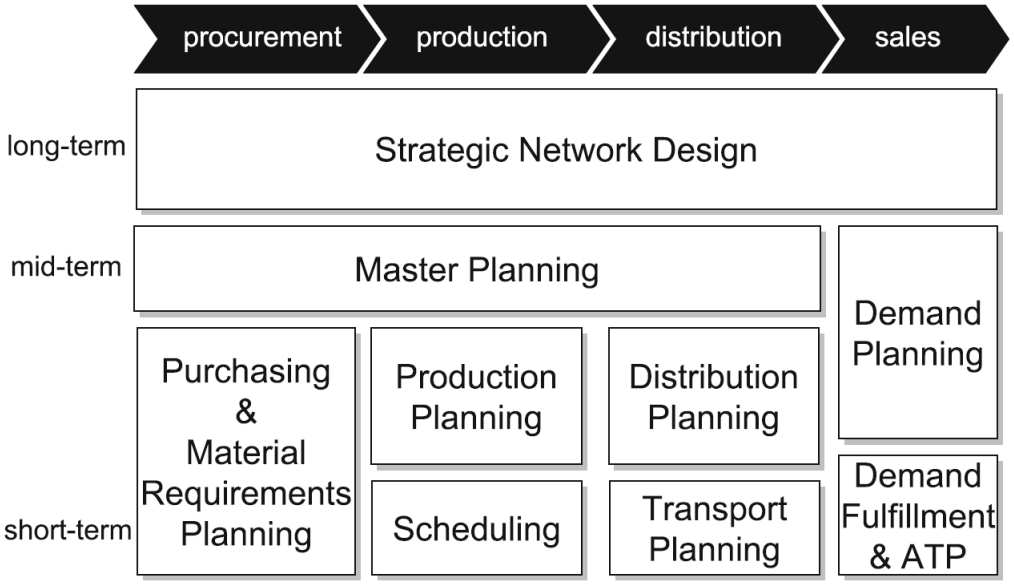


FIGURE 2.1 – Matrice de planification de la chaîne logistique (Meyr et al. (2015))

On remarquera que certains modules couvrent plusieurs champs de la chaîne logistique, tels que le *Strategic Network Design* ou le *Master Planning*. Ceci est logique étant donné que ceux-ci agissent sur des horizons de planification assez longs et sont responsables de coordonner différents aspects tels que l'approvisionnement, la production et la distribution des produits.

Cette matrice générique des différents modules de planification peut varier selon le secteur considéré, avec certains blocs pouvant fusionner alors que d'autres pourraient être encore décomposés en sous-modules plus spécifiques. Dans le cas de l'industrie des semi-conducteurs, cette décomposition de la chaîne logistique varie effectivement par rapport au cas générique, ce qui est illustré par la matrice de Mönch et al. (2018a) via la figure 1.5 présentée dans le chapitre 1.

Dans le cadre de cette thèse, nous nous intéressons au champ "Production" de la chaîne logistique. La section suivante présente un bref historique des principales méthodes développées au cours des dernières décennies afin de traiter ces problématiques de planification de la production.

2.3 Une brève histoire de la planification de la production

Les entreprises manufacturières ont toujours cherché des moyens d'améliorer leur compétitivité. Au cours de la première moitié du XXe siècle, l'efficacité interne de la fabrication en ateliers était largement suffisante pour assurer le succès des opérations. Cependant, avec l'intensification de la concurrence, les entreprises ont été obligées de trouver de nouvelles façons d'améliorer leurs opérations et de regarder au-delà des murs de l'usine. Aujourd'hui, les entreprises doivent être compétitives dans de nombreux domaines, tels que la qualité, la distribution et la flexibilité, et doivent planifier et contrôler leurs opérations en conséquence. Les tâches de planification et de contrôle sont devenues plus complexes, les délais de livraison sont plus courts, les catalogues de produits plus diversifiés, l'utilisation des ressources est plus optimisée, etc.

Au cours des 50 dernières années, de nombreux concepts et nouvelles approches ont été pensés et intégrés dans les entreprises manufacturières, certains avec plus de succès que d'autres. L'évolution des technologies de l'information et de la communication a facilité l'amélioration progressive des systèmes informatisés de planification et de contrôle des opérations. Ces évolutions peuvent être décomposées en 5 grandes périodes, s'étalant du début du XXe siècle jusqu'à nos jours (Olhager (2013)).

Les premières approches qui ont été élaborées au début des années 1900 analysaient de manière très détaillée les activités de fabrication. Tout d'abord Taylor (Taylor (1911)) a cherché à décomposer le processus de fabrication en sous parties, chacune analysable, amenant à la *Method-Time-Measurement* (MTM), une approche systématique de mesure des opérations (Maynard et al. (1948)). Certaines méthodes de planification et de contrôle des opérations industrielles, bien que basées sur des méthodes relativement simples, sont encore utilisées aujourd'hui. C'est par exemple le cas du *Economic Order Quantity* (EOQ), utilisé pour mieux gérer les décisions sur les stocks (quantités à acheter et périodicité) pour une gestion économique (Harris (1913)). On peut également citer le diagramme de Gantt pour afficher et planifier graphiquement les ordres de fabrication dans le temps et les ressources, prenant en compte la séquence de production des différents ordres (Gantt (1919)). Wilson (1934) a également introduit la notion de stock de sécurité et de seuil de commande automatique afin de limiter les risques de rupture de stock. Avec l'introduction des ordinateurs à la fin des années 1950 et au début des années 1960, il est devenu possible d'automatiser ces systèmes de commande (Orlicki (1975)). Des méthodes informatisées de contrôle en atelier ont été mises au point pour faciliter l'utilisation des méthodes d'ordonnancement (par exemple : Johnson (1954); Arrow et al. (1958); Maxwell and Miller (1967)).

Au cours des années 1970, le concept de *Material Requirements Planning* (MRP) a été introduit et fut largement accepté en peu de temps. L'introduction du MRP, où la production est déterminée en fonction de la demande et la connaissance de la nomenclature des produits (BOM pour *Bill Of Materials* en anglais), a alors grandement changé la logique de gestion de la production au sein de l'usine ainsi que de la planification des achats. Au milieu des années 1970, on estimait qu'il y avait environ 700 utilisateurs de systèmes MRP informatisés (Orlicki (1975)).

Ensuite, la décennie 1980/1990 a connu un changement de paradigme fondamental dans l'amélioration de l'efficacité des processus de production. Au début des années 80, des rapports sur l'approche juste-à-temps (JIT pour *Just In Time* en anglais) et sur la méthode de la théorie des contraintes (TOC pour *Theory Of Constraints* en anglais) font leur apparition. Plutôt que d'optimiser les opérations en tenant compte des contraintes actuelles, l'accent est mis sur l'amélioration des caractéristiques de base du système de production,

telles que l'identification et l'amélioration des ressources critiques (dites aussi limitantes ou goulots, *bottlenecks* en anglais). Le système de production érigé en modèle est celui de Toyota (Sugimori et al. (1977)) avec l'application de la méthode JIT et son système kanban qui montrèrent des réductions spectaculaires sur des indicateurs tels que le temps de cycle ou le niveau moyen d'en-cours. Dans la même période, Goldratt développe la TOC (Goldratt (1990)), dont l'approche vise à identifier les ressources goulots, donc limitant le flux de production, afin d'améliorer leur rendement dans le but d'augmenter la productivité globale du système. La méthode MRP a peu à peu évolué pour devenir un outil plus global tenant compte notamment des aspects financiers. Un nouveau terme est alors utilisé afin de définir cette fonction plus globale, le *Manufacturing Ressource Planning*. Afin de différencier les deux sigles, le terme MRP II a été introduit par Wight (voir par exemple : Wight (1995)) et désigne ce nouveau système englobant un plus large spectre de la production dont les aspect financiers et ressources humaines, mais aussi en intégrant mieux la capacité de production du système.

Le concept d'ERP (*Enterprise Ressource Planning*) a été introduit par Gartner Group en 1990 (Wylie (1990)). De nombreux systèmes MRP II ont été rebaptisés ERP dans les années 90 et ont bénéficié des améliorations de l'époque notamment concernant les technologies de l'information, leur permettant de remplir un éventail plus large de fonctions au sein de l'entreprise. Parmi ces nouvelles fonctions, on trouve notamment la planification à long terme de l'approvisionnement, de la capacité requise et de la production en fonction de la demande prévue, correspondant globalement au problème de *Master Planning* si l'on se réfère à la matrice en figure 2.1. Les données considérées sont agrégées (familles de produits) et l'horizon de planification est généralement de 15 à 18 mois avec des périodes de planification mensuelles. Cependant, la logique de la planification de production reste souvent basée sur celle du MRP amenant à une optimisation limitée des plans de productions fournis.

D'autre concepts ont vu le jour dans ces années 90, la plus importante étant peut-être l'approche *lean* (Krafcik (1988), Womack et al. (1990)), une approche de gestion de la production, dans la continuité du taylorisme, du système Toyota et de la méthode JIT, recherchant l'amélioration de la productivité et de la qualité, la réduction des délais et des coûts via une production "au plus juste" et l'élimination systématique des "gaspillages".

Dans les années 2000, l'ERP était devenu le modèle standard implanté dans grand nombre d'entreprises (Mabert et al. (2000), Olhager and Selldin (2003), Katerattanakul et al. (2006)), permettant une gestion assez complète de la plupart des aspects de l'entreprise notamment du fait de la centralisation des différents flux d'information. Ces systèmes finiront de s'étendre à des problématiques plus avancées de prise de décision par l'ajout d'*Advanced Planning Systems* (APS), proposant généralement une meilleure planification de la production via des outils de Recherche Opérationnelle et/ou d'Intelligence Artificielle (Mönch et al. (2012)).

Ces progiciels de planification ERP/APS se sont démocratisés au fil des années et sont aujourd'hui très répandus, notamment dans les grands groupes où la gestion et la coordination de l'ensemble des dimensions de l'entreprise s'avère complexe. Cependant, il reste certains secteurs où l'utilisation de ces progiciels reste difficile. C'est par exemple le cas de l'industrie des semi-conducteurs et notamment les blocs de planification de la production moyen et court termes (Mönch et al. (2012)), où la complexité des flux de production, principalement dans la partie *front end*, limite les performances de ces progiciels. Lin et al. (2006) ont notamment montré que dans le contexte de la fabrication des semi-conducteurs, l'utilisation d'APS n'amenait pas à de meilleures performances par rapport à un utilisateur prenant des décisions via des méthodes "manuelles" avec l'ordinateur comme support d'information. Ces résultats restent toutefois à nuancer, car des avancées ont probablement été réalisées depuis

plus d'une décennie.

2.4 La planification de production dans l'industrie des semi-conducteurs

Nous venons de voir dans la section précédente que la planification de production a fait l'objet de nombreux travaux depuis le début du XXème siècle, se concentrant d'abord sur l'optimisation locale de machines ou d'usines, jusqu'à de nos jours la gestion globale de plusieurs sites de production, voire de l'ensemble de la chaîne logistique en incluant clients et fournisseurs. Au fur et à mesure des décennies, des méthodes de planification et de contrôle de la production se sont popularisées, du MRP dans les années 60 jusqu'aux ERP et APS qui sont aujourd'hui intégrés dans de nombreuses entreprises et dans la plupart des secteurs industriels. Il est en revanche certains secteurs où ces outils de gestion globale restent limités, et l'industrie des semi-conducteurs en fait partie (Lin et al. (2006)). Plus particulièrement, les difficultés concernent la fabrication des circuits intégrés sur les plaquettes de silicium, c'est à dire les usines *front end*.

Les usines *front end* sont de type *job-shop*, c'est à dire que chaque produit suit une séquence d'opérations sur les machines, avec un ordre différent, à la différence des systèmes de production de type *flow-shop*, où chaque produit suit la même séquence de machines. D'autres part, les usines *front end* tombent dans les catégorie des *job-shop* flexibles, où plusieurs machines sont souvent qualifiées pour réaliser la même recette (Mönch et al. (2012)). Enfin, compte tenu des caractéristiques des systèmes de production rencontrés dans les usines *front end*, ces dernières sont souvent qualifiées de *job-shop* complexes, ou *complex job-shop* (Ovacik and Uzsoy (1997), Mason et al. (2002), Knopp et al. (2017)). Selon Mason et al. (2002), un *job-shop* est considéré comme complexe lorsqu'il possède les caractéristiques suivantes :

— Tâches avec des dates de disponibilité différentes,
— Temps de setup entre certaines tâches,
— Dates de livraison pour chaque lot,
— Flux ré-entrants,
— Types de process différents, tels que par *batch* ou non,
— Variabilité importante des paramètres du système (pannes machines, rendement, ...),

À cause de ces caractéristiques, ainsi que d'autres telles que la pluralité des objectifs à optimiser ou la taille importante des problèmes considérés, la tâche de planification de la production (allant de la planification du lancement de nouveaux lots dans l'usine jusqu'à l'ordonnancement des étapes de process sur les machines) est très complexe. Ces difficultés sont encore plus exacerbées dans le contexte des usines de type *High-Mix Low-Volume* (beaucoup de produits différents en faible quantité), expliquant que des solutions développées en interne sont souvent préférées aux modules de planification fournis par logiciels commerciaux de type ERP/APS.

Les premiers travaux traitant de la planification de production au sein de l'industrie des semi-conducteurs sont apparus au milieu des années 80. Ces travaux s'appuyaient surtout sur l'utilisation de la simulation afin d'étudier le comportement du système de production, et notamment d'indicateurs tels que le *throughput* (productivité) ou le temps de cycle moyen de l'usine, face à différents paramètres tels que le choix des règles de *Dispatching* des lots, l'ajout ou la perte de capacité, ou la politique de lancement des lots en production. Parmi ces

travaux, nous pouvons citer ceux de Dayhoff and Atherton (1984a), Dayhoff and Atherton (1984b), Burman et al. (1986), Atherton and Dayhoff (1986), Spence and Welter (1987), Wein (1988), Glassey and Resende (1988) ou Miller (1989).

Les problématiques de planification de production dans l'industrie des semi-conducteurs ont depuis connu un intérêt croissant de la part des chercheurs. Uzsoy et al. (1992) furent les premiers à présenter une description des caractéristiques de ces milieux de production, aujourd'hui devenue une référence lorsqu'il s'agit de souligner la complexité du secteur. Leur étude est aussi l'occasion de présenter une première revue des différents travaux proposés sur les problèmes de *Production Planning* et l'évaluation de performances. Deux ans plus tard, les même auteurs présentent une revue de la littérature (Uzsoy et al. (1994)) sur les systèmes de pilotage de la production (*shop-floor control*). L'année précédente, Johri (1993) a quant à lui proposé une revue des principaux problèmes de *Dispatching* et de *Production Scheduling* associés à la fabrication des circuits intégrés. Il faudra attendre ensuite 20 ans, avec Mönch et al. (2012), pour qu'une nouvelle revue de la littérature très complète des problématiques de planification de production en fabrication de semi-conducteurs soit proposée. Une nouvelle revue des différents problèmes d'ordonnancement et des travaux associés, ainsi que de futurs axes de recherche prometteurs, a été proposée par Mönch et al. (2011). Les revues de la littérature citées précédemment sont essentiellement concentrées sur la planification au sein d'une seule usine de fabrication. Ainsi, une revue des problématiques de planification (dont celles de production) au niveau de la chaîne logistique micro-électronique a été proposée via trois articles : Mönch et al. (2018a), Uzsoy et al. (2018) et Mönch et al. (2018b). À travers ces trois articles, les auteurs passent en revue les principaux travaux autour de nombreux problèmes allant de la planification de la demande à l'approvisionnement, en passant par la gestion de la capacité de production des unités de fabrication ou la gestion des stocks. Dans le cadre de cette thèse, nous nous intéressons spécifiquement à la partie liée à la planification de production. De plus, bien que notre travail s'inscrive dans la planification moyen/court terme, nous passerons succinctement en revue les différents modules rencontrés dans l'industrie des semi-conducteurs.

2.4.1 Les différents blocs de planification de production

Nous avons pu voir dans la section 2.2 que les différentes décisions de planification pour la chaîne logistique pouvaient être décomposées selon le type (approvisionnement, production, distribution et ventes) et l'échelle de temps considérés, sous forme d'une matrice telle que celle présentée en figure 2.1. Dans leur récente revue de la littérature, Mönch et al. (2018a) ont présenté une version modifiée de cette matrice adaptée au secteur de la micro-électronique, visible en figure 1.5 dans le chapitre 1.

Strategic Network Design. Une fois qu'une entreprise a déterminé quels produits fabriquer pour quels marchés au cours des prochaines années, elle doit concevoir/modifier sa chaîne logistique. Le *Strategic Network Design* tient compte de ces décisions à long terme et fournit l'environnement dans lequel devront s'inscrire les décisions relatives aux différents blocs de planification de production. Cet environnement doit notamment définir les sites de production à construire, détruire ou déplacer, les principaux équipements à acquérir ou bien la part de la demande devant être satisfaite, externalisée ou abandonnée. Cette tâche de planification stratégique est notamment complexe dans l'environnement des semi-conducteurs pour plusieurs raisons (Karabuk and Wu (2003)). La première est le coût extrêmement élevé des équipements (plus de $100 millions aujourd'hui pour certaines machines de photolithographie) et le fait que la construction d'un nouvel environnement de fabrication est très

long (généralement au moins 15 mois). Il faut aussi beaucoup de temps, souvent plus d'un an, pour obtenir, installer et qualifier une machine une fois que celle-ci a été commandée. Comme c'est le cas pour tout autre investissement de grande ampleur, il est nécessaire de tenir compte d'un certain nombre de paramètres contextuels, tels que la conjoncture économique mondiale, la situation des régions où des installations pourraient être construites et la disponibilité d'un vivier de main-d'œuvre et d'ingénieurs compétents pour le bon fonctionnement d'un site.

Capacity Planning. La planification de la capacité consiste à estimer le nombre d'équipements nécessaires pour répondre à une demande donnée ou, de façon symétrique, la quantité de produits qu'un ensemble donné d'équipements peut fabriquer tout en maintenant un rendement acceptable (par exemple un temps de cycle compétitif). Dans la fabrication de semi-conducteurs, il s'agit généralement d'une décision à moyen/long terme sur un horizon d'un à trois ans en raison du long délai d'obtention de nouveaux équipements (Cakanyildirim and Roundy (1999)). L'approche la plus courante consiste à utiliser un calcul déterministe en accordant à chaque groupe d'équipements un taux de production moyen, en tenant compte d'une marge "d'inefficacité" liée à des éléments tels que la maintenance préventive, les temps de setup ou les périodes où les machines sont inutilisées (i.e. disponibles mais sans aucune plaquette disponible pour processer). Pour une revue des enjeux et travaux sur la planification stratégique de la capacité dans l'industrie des semi-conducteurs, les lecteurs peuvent se référer à Wu et al. (2005), Geng and Jiang (2009) ou Uzsoy et al. (2018).

Master Planning. Que ce soit dans les travaux de Meyr et al. (2015) ou Mönch et al. (2018b), le *Master Planning* se place sur une échelle à moyen terme au niveau de l'entreprise (donc multi-sites). L'objectif de cette planification et de définir les quantités à produire, pour chaque période (généralement mois ou semaine) et pour chaque site, ainsi que les niveaux de stocks. Les données sont le plus souvent agrégées sous forme de groupes de machines potentiellement limitantes et de familles de produits. Le *Master Planning* doit pouvoir déterminer comment la chaîne d'approvisionnement réagira aux fluctuations saisonnières de la demande (par exemple en accumulant des stocks avant les pics de demande). Elle doit aussi définir comment la capacité de production doit être répartie entre les différentes familles de produits et les marchés concurrents. Par conséquent, pour être efficace, le *Master Planning* doit considérer un horizon de planification englobant au moins un cycle de demande saisonnier complet, ce qui en fait une activité à moyen terme. Au final, le *Master Planning* prend en entrée les prévisions de demande, et les transforme en objectifs de production pour chaque site. Ces objectifs sont le point d'entrée du module de *Production Planning*.

Production Planning. À partir d'objectifs de production fournis (par semaine ou par mois) par le *Master Planning*, le principal rôle du *Production Planning* est de définir un plan de lancement (*release*) en production de nouveaux lots, pour chaque produit et pour chaque période (généralement la semaine). Les décisions étant à un niveau tactique inférieur, l'horizon de planification est généralement de plusieurs mois, car il faut tenir compte du temps de cycle des lots (1 à 3 mois) afin de lier l'entrée des lots dans l'usine avec le planning de livraison. Les données considérées sont plus détaillées que pour le *Master Planning*, notamment la capacité des machines est plus détaillée et la planification se concentre généralement sur une seule usine. Le but est de fournir un plan de lancement des lots tenant compte, de manière suffisamment détaillée, de la capacité de l'usine afin de respecter les objectifs de production, sans dégrader des indicateurs tels que le niveau de stock ou le temps de cycle moyen des produits.

Production Scheduling. Une fois qu'ont été définis 1) un plan de lancement des lots, spécifiant les quantités devant commencer leur fabrication chaque semaine dans l'usine et 2)

un plan de livraison, spécifiant ce qui doit en sortir, il est nécessaire de piloter l'évolution des flux dans l'usine afin de lier ces deux plans. Ce pilotage du flux de production, c'est le rôle de l'ordonnancement de production (ou *Production Scheduling* en anglais). Son but est globalement de définir la machine sur laquelle chaque lot à chaque étape est affecté. Le *Scehduling* doit également définir l'ordre de passage des lots sur la machine (la séquence) ainsi que l'instant de démarrage prévu (l'ordonnancement.) Du fait du statut complexe des flux du système de production, et notamment dans les usines *front end* (*complex job shop*, Mason et al. (2002)), cet ordonnancement n'est pas fait sur l'ensemble de l'usine mais par atelier, ou un sous-ensemble d'un atelier lorsque les flux sont particulièrement complexes.

2.4.2 À la frontière entre planification et ordonnancement

Nous venons de voir dans la section 2.4.1 les différents blocs de la planification de production rencontrés dans l'industrie des semi-conducteurs et leurs liens. Nous avons conclu cette section en présentant les modules de *Production Planning* et de *Production Scheduling*, l'un définissant les lots à lancer en production chaque semaine, l'autre définissant l'affectation et l'ordonnancement des lots sur les machines afin de respecter les délais de livraison fixés par le *Master Planning*. Or, si l'on se place dans le cadre de la fabrication des circuits intégrés (partie *front end*), un lot comprend en moyenne 700 étapes dans sa gamme de fabrication (route). La durée opératoire d'une étape varie d'une dizaine de minutes à plus de 12 heures. Considérant les milliers de lots présents au même instant dans l'usine, cela donne des millions de décisions à prendre par la fonction de *Production Scheduling* afin de piloter ce flux de production, avec pour seules principales consignes, les dates d'entrée et de sortie des lots. Il y a donc un écart important dans la prise de décision entre ces deux modules de planification que sont le *Production Planning* et le *Production Scheduling*. Ce dernier ne peut en effet pas optimiser la gestion de l'ensemble des étapes des lots dans l'usine sur la seule base des entrées et des sorties des lots de cette même usine. Par conséquent, les outils d'ordonnancement ne considèrent généralement que des sous-ensembles de machines et n'optimisent que des critères locaux tels que le temps de cycle moyen des lots passant dans un atelier, le taux d'utilisation des machines, ou leur productivité (*throughput*).

Il est donc essentiel de se doter d'une fonction supplémentaire, telle que proposée par Govind et al. (2008), dont le but est d'assurer le traitement des produits au bon moment dans chaque atelier afin de respecter les objectifs de production définis par le *Master Planning*. Cette prévision des flux de production permet aussi par exemple de planifier des maintenances préventives ou au contraire de s'assurer du maintien du fonctionnement d'une machine pendant une période de temps donnée.

C'est afin de combler cet écart qu'un outil d'aide à la décision a été développé au sein du site de Crolles de l'entreprise STMicroelectronics. Le but de cet outil est d'amener à un meilleur pilotage global des flux de production de l'usine en prenant en entrée les plans de lancement et de livraison des lots, et fournissant en sortie des objectifs plus précis aux outils d'ordonnancement. C'est ce problème "frontière" que nous avons étudié durant cette thèse. Par conséquent, dans la prochaine section, nous nous intéresserons à ces deux modules que sont le *Production Planning* dans la section 2.5.1 et le *Production Scheduling* dans la section 2.5.2. Nous verrons ensuite dans la section 2.5.3, comment certains travaux ont essayé d'intégrer ces deux niveaux de planification. Cette section se termine par un recensement des approches considérées à la frontière entre le *Production Planning* et le *Production Scheduling*, que nous appellerons par la suite *Operational Production Planning*.

2.5 La planification de production et l'ordonnancement dans l'industrie micro-électronique

2.5.1 Moyen Terme: la planification de production

Dans cette section nous traitons des problématiques de *Production Planning*, c'est-à-dire la planification tactique, dont le but est de déterminer les plans de lancement des lots en production à partir des plans de livraison fournis par le *Master Planning*. La littérature étudiant les problèmes de *Production Planning* est significative, avec une grande variété d'approches allant du développement de programmes linéaires à l'utilisation de méthodes d'analyse basées sur les files d'attente. Les principales revues de la littérature considérant (uniquement ou en partie) les problèmes de *Production Planning* dans l'industrie des semi-conducteurs sont celles d'Uzsoy et al. (1992), de Mönch et al. (2012) et Mönch et al. (2018b). Dans la thèse de Mhiri (2016), une présentation des différentes approches de la littérature pour ce problème est proposée, classées selon les approches utilisées telles que les méthodes de programmation linéaire ou de programmation linéaire en nombres entiers, les (méta-)heuristiques ou les méthodes basées sur l'utilisation de modèles de simulation. Dans leur récente revue de la littérature, Mönch et al. (2018b) dédient une section aux problèmes de *Production Planning*, tout en rappelant que la frontière avec le *Master Planning* reste floue et peut parfois varier selon les auteurs.

L'un des principaux challenges du *Production Planning* est de faire le pont entre le plan de livraison établi par le *Master Planning*, et le plan de lancement des lots en production qui est à définir. L'un des principaux paramètres à considérer est donc le temps de cycle des lots, c'est à dire le temps passé entre l'entrée d'un lot et sa sortie de l'usine. Ce paramètre est fortement corrélé à la notion de capacité du système de production. Dans leur revue, Mönch et al. (2018b) décomposent les différents travaux liés à la planification de production en fabrication des semi-conducteurs selon la façon dont est considérée ce temps de cycle, aussi appelé *lead time* (délai d'obtention). Dans la plupart des travaux, le *lead time* est traité comme une donnée d'entrée fixe et extérieure au problème, c'est à dire considérée comme une donnée exogène. Cette hypothèse permet de simplifier les modèles mais ne tient pas compte de la relation entre le *lead time* et d'autres paramètres tels que la capacité de l'usine ou son taux de charge (i.e. l'en-cours moyen dans l'usine). Certains travaux ont cherché à tenir compte de cet aspect variable du *lead time*, à travers différentes méthodes.

Ainsi, dans la suite de cette section, nous suivons la structure proposée dans Mönch et al. (2018b), en mentionnant les différents travaux liés à la planification de production dans l'industrie des semi-conducteurs, selon leur façon de traiter le *lead time*. Cette section est donc un enrichissement de Mönch et al. (2018b), intégrant les travaux plus récents. Dans un souci de concision, notre revue de la littérature se veut en partie plus synthétique que celle de Mönch et al. (2018b), et nous invitons le lecteur à se référer à cette dernière s'il souhaite avoir davantage de détails. Nous abordons d'abord les travaux considérant le *lead time* comme une valeur exogène. Puis nous nous intéresserons aux travaux ayant tenté de lier le *lead time* à d'autres paramètres du problème, notamment au travers de deux approches que sont les méthodes itératives à base de simulation, et l'utilisation de *clearing function*.

Approches avec *lead time* fixe

L'approche la plus couramment rencontrée dans la littérature est de considérer le *lead time* comme une valeur fixe, le plus souvent calculée en amont de la planification via l'uti-

lisation d'historiques ou de modèles de simulation. L'un des premiers travaux concernant la planification de la production dans l'industrie des semi-conducteurs est celui de Smith (1965) qui proposa un Programme Linéaire (PL) pour la compagnie Raytheon, et dans lequel le *lead time* était considéré comme nul. Hormis le cas particulier de ne pas tenir compte du *lead time*, le modèle le plus simple est de considérer celui-ci comme un multiple de la taille des périodes utilisées pour la planification. Par exemple, si la planification est faite sur 2 mois avec des périodes d'une semaine, le *lead time* d'un produit sera défini en nombre de semaines. On parle alors de *lead time* entier, ou *integer lead time*. Parmi les travaux ayant eu recours à cette approche, nous pouvons par exemple citer ceux basés sur des méthodes de PL ou de Programmation Linéaire en Nombre Entiers (PLNE) de Billington et al. (1983), Johnson et al. (1974), Pochet and Wolsey (2006) ou Voß and Woodruff (2006). Une autre approche consiste à permettre au *lead time* de prendre des valeurs autres qu'un multiple de la taille des périodes, on parle alors de *non integer lead time*. Bien que l'utilisation de tels *lead time* soit d'aspect moins simple à intégrer dans un PL, Hackman and Leachman (1989) et Hackman (2007) ont montré que cette approche est de même complexité que celle utilisant des *lead time* entiers.

Dans les usines *front end*, il y a presque toujours plusieurs machines alternatives capables d'effectuer une étape de fabrication. Étant donné que les durées opératoires peuvent varier considérablement d'une machine à une autre, un modèle de planification nécessite une représentation précise de la capacité disponible et de la charge à allouer aux machines disponibles. Cependant, cette affectation spécifique d'étapes de process à des ressources relève plutôt du domaine de l'ordonnancement à court terme, tenant mieux compte du détail et de la situation courante de l'usine. Johri (1994) a étudié le problème de l'affectation des opérations individuelles à des sous-ensembles spécifiques de machines. Toktay and Uzsoy (1994), Akçalı et al. (2005) et Ignizio (2009) ont traité des problèmes similaires.

Leachman and Carmon (1992) présentent un programme linéaire qui établit des allocations spécifiques pour chaque période de planification et s'en servent pour élaborer une formulation globale qui assure des charges de capacité réalisables sans utiliser de variables d'allocation spécifiques en supposant que les durées opératoires sur les machines diffèrent d'un ratio fixe pour toutes les étapes du processus partagé. Pour ce faire, ils identifient les ensembles d'opérations et les machines alternatives dont les contraintes de capacité sont susceptibles d'être bloquantes dans la solution optimale (ensembles opération-machine dominantes) et écrivent les contraintes de capacité uniquement pour ces ensembles. Bermon et al. (1995), Hung and Cheng (2002) et Liberopoulos (2002) proposent des approches de modélisation similaires considérant des ateliers composés de machines non identiques.

Un problème intéressant et qui semble avoir fait l'objet de peu de recherches jusqu'à présent (Mönch et al. (2018b), est de déterminer non pas le meilleur plan en fonction d'un *lead time* pré-défini, mais plutôt quel *lead time* permettrait d'obtenir les meilleurs résultats? Dans leur article, Milne et al. (2015) proposent un PLNE pour déterminer les meilleures valeurs de *lead time* à intégrer dans un MRP afin de minimiser certains coûts et pénalités. Des tests menés à partir de données fournies par un fabricant de DRAM ont montré la pertinence de cette approche, améliorant les performances de la planification faite par l'outil MRP. Albey and Uzsoy (2015) abordent cette question à l'aide d'un modèle réduit d'une usine de fabrication *front end*, en utilisant simulation et optimisation afin d'estimer le *lead time* donnant les meilleures performances. Kriett et al. (2017) proposent un PL pour la planification de la production dans une usine également *front end*. L'objectif du PL est de fournir d'une part les quantités à lancer en production, mais aussi pour chaque lot un *lead time* objectif qui va permettre de définir un ensemble d'ODD (operation due dates),

c'est à dire des dates avant lesquelles certains couples lot/opération doivent être réalisés. Récemment, Beraudy et al. (2018) ont proposé un PL intégrant (en plus de la minimisation classique des coûts de production, de stockage et de retard) un terme financier visant à maximiser les profits. Pour cela, ils s'inspirent du critère financier *Net Present Value* (NPV), servant à évaluer le retour sur investissement, considérant que les bénéfices d'aujourd'hui valent davantage que les mêmes bénéfices demain. Les résultats montrent que ce nouveau modèle assure une meilleure productivité de l'usine, comparé aux modèles ne considérant que les coûts, mais que ce nouveau modèle peut également amener à des cas de surproduction, notamment vers la fin de l'horizon de planification.

Bien que les *lead time* exogènes, indépendants de la charge de travail, aient été l'approche prédominante de la modélisation des *lead time* dans la planification de la production, cette approche présente des inconvénients. Que ce soit via des modèles de simulation (Ankenman et al. (2011)), la théorie des files d'attente (Curry and Feldman (2010)), ou bien simplement l'expérience industrielle, tous suggèrent que la durée moyenne du *lead time* dans un site de production augmente non linéairement par rapport au taux moyen de remplissage de l'usine, ce dernier étant lui-même dépendant des plans de lancement des lots déterminés par la fonction de *Production Planning*. Notamment, on constate qu'à des niveaux élevés d'utilisation des machines, de petites fluctuations de la charge de travail peuvent entraîner des changements importants du *lead time*. Dans ces conditions, l'utilisation de délais exogènes indépendants de la charge de travail dans les modèles de planification peut conduire à des performances médiocres. La façon de modéliser cette dépendance entre le *lead time* et le niveau de charge a fait l'objet de nombreuses recherches (Pahl et al. (2005b), Pahl et al. (2005a), Pahl et al. (2007)). Mönch et al. (2018b) décomposent les différents travaux associés à cette modélisation selon deux grandes approches. La première, qu'ils nomment *méthodes multi-modèles*, décompose le problème de planification de la production en deux sous-problèmes. Un sous-problème, habituellement mis en œuvre à l'aide d'un PL, détermine un plan optimal de lancement des lots basé sur des valeurs fixées de *lead time*, tandis que l'autre modèle estime l'influence sur ce même *lead time*, du plan de lancement des lots en production, et donc de la charge prévisionnelle. La deuxième approche, quant à elle, a recours à des *clearing function*, qui sont des fonctions mettant en relation le *throughput* moyen d'un équipement (ou d'un groupe d'équipements, voire de l'usine complète), avec son niveau de charge. Les deux paragraphes suivants traitent de ces deux approches dans le contexte de l'industrie des semi-conducteurs.

Approches multi-modèles

La première des approches multi-modèles largement discutée a été celle de Hung and Leachman (1996). Les auteurs proposent un algorithme itératif qui estime les *lead time* résultant d'un premier plan de lancement évalué avec un modèle de simulation du système de production afin de les transmettre à un PL. Ce dernier calcule ensuite un nouveau plan de lancement basé sur ces estimations de *lead time*, et le modèle de simulation est de nouveau exécuté pour estimer les nouveaux *lead time*. Le processus iétarif se poursuit jusqu'à ce qu'un critère de convergence soit satisfait. Un schéma d'une procédure classique de ce type d'approche itérative est présentée en figure 2.2.

Des approches similaires sont proposées par Kim and Kim (2001), Byrne and Bakir (1999), Byrne and Hossain (2005), Kim et al. (2014), Albey et al. (2014) et Bang and Kim (2010). Hung and Hou (2001) examinent l'utilisation de modèles de file d'attente et de régression comme substitut au modèle de simulation afin de réduire la charge de calcul. Manda and

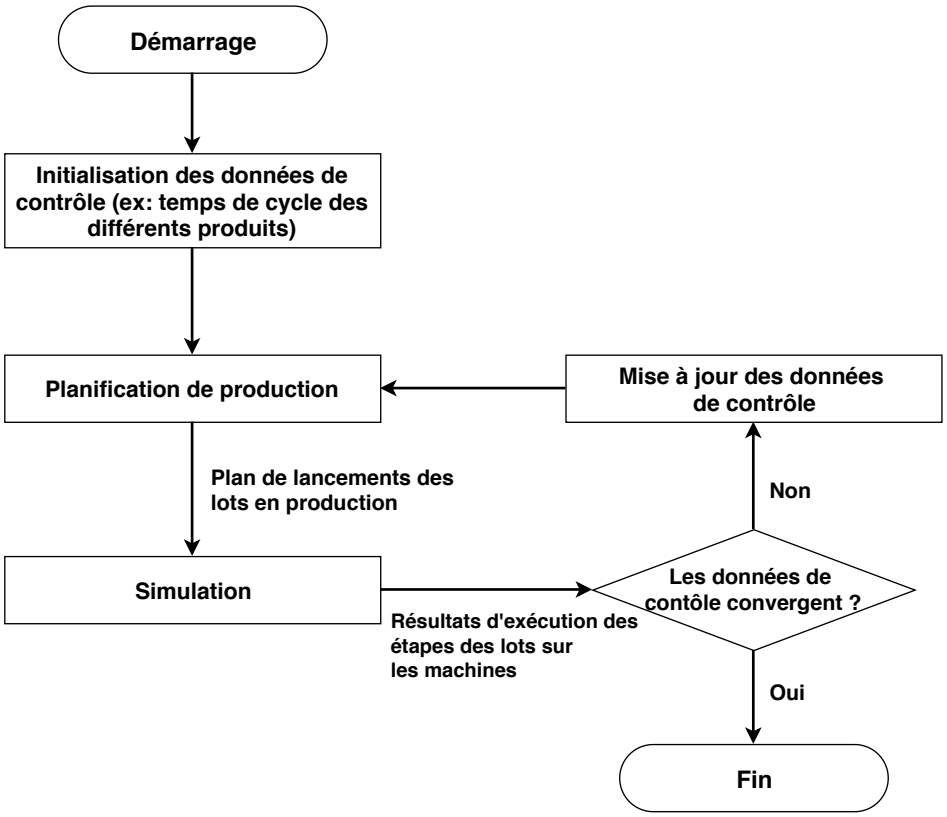


FIGURE 2.2 – Procédure classique d'une méthode itérative multi-modèle de planification de la production (adaptée de Kim and Lee (2016))

Uzsoy (2018) développent un modèle de simulation couplé à un modèle d'apprentissage afin de déterminer les meilleures politiques de lancement de lots en production dans le cadre d'introduction de nouveaux produits parmi d'autres produits plus matures. Irdem et al. (2010) étudient les procédures de Hung and Leachman (1996) et Kim and Kim (2001) et constatent que le comportement de convergence de ces méthodes diffère considérablement. Une récente revue des méthodes de simulation-optimisation appliquées à l'industrie des semi-conducteurs est proposée par Ghasemi et al. (2018).

Ces approches sont intéressantes par leur capacité à allier l'optimisation (via un PL ou un PLNE) à une modélisation réaliste du système de production permettant d'en déduire assez fidèlement des paramètres tels que le *lead time*. Elles souffrent toutefois de plusieurs limites. La première est que les approches multi-modèles itératives combinent deux méthodologies : programmation linéaire et simulation, qui sont bien comprises par de nombreux praticiens et pour lesquelles d'excellents outils et logiciels commerciaux sont disponibles. En revanche, leur convergence n'est pas bien comprise. Certaines méthodes, comme celle de Kim and Kim (2001), semblent converger assez régulièrement, mais Albey et al. (2014) constatent que la solution finale dépend fortement de la solution initiale. La deuxième limite est la charge de calcul élevée des modèles de simulation détaillés. Les efforts visant à remplacer le modèle de simulation par des modèles de file d'attente en régime permanent ont pour le moment des résultats mitigés, donnant lieu à des prévisions très imprécises du *lead time* (Hung and Hou (2001)). Le remplacement du modèle de simulation par un méta-modèle efficace comme dans Li et al. (2016) pourrait constituer une réponse à cette difficulté.

Approches à base de *clearing function* (CF)

Les approches de planification de la production décrites ci-dessus supposent des *lead time* indépendants de la charge du système de production, ou bien nécessitent des simulations relativement exigeantes sur le plan informatique afin d'estimer les *lead time* découlant des plans de lancement de lots proposés par les modèles de PL.

La troisième et dernière grande catégorie d'approches est basée sur les *clearing function* (CF). Ces fonctions ont pour but de mettre en relation le throughput (TP), c'est à dire le nombre de produits sortant d'un système (une machine, un groupe de machines, ou une usine complète) pendant un temps donné, avec une estimation de son taux de charge (Missbauer and Uzsoy (2011)).

La figure 2.3 présente différentes formes classiques de CF ayant été proposées dans la littérature. Une comparaison de l'influence de ces différents modèles sur le processus de planification de production a été proposée par Orcun et al. (2006).

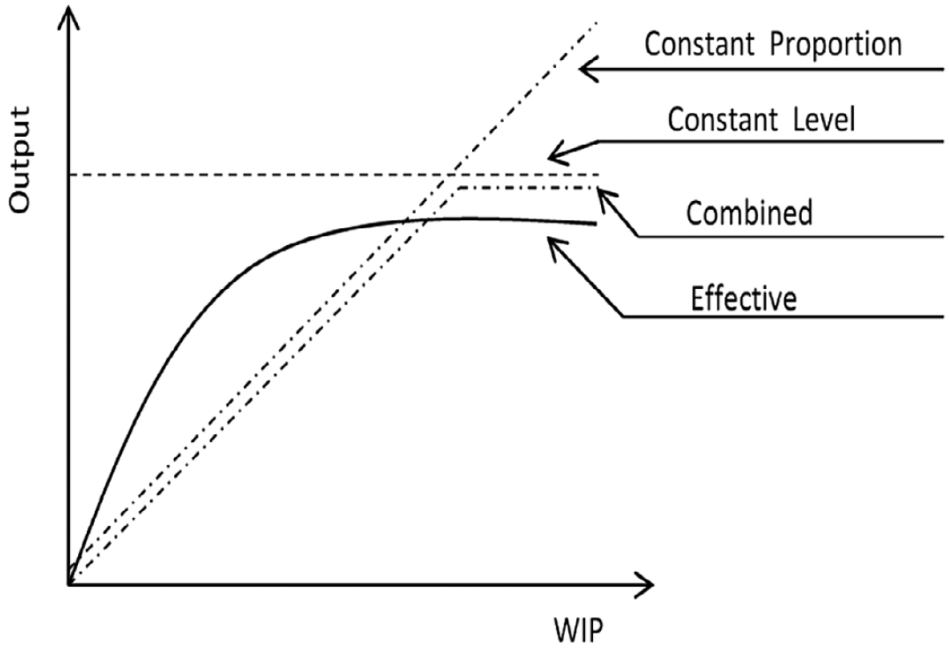


FIGURE 2.3 – Différentes formes de *clearing function* (extrait de Mönch et al. (2018b))

Parmi les courbes présentées, la plus simple est celle *Constant Level*, où la vitesse de sortie des lots (l'*output*) est supposée indépendante du niveau de charge du système (WIP). Ce modèle correspond aux travaux présentés dans la première section, concernant les *lead time* indépendants. Le second modèle *Constant Proportion*, proposé pour la première fois par Graves (1986), suppose une évolution linéaire entre le taux de sortie des lots et la charge du système. Cette approche a l'intérêt d'illustrer que plus une usine a de lots en-cours de production, plus le nombre de lots livrés par unité de temps sera important. Cette relation ne fonctionne cependant que dans une certaine limite, pour des systèmes peu chargés. En effet, à un certain niveau, les encours sont tels que certaines machines atteignent leur capacité limite (100% de leur temps disponible est utilisé). Dans ce cas, augmenter davantage le WIP ne permettra pas d'augmenter le niveau de sortie, ce que ne traduit pas la courbe *Constant Proportion*. Cette représentation linéaire peut-être complétée par une limite de production, donnant alors une courbe combinée. Le quatrième courbe, dite *Effective* dans Mönch et al.

(2018b), modélise la relation entre productivité et charge via une fonction continue strictement croissante et concave. Ce type de courbe a été indépendamment présentée et étudiée pour la première fois par Srinivasan et al. (1988) et Karmarkar (1989). Une introduction assez complète des principaux modèles existant est disponible dans Pahl et al. (2005b).

Une fois le modèle de courbe sélectionné, il faut alors l'estimer afin qu'elle se rapproche le plus possible de la réalité. Les principales méthodes utilisées pour estimer ces courbes peuvent être soit via l'utilisation de la théorie des files d'attente (Missbauer and Uzsoy (2011)), soit via l'utilisation de modèles de simulation (Kacar and Uzsoy (2014)) ou de données historiques. Kalir and Rozen (2018) par exemple, utilisent la formule de Pollaczek-Khinchine afin de lier temps de cycle et charge (Hopp and Spearman (2011)) et l'intègrent dans un outil de *Production Planning* dans une usine *front end* en phase de montée en volume de production (*ramp up*). Les résultats montrent que garder une marge de capacité dans la phase de *ramp up* augmente les volumes de sortie de l'usine. De la même façon que Milne et al. (2015) cherchaient à déterminer les meilleurs valeurs de *lead time* fixe à choisir pour optimiser les plans fournis par un modèle MRP, Kacar and Uzsoy (2015) étudient le problème symétrique en cherchant à définir, non pas les CF les plus proches des observations, mais plutôt celles permettant d'optimiser les performances des plans de lancement.

Les premiers modèles de CF avaient cependant une grande difficulté à modéliser le comportement de systèmes comprenant plusieurs produits différents, étant donné qu'il était possible d'artificiellement augmenter le throughput d'un produit A en augmentant la charge d'un produit B. Pour répondre à cette difficulté, Asmundsson et al. (2009) proposent l'*Allocated Clearing Function* (ACF), où le throughput d'une ressource (machine, atelier, usine) est défini selon une mesure agrégée de la charge, puis redistribué aux différents produits. Dans le cas où les CF sont modélisées sous la forme de fonctions continues par morceaux, l'utilisation des ACF aboutit à une formulation de type PL. Beaucoup de travaux ont été menés sur la base de ces ACF, montrant leur bonnes performances face aux méthodes itératives (Kacar et al. (2012)) ou utilisant des *lead time* fixes (Kacar et al. (2013), Ziarnetzky et al. (2015)). Ziarnetzky et al. (2018) ont utilisé les ACF dans le cadre d'un problème de *Production Planning* considérant des stocks de sécurité de produits finis et des demandes clients variables. Récemment, Guhlich et al. (2018) ont exploité des CF dans le cadre d'un problème intégré de *Production Planning* et de gestion des stocks, considérant de plus les aspects stochastiques de la demande client et de la productivité de l'usine.

En plus de ces performances prometteuses, les CF, parce-qu'elles modélisent la relation entre charge et capacité, permettent d'estimer assez directement le gain de productivité d'une ressource si on augmentait son taux d'utilisation d'une unité.

En revanche, les méthodes basées sur les CF présentent tout de même actuellement des limites. Les CF sophistiquées multi-variées, cherchant à modéliser avec précision la relation throughput/charge de systèmes à plusieurs produits, telles que proposées dans Albey et al. (2014) et Albey et al. (2017), tendent à produire des modèles d'optimisation non convexes qui sont assez difficiles à résoudre. Même la formulation à base d'ACF nécessite la définition de variables de décision et de contraintes pour toutes les opérations effectuées sur chaque groupe de machines, ce qui donne des formulations beaucoup plus grandes qu'une formulation conventionnelle (Kacar et al. (2016)). Une possibilité serait de ne modéliser par les CF qu'un sous-ensemble des ateliers les plus critiques, afin de réduire la taille des modèles. Cependant, Kacar et al. (2018) ont montré que lorsque les ateliers considérés représentent 75% de la moyenne du WIP présent dans le système, les modèles réduits amènent à des résultats significativement inférieurs à ceux obtenus par les modèles complets. Ces résultats suggèrent des interactions complexes entre les différents ateliers, et que les modèles se

concentrant sur un nombre limité de machines peuvent donner des estimations trompeuses des performances du système. De plus, Kacar et al. (2016) ont souligné le fait que dans des conditions stables, les méthodes à bases de CF peuvent s'avérer moins performantes que des méthodes basées sur des *lead time* fixes telles que dans Leachman (1993). Enfin, il n'existe pas encore de méthodologie rigoureuse et générale permettant de construire ces CF à partir de données simulées ou réelles, ce qui constitue un obstacle important à une adoption généralisée (Gopalswamy and Uzsoy (2017)).

Face à ces limites, la solution viendra peut-être de l'utilisation de méta-modèles plus riches que ceux finalement relativement simples des CF. Nous pourrions par exemple citer Li et al. (2016), qui ont proposé une approche de type simulation/optimisation, où la simulation est remplacée par un méta-modèle, et dont les tests sur des problèmes de tailles modérées montrent des résultats prometteurs. Le développement du Big-Data et des capacités de calcul conduit à de nouveaux travaux mettant à profit les grandes quantités de données disponibles dans l'industrie des semi-conducteurs, afin de prévoir toujours plus précisément l'évolution de certains paramètres tels que le temps de cycle (Wang et al. (2018)). Récemment, Gopalswamy and Uzsoy (2018) a mené une étude comparative entre les méthodes à base d'ACF et une alternative récemment proposée par Omar et al. (2017) nommée *data-driven approach*. Le principe de cette approche, qui a l'avantage par rapports aux ACF de ne pas agréger les données ni de devoir configurer certains paramètres, est d'estimer le volume de sorties pour un ensemble fini d'états possibles du système, définis par le niveau de WIP de chaque produit. Des tests ont montré que les ACF amenaient à de meilleures solutions dans des contextes à forte variabilité, mais l'auteur souligne le fait que les méthodes à base de CF ont déjà bénéficié de plus de 10 années de recherche, et que l'approche *data-driven* de Omar et al. (2017) reste intéressante et mérite des études plus approfondies.

Pour résumer, de nombreux travaux ont été proposés pour traiter des problématiques de *Production Planning*, même dans le cas particulier de l'industrie des semi-conducteurs. Ces travaux ont fait appel à une grande variété d'approches, dont la façon de considérer le *lead time* des produits est une composante importante. Dans toutes ces méthodes, l'objectif reste globalement le même, à savoir définir les plans de lancement des lots en production afin de respecter les objectifs de livraison. Une fois ces plans de lancement définis, la dernière tâche de planification est celle de *Production Scheduling*, où l'objectif est cette fois de définir précisément l'allocation et l'ordonnancement des lots sur les différentes machines, tâche particulièrement complexe dans les usines *front end*.

2.5.2 Très court terme: l'ordonnancement des lots

De nos jours, les méthodes d'ordonnancement ne sont pas encore largement implantées dans les usines *front end* (Mönch et al. (2012), à l'inverse des règles de *Dispatching*. Une règle de *Dispatching* permet de choisir le prochain lot à passer sur une machine parmi une liste de lots en attente. Chaque lot possède un poids calculé selon des règles plus au moins avancées. Parmi les plus connues nous pouvons par exemple citer les règles *Earliest Due Date* (EDD) et *Shortest Processing Time* (SPT) qui donnent le plus grand poids au taches ayant respectivement la date de livraison la plus proche, et la durée opératoire la plus courte. La principale raison de leur succès est leur relative simplicité facilitant leur développement, leur intégration mais aussi leur compréhension une fois mises en oeuvre.

Cependant, ces approches ont aussi le désavantage d'être relativement myopes, étant donné qu'elles ne considèrent généralement qu'une machine (ou un petit groupe de machines), et non le problème plus général, et que ces décisions sont prises en ne regardant généralement

pas très loin dans le futur. De plus, ces règles sont difficiles à adapter aux spécificités de chaque atelier et au caractère changeant de l'usine (notamment dans la constitution du WIP). En conséquence, les managers peuvent chercher à complexifier les règles de *Dispatching* en vue de gagner en performance, mais probablement au détriment de l'intelligibilité du modèle.

Face à cela, les méthodes d'ordonnancement optimisé sont quant à elles prometteuses pour le management des ateliers de production (Mönch et al. (2011)). La quantité considérable de données dont ces systèmes ont besoin n'est plus un obstacle grâce au haut degré d'automatisation et à la grande capacité de stockage des bases de données permettant la collecte de données en temps réel. De plus, la puissance de calcul croissante des ordinateurs permet de proposer des ordonnancements de grande qualité dans un temps de calcul court, ce qui permet d'utiliser les solutions de planification optimisées dans un environnement dynamique.

Alors que les règles de *Dispatching* visent à assigner la tâche suivante à réaliser à partir d'un ensemble de tâches en attente, le *Production Scheduling* est le processus plus global d'affectation de tâches à des ressources au fil du temps dans le but d'optimiser un ou plusieurs objectifs (Pinedo (2016)). Axé sur des objectifs plutôt que sur des règles, le *Production Scheduling* facilite la résolution de problèmes englobant plusieurs machines, pourvu que l'ensemble des contraintes pertinentes soient prises en compte, au lieu d'être myope dans l'espace et de limiter les décisions à un groupe particulier de machines. De plus, le paradigme orienté objectif des algorithmes d'ordonnancement les rend plus robustes pour faire face aux situations changeantes dans l'atelier.

L'ordonnancement a fait l'objet d'un très grand nombre de travaux, dont une introduction est disponible dans les livres de Błażewicz et al. (2007), Brucker (2007), Framinan et al. (2014) et Pinedo (2016).

Les usines *front end* sont de type *job shop*, c'est à dire que chaque produit suit une séquence d'opérations sur les machines, avec un ordre différent, à la différence des systèmes de production de type *flow shop*, où chaque produit suit la même séquence de machines. Le problème d'ordonnancement dans un atelier de type *job shop* (JSS pour *Job Shop Scheduling*) est un problème classique puisque les premières formulations ont été proposées dans les années 50 par Bowman (1959), et réputé NP-difficile (Garey (1979)). Les méthodes de résolution les plus efficaces sont souvent basées sur la représentation du problème sous forme de graphes disjonctifs, introduit par Roy and Sussmann (1964), modélisant de manière concise les dépendances entre les opérations.

Dans le cas de la fabrication des semi-conducteurs, ce problème intègre de nombreuses contraintes additionnelles, dont une description est par exemple disponible dans Mason et al. (2002), Mönch et al. (2011) ou Yugma et al. (2015). Dans la suite de cette section, nous mentionnons ces contraintes et passons en revue certains travaux prenant en compte celles-ci dans un environnement de type *job shop*.

Flux ré-entrants. Dans un JSS classique, un produit ne passe généralement qu'une seule fois sur une machine. Dans l'industrie des semi-conducteurs, et notamment dans les usines *front end*, le processus répété de dépôt de couches similaires implique qu'un même lot passe souvent plus d'une trentaine de fois dans le même atelier (Ovacik and Uzsoy (2012)). L'influence de ces flux ré-entrants par rapport à un JSS classique est par exemple étudiée dans Zoghby et al. (2005).

Temps de Setup. Dans le problème classique d'ordonnancement d'atelier, on suppose qu'une opération peut commencer sur une machine dès que cette machine termine l'opération précédente. Dans la pratique, une machine peut cependant nécessiter des réglages, un nettoyage ou des essais avant de commencer l'opération suivante. Le temps nécessaire à cette

préparation est appelé temps de *setup*. Une étude approfondie sur l'influence de l'intégration des temps de setup est par exemple disponible dans Allahverdi et al. (2008) ou Allahverdi (2015). Dans la fabrication de semi-conducteurs, certains ateliers possèdent des machines nécessitant des temps de préparation dépendant de la séquence de produits, c'est par exemple le cas de l'implantation ionique et de la photolithographie. De nombreux travaux ont proposé des approches pour traiter ce problème, nous pouvons citer par exemple les méthodes par recherche taboue de Shen (2014) et de González et al. (2013), ou bien par recuit simulé de Naderi et al. (2010).

Contraintes de disponibilité. La disponibilité continue des machines pendant toute la durée de l'ordonnancement est une hypothèse pouvant être justifiée dans certains cas, mais qui ne peut s'appliquer à tous les environnements industriels. La fabrication de semi-conducteurs est un exemple où il est essentiel de tenir compte des contraintes de disponibilité des machines. Dans cette industrie, les machines sont complexes, nécessitant un entretien préventif fréquent. Elles sont par ailleurs souvent très coûteuses, et il est commun de chercher à les utiliser le plus souvent possible (Bureau et al. (2006)). Une revue des problèmes d'ordonnancement avec contraintes de disponibilités est disponible dans Schmidt (2000) et Ma et al. (2010). Comme pour la plupart des problèmes d'ordonnancement complexes, une grande variété d'approches (généralement heuristiques) a été proposée pour traiter ce problème. Nous pourrions citer par exemple les méthodes à base d'algorithme génétique de Gao et al. (2006), par *tabu thresholding* (variante de la recherche taboue) de Mati (2010), ou bien par recuit simulé (Tamssaouet et al. (2018). Dans ce dernier, une évaluation de l'influence des périodes d'indisponibilité est proposée.

Machines à batch. Certaines machines sont capables de traiter plusieurs lots en même temps, on parle de machines à *batch* (Brucker et al. (1998)). Une revue de la littérature sur les problème d'ordonnancement avec batching peut être trouvée dans Potts and Kovalyov (2000) pour l'industrie en général, et spécifiquement dans l'industrie des semi-conducteurs dans Mathirajan and Sivakumar (2006). La plupart des approches rencontrées pour traiter le JSS avec contraintes de batching reposent sur une représentation sous forme de graphes disjonctifs tels que proposé dans Ovacik and Uzsoy (2012). Cette représentation introduit des nœuds dédiés afin de représenter explicitement le choix des *batch*. Plus récemment, Knopp et al. (2017) ont proposé une nouvelle approche appelée *batch-oblivious*, où les *batch* ne sont plus représentés par des arcs supplémentaires mais sont "encodés" via les poids des arcs existants. L'approche a été reprise par Tamssaouet et al. (2018) et a montré de très bonnes performances tout en simplifiant la complexité du graphe.

Contraintes temporelles. Dans le problème JSS classique, un produit peut attendre indéfiniment devant une machine avant d'être traité. Dans l'industrie des semi-conducteurs, on retrouve des contraintes de temps (souvent maximales), généralement qualifiées de *maximum time lag* dans la littérature, instaurant un délai maximum entre deux étapes de process. Beaucoup de ces contraintes peuvent être rencontrées dans les zones de gravure ou d'oxydation/déposition/diffusion (Mönch et al. (2011), Lima et al. (2019)). L'ajout de telles contraintes rend parfois difficile le fait même de trouver une solution faisable. Klemmt and Mönch (2012) donnent une vue d'ensemble et une classification des différentes contraintes temporelles qui apparaissent en fabrication des semi-conducteurs.

Complexité des machines. En plus d'avoir une grande diversité de fonctionnement (*batch* ou non, temps de setup ou non, ...) certaines machines présentent d'autres caractéristiques complexes. Par exemple, dans la littérature, le temps de traitement d'une étape de process sur une machine est généralement considéré comme fixé et connu à l'avance. Cette modélisation peut s'avérer irréaliste si l'on considère des machines qui présentent un compor-

tement interne complexe. C'est par exemple le cas des machines de type *cluster*, composées de plusieurs chambres (souvent possédant des différences entre elles) pouvant chacune traiter une plaquette, ces dernières étant manipulées par un bras robot interne avec sa propre logique d'ordonnancement. Cette logique d'ordonnancement est interne à la machine et rarement disponible, il en résulte que la séquence des lots envoyés a une influence sur le délai entre leur entrée et leur sortie de la machine (i.e. le temps de traitement est variable). Lee (2008) donne un aperçu de la littérature sur la planification des outils de type *cluster*.

Critère d'optimisation. Il existe une variété de critères pouvant être considérés dans les problèmes d'ordonnancement (Pinedo (2016)). Parmi les plus connus, nous pouvons citer la minimization du *makespan* (C_{max}), du retard maximum (l_{max}), ou du retard total pondéré (TWT pour *Total Weighted Tardiness* en anglais). Dans le contexte de la fabrication des semi-conducteurs, les critères utilisés pour les problèmes d'ordonnancement sont habituellement le temps de cycle, le *throughput* et le respect des délais de livraison (Mönch et al. (2011)). Une difficulté supplémentaire est que ces différents critères, regardés par différents acteurs, sont parfois antinomiques. C'est par exemple le cas du temps de cycle moyen des lots et du taux d'utilisation des machines. L'amélioration de l'un entraîne généralement la dégradation de l'autre (voir section 2.5.1 sur les *clearing function*). Les services chargés de la définition et du suivi des plans de livraison sont par exemple très attentifs au temps de cycle moyen des produits, tandis que les managers, plus opérationnels, regarderont davantage le taux d'utilisation de leurs machines.

Pour résumer, l'ordonnancement dans les ateliers de type *job shop* est un problème classique ayant fait l'objet d'un très grand nombre d'études. Ce problème NP-Difficile se voit être encore plus complexe lorsqu'il est appliqué à l'industrie des semi-conducteurs, du fait de l'ajout de contraintes supplémentaires. Cette complexité implique que ces outils d'ordonnancement, bien que prenant peu à peu le pas sur les règles plus simples de *Dispatching*, ne peuvent s'appliquer qu'à un sous-ensemble de l'usine de production, généralement un atelier. Pour un pilotage automatique d'une usine, il est donc nécessaire de se doter de plusieurs outils d'ordonnancement. Or, comme nous l'avions souligné dans la section 2.4.2, chacun de ces ordonnanceurs cherchera à optimiser un certain objectif lié à la zone qu'il considère, son atelier. Ce manque de vision globale risque d'amener dans cet environnement complexe à une mauvaise gestion du flux des produits, aboutissant à des lots qui ne sont pas livrés dans les temps. Face à cette situation, à cet écart entre l'optimisation locale de l'ordonnancement et la décision très haut niveau des plans de lancement du *Production Planning*, certains travaux, faisant l'objet de la prochaine section, ont cherché à faire le lien entre ces fonctions à travers différentes approches.

2.5.3 Les approches intégrées

Nous venons de voir dans les sections 2.5.1 et 2.5.2 les principaux travaux liés aux problèmes de planification de production et d'ordonnancement d'atelier dans l'industrie des semi-conducteurs. Ces deux problèmes sont généralement traités séparément, d'abord en décidant les lots à lancer en production grâce aux outils de planification de la production, puis en optimisant localement l'ordonnancement des lots sur les machines.

Cette résolution hiérarchique peut amener à des incohérences entre ces deux niveaux, du fait de l'approche souvent agrégée utilisée par les méthodes de *Production Planning* qui ne modélise pas correctement la capacité des machines et donc de l'usine. Dauzère-Pérès and Lasserre (2002) soulignent déjà l'importance d'intégrer les aspects opérationnels dans la planification tactique sous peine de fournir des objectifs inatteignables par les outils de

planification opérationnelle. Partant de ce principe, un nombre non négligeable de travaux ont cherché à mieux lier ces deux niveaux de planification.

Une première solution est d'assurer que les contraintes rencontrées dans les problèmes d'ordonnancement soient suffisamment bien représentées dans les problèmes tactiques de *Production Planning*. Les méthodes à base de *clearing function* ou celles utilisant des modèles de simulation vont dans ce sens. En effet, l'objectif des CF est de rendre compte au mieux du temps de cycle des lots, directement lié à la charge de travail (liée aux en-cours, le WIP) des machines et à leur capacité de production. Une bonne estimation de cette relation permet d'évaluer correctement la capacité de production des machines, et donc améliore les chances de fournir des directives faisables pour les outils d'ordonnancement. De la même manière, les méthodes itératives à base de modèles de simulation ont pour but de modéliser le système de production afin d'intégrer au mieux la réalité des contraintes de production et proposer des plans de production cohérents.

La troisième solution à ce problème de cohérence entre *Production Planning* et *Production Scheduling*, est de traiter ces deux fonctions en même temps. De nombreux travaux ont traité des problèmes intégrés de *Production Planning* (pour être précis, la littérature considère plutôt le terme *Lot-Sizing*) et de *Production Scheduling*, dont une récente revue de la littérature est présentée par Copil et al. (2017). Cependant, ces approches traitent généralement de problèmes soit théoriques, soit assez éloignés des contraintes rencontrées dans l'industrie des semi-conducteurs. On retrouve ainsi certains travaux liés à cette industrie, tels que ceux de Quadt and Kuhn (2005), Quadt and Kuhn (2009), Xiao et al. (2013) et Xiao et al. (2015), mais ceux-ci ne considèrent généralement qu'une partie des contraintes couramment rencontrées (telles que celles de setup) et n'ont été appliqués que sur des problèmes de taille relativement réduite.

2.5.4 À la frontière entre les deux problèmes

Une dernière approche consiste à modéliser le problème à la frontière entre les problèmes de *Production Planning* et de *Production Scheduling*. Le but n'est pas d'affecter les étapes de process sur les ateliers tout en décidant des lots à lancer en production comme le ferait une approche intégrée. Au contraire, le plan de lancement des lots en production est une donnée d'entrée, et les résultats sont une entrée aux outils d'ordonnancement qui, eux, devront affecter et ordonnancer les lots sur les machines.

Nous décomposons ces approches en deux catégories : celles considérant le problème de *WIP Control* (WC), et celles traitant ce que nous nommerons dans cette thèse le problème d'*Operational Production Planning*. Ces deux approches ont pour même objectif de traiter un problème à la frontière entre *Production Planning* et *Production Scheduling* afin de faire le pont entre les plans de livraison, de lancement en production et les outils d'affectation et d'ordonnancement de chaque atelier. Les problèmes de *WIP Control* et d'*Operational Production Planning* varient cependant dans leur modélisation des quantités de production et en conséquences sur les consignes pouvant être transmises pour aider au pilotage de l'usine.

Travaux autour du problème de *WIP Control*

L'objectif du *WIP Control* (ou *WIP Balancing Control*, ou équilibrage des en-cours) est de veiller à la bonne répartition des en-cours (le WIP) sur l'ensemble de la ligne de production. Cette approche est considérée comme étant particulièrement efficace afin d'améliorer les performances du système, telles que la productivité ou le temps de cycle moyen (Lee and

Lee (2003)) , Lee et al. (2002)) ont montré les très bonnes performances des règles de *dispatching* lorsque celles-ci tiennent compte de valeurs cibles (telles que le WIP) déterminées par une approche globale. Les méthodes de *WIP Control* peuvent être décomposées en deux catégories (Barhebwa-Mushamuka et al. (2019)) :

— Les méthodes orientées opérations, où les priorités des lots et des règles d'ordonnancement sont utilisées afin d'équilibrer le WIP selon différentes opérations des gammes de fabrication.

— Les méthodes orientées machines, où des objectifs de WIP sont définis pour certaines machines (généralement goulots). L'équilibrage est réalisé en minimisant l'écart entre le niveau de WIP devant une machine et le niveau de WIP cible.

Fordyce et al. (1992) proposent une planification quotidienne de la production en utilisant un WIP cible pour chaque opération d'un produit. L'objectif est de fournir des quantités de lots qui doivent être traités dans chaque opération à une période donnée afin de répondre à la demande immédiate ou d'anticiper la demande future.

Leachman et al. (2002) proposent, à travers l'outil SLIM (pour *Short Cycle Time and Low Inventory Management*) un ensemble de méthodes et d'applications de gestion du temps de cycle en fabrication de semi-conducteurs. Ces méthodes, allant de la définition de WIP cibles à l'ordonnancement de tâches sur les machines de photolithographie ou de traitement thermique, ont été intégrées sur l'ensemble des sites de Samsung Electronics Corp à la fin des années 90, amenant à une importante réduction du temps de cycle des produits.

Dans Lee and Lee (2003), les auteurs développent plusieurs programmes linéaires (PL) selon différentes approches (flux poussés, flux tirés ou une combinaison des deux) pour un problème de *WIP control* dont le but est de définir des volumes de production cibles pour chaque produit à certains *managing points* (étapes de process liées à un groupe de machines goulots). Les différents PL ont été comparés sur des instances générées aléatoirement et ont montré que les méthodes en flux tirés (*pull*) permettent d'avoir une productivité stable et une bonne satisfaction client avec un coût de production réduit, tandis que l'approche en flux poussés (*push*) tend à maximiser la productivité de l'usine et réduire le temps de cycle moyen.

Lee et al. (2008) font suite aux travaux de Lee and Lee (2003) et développent de nouveaux modèles pour le *WIP Control* intégrant à la fois le maintien de l'équilibrage du WIP sur la ligne de production ainsi que le respect des dates de livraison. Une analyse comparative basée sur des données fictives montre l'intérêt d'intégrer les demandes clients dans les modèles d'optimisation servant à la définition des niveaux de WIP cibles.

Bureau et al. (2007) présentent une approche d'ordonnancement globale en fabrication de semi-conducteurs avec une méthode originale d'évaluation des performances. Dans les approches précédentes, certains paramètres sont déterminés via une approche de résolution globale (généralement par programmation linéaire) et les performances de ces approches sont comparées en simulant le fonctionnement de l'usine sur l'horizon de planification complet. Dans Bureau et al. (2007), le principe est de déterminer les paramètres servant à piloter la production à l'aide de l'approche d'optimisation globale (ici des priorités pour chaque type de produit), de simuler le fonctionnement de l'usine sur un temps donné (une heure, une journée, ...), puis de relancer l'approche globale avec la nouvelle situation de l'usine, de simuler le fonctionnement de l'usine, et de répéter le processus jusqu'à la fin de l'horizon de planification. De cette manière, la simulation tient compte du phénomène réel de la mise à jour régulière des paramètres d'ordonnancement global (ici les priorités) en fonction de la situation changeante de l'usine (par exemple si une machine tombe en panne). Des tests

préliminaires ont été menés à l'aide de données réelles issues d'une usine *front end*, et des résultats prometteurs d'une révision régulière des priorités dans la gestion globale des flux de production sont présentés.

Chung and Jang (2009) proposent une approche pour le *WIP Control* basée sur l'utilisation d'un programme linéaire en nombres entiers (PLNE). L'objectif est de définir des objectifs de production dans l'atelier de photolithographie, pour chaque type de produit et chaque couche. Pour rappel, le procédé de fabrication des semi-conducteurs consiste en un dépôt successif de plusieurs couches de matière sur une plaquette de silicium. Pour chaque nouvelle couche, chaque plaquette doit repasser dans les mêmes ateliers de production (ce sont les flux ré-entrants) et notamment dans l'atelier de photolithographie très souvent considéré comme limitant. Ainsi, sur la base des plans de livraison et du temps de cycle théorique entre deux couches pour un produit donné, le PLNE détermine les quantités à produire par chaque groupe de machines pour chaque couche de chaque produit, l'objectif étant de maximiser la productivité de l'usine.

Bard et al. (2010) se sont intéressés à ce qu'ils nomment *Manufacturing Problem* où, considérant pour une usine "un nombre donné de démarrages de plaquettes par jour et un ensemble d'objectifs de production par produit, le but principal est de développer un modèle permettant de déterminer quand traiter les plaquettes à chaque opération afin d'assurer lesdits objectifs de production". Bard et al. (2010) modélisent d'abord le problème comme un PLNE, puis le résolvent en utilisant soit une méthode de relaxation lagrangienne, soit une *décomposition de Benders* (Benders (1962)), mais sans résultats concluants. Les auteurs ont alors proposé de résoudre le problème de planification, s'étalant sur un horizon d'un à trois mois, en le décomposant en sous-problèmes d'une semaine, chacun d'entre eux pouvant être résolus rapidement par programmation linéaire. Cette approche itérative a cependant le désavantage des approches séquentielles gloutonnes. Bard et al. (2010) soulignent entre autres que le fait d'optimiser sur une semaine empêche de considérer la demande des semaines suivantes et donc ne permet pas au modèle de traiter des lots en avance dans l'optique de mieux servir la demande à venir. Pour pallier à cette faiblesse, les auteurs proposent une heuristique, elle aussi basée sur la résolution d'un programme linéaire, exécutée pour chaque semaine et dont le but est de réajuster les niveaux de WIP de chaque produit afin de tirer profit de la capacité inutilisée de certaines machines suite au premier modèle. L'approche a été testée sur des instances issues d'une usine de Texas Instruments de type *High-Mix Low-Volume* contenant près de 600 machines, 80 produits différents et plus de 650 étapes de fabrication par plaquette. Les tests ont montré que l'approche est capable de fournir des solutions de qualité généralement en moins d'une heure, mais à condition d'agréger les produits en seulement trois grandes familles. Les principales limites (soulignées par les auteurs) de l'approche présentée pour une intégration en production sont d'une part le manque de détails sur les familles de produits prises en compte, et d'autre part que les flux sont considérés sous forme de quantités de produits (quantité de WIP à chaque période) et non sous forme discrète comme des lots. Ces agrégations rendent plus difficiles la considération de contraintes détaillées liées à l'ordonnancement sur les machines, mais aussi implique un effort supplémentaire de développement afin de traduire ces *Manufacturing Plans* en consignes pour les outils opérationnels.

Récemment, Barhebwa-Mushamuka et al. (2019) ont proposé une approche d'ordonnancement global dans le but de minimiser la dispersion du temps de cycle et de la production entre les différents produits d'une usine. L'approche fait suite à celle de Bureau et al. (2007), où l'alternance régulière entre un modèle d'optimisation globale (sous forme d'un PLNE) et un modèle de simulation (combinant à la fois simulation à évènements discrets et simulation

multi-agents) permet de considérer la mise à jour régulière des paramètres de pilotage dans un environnement changeant. L'approche de Barhebwa-Mushamuka et al. (2019) varie cependant de celle de Bureau et al. (2007), d'une part par le pilotage de l'usine via l'utilisation de quantités cibles de production (*production targets)* plutôt que la gestion de la priorité des lots, et d'autre part du fait d'une considération exhaustive de l'ensemble des opérations et de l'ensemble des machines de l'usine. Deux versions du modèle d'optimisation globale sont étudiées, une où le PLNE vise à uniquement minimiser la quantité totale d'en-cours (WIP) dans l'usine, et une autre version pénalisant en plus le déséquilibre du WIP sur l'ensemble de la ligne de production. Les tests sont menés à l'aide de données industrielles avec un nombre réaliste de machines et d'opérations par gamme de fabrication, mais un nombre réduit (cinq) de produits. Les résultats montrent l'importance du contrôle du niveau d'en-cours sur l'ensemble de la ligne de production afin de maintenir une répartition équilibrée et stable de la production et du temps de cycle entre les différents produits de l'usine.

Travaux autour du problème d'*Operational Production Planning*

L'objectif du *WIP Control* est de définir pour les outils d'ordonnancement et d'affectation les quantités de WIP à traiter ou à positionner devant chaque opération ou machine afin d'optimiser un certain critère et notamment le respect des plans de livraison des lots. Les approches traitant ce type de problèmes ne modélisent pas chaque lot individuellement, ce qui peut amener à une perte de précision sur des aspects tels que la modélisation des flux de production ou la modélisation de la charge des machines, mais aussi limite le niveau de précision des consignes pouvant être données. Par exemple, il peut être important de spécifier précisément pour certains lots très prioritaires, les périodes de réalisation de certaines étapes critiques afin de garantir la livraison du lot dans les temps.

Une dernière approche, qui est celle développée dans cette thèse et qui a été assez peu explorée par le passé, a pour objectif de définir la période durant laquelle (ou avant laquelle) chaque étape (ou opération) de chaque lot doit être réalisée. Ce problème, que nous nommons dans cette thèse *Operational Production Planning*, permet notamment un pilotage plus fin des flux de production que l'approche par quantités de plaquettes du *WIP Control*.

Comme cela est souligné dans la récente revue de la littérature de Mönch et al. (2018a), le nombre de travaux ayant exploré cette interface entre planification de production et *Shop-Floor Scheduling*, reste très faible et mériterait d'être davantage étudié à l'avenir.

Les premiers travaux que nous avons recensés sont ceux d'Horiguchi et al. (2001). Dans cet article, les auteurs cherchent à définir, la période de temps dans laquelle chaque couple (lot,opération) doit être exécuté. Toutes les opérations ne sont pas considérées mais seules celles liées aux machines limitantes ou potentiellement limitantes (*Near-Bottleneck* dans leur papier). Pour définir ces plans de production court-terme, les auteurs utilisent une heuristique classique de projection en flux poussés pour les lots du WIP, et de projection en flux tirés pour les demandes ne pouvant être satisfaites par ce WIP. Des tests via un modèle de simulation ont montré l'intérêt de piloter l'ordonnancement des ateliers sur la base des données [lot,opération,période] plutôt que sur la base des dates de livraison des lots, notamment en réduisant le retard global. Horiguchi et al. (2001) ont également mis en avant l'importance de ne pas considérer de dates trop précises de passage d'un lot à une opération du fait de la variabilité trop importante inhérente à l'industrie des semi-conducteurs, et que la définition d'une période de passage était suffisante. Enfin, les auteurs ont aussi montré que modéliser explicitement la capacité des ateliers amène à de meilleurs résultats. Ils ont également montré qu'une fois considérée la capacité d'un atelier limitant, il ne semble pas y

avoir de gain significatif à considérer d'autres ateliers limitants, sauf si l'ensemble des ateliers sont considérés. Pour les auteurs, ces résultats font sens étant donné que, dans le système complexe qu'est la fabrication des semi-conducteurs, avec notamment ces flux ré-entrants, la performance de l'usine est fortement liée à l'interaction entre les ateliers/machines limitants ou potentiellement limitants. Ainsi, le fait d'omettre un seul de ces éléments peut empêcher de capturer la réalité de la capacité de l'usine, ce qui a par ailleurs été rappelé récemment par Kacar et al. (2018).

Habenicht and Mönch (2002) ont étudié un problème similaire où l'objectif est de définir des dates de début et de fin pour chaque opération de chaque lot, ceci en considérant la capacité des machines mais sans pour autant allouer précisément les opérations aux machines. Pour cela, les auteurs ont utilisé une méthode de *recherche en faisceau* (*Beam Search* en anglais, Fox (1983)). Cependant, le temps de calcul restant important, les étapes élémentaires de fabrication sont agrégées en opérations afin de réduire la taille des problèmes. Là encore, l'approche est évaluée via un modèle de simulation et montre l'intérêt pour les outils opérationnels de suivre les indications fournies par le plan de production court terme.

Habla et al. (2007) ont considéré le même problème, d'abord en modélisant ce dernier sous la forme d'un programme linéaire en nombres mixtes (*Mixed Integer Programm* en anglais) et en le résolvant par une méthode de *relaxation lagrangienne* (Held and Karp (1970), Held and Karp (1971)) . Cette méthode a pour but de relacher certaines contraintes (en l'occurrence celles de capacité des machines) et de les intégrer au sein de la fonction objectif afin d'accélérer la méthode de résolution et donc la taille des problèmes résolus. Cependant, Habla et al. (2007) limitent le nombre d'étapes considérées à une vingtaine par lot, correspondant à celles passant sur les groupes de machines considérées comme limitantes, permettant ainsi à l'approche d'être testée sur des instances d'environ 600 lots. Le programme linéaire présenté est assez proche de celui proposé dans cette thèse. La différence majeure est cependant que Habla et al. (2007) ne considèrent pas l'influence de la répartition des produits sur la charge des machines. Cet aspect est pourtant critique dans l'industrie des semi-conducteurs où des machines de générations différentes et donc hautement hétérogènes se partagent les même étapes de process.

Enfin, Mhiri et al. (2018) ont récemment présenté un outil pour la projection du WIP au sein d'une usine de fabrication *front end*, également de type *High-Mix Low-Volume*. Les auteurs présentent une modélisation du problème sous forme d'un programme linéaire en nombres entiers où l'objectif est de minimiser le retard global pondéré (TWT) de l'ensemble des lots. Les principales variables de décision déterminent la période dans laquelle chaque étape de chaque lot est exécutée, ainsi que la répartition de la charge induite par ces étapes sur les machines. Les auteurs ont utilisé ILOG CPLEX pour résoudre le problème, modélisé sous forme d'un PLNE, et ont montré que celui-ci ne pouvait être résolu en des temps raisonnables que pour de petites instances de l'ordre de quelques dizaines de lots et centaines d'étapes de fabrication par lot. Afin de pouvoir résoudre des problèmes de taille réelle, Mhiri et al. (2018) proposent une heuristique en trois étapes capable de fournir des solutions en moins d'une minute. Des analyses numériques montrent les bonnes performances de l'approche heuristique sur de petites instances en comparaison avec le modèle exact, et sur de grandes instances où l'approche trouve des solutions proches (selon différents indicateurs) de ce qui est réalisé dans l'usine. Ainsi, l'article présente davantage un outil de prévision, capable de simuler l'évolution de l'en-cours (WIP) dans l'usine en fonction de la capacité des machines et dans le but de minimiser les retards de livraison. La modélisation reste cependant proche des autres travaux présentés dans cette section.

L'article de Mhiri et al. (2018) est en lien avec le travail de thèse de Mhiri (2016), auquel

cette thèse fait suite. Notre travail se différencie cependant sur plusieurs aspects par rapport aux autres travaux cités dans cette section, et c'est sur ces différences que nous concluons ce chapitre.

2.6 Positionnement de notre problématique

Dans cette thèse, nous considérons le problème à la frontière entre *Production Planning* et *Production Scheduling*, que nous appellerons dans la suite *Operational Production Planning* (OPP). L'objectif du problème *OPP* est de faire le pont en aidant les outils de *Production Scheduling*, par le biais de consignes, à optimiser le fonctionnement des différents ateliers de l'usine, tout en garantissant à terme la livraison à temps des clients.

Ainsi, l'objectif de notre problème est de définir la période de temps dans laquelle chaque étape de process de chaque lot doit être réalisée. Contrairement à la considération sous forme d'en-cours des travaux présentés sur le *WIP Control*, les étapes de process des lots sont considérées individuellement, permettant une meilleure modélisation des contraintes du problème et notamment du temps de cycle. De plus, les étapes ne sont pas agrégées par opération comme dans Habenicht and Mönch (2002), car différentes étapes d'une même opération peuvent avoir lieu dans des ateliers différents, ce qui poserait des difficultés dans la modélisation de la charge des machines.

Étant donné le contexte *High-Mix Low-Volume* des deux usines de Crolles (à l'instar de beaucoup d'autres usines *front end*, notamment européennes), la variabilité importante du mix produit, couplé à une capacité limitée de l'usine afin de faire face à l'ensemble des demandes, oblige à une modélisation suffisamment précise de la capacité des machines et de la charge induite par le plan de production. Il est donc important de considérer l'allocation des différentes étapes de process sur des machines n'ayant pas toutes les mêmes vitesses de traitement, ce qui n'est par exemple pas considéré dans Habla et al. (2007), où seule une capacité globale par groupe de machines est considérée. De plus, une évaluation précise de la charge de travail ne peut se faire en agrégeant les ressources par atelier, comme dans Horiguchi et al. (2001), mais nécessite de considérer explicitement la diversité des machines.

Cette forte variabilité dans le mix produit au fur et à mesure des semaines, implique également des changements potentiels dans les machines (ou groupe de machines) limitantes. Ne considérer que les machines généralement les plus saturées (typiquement certaines machines de l'atelier de photolithographie) risque d'induire une mauvaise évaluation de la capacité globale en laissant échapper les interrelations entre les différents ateliers, comme souligné dans Horiguchi et al. (2001) etKacar et al. (2018). C'est pour cette raison que notre problème considère l'ensemble des machines de l'usine, de même que l'ensemble des étapes de process sont considérées, et non pas seulement celles passant sur les (groupes de) machines limitantes, comme dans Habla et al. (2007) ou Horiguchi et al. (2001).

Ensuite, la plupart des approches traitant du problème d'*Operational Production Planning* se concentrent sur le suivi d'un plan de livraison, et donc sur l'optimisation d'objectifs tels que la minimisation de la déviation par rapport à ce plan de livraison (Bard et al. (2010)) ou bien la minimisation des retards de livraison, comme par exemple dans Habenicht and Mönch (2002) ou Mhiri et al. (2018). Or, bien qu'il semble assez naturel que l'objectif final reste la satisfaction des clients, d'autres critères sont très souvent considérés dans l'industrie des semi-conducteurs. Ainsi, nous considérerons dans cette thèse d'autres critères de performance, et notamment nous évaluerons dans le chapitre 5 différentes approches de résolution selon différents critères tels que le temps de cycle moyen ou le *throughput* moyen de l'usine.

Un autre élément important, lié au contexte industriel, est qu'en plus du problème étudié, l'approche proposée doit pouvoir être intégrée dans un environnement de production, et notamment être capable d'apporter des solutions à des instances de taille industrielle (donc très grandes). Parmi les travaux antérieurs, certains ont proposé des approches via des modèles de PLNE tels que dans Habla et al. (2007) difficiles à étendre à des cas avec des dizaines de produits, des milliers de lots et des centaines de machines différentes. D'autres auteurs ont proposé des approches heuristiques (Horiguchi et al. (2001), Habenicht and Mönch (2002)), mais les tests étaient limités à des instances de petite taille.

Le travail le plus proche de notre problème reste celui de Mhiri et al. (2018) étant donné que notre thèse fait suite à celle de Mhiri (2016). Toutefois, l'étude de cas et l'approche globale sont sensiblement différentes. Premièrement, notre approche intègre mieux la charge de travail des machines, et nous proposons une évaluation de la complexité de notre approche heuristique. Deuxièmement, des modifications sont apportées au modèle mathématique, telles que les tailles variables de lots et les dates de lancement correspondant à des périodes fixes et non à des dates minimales. Troisièmement, dans Mhiri et al. (2018), seuls les critères clients sont pris en compte, liés au retard des lots, tandis que nous présenterons différentes variantes de notre approche afin d'optimiser différents indicateurs. Quatrièmement, l'objectif de Mhiri et al. (2018) est de simuler l'évolution d'un très grand nombre de lots dans l'usine en intégrant la capacité des machines. Notre objectif est de fournir un plan de production qui peut être suivi (dans le respect des capacités), mais qui est aussi la meilleure solution possible (selon les différents indicateurs analysés) et non pas seulement la plus proche possible de la réalité. Contrairement aux travaux précédents, l'objectif n'est donc pas que le plan soit proche de ce qui est observé, mais qu'il soit le meilleur possible afin de servir de plan de référence.

Enfin, un dernier aspect qui cette fois n'a été traité par aucun des travaux cités dans la section 2.5.4, et qui de façon générale n'est que très peu considéré dans la littérature, concerne l'intégration et l'utilisation des solutions proposées dans un système donné. C'est également le cas en fabrication de semi-conducteurs, où on trouve une littérature très riche sur différentes méthodes permettant de résoudre des problèmes de planification de production, mais beaucoup moins sur le développement d'outils d'aide à la décision pour accompagner les acteurs en charge de ces planifications. Dans le cadre de cette thèse CIFRE, le but était de développer un outil d'aide à la décision pour accompagner la réalisation de plans de production directeurs. Dans ce cadre, à la composante optimisation cherchant à définir le meilleur plan de production possible, s'ajoute une composante pratique qui vise à concevoir et à développer une interface permettant aux utilisateurs de tirer le maximum d'information des solutions proposées par l'outil d'aide à la décision.

Le tableau 2.4 résume les principales différences entre les travaux que nous avons présentés dans la section 2.5.4. Les différents critères de comparaison sont les suivants :

— *Modélisation des étapes.* Est-ce que chaque étape de process est considérée individuellement ?

— *Modélisation des machines.* Les machines sont-elles considérées individuellement ?

— *Ensemble des ressources.* Est-ce que toutes les machines de l'usine sont considérées ?

— *Équilibrage avec qualifications.* Les différences de qualifications, de temps opératoire et de capacité entre les machines sont-elles considérées ?

— *Modélisation des lots.* Chaque lot est-il modélisé individuellement ?

— *Différents Objectifs.* Est-il possible d'optimiser différents critères de performance selon les objectifs de l'utilisateur ?

— *Instances réelles*. L'approche proposée permet-elle de traiter des instances réelles en fabrication de semi-conducteurs ?

— *Application industrielle*. L'approche traite t-elle de l'intégration d'un outil d'aide à la planification au sein d'un système de production ?

FIGURE 2.4 – Revue des travaux de la littérature liés à la planification de production opérationnelle et au *WIP Control*, appliquée à l'industrie des semi-conducteurs

	Modélisation des étapes	Modélisation des machines	Ensemble des ressources	Équilibrage avec qualifications	Modélisation des lots	Différents Objectifs	Instances réelles	Application industrielle
Fordyce et al. (1992)							✓	✓
Leachman et al. (2002)	✓	✓		✓			✓	✓
Lee and Lee (2003)	✓	✓		✓		✓		
Bureau et al. (2007)	✓	✓		✓		✓		
Lee et al. (2008)	✓	✓		✓		✓		
Chung and Jang (2009)	✓	✓				✓		
Bard et al. (2010)	✓	✓	✓				✓	
Barhebwa-Mushamuka et al. (2019)	✓	✓	✓			✓		
Horiguchi et al. (2001)		✓			✓			
Habenicht and Mönch (2002)		✓			✓			
Habla et al. (2007)		✓			✓			
Mhiri et al. (2018)	✓	✓	✓	✓	✓		✓	
Nos travaux	✓	✓	✓	✓	✓	✓	✓	✓

2.7 Conclusion

Ce chapitre a été l'occasion d'aborder différents aspects de la planification de production et des travaux qui y sont associés. Après une brève section où nous avons introduit le concept de planification de la chaîne logistique et de sa décomposition selon l'échelle de temps et du maillon (approvisionnement, production, transport, vente) considéré, nous nous sommes intéressés à l'histoire de la planification de production, des premières méthodes théoriques ayant émergé au début du siècle dernier, jusqu'aux progiciels de gestion intégrée actuels. Nous avons ensuite abordé la planification de production dans l'industrie des semi-conducteurs, définissant les principaux blocs principalement délimités selon l'échelle de temps considérée, puis nous avons relevé les différents facteurs de complexité pour la tâche de planification de production, notamment dans les usines *front end*.

Étant donné le problème considéré dans la thèse, à la frontière entre *Production Planning* et *Production Scheduling*, une revue de la littérature pour chacun de ces deux blocs de planification a été présentée. Bien que la littérature liée aux problèmes de *Production Planning*

et de *Production Scheduling* dans l'industrie des semi-conducteurs soit plutôt riche, peu de travaux ont traité de la relation entre ces deux blocs. Notamment, peu de travaux ont cherché à modéliser explicitement le problème intermédiaire, visant à donner des consignes globales aux outils d'ordonnancement, plus précises que les plans de lancement définis par le *Production Planning*, afin de respecter les plans de livraison définis par le *Master Planning*. Nous présentons les approches de *WIP Control* dont le but est de maximiser les performances de l'usine (notamment la productivité et le respect des plans de livraison) via un équilibrage des en-cours sur l'ensemble des lignes de production. Ces approches pilotent les outils locaux d'optimisation (outils d'ordonnancement et de *dispatching*) en fournissant généralement des objectifs de quantités à réaliser (ou à maintenir devant un groupe de machines) pour différentes opérations de chaque produits. Dans cette thèse, nous nous intéressons à un autre problème frontière, plus détaillé, où l'objectif est de définir pour une période de temps dans laquelle chaque étape de chaque lot doit être réalisée. Nous appelons ce problème, qui a été peu exploré par le passé, *Operational Production Planning*.

Une revue de la littérature détaillée des différents travaux considérant ce problème frontière a été présentée, soulignant les différences par rapport à notre approche. Les principales différences sont le niveau de détail choisi dans la modélisation des produits et des machines, la taille des problèmes considérés, la variété des critères de performances étudiés, ainsi que la volonté de développer un outil à intégrer dans le système de production et pensé pour aider au mieux les utilisateurs opérationnels.

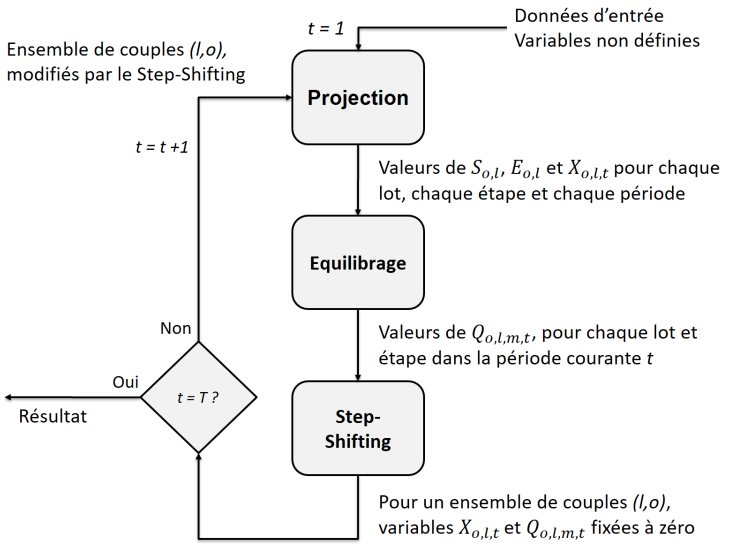
Dans le prochain chapitre, une modélisation formelle du problème sera présentée, ainsi qu'une évaluation théorique et expérimentale témoignant de sa forte complexité, ce qui nous amènera à proposer une approche heuristique en trois étapes pour traiter les problèmes de taille industrielle.

Chapitre 3

Modélisation du problème et structure globale de l'approche de résolution

.

Le problème de planification opérationnelle considéré dans cette thèse a pour but de faire le pont entre les consignes haut niveau décidées par le *Production Planning* et l'optimisation bas niveau du *Production Scheduling*. L'objectif est de fournir des consignes à l'ensemble de l'usine, sans pour autant tomber dans la complexité de l'ordonnancement de tâches. Cependant, la complexité et la taille des problèmes à résoudre rendent impossible l'utilisation de méthodes exactes pour une utilisation industrielle, ouvrant la voie à des méthodes heuristiques plus adaptées.



3.1 Introduction

Nous avons présenté dans le chapitre 1 le secteur des semi-conducteurs ainsi que sa chaîne logistique, et plus spécifiquement la partie liée à la fabrication des circuits intégrés qui est réalisée dans des usines dites *front end*. Ce chapitre, ainsi que le chapitre 2, ont également permis de mettre en évidence les caractéristiques de ces systèmes de fabrication, soulignant la complexité de la tâche de planification de la production. Le chapitre 2 a également mis en évidence un écart important entre les consignes données sous forme d'un plan de lancement des lots en production par le module de *Production Planning* et la tâche importante allouée au *Production Scheduling* de gérer l'ensemble des flux de production afin de livrer les lots à temps, sur la seule base de ces plans de lancement. Peu de travaux dans la littérature se sont intéressés à la modélisation de ce problème intermédiaire et à sa résolution. Ce problème, que nous désignons sous le terme d'*Operational Production Planning*, fait l'objet de cette thèse. Dans ce chapitre, nous définissons le modèle mathématique associé au problème de planification de production que nous étudions. Nous évaluons la complexité théorique du modèle en montrant que celui-ci est NP-Difficile au sens fort, puis nous évaluons cette complexité empiriquement via une étude numérique. Enfin, après avoir montré que le problème est insolvable pour des instances réelles, nous introduisons la structure globale d'une heuristique à trois étapes visant à fournir des solutions de qualité en des temps de calcul de quelques minutes, pouvant ainsi être intégrée dans un outil de planification de la production utilisé quotidiennement.

3.2 Modélisation du problème

3.2.1 Rappel des hypothèses

Avant la définition formelle du problème présentée dans les sections 3.2.2 et 3.2.3, rappelons l'objectif de notre problème d'*Operational Production Planning* (OPP) et ses principales hypothèses.

Tout d'abord, le problème OPP a pour but de faire le lien entre la fonction de *Production Planning* et les fonctions de *Production Scheduling*. Ainsi, les principales données d'entrée sont le plan de lancement des lots en production donné par le *Production Planning*, et le plan de livraison donné par le *Master Planning*. Sur la base de ces plans d'entrée et de sortie, l'objectif de l'OPP est de donner des consignes aux outils de *Production Scheduling* de chaque atelier, afin de les diriger dans les étapes de process à réaliser afin de respecter les livraisons, tout en leur laissant une marge de manoeuvre afin d'optimiser au mieux les contraintes de *Production Scheduling* très bas niveau (batch, temps de setup, cluster, ...). Le but n'est donc pas d'ordonnancer les lots en affectant chaque étape de process de chaque lot à une machine en particulier (ce qui ne pourrait probablement pas être fait correctement à l'échelle de toute l'usine et sur des horizons de plusieurs jours ou semaines).

Le but de l'OPP est *de définir pour chaque étape de chaque lot, la période (principalement jour ou semaine) durant laquelle celle-ci doit être réalisée, ainsi que d'évaluer l'impact sur la charge de travail des différentes machines de l'usine.*

Afin de réaliser ce plan de production, certaines contraintes et hypothèses sont considérées :

— Afin de permettre une gestion globale du flux de production, le problème OPP doit considérer l'ensemble de l'usine, et non pas seulement certains ateliers. De plus, étant

donné le temps de cycle assez long des différents produits (en général de 6 à 10 semaines), la planification doit être faite sur plusieurs semaines afin de ne pas optimiser seulement les livraisons à court terme au détriment des livraisons à plus long terme.

— Chaque lot est modélisé individuellement. Le raisonnement n'est pas conduit sur des quantités, mais considère chaque lot individuellement. Ce choix, proche de la modélisation d'un problème d'ordonnancement, permet d'intégrer assez fidèlement les contraintes d'ordre de réalisation des étapes de process, mais aussi de mieux considérer les temps de process et le temps de cycle des lots (plutôt que de considérer une même valeur de temps de cycle pour chaque lot d'un même produit).

— Étant donné que le problème OPP doit fournir des consignes de production aux approches de *Production Scheduling*, il est important que la modélisation de la capacité des machines soit suffisamment détaillée pour être proche des capacités réelles de production. Ainsi, en plus de modéliser chaque machine individuellement, l'évaluation de la charge induite par le WIP (i.e. les en-cours de production) sur les machines doit tenir compte des qualifications et des performances variables de ces dernières.

— Les machines sont considérées individuellement et ne sont donc pas rassemblées en groupes plus grands tels qu'un atelier comme dans beaucoup de problèmes de *Production Planning* ou ceux proches de notre problème OPP présentés dans la section 2.5.4 du chapitre 2. De même, du fait de la variabilité du mix-produit (pour rappel, la proportion de chaque produit pour un volume donné de WIP) au cours des semaines, toutes les machines sont considérées et non pas seulement celles généralement considérées comme limitantes ou potentiellement limitantes.

— La conséquence du point précédent est que toutes les étapes de process sont considérées et non pas seulement celles passant sur les machines limitantes ou potentiellement limitantes.

— Afin de ne pas tomber dans la complexité opérationnelle gérée par la fonction de *Production Scheduling*, certains aspects tels que les temps de setup ou le fonctionnement par batch sont considérés de façon simplifiée. Ainsi, le temps de setup, responsable d'une baisse du temps disponible de la machine pour traiter des lots, est intégré sous forme d'un pourcentage de temps d'indisponibilité de la machine. Dans le cas des machines à batchs, nous considérons un taux de batch moyen de chaque machine permettant de définir un nombre moyen de plaquettes passant par unité de temps. Ces simplifications ont pour but de s'extraire d'une partie de la complexité du système de production, qui est laissée à la fonction de *Production Scheduling*.

— Les contraintes et coûts de stockage ne sont pas considérées, étant donné qu'ils ne représentent pas un paramètre déterminant dans les décisions de production à notre niveau opérationnel.

Après avoir passé en revue les principaux objectifs, contraintes et hypothèses considérés dans notre problème frontière d'*Operational Production Planning*, une modélisation formelle de ce dernier est proposée dans les sections 3.2.2 et 3.2.3.

3.2.2 Définition du problème et notations

Notre problème d'OPP s'étend sur un horizon décomposé en T périodes temporelles, chacune de ces périodes $t \in \{1, \ldots, T\}$ ayant une longueur p_t. On considère un ensemble $\mathcal{M}$ de machines, ainsi qu'un ensemble $\mathcal{L}$ de lots qui sont soit actuellement dans l'usine (i.e.

les en-cours de production, ce que l'on nomme habituellement WIP), soit faisant partie du plan de lancement des lots en production. Chaque lot l possède une période de livraison d_l (période jusqu'à laquelle le lot peut-être livré sans pénalités), une taille q_l (nombre de plaquettes dans le lot), un poids w_l (sa priorité), et une période de lancement r_l. Chaque lot requiert une séquence $\mathcal{S}_l$ d'étapes de process à réaliser, appelée *route*. Chaque étape de process s d'un lot l peut être réalisée par un sous-ensemble $\mathcal{M}_{s,l,t}$ de machines qualifiées durant la période t, avec un temps de process $a_{s,l,m}$ dépendant de la machine m. Chaque machine m possède une capacité temporelle $c_{m,t}$ durant la période t, qui est une fraction de la longueur de cette période.

La période de fin C_l du lot l correspond à la période de livraison de ce dernier et est une variable dépendante de la solution. Le retard T_l du lot l est défini par la formule $\max(C_l - d_l, 0)$. Du fait du marché très concurrentiel de la micro-électronique, l'un des principaux objectifs à optimiser est le respect des dates de livraison aux clients, dont un critère classiquement associé est le retard total pondéré, ou *Total Weighted Tardiness* (TWT). Ce critère sera utilisé dans la modélisation présentée dans la section 3.2.3, mais d'autres objectifs seront étudiés dans le chapitre 5.

Une solution est un plan de production définissant pour chaque étape s de chaque lot l, la période t durant laquelle l'étape s est réalisée, ce qui est définit par la variable $X_{s,l,t} \in \{0, 1\}$. Un certain nombre de contraintes sont considérées, les plus importantes étant les contraintes de capacité. Les contraintes liées au stockage des lots ne sont pas considérées, n'étant pas critiques à notre échelle de planification et peu pertinentes en fabrication de semi-conducteurs. Un aspect important de cette modélisation est qu'elle considère chaque lot individuellement et ne raisonne pas en termes de volumes, ce qui a l'avantage de mieux modéliser le système de production et notamment les contraintes de capacité des machines ainsi que le temps de cycle des lots. En revanche, l'objectif n'est pas d'effectuer un réel ordonnancement en définissant l'affectation et l'ordre de passage des lots sur les machines, mais seulement de définir dans quelle période chaque étape de process devrait être réalisée. L'intérêt est de réduire la complexité du problème et donc de traiter des problèmes de taille raisonnable, tout en garantissant un meilleur niveau de détail qu'avec la fonction de *Production Planning*.

La tableau 3.1 résume les notations utilisées dans l'écriture du modèle mathématique de dans la section 3.2.3.

TABLE 3.1 – Notations du problème

Ensembles, indices, paramètres	**Description**
$\mathcal{L}$	Ensemble des lots
$l \in \mathcal{L}$	Indice d'un lot
$\mathcal{S}_l$	Ensemble des étapes de process à réaliser pour terminer le lot l
$s \in \mathcal{S}_l$	Indice d'une étape de process du lot l
$\mathcal{M}$	Ensemble des machines
$m \in \mathcal{M}$	Indice d'une machine
T	Nombre de périodes de l'horizon de planification
$t \in \{1, \ldots, T\}$	Indice d'une période
$\mathcal{M}_{s,l,t}$	Machines capables de traiter l'étape s du lot l durant la période t
s_t	Date de début de la période t ($s_0 = 0$)
q_l	Nombre de plaquettes dans le lot l (i.e., la taille du lot), au plus 25
$r_l \in \{1, \ldots, T\}$	Période de lancement du lot l
w_l	Poids (priorité) du lot l
$d_l \in \{1, \ldots, T\}$	Période de livraison du lot l, i.e. la période jusqu'à laquelle le lot peut être livré sans pénalités de retard
$a_{s,l,m}$	Temps requis par la machine m pour traiter une plaquette du lot l à l'étape de process s
$c_{m,t}$	Capacité de la machine m à la période t
Variables	**Description**
$S_{s,l} \in \mathbb{R}^+$	Date de début de l'étape de process s du lot l
$E_{s,l} \in \mathbb{R}^+$	Date de fin de l'étape de process s du lot l
$T_l \in \mathbb{N}^+$	Retard du lot l (en nombre de périodes)
$Q_{s,l,m,t} \in \mathbb{R}^+$	Nombre de plaquettes du lot l à l'étape de process s traitées par la machine m durant la période t
$X_{s,l,t} \in \{0, 1\}$	Egale à 1 si l'étape de process s du lot l est réalisée en période t, et 0 sinon

3.2.3 Modèle Mathématique

En utilisant les notations présentées dans le tableau 3.1, le problème étudié peut être modélisé sous la forme du Programme Linéaire en Nombres Entiers (PLNE) suivant:

$$(P) \quad min \quad \sum_{l=1}^{L} w_l T_l \tag{3.1}$$

$$s.c. \quad X_{1,l,r_l} = 1 \qquad l \in \mathcal{L} \tag{3.2}$$

$$S_{s,l} \geq E_{s-1,l} \qquad l \in \mathcal{L} \quad s \in \mathcal{S}_l \tag{3.3}$$

$$S_{s,l} + \sum_{i \in \mathcal{M}_{s,l,t}} (a_{s,l,m} Q_{s,l,m,t}) \leq E_{s,l} \qquad t \in \{1, \ldots, T\} \quad l \in \mathcal{L} \quad s \in \mathcal{S}_l \tag{3.4}$$

$$\sum_{t=r_l}^{T} X_{s,l,t} = 1 \qquad l \in \mathcal{L} \quad s \in \mathcal{S}_l \tag{3.5}$$

$$S_{s,l} \geq \sum_{t=r_l}^{T} (s_t X_{s,l,t}) \qquad l \in \mathcal{L} \quad s \in \mathcal{S}_l \tag{3.6}$$

$$E_{s,l} \leq \sum_{t=r_l}^{T} (s_{t+1} X_{s,l,t}) \qquad l \in \mathcal{L} \quad s \in \mathcal{S}_l \tag{3.7}$$

$$T_l \geq \sum_{t=r_l}^{T} (t X_{S_l,l,t}) - d_l \qquad l \in \mathcal{L} \tag{3.8}$$

$$\sum_{l=1}^{L} \sum_{s=1}^{S_l} (a_{s,l,m} Q_{s,l,m,t}) \leq c_{m,t} \qquad t \in \{1,\dots,T\} \quad m \in \mathcal{M} \tag{3.9}$$

$$\sum_{m \in \mathcal{M}_{s,l,t}} Q_{s,l,m,t} = q_l X_{s,l,t} \qquad t \in \{1,\dots,T\} \quad l \in \mathcal{L} \quad s \in \mathcal{S}_l \tag{3.10}$$

$$S_{s,l},\ E_{s,l},\ Q_{s,l,m,t} \geq 0 \qquad t \in \{1,\dots,T\} \quad l \in \mathcal{L} \quad s \in \mathcal{S}_l \quad m \in \mathcal{M} \tag{3.11}$$

$$X_{s,l,t} \in \{0,1\} \qquad t \in \{1,\dots,T\} \quad l \in \mathcal{L} \quad s \in \mathcal{S}_l \tag{3.12}$$

$$T_l \in \mathbb{N}^+ \qquad l \in \mathcal{L} \tag{3.13}$$

L'objectif (3.1) est de minimiser la somme des retards pondérés sur l'ensemble des lots. Les contraintes (3.2) assurent que chaque lot est lancé à la période qui lui est imposée. Les contraintes (3.3) assurent la précédence entre deux étapes de process consécutives d'un même lot. Les contraintes (3.4) garantissent que la date de fin d'une étape de process est supérieure à la date de début de l'étape de process précédente plus le temps requis pour la réaliser. Chaque étape de process est réalisée dans une et une seule période grâce aux contraintes (3.5), tandis que les contraintes (3.6) et (3.7) obligent chaque étape de process à finir dans la même période où elle a commencé. Ces contraintes permettent également de faire le lien entre les variables $S_{s,l}$, $E_{s,l}$ et $X_{s,l,t}$. Les contraintes (3.8) permettent de définir le retard de chaque lot. La capacité de chaque machine durant chaque période est respectée grâce aux contraintes (3.9). Comme la préemption est permise, les contraintes (3.10) assurent que toutes les plaquettes de chaque lot sont réalisées pour chaque opération. Les contraintes (3.11) assurent la non-négativité des dates de début et de fin de chaque étape de process ainsi que les quantités traitées par les machines. Les contraintes (3.12) et (3.13) sont les contraintes d'intégrité.

Un modélisation similaire sous forme d'un PLNE à été proposée dans Mhiri (2016), mais notre modèle se distingue de ce dernier sur plusieurs points :

— Dans l'industrie des semi-conducteurs, les livraisons ne sont généralement pas réalisées chaque jour de la semaine. Ces livraisons sont souvent hebdomadaires, ce qui implique que si les livraisons vers les usines de *back end* sont faites par exemple le dimanche, un lot disponible à livraison le vendredi ne sera pas plus en retard qu'un lot disponible dès le mardi. Cette caractéristique est intégrée dans notre formulation du retard T, basée sur la différence entre la période de livraison, et la période où le lot termine sa dernière étape de process. Cet aspect n'est pas considéré dans Mhiri (2016).

— Les plans de lancement des lots en production, déterminés par le *Production Planning*, sont le plus souvent donnés par semaine. C'est à dire que, pour chaque semaine et chaque produit, est définie une quantité à lancer en production, ce qui se traduit pour le problème d'OPP par une liste de lots de chaque produit à lancer en production chaque semaine. Ainsi, dans le modèle PLNE que nous présentons, chaque lot à lancer en production possède une période de lancement, ce qui est moins bien retranscrit dans Mhiri (2016), où chaque lot possède une *date minimum* de lancement en production.

— Dans les usines *front end*, et notamment dans celles de type *High-Mix Low-Volume*, la capacité d'une machine à exécuter telle ou telle étape de process peut varier au fil du temps, notamment par le biais de qualifications permettant de répondre de façon dynamique au différents flux de produits dans l'usine. Cette caractéristique est déjà mentionnée dans la section 1.3.2 du chapitre 1, à propos des facteurs de complexité des systèmes de production des usines *front end*. Dans notre modèle mathématique, l'ensemble des machines capables de réaliser une étape de process donnée d'un produit donné peut varier avec la période considérée. Cet aspect changeant des qualifications au fil du temps n'est pas considéré dans Mhiri (2016).

— Enfin, un dernier aspect qui distingue notre travail de thèse par rapport aux précédents travaux proches du problème d'OPP, est le fait de considérer l'hétérogénéité des machines en termes de vitesse de réalisation des étapes de process. En effet, seules certaines machines sont qualifiées pour réaliser une étape de process d'un produit donné, et le temps de process dépend de la machine considérée. Ainsi, l'allocation des étapes de process aux différentes machines est important afin d'optimiser la charge de travail. Or, dans Mhiri (2016), la quantité de plaquettes de chaque lot pouvant être traitées par chaque machine est définie. Ceci implique une répartition pré-établie, statique, de la charge sur les machines, ce qui supprime un des principaux leviers de décision du problème OPP ! Notons toutefois que dans l'approche heuristique également proposée dans Mhiri (2016), la répartition des quantités sur les machines est cette fois considérée comme une variable. Cela reste cependant un important point de divergence entre le PLNE que nous proposons et celui présenté dans Mhiri (2016).

3.3 Analyse de la complexité du problème

Maintenant que le problème d'OPP a été formellement défini, nous souhaitons évaluer la complexité de ce dernier. Ainsi, dans cette section, nous proposons tout d'abord une analyse théorique de la complexité du problème OPP dans la section 3.3.1, que nous complétons par une analyse des performances du modèle mathématique dans la section 3.3.2.

3.3.1 Complexité théorique

Theorem 1. *Le problème OPP est NP-Difficile au sens fort, même dans le cas d'une seule machine et de deux étapes de process par lot.*

Démonstration. La preuve est faite via une réduction à partir du problème de *Bin Packing* (BBP), un problème NP-Difficile (Korte et al. (2012)) qui, considérant un ensemble de $i \in \{1, \ldots, n\}$ objets chacun avec une taille positive ou nulle s_n et de boîtes de capacité B, consiste à déterminer le nombre minimal de boîtes nécessaires pour ranger les n objets. Pour le problème de décision associé (que l'on nomme BBP_N), prenant une instance du BBP ainsi qu'un entier N, la question est de savoir s'il existe une solution du BBP avec seulement N boîtes. De la même manière, le problème de décision OPP_M associé à notre problème OPP, consiste à déterminer s'il existe une solution de planification nécessitant seulement M périodes.

L'idée de la réduction est de représenter chaque objet du problème BBP par un lot l ayant deux étapes de process. La première étape de process a un temps de traitement total p_l^1 égal à 0, et doit être réalisée dans la première période (période de lancement $r_l = 0$. La

seconde étape de process du lot l possède un temps de traitement total p_l^2 égal à la taille de l'objet i correspondant. Mis à part le temps de process, tous les lots sont identiques. Le problème est réduit à une seule machine, traitant de façon identique tous les lots. La machine a une capacité C par période égale à la capacité B des boîtes b. À chaque boîte est associée une période. Trouver un agencement des objets dans les différentes boîtes correspond alors à trouver un agencement des lots parmi les différentes périodes respectant la capacité de la machine.

Plus formellement, le tableau 3.2 résume les paramètres d'une Instance I du problème BBP_N et donne sa correspondance avec une instance J de notre problème OPP_M.

TABLE 3.2 – Transcription instance de BBP_N vers instance de OPP_M

BBP_N	OPP_M
n objets	n lots
Taille de l'objet i, s_i	Temps de process du lot l, $p_l^2 = s_i$
Capacité des boîtes = B	Capacité de la machine par période = B
	Nombre d'étapes de process par lot = 2
	$p_l^1 = 0 \; \forall \; l$
	$r_l = 0 \; \forall \; l$
	$w_l = 1 \; \forall \; l$
	$d_l = N \; \forall \; l$
$X_{i,b} = 1$ si l'objet i est placé dans la boîte b, 0 sinon	$X_{l,t} = 1$ si le lot l est placé dans la période t, 0 sinon

On veut montrer qu'une instance I du problème de BBP_N est positive si et seulement si il est possible de planifier tous les lots sans qu'aucun d'entre eux ne soit en retard.

Si I est une instance positive, cela signifie qu'il est possible d'agencer les objets dans N boîtes tout en respectant la capacité de ces dernières. Pour chaque objet i placé dans la première boîte, on prend les lots l associés et exécute leur seconde étape de process dans la première période de planification (les premières étapes de process étant toutes réalisées dans la première période). Comme il existe au moins une solution en N boîtes respectant les contraintes de capacité, donc a fortiori pour la première boîte $\sum_i (s_i X_{i,1}) \leq B$, la planification des lots associés aux objets dans la première période respecte également la capacité de la machine. Le raisonnement est répété pour toutes les autres périodes, on en conclut qu'il existe un plan de production faisable en seulement N périodes, respectant donc les périodes de livraison, c'est à dire avec aucun retard.

Inversement, considérons un plan de production faisable où aucun lot n'est traité en retard. À chaque période t est associée une boîte b. On place alors dans la boîte b, l'ensemble des objets i dont la seconde étape de process du lot l correspondant est réalisée dans la période t associée. Comme la solution est faisable, elle respecte notamment la contrainte de capacité $\sum_i (s_i X_{i,b}) \leq B \; \forall b$. Enfin, comme la solution n'admet aucun lot en retard, cela signifie que la solution nécessite au plus N périodes. Il existe donc une solution de rangement des objets en seulement N boîtes. L'instance I est donc positive.

□

3.3.2 Analyse expérimentale

Dans cette section, nous souhaitons compléter l'étude théorique de complexité par une analyse numérique des performances du PLNE présenté dans la section 3.2.3. Cette étude se

veut assez succincte pour deux raisons. La première, est qu'une étude similaire assez complète a déjà été proposée dans la thèse de Mhiri (2016), et nous invitons le lecteur à s'y référer s'il souhaite approfondir les résultats de nos travaux. Ensuite, contrairement à Mhiri (2016) qui a mené des travaux approfondis sur la résolution exacte du problème, notamment via l'utilisation d'une méthode de relaxation lagrangienne, le but premier de notre thèse était de travailler sur une approche performante et suffisamment rapide pour pouvoir s'intégrer dans le processus de planification de l'entreprise. Or, les conclusions de cette section montreront que, bien que l'utilisation de méthodes sophistiquées puisse améliorer la taille des problèmes solvables, ces derniers restent tout de même bien trop petits par rapport à ceux rencontrés dans notre application industrielle.

Ainsi, l'étude numérique effectuée a été la suivante. Le modèle mathématique présenté dans la section 3.2.3 a été résolu avec IBM ILOG CPLEX 12.7.1 sur de très petites instances obtenues à partir de données industrielles simplifiées. Les caractéristiques des ces instances sont résumées dans le tableau 3.3.

TABLE 3.3 – Caractéristiques des instances de test simplifiées

Paramètres	**Valeurs**
Nombre de lots	1 à 7
Nombre moyen de machines	357
Nombre de produits	3
Nombre moyen d'opérations/lot	350
Période de livraison des lots	Distantes de 1, 1.5 ou 2 fois le temps de cycle moyen
Longueurs et nombre de périodes	8 périodes d'une semaine chacune

Les instances ont été tirées de cas réels, mais certains paramètres ont été fortement réduits. Ainsi, le nombre de machines et le nombre moyen d'étapes de process sont très proches de ceux trouvés dans l'usine. En revanche, le nombre de lots a été le principal paramètre permettant de réduire la taille des instances considérées allant de 1 à 7 lots, contrairement aux milliers rencontrés habituellement dans l'usine. La longueur et le nombre des périodes (8 semaines) est un cas classique pouvant être rencontré dans notre contexte industriel. Les périodes de livraison ont été réajustées selon trois configurations possibles. Les instances les plus simples considèrent des lots ayant des périodes de livraison telles que le délai disponible restant soit deux fois supérieur au temps de cycle théorique du lot. Des instances plus difficiles ont également était proposées avec des périodes de livraison laissant un délai valant exactement le temps de cycle moyen du lot considéré.

Un résumé des résultats est proposé dans le tableau 3.4, indiquant le temps moyen requis par l'ordinateur (Intel Core i5 1.60GHz, 8 Go de RAM) pour résoudre les différentes instances. Pour éviter des problèmes de variabilité, chaque cas représente la valeur moyenne issue de 10 résolutions. De plus si au bout de 3 heures, la solution optimale n'était pas trouvée par IBM ILOG CPLEX, celui-ci était arrêté. D'après le tableau 3.4, on constate que le modèle a été capable de résoudre en moins d'une heure des instances considérant 4 ou 5 lots, selon la difficulté des instances.

Suite à ce constat, nous avons voulu améliorer les performances du modèle initial par l'utilisation de deux techniques rapidement implémentables.

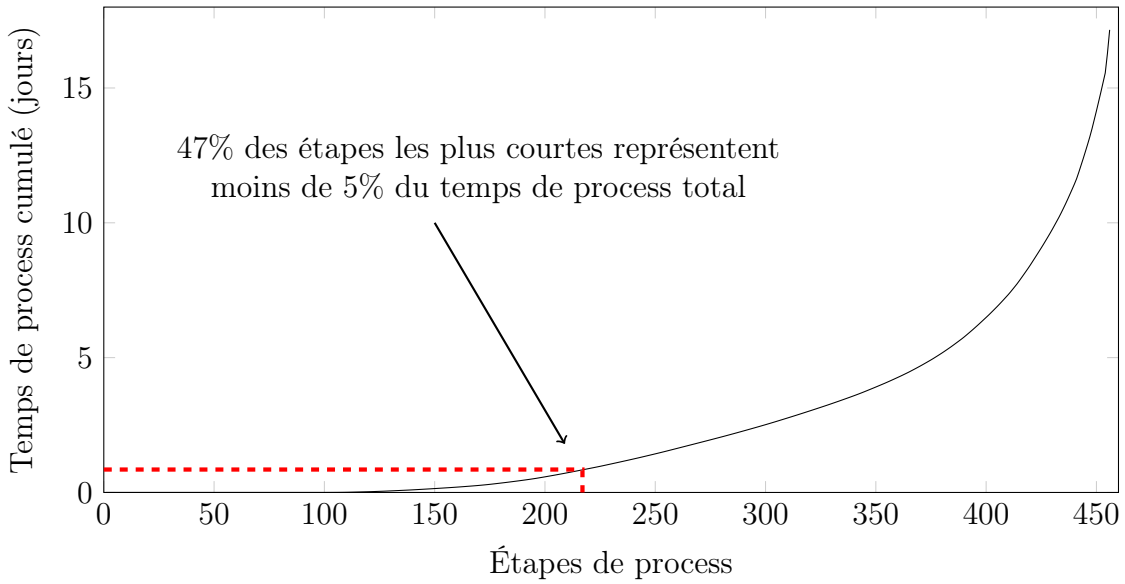
Premièrement, comme les instances ont souvent un grand nombre d'étapes de process avec des temps de traitement très courts, il est possible d'agréger ces étapes de process en un temps total fixe qui resterait tout de même petit par rapport à l'horizon de planification

TABLE 3.4 – Temps de calcul moyen (en secondes) pour la résolution exacte du PLNE et du PLNE amélioré sur de très petites instances de tailles et difficultés variables

Taille des instances	**PLNE initial (sec.)**		**PLNE amélioré (sec.)**		**Réduction du temps (%)**	
	Facile	**Difficile**	**Facile**	**Difficile**	**Facile**	**Difficile**
1 lot	13	12	12	11	-12%	-5%
2 lots	26	29	22	19	-13%	-34%
3 lots	212	523	73	56	-66%	-89%
4 lots	464	1572	147	424	-68%	-73%
5 lots	4100	9645	413	1547	-90%	-84%
6 lots	-	-	23010	3404	-	-
7 lots	-	-	7234	14228	-	-

global. Cette agrégation a donc peu d'impact sur la solution finale. La figure 3.3.2 illustre le temps de process cumulé des différentes étapes de process pour un produit classique de l'usine Crolles 300. Le temps total des étapes de process représente environ 17 jours, pour un total d'environ 450 étapes. Or, on constate que le cumul des 217 étapes les plus courtes (environ 47%) représente moins de 5% du temps total de process de la route. L'agrégation permet donc de réduire considérablement le nombre de variables associées aux étapes de process, et donc d'améliorer le temps de calcul, sans pour autant impacter la solution optimale.

FIGURE 3.1 – Temps de process cumulé par ordre croissant des étapes de process d'une gamme de fabrication classique



Une deuxième approche que nous avons utilisée consiste à résoudre d'abord le problème OPP en relâchant les contraintes de capacité. L'intérêt est que chaque lot peut alors être planifié, d'une part indépendamment des autres, et d'autre part de façon très simple en sélectionnant pour chaque étape de process la date de démarrage au plus tôt. Ce pré-traitement est donc réalisé très rapidement via une heuristique gloutonne et fournit une solution qui servira de borne inférieure pour la résolution du problème général.

Ces améliorations ont conduit à une réduction globale du temps de calcul du modèle, rendant le PLNE amélioré capable de résoudre en moins d'une heure des cas allant cette

fois de 6 à 7 lots. Cependant, malgré ces améliorations, plusieurs heures sont encore nécessaires pour résoudre des problèmes avec seulement 10 lots, rendant difficilement envisageable l'utilisation d'une telle approche pour résoudre les problèmes de planification opérationnelle rencontrés dans l'usine.

Comme nous l'avions mentionné en ce début de section, une étude d'un PLNE proche de celui que nous présentons a été proposée dans Mhiri (2016) avec toutefois l'application d'une méthode de relaxation lagrangienne afin de pousser encore plus loin la taille des instances pouvant être résolues de façon optimale. Cependant, les conclusions présentées restent les mêmes que les nôtres, à savoir qu'une méthode basée sur une résolution exacte du problème a peu de chance d'être pertinente pour faire face à la réalité industrielle. C'est pour cette raison qu'une méthode heuristique a été développée afin d'être au coeur d'un outil d'aide à la décision pour ce problème d'*Operational Production Planning*. Nous allons introduire cette heuristique dans le reste de ce chapitre.

3.4 Approche heuristique en trois étapes pour la résolution du problème

Nous venons de voir dans la section précédente, à travers l'analyse théorique et expérimentale de la complexité du problème, que celui-ci ne pouvait guère traiter des problèmes considérant plus d'une dizaine de lots. Or, comme cela a été déjà mentionné, la fabrication de semi-conducteurs implique la gestion de problèmes considérants des milliers de lots, chacun nécessitant de passer par des centaines d'étapes de process. Il est donc peu probable que des méthodes exactes puissent être mises en œuvre pour résoudre des cas réels de planification opérationnelle. Nous avons donc développé un outil de planification basé sur l'utilisation d'une heuristique en trois étapes, que nous appellerons par la suite *approche TSH* pour *Three-Step Heuristic*. Cette approche TSH est décomposée en trois modules principaux schématisés dans la figure 3.2.

3.4.1 Structure globale

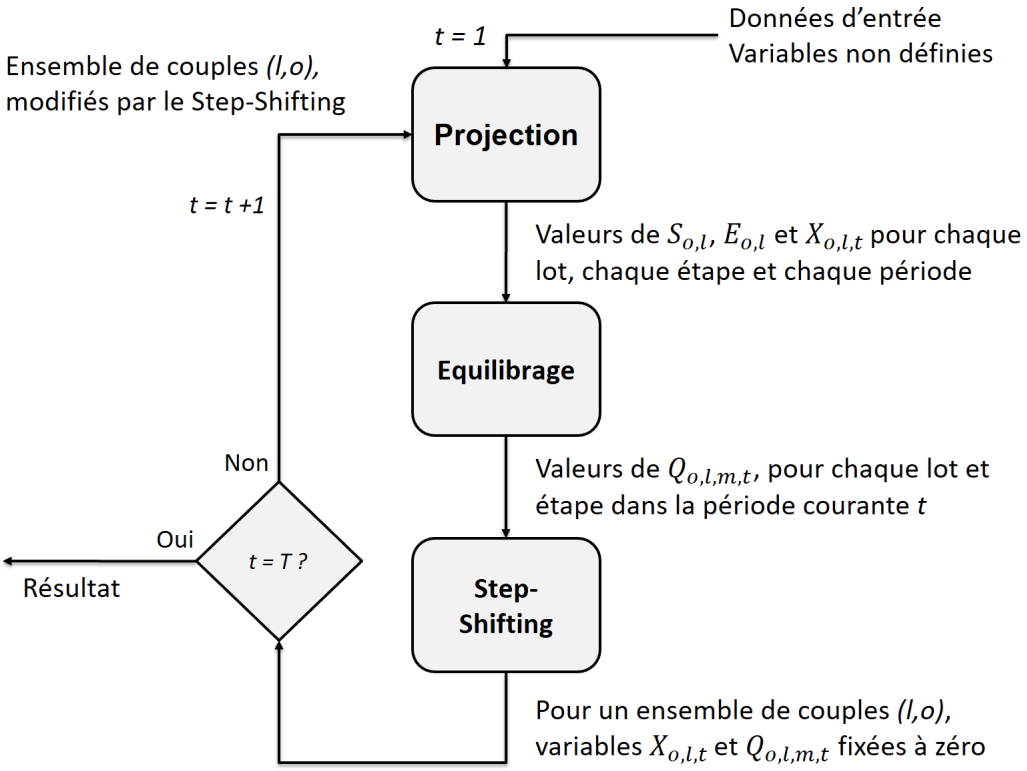


FIGURE 3.2 – Schéma de l'approche TSH

L'horizon de planification est discrétisé en périodes, et l'approche TSH utilise alternativement chacun des trois modules sur chaque période de l'horizon de planification.

1. Le premier module, appelé module de projection et qui est détaillé dans la section 3.4.2, prend en entrée différentes informations et notamment des données historiques de temps de cycle. À partir de ces informations, le module de projection fixe pour chaque étape de process de chaque lot, des premières dates de début et de fin, fixant donc des valeurs pour l'ensemble des variables $S_{s,l}$ et $E_{s,l}$. Le module fixe également logiquement la période pendant laquelle chaque étape de chaque lot doit être réalisée, déterminant des valeurs pour chaque variable $X_{s,l,t}$. Cette première étape ne tient pas compte des contraintes de capacité.
2. Connaissant les étapes de process devant être réalisées dans chaque période, le module d'équilibrage, détaillé dans la section 3.4.3, a pour but but de répartir les lots sur les machines disponibles (toujours sans tenir compte des contraintes de capacité), ce qui donne une estimation de la charge de travail de ces dernières. Comme les contraintes de capacité ne sont pas prises en compte, certaines machines peuvent être surchargées et la solution peut donc ne pas être réalisable.
3. Ainsi, un troisième module, appelé module de *step-shifting* et décrit dans la section 3.4.4, a pour tâche de décaler certains lots à des périodes ultérieures afin de lisser la charge de travail des machines, et donc de satisfaire les contraintes de capacité.

Après avoir itéré sur l'ensemble des périodes de l'horizon de planification, le résultat final est une solution réalisable, c'est-à-dire un plan définissant pour chaque étape de process

de chaque lot la période dans laquelle l'étape est réalisée et respectant les contraintes de capacité des machines.

3.4.2 Module de projection des lots

Dans ce premier module, appelé module de projection, l'objectif est d'attribuer des dates de début et de fin pour chaque étape de process de chaque lot. Afin de ne pas tomber dans la complexité du problème initial, les contraintes de capacité ne sont pas prises en compte dans un premier temps.

Afin de définir des dates de début et de fin, nous devons prendre en compte de nouvelles données, à savoir pour chaque étape de process des temps de traitement minimum et des temps d'attente minimum. Le temps de traitement minimum $pt_{s,l}^{min}$ correspond au temps minimum requis pour traiter l'étape s du lot l. Le temps d'attente minimum $wt_{s,l}^{min}$ correspond au temps minimum entre la fin de l'étape s du lot l et le début de la prochaine étape. Elle est principalement due au temps de transport et au temps de chargement/déchargement sur les machines. Dans un environnement industriel, ces temps de traitement et d'attente théoriques peuvent être obtenus à partir de plusieurs mois de données historiques. La procédure de projection est détaillée dans l'algorithme 1.

Algorithme 1 : Procédure du module de projection

Input : $init\_op_l$: Pour chaque lot, indice de l'étape de process au début de la projection
d_l: Période de livraison du lot l
t: Date de début de projection
$pt_{s,l}^{min}$: Temps de process minimum pour chaque étape de process de chaque lot
$wt_{s,l}^{min}$: Temps d'attente minimum pour chaque lot entre les étapes de process s et $s+1$

Output : Variables $S_{s,l}$ et $E_{s,l}$, $\forall l \in \mathcal{L}, \forall s \in \mathcal{S}_l$

$rem\_wt_l^{min} = \sum_{s \in [init\_op_l, ..., S_l]} wt_{s,l}^{min}$;
$rem\_pt_l^{min} = \sum_{s \in [init\_op_l, ..., S_l]} p_{s,l}^{min}$;
for $l \in \mathcal{L}$ **do**
    `// Si le temps restant disponible est supérieur au temps de cycle minimum restant`
    if $(d_l - t \leq rem\_pt_l^{min} + rem\_wt_l^{min})$ **then**
        $\alpha = 1$
    else
        $\alpha = \frac{d_l - t - rem\_pt_l^{min}}{rem\_wt_l^{min}}$
    end
    for $s \in [init\_op_l, ..., S_l]$ **do**
        $E_{s,l} = S_{s,l} + p_{s,l}^{min}$
        $S_{s+1,l} = E_{s,l} + \alpha wt_{s,l}^{min}$
    end
end

Le module de projection calcule séquenciellement pour chaque lot, indépendamment les uns des autres, ce qui se traduit par les deux boucles "for". Puisque le nombre maximum

d'étapes de process par lot est limité par $K = \max_{l \in \mathcal{L}}(|\mathcal{S}_l|)$, l'algorithme est donc rapide avec une complexité linéaire en $O(K|\mathcal{L}|)$. Pour chaque lot, les temps minimums de traitement et d'attente cumulés $(rem\_pt_l^{min}, rem\_wt_l^{min})$ avant que le lot quitte l'usine (lignes 1 et 2 dans l'algorithme 1), sont calculés. Ce temps est appelé *temps de cycle minimum restant* du lot. À noter que, dans Mhiri et al. (2018), le temps d'attente minimum n'est pas pris en compte et le temps de cycle minimum restant d'un lot est considéré comme étant la somme des temps de traitement des étapes de process restantes.

Puisque tous les lots de l'usine ne se déplacent pas à la même vitesse, principalement du fait de dates de livraison différentes, la phase suivante consiste à déterminer un coefficient α pour ajuster les temps d'attente des lots. Pour cela, deux cas sont considérés :

Cas 1 Le lot ne pourra sortir avant sa période de livraison (le temps disponible restant $(d_l - t)$ est plus petit que le temps de cycle minimum restant $(rem\_pt_l^{min} + rem\_wt_l^{min})$), le coefficient α est alors égal à 1.

Cas 2 Le temps disponible restant est supérieur au temps de cycle minimum restant, le coefficient α est alors une fonction du temps d'attente disponible $(d_l - t - rem\_pt_l^{min})$ et de la somme du temps minimum restant de traitement et d'attente.

L'étape suivante consiste à définir, pour chaque lot, les dates de début et de fin de chaque opération, en commençant itérativement par l'étape de process en cours au début de la projection. La ligne 10 exige que la date de fin d'une étape de process soit égale à sa date de début plus le temps de traitement minimal. La ligne 11 indique que le temps d'attente entre deux étapes de process est égal au temps d'attente minimum avant la prochaine opération, pondéré par le coefficient α. Plus la marge du lot est grande par rapport à sa période de livraison, plus le temps d'attente est long pour refléter le fait que le lot a une priorité moindre et qu'il passera donc plus de temps à attendre devant les machines. Le coefficient α est défini de telle sorte que, pour chaque lot, la somme des temps de traitement et des temps d'attente fait sortir ce lot à sa période de livraison. Une illustration de cet ajustement du temps d'attente se trouve dans Mhiri et al. (2018).

À la fin de ce module, nous avons des dates de début et de fin ($E_{s,l}$ et $S_{s,l}$) pour chaque étape de process de chaque lot. Par conséquent, des valeurs sont également obtenues pour les variables $X_{s,l,t}$, indiquant dans quelle période chaque étape de process est traitée. À noter que, lorsqu'une étape de process chevauche deux périodes, le lot est considéré comme étant dans la période dans laquelle la majeure partie de son temps de traitement est affectée. Nous disposons alors d'une première solution indiquant dans quelle période traiter chaque étape de process de chaque lot. Cependant, cette solution n'est pas complète car les informations des machines sur lesquelles effectuer ces étapes de process manquent. En outre, les contraintes de capacité ne sont pas encore prises en compte (sauf indirectement par l'intégration de temps d'attente minimaux), et il n'est donc pas garanti que la solution soit réalisable. Dans l'étape suivante, nous répartissons donc les lots sur les différentes machines et évaluons l'impact de cette première solution sur la charge de travail.

3.4.3 Module d'équilibrage des charges

Le second module vise à répartir les quantités à traiter en période t sur les différentes machines disponibles, c'est-à-dire à déterminer les variables $Q_{s,l,m,t}$, $(\forall\ l, s, m, t)$ dans le problème de programmation linéaire en nombres entiers (P). La première étape consiste à décomposer notre problème en séparant les machines en différents groupes appelés *Isolated Group* (IG). Cette décomposition est basée sur la notion de qualification d'une machine

pour une recette (voir description chapitre 1, section 1.2.2) donnée, c'est-à-dire la capacité de cette dernière à réaliser une certaine étape de process pour un certain produit. Plus précisément, la machine m est qualifiée pour traiter l'étape de process s du lot l dans la période t si $m \in \mathcal{M}_{s,l,t}$. Dans un contexte de production de type *High-Mix Low-Volume*, avec de nombreuses machines non identiques, la diversité des qualifications est très élevée. Les IG sont construits de telle sorte que deux machines de deux IG différents ne partagent aucune qualification commune, que ce soit directement ou par transitivité. Nous pouvons formaliser cette propriété sous la forme d'un graphe où les machines correspondent à des nœuds et où deux noeuds sont liés par une arête si et seulement si les deux machines correspondantes partagent au moins une qualification pour réaliser une même recette. La figure 3.3 illustre cette représentation sous forme de graphe des qualifications partagées. Dans un tel graphe, l'ensemble des IG correspond à l'ensemble des différents graphes connexes.

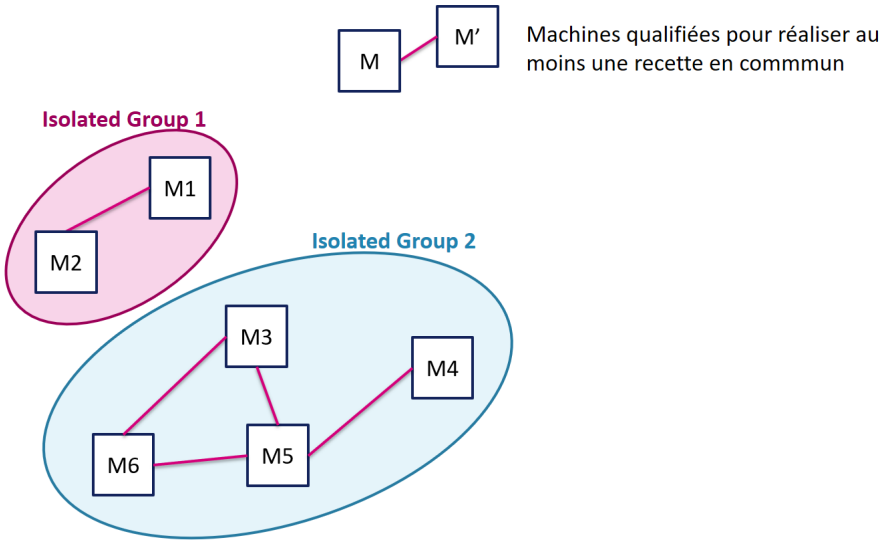


FIGURE 3.3 – Construction des *Isolated Group* par la répartition des machines selon leurs qualifications communes

En groupant les machines de cette façon, l'équilibrage des machines de chaque IG peut être traité indépendamment de l'équilibrage des autres groupes, puisque par construction deux machines de groupes différents ne peuvent pas s'influencer mutuellement pour s'équilibrer. Ainsi, la procédure suivante est répétée en parallèle pour chaque IG.

Chaque sous-problème est modélisé comme un programme linéaire. En utilisant les notations utilisées pour le modèle mathématique de la section 3.2.3, le programme linéaire peut être écrit comme suit pour la période t et un groupe isolé de machines IG.

$$min \quad \alpha \max_{m \in \mathcal{M}} W_m - \beta \min_{m \in \mathcal{M}} W_m + \gamma \sum_{m \in \mathcal{M}_p} c_m W_m \tag{3.14}$$

$$(LP) \qquad W_{m,t} = \frac{\sum_{l=1}^{L} \sum_{s=1}^{S_l} (a_{s,l,m} Q_{s,l,m,t})}{c_{m,t}} \qquad m \in IG \tag{3.15}$$

$$\sum_{m \in \mathcal{M}_{s,l,t}} Q_{s,l,m,t} = q_l x_{s,l,t} \qquad l \in \mathcal{L} \quad s \in \mathcal{S}_l \quad \text{s.t. } \mathcal{M}_{s,l,t} \in IG \tag{3.16}$$

$$Q_{s,l,m,t} \geq 0 \qquad l \in \mathcal{L} \quad s \in \mathcal{S}_l \qquad m \in IG \tag{3.17}$$

Dans ce problème, la fonction objectif présentée est celle utilisée au démarrage de la thèse, qui est également celle proposée par Mhiri et al. (2018). Le but est d'optimiser une combinaison de trois termes, à savoir, pour les deux premiers termes, minimiser le taux de charge (le rapport entre le temps de traitement total assigné à la machine et sa capacité) de la machine la plus chargée, puis maximiser le taux de charge de la machine la moins chargée, afin de réduire la dispersion des taux de charge. Enfin, le troisième terme à pour but de minimiser le temps total de process, et donc vise en particulier à faire passer les lots sur les machines les plus rapides. Les coefficients (α, β, γ) pondèrent les différents termes de la fonction objectif. Dans l'implémentation initiale de l'outil, les coefficients étaient choisis tels que $\alpha >> \beta >> \gamma$, privilégiant ainsi dans l'ordre la minimisation du taux de charge maximal, la maximisation du taux de charge minimal, puis la minimisation du temps de process global. Les contraintes (3.15) définissent le taux de charge de chaque machine, et il est à noter que, bien que la capacité de la machine soit prise en compte, elle n'est utilisée que pour évaluer le taux de charge de la machine et ne constitue pas une contrainte de capacité réelle. Les contraintes (3.16) garantissent que chaque étape de process planifiée dans la période considérée est traitée. À noter également que les $x_{s,l,t}$ ne sont pas des variables, ayant été fixés dans le module de projection. Ainsi, les seules variables restantes sont $Q_{s,l,m,t} \in \mathbb{R}^+$. Ces variables étant réelles, cela signifie que la préemption est autorisée.

Ce problème d'équilibrage de la charge de travail initialement développé a l'avantage d'être simple de compréhension avec une fonction objectif semblant assez naturelle, mais elle présente des limites qui seront présentées et illustrées dans le chapitre 4. Pour pallier à ces limites, ce chapitre introduira une nouvelle approche, ayant fait l'objet d'un article (Christ et al. (2019)), fournissant des solutions avec certaines propriétés et étant plus efficace que l'approche initiale utilisée dans Mhiri et al. (2018).

Lorsque le processus est terminé, on obtient une allocation des étapes de process de chaque lot dans la période t sur les différentes machines, c'est-à-dire des valeurs pour $Q_{s,l,m,t}$. Étant donné que le problème ne tient pas compte des contraintes de capacité, cette solution est potentiellement irréalisable. Néanmoins, ce module nous permet d'estimer une bonne répartition des étapes de process sur les machines dans la période t et l'impact sur leur charge de travail dans chaque période. À la suite de ce module, l'objectif est d'adapter la solution si elle n'est pas réalisable. Pour tenir compte des contraintes de capacité des machines, nous avons donc développé un troisième module qui inclut une procédure de lissage avant pour déplacer les étapes de process d'une période à la suivante.

3.4.4 Module de lissage de la capacité *(step-shifting)*

À l'entrée de cette étape, des valeurs ont déjà été assignées à toutes les variables $S_{s,l}$, $E_{s,l}$, $X_{s,l,t}$ et $Q_{s,l,m,t}$. Ainsi, un plan de production complet est défini pour la période t, qui cependant peut ne pas répondre aux contraintes de capacité. Le module de *step-shifting* construit une solution réalisable à partir de ce plan initial, en utilisant un processus de lissage vers l'avant, une approche utilisée par exemple dans les problèmes de *Capacited Lot-Sizing* (Trigeiro et al. (1989); Brahimi et al. (2006); Lu et al. (2013)). L'idée est de sélectionner certains lots ayant des étapes de process à réaliser dans la période courante et de reporter certaines étapes à la période suivante afin de réduire la charge induite sur les machines. L'algorithme 2 décrit la procédure générale du module de *step-shifting*.

Algorithme 2 : Procédure du module de *step-shifting*

Input : t = Période courante
$X_{s,l,t}$, $Q_{s,l,m,t}$ = Valeurs définies pour chaque étape de process de chaque lot et chaque machine
$W_{m,t}$ = Taux de charge de la machine m durant la période t
Output : $\mathcal{O}_{shifted}$ = Ensemble des étapes de process décalées à la période suivante
`// Tant qu'il existe une machine surchargée`
$\mathcal{O}_{shifted} \leftarrow \emptyset$
while $\max_{m \in \mathcal{M}} W_{m,t} > 1$ **do**
  `// Sélectionne la machine la plus chargée`
  $m' \leftarrow \text{argmax}_{m \in \mathcal{M}} W_{m,t}$
  $\mathcal{O}^{m'} \leftarrow \{s \in \mathcal{O}_l; \forall l \in \mathcal{L} \mid Q_{s,l,m',t} > 0\}$
  `// Sélectionne parmi les lots passant sur la machine, celui dont la période de livraison est la plus éloignée`
  $s' \leftarrow \text{argmax}_{s \in \mathcal{O}^{m'}}(d_l)$
  for $s \in [s', ..., S_l]$ **do**
    if $X_{s,l,t} = 1$ **then**
      $X_{s,l,t} \leftarrow 0$
      $\mathcal{O}_{shifted} \leftarrow \mathcal{O}_{shifted} \cup \{s\}$
    for $m \in \mathcal{M}$ **do**
      $Q_{s,l,m,t} \leftarrow 0$
    end
  end
  for $m \in \mathcal{M}$ **do**
    $W_{m,t} \leftarrow \frac{\sum_{s \in \mathcal{S}_l}(a_{s,l,m} Q_{s,l,m,t})}{c_{m,t}}$
  end
end

L'algorithme 2 prend comme entrées la période, le plan de production et l'équilibrage sur les machines, c'est-à-dire les variables $X_{s,l,t}$ et $Q_{s,l,m,t}$. Tant qu'il reste des machines surchargées, l'algorithme va d'abord sélectionner la machine la plus surchargée. Ensuite, parmi les étapes de process qui lui sont assignées, l'algorithme sélectionne celle du lot ayant la priorité la plus faible. Cette priorité peut être évaluée de plusieurs façons, ce qui fait l'objet d'une étude proposée dans le chapitre 5. Dans l'algorithme 2 utilisé en exemple, la priorité des lots est définie en fonction de leur date de livraison. Ainsi, un lot ayant une due date proche de la date courante aura une priorité plus élevée qu'un lot ayant une due date plus éloignée, ce qui équivaut à la règle classique d'affectation *Earliest Due Date* (EDD).

Une fois que le lot le moins prioritaire (à une étape de process donnée) est sélectionné, il est reporté (ainsi que toutes les étapes de process suivantes du lot) à la période suivante et la charge de travail correspondante est retirée de toutes les machines auxquelles le lot a été affecté.

Pour évaluer la complexité de l'algorithme, analysons l'ensemble de ses composants :

- A chaque itération du module de *step-shifting*, il y a au moins une étape de process qui est repoussée à la période suivante. Il y a donc, au plus, autant d'itération qu'il y a d'étape dans la période, hypothétiquement toutes : $O(|\mathcal{S}|)$.
  - La machine la plus chargée doit être déterminée : $O(|\mathcal{M}|)$.
  - Pour cette machine, il faut déterminer parmi l'ensemble des étapes, celle correspondant au lot le moins prioritaire : $O(|\mathcal{S}|)$.
  - Une fois l'étape sélectionnée, elle est décalée, ainsi que tous ses successeurs initialement prévus dans la période (nombre majoré par $K \leq S$) : $O(\mathcal{K})$
  - Il faut finalement mettre à jour la charge de chaque machine : $O(|\mathcal{M}|)$.

On obtient donc pour le module de *step-shifting*, pour une période, une complexité globale $O(|\mathcal{S}|(|\mathcal{S}|+2|\mathcal{M}|+\mathcal{K}))$. Le terme le plus important est le nombre d'étapes dans une période donnée (au maximum $|\mathcal{S}|$, soit plusieurs dizaines de milliers), en comparaison du nombre de machines (plusieurs centaines) et le nombre maximum d'étapes dans une même période pour un lot (au maximum un millier). Nous retenons donc que l'algorithme de *step-shifting* est quadratique avec le nombre d'étapes de process à réaliser dans la période, soit $O(|\mathcal{S}|^2)$.

Le module de *step-shifting* se termine par un plan de production pour la période en cours qui est réalisable en termes de capacité. Il fournit également la liste des lots ayant été reportés (et à partir de quelle étape de process). Si les lots sont décalés, il est nécessaire de projeter à nouveau leurs étapes de process à partir de la période suivante. Nous utilisons donc à nouveau dans le module de projection qui prendra en entrée le plan initial, et où les nouvelles dates de début et de fin ne seront recalculées que pour les lots qui ont été décalés (affectation de nouvelles valeurs à $S_{s,l}$, $E_{s,l}$ et $X_{s,l,t}$).

Une fois la projection terminée, le nouveau plan de production est envoyé au module d'équilibrage et l'approche en trois étapes est répétée jusqu'à ce que toutes les périodes aient été traitées. Le résultat du module de *step-shifting* au cours de la dernière période est un plan de production à capacité finie qui vise à minimiser le retard des lots, en donnant pour chaque étape de process de chaque lot la période pendant laquelle il doit être traité, ainsi qu'une estimation de l'équilibrage de la charge de travail et l'impact sur les machines.

3.5 Performances de l'approche

Pour rappel, une évaluation expérimentale de la complexité du problème (section 3.3.2) a montré que le modèle mathématique n'était capable de trouver la solution optimale en seulement quelques heures que pour de très petites instances, jusqu'à 6 à 7 lots.

Contrairement au PLNE, l'approche TSH a l'avantage d'être très rapide, ne nécessitant en moyenne que 7 secondes pour résoudre de petits cas. Pour les cas réels, impliquant des milliers de lots, ayant chacun des centaines d'étapes de process sur une dizaine de semaines, l'approche permet toujours de proposer les plans en moins de 5 minutes. Cependant, il est important d'évaluer la qualité des solutions qui en résultent. Ainsi, une étude comparative a été menée entre les solutions optimales fournies par le modèle exact (ou la meilleure borne supérieure si le PLNE n'est pas résolu en moins d'une heure) et les solutions fournies par notre approche. Les résultats sont résumés dans le tableau 3.5. Pour chaque instance, nous présentons le *Total Weighted Tardiness* moyen sur 5 exécutions des différentes méthodes.

TABLE 3.5 – Comparaison entre les solutions déterminées par l'approche TSH et celles déterminées par le PLNE après une heure d'exécution (* si la solution optimale est obtenue).

Nb Jobs	1	2	3	4	5	6	7	8	9	10
MILP	0,1*	0,5*	4,5*	6,7*	7,3*	14,0*	18,0*	26,0	45,0	45,0
Heuristic	0,1	0,5	4,5	6,7	8,0	14,0	19,0	22,0	28,3	30,3
Gap (%)	0%	0%	0%	0%	10%	0%	6%	-15%	-37%	-33%
Nb Jobs	11	12	13	14	15	16	17	18	19	20
MILP	47,0	55,0	65,0	80,0	64,0	100,0	117,0	120,0	152,0	160,0
Heuristic	34,0	37,3	41,3	42,3	47,0	50,3	52,7	61,0	64,7	71,3
Gap (%)	-28%	-32%	-36%	-47%	-27%	-50%	-55%	-49%	-57%	-55%

D'après les résultats du tableau 3.5, on remarque que, dans le cas où le modèle exact détermine une solution optimale, notre approche atteint le même objectif dans presque tous les cas. Il est cependant à noter que notre approche détermine une solution de moins bonne qualité dans certains cas avec 5 et 7 lots. Ceci est en fait dû au comportement glouton de l'algorithme de lissage utilisé dans le module de *step-shifting*, ce qui est une faiblesse classique de ce type d'approches. Néanmoins, nous noterons que l'approche TSH trouve la solution optimale dans la plupart des cas et, dans les cas où le PLNE fournit seulement des bornes supérieures, des solutions toujours de meilleure qualité.

Pour plus d'informations sur l'approche heuristique et ses performances, une étude a également été menée par Mhiri et al. (2018). Les auteurs ont comparé les plans déterminés par l'approche et la production réelle dans l'usine. L'étude vise à évaluer la capacité de l'approche à simuler de façon fiable l'évolution du WIP dans l'usine.

Cependant, ce n'est pas le but de notre approche, qui est utilisée quotidiennement pour définir les objectifs de production. Les plans ne peuvent se contenter de prédire l'évolution de certains paramètres de l'usine, mais doivent prescrire la voie à suivre pour maximiser certains objectifs. Ainsi, dans les trois chapitres suivants de cette thèse, nous présentons les différents travaux menés afin d'améliorer la version initiale de l'approche TSH, non pas afin de mieux prédire l'évolution de certains paramètres, mais afin de fournir les meilleurs (selon différents critères de performances) plans de production possibles.

3.6 Conclusion

Dans ce chapitre, nous avons proposé un modèle mathématique afin de formaliser le problème de planification de production opérationnelle OPP. Nous avons par la suite présenté une analyse théorique puis expérimentale de sa complexité, montrant qu'il n'était pas envisageable de résoudre de façon exacte en des temps raisonnables des problèmes même de taille moyenne. Nous avons alors proposé une approche heuristique en trois étapes, appelée TSH, afin de résoudre ce problème. Cette approche itérative construit, période par période, d'abord une solution initiale potentiellement non faisable du fait de la capacité limitée des machines, et ensuite adapte cette solution pour déterminer un plan respectant cette capacité. Une étude expérimentale des performances (en temps de calcul) de l'heuristique a été présentée, montrant la grande rapidité de résolution dans le cas d'instances réelles. Une étude comparative entre le modèle exact et l'heuristique a également validé le bon fonctionnement de cette dernière.

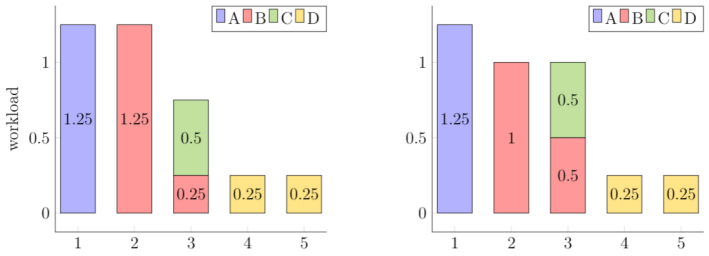
Dans les deux chapitres suivants, nous allons entrer plus en profondeur dans les modules d'*équilibrage* (chapitre 4) et de *step-shifting* (chapitre 5). Pour chaque module, nous décrirons leur fonctionnement et leur importance dans le fonctionnement global de l'heuristique. Nous présenterons les différents travaux menés sur chaque module, les ajouts et améliorations, et nous évaluerons l'impact de nos réalisations sur les performances de l'approche TSH, en comparaison de celle initialement présente dans le système de production.

Chapitre 4

Module d'équilibrage des charges

.

Le module de *projection* de l'approche TSH définit dans quelle période doit être réalisée chaque étape de process de chaque lot. Suite à ce premier module, l'étape suivante consiste à assigner la charge liée à ces étapes de process sur les différentes machines de l'usine, c'est le problème d'*équilibrage des charges*. Cette tâche n'est pas triviale du fait de l'hétérogénéité des machines en termes de qualifications et de durées opératoires. Le problème d'équilibrage des charges est très important dans la détection des machines limitante et pour orienter les décisions de management de la capacité. La version initiale de l'approche TSH a des difficultés dans l'accomplissement de ces tâches, et une nouvelle méthode, inspirée des problèmes de partage équitable des ressources, a été développée afin de fournir des solutions riches de sens pour les utilisateurs.



(a) $[X_{B,2}, X_{B,3}] = [10, 2]$ (b) $[X_{B,2}, X_{B,3}] = [8, 4]$

4.1 Introduction

Pour rappel, la réalisation d'une certaine étape de process d'un certain produit requiert une configuration particulière que l'on appelle *recette*. Si une machine est capable de réaliser une étape de process donnée d'un produit donné, on dit qu'elle est qualifiée pour réaliser la recette associée. Dans le chapitre 1, nous avons rappelé que, dans l'industrie des semi-conducteurs, et notamment dans les usines *front end*, les machines d'un même atelier ont souvent des qualifications différentes Johnzén et al. (2011); Rowshannahad et al. (2015). Cela signifie que, pour une recette donnée, seule une partie des machines de l'usine est capable (qualifiée) pour réaliser cette recette. Inversement, chaque machine n'est qualifiée que pour exécuter un certain nombre de recettes différentes. En plus de cette variété des qualifications, le temps de process d'une plaquette donnée peut varier d'une machine qualifiée à une autre. Par conséquent, l'allocation optimale des étapes de process sur les machines (que l'on nomme équilibrage des charges dans ce chapitre), pour un critère donné, n'est généralement pas un problème trivial. Ce problème d'équilibrage est d'autant plus important dans les usines *front end* de type *High-Mix Low-Volume*, où le choix d'allocation peut grandement affecter la charge globale. Dans le contexte de problèmes de planification tels que le *Production Planning* ou le *Capacity Planning*, une mauvaise évaluation de la charge des machines peut par exemple amener à une mauvaise détection des ressources limitantes, et donc ne pas orienter les managers vers les bonnes décisions afin de gérer les flux de production.

Dans les travaux précédents autour du problème d*Operational Production Planning*, l'hétérogénéité des machines en termes de vitesse de réalisation d'une même recette est rarement considérée. Cela permet de simplifier le problème en supprimant le choix de l'allocation des quantités sur les machines, mais comme nous l'avons dit précédemment cela se fait au détriment d'une évaluation fiable des charges des machines.

Dans cette thèse, que ce soit dans le modèle mathématique sous forme de PLNE ou l'approche TSH, tous deux présentés dans le chapitre 3, ce choix de l'allocation des étapes de process sur les machines est considéré. Dans l'approche TSH, cette allocation est réalisée dans le module d'*équilibrage*, introduit dans le chapitre 3, et auquel ce chapitre est consacré.

Le présent chapitre reprend les travaux présentés dans Christ et al. (2019), et est structuré comme suit. Le problème d'équilibrage des charges dans les systèmes de fabrication flexibles est défini et motivé dans la section 4.2. La section 4.3 traite du module d'*équilibrage* initialement implémenté dans l'approche heuristique développée par Mhiri (2016), ainsi que ses limites. Dans la section 4.4, nous rappelons le concept de *Min-Max Fairness* ainsi que la littérature associée, et nous présentons son application à notre problème d'équilibrage des charges. Nous introduisons une nouvelle procédure nommée *Iterated Min-Max* (IMM) dans la section 4.5, et nous montrons qu'elle fournit des solutions plus appropriées aux attentes industrielles. Les résultats de tests numériques sur des instances industrielles sont discutés dans la section 4.6 et la section 4.7 propose un bilan de l'approche développée.

4.2 Définition du problème

4.2.1 Problème d'équilibrage des charges dans l'approche globale

Avant d'entamer cette section, nous souhaitons préciser certains points permettant de mieux comprendre l'intégration du module d'équilibrage des charges dans l'approche TSH.

Tout d'abord, le module de projection détermine pour chaque période, les étapes de

process devant être réalisées pour chaque lot. Chaque lot, comprenant jusqu'à 25 plaquettes, correspond à un produit donné, parmi des dizaines possibles. La réalisation d'une certaine étape de process d'un certain produit requiert une configuration particulière que l'on appelle *recette*. Les recettes sont essentielles au problème d'équilibrage des charges, étant donné que les machines sont qualifiées pour réaliser des recettes, c'est à dire des étapes de process données pour des produits donnés. Ainsi, les quantités à répartir sur les machines dans notre problème d'équilibrage des charges, ne sont pas regroupées par produit, mais par recette. Notons également, qu'alors que le nombre de produits différents dans une usine *front end* (*High-Mix Low-Volume*) est généralement de quelques dizaines, le nombre de recettes différentes s'élève à plusieurs milliers.

Le deuxième point, qui a déjà été abordé dans la section 3.4.3 du chapitre 3, concerne la taille des instances pour chaque problème d'équilibrage des charges. En effet, nous venons de mentionner que dans une usine *front end* il existait plusieurs milliers de recettes différentes, et nous savons également que ces usines abritent plusieurs centaines de machines. Cependant, pour résoudre le problème d'équilibrage, qui est modélisé sous forme d'un Programme Linéaire (PL), il n'est pas nécessaire d'utiliser un seul modèle avec l'ensemble des recettes sur l'ensemble des machines. En effet, nous avions introduit dans la section 3.4.3 du chapitre 3 le concept d'*Isolated Group* (IG), un ensemble de machines partageant des qualifications de recettes, de telle sorte que deux machines situées dans des IG différents ne peuvent s'influencer dans leur équilibrage des charges. Un IG est généralement constitué de 1 à 20 machines, et peut être résolu indépendamment des autres. Par conséquent, la résolution d'un PL considérant l'ensemble des machines est remplacée par la résolution, en parallèle, de plusieurs dizaines (généralement entre 30 et 50) de PL plus petits.

Enfin, cette résolution de différents PL est à répéter pour chaque période de l'horizon. En effet, comme cela a été expliqué dans la section 3.4 du chapitre 3, le module de *projection* projette l'évolution des lots sur l'ensemble de l'horizon de planification. Puis, le module d'*équilibrage* répartit la charge liée aux étapes de process sur les machines, pour la *première période*. Ensuite le module de *step-shifting* déplace certains lots de cette période afin de respecter la capacité. Puis le module de *projection* est de nouveau exécuté à partir de la période suivante, suivi du module d'*équilibrage*, etc. Ainsi, le module d'*équilibrage* est exécuté pour chaque période, ce qui signifie que, pour une planification de 10 périodes (que ce soit des périodes d'une journée, une semaine, ou un mois) par exemple, ce sont plusieurs centaines de PL qui doivent être résolus. Cette information est très importante, car elle justifie l'obligation de développer une approche suffisamment rapide pour résoudre des centaines de problèmes d'équilibrage dans un outil de planification devant fournir des plans de production en seulement quelques minutes.

4.2.2 Modélisation du problème

Dans la suite de ce chapitre, nous étudions le problème d'équilibrage des charges pour un *Isolated Group*.

On considère un ensemble de recettes $\mathcal{R} = \{1, \ldots, R\}$, et des quantités $q_r \in \mathbb{R}^+$ pour chaque recette $p \in \mathcal{R}$. On dispose d'un ensemble de machines $\mathcal{M} = \{1, \ldots, M\}$, et chaque recette r peut être réalisée sur un sous-ensemble de machines $\mathcal{M}_r \subseteq \mathcal{M}$, avec un temps de process $a_{r,m}$ par plaquette de la recette r sur la machine $m \in \mathcal{M}_r$. Chaque machine m a une capacité c_m strictement positive. On définit $Q_{r,m}$ la quantité de plaquette de la recette r allouée à la machine $m \in \mathcal{M}_r$. Le taux de charge de la machine m est alors définie comme:

$$W_m = \frac{\sum_{r \in \mathcal{R}} a_{r,m} X_{r,m}}{c_m} \tag{4.1}$$

Notons que le taux de charge W_m tient compte de la capacité de la machine m. La détermination d'un ensemble $X = \{X_{r,m};\ \forall (r,m) \in \mathcal{R} \times \mathcal{M}_r\}$ constitue une solution d'équilibrage des charges. L'objectif est donc de définir, pour chaque machine, la quantité de chaque recette à traiter afin d'optimiser un certain objectif. Ce qui peut être modélisé par le PL suivant:

$$(P) \quad \min \ f(X) \tag{4.2}$$

$$\sum_{m \in \mathcal{M}_r} X_{r,m} = q_r \quad r = 1, \ldots, \mathcal{R} \tag{4.3}$$

$$X_{r,m} \in \mathbb{R}^+ \qquad r = 1, \ldots, \mathcal{R},\ m = 1, \ldots, \mathcal{M}_r \tag{4.4}$$

La fonction objectif $f(.)$ prend en entrée une solution d'équilibrage des charges, et dépend des critères que nous souhaitons optimiser. Les contraintes (4.3) assurent que, pour chaque recette r, la quantité q_r est assignée aux machines dans $\mathcal{M}_r$. La préemption est autorisée étant donné que les variables $Q_{r,m}$ sont continues. Notons que la capacité de chaque machine est considérée dans la définition du taux de charge W_m, mais pas en tant que contrainte à ne pas dépasser. Si la capacité de la machine est dépassée, le taux de charge associé sera strictement supérieure à 1.

Soulignons encore une fois qu'il est important que le temps de résolution de (P) soit très court, puisque des centaines de problèmes (voire les tests expérimentaux sur les données industrielles dans la section 4.6) sont résolus pour chaque exécution de l'approche TSH. C'est pourquoi la fonction $f(.)$ est généralement linéaire, et les variables de décision $Q_{r,m}$ sont dans $\mathbb{R}^+$.

4.3 Le modèle initial et ses limites

Dans le modèle initialement implémenté dans l'outil de planification opérationnelle, la fonction objectif était composée de trois termes, chacun pondéré par un poids (α, β, γ). La fonction objectif $f_c(.)$ du problème (P) était de la forme suivante:

$$f_c(X) = \alpha \max_{m \in \mathcal{M}} W_m - \beta \min_{m \in \mathcal{M}} W_m + \gamma \sum_{m \in \mathcal{M}_r} c_m W_m$$

$$= \alpha \max_{m \in \mathcal{M}} \frac{\sum_{r \in \mathcal{R}} a_{r,m} X_{r,m}}{c_m} - \beta \min_{m \in \mathcal{M}} \frac{\sum_{r \in \mathcal{R}} a_{r,m} X_{r,m}}{c_m} + \gamma \sum_{r \in \mathcal{R}} \sum_{m \in \mathcal{M}_r} a_{r,m} X_{r,m} \tag{4.5}$$

Le but est d'optimiser une combinaison de trois termes, à savoir, pour les deux premiers termes, minimiser le taux de charge (le rapport entre le temps de traitement total assigné à la machine et sa capacité) de la machine la plus chargée, puis maximiser le taux de charge de la machine la moins chargée, afin de réduire la dispersion des charges. Enfin, le troisième terme a pour but de minimiser le temps total de process, et donc vise en particulier à faire passer les lots sur les machines les plus rapides. Les coefficients (α, β, γ) pondèrent les différents termes de la fonction objectif. Dans l'implémentation initiale de l'approche, les coefficients

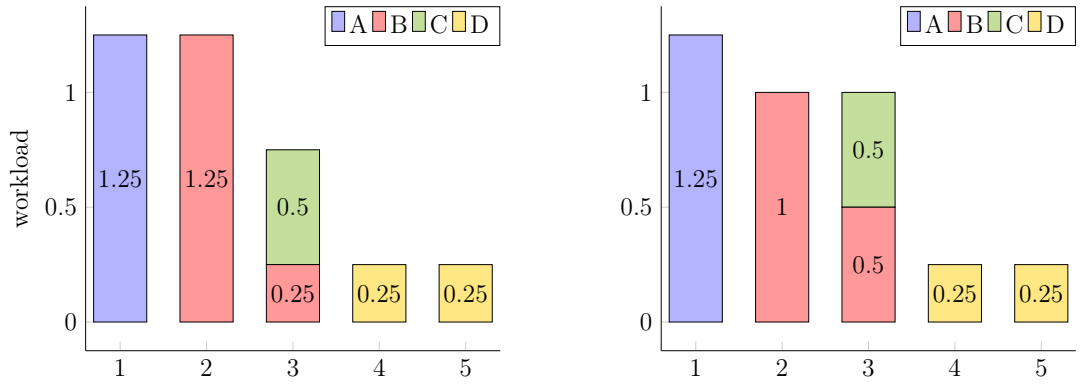
étaient choisis tels que $\alpha >> \beta >> \gamma$, privilégiant ainsi dans l'ordre la minimisation du taux de charge maximum, la maximisation du taux de charge minimum, puis la minimisation du temps de process global. Une autre variante de cette fonction objectif, également assez naturelle, consiste à choisir les coefficients $\alpha >> \gamma >> \beta$, afin de privilégier la minimisation du temps de process global par rapport à la maximisation du taux de charge de la machine la moins chargée. Cette variante sera également étudiée dans la section 4.6. Bien qu'assez naturelle, la fonction objectif initiale (et plus largement ce modèle mathématique d'allocation des charges) a certaines limites, que nous présentons dans la section suivante.

4.3.1 Limites et illustration

Le principal objectif du problème d'OPP, que nous avons présenté dans les précédents chapitres, est la définition d'un plan de production, définissant les périodes où exécuter chaque étape de chaque lot, afin d'orienter les outils d'ordonnancement dans l'usine. Cependant, l'outil de planification utilisé en production, et basé sur l'approche TSH, est aussi utilisé pour d'autres fonctions. Parmi elles, l'outil de planification permet d'effectuer une évaluation prévisionnelle du taux de charge des machines dans les périodes à venir. En effet, le module d'*équilibrage* permet de répartir la charge associée aux étapes de process devant passer dans chaque période (les valeurs des variables $Q_{s,l,t}$ déterminées par le module de *projection*), sur les différentes machines. Cette répartition n'est pas limitée par la capacité des machines, mais nous avons vu dans la section 4.2 que si cette charge W_m est supérieure à 1, cela informe que la machine est surchargée. Lorsqu'une solution est fournie par l'approche TSH, les utilisateurs regardent cette charge prévisionnelle afin de détecter les machines limitantes ($W_m > 1$) ou proches de l'être (W_m proche de 1). Le chapitre 6 détaille les utilisations de l'outil de planification et montre certaines interfaces permettant de répondre à ce type de questions sur le taux de charge des machines.

Dans ce chapitre, nous illustrons ces problématiques d'équilibrage des charges au travers d'une instance fictive de petite taille. Cette instance est composée de 5 machines et de 4 recettes A, B, C et D. Afin de simplifier la compréhension du problème, nous supposons que le temps de process est identique pour chaque machine ($a_{r,m} = 1$, $\forall (r,m) \in \mathcal{R} \times \mathcal{M}_r$), ainsi que la capacité ($c_m = 8$, $\forall m \in \mathcal{M}_r$). De plus, afin de normaliser le problème, nous assignons les valeurs $(\alpha, \beta, \gamma) = (1; 1; 0,01)$ aux facteurs pondérateurs. Les quantités à traiter pour chaque recette sont $\{q_A, q_B, q_C, q_D\} = \{10; 12; 4; 4\}$. Toutes les machines ne sont pas qualifiées pour toutes les recettes. Les machines 4 et 5 ne peuvent réaliser que la recette D, les machines 2 et 3 peuvent réaliser les recettes B, C et D, tandis que la machine 1 est qualifiée pour toutes les recettes.

La figure 4.1 présente deux solutions possibles d'équilibrage des charges. Les deux solutions ne diffèrent que par l'affectation de la recette B sur les machines 2 et 3. Dans la solution (a), la machine 2 prend 10 unités de la recette B pour un temps de traitement total de 10 heures, et est donc équilibrée avec la machine 1. En revanche, dans la solution (b), 2 unités de la recette B sont déplacées de la machine 2 vers la machine 3, pour équilibrer le taux de charge entre les machines 2 et 3. Les deux solutions conduisent à la même valeur pour la fonction objectif. En effet, dans les deux cas, le taux de charge maximum est fixé par la machine 1 qui, étant la seule à pouvoir traiter la recette A, a un taux de charge égal à $(a_{A,1}X_{A,1})/c_1 = (1 \times 10)/8 = 1,25$. De l'autre côté, les machines 4 et 5 ne sont qualifiées que pour traiter la recette D. Chaque machine prend 2 unités de la recette D, ce qui conduit au taux de charge minimum $(1 \times 2)/8 = 0,25$. Enfin, comme le temps de process est le même pour toutes les machines, le troisième terme ne dépend pas des quantités allouées et est égal



(a) $[X_{B,2}, X_{B,3}] = [10, 2]$ (b) $[X_{B,2}, X_{B,3}] = [8, 4]$

FIGURE 4.1 – Deux solutions "équivalentes" d'équilibrage des charges

à $(q_A + q_B + q_B + q_C + q_D) = (10 + 12 + 4 + 4) = 30$. Par conséquent, la fonction objectif des deux solutions (a) et (b) est égale à $f_c = \alpha \times 1,25 + \beta \times 0,25 + \gamma \times 30 = 1,80$. Toutefois, les deux solutions ne fournissent pas les mêmes informations à l'utilisateur et la solution (b) fournit en réalité des informations plus pertinentes que la solution (a).

En effet, si l'on regarde l'allocation de la solution (a), les machines 1 et 2 semblent avoir un taux de charge de 1,25, ce qui signifie que ces machines sont critiques et doivent être analysées. Sur la base de ces informations, les utilisateurs pourraient être tentés de prendre des mesures telles que, par exemple, retarder les opérations de maintenance préventive pour fournir une capacité supplémentaire à la ressource temporairement surchargée. Cependant, la solution (b) montre que seule la machine 1 est réellement critique car la machine 2 peut être équilibrée avec la machine 3. Par conséquent, il serait inutile et coûteux de fournir une capacité supplémentaire à la machine 2, car cela n'empêcherait pas la machine 1 de limiter les flux. À l'inverse, comme le taux de charge de la machine 3 dans la solution (a) est égal à 0,75, les utilisateurs pourraient conclure que la machine 3 ne nécessite pas d'attention particulière, et même qu'une perte de productivité (c'est-à-dire une augmentation du temps de process $a_{r,m}$ ou une diminution du paramètre de capacité c_m), pour la machine 3, ne serait pas critique. Toutefois, la solution (b) souligne le fait que la machine 3 est en fait importante car elle peut s'équilibrer avec la machine 2, étant donné que les deux machines sont qualifiées pour traiter le produit B. Par conséquent, une perte de productivité sur la machine 3 entraînerait un taux de charge supérieur à 1 pour les machines 2 et 3, ce qui signifie qu'elles seraient surchargées et ne seraient donc pas capables de traiter le plan de production.

Cet exemple illustre les multiples risques d'une prévision inexacte des machines critiques ou sous-chargées. Des décisions inutiles pourraient être prises, et l'importance de certaines machines pourrait être sous-estimée. En outre, certaines solutions d'équilibrage des charges peuvent ne pas mettre en évidence les interactions pertinentes entre les machines, comme la solution (a) pour les machines 2 et 3. Ces problèmes ont été observés dans les solutions fournies par la version initiale de l'outil de planification, dont le module d'*équilibrage* était basé sur la fonction (4.5) définie précédemment, et qui est assez similaire à celle présentée dans Mhiri et al. (2018). Lorsque les utilisateurs analysaient en détail les résultats fournis par cette version initiale de l'outil de planification, ces derniers se plaignaient parfois de ne pas comprendre la répartition de la charge sur certaines machines. Par exemple, avec des cas évidents de mauvais équilibrage comme dans la solution (a), il est clair pour les utilisateurs

qu'il est possible de transférer de la charge liée à la recette B de la machine 2 vers la machine 3. De plus, le fait de proposer des cas d'équilibrage clairement sous-optimaux amenait les utilisateurs à douter de la validité de l'outil de planification.

Afin de différencier les solutions (a) et (b) par la fonction objectif, et notamment privilégier (b), une idée serait d'ajouter d'autres termes à la fonction objectif (comme cela est d'ailleurs fait dans Mhiri et al. (2018)). Cependant, l'ajout de nouveaux termes n'est pas nécessairement une bonne alternative, en particulier parce que l'utilisation d'une fonction objectif linéaire combinant différents critères entraîne des difficultés dans le réglage des poids, ainsi qu'une perte de clarté. C'est pourquoi nous avons décidé de développer une nouvelle approche, plus pertinente.

4.4 Le problème *Min-Max Fairness Workload Balancing* (MMFWB)

Dans cette section, nous introduisons le problème, très étudié, de *Min-Max Fairness* (MMF), en particulier dans divers domaines autour des problématiques de routage dans les réseaux de communication. Puis, nous expliquons l'intérêt d'étendre ce concept à notre problème d'équilibrage des charges pour les systèmes de fabrication, en dégageant plusieurs propriétés des solutions fournies.

4.4.1 Le problème de *Min-Max Fairness*

D'une manière générale, le problème de *min-max fairness* est applicable dans les situations où il est souhaitable de proposer une répartition équitable de certaines ressources, partagée par des demandes concurrentes Nace and Pióro (2008). Intuitivement, une solution *min-max fair* (respectivement *max-min fair*) est une solution où la diminution (respectivement l'augmentation) des ressources allouées à une demande entraîne nécessairement une augmentation (respectivement une diminution) des ressources allouées à des demandes déjà plus importantes (respectivement plus faibles) ou également allouées. Il est à noter que le problème de *min-max fairness* (respectivement *min-max*) était à l'origine défini comme le problème lexicographique minimax (respectivement maximin) et que certains articles utilisent cette formulation. Bien que ces termes ne soient équivalents que dans le cas d'ensembles convexes atteignables Radunovic and Le Boudec (2007), comme cela est le cas pour notre problème, nous faisons référence dans le reste de cette section à la fois aux travaux liés au problème de *min-max fairness* et aux problèmes lexicographique minimax. De plus, dans la suite de cet article, nous utilisons l'acronyme MMF pour mentionner à la fois les problèmes *Min-Max Fairness* et *Max-Min Fairness*.

Le concept de MMF a été largement étudié dans divers contextes, notamment pour des problèmes de réseaux de communication Bertsekas et al. (1987); Radunovic and Le Boudec (2007); Nace and Pióro (2008); Yaakob and Khalil (2016); Sadeghi et al. (2018); Zhu et al. (2018). Pour plus d'informations sur les applications du MMF dans les problèmes de réseaux, les lecteurs sont invités à consulter Ogryczak et al. (2014). Le concept de MMF a également été appliqué dans d'autres domaines tels que les services publics pour l'allocation équitable des ressources en eau Wang et al. (2008) ou dans les problèmes de transport aérien comme dans Murça (2018) pour une gestion équitable des flux de trafic aérien. Plus récemment, Qi (2016) a mis au point un nouvel indicateur de performance pour des problèmes d'affectation dans le contexte de cliniques ambulatoires pour décrire l'insatisfaction des médecins et

des patients. Elle utilise ensuite une approche lexicographique minimax pour améliorer la conception du système de prise de rendez-vous.

Considérant les applications aux problèmes de fabrication, Luss and Smith (1986) ont proposé un algorithme en temps polynomial qui peut être utilisé en planification de la production pour équilibrer l'écart pondéré par rapport aux demandes de produits. Tang (1988) étudie l'application du MMF pour résoudre des problèmes de MRP afin de minimiser le coût des pénalités lorsque les demandes ne sont pas satisfaites. Il a également présenté une application du MMF afin de déterminer quand produire afin de maximiser le temps entre deux ordres de production. King (1989) proposent une application industrielle de l'algorithme de Luss and Smith (1986). Il utilise une procédure de minimisation lexicographique pour développer un outil d'aide à la décision afin d'aider les planificateurs à choisir des plans de production alternatifs lorsque le plan de production initial devient irréalisable en raison de la variabilité du système de fabrication, comme des pannes de machines ou des changements de commandes clients. Il a étendu l'algorithme initial à un problème de production multi-périodes, complété par une heuristique pour garantir des valeurs de production entières. Luss (1999) examine divers problèmes d'affectation des ressources en utilisant l'approche lexicographique minimax et a souligné l'intérêt d'utiliser cette méthode dans la planification de la production. Plus récemment et à une plus grande échelle, Liu and Papageorgiou (2013) exploitent la méthode lexicographique minimax pour résoudre l'optimisation multi-objectif de la chaîne logistique. Enfin, Liu and Papageorgiou (2018) ont également recours à une approche MMF pour équilibrer équitablement le profit entre les différents acteurs d'une chaîne d'approvisionnement à trois niveaux. À notre connaissance, il n'y a aucune référence d'application du problème de *min-max fairness* pour l'équilibrage de la charge sur des machines dans un système de fabrication.

Plusieurs définitions ont été proposées pour caractériser une solution *min-max fair*. Dans ce chapitre, nous utilisons celle utilisée dans Radunovic and Le Boudec (2007). Considérons un ensemble $\chi \subset \mathbb{R}^N$ $(N \in \mathbb{N})$ et un vecteur $x \in \chi$. Le vecteur x est dit *min-max fair* si et seulement si :

$$\forall\, y \in \chi \quad \exists s \in (1,\dots,N) \quad y_s < x_s \implies \exists\, t \in (1,\dots,N) \quad s.t. \quad y_t > x_t \geq x_s \tag{4.6}$$

Cela signifie que diminuer la valeur de Q_s implique nécessairement l'augmentation d'un autre élément Q_t d'une valeur égale ou supérieure.

Pour relier cette définition à notre problème d'équilibrage des charges, nous considérons $x \in \mathbb{R}^M$ comme une solution d'équilibrage des charges où chaque composant y_m est le taux de charge assigné à la machine m, et $\chi \subset \mathbb{R}^M$ l'ensemble des allocations possibles. Ensuite, x est une solution *min-max fair* s'il n'est pas possible de réduire le taux de charge sur une machine sans augmenter celui d'une autre machine déjà plus ou également chargée.

Nous définissons la recherche de la solution *min-max fair* pour notre problème d'équilibrage des charges comme le problème *Min-Max Fair Workload Balancing* (MMFWB) et définissons une solution optimale de ce problème comme une solution *min-max fair* d'équilibrage des charges.

Si l'on reprend l'exemple de la figure 4.1, l'équilibrage des charges proposé dans la solution (a) n'est pas une solution optimale pour le problème MMFWB, c'est-à-dire que la solution du problème d'équilibrage des charges n'est pas *min-max fair*. En effet, il est possible de diminuer le taux de charge de la machine 2 sans augmenter celle de la machine 1. Cette réaffectation ne fait qu'augmenter le taux de charge de la machine 3 qui est initialement moins chargée. En revanche, il ne semble pas possible de procéder à une telle réduction du

taux de charge dans la solution (b). En fait, cette solution est *min-max fair*, ce que nous montrerons dans la section 4.5.

4.4.2 Propriétés des solutions du problème MMFWB

Dans cette section, nous présentons certaines propriétés des solutions optimales du problème MMFWB, dérivées directement de la structure des solutions *min-max fair*.

Détection des machines critiques

Tout d'abord, rappelons la définition utilisée pour caractériser les machines critiques.

Definition 4.4.1. Une machine est dite **critique** s'il est impossible de réduire son taux de charge sans augmenter le taux de charge d'une autre machine ayant un taux de charge supérieur ou égal.

De cette définition, nous pouvons énoncer la proposition suivante.

Proposition 2. *Dans une solution optimale du problème MMFWB, toute machine est critique.*

Cette propriété est directement dérivée de la définition d'une solution *min-max fair*, car il est impossible de réduire un composant sans augmenter un autre composant qui est déjà supérieur ou égal. Ainsi, dans une solution du problème MMFWB, pour toute machine, il n'est pas possible de réduire son taux de charge sans augmenter celui d'une autre machine ayant un taux de charge équivalent ou supérieur. Grâce à cette propriété, les utilisateurs peuvent déterminer avec plus de précision les machines critiques (et en particulier celles dont le taux de charge est le plus important), et donc mieux planifier les actions préventives.

Regroupement des machines équilibrées

Ici nous présentons une propriété pour les machines pouvant s'équilibrer entre elles.

Proposition 3. *Dans une solution optimale du problème MMFWB, si deux machines* $m1$ *et* $m2$ *traitent au moins une recette en commun, c'est-à-dire* $\exists r$ *tel que* $Q_{r,m1} > 0$ *et* $Q_{r,m2} > 0$*, alors elles ont le même taux de charge, soit :* $W_{m1} = \frac{\sum_{r\in\mathcal{R};\ m1\in\mathcal{M}_r} a_{r,m1}X_{r,m1}}{c_{m1}} = W_{m2} = \frac{\sum_{r\in\mathcal{R};\ m2\in\mathcal{M}_r} a_{r,m2}X_{r,m2}}{c_{m2}}$.

La preuve peut être faite via un raisonnement par l'absurde. Considérons une solution optimale du problème MMFWB avec deux machines $m1$ et $m2$, ayant des taux de charge différents, et qui traitent au moins une recette en commun r. Supposons que la machine $m1$ est plus chargée que la machine $m2$. Il est toujours possible de transférer une partie de la charge de $m1$ à celle de $m2$ en réduisant $Q_{r,m1}$ et en augmentant $Q_{r,m2}$. Ainsi, il est possible de réduire la charge d'une machine sans augmenter le taux de charge d'une machine dont le taux de charge est supérieur ou égal, ce qui contredit la propriété fondamentale des solutions optimales du problème MMFWB.

En raison de la proposition 3, dans une solution optimale du problème MMFWB, il n'y a donc pas de déséquilibre possible entre les machines où une charge peut être transférée d'une machine chargée à une autre ayant un taux de charge moindre. Dans l'exemple illustratif de la section 4.3.1, la solution (a) de la figure 4.1 n'est donc clairement pas une solution *min-max fair*, alors que la solution (b) semble l'être. Le fait de fournir des solutions *min-max fair* donne plus de confiance aux planificateurs.

Existence et unicité

Dans leur livre, Bertsekas et al. (1987) énoncent une propriété importante pour des vecteurs *min-max fair*.

Lemma 1. *S'il existe un vecteur min-max fair pour un ensemble donné, alors il est unique.*

Ensuite, en plus de fournir un cadre général pour définir les problèmes MMF, Radunovic and Le Boudec (2007) soulignent une condition suffisante d'existence d'un vecteur *min-max fair*, dont une version simplifiée est donnée ci-dessous.

Lemma 2. *Si l'ensemble $\chi \subset \mathbb{R}^N$ est convexe et compact, alors il existe un vecteur min-max fair sur cet ensemble.*

À partir de ces deux propriétés, nous pouvons en déduire la proposition suivante:

Proposition 4. *Pour toute instance d'un problème MMFWB, il existe une solution optimale à ce problème, et elle est unique.*

Démonstration. L'idée est de prouver que, pour toute instance du problème MMFWB, l'ensemble des configurations possibles du taux de charge des machines $\chi \subset \mathbb{R}^M$, est toujours compact et convexe. Cependant, cette propriété n'est pas évidente dans notre cas, et nous devons d'abord définir l'ensemble des solutions réalisables pour équilibrer la charge de travail ψ du problème (P).

Une solution est définie par un vecteur :

$$X \in \mathbb{R}^{+(R\times M)} = \{\dots, X_{r,m}, \dots\}$$

qui résume les quantités de recettes affectées aux machines. Définissons ensuite l'ensemble des solutions réalisables $\psi \subset R^{+(R\times M)}$, c'est-à-dire l'ensemble des solutions qui satisfont la réalisation des quantités de recettes et les contraintes de qualification :

$$\begin{aligned}\psi = \{\ & X \in \mathbb{R}^{+(R\times M)} \\ & \text{s.t.} \quad \forall r \in \mathcal{P}, \sum_{m\in\mathcal{M}_r} X_{r,m} = q_r \quad \wedge \quad \forall r \in \mathcal{P}, \forall m \notin \mathcal{M}_r, X_{r,m} = 0\}\end{aligned} \tag{4.7}$$

Comme aucune relation d'inégalité stricte n'est utilisée et que l'intersection d'ensembles fermés est fermée, nous pouvons affirmer que l'ensemble ψ est fermé. Puisque $\sup(\psi) = \max_{r\in\mathcal{P}} q_r$, l'ensemble ψ est aussi fermé. Or, d'après le théorème de Borel-Lebesgue, dans une topologie $\mathbb{R}^N$, tous les ensembles fermés et bornés sont compacts. Par conséquent ψ est compact. De plus, l'application ψ est convexe. En effet, si $(x, y) \in \psi^2$ alors, tout vecteur $z = \lambda x + (1-\lambda)y$ avec $\lambda \in [0, 1]$ est aussi dans ψ.

Nous définissons l'application ϕ prenant en entrée un vecteur d'allocation des quantités de recettes, et générant le vecteur de charges γ indiquant le taux de charge résultant pour chaque machine :

$$\phi : \qquad \psi \subset R^{+(R\times M)} \qquad \rightarrow \qquad \phi(\psi) \subset \mathbb{R}^{+M}$$

$$X = \{\dots, X_{r,m}, \dots\} \ \mapsto \ \gamma = \{\dots, \textstyle\sum_{r\in\mathcal{P};\ m\in\mathcal{M}_r} \frac{a_{r,m}}{c_m} X_{r,m}, \dots\}$$

L'application ϕ est linéaire car $\forall(\lambda, \mu) \in \mathbb{R}^2$ et $\forall(x, y) \in \chi$, $\phi(\lambda x + \mu y) = \lambda\phi(x) + \mu\phi(y)$. L'ensemble χ est défini par : $\chi = \phi(\psi)$. L'ensemble χ représente donc l'ensemble

des configurations possibles de taux de charge des machines pour un problème MMFWB donné. Comme l'ensemble ψ est convexe et l'application ϕ est linéaire, χ est aussi convexe. De plus, $\mathbb{R}^{+(R\times M)}$ et $\mathbb{R}^{+M}$ sont des espaces de dimension finie, l'application ϕ préserve donc la compacité. Comme ψ est compact, alors χ est aussi compact.

On en conclut donc que l'ensemble χ est convexe et compact. À partir du lemme 2, nous en concluons donc qu'il existe pour toute problème MMFWB un vecteur *min-max fair* (donc une solution optimale), et d'après le lemme 1 que celui-ci est unique.

□

Le résultat important de la proposition 4 garantit que, pour tout problème MMFWB, il est possible de trouver une solution *min-max fair* qui satisfait les propriétés présentées dans cette section. Elle garantit également que la solution optimale du problème MMFWB est unique, contrairement au modèle initial de la section 4.3, pour lequel plusieurs solutions optimales peuvent exister.

4.5 La méthode *Iterated Min-Max* (IMM)

Pour le moment, nous avons dégagé certaines propriétés des solutions *min-max fair* et montré l'intérêt de leur utilisation dans le cadre du problème d'équilibrage des charges de machines dans un système de fabrication. La prochaine étape est désormais de définir une méthode permettant de déterminer, pour tout problème d'équilibrage des charges, la seule et unique solution *min-max fair*. Ainsi, dans cette section, nous présentons une méthode pour déterminer les solutions optimales du problème MMFWB. La procédure Iterated Min-Max (IMM) est basée sur la résolution itérative d'une version réduite du programme linéaire du problème d'équilibrage et sur l'utilisation du théorème des écarts complémentaires, qui est une propriété importante de la théorie de la programmation linéaire. Nous prouvons que la procédure IMM détermine la solution optimale du problème MMFWB, et nous illustrons sont fonctionnement à l'aide de notre exemple.

4.5.1 Description

Afin de présenter la procédure IMM, nous réécrivons le problème d'équilibrage de la façon suivante. Soit $\mathcal{B} \subseteq \mathcal{M}$ un sous-ensemble de machines, et soit $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_m\} \in \mathbb{R}^M$ un vecteur de charge où γ_m est le taux de charge de la machine m. Ainsi, nous pouvons écrire le programme linéaire $P(M, B, \gamma)$:

$$
\begin{aligned}
P(M,B,\gamma) \quad & min && S && (4.8)\\
s.c. \quad & W_m \leq S && \forall m \in \mathcal{M} \setminus \mathcal{B} && (4.9)\\
& W_m = \gamma_m && \forall m \in \mathcal{B} && (4.10)\\
& W_m = \sum_{r\in\mathcal{R};\ m\in\mathcal{M}_r} \frac{a_{r,m}}{c_m} X_{r,m} && \forall m \in \mathcal{M} && (4.11)\\
& \sum_{m\in\mathcal{M}_r} X_{r,m} = q_r && \forall r \in \mathcal{R} && (4.12)\\
& X_{r,m} \geq 0 && \forall r \in \mathcal{R},\ \forall m \in \mathcal{M}_r && (4.13)
\end{aligned}
$$

Les contraintes (4.12) et (4.13) sont respectivement les contraintes de quantité et de non-négativité. L'objectif est de minimiser le taux de charge d'un sous-ensemble de machines, tandis que le taux de charge des autres machines est fixe. La variable S définit le taux de charge de la (ou des) machine(s) la plus chargée dans $\mathcal{M} \setminus \mathcal{B}$ et est déterminée par les contraintes (4.9). Notez que S est positif puisque les variables W_m sont positives. Les contraintes (4.10) fixent le taux de charge des machines dans $\mathcal{B}$. Les contraintes (4.11) définissent comment la charge de chaque recette est affectée aux machines.

Au début de l'IMM, $\mathcal{B} = \emptyset$ et la procédure minimise le taux de charge maximum parmi l'ensemble des machines. Ensuite, le vecteur de charges γ est construit itérativement, de sorte qu'à la fin de la procédure, le taux de charge affecté à chaque machine dans γ est une solution du problème d'équilibrage. Définissons λ_m comme étant la variable duale associée à la contrainte de charge (4.9) de la machine m. L'algorithme 3 résume la procédure IMM.

Algorithme 3 : Procédure de la méthode IMM

Data : Une instance de M machines et R recettes
Result : Une solution γ au problème MMFWB

```
1  B := ∅, γ_m = 0 ∀m ∈ M, S* = 0;
2  while B ≠ M do
3  |   Résoudre P(M, B, γ) et déterminer S*;
4  |   for m ∈ M \ B do
5  |   |   if λ_m < 0 then
6  |   |   |   γ_m := S*;
7  |   |   end
8  |   end
9  |   B ← B ∪ {m ∈ M \ B; λ_m > 0};
10 end
```

L'algorithme prend en entrée une instance du problème d'équilibrage des charges avec un ensemble de machines M et de recettes R. Ensuite, le programme linéaire $P(\mathcal{M}, \mathcal{B}, \gamma)$ est résolu pour déterminer la fonction objectif optimale S^*. Puis, les contraintes (4.9) limitantes sont identifiées, c'est-à-dire les contraintes correspondant aux machines limitant S^* à sa valeur courante. Ceci est fait en analysant les valeurs duales λ des contraintes (4.9), et en utilisant le théorème des écarts complémentaires, qui établit que : Dans une solution optimale, le produit de la variable d'écart et de la variable duale associée est égal à 0. Donc, si $\lambda_m < 0$, alors la contrainte (4.9) associée à la machine m est limitante, c'est-à-dire qu'il n'est pas possible de réduire le taux de charge de m sans dégrader S^* en augmentant le taux de charge des autres machines dans $\mathcal{M} \setminus \mathcal{B}$. Une fois que les contraintes (4.9) contraignantes sont identifiées, le taux de charge des machines correspondantes est fixé à la valeur de la fonction objectif optimale courante S^* (passant de contraintes (4.9) à des contraintes (4.10)) et le vecteur de charge γ est actualisé. Ainsi, le programme linéaire, à la prochaine itération, minimise le taux de charge maximum sur les machines restantes, tout en empêchant l'augmentation du taux de charge des machines dans $\mathcal{B}$. La procédure est répétée jusqu'à ce que le taux de charge de toutes les machines ait été fixé.

Notons que le problème ***Max-Min*** *Fairness Workload Balancing*, qui peut être considéré comme une version duale du problème MMFWB, peut être résolu en adaptant la procédure IMM, qui devient alors la procédure *Iterated Max-Min*. Plus précisément, dans l'algorithme 3, le programme linéaire $P(M, B, \gamma)$ est modifié en maximisant S dans (4.8) et en remplaçant

(4.9) par $W_m \geq S$.

4.5.2 Preuve d'exactitude

Divers articles de la littérature sur les problèmes de *min-max fairness* (ou leur équivalent lexicographique) ont proposé des méthodes de résolution. Luss and Smith (1986) et Tang (1988) proposent des algorithmes en temps polynomiaux pour résoudre des problèmes particuliers de planification de production. Bertsekas et al. (1987) proposent également un algorithme pour résoudre le problème MMF. Radunovic and Le Boudec (2007) montrent qu'un cas particulier du problème de *max-min fairness* peut être résolu très rapidement par un algorithme de *Water Filling*, et proposent une procédure utilisant des programmes linéaires de façon itérative pour résoudre le problème général. De plus, Behringer (1981) détaille l'utilisation d'une méthode basée sur le Simplexe pour résoudre un problème étendu de maximin lexicographique. Dans ces deux derniers articles, il n'est pas fait mention de l'utilisation de variables duales. Cependant, d'autres travaux de recherche tiennent explicitement compte de la dualité. Dans son livre, Luss (2012) considère les variables duales pour détecter les contraintes saturées, notamment dans le cas de fonctions objectifs non séparables. Nace and Pióro (2008) ont développé une procédure à base de programmation linéaire pour résoudre un problème de routage *min-max fair* dans les réseaux de communication et mentionnent également l'utilisation du concept de *min-max fairness* pour équilibrer lexicographiquement la charge dans un réseau donné. Enfin, Nace and Orlin (2007) introduisent ce qu'ils appellent des *lexicographically minimum load linear programming problems* pour une application dans des réseaux multiproduits à capacité. Ils présentent une procédure basée sur l'utilisation d'un programme linéaire et apportent la preuve de son exactitude. Ainsi, bien que nous n'ayons pas pu trouver d'autres algorithmes pour résoudre notre problème MMFWB, l'analogie entre la minimisation lexicographique et les problèmes *min-max fair* est très forte. C'est pourquoi nous nous basons sur Nace and Orlin (2007) pour énoncer la proposition 5.

Proposition 5. *La solution apportée par la procédure IMM au problème MMFWB est optimale et est obtenue en temps polynomial en résolvant au maximum* $|\mathcal{M}|$ *programmes linéaires, où* $\mathcal{M}$ *est l'ensemble des machines.*

La preuve de la proposition 5 suit la preuve présentée dans Nace and Orlin (2007), avec néanmoins une différence pour l'analyse du temps polynomial.

Tout d'abord, soulignons l'analogie entre l'approche de Nace and Orlin (2007) et la procédure IMM. Le programme linéaire (P_1) dans Nace and Orlin (2007) peut être transformé en problème $P(\mathcal{M}, \mathcal{B}, \gamma)$, en considérant que les contraintes (1) dans (P_1) sont les contraintes (4.9) dans $P(\mathcal{M}, \mathcal{B}, \gamma)$ et les contraintes (2) dans (P_1) sont les contraintes (4.10) et (4.12) dans $P(\mathcal{M}, \mathcal{B}, \gamma)$. En outre, l'étape 1 de l'algorithme de Nace and Orlin (2007), dans lequel un programme linéaire est résolu, correspond à la résolution de $P(\mathcal{M}, \mathcal{B}, \gamma)$ dans l'algorithme 3. L'étape 2 de l'algorithme de Nace and Orlin (2007), qui vise à trouver les contraintes limitantes et à mettre à jour le nouveau programme linéaire, correspond alors aux étapes restantes de l'algorithme 3. Enfin, les deux algorithmes se terminent lorsqu'il ne reste plus de contrainte d'inégalité, et le vecteur de charge γ résultant est *min-max fair* (*leximax minimal* dans Nace and Orlin (2007)).

Nace and Orlin (2007) montrent aussi que leur algorithme est polynomial en s'appuyant sur deux points principaux : (1) Le problème linéaire peut être résolu en temps polynomial et (2) Au plus $2|\mathcal{M}|-1$ programme linéaires sont à résoudre. Le premier point n'apporte aucune

difficulté car il existe de nombreuses méthodes pour résoudre les programmes linéaires en temps polynomial (Cook et al. (1995)).

Pour le second point, nous souhaitons aller un peu plus loin que Nace and Orlin (2007), qui garantissent que le programme linéaire est résolu au maximum $|\mathcal{M}|$ fois, pour peu que l'on utilise une méthode de résolution fournissant une solution strictement complémentaire. Les auteurs citent par exemple la méthode des points intérieurs de Freund and Mizuno (2000). Dans le cas de l'utilisation d'une méthode qui ne garantirait pas de fournir des solutions strictement complémentaires, Nace and Orlin (2007) proposent une étape supplémentaire, conduisant à la résolution d'au plus $2|\mathcal{M}||-1$ programmes linéaires. Cependant, nous affirmons qu'en utilisant l'IMM, au plus $|\mathcal{M}|$ programmes linéaires doivent être résolus, et ce quelle que soit la méthode utilisée pour résoudre le programme linéaire. Pour le prouver, il faut montrer qu'à chaque itération, au moins une valeur duale est strictement négative. Cette hypothèse n'est pas évidente dans le cas d'une solution optimale qui n'est pas strictement complémentaire, car le théorème des écarts complémentaires indique que, pour une contrainte donnée, "au moins" la variable d'écart ou la variable duale associée est égale à 0. Ainsi, bien qu'une contrainte donnée soit limitante, la variable duale associée peut prendre la valeur 0, et si toutes les variables duales sont nulles, l'algorithme peut alors effectuer un cycle sans modifier $\mathcal{B}$. Cependant, la solution vient du fait que la variable primale S des contraintes (4.9) conduit à la contrainte associée dans le problème dual $\sum_{m\in\mathcal{M}.\setminus\mathcal{B}} \lambda_m \leq -1$. Pour plus de détails, un raisonnement similaire peut être trouvé dans Luss (2012) (Chapitre 3, page 115). Par conséquent, puisque $\lambda_m \leq 0$, $\forall m \in \mathcal{M} \setminus \mathcal{B}$, la contrainte duale implique que $\lambda_m < 0$ pour au moins une machine m à chaque itération. Par conséquent, il y a au maximum M programmes linéaires à résoudre.

4.5.3 Illustration de la procédure

Pour conclure cette section de présentation de l'IMM, nous illustrons le fonctionnement de la méthode à l'aide de l'exemple de la section 4.3.1. Rappelons que l'instance comprend 5 machines, chacune avec une capacité $c_m = 8$, et 4 recettes A, B, C et D. De plus, pour simplifier, tous les temps de process sont identiques. Les quantités à traiter pour chaque recette sont $\{q_A, q_B, q_C, q_D\} = \{10; 12; 4; 4; 4\}$ et toutes les machines ne sont pas qualifiées pour toutes les recettes. Les machines 4 et 5 ne peuvent réaliser que la recette D, les machines 2 et 3 peuvent réaliser les recettes B, C et D, tandis que la machine 1 est qualifiée pour toutes les recettes.

— **Initialisation**: $\mathcal{B} := \emptyset$ et $\gamma = \{0, \ldots, 0\}$. Tout d'abord, nous minimisons le taux de charge maximum sur l'ensemble des machines de l'instance.

— **Step 1.1** Résoudre le programme linéaire $P(\mathcal{M}, \emptyset, \{0, \ldots, 0\})$. La fonction objectif de la solution optimale vaut $S^* = 10/8 = 1,25$ avec la solution $[q_{A,1}, q_{B,2}, q_{B,3}, q_{C,3}, q_{D,4}, q_{D,5}] = [10; 10; 2; 4; 4; 0]$. Les machines 1 et 2 semblent être limitantes.

— **Step 1.2** Analyser les variables duales associées aux contraintes (4.9). $[\lambda_1, \ldots, \lambda_5] = [-12,5; 0; 0; 0; 0]$. Par conséquent, bien que les contraintes de charge associées aux machines 1 et 2 semblent être limitantes, l'analyse des valeurs duales associées montre que seule la machine 1 est réellement critique.

— **Step 1.3** Soit $\gamma_1 = S^* = 1,25$ et $\mathcal{B} := \{1\}$. Le taux de charge de la machine 1 est fixé à 1.25, et la machine 1 n'est plus considérée dans le problème de minimisation du taux de charge maximum.

— **Step 2.1** Résoudre le programme linéaire $P(\mathcal{M}, \{1\}, \{1,25; 0; 0; 0; 0\})$. La fonction objectif de la solution optimale vaut $S^* = 8/8 = 1$ avec la solution $[q_{A,1}, q_{B,2}, q_{B,3}, q_{C,3}, q_{D,4}, q_{D,5}] = [10; 8; 4; 4; 4; 0]$. Par rapport à la première solution, une partie de la quantité de la recette B a été transférée de la machine 2 à la machine 3, permettant à ces deux machines de s'équilibrer entre elles avec une charge commune de 8/8=1.

— **Step 2.2** Analyser les variables duales associées aux contraintes (4.9). $[\lambda_2, \ldots, \lambda_5] = [-6,25; -6,25; 0; 0]$. Les machines 2 et 3 ont toute deux des variables duales, associées à leur contrainte de charge, strictement négatives. Ce qui signifie que les machines 2 et 3 sont limitantes.

— **Step 2.3** Soit $\gamma_1 = \gamma_2 = S^* = 1$ et $\mathcal{B} := \{1; 2; 3\}$. Le taux de charge des machines 2 et 3 est fixé à 1, et celles-ci ne sont plus considérées dans le problème de minimisation du taux de charge maximum.

— **Step 3.1** Résoudre le programme linéaire $P(\mathcal{M}, \{1; 2; 3\}, \{1,25; 1; 1; 0; 0\})$. La fonction objectif de la solution optimale vaut $S^* = 2/8 = 0,25$ avec la solution $[q_{A,1}, q_{B,2}, q_{B,3}, q_{C,3}, q_{D,4}, q_{D,5}] = [10; 8; 4; 4; 2; 2]$. Les machines 4 and 5 ne peuvent réaliser que la recette D, et donc partagent q_D pour s'équilibrer entre elles, amenant à une faible charge de 0.25.

— **Step 2.2** Analyser les variables duales associées aux contraintes (4.9). $[\lambda_4, \lambda_5] = [-6,25; -6,25]$. Les machines 4 et 5 ont toute deux des variables duales, associées à leur contrainte de charges, strictement négatives. Ce qui signifie que les machines 4 et 5 sont limitantes.

— **Step 2.3** Soient $\gamma_4 = \gamma_5 = S^* = 0,25$ et $\mathcal{B} := \{1; 2; 3; 4; 5\}$. Le taux de charge des machines 2 et 3 est fixé à 0,25, et celles-ci ne sont plus considérées dans le problème de minimisation du taux de charge maximum.

— **Fin**, car $\mathcal{B} = \mathcal{M}$. Retourner $\gamma = [1,25; 1; 1; 0,25; 0,25]$ avec l'allocation des quantités $[q_{A,1}, q_{B,2}, q_{B,3}, q_{C,3}, q_{D,4}, q_{D,5}] = [10; 8; 4; 4; 2; 2]$. Toutes les machines ont une charge fixée.

4.6 Résultats expérimentaux

Nous avons montré dans les sections précédentes que les caractéristiques des solutions *min-max fair* conduisent à des propriétés utiles pour notre problème d'équilibrage des charges, que la procédure IMM permet d'obtenir ces solutions *min-max fair*, et donc des solutions d'équilibrage des charges avec les propriétés souhaitées. Dans cette section, notre objectif est d'évaluer expérimentalement la performance de la procédure IMM par rapport au modèle d'équilibrage initial (avec les coefficients pondérateurs dans $f_c(.)$ ($\alpha >> \beta >> \gamma$), ainsi que sa variante dans laquelle ($\alpha >> \gamma >> \beta$). Les trois approches sont appelées respectivement, *MinFirst* (la maximisation du taux de charge minimum est privilégiée par rapport à la minimisation de la charge globale), *AvgFirst* (inverse du premier cas) et IMM. Nous utilisons un ensemble de 30 instances industrielles réelles provenant de l'usine 300mm de Crolles, dont les principaux paramètres pour notre problème sont le nombre de machines (environ 350) et le nombre de recettes (variant de 4000 à plus de 8000). Ces caractéristiques sont résumées dans le tableau 4.1.

Les indicateurs de performance permettant de comparer le modèle d'équilibrage initial (et sa variante) avec la procédure IMM sont examinés dans la section 4.6.1. La comparaison

proprement dite est effectuée dans la section 4.6.2.

TABLE 4.1 – Caractéristiques des instances industrielles

Instance	Nb de machines	Nb de recettes
Maximum	362	8540
Average	359	7622
Minimum	340	4053

4.6.1 Indicateurs de performance

L'indicateur *Nb de machines inutilement chargées* vise à quantifier le rapport des machines pour lesquelles le modèle initial donne un taux de charge supérieur à celui déterminé par la procédure IMM. Compte tenu des propriétés des solutions *min-max fair*, nous savons que, si une machine a un taux de charge supérieur à celui de la solution MMF, il devrait être possible de diminuer ce taux de charge en n'augmentant que celui des machines moins chargées, ce qui est préférable. Cet indicateur permet d'évaluer la proportion de machines que la procédure IMM ne charge pas inutilement. En utilisant l'exemple de la figure 4.1, la solution fournie par le modèle initial et présentée en (a) montre que la machine 2 est inutilement surchargée. En effet, la machine 2 a un taux de charge plus important qu'en (b). Cependant, selon les propriétés des solutions *min-max fair* déterminées par l'IMM, nous savons qu'il est possible de réduire le taux de charge de la machine 2 (de la solution (a) à la solution (b)), en augmentant seulement le taux de charge des machines moins chargées, c'est-à-dire la machine 3 dans notre cas. Dans cet exemple, nous avons une machine sur cinq qui est surchargée inutilement, ce qui conduit à un ratio de 20%. Notons que la charge de la machine 3 est plus faible avec la méthode initiale (solution (a)) qu'avec l'IMM (solution (b)). Cela ne signifie pas pour autant que la solution (a) est meilleure, car nous savons que, encore une fois par la propriété des solutions *min-max fair*, il serait possible d'augmenter le taux de charge de la machine 3 pour réduire celui d'une machine déjà plus chargée (ici machine 2), ce qui est toujours préférable.

Ensuite, comme indiqué dans les sections précédentes, la procédure IMM détermine des ensembles indépendants de machines et de recettes. L'indicateur *Nb de machines déséquilibrées* vise à analyser les machines équilibrées dans la solution donnée par l'IMM, puis évalue la proportion de ces machines dont l'équilibre a été rompu dans les solutions fournies par la méthode initiale. Ces cas de rupture d'équilibre ne sont pas souhaités car ils ne renseignent pas sur l'éventuelle relation de "support mutuel" entre les machines. Encore une fois, en illustrant avec l'exemple de la figure 4.1, nous notons que les taux de charge des machines 2 et 3 ne sont pas les mêmes en solution (a) et en solution (b). Mais nous savons, par propriété des solutions *min-max fair*, que si une méthode fournit une solution avec un taux de charge différent pour une machine, alors il y a deux cas possibles :

(1) La machine a un taux de charge plus important que dans la solution de l'IMM, c'est-à-dire qu'il est possible de transférer une partie de sa charge vers une machine moins chargée (machine 2 dans l'exemple).

(2) La machine a un taux de charge plus faible que dans la solution de l'IMM, c'est-à-dire qu'il est possible d'ajouter une partie de la charge d'une machine plus chargée (machine 3 dans l'exemple). Dans les deux cas, ces machines ne sont pas bien équilibrées. Dans l'exemple, il y a deux machines asymétriques, soit un rapport de 40%.

L'indicateur *Charge moyenne* est utilisé pour évaluer l'impact de la procédure IMM sur la charge totale des machines. En effet, nous avons observé que, dans le modèle d'équilibrage initial, le troisième terme de la fonction objectif (4.5) vise à minimiser la charge totale. L'objectif est d'éviter que le lissage du taux de charge entre les machines (minimiser le taux de charge maximum et maximiser le taux de charge minimum) ait un impact trop négatif sur la charge totale, en assignant par exemple des recettes sur des machines avec un temps de process élevé. Comme la procédure IMM ne tient pas explicitement compte de la charge totale, il est intéressant d'analyser l'impact de cette méthode sur cet indicateur.

Enfin, comme la procédure IMM résout généralement plusieurs programmes linéaires contre un seul avec le modèle d'équilibrage initial, il est également intéressant d'évaluer l'impact sur les temps de calcul. La colonne *Gap* montre la différence en pourcentage entre le temps requis en utilisant la procédure IMM et la méthode la plus rapide entre *AvgFirst* et *MinFirst*, c'est-à-dire:

$$Gap = \frac{CPU\_TIME_{IMM} - Min(CPU\_TIME_{AvgFirst}, CPU\_TIME_{MinFirst})}{Min(CPU\_TIME_{AvgFirst}, CPU\_TIME_{MinFirst})}$$

4.6.2 Comparaison avec le modèle d'équilibrage initial

Chaque instance a été exécutée avec l'outil de planification basé sur l'approche TSH, et les résultats sont résumés dans les tableaux 4.2 et 4.3. Dans le tableau 4.3, la deuxième colonne indique le nombre de problèmes résolus pour chaque cas. Notons que, pour chaque problème de planification, l'approche TSH a résolu en moyenne 365 problèmes d'équilibrage (deuxième colonne). Pour chaque problème d'équilibrage, les méthodes *AvgFirst*, *MinFirst* et IMM sont exécutées et comparées en utilisant les indicateurs présentés dans la section précédente. Chaque problème d'équilibrage comprend plusieurs machines non identiques et parallèles (8 en moyenne mais variant entre 1 et 20 machines selon les problèmes) et plusieurs dizaines de recettes différentes.

Plusieurs remarques peuvent être faites lors de l'analyse du tableau 4.2. Premièrement, comme prévu, les colonnes correspondant à la procédure IMM ne sont composées que de zéros. Ceci est normal car l'objectif est de comparer les solutions avec la solution *min-max fair* souhaitée, qui est celle fournie par l'IMM (voir propriété 5). Ensuite, nous remarquons des différences entre les solutions fournies avec le modèle initial et sa variante. Le *Nb de machines chargées inutiles* est égal à 15.2% en moyenne avec *AvgFirst* alors qu'il augmente à 17.8% avec *MinFirst*. Cela montre qu'un nombre important de machines (55 en moyenne, au moins 36 dans chaque cas) ont un taux de charge qui pourrait être réduit sans affecter le taux de charge des machines plus chargées.

Là encore, une remarque importante doit être faite sur les machines pour lesquelles la procédure IMM semble augmenter le taux de charge par rapport au modèle d'équilibrage initial. En effet, il y a des machines pour lesquelles le taux de charge total déterminé par la procédure IMM est supérieur au taux de charge déterminé par le modèle initial. Ce type de situations est possible mais implique nécessairement qu'en retour la procédure IMM est capable de réduire le taux de charge d'autres machines qui sont déjà plus chargées. Des analyses ont montré qu'à chaque fois qu'une machine est plus chargée dans la solution IMM que dans la solution obtenue avec le modèle initial, le phénomène inverse est observé pour une machine déjà plus chargée.

La procédure IMM peut donc avoir un impact significatif sur l'analyse de la capacité de l'usine, en évitant aux utilisateurs d'avoir une vision erronée du taux de charge de dizaines de machines.

TABLE 4.2 – Proportion de cas d'équilibrage non souhaité entre les solutions obtenues avec le modèle initial et celles obtenues avec la procédure IMM

Instance	Nb de machines inutilement chargées			Nb de machines déséquilibrées		
	AvgFirst	MinFirst	IMM	AvgFirst	MinFirst	IMM
1	12,1%	14,9%	0,0%	16,7%	8,8%	0,0%
2	12,5%	17,3%	0,0%	16,8%	8,6%	0,0%
3	12,8%	15,6%	0,0%	17,5%	7,8%	0,0%
4	11,3%	14,6%	0,0%	15,1%	7,4%	0,0%
5	9,9%	13,1%	0,0%	12,1%	6,2%	0,0%
6	11,2%	13,3%	0,0%	10,8%	5,0%	0,0%
7	10,5%	13,4%	0,0%	14,6%	7,1%	0,0%
8	13,4%	16,1%	0,0%	18,1%	10,6%	0,0%
9	12,5%	16,0%	0,0%	18,0%	8,4%	0,0%
10	13,9%	17,2%	0,0%	14,9%	9,7%	0,0%
11	13,9%	19,6%	0,0%	16,0%	11,5%	0,0%
12	13,2%	19,5%	0,0%	16,4%	11,6%	0,0%
13	12,8%	14,8%	0,0%	14,1%	8,9%	0,0%
14	16,0%	17,9%	0,0%	16,4%	12,1%	0,0%
15	17,5%	18,7%	0,0%	16,5%	12,7%	0,0%
16	16,4%	18,1%	0,0%	15,1%	11,6%	0,0%
17	16,7%	18,5%	0,0%	16,0%	12,0%	0,0%
18	17,2%	19,8%	0,0%	16,8%	11,1%	0,0%
19	17,0%	18,0%	0,0%	14,2%	9,2%	0,0%
20	17,4%	18,1%	0,0%	16,4%	11,3%	0,0%
21	16,8%	19,6%	0,0%	15,3%	8,6%	0,0%
22	17,6%	18,7%	0,0%	14,6%	10,2%	0,0%
23	17,2%	18,1%	0,0%	16,4%	11,5%	0,0%
24	17,9%	19,6%	0,0%	18,1%	12,8%	0,0%
25	17,6%	21,3%	0,0%	17,0%	10,5%	0,0%
26	17,5%	20,0%	0,0%	17,0%	11,0%	0,0%
27	19,5%	21,1%	0,0%	17,4%	11,3%	0,0%
28	17,5%	18,2%	0,0%	17,6%	11,7%	0,0%
29	18,1%	21,4%	0,0%	17,8%	11,4%	0,0%
30	18,7%	20,3%	0,0%	16,8%	9,8%	0,0%
Avg	15,2%	17,8%	0,0%	16,0%	10,0%	0,0%
Max	19,5%	21,4%	0,0%	18,1%	12,8%	0,0%
Min	9,9%	13,1%	0,0%	10,8%	5,0%	0,0%

Considérons maintenant le deuxième indicateur, le *Nb de machines déséquilibrées*, qui correspond au rapport des machines qui sont équilibrées les unes avec les autres dans une solution *min-max fair* mais ne le sont pas dans les solutions déterminées par le modèle initial. Les résultats montrent qu'en moyenne 16% des machines avec *AvgFirst* et 10% des machines avec *MinFirst* ne sont pas correctement équilibrées. Ce type de déséquilibre est similaire à celui présenté dans notre exemple illustratif, où les machines 2 et 3 ne sont pas équilibrées dans la solution (a) de la figure 4.1 alors qu'elles pourraient l'être. Ces résultats montrent que l'utilisation de la procédure IMM pour remplacer le modèle d'équilibrage initial permet de mettre en évidence les relations d'interdépendance pour un nombre significatif de machines.

TABLE 4.3 – Charge totale des machines et temps de calcul avec le modèle initial et la procédure IMM

Instance	Nb de problèmes MMFWB	Charge moyenne			Temps de calcul / problème (ms)			
		AvgFirst	MinFirst	IMM	AvgFirst	MinFirst	IMM	Gap
1	386	0,391	0,398	0,397	130	120	130	8%
2	384	0,418	0,419	0,418	130	130	130	0%
3	384	0,318	0,320	0,320	120	130	120	0%
4	380	0,302	0,303	0,303	130	120	130	8%
5	379	0,346	0,344	0,344	130	130	120	-8%
6	379	0,207	0,206	0,206	150	110	120	9%
7	380	0,306	0,310	0,309	120	120	140	17%
8	378	0,314	0,321	0,321	120	110	110	0%
9	376	0,332	0,340	0,339	120	120	110	-8%
10	375	0,329	0,333	0,330	120	120	120	0%
11	372	0,262	0,267	0,265	120	110	110	0%
12	375	0,263	0,266	0,264	130	110	120	9%
13	376	0,300	0,303	0,302	120	120	120	0%
14	368	0,231	0,234	0,237	370	370	380	3%
15	335	0,251	0,254	0,258	380	390	390	3%
16	349	0,239	0,243	0,247	370	370	380	3%
17	369	0,308	0,312	0,317	380	380	420	11%
18	365	0,246	0,247	0,251	370	370	380	3%
19	382	0,250	0,250	0,251	380	400	450	18%
20	368	0,278	0,283	0,286	370	370	370	0%
21	339	0,215	0,217	0,219	120	110	120	9%
22	345	0,252	0,253	0,255	370	400	390	5%
23	365	0,264	0,263	0,265	430	450	420	-2%
24	378	0,298	0,299	0,301	420	510	430	2%
25	344	0,249	0,253	0,256	430	410	430	5%
26	345	0,278	0,282	0,285	400	450	430	8%
27	348	0,296	0,300	0,303	430	470	450	5%
28	345	0,329	0,336	0,340	430	460	440	2%
29	354	0,324	0,332	0,335	440	440	450	2%
30	339	0,246	0,248	0,250	420	430	440	5%
Avg	365	0,288	0,291	0,293	272	278	279	3,9%
Max	384	0,418	0,419	0,418	44,00	51,00	45,00	18,4%
Min	335	0,207	0,206	0,206	12,00	11,00	11,00	-8,3%

Les résultats sur la charge totale, résumés dans le tableau 4.3, montrent que la variante *AvgFirst* de l'approche initiale, parce qu'elle donne la priorité à la charge totale sur le taux de charge minimum, détermine des solutions avec une charge moyenne inférieure par rapport aux autres approches. Cependant, il est à souligner que la différence reste relativement faible avec *MinFirst* et la procédure IMM. Par ailleurs, il est à noter que certaines des solutions déterminées par la procédure IMM sont meilleures que celles du modèle initial. Par conséquent, les avantages des solutions déterminées par la procédure IMM ne semblent pas se faire au détriment de la charge totale.

De plus, malgré la taille importante des instances, et parce que seules des variables continues sont utilisées, les trois approches fonctionnent très rapidement, avec des temps de calcul généralement de l'ordre de quelques secondes pour résoudre des centaines de problèmes d'équilibrage des charges. Une légère augmentation du temps de calcul de 3,9% en moyenne est observée avec la procédure IMM, ce qui est peu significatif surtout lorsque l'on rapporte cela au temps total d'exécution de l'approche TSH.

4.7 Conclusions et perspectives

Dans ce chapitre, nous avons abordé le problème d'équilibrage des charges dans des systèmes de fabrication sur des machines parallèles non-identiques. Nous avons ensuite rappelé la notion de solutions *min-max fair*, et montré que, appliquée à notre problème, elle peut apporter des solutions riches de sens pour les utilisateurs, notamment afin de détecter les machines critiques, ou celles "connectées" du fait du type de recettes à réaliser et de leurs quantités. La procédure Iterated Min-Max (IMM) est proposée et, sur la base des travaux de Nace and Orlin (2007), a été prouvé capable de trouver des solutions optimales à notre problème d'équilibrage des charges nommé problème *Min-Max Fair Workload Balancing*.

Nous avons montré que la procédure IMM permet d'obtenir des solutions de meilleure qualité que celles déterminées par la méthode implémentée dans la version initiale de l'outil de planification. Ainsi, suite à ces travaux, la méthode initiale d'équilibrage a été remplacée par l'approche IMM, qui est désormais la méthode par défaut pour les utilisateurs de l'outil d'aide à la planification de production opérationnelle.

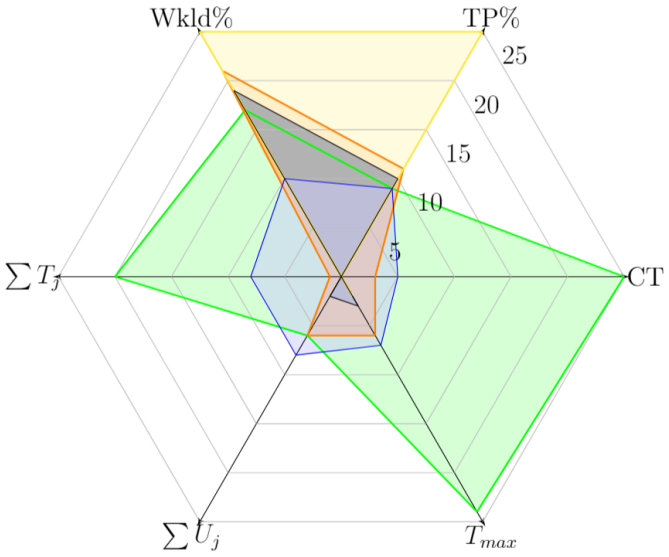
Nous avons donc amélioré l'approche TSH grâce à un meilleur équilibrage des étapes de process sur les nombreuses machines de l'usine. Cependant, cet équilibrage reste un problème à capacité infinie, ne considérant par la capacité des machines comme de réelles contraintes. Ainsi, à la sortie du module d'*équilibrage*, l'approche TSH détermine un plan de production à capacité infinie, certes riche en information, mais qui est potentiellement non réalisable à l'égard de la capacité de l'usine. La transformation de ce plan initial en un plan réalisable est la tâche du troisième composant de l'approche TSH, le module de *step-shifting*, sur lequel nous avons également travaillé durant cette thèse et dont il est question dans le prochain chapitre.

Chapitre 5

Module de lissage de la capacité

.

Les modules de *projection* et d'*équilibrage* de l'approche TSH ont défini dans quelle période, et sur quelle machine, réaliser chaque étape de process de chaque lot. Ce plan de production initial n'est cependant potentiellement pas réalisable et doit être ajusté par le module de *step-shifting* via un lissage des charges afin de respecter les contraintes de capacité. Dans cette thèse sont développées différentes règles de lissage permettant d'optimiser la qualité des plans de production selon différents critères de performances. Nous introduisons également une nouvelle approche d'évaluation du taux de charge tenant compte des relations entre machines équilibrées. Enfin, face à ces approches ne pouvant adapter la solution courante qu'en la dégradant, une approche complémentaire de lissage par anticipation est introduite afin de tirer profit de cas de surcapacité de certaines machines.



5.1 Introduction

Dans le chapitre 4, nous nous sommes intéressés aux problématiques d'équilibrage des charges sur des machines non identiques au travers du module d'*équilibrage* de l'approche TSH. Une nouvelle approche d'équilibrage, nommée méthode *Iterated Min-Max* (IMM), a été présentée et détermine des solutions riches de sens pour les utilisateurs, notamment afin de détecter, en fonction du mix-produit prévu, les machines potentiellement limitantes ou celles "liées", car partageant les même files d'attente de produits.

Ce module d'*équilibrage* prend comme entrée le plan de production déterminé par le module de *projection*, sous forme d'une liste d'étapes de process de chaque lot devant être traitées dans la période considérée. Bien que l'équilibrage tienne compte de la capacité des machines dans le calcul du taux de charge, cette contrainte n'est pas "dure" et le plan de production potentiellement non réalisable, défini par le module de *projection*, n'est pas remis en question. Or, dans notre contexte d'*Operational Production Planning* (OPP), où des consignes globales doivent être données afin d'orienter les différents outils d'ordonnancement dans l'usine, le fait de ne pas considérer la capacité risque d'aboutir à des objectifs non atteignables par les différents ateliers (Dauzère-Pérès and Lasserre (2002)), risquant de dégrader le respect du planning de livraison.

Le respect des contraintes de capacité est assuré par le troisième module de l'approche TSH, nommé module de *step-shifting*. Ce module tire son nom de son principe de fonctionnement, qui consiste à "décaler" certaines étapes de process de certains lots, d'une période à une autre, afin de réduire le taux de charge des machines surchargées. Ce décalage peut être vu comme une simulation d'un lot peu prioritaire arrivant à une période (jour, semaine) donnée devant une machine, et pour lequel l'étape de process sera seulement réalisée à la période suivante.

Le choix du lot est un élément important. Dans la section 3.4.4 du chapitre 3, la présentation du module de *step-shifting* est faîte en décalant les lots dont la date de livraison est la plus éloignée, ce qui équivaut à la règle classique de *Dispatching Earliest Due Date* (EDD). Dans les travaux de Mhiri et al. (2018), qui sont à la base de la version initiale de l'outil de planification opérationnelle, est présentée une règle en partie basée sur les dates de livraison. Cette règle a pour principal objectif de faire attendre les lots les moins en retard afin de minimiser les retards de livraison, représentés notamment par le *Total Weighted Tardiness*. Mais bien qu'elle tienne compte des dates de livraison, cette règle est-elle la meilleure afin de respecter le plan défini par le *Master Planning*? De plus, les critères d'optimisation dans l'industrie des semi-conducteurs, et dans les usines *front end*, sont variés et parfois antinomiques, allant du respect des commandes clients à la maximisation du nombre de plaquettes lancées en production chaque semaine, en passant par la minimisation du temps de cycle global des produits (Pfund et al. (2006)). Il est donc intéressant d'évaluer les performances de la règle proposée dans Mhiri et al. (2018), mais aussi de proposer de nouvelles règles, en fonction de différents critères de performance classiques de l'industrie des semi-conducteurs. Ainsi, la section 5.2 de ce chapitre sera axée sur l'analyse de différentes règles de lissage, leur comparaison selon différents critères de performance, et la perspective de leur utilisation dans le cadre industriel.

Nous étudierons ensuite d'autres approches afin d'améliorer la qualité des plans de production réalisables calculés par le module de *step shifting*. Tout d'abord, nous chercherons dans la section 5.3 à limiter les opérations de décalage par une meilleure évaluation du taux de charge de machines dites "liées" et définies grâces à notre nouvelle approche d'*équilibrage*. Une approche de lissage arrière, ou *preponing*, sera ensuite présentée dans la section 5.4.

Contrairement au lissage vers l'avant classique qui dégrade la solution initiale en décalant des lots afin de créer un plan de production réalisable, la nouvelle approche a pour but de tirer profit de cas de surcapacité, afin d'améliorer les solutions courantes.

5.2 Étude de règles de lissage

Comme déjà précisé, le module de *step-shifting* consiste à sélectionner certaines étapes de certains lots et à les décaler à la période suivante, c'est à dire faire attendre la réalisation de ces étapes de process jusqu'à la période suivante. Le choix du lot à décaler a une influence importante sur la qualité des solutions obtenues, comme nous le montrerons dans l'analyse numérique de cette section. Le choix du lot à décaler se fait selon une règle qui choisit, parmi une liste de lots potentiels, celui étant le moins prioritaire. Cette démarche est très similaire au processus de *Dispatching* (heuristique de liste en ordonnancement), où une règle est utilisée afin de choisir, parmi une liste de lots, le prochain devant passer sur une machine. C'est d'ailleurs pour cette raison que certaines des règles de lissage, utilisées dans le module de *step-shifting* de l'approche TSH et introduites dans cette section, ont le même nom que certaines règles classiques de *Dispatching*, telles que la règle *Earliest Due Date*, qui est la première règle de lissage que nous introduisons.

Les différentes notations introduites et utilisées dans la section 5.2 sont résumées dans le tableau 5.1.

TABLE 5.1 – Notations utilisées dans la définition des règles de lissage

Notation	**Description**
$\mathcal{L}$	Ensemble des lots
$l \in \mathcal{L}$	Indice d'un lot
$\mathcal{S}_{s,l}$	Ensemble des étapes de process à réaliser à partir de l'étape de process s pour terminer le lot l
$s \in \mathcal{S}_l$	Indice d'une étape de process du lot l
$t_{s,l}$	Date à laquelle l'étape de process s du lot l commence à être réalisée
d_l	Date de livraison du lot l (correspond à la date de fin de la période de livraison du lot).
$p_{s,l}$	Temps moyen requis pour réaliser l'étape s du lot l
$CT^{th}_{s,l}$	Temps de cycle théorique (process + attente) restant pour le lot l à partir de l'étape s

5.2.1 Règles orientées livraisons

Parmi les règles de lissage proposées, la plus simple est celle de la date de livraison au plus tôt, ou *Earliest Due Date* (EDD), basée uniquement sur les dates de livraison des lots. Ainsi, une fois que la machine la plus surchargée est déterminée, le lot ayant la date de livraison la plus lointaine (parmi les lots ayant une étape de process traitée par cette machine dans la période considérée) est décalé à la période suivante (c'est à dire, attendra devant la machine jusqu'à la période suivante).

Le problème de la règle EDD est qu'elle ne tient pas compte de la position du lot dans sa gamme opératoire (c'est-à-dire la séquence fixe des étapes de process restant à réaliser avant que le lot puisse sortir de l'usine). En effet, deux lots l_1 et l_2 ayant la même date de

livraison, auront la même priorité avec la règle EDD. Cependant, si l_1 nécessite encore 100 étapes de process avant d'être terminé alors que l_2 n'a plus qu'une étape, on comprend qu'il sera plus difficile de respecter l'échéance de l_1 que celle de l_2. Le lot l_1 devrait donc avoir une priorité plus élevée que le lot l_2, ce qui n'est pas possible avec la règle EDD. C'est pour répondre à ce type de cas que nous avons introduit la deuxième règle, qui est une variante de la règle classique *minimum slack time* (MST, Baker (1974)), où le *slack time* (la marge), est la différence entre le temps restant jusqu'à la date de livraison et le temps cumulé des étapes de process restantes. Définissons par $t_{s,l}$ l'instant où l'étape de process s du lot l commence à être réalisée, d_l la date de livraison du lot l, S_l l'ensemble des étapes de process restant à faire pour le lot l, et $p_{s,l}$ le temps requis pour réaliser l'étape s du lot l. Alors le *slack time* du lot l à l'étape s, nommé $ST_{s,l}$, est défini par :

$$ST_{s,l} = d_l - t_{s,l} - \sum_{s \in S_l} p_{s,l} \tag{5.1}$$

La règle MST de *Dispatching* consiste ensuite à privilégier les lots ayant le plus faible *slack time*, c'est à dire la plus faible marge entre le temps disponible et le temps cumulé de process restant. Dans le cas du lissage, le lot à décaler à la période suivante sera celui ayant la plus grande valeur de *slack time*. Cette règle est cependant légèrement modifiée afin d'inclure le temps d'attente des lots devant chaque étape $w_{s,l}$, en plus des temps de process $p_{s,l}$. En effet, dans une usine *front end*, l'attente devant les machines constitue la majeure partie du temps de cycle total des lots, largement plus que le temps de process. Ainsi, une bonne évaluation de la marge disponible, pour un lot donné, doit tenir compte non seulement du temps nécessaire restant pour réaliser toutes les étapes, mais aussi du temps d'attente estimé devant chacune de ces étapes. Le terme $\sum_{s \in S_l} p_{s,l}$ est donc remplacé par le temps de cycle théorique restant pour le lot l à partir de l'étape s, nommé $CT^{th}_{s,l}$. Ce temps de cycle théorique inclut les temps de process et d'attente des étapes restantes, et est évalué statistiquement à l'aide des données historiques de l'usine. On obtient alors la nouvelle formule du *slack time* :

$$ST_{s,l} = d_l - t_{s,l} - CT^{th}_{s,l} \tag{5.2}$$

Grâce à la règle de lissage ST, parmi les lots pouvant potentiellement être décalés, le choix portera sur ceux ayant la plus grande marge pour être livrés à temps.

En plus de la règle ST présentée ci-dessus, une variante est proposée, nommée "*Slack Time If Postponed*" (ST_{post}). Afin de motiver cette variante, rappelons que nous considérons un horizon temporel discrétisé, et que chaque période est traitée itérativement l'une après l'autre. La longueur des périodes est généralement de l'ordre d'une semaine. Cela signifie que, si une étape d'un lot est repoussée à la période suivante, le décalage ne sera pas le même selon que l'étape était prévue au début ou à la fin de la période actuellement considérée. Si l'étape était planifiée au début de la période, la reporter à la période suivante signifie retarder le lot de plusieurs jours (potentiellement jusqu'à 7 dans le cadre de périodes d'une semaine). Prenons par exemple un lot l_1 ayant plus de marge qu'un lot l_2 (par exemple une marge de 10 jours pour l_1 contre une marge de 6 jours pour l_2). Supposons que l_1 et l_2 sont des candidats pour être décalés à la période suivante, car ils ont tous deux une étape de process à réaliser sur la machine la plus chargée pendant la période considérée. Le lot l_1 a en théorie une priorité inférieure à l_2. Cependant, comme l_1 a son étape de process prévue au début de la période actuellement considérée, reporter l_1 le retarderait de 7 jours, alors que l'étape de process de l_2 sur la machine n'est prévue qu'un jour avant la fin de la période. Dans ce cas, l_1 aurait une nouvelle marge ST_{l_1} de $10 - 7 = 3$ jours, tandis que l_2 aurait une

marge de 5 jours. Cela signifie que décaler l_2 est en fait préférable car cela laisserait une plus grande marge que si l_1 était décalé. Ainsi, la variante $ST_{post,s,l}$ calcule la marge que le lot l *aurait* s'il était déplacé à la période suivante. Si l'on définit toujours $t_{s,l}$ l'instant où le lot l attend devant la machine pour que soit réalisée son étape de process s, et t^{next} la date de début de la prochaine période, on obtient pour le calcul de ce nouvel indicateur l'expression ci dessous :

$$ST_{post,s,l} = ST_{s,l} - (t^{next} - t_{s,l}) \tag{5.3}$$

En combinant avec l'expression (5.2), on obtient alors :

$$ST_{post,s,l} = d_l - t^{next} - CT^{th}_{s,l} \tag{5.4}$$

Ainsi, l'indicateur $ST_{post,s,l}$ n'est autre que l'indicateur $ST_{s,l}$ en considérant l'étape de process s du lot l réalisée au début de la période suivante t^{next}.

Les règles ST et ST_{post} ont donc l'avantage, par rapport à la règle EDD, d'être basées sur la marge qu'ont les lots afin d'être livrés à temps au client, compte tenu des dates de livraison et de la connaissance du temps de cycle théorique des produits dans l'usine. Toutefois, ces règles ne font pas de distinction entre deux lots ayant la même marge, mais situés à des positions différentes sur leur gamme de fabrication. Or, si un lot est considéré en retard, il sera plus facile de rattraper ce retard si le lot est au début de sa gamme de fabrication (par exemple s'il lui reste théoriquement deux mois de temps de cycle) que s'il est censé être livré dans la semaine. Du fait de cette limite des règles précédentes, nous avons implémenté une quatrième règle de lissage qui est une variante de la classique règle de *Dispatching Critical Ratio* (CR, Baker (1974)). En reprenant les notations utilisées pour définir l'indicateur ST, nous pouvons définir le *critical ratio* $CR_{s,l}$ d'un lot l à une étape s, par l'expression suivante :

$$CR_{s,l} = \frac{d_l - t_{s,l}}{CT^{th}_{s,l}} \tag{5.5}$$

Cet indicateur est donc le rapport entre le temps restant jusqu'à la date de livraison du lot et le temps restant théorique. La règle CR aura donc tendance à donner plus de poids aux lots proches de la fin de leur gamme de fabrication.

De la même manière que pour la règle de lissage ST, une variante de la règle *Critical Ratio* est proposée, appelée *Critical Ratio If Postponed* ($CR_{post,s,l}$), considérant la valeur de l'indicateur $CR_{s,l}$ du lot l à une étape s, si la réalisation de cette étape est reportée au début de la période suivante. Cette variante est donc la cinquième règle de lissage.

Enfin, nous intégrons dans cette analyse la règle de lissage présentée dans Mhiri et al. (2018).Cette règle de priorité, nommée *RankingCoeff*, évalue la priorité du lot selon sa position dans la file d'attente devant la machine où l'étape de process doit être réalisée ainsi que son "urgence de livraison". Dans ce manuscrit, nous nous référons à cette règle comme étant la règle "Critical Ratio and Position on Machine" (CRPM). Adaptée à notre formulation, avec $t_{s,l}$ la date de début de l'opération s du lot l et t^{next} la date de début de la période suivante, l'indicateur $CRPM$ peut être défini par l'expression suivante :

$$CRPM_{s,l} = CR_{s,l} + \frac{t_{s,l}}{t^{next}} \tag{5.6}$$

Le premier terme est le *critical ratio* et vise à évaluer l'urgence de livraison du lot, tandis que le second terme est une valeur normalisée de la position du lot dans la file d'attente de la machine. La règle de lissage $CRPM$ tend donc à retarder les lots qui, d'une part, sont les plus en avance par rapport à leur date de livraison et qui, d'autre part, partagent la machine considérée avec beaucoup d'autres lots prévus de passer plus tôt.

5.2.2 Règles orientées machines

Les règles précédemment introduites sont basées sur des considérations temporelles (dates de livraison, temps de cycle restant) et visent principalement à minimiser le retard global (TWT). Cependant, il existe d'autres indicateurs en fabrication de semi-conducteurs, tels que le temps de cycle moyen, la productivité de l'usine (le *throughput*) ou le taux d'utilisation des machines. Ces indicateurs sont suivis par les managers avec autant d'attention (parfois plus) que les indicateurs orientés client, mais ne sont pas réellement considérés par les règles de lissage introduites dans la section précédente.

Il semble donc pertinent de proposer de nouvelles règles, prenant en compte d'autres indicateurs non orientés client (exclusivement ou partiellement). Il sera ainsi possible d'évaluer dans quelle mesure les règles de la section 5.2.1 peuvent dégrader des indicateurs tels que la productivité ou le taux d'utilisation des machines. L'implémentation de nouvelles règles de lissage permet également de diversifier les possibilités d'optimisation pour l'approche de résolution TSH.

Nous proposons par conséquent une sixième règle, appelée *Machine Impact* (MI), et qui prend en compte la charge générée par les lots sur les machines. Comme pour les règles précédentes, la première étape consiste à déterminer la machine la plus surchargée, puis à identifier tous les lots ayant au moins une étape de process à réaliser sur cette machine pendant la période considérée. Ensuite, parmi ces lots, nous recherchons celui à reporter générant la plus faible charge sur des machines qui ne sont pas surchargées. En effet, notre objectif est de réduire la charge des machines surchargées, en minimisant la perte de charge sur les machines qui n'ont pas besoin d'être déchargées davantage. Nous essayons donc de décaler à la période suivante le lot générant le moins de charge sur les machines non surchargées. Pour cela, pour chaque couple (lot, étape), est calculée la charge *cumulée* générée par ce lot pour cette étape, *ainsi que les étapes ultérieures dans la période* (uniquement sur les machines non surchargées). Cette évaluation de la charge cumulée est nécessaire parce qu'un lot passe généralement par plusieurs étapes de process durant la même période. Ainsi, décaler un lot à partir d'une certaine étape de process implique de décaler également les étapes de process suivantes et également prévues dans la période considérée.

La règle de lissage MI, ne tient pas compte de la dimension client. Cette règle favorise des objectifs tels que la maximisation de la charge (ou taux d'utilisation des machines) ou la maximisation de la productivité de l'usine, plutôt que des objectifs orientés sur les livraisons. Afin de concilier ces deux classes d'indicateurs, nous avons combiné les règles de lissage présentées précédemment en deux nouvelles règles appelées "Machine Impact and Slack Time" (MI_{ST}) et "Machine Impact and Critical Ratio" (MI_{CR}). Ces deux règles suivent le principe suivant : La machine la plus surchargée est toujours identifiée en premier, puis tous les lots avec au moins une opération traitée par cette machine pendant la période sont identifiés. Toutefois, seuls les lots considérés comme étant en avance sont pris en compte. Un lot est considéré en avance si l'indicateur ST (leur marge) est positif, c'est à dire que le temps restant disponible avant la date de livraison du lot est supérieur au temps théorique restant pour finir l'ensemble de ses étapes de process. Notons que ces lots en avance sont également ceux ayant un *critical ratio* supérieur à 1. Une fois que les lots considérés comme en avance sont identifiés, la règle de lissage MI est appliquée en reportant le lot ayant le moins d'impact sur les machines non surchargées. Si tous les lots traités par la machine la plus surchargée sont considérés en retard, la règle MI n'est pas appliquée, et les règles de lissage orientées client sont utilisées, avec respectivement ST_{Post} pour la règle MI_{ST} et CR_{Post} pour la règle MI_{CR}.

Le tableau 5.2 synthétise les différentes règles décrites dans les sections 5.2.1 et 5.2.2. Notons que dans la colonne "Description", les lots à décaler sont dans le sous-ensemble des lots ayant au moins une étape de process réalisée dans la période par la machine la plus surchargée.

TABLE 5.2 – Règles de lissage étudiées

Notation	**Nom**	**Description**
EDD	*Earliest Due Date*	Repousse le lot ayant la date de livraison la plus éloignée.
ST	*Slack Time*	Décale le lot dont la marge (différence entre le temps restant jusqu'à la date de livraison et le temps de cycle théorique requis par le lot pour finir sa gamme de fabrication) est la plus grande.
ST_{Post}	*Slack Time if postponed*	Considère la marge (ST) que le lot *aurait* s'il était décalé au début de la période suivante.
CR	*Critical Ratio*	Décale le lot dont le ratio entre le temps restant jusqu'à la date de livraison et le temps de cycle théorique requis par le lot pour finir sa gamme de fabrication, est le plus grand.
CR_{Post}	*Critical Ratio if postponed*	Considère le *critical ratio* (CR) que le lot *aurait* s'il était décalé au début de la période suivante.
$CRPM$	*Critical Ratio and Position on Machine*	Définit la priorité des lots selon leur niveau de retard, combiné à leur position dans la file d'attente de la machine considérée. Tend à décaler les lots les plus en avance et les plus éloignés dans la file d'attente.
MI	*Machine Impact*	Repousse le lot dont la charge induite dans la période (à partir de l'étape considérée), sur des machines qui ne sont pas surchargées, est la plus petitz.
MI_{ST}	*Machine Impact with Slack Time if postponed*	Repousse le lot (parmi ceux en avance) dont la charge induite dans la période (à partir de l'étape de process considérée) sur les machines non surchargées, est la plus petite. Si aucun lot n'est en avance, la règle ST_{Post} est appliquée
MI_{CR}	*Machine Impact with Critical Ratio if postponed*	Repousse le lot (parmi ceux en avance) dont la charge induite dans la période (à partir de l'étape de process considérée) sur les machines non surchargées, est la plus petite. Si aucun lot n'est en avance, la règle CR_{Post} est appliquée

5.2.3 Évaluation des performances

Instances et critères de comparaison

Dans le chapitre 3, nous avons comparé l'approche heuristique TSH avec un Programme Linéaire en Nombres Entiers sur de très petites instances, montrant la grande rapidité et les bonnes performances de l'approche TSH.

Dans cette section, nous comparons sur 25 instances réelles l'approche TSH avec différentes règles de lissage. Les principaux paramètres des instances sont le nombre de machines et le nombre moyen de lots dans l'usine. Le tableau 5.3 résume la plage de ces paramètres.

TABLE 5.3 – Caractéristiques des instances industrielles

Instance	Nb Machines	Moyenne Nb Lots
Maximum	374	3172
Moyenne	357	3040
Minimum	334	2830

Pour chaque instance, l'outil de planification, basé sur l'approche TSH, a été exécuté sur un horizon de 12 périodes d'une semaine. L'outil a été lancé pour chacune des 9 variantes des règles de lissage, résumées dans le tableau 5.2, et les résultats ont été comparés selon 6 indicateurs de performances (KPI pour *Key Performance Indicators*), résumés dans le tableau 5.4.

TABLE 5.4 – Indicateurs de performance considérés

Indicateurs	**Description**
$\sum T_l$	Somme des retards (en nombre de semaines) de l'ensemble des lots.
$\sum U_l$	Nombre total de lots livrés en retard.
T_{max}	Retard maximum de livraison sur l'ensemble des lots.
CT	Temps de cycle moyen des lots.
Wkld%	Taux de charge moyen des machines. Le résultat est donné en pourcentage d'écart par rapport à la charge moyenne obtenue avec la règle de lissage EDD.
TP%	Écart de productivité de l'usine obtenu avec la règle de lissage EDD. Une valeur positive signifie que la règle de lissage considérée fournit des solutions avec une productivité plus élevée et permet de planifier plus d'étapes de process dans le même horizon.

Analysons d'abord ces indicateurs. Tout d'abord, notons que les trois premiers indicateurs ($\sum T_l$, $\sum U_l$ et T_{max}) sont basés sur les dates de livraison des lots, et évaluent si les approches fournissent des solutions de qualité vis à vis des clients. Les autres indicateurs, se concentrent sur d'autres aspects, afin d'évaluer l'impact des règles de lissage sur le temps de cycle des lots (CT), la charge des machines (Wkld%) et la productivité de l'usine (TP%).

Deuxièmement, bien que les indicateurs $\sum T_l$, $\sum U_l$ et T_{max} soient orientés clients, les résultats peuvent probablement différer entre eux. En particulier, si une règle tend à répartir équitablement le retard entre les lots, elle aura tendance à minimiser le retard maximum, c'est-à-dire T_{max}. Cela se fera généralement au détriment du nombre total de lots en retard, c'est-à-dire $\sum U_l$. Inversement, une règle qui limite le nombre de lots en retard aura de

bons résultats concernant $\sum U_l$, mais pas forcément sur T_{max}, car elle peut avoir tendance à accumuler le retard sur quelques lots. Quant à l'indicateur $\sum T_l$, il n'est pas nécessairement fortement corrélé avec $\sum U_l$ ou T_{max}. Cependant, on peut s'attendre à ce qu'une mauvaise solution pour $\sum U_l$ et T_{max}, soit également mauvaise pour $\sum T_l$.

De plus, une corrélation positive entre TP% et Wkld% peut également être attendue, puisqu'une augmentation de la productivité TP% implique plus d'étapes à traiter dans une période donnée, et donc plus de charge sur les machines. Toutefois, il convient de noter que cette corrélation n'est pas garantie. En effet, les temps de traitement variant d'une étape de process à l'autre, et d'une machine à l'autre, il peut être tentant, par exemple, de traiter de longues étapes de process sur des machines afin d'augmenter leur taux d'utilisation, au détriment d'autres étapes plus rapides. Ce choix montre qu'il est possible d'augmenter la charge des machines sans augmenter la productivité globale de l'usine. De façon symétrique, il est possible de favoriser des étapes de process rapides afin d'augmenter la productivité, sans augmenter la charge des machines.

En ce qui concerne le temps de cycle, nous avons vu dans le chapitre 2 qu'il existe une relation non linéaire entre le temps de cycle moyen des produits dans un système, et la charge de ce système. Plus précisément, les travaux de la littérature soulignent que le temps de cycle moyen des lots dans l'usine augmente (de façon non linéaire) avec le niveau d'en-cours (le WIP) dans l'usine. On peut donc s'attendre à ce que, si une règle augmente de manière significative le taux d'utilisation de la machine et/ou la productivité, cela se fasse au détriment du temps de cycle global. De plus, une corrélation positive entre CT et $\sum T_l$ peut également être attendue, puisqu'une réduction du temps de cycle moyen devrait permettre aux lots d'atteindre plus rapidement leur dernière étape de process, et donc de respecter plus facilement leurs dates de livraison.

5.2.4 Comparaison des règles de lissage des charges

Performances moyennes

Dans cette section, nous analysons les performances moyennes des différentes règles de lissage, selon les différents critères de comparaison présentés dans la section 5.2.3. Les résultats sont résumés dans les tableaux 5.5 et 5.6, montrant pour les 25 instances respectivement la valeur moyenne des solutions et la valeur maximale. Ainsi, le tableau 5.5 montre la performance moyenne de chaque règle, tandis que le tableau 5.6 montre les pires résultats pour les quatre premiers indicateurs, ainsi que les meilleures solutions trouvées (par rapport à la règle EDD) pour les indicateurs TP% et Wkld%. Les détails des résultats pour chaque critère sont disponibles dans l'annexe A.

TABLE 5.5 – Comparaison des règles de lissage (moyenne)

Règles	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	$CRPM$
$\sum T_l$	828,8	1475,7	1215,5	1496,5	**624,1**	1205,6	1125,6	946,9	1366,8
$\sum U_l$	363,8	333,4	**301,3**	496,3	405,7	442,6	324,2	312,6	479
T_{max}	9,8	11,4	10,6	10,2	**6,0**	12	10,1	9,1	11,5
CT	43,4	45,9	45,00	45,5	**42,6**	44,5	44,9	43,8	44,5
TP%	0,00%	0,78%	0,82%	0,58%	0,62%	**1,01**%	0,91%	0,81%	0,23%
Wkld%	0,00%	0,81%	1,19%	0,57%	0,93%	**1,75**%	1,62%	1,42%	0,21%

TABLE 5.6 – Comparaison des règles de lissage (maximum)

Règles	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	$CRPM$
$\sum T_l$	2528	3353	2981	3444	**1625**	2768	2897	2453	2711
$\sum U_l$	955	738	**682**	957	1030	991	718	775	865
T_{max}	22	18	18	18	**14**	21	17	16	24
CT	53,9	56,5	55,8	55,4	**51,9**	54,8	55,4	54,1	56,8
TP%	0,00%	2,80%	2,80%	2,60%	1,90%	**2,90%**	2,80%	2,50%	2,25%
Wkld%	0,00%	3,13%	3,13%	3,13%	3,13%	**5,13%**	3,13%	3,33%	3,32%

Tout d'abord, nous pouvons souligner les très bons résultats de la version de l'approche TSH basée sur la règle de lissage CR_{Post}. En effet, dans un grand nombre de cas, le lissage basé sur le *critical ratio prédit* donne les meilleurs résultats. Ceci est notamment visible sur les indicateurs $\sum T_l$ et T_{max}. Notons que la règle EDD, qui est la règle de base, donne également des résultats raisonnables. Cependant, malgré l'apparente efficacité de la règle CR_{Post}, les résultats sont plus mitigés pour l'indicateur $\sum U_l$. Ainsi, la règle CR_{Post} semble réussir à réduire le retard moyen et le retard maximal, mais aux dépens d'un plus grand nombre de lots livrés en retard. Concernant l'indicateur $\sum U_l$, les meilleurs résultats sont obtenus avec les règles ST et ST_{Post}. Ainsi, ne considérer les lots que du point de vue de leur retard absolu (et non par rapport à leur position) tend à minimiser le nombre de lots en retard. Mais cela se fait au détriment des retards globaux et maximums des lots. D'autre part, il n'est pas surprenant de constater des performances moyennes pour les indicateurs orientés clients, des règles considérant (uniquement ou en partie) l'impact des lots sur la charge des machines. On peut néanmoins souligner les performances correctes de la règle MI_{CR}, qui réussit parfois à obtenir les meilleures solutions, mais surtout permet d'obtenir des résultats généralement corrects et rarement très mauvais.

Pour les indicateurs non orientés clients, notons la bonne performance globale des règles considérant l'impact des lots sur la charge des machines (MI, MI_{ST} et MI_{CR}). Les meilleurs résultats sont obtenus par la règle totalement orientée machine, MI, avec une productivité moyenne supérieure de 1% à celle de la règle EDD, et de 1,75% supérieur concernant la charge moyenne des machines. Cependant, cette domination est limitée aux indicateurs TP% et Wkld%. En effet, pour le temps de cycle moyen des lots, c'est la règle CR_{Post} qui se démarque de nouveau, en fournissant les meilleurs résultats sur toutes les instances. Ce dernier résultat est d'ailleurs cohérent avec l'analyse des indicateurs présentée dans la section 5.2.3, où nous avions mentionné que la réduction du temps de cycle moyen des lots engendre assez mécaniquement la réduction du retard global des livraisons.

Notons également encore une fois la qualité des résultats obtenus à l'aide de la règle MI_{CR}, avec un bon compromis entre les trois indicateurs.

En revanche, force est de constater les faibles performances de la règle $CRMP$, présentée dans Mhiri et al. (2018), toujours dominée par d'autres règles, avec de plus les moins bons résultats sur les indicateurs T_{max}, TP% et Wkld%.

Lorsque l'on considère tous les indicateurs, les règles ST et CR sont globalement dominées par leurs variantes MI_{ST} et MI_{CR}. Cela signifie que la prise en compte de l'influence du déplacement d'un lot vers une nouvelle période améliore réellement la qualité des solutions. Par conséquent, les règles ST et CR ne seront plus prises en compte, de même que la règle $CRPM$. De plus, nous ne considérons pas dans la suite la règle MI_{ST}, dont les performances sont globalement moins bonnes que celles de la règle MI_{CR}.

Meilleures et pires performances

Nous venons de voir dans la section 5.2.4, qu'aucune règle de lissage ne domine totalement les autres sur les six indicateurs. Cependant, les résultats présentés précédemment sont agrégés sur l'ensemble des instances de test. Une information intéressante serait par exemple de savoir si, bien que certaines règles ne sont pas les meilleures pour certains critères, elles restent tout de même correctes car ne fournissent jamais de solution particulièrement mauvaises. À l'inverse, certaines règles peuvent présenter de bonnes performances moyennes sur certains critères, tout en fournissant les pires résultats (parmi l'ensemble des règles) sur d'autres critères. Ainsi, il nous a semblé intéressant d'évaluer la propension des différentes règles à fournir les meilleures, ou bien les pires solutions. Les résultats sont représentés dans les figures 5.1 et 5.2, à travers deux diagrammes de Kiviat (ou digramme "toile d'araignée"), où chaque axe correspond à un des indicateurs et où chaque règle est représentée par une couleur différente. La figure 5.1 indique le nombre de fois qu'une règle trouve la meilleure solution pour un indicateur (parmi les solutions des 5 règles de lissage candidates). Ainsi, une bonne règle est supposée couvrir une grande surface. Le maximum atteignable sur un axe est 25, ce qui signifie que la règle fournit toujours la meilleure solution pour l'indicateur associé. En revanche, La figure 5.2, indique le nombre de fois qu'une règle trouve la pire solution pour un indicateur. Une bonne règle est donc sensée couvrir une surface relativement réduite. Notons également que la somme sur chaque axe n'est pas nécessairement 25, puisque certaines des meilleures et pires solutions peuvent être obtenues par plusieurs règles.

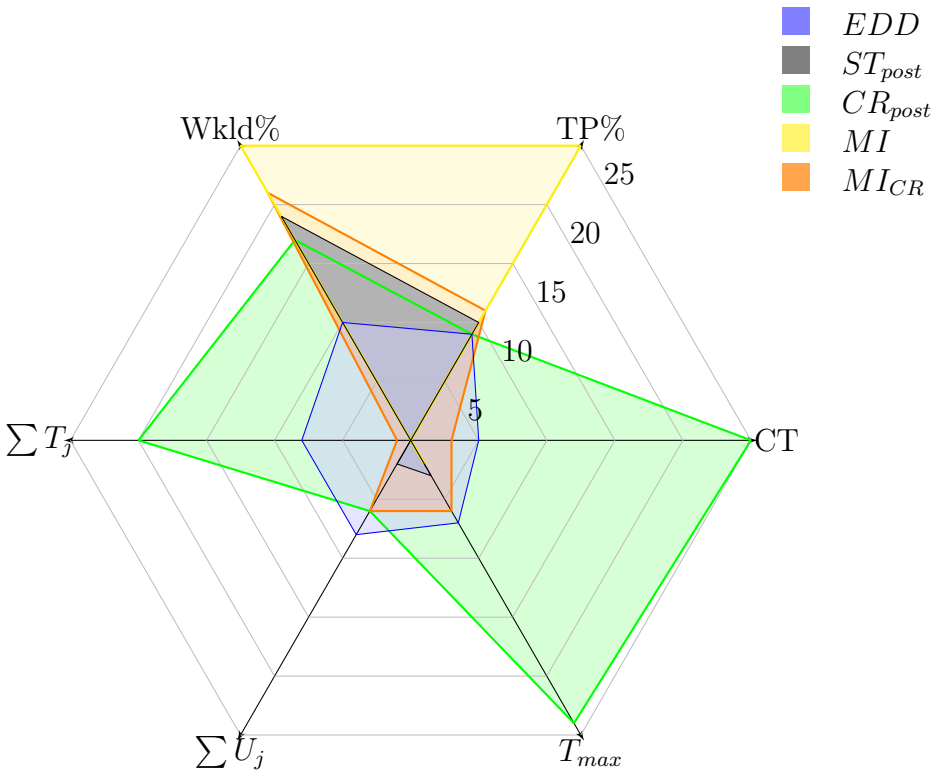


FIGURE 5.1 – Nombre de meilleures solutions par indicateur et par règle de lissage

La figure 5.1 permet de souligner à nouveau les bonnes performances de la règle CR_{Post}, qui conduit souvent aux meilleures solutions pour certains indicateurs. De plus, la figure 5.2 montre que la règle CR_{Post} aboutit rarement aux pires solutions. Cependant, il est intéressant de remarquer que, pour les indicateurs pour lesquels la règle CR_{Post} obtient les

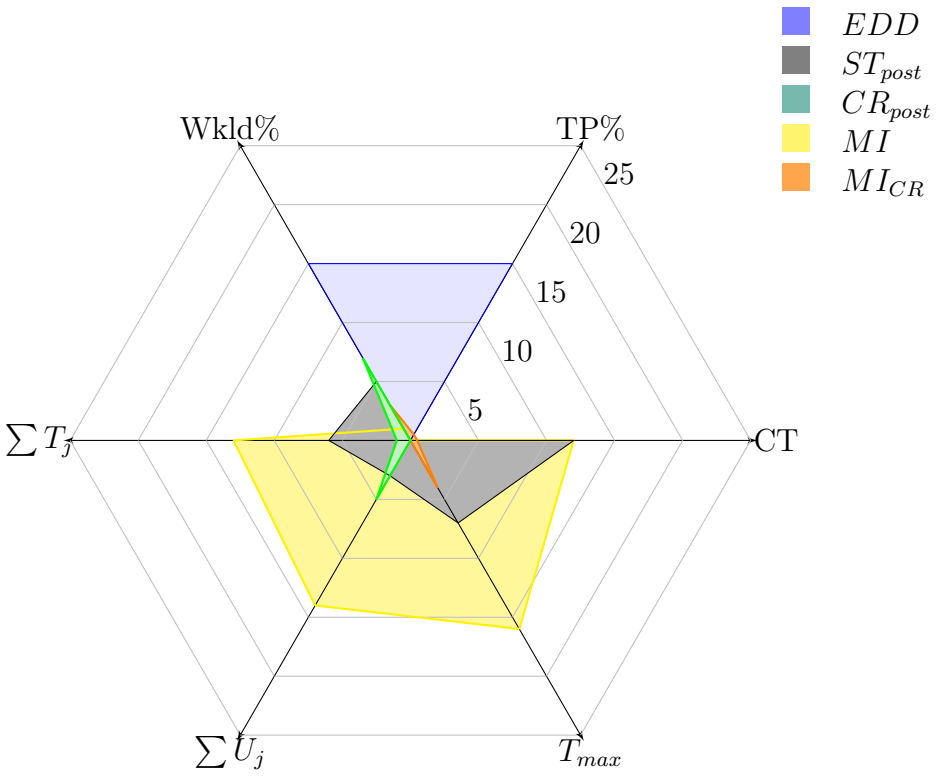


FIGURE 5.2 – Nombre de pires solutions par indicateur et par règle de lissage

pires solutions, celle-ci obtient souvent également les meilleures solutions, ce qui peut refléter une certaine variabilité dans la qualité des résultats obtenus par l'approche basée sur la règle de lissage CR_{Post}.

La figure 5.1 montre également les bons résultats de la règle EDD qui, pour chaque indicateur, arrive parfois aux meilleures solutions. Cependant, la figure 5.2 montre que cette règle EDD aboutit également 15 fois à la pire solution pour les indicateurs Wkld% et TP%, ce qui amène à un bilan mitigé.

Selon la figure 5.1, la règle MI obtient également de bons résultats en dominant les autres règles pour les indicateurs Wkld% et TP%. Cependant, la figure 5.2 souligne le fait que MI, qui ne considère pas la dimension clients, obtient régulièrement les pires solutions pour les 4 autres critères.

Enfin, l'analyse montre que la règle MI_{CR} donne de bons résultats pour les indicateurs Wkld% et TP% (même si elle n'est pas aussi efficace que la règle MI qui ne tient pas compte du retard des lots), mais elle obtient rarement les meilleures solutions pour les autres indicateurs. Cependant, la figure 5.2 montre que la règle MI_{CR} obtient très rarement les pires solutions (trois fois pour T_{max} et Wkld%) et n'est donc jamais dominée en termes de pires performances.

Influence des règles sur le temps d'exécution

La section précédente a permis de comparer les différentes règles de lissage selon la qualité des plans de production auxquelles elles permettent d'aboutir, qualité définie selon différents indicateurs de performance. Dans cette section, nous souhaitons évaluer l'influence de ces règles de lissage sur le temps d'exécution de l'approche TSH, qui a été exécutée sur un horizon de 12 périodes d'une semaine. Le tableau 5.7 résume le temps de calcul moyen de

l'approche en trois étapes sur les 25 instances et selon la règle de lissage. La première colonne donne le temps de calcul moyen et la deuxième colonne l'écart, en pourcentage, par rapport à la règle EDD, que nous considérons comme la règle de référence.

TABLE 5.7 – Temps d'exécution moyen de l'approche TSH en fonction de la règle de lissage

Règle de lissage	Temps de calcul (sec)	Écart(%)
EDD	104	+0%
ST	90	-13%
ST_{Post}	95	-9%
CR	88	-15%
CR_{Post}	104	+0%
MI	119	+15%
MI_{ST}	101	-2%
MI_{CR}	106	+2%
$CRMP$	109	+5%

Tout d'abord, notons que les règles de lissage ST et CR sont en moyenne plus rapides que la règle EDD, ainsi que leurs variantes respectives ST_{Post} et CR_{Post}. Ce dernier point peut s'expliquer de deux façons. Premièrement, l'évaluation à priori des indicateurs nécessite davantage de calculs. Ensuite, et surtout, parce que les règles ST_{Post} et CR_{Post} ont tendance à prioriser le déplacement des lots dont la réalisation de l'étape est prévue en fin de période. Ces étapes en fin de période ont tendance à engendrer moins de réduction de charge à chaque décalage, car peu d'étapes consécutives sont impliquées. Les règles ST_{Post} et CR_{Post} nécessitent donc généralement plus d'itérations pour assurer le respect de la capacité, ce qui les rend plus lentes en moyenne que leur version standard. Nous rappelons cependant que, d'après les analyses faites dans les sections 5.2.4 et 5.2.4, cette rapidité des méthodes de lissage ST et CR se fait souvent au détriment de la qualité des solutions. La règle $CRMP$ augmente légèrement le temps de calcul mais celui-ci reste tout de même proche de la moyenne générale. Ensuite, notons que la règle MI est celle qui nécessite le temps de calcul le plus long. Cela peut s'expliquer par le fait que la règle MI exige un prétraitement afin d'évaluer l'impact que chaque lot aurait sur la charge des machines s'il était reporté. Il convient également de noter que les règles MI_{ST} et MI_{CR} ne subissent pas d'augmentation significative du temps de calcul par rapport à la règle EDD. Le fait que les règles MI_{ST} et MI_{CR} ont un temps de calcul beaucoup plus court que la règle MI originale peut s'expliquer par le fait que, dans de nombreuses situations, aucun lot candidat pour être décalé n'est considéré comme en avance. Dans ce cas, la phase de prétraitement, qui prend beaucoup de temps pour évaluer l'impact des lots sur la charge des machines, n'est pas effectuée, et les règles MI_{ST} et MI_{CR} utilisent uniquement les règles ST_{Post} et CR_{Post}, qui sont plus rapides.

5.2.5 Recommandations

Les sections précédentes ont permis de montrer l'influence des règles de lissage sur la qualité des solutions fournies par l'approche TSH. Nous avons vu que certaines règles, telles que la règle CR_{post} basée sur un calcul prévisionnel de l'état d'avance/retard relatif des lots, semble globalement plus performante que d'autres. Mais les résultats montrent cependant qu'aucune règle ne domine totalement les autres sur l'ensemble des critères de comparaison. La section précédente a également permis de souligner l'influence des différentes règles de

lissage sur le temps d'exécution de l'approche TSH, montrant cependant que cette influence reste relativement limitée. Dans cette section, nous souhaitons faire le pont entre les résultats précédents et l'application industrielle de l'approche TSH et des règles de lissage dans un outil d'aide à la décision. Ainsi, quelles sont les règles de lissage recommandées pour la création de plans de production opérationnel? Selon les résultats, les trois règles à préférer sont CR_{Post}, MI et MI_{CR}. La règle CR_{Post} fournit les meilleurs résultats moyens pour les indicateurs orientés clients $\sum T_l$ et T_{max}, ce qui en fait une règle privilégiée si les gestionnaires sont principalement concernés par les engagements clients. Notons que la règle ST_{Post} est meilleure pour limiter le nombre de lots en retard $\sum U_l$, ce qui conduit cependant à une augmentation significative du temps du cycle moyen par rapport à la règle CR_{Post} (voir tableau 5.5), ce qui rend généralement la règle ST_{Post} moins préférable.

L'objectif principal n'est cependant pas toujours de minimiser les retards de livraison, mais peut être aussi de maximiser la productivité de l'usine, ou l'utilisation des machines. C'est notamment le cas lorsque les demandes sont très élevées et que la capacité de production est trop faible. Dans ce contexte, les retards peuvent devenir inévitables et les gestionnaires peuvent choisir de se concentrer sur la maximisation de la productivité globale de l'usine, en maximisant la productivité de l'usine et l'utilisation des machines. Dans ce cas, il est préférable d'utiliser la règle MI qui, bien qu'elle ne soit pas la meilleure pour minimiser les retards des clients (voir figure 5.2), domine les autres règles quand il s'agit de maximiser la productivité de l'usine et l'utilisation moyenne des machines. Cependant, cette règle induit une augmentation du temps de calcul par rapport aux autres règles. Ce temps de calcul reste toutefois court (environ deux minutes), et largement acceptable pour la création de plans de production de plusieurs semaines.

Enfin, l'utilisation de la règle MI_{CR} peut être recommandée en raison de sa bonne performance globale. En effet, bien qu'en moyenne cette règle ne soit jamais la meilleure pour aucun des indicateurs, elle reste généralement la deuxième ou la troisième meilleure. Seul le retard maximum semble poser des difficultés, mais la règle MI_{CR} reste efficace pour le retard global et le nombre de lots en retard. De plus, cela ne se fait pas au détriment du temps de calcul, ce qui fait de cette règle une bonne option pour l'équilibre entre tous les indicateurs.

5.2.6 Bilan de l'étude des méthodes de lissage de charge

La module de *step-shifting* de l'approche TSH a pour but de retarder la réalisation de certaines étapes de process de certains lots à une période ultérieure, afin de réduire la charge sur les machines ne respectant pas leur contrainte de capacité. Cette procédure peut être vue comme une simulation de l'attente de certains lots, non prioritaires, devant des machines limitantes. Le niveau de priorité d'un lot est défini à partir d'une règle de lissage, de façon tout à fait analogue aux règles de priorités utilisées dans les méthodes de *Dispatching*. Le choix de la règle influe sur l'ordre de passage des lots sur les machines, et a donc une influence sur le plan de production déterminé par l'approche TSH. Dans cette section, nous avons évalué l'influence de ces règles de lissage des charges sur la qualité des plans de production. Cette qualité n'est pas unique, et peut être analysée sous l'angle de différents indicateurs, parfois antinomiques, tels que la minimisation des retards de livraison ou la maximisation du taux d'utilisation des machines. Nous avons développé un ensemble de 9 règles, parfois proches de règles connues de *Dispatching*, telles que les règles *Earliest Due Date* ou *Minimum Slack Time*, et avons comparé leur performances selon les différents indicateurs de performances sur 25 instances réelles. Les conclusions sont que, bien qu'aucune règle de lissage ne domine les autres sur l'ensemble des critères de performances, certaines

se démarquent par leur très bonnes performances moyennes et par leur capacité à rarement fournir de très mauvaises solutions. C'est par exemple le cas de la règle CR_{post}, basée sur un calcul prévisionnel de l'état d'avance/retard relatif des lots, ou bien de la règle MI_{CR}, considérant en plus, partiellement, la charge des machines. Une analyse des temps d'exécution de l'approche TSH, selon les différentes règles de lissage, a montré que, bien que le choix de la règle a une influence sur ce temps d'exécution, ce dernier reste cependant toujours raisonnables dans le contexte d'utilisation prévu. Actuellement, la règle CR_{post} est intégrée dans l'approche comme règle par défaut. Mais nous travaillons à permettre aux gestionnaires de sélectionner facilement d'autres règles (en particulier les règles MI et MI_{CR}) en fonction des indicateurs qu'ils considèrent comme les plus importants quand l'approche est exécutée.

5.3 Lissage basé sur les *Balanced Group*

L'étude des règles de lissage de charge présentée dans la section 5.2 a permis de démontrer l'influence de ces règles dans la qualité des solutions déterminées par l'approche TSH, et de l'intérêt de bien choisir la règle à utiliser selon les critères que l'on souhaite optimiser. Ces règles sont intégrées dans le module de *step-shifting*, dont la procédure (présentée dans l'algorithme 2 de la section 3.4.4 du chapitre 3) possède cependant une limite dans l'évaluation de la charge et la capacité des machines, ce dont il est question dans cette section.

5.3.1 Limites de l'approche par machine

Le module de *step-shifting* de l'approche TSH considère individuellement le taux de charge de chacune des machines de l'usine. C'est d'ailleurs un des points principaux de notre travail de thèse afin de traiter le problème d'*Operational Production Planning*, en veillant à considérer l'ensemble des machines de l'usine pour une meilleure modélisation de la capacité (Horiguchi et al. (2001)). En plus d'une modélisation réaliste de la capacité, le besoin est de déterminer des plans de production détaillés, afin de donner des consignes suffisamment précises pour le pilotage global de l'usine, d'où la nécessité de modéliser la capacité de chaque machine individuellement.

Cependant, cette modélisation individuelle amène à une limite dans le fonctionnement du module de *step-shifting* quant au respect de la capacité des machines. Afin d'illustrer notre propos, la figure 5.3 représente un cas fictif d'équilibrage entre deux machines m_1 et m_2. Supposons que ces deux machines soient les deux seules d'un même *Isolated Group* (voir définition dans la section 3.4.3 du chapitre 3), noté $\mathcal{IG}$, et que nous nous plaçons dans une période quelconque de l'horizon de planification. Le module de *projection* a permis de définir les étapes de process à réaliser dans chaque période, et le groupe $\mathcal{IG}$, de par les qualifications des machines qui le compose, doit traiter trois recettes A, B et C, avec les quantités $\{q_A, q_B, q_C\} = \{25, 25, 75\}$. Ces quantités sont des multiples de 25, étant donné que, en fabrication de semi-conducteurs, les plaquettes sont généralement transportées dans des lots (appelés FOUP, voir figure 1.3 du chapitre 1) de capacité 25. La machine m_1 est qualifiée pour réaliser les recettes A et C avec un temps opératoire de 10 minutes par plaquette pour la recette A, et 5 minutes par plaquette pour la recette C. La machine m_2 quant à elle, peut réaliser les recettes B et C avec un temps opératoire de 5 minutes par plaquette. Les deux machines ont par ailleurs une capacité de 300 minutes durant la période considérée. Ces informations sont le point d'entrée du module d'*équilibrage* de l'approche TSH, qui détermine une répartition de la charge (les plaquettes) sur les machines. En l'occurrence, le

choix porte dans ce cas fictif sur la répartition des plaquettes de la recette C sur les machines m_1 et/ou m_2, dont une solution pourrait être celle proposée dans (a) dans la figure 5.3. Une étude approfondie du fonctionnement du module d'*équilibrage* est présentée dans le chapitre 4.

La solution d'équilibrage de (a) est un des points d'entrée du module de *step-shifting*, dont la procédure de lissage des charges est illustrée dans la suite de la figure 5.3. L'équilibrage des charges attribue à chaque machine une charge de 375 minutes. Or, chacune des machines possède une capacité de 300 minutes dans la période, les machines sont donc (également) surchargées. La procédure du module de *step-shifting* sélectionne la plus surchargée (nous choisissons arbitrairement ici la machine m_1) et choisit ensuite parmi les lots devant passer sur cette machine, celui étant le moins prioritaire suivant la règle de lissage utilisée (voir section 5.2). Supposons que le lot le moins prioritaire soit un lot associé à la recette A, la réalisation de l'étape de process du lot est alors repoussée à la période suivante, ce qui réduit la charge de la machine m_1 à 125 minutes, respectant ainsi sa capacité. En revanche, la machine m_2 est quant à elle toujours surchargée ((b) de la figure 5.3). La procédure de décalage est donc répétée, repoussant par exemple le lot associé à la recette B ((c) de la figure 5.3). Suite à ce décalage, toutes les machines ont une charge qui respecte leur capacité, et en répétant cette procédure pour l'ensemble des machines de l'usine, la procédure de *step-shifting* aboutit à un plan d'étapes de process à réaliser dans la période courante respectant les contraintes de capacité.

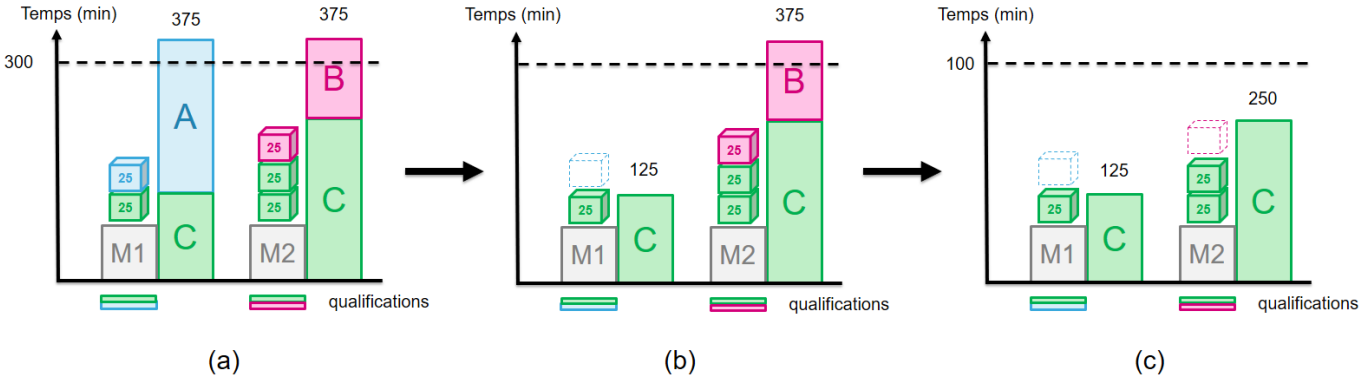


FIGURE 5.3 – Illustration de la procédure de lissage des charges du module de *step-shifting*

Cependant, le module de *step-shifting* a certes permis le respect de la capacité des machines, mais la solution finale ((c) de la figure 5.3) semble très en deçà de la solution initialement surchargée ((a) de la figure 5.3). En effet, force est de constater que la procédure de lissage aboutit à une solution où chacune des machines se retrouve avec une importante capacité inutilisée (175 minutes pour la machine m_1, 50 minutes pour la machine m_2), alors qu'il y avait a priori (d'après la solution initiale), suffisamment d'étapes de process à réaliser pour "occuper" ces machines durant la période ! En fait, une meilleure solution aurait été de ne pas procéder au lissage des charges de la machine m_2, car il était possible de transférer un peu de cette charge sur la machine m_1 nouvellement sous-chargée. Ainsi, en répartissant de nouveau la charge liée à la recette C à partir de la solution de l'image (b) de la figure 5.3, nous pouvons aboutir à une solution illustrée par l'image (c') de la figure 5.4, avec 25 plaquettes de recette B en plus par rapport à la solution (c) de la figure 5.3, obtenue via la

procédure de lissage illustrée.

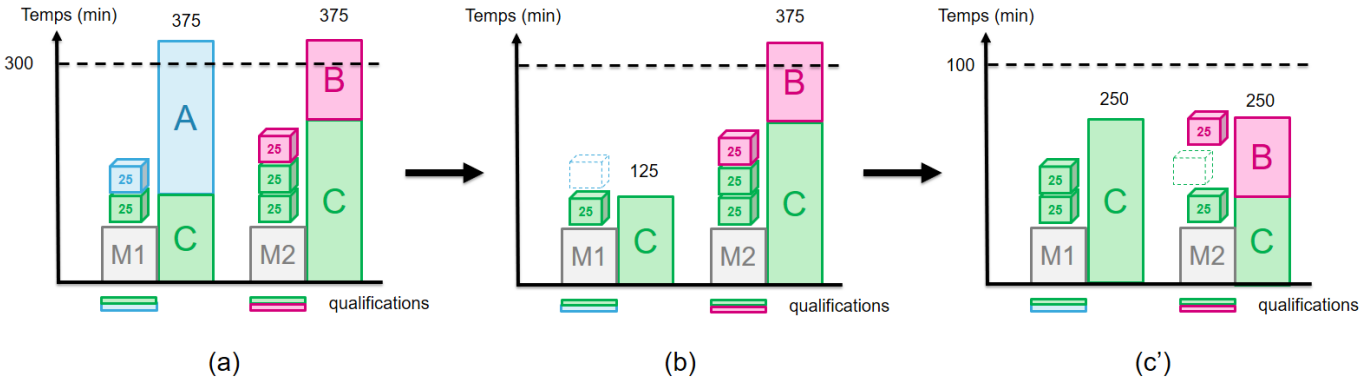


FIGURE 5.4 – Solution alternative par rééquilibrage, par rapport à la procédure de lissage des charges initiale du module de *step-shifting*

Une première solution évidente serait d'exécuter, après chaque décalage d'un lot, une procédure de rééquilibrage des charges. Cette procédure devrait être exécutée pour l'*Isolated Group* contenant la machine la plus chargée et qui est à l'origine du décalage, mais aussi pour l'ensemble des *Isolated Group*, potentiellement impactés par d'autres étapes de process du lot, et également réalisées dans la période. Cette solution de rééquilibrage systématique n'est cependant pas viable. En effet, si le nombre d'itérations réalisées dans la procédure de *step-shifting* était assez faible, par exemple quelques dizaines, exécuter le Programme Linéaire (PL) d'équilibrage à chaque itération serait éventuellement envisageable. Cependant, le nombre d'actions de décalage pour une période dans une exécution classique de l'approche *TSH* pour un problème industriel, varie généralement entre 500 et 3000. Ce nombre varie principalement par rapport au niveau de charge de l'usine par rapport à sa capacité. Ainsi, les instances issues de l'usine 200mm de Crolles amènent généralement le module de *step-shifting* de l'approche *TSH* à réaliser moins de 1000 itérations par période, tandis que les instances issues de l'usine 300mm de Crolles obligent parfois le module à itérer jusqu'à 3000 fois cette opération de décalage afin d'obtenir des plans respectant la capacité des machines. Par conséquent, effectuer un équilibrage des charges après chaque opération de décalage d'un lot reviendrait à exécuter plusieurs milliers de PL, ce qui augmenterait considérablement le temps de calcul, et ne serait pas en adéquation avec nos objectifs d'avoir un outil de planification opérationnelle très rapide.

Face à ce constat, nous avons voulu développer une nouvelle approche pour le module de *step-shifting*, basée sur la considération de groupes de machines nommés *Balanced Group*.

5.3.2 *Balanced Group* et règle de lissage associée

Le point de départ de l'approche est le constat que dans (b) de la figure 5.3, bien que la machine m_2 semble surchargée, la capacité globale des deux machines m_1 et m_2 (600 minutes) devrait pouvoir absorber la charge liée aux recettes B et C. L'idée de l'approche est donc de se concentrer sur la capacité *agrégée* des machines. Mais quelle agrégation de machines utiliser ? Utiliser des groupements par atelier amènerait à une approche trop agrégée de la capacité, ce que nous souhaitons justement éviter dans le problème d'*Operational*

Production Planning. Une autre possibilité serait de considérer la capacité au niveau des *Isolated Group*. Cependant, bien que cette agrégation a l'avantage de tenir compte des qualifications partagées des machines, l'équilibrage de ces dernières entre elles peut fortement varier en fonction des produits constituant l'en-cours de production (le mix-produit). Pour illustrer notre propos, reprenons l'exemple simplifié à 5 machines présenté dans le chapitre 4 et illustré dans la figure 4.1. Ces 5 machines sont dans le même IG, du fait du partage de certaines qualifications. Le ratio entre la charge totale à distribuer sur les machines, et la capacité cumulée de ces dernières, devrait donner une valeur inférieure à 1, car les machines 4 et 5 sont fortement sous-chargées. Pourtant, nous avons vu dans la chapitre 4 que la machine 1 est nécessairement surchargée, du fait de la grande quantité de plaquettes de recette A que la machine est chargée de traiter. Mais dans d'autres contextes, par exemple dans un cas où les en-cours amèneraient à devoir traiter davantage de recette D, et moins de recette A, alors l'évaluation d'un ratio charge/capacité inférieur à 1 serait réaliste car les machines pourraient s'équilibrer pour ne pas être surchargées. Cependant, une telle évaluation nécessite de tenir compte des jeux de qualifications, des volumes spécifiques de chaque recette, de la variété des capacités et des durées opératoires des machines, ce qui est difficilement faisable sans l'utilisation de méthodes telles que la Programmation Linéaire.

Or, nous souhaitons une approche qui ne requiert pas l'utilisation d'un PL pour chaque modification des quantités à équilibrer du fait des décalages opérés par le module de *step-shifting*. Pour cela, nous nous sommes intéressés à un autre groupement des machines, basé sur la notion de file d'attente partagée. Si l'on reprend l'exemple présenté dans la figure 4.1, une vision simplifiée serait d'imaginer que tous les lots générant la charge attendent devant un groupe unifié de 5 machines. Or, si l'on regarde l'équilibrage en (b) (que nous avions prouvé être la meilleure solution d'équilibrage), on remarque que la machine 1 n'est supposée réaliser que la recette A. De même, bien que toutes les machines puissent réaliser la recette D, seules les machines 4 et 5 sont chargées de le faire. Ainsi, une vue plus affinée de la répartition des files d'attente de recettes consiste à imaginer une file d'attente constituée de plaquettes de la recettes A devant la machine 1, des recettes B et C devant les machines 2 et 3, et de le recette D devant les machines 4 et 5. Cette vision par file d'attente est intéressante car, d'un point de vue capacité, on comprend que si la file d'attente liée à la recette D diminue, cela ne doit pas affecter les machines 1, 2 et 3 qui sont déjà "occupées" par leur files d'attente dédiées. Ainsi, la charge liée à la recette D ne doit pas influer dans la mesure charge/capacité pour les machines 1, 2 et 3. Ce groupement des machines par files d'attente communes semble donc être une base intéressante afin d'agréger la capacité.

Mais comment regrouper rapidement les machines partageant la même file d'attente ? Pour cela, nous nous basons directement sur les propriétés de la méthode d'équilibrage IMM, présentée dans la chapitre 4. Dans ce chapitre, nous avons dégagé plusieurs propriétés des solutions fournies par l'IMM, c'est-à-dire les solutions *min-max fair*, dont une particulièrement intéressante pour notre cas qui est que : Dans une solution d'équilibrage *min-max fair*, si deux machines traitent au moins une recette en commun, alors elles ont la même charge. Cela signifie donc que, suite à l'équilibrage effectué par la méthode IMM utilisée par le module *d'équilibrage*, si deux machines traitent au moins une recette en commun (partagent donc la même file d'attente), alors elles ont la même charge. Ainsi, le regroupement des machines partageant les mêmes files d'attente sont les machines ayant la même charge à l'issue du processus d'équilibrage (sous réserve d'utiliser la méthode IMM). Ces groupes de machines équilibrées entre elles à l'issue de la méthode IMM sont nommés *Balanced Group* (BG). À noter toutefois qu'il existe un cas particulier où deux machines peuvent avoir la même charge sans pour autant s'équilibrer entre elles (la propriété de l'IMM énoncée n'est

pas une équivalence mais une implication). Dans ce cas, l'agrégation de la capacité de ces machines pourrait amener à des conclusions erronées. Cependant, ce cas est extrêmement rare dans la pratique et le regroupement des machines ayant la même charge reste une bonne approximation pour regrouper les machines partageant les mêmes files d'attente.

Ainsi, nous avons défini les *Balanced Group* comme le groupement des machines (parmi les machines d'un même *Isolated Group*) ayant la même charge à l'issue de l'équilibrage effectué par la méthode IMM. Notre approche consiste alors à évaluer la capacité et la charge dans la procédure de *step-shifting*, sous le prisme de ce nouveau regroupement des machines. Nous rappelons d'abord la définition du taux de charge d'une machine m introduite dans le chapitre 4 :

$$W_m = \frac{\sum_{r \in \mathcal{R}} a_{r,m} X_{r,m}}{c_m}$$

Où $X_{r,m}$ est la quantité de la recette r affectée à la machine m, $a_{r,m}$ est le temps requis pour réaliser une plaquette de la recette r par la machine m, et c_m la capacité de la machine m. On prolonge alors cette formule pour définir le taux de charge d'un ensemble de machines appartenant à un même *Balanced Group* (BG) :

$$W_{BG} = \frac{\sum_{m \in BG;\ r \in \mathcal{R}} a_{r,m} X_{r,m}}{\sum_{m \in BG} c_m}$$

La charge associée à un *Balanced Group* (BG) est donc le rapport entre le cumul des charges de l'ensemble des machines qui composent ce groupe, et la somme des capacités de ces mêmes machines.

Prenons pour exemple la figure 5.3 présentée dans ce chapitre, plus précisément l'équilibrage initial (avant processus de lissage) présenté dans (a). Les deux machines m_1 et m_2 ont la même charge, elles font donc partie du même *Balanced Group* (BG). Le calcul du taux de charge du groupe (BG), avant lissage, donne $W_{BG} = (10 \times q_a + 5 \times q_b + 5 \times q_c)/((c_1 + c_2) = (250 + 125 + 375)/(2 \times 300) = 1,25 > 1$. Le *Balanced Group* est donc considéré comme étant surchargé.

Comme pour l'approche initiale du module de *step-shifting* basée sur la charge de chaque machine, l'objectif de la nouvelle approche est que le taux de charge de chaque *Balanced Group* soit inférieur ou égale à 1. Dans cette procédure, lorsque la réalisation de l'étape d'un lot est reportée à la période suivante, au lieu de réduire la charge d'une machine spécifique, le taux de charge global du *Balanced Group* est réduit. Par exemple, si l'étape de process du lot décalé générait (d'après l'équilibrage initial) une charge y_l (dont la dimension est un temps), alors la nouvelle charge W'_{BG} du *Balanced Group* (BG) est :

$$W'_{BG} = \frac{\sum_{m \in BG;\ r \in \mathcal{R}} (a_{r,m} X_{r,m}) - y_l}{\sum_{m \in BG} c_m} = W_{BG} - \frac{y_l}{\sum_{m \in BG} (c_m)} = W_{BG} - w_l$$

Où w_l est la portion de capacité du *Balanced Group* (BG) qui était allouée pour réaliser l'étape du lot l.

5.3.3 Modification de la procédure de lissage

La procédure de lissage des charges basée sur les *Balanced Group* reste proche de la procédure initiale considérant individuellement le taux de charge de chaque machine, et nous présentons les principales différences dans cette section. L'algorithme 4 présente la

Algorithme 4 : Procédure modifiée du module de *step-shifting* par *Balanced Group*

Input : t = Période courante
$X_{s,l,t}$, $Q_{s,l,m,t}$ = Valeurs définies pour chaque étape de process de chaque lot et chaque machine
$W_{BG,t}$ = Charge du *Balanced Group* BG durant la période t
Output : $\mathcal{O}_{shifted}$ = Ensemble des étapes de process décalées à la période suivante

$\mathcal{O}_{shifted} \leftarrow \emptyset$
while $\max_{BG \in \mathcal{BG}} W_{BG,t} > 1$ **do**
  // Sélectionne le *Balanced Group* le plus chargé
  $BG' \leftarrow \text{argmax}_{BG \in \mathcal{BG}} W_{BG,t}$
  $\mathcal{O}^{BG'} \leftarrow \{s \in \mathcal{O}_l; m \in BG; \forall l \in \mathcal{L} \mid Q_{s,l,m,t} > 0\}$
  // Sélectionne parmi les lots associés au *Balanced Group*, celui dont la période de livraison est la plus éloignée
  $s' \leftarrow \text{argmax}_{s \in \mathcal{O}^{BG}}(dd_l)$
  for $s \in [o', ..., S_l]$ **do**
    if $X_{s,l,t} = 1$ **then**
      $X_{s,l,t} \leftarrow 0$
      $\mathcal{O}_{shifted} \leftarrow \mathcal{O}_{shifted} \cup \{s\}$
    for $m \in \mathcal{M}$ **do**
      $Q_{s,l,m,t} \leftarrow 0$
    end
  end
  for $BG \in \mathcal{BG}$ **do**
    $W_{BG,t} \leftarrow \frac{\sum_{s \in \mathcal{S}_l, m \in BG}(a_{s,l,m} Q_{s,l,m,t})}{\sum_{m \in BG} c_m}$
  end
end

procédure modifiée du module de *step-shifting*, en soulignant (en vert) les différences par rapport à la version initiale.

Comme dans la version initiale de la procédure de *step-shifting*, les charges associées aux différentes étapes de process des différents lots sont affectées au différentes machines (variables $Q_{s,l,m,t}$). Cependant, le taux de charge W est calculé par *Balanced Group* et non pas par machine. Le processus itératif de lissage consiste à d'abord évaluer si le taux de charge du *Balanced Group* le plus chargé est supérieur à 1. Si ce n'est pas le cas, la procédure de lissage s'arrête car le plan est considéré comme respectant les contraintes de capacité. Si en revanche, le *Balanced Group* le plus chargé possède un taux de charge supérieur à 1, alors la procédure détermine l'ensemble des étapes de process passant sur l'une des machines du *Balanced Group* durant la période considérée. Parmi cet ensemble d'étapes de process, la procédure va alors sélectionner celle dont le lot associé est le moins prioritaire. Cette priorité est définie par la règle de lissage utilisée (voir détail dans la section 5.2). Dans le cas de l'algorithme 4 présenté comme exemple, le lot le moins prioritaire est celui dont la date de livraison est la plus éloignée (règle *Earliest Due Date*). Une fois le lot le moins prioritaire sélectionné, la réalisation de son étape de process (et de toutes les étapes suivantes) est reportée à la période suivante. La charge engendrée par la réalisation de cette étape de process (et des suivantes) est alors retirée des *Balanced Group* concernés, mettant à jour le taux de charge de tous les *Balanced Group*. Une fois le décalage effectué, le processus est répété jusqu'à ce que plus aucun *Balanced Group* ne soit surchargé.

Si nous appliquons ce processus à notre exemple de la figure 5.3, après avoir décalé le premier lot (situation (b)), la nouvelle charge agrégée sera $W'_{BG} = 1,25 - (250/600) \simeq 0,83$. La valeur du taux de charge du *Balanced Group* étant inférieure à 1, ce dernier n'est plus considéré comme surchargé et le processus de lissage cesse pour ce groupement de machines. La charge globale étant en effet inférieure à la capacité cumulée des deux machines, nous pouvons supposer qu'il existe un nouvel équilibrage (que nous n'exécutons cependant pas pour des raisons de temps de calcul) évitant aux machines d'être surchargées. Cette possibilité de rééquilibrage est illustrée par (c') de la figure 5.4.

5.3.4 Avantages de l'approche de lissage par *Balanced Group*

L'approche de lissage des charges par *Balanced Group* présente plusieurs avantages par rapport à l'approche initiale par machine. Premièrement, l'agrégation des machines en groupes permet de réduire un peu la complexité de l'algorithme de *step-shifting* quant à la recherche du groupe le plus chargé (ligne 3 de l'algorithme 4), en passant d'un facteur $|\mathcal{M}|$ à un facteur $|\mathcal{B}| \leq |\mathcal{M}|$. Cette réduction de temps est cependant limitée, car le principal facteur limitant est le nombre d'étapes de process décalées (algorithme de *step-shifting* quadratique en $O(|\mathcal{S}|^2)$). Ainsi, le principal facteur de la réduction du temps de calcul vient de la réduction du nombre d'étapes de process subissant un décalage. Car nous avons vu que cette approche tend à réduire le nombre d'étapes de process décalées, c'est à dire le nombre d'itérations requises pour obtenir un plan de production qui respecte les contraintes de capacité. Cette réduction du nombre d'itérations permet donc de réduire mécaniquement le temps d'exécution du module de *step-shifting*.

Ensuite, en plus de la réduction du temps de calcul, cette réduction du nombre de lots décalés a également un impact positif sur la qualité des solutions obtenues par l'approche TSH. En effet, chaque opération de décalage ne peut que dégrader la solution courante. Décaler un lot signifie moins d'étapes de process réalisées dans la période, donc une dégradation de la productivité et de la charge moyenne de l'usine. De plus, décaler un lot signifie égale-

ment le retarder, donc amène à un risque de dégradation des indicateurs liés aux retards de livraisons, mais aussi le temps de cycle moyen des lots. Par conséquent, réduire le nombre de lots décalés engendre également une amélioration de la qualité des solution obtenues. Cette nouvelle version du lissage des charges présente donc des gains à la fois en termes de temps de calcul et en termes de qualité des solutions, ce que nous allons évaluer dans la section suivante.

5.3.5 Résultats expérimentaux

Dans cette section, nous souhaitons évaluer expérimentalement les performances de la procédure de lissage basée sur les *Balanced Group* par rapport à la procédure initiale considérant chaque machine individuellement. Pour ce faire, nous nous sommes appuyés sur un ensemble de 15 instances industrielles de taille réelle. Les principaux paramètres sont le nombre de machines et le nombre de lots dans l'usine, dont le tableau 5.8 résume la plage de variation.

TABLE 5.8 – Caractéristique des instances industrielles

Instance	Nb de Machines	Nb de Lots
Maximum	374	3172
Moyenne	357	3040
Mininimum	334	2830

Pour chaque instance et pour chacune des deux procédures (par *Balanced Group* ou par machine), nous exécutons sur huit périodes d'une semaine l'approche TSH. Les solutions sont comparées selon différents critères, et les résultats sont présentés dans le tableau 5.9. Le premier indicateur, nommé "Nb Décalages", représente le nombre d'opérations de décalage effectuées par le module de *step-shifting*, sur l'ensemble de l'horizon, afin de lisser les charges. Le but de cet indicateur est d'évaluer dans quelle mesure l'approche par *Balanced Group* permet de réduire le nombre d'itérations dans le module de *step-shifting*. Notons par ailleurs que plusieurs opérations de décalage peuvent correspondre à un même lot. La colonne "Gap" indique la réduction (ou l'augmentation) du nombre d'opérations de décalage, en utilisant la procédure de lissage par *Balanced Group* plutôt que la méthode initiale par machine.

Le deuxième indicateur est le retard total pondéré (TWT), qui cumule sur chaque lot le nombre de périodes de retard. L'idée est ici d'évaluer si la nouvelle approche permet d'améliorer la qualité des plans de production fournis.

Le troisième indicateur, appelé "Rupture Capacité", vise à évaluer la capabilité de l'approche par *Balanced Group* à effectivement respecter les contraintes de capacité. En effet, nous rappelons que ces méthodes reposent sur l'hypothèse que, si une machine d'un *Balanced Group* est surchargée mais que le taux de charge du *Balanced Group* est inférieur à 1, alors il doit exister un arrangement des charges permettant de respecter la capacité de toutes les machines. Comme les *Balanced Group* sont construits de telle sorte que les machines sont supposées s'équilibrer, cette hypothèse devrait être valide, c'est ce que l'indicateur "Rupture Capacité" doit évaluer. Pour ce faire, à la fin de chaque module de *step-shifting*, un nouveau module d'équilibrage est exécuté afin de répartir de nouveau la charge entre les machines. L'indicateur "Rupture Capacité" mesure la proportion de machines qui, après ré-équilibrage, possède encore un taux de charge supérieur à 1.

Enfin, l'indicateur "CPU Times" donne le temps de traitement total en secondes requis par le module de *step-shifting* pour les huit périodes.

TABLE 5.9 – Comparaison des performances entre l'approche de lissage des charges par machine (Ind) et par *Balanced Group* (BcdGrp)

	Nb Décalages			TWT		Rupture capacité		CPU Time (sec)	
	Ind	BcdGrp	Gap	Ind	BcdGrp	Ind	BcdGrp	Ind	BcdGrp
1	14757	**14011**	-5%	10	**0**	0	4,55%	8,80	**8,07**
2	14910	**11536**	-23%	**0**	**0**	0	3,90%	9,22	**7,64**
3	13867	**8531**	-38%	5	**0**	0	6,07%	9,41	**7,13**
4	15303	**9525**	-38%	**0**	**0**	0	3,97%	6,76	**5,87**
5	**19212**	20517	7%	1	**0**	0	9,31%	**8,58**	16,99
6	17468	**11887**	-32%	**0**	**0**	0	11,99%	11,34	**7,64**
7	**18095**	18977	5%	2	**1**	0	11,18%	**12,96**	14,29
8	16755	**16034**	-4%	**1**	**1**	0	5,02%	9,89	**9,41**
9	12303	**7389**	-40%	**0**	**0**	0	3,57%	6,64	**6,29**
10	10535	**3451**	-67%	**0**	**0**	0	12,60%	4,16	**1,80**
11	**35405**	54515	54%	2	**0**	0	21,49%	**31,99**	33,98
12	11358	**7279**	-36%	**0**	**0**	0	8,64%	7,21	**5,15**
13	35948	**14637**	-59%	787	**17**	0	1,83%	141,92	**13,86**
14	12165	**6718**	-45%	1	**0**	0	5,18%	8,01	**3,30**
15	12943	**2977**	-77%	337	0	0	5,47%	4,69	**1,61**

Considérant les résultats du tableau 5.9, nous notons d'abord que la procédure de lissage par *Balanced Group* permet de réduire le nombre moyen d'opérations de décalage. En effet, bien que dans certains cas la méthode de lissage initiale (par machine) nécessite moins d'étapes de déplacement que la nouvelle approche, cette dernière permet en moyenne une réduction de 27% du nombre de décalages. Cet effet permet notamment d'expliquer la réduction globale du temps de calcul avec un temps moyen inférieur de 19% avec la procédure de lissage par *Balanced Group* par rapport à l'approche initiale. La diminution du nombre de lots décalés implique également un meilleur respect des dates de livraisons, et conduit donc à des solutions toujours meilleures en matière de TWT, parfois avec des différences importantes comme pour les instances 13 et 15.

Cependant, ces résultats doivent être nuancés en ce qui concerne le nombre de machines (7,5%, soit 27 machines en moyenne) pour lesquelles la charge reste supérieure (+7% en moyenne) à leur capacité, même après ré-équilibrage de la charge. Cet effet est dû au fait qu'à la suite de la procédure de lissage, les changements de proportions de produits à traiter (le mix-produit) ne permettent plus à certaines machines de s'équilibrer entre elles, et donc de décharger certaines des machines les plus chargées.

Car, nous le rappelons, l'approche de lissage par *Balanced Group* repose notamment sur l'hypothèse forte que les machines peuvent toujours s'équilibrer même après une modification de la charge suite au décalage de lots. C'est grâce à cette hypothèse que l'on peut se permettre d'évaluer le respect de la capacité des machines composant un *Balanced Group*, en divisant simplement la charge totale par la capacité de l'ensemble des machines. Toutefois, cette hypothèse d'équilibre n'est pas toujours garantie. Afin d'illustrer notre propos, prenons l'exemple de la figure 5.5, qui est une variante de l'exemple présenté dans les figures 5.3 et 5.4. Ici, nous avons toujours un *Balanced Group* composé de deux machines m_1 et m_2, avec cette fois une capacité de 200 minutes chacune pour la période, mais toujours avec le même équilibrage initial. Supposons qu'à l'issue du processus de lissage opéré par le module de *step-shifting*, le *Balanced Group* passe de la configuration (a) à la configuration (b). La

machine m_1 est surchargée, mais si l'on considère le taux de charge du *Balanced Group*, on obtient alors $W_{BG} = (250 + 125)/(2 \times 200) \simeq 0,94 < 1$. Par conséquent, le lissage pour ce *Balanced Group* n'est a priori plus nécessaire car le taux de charge est inférieur à 1, il doit donc y avoir une nouvelle répartition de la charge permettant de respecter la capacité des machines. Or, on constate que dans (b) les machines ne peuvent plus s'équilibrer et qu'il n'est pas possible de transférer une partie de la charge de la machine m_1 vers m_2 afin d'éviter à m_1 d'être surchargée. Les machines constituant le *Balanced Group* ne peuvent plus s'équilibrer entre elles, cette hypothèse est pourtant au coeur de l'approche de lissage par *Balanced Group*, ce qui amène à des cas de non respect de la capacité de certaines machines. Mais pourquoi les machines de certains *Balanced Group* ne sont plus capables de s'équilibrer entre elles ? La raison principale est le changement du mix-produit, c'est à dire les proportions de chaque produit (plus précisément de chaque recette dans le cas de l'équilibrage) dans l'en-cours de production. Dans le cas de la figure 5.5, les décalages successifs ont rendu la recette A prépondérante en termes de quantités et la recette C, qui liait les deux machines, est désormais en trop faible quantité, empêchant les deux machines de s'équilibrer. Ainsi, les variations importantes de mix-produit a un impact sur l'équilibre des machines en diminuant la proportion des recettes dont les qualifications sont partagées par les machines. Les variations du mix-produit impactent également cette capabilité d'équilibre des machines du fait de la variété des temps de process entre les machines. En effet, suite au décalage d'un lot, si le taux de charge du *Balanced Group* considéré est inférieur à 1, cela suppose qu'on peut transférer la charge (donc du temps de process) de certaines machines vers d'autres. Cette charge, liée aux quantités de recettes et au temps de process requis par les machines, n'est pas réévalué. Or, si l'on transfère des quantités de recettes vers une machine qualifiée pour la réaliser, mais avec un temps de process extrêment long, ce rééquilibrage risque de beaucoup augmenter la charge totale. Cet effet n'est pas pris en compte lors de l'évaluation du taux de charge des Balanced Group, et est d'autant plus significatif que les variations du mix-produit sont importantes.

Ces variations de mix-produit sont d'autant plus importantes que l'usine est très chargée par rapport à sa capacité de production. En effet, dans les cas de sous-capacité, le module d'équilibrage fournit des solutions où le taux de charge est très supérieur à 1, et où le *step-shifting* doit donc effectuer beaucoup d'opérations de décalage afin de faire en sorte que le plan respecte la capacité des machines. Ce nombre important d'opérations de décalage entraine une variation importante du mix-produit, et c'est pour cela que ces cas de rupture de capacité sont plus nombreux dans les instances où la capacité de l'usine est particulièrement limitée.

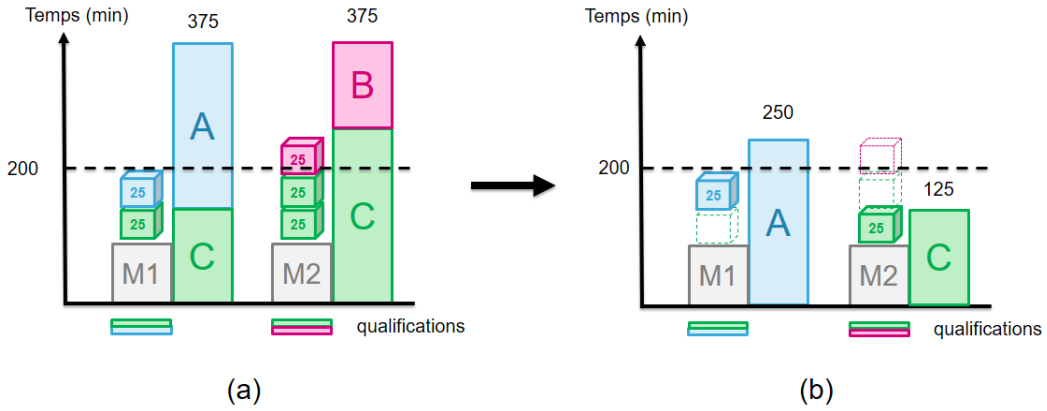


FIGURE 5.5 – Cas problématique pour l'approche de lissage par *Balanced Group*

5.3.6 Bilan de l'approche de lissage des charges par *Balanced Group*

Nous avons mis en évidence dans la section 5.3 les limites de l'approche de lissage par machine, en cela qu'elle a parfois tendance à réduire la charge sur les machines de façon exagérée par rapport à leur capacité réelle. Nous avons souligné que cette faiblesse venait du fait que le lissage par machine ne considérait pas la possibilité de réduire le taux de charge d'une machine en transférant un peu de sa charge sur une autre machine moins chargée. Bien que la solution idéale serait d'effectuer un équilibrage après chaque opération de décalage d'un lot, nous avons souligné que cette solution n'était pas possible car trop gourmande en temps, et par conséquent nous avons introduit une nouvelle approche de lissage basée sur les *Balanced Group*. Ces derniers sont des groupes de machines partageant les mêmes files d'attente, et qui s'avèrent être celles ayant le même taux de charge à l'issue de l'équilibrage des charges par la méthode IMM. Une étude expérimentale a montré les performances prometteuses de l'approche par *Balanced Group*, que ce soit en termes de qualité des solutions ou de temps de calcul, par rapport à la méthode initiale basée sur le taux de charge individuel des machines. Ces résultats sont cependant à nuancer, car certains cas de machines encore surchargées, suite au processus de lissage, sont constatés. Ces situations sont principalement causées par les variations importantes du mix-produit qui perturbe la capabilité des machines à s'équilibrer entre elles, et ces variations sont d'autant plus importantes que l'usine de production est dans une situation de sous-capacité. Par conséquent, bien que l'approche de lissage par *Balanced Group* semble prometteuse, des études complémentaires doivent être menées afin de vérifier que les gains ne sont pas seulement liés aux excès de charge de certaines machines.

5.4 Lissage des charges par anticipation

5.4.1 Améliorer la solution courante

Dans ce chapitre dédié au module de *step-shifting* de l'approche TSH, nous avons vu dans un premier temps dans la section 5.2 plusieurs règles de lissage permettant d'optimiser différents critères de performance. Puis, nous avons étudié dans la section 5.3 comment éviter un processus de lissage trop fort grâce à une évaluation intelligente du taux de charge sous forme de groupes de machines. Cependant, les approches de lissage présentées dans ces deux

sections ont toutes le même défaut, elles ne peuvent que dégrader la solution courante. En effet, que ce soit par l'approche de lissage par *Balanced Group* ou l'approche de lissage par machine, quelle que soit la règle utilisée, l'opération élémentaire est le décalage de lots à une période ultérieure. Ce décalage dégrade nécessairement la qualité de la solution courante, que ce soit du point de vue des critères orientés clients (retard total, maximum, ...) ou ceux orientés production (productivité, taux de charge, ...). Par conséquent, le processus de *step-shifting*, ne peut adapter une solution initiale non réalisable en une solution finale réalisable qu'en dégradant continuellement la solution de départ.

Ainsi, nous souhaitons aborder dans cette section une dernière approche permettant de modifier une solution courante, mais dans le but de l'améliorer. L'idée est en fait symétrique à l'approche de lissage que nous avons jusque-là présentée et que l'on nomme généralement *lissage des charges vers l'avant* (ou lissage avant). Dans le *lissage des charges vers l'avant*, l'idée est de repousser des étapes de process à des périodes ultérieures, afin de lisser la charge entre les périodes et donc de respecter les contraintes de capacité. Dans le module de *step-shifting* cela signifie que, pour une période donnée, si une machine est surchargée, alors un lot sera décalé de la période courante vers la période suivante. À l'inverse, l'objectif de notre approche est de détecter, pour une période considérée, des situations où des machines ont un excès de capacité, c'est à dire des machines pour lesquelles de la capacité est disponible et pour lesquelles il serait possible de *ramener* des lots de la période suivante vers la période courante, sans excéder les contraintes de capacité. Nous appelons cette approche *lissage par anticipation*. Nous faisons cette distinction avec le terme *lissage vers l'arrière* car ce dernier sous-entend l'idée d'un lissage des charges d'une période donnée en décalant la charge vers une période antérieure. Dans notre approche, les périodes antérieures à la période courante ne sont jamais remises en question. Le terme lissage par anticipation est donc plus approprié, soulignant le fait que l'on souhaite augmenter la charge de la période courante en *anticipant* la réalisation d'étapes de process initialement prévues dans une période ultérieure.

5.4.2 Le processus de lissage des charges par anticipation

L'objectif de cette nouvelle approche de lissage par anticipation est de pouvoir améliorer la solution courante en anticipant la réalisation de certaines étapes de process initialement prévues dans la période suivante. Elle ne permet cependant pas de réduire la charge des machines dans la période courante, et notamment les machines surchargées. L'approche de lissage par anticipation n'est donc pas une alternative à l'approche de lissage vers l'avant, mais plutôt une procédure complémentaire. Ainsi, en reprenant l'algorithme 2 du module de *step-shifting* présenté dans le chapitre 3, nous présentons l'algorithme modifié 5 incluant la composante de lissage par anticipation (surlignée en vert).

La procédure de lissage par anticipation requiert certaines données en entrée, en plus de celles déjà utilisées pour le lissage vers l'avant. Premièrement, la nouvelle procédure de lissage nécessite la liste des étapes de process prévues dans la période suivante $t + 1$, et pouvant être réalisées dans la période courante t. Cela signifie que si l'on avance la réalisation d'une étape de process une période plus tôt (donc en réduisant le temps d'attente devant la machine), il faut vérifier que la nouvelle date de démarrage de l'étape de process respecte le délai minimum d'attente avec la fin de l'étape de process précédente. Dans cette étude préliminaire, visant avant tout à évaluer le gain potentiel lié à l'ajout de cette procédure de lissage par anticipation (au regard du temps additionnel d'exécution requis), ces étapes de process candidates sont celles les plus proches de la frontière entre les deux périodes. Plus précisément, nous considérons comme candidate à l'anticipation, la première étape de process

Algorithme 5 : Procédure du module de *step-shifting* avec composante de lissage par anticipation

Input : t = Période courante
$X_{s,l,t}$, $Q_{s,l,m,t}$ = Valeurs définies pour chaque étape de process de chaque lot et chaque machine
$W_{m,t}$ = Charge de la machine m durant la période t
$\mathcal{M}^{s,l}$ = Ensemble des machines qualifiées pour réaliser la recette associée à l'étape de process s du lot l
$\mathcal{O}^{next}$ = Ensemble des étapes de process prévues à la période suivante et pouvant être réalisées à la période courante en respectant les contraintes de précédence
Output : $\mathcal{O}_{shifted}$ = Ensemble des étapes de process décalées à la période suivante

$\mathcal{O}_{shifted} \leftarrow \emptyset$
while $\max_{m \in \mathcal{M}} W_{m,t} > 1$ **do**
  `// Sélectionne la machine la plus chargée`
  $m' \leftarrow \text{argmax}_{m \in \mathcal{M}} W_{m,t}$
  $\mathcal{O}^{m'} \leftarrow \{s \in \mathcal{O}_l; \forall l \in \mathcal{L} \mid Q_{s,l,m',t} > 0\}$
  `// Sélectionne parmi les lots passant sur la machine, celui dont la période de livraison est la plus éloignée`
  $s' \leftarrow \text{argmax}_{s \in \mathcal{O}^{m'}}(dd_l)$
  for $s \in [o', ..., S_l]$ **do**
    if $X_{s,l,t} = 1$ **then**
      $X_{s,l,t} \leftarrow 0$
      $\mathcal{O}_{shifted} \leftarrow \mathcal{O}_{shifted} \cup \{s\}$
    for $m \in \mathcal{M}$ **do**
      $Q_{s,l,m,t} \leftarrow 0$
    end
  end
  for $m \in \mathcal{M}$ **do**
    $W_{m,t} \leftarrow \frac{\sum_{s \in \mathcal{S}_l}(a_{s,l,m} Q_{s,l,m,t})}{c_{m,t}}$
  end
end
$\mathcal{M}^- \leftarrow \{m \in \mathcal{M} \mid W_{m,t} < 1\}$
while $\mathcal{M}^- \neq \emptyset \mathcal{O}^{next} \neq \emptyset$ **do**
  $m \leftarrow \text{argmin}_{m \in \mathcal{M}} W_{m,t}$
  $\mathcal{O}_m^{next} \leftarrow \{ s \in \mathcal{O}^{next} \mid (s \in \mathcal{M}^{s,l} \quad W_{m,t} + \frac{a_{s,l,m}}{c_m} \leq 1\}$
  if $\mathcal{O}_m^{next} \neq \emptyset$ **then**
    $s' \leftarrow \text{argmin}_{s \in \mathcal{O}_m^{next}}(dd_l)$
    $W_{m,t} \leftarrow W_{m,t} + \frac{a_{s',l,m}}{c_m}$
    $X_{s,l,t} \leftarrow 1$, $X_{s,l,t+1} \leftarrow 0$
    $Q_{s,l,m,t} \leftarrow q_l$
    $\mathcal{O}^{next} \leftarrow \mathcal{O}^{next} - \{s'\}$
    if $W_{m,t} > 1$ **then**
      $\mathcal{M}^- \leftarrow \mathcal{M}^- - \{m\}$
  else
    $\mathcal{M}^- \leftarrow \mathcal{M}^- - \{m\}$
  end
end

de chaque lot réalisée dans la période $t+1$. Ensuite, la deuxième information supplémentaire requise par le module de lissage par anticipation est la liste des qualifications des machines pour les différentes recettes (et donc étapes de process de chaque lot) afin d'affecter aux machines la charge associée aux étapes de process anticipées.

La procédure de lissage vers l'avant est d'abord exécutée afin de garantir que les contraintes de capacité sont bien respectées. Notons qu'en réalité, la procédure de lissage vers l'avant peut être exécutée après celle de lissage par anticipation. Nous comparons ces deux configurations dans la section 5.4.3.

Ensuite, la procédure de lissage par anticipation liste dans un premier temps l'ensemble des machines qui ne sont pas surchargées. La machine m ayant le taux de charge dans la période courante $W_{m,t}$ le plus faible, est ensuite sélectionnée. La procédure évalue ensuite s'il existe au moins une étape de process, parmi les étapes dont la réalisation peut être anticipée, pouvant être réalisée par la machine m et n'amenant pas cette dernière à être surchargée.

(1) Il n'existe pas d'étape de process disponible pour augmenter la charge de la machine m, cette dernière est retirée des machines candidates et la procédure est relancée pour la machine la moins chargée, parmi celles restantes.

(2) Il existe au moins une étape de process disponible pour augmenter la charge de la machine m. La procédure sélectionne parmi les étapes candidates, celle dont le lot est considéré comme le *plus* prioritaire. Notons que, tout comme pour la procédure de lissage avant, la nouvelle approche requiert également la définition d'une règle de lissage. Dans l'exemple de l'algorithme 5, la règle est celle basée sur la date de livraison des lots (*Earliest Due Date*).

Une fois l'étape de process s du lot l sélectionnée, cette dernière est alors déplacée de telle sorte que sa date de fin coïncide avec la date de fin de la période courante. Les valeurs des variables $X_{s,l,t}$ et $Q_{s,l,m,t}$ sont donc modifiées en conséquence pour la machine m et l'étape s du lot l, ainsi que le taux de charge $W_{m,t}$ qui est mis à jour.

La procédure est répétée jusqu'à ce qu'il n'y ait plus de machine candidate ou d'étape de process pouvant être anticipée. À l'issue de la procédure de lissage par anticipation, on obtient une version modifiée d'un plan de production initial, nécessairement de meilleure qualité. Cette amélioration de la qualité des solutions est évaluée expérimentalement dans la section suivante.

5.4.3 Études expérimentales

Dans cette section, nous évaluons dans quelle mesure l'approche complémentaire de lissage par anticipation permet d'améliorer la qualité des plans de production obtenus.

Tout d'abord, introduisons les différentes versions considérées du module de *step-shifting* et qui sont résumées dans le tableau 5.10. Pour les quatre premières versions EDD, CR_{post}, MI, et RND, la procédure de lissage par anticipation est réalisée *après* la procédure classique de lissage vers l'avant. La différence entre les versions concerne les règles de lissage utilisées par la procédure de lissage par anticipation, qui sont pour la plupart analogues à celles présentées dans la section 5.2. Ainsi, on retrouve la règle EDD dont le but est cette fois d'anticiper la réalisation de l'étape de process du lot dont la date de livraison est la plus proche de la date courante. La règle CR_{post} privilégie le lot dont le temps disponible avant livraison, relativement au temps de cycle théorique restant, est le plus faible. La règle MI sélectionne le lot dont la charge induite, par anticipation de son étape de process, est la plus faible. Le but de cette règle est donc de maximiser le nombre d'étapes à exécuter dans la période courante en évitant de charger trop rapidement les machines. Une nouvelle règle de

lissage que nous introduisons est la règle *Random* (RND), qui sélectionne aléatoirement une étape de process parmi celles candidates pour charger la machine considérée.

Ensuite, la version *Only Postpone* (OP) comme son nom l'indique, n'exécute par la procédure de lissage par anticipation. Enfin, dans la version *Prepone First* (PF) le module de lissage des charges par anticipation est exécuté *avant* le module de lissage vers l'avant classique. Par ailleurs, dans cette version, le module de lissage par anticipation utilise la règle MI.

Dans toutes les versions présentées, la procédure de lissage vers l'avant a été exécutée avec la règle de lissage MI_{CR} qui, comme montré dans la section 5.2, permet d'obtenir des résultats corrects, voire très bons, sur l'ensemble des critères de performance.

TABLE 5.10 – Règles et procédures de lissage étudiées

Notation	**Nom**	**Description**
EDD	*Earliest Due Date*	Avance le lot ayant la date de livraison la plus proche.
CR_{Post}	*Critical Ratio if postponed*	Considère le *critical ratio* (CR) que le lot *aurait* s'il était avancé à la fin de la période courante et avance le lot ayant le critical ratio (a priori) le plus grand.
MI	*Machine Impact*	Avance le lot dont la charge induite dans la période courante est la plus faible.
RND	*Random*	Choisit aléatoirement l'étape à avancer (parmi les étapes pouvant être avancées sur la machine considérée)
OP	*Only Postpone*	Exécute seulement la procédure de lissage vers l'avant
PF	*Prepone First*	Exécute d'abors la procédure de lissage par anticipation (règle MI), puis la procédure de lissage vers l'avant

Pour chaque version du module de *step-shifting*, l'approche TSH a été exécutée sur 10 périodes d'une semaine et testée sur 40 instances réelles tirées des usines de fabrication de STMicroelectronics. La moyenne des performances pour chaque indicateur est présentée dans le tableau 5.11.

TABLE 5.11 – Comparaison de différentes versions du module de *step-shifting*

Indicateurs \ Règles	OP	EDD	CR_{Post}	MI	RND	PF
$\sum T_l$	0%	-2,18%	**-2,83%**	-2,69%	-2,13%	-1,51%
$\sum U_l$	0%	-1,74%	-2,24%	**-2,73%**	-2,23%	-1,68%
T_{max}	0%	-1,51%	-1,65%	0,30%	-1,03%	**-1,76%**
CT%	0%	-0,45%	-0,53%	**-0,57%**	-0,46%	-0,35%
TP%	0%	0,28%	0,27%	**0,32%**	0,30%	0,17%
CPU Time%	**0%**	3,98%	5,53%	3,67%	7,18%	2,87%

Tout d'abord, nous constatons que peu importe la version de l'approche de lissage utilisée,

le fait d'inclure la règle de lissage par anticipation améliore la qualité des solutions, et ce quel que soit le critère considéré.

Ensuite, notons que cette amélioration de la qualité des solutions s'accompagne logiquement d'un accroissement du temps de calcul, mais que cette augmentation reste cependant raisonnable. En effet, l'objectif de l'outil de planification de production opérationnelle basé sur l'approche TSH est de pouvoir fournir des solutions en moins de 5 minutes. L'augmentation maximale constatée étant de 7,18% avec l'approche RND, soit moins de 30 secondes supplémentaires, cet accroissement reste donc acceptable étant donné les gains de performance constatés.

Concernant la qualité des solutions fournies, nous pouvons remarquer qu'aucune version ne domine totalement les autres. Nous noterons cependant les bonnes performances des versions utilisant la procédure de lissage par anticipation (après la procédure de lissage vers l'avant) basée sur les règles CR_{Post} et MI. Ces règles s'étaient d'ailleurs déjà démarquées dans la section 5.2 pour leur bonnes performances dans la procédure de lissage vers l'avant. Ainsi, alors que l'approche de lissage par anticipation basée sur la règle CR_{Post} semble particulièrement intéressante afin de réduire le retard total ($-2,83\%$ en moyenne), la règle MI se démarque particulièrement par sa capacité à optimiser le nombre de lots en retard ($-2,73\%$), la productivité de l'usine ($+0,32\%$) ainsi que le temps de cycle moyen ($-0,57\%$). Afin d'expliquer les très bonnes performances de la règle MI, rappelons qu'à chaque itération est anticipée l'étape de process augmentant le moins (mais augmentant tout de même) le taux de charge de la machine considérée. Par conséquent, cette approche a tendance à assigner les étapes de process aux machines ayant les plus faible temps opératoires pour les réaliser, mais aussi à maximiser le nombre d'étapes de process pouvant être avancées avant que la machine considérée soit totalement chargée. La maximisation du nombre d'étapes de process avancées à la période courante explique donc assez naturellement les bonnes performances de la règle MI, notamment concernant le nombre de lots en retard ainsi que la productivité moyenne de l'usine.

Ensuite, concernant la règle *Prepone First* (PF), utilisant d'abord la procédure de lissage par anticipation, nous remarquons que celle-ci obtient les meilleurs résultats dans la minimisation du retard maximum. Cependant, rappelons que cette approche utilise la règle de lissage par anticipation *Machine Impact*, elle est donc le symétrique de la version MI que nous venons de traiter, en inversant seulement l'ordre d'utilisation des deux approches de lissage. En comparant les versions de *step-shifting* MI et PF, on constate que la première domine globalement la seconde en termes de qualité des solutions, pour un temps de calcul similaire. Nous pouvons alors conclure que l'utilisation de l'approche de lissage par anticipation *après* l'utilisation du lissage classique vers l'avant (donc version MI) est plus efficace que dans l'ordre inverse. Cela peut s'expliquer notamment par le fait que dans le cas où le lissage par anticipation est réalisé *avant* le lissage vers l'avant, ce dernier peut, en décalant certaines étapes de process à la période suivante, faire perdre les gains de performances obtenus par la méthode de lissage par anticipation. De plus, l'utilisation de l'approche de lissage vers l'avant en premier, permet de révéler des machines sous-chargées qui peuvent ensuite bénéficier de l'approche de lissage par anticipation.

5.4.4 Bilan de la procédure de lissage par anticipation

Alors que les approches de lissage par *Balanced Group* (section 5.3) ou par machine (selon différentes règles de lissage, section 5.2) ont montré des gains réels ou potentiels en termes de qualité et de performances de calcul, celles-ci obligent le module de *step-shifting* à

continuellement dégrader la solution courante afin de garantir le respect des contraintes de capacité. Dans la section 5.4, nous avons présenté une nouvelle approche, appelée procédure de lissage *par anticipation*, dont le but est de permettre à des étapes de process, dont la réalisation est prévue à la période suivante, d'avancer leur réalisation dans la période courante. À la différence de l'approche par lissage des charges vers l'avant proposée dans les sections 5.2 et 5.3, cette procédure *complémentaire* permet d'améliorer la solution courante en tirant profit de certaines machines sous-chargées. Une étude expérimentale menée sur des instances industrielles a mis en lumière l'intérêt d'inclure la procédure de lissage des charges par anticipation dans le module de *step-shifting*, après la procédure de lissage vers l'avant. Cette procédure complémentaire permet d'améliorer la qualité des solutions fournies selon différents indicateurs, tout en n'augmentant que légèrement le temps d'exécution global de l'approche TSH. Face à ces résultats prometteurs, deux actions sont prévues. La première consiste à intégrer la procédure de lissage par anticipation dans l'outil d'aide à la planification opérationnelle. Ensuite, nous projetons de perfectionner l'approche en permettant l'anticipation de plus d'étapes de process que celles actuellement considérées (plus d'une étape de process candidate par lot).

5.5 Conclusions et perspectives

Dans ce chapitre, nous avons présenté plusieurs travaux menés sur le module de *step-shifting*. Ce module est la troisième composante de l'approche TSH et a pour objectif de garantir la faisabilité des plans de production fournis du point de vue des contraintes de capacité de l'usine. Pour cela, le module de *step-shifting* reçoit en entrée un plan de production (pour une période donnée) potentiellement non faisable et l'adapte jusqu'à respecter la capacité maximale de chaque machine. Les changements consistent à décaler certaines étapes de process de certains lots vers une période ultérieure (c'est à dire en fait retarder certaines dates de réalisation) afin de réduire la charge des machines initialement assignées à la réalisation de ces étapes de process.

Le choix des étapes de process à décaler se fait par l'utilisation de règles de lissage permettant d'assigner une priorité à chaque lot. Ce module, déjà présenté dans Mhiri et al. (2018), a des limites. Tout d'abord, la règle de lissage des charges utilisée a pour principal objectif de minimiser les retards de livraison. Or, aucune évaluation n'a été entreprise afin de valider les performance de la règle de lissage afin d'optimiser cet indicateur orienté client. De plus, les critères d'optimisation dans l'industrie des semi-conducteurs sont variés et parfois antinomiques, allant du respect des commandes clients à la maximisation du nombre de plaquettes lancées en production chaque semaine. Nous avons présenté dans la section 5.2 un ensemble de 9 règles, parfois proches de règles connues de *Dispatching*, telles que les règles *Earliest Due Date* ou *Minimum Slack Time*, et avons comparé leurs performances selon les différents indicateurs sur 25 instances réelles. Les conclusions sont que, bien qu'aucune règle de lissage ne domine les autres sur l'ensemble des critères de performances, certaines se démarquent par leurs très bonnes performances moyennes et par leur capacité à rarement fournir de très mauvaises solutions. C'est par exemple le cas de la règle CR_{post}, basée sur un calcul prévisionnel de l'état d'avance/retard relatif des lots, ou bien de la règle MI_{CR}, considérant en plus, partiellement, la charge des machines. Une analyse du temps de calcul a montré que ce dernier varie légèrement selon la règle de lissage utilisée, mais reste toujours raisonnables dans le contexte d'utilisation prévu. La règle CR_{post} est actuellement intégrée dans l'approche comme règle par défaut, et nous travaillons à permettre aux gestionnaires

de sélectionner facilement d'autres règles (en particulier les règles MI et MI_{CR}) en fonction des indicateurs qu'ils considèrent comme les plus importants quand l'approche est exécutée

Ensuite, nous avons mis en évidence dans la section 5.3 les limites de l'approche de lissage par machine, en cela qu'elle a tendance à réduire la charge sur les machines de façon exagérée par rapport à leur capacité réelle. Nous avons souligné que cette faiblesse venait du fait que le lissage par machine ne considérait par la possibilité de réduire le taux de charge d'une machine en transférant un peu de sa charge sur une autre machine moins chargée. Nous avons introduit une nouvelle approche de lissage basée sur les *Balanced Group*, des groupes de machines partageant les mêmes files d'attente, et qui s'avèrent être celles ayant le même taux de charge à l'issue de l'équilibrage des charges par la méthode IMM. Une étude expérimentale a montré les performances prometteuses de l'approche par *Balanced Group*, que ce soit en termes de qualité des solutions ou de temps de calcul, par rapport à la méthode initiale basée sur le taux de charge individuel des machines. Ces résultats sont cependant à nuancer, car certains cas de machines encore surchargées, suite au processus de lissage, sont constatés. Ces situations sont principalement causées par les variations importantes du mix-produit qui perturbe la capabilité des machines à s'équilibrer entre elles, et ces variations sont d'autant plus importantes que l'usine de production est dans une situation de sous-capacité. Par conséquent, bien que l'approche de lissage par *Balanced Group* semble prometteuse, des travaux complémentaires doivent être menés pour vérifier que les gains ne sont pas seulement liés aux excès de charge de certaines machines, mais aussi pour corriger de tels excès, qui ne sont pas acceptables pour un plan de production opérationnel.

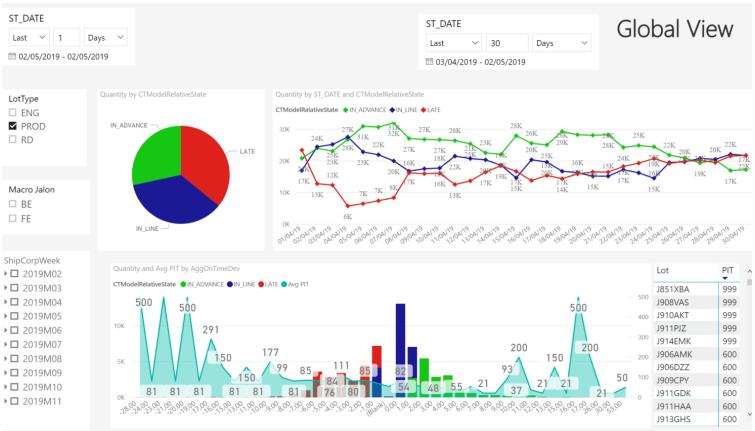
Enfin, alors que les approches de lissage par *Balanced Group* (section 5.3) ou par machine (selon différentes règles de lissage, section 5.2) ont montré des gains réels ou potentiels en terme de qualité et de performances de calcul, celles-ci obligent le module de *step-shifting* à continuellement dégrader la solution courante afin de garantir le respect des contraintes de capacité. Dans la section 5.4, nous avons présenté une nouvelle approche, appelée procédure de lissage *par anticipation*, dont le but est de permettre à des étapes de process, dont la réalisation est prévue à la période suivante, d'avancer leur réalisation dans la période courante. À la différence de l'approche par lissage des charges vers l'avant proposée dans les sections 5.2 et 5.3, cette procédure *complémentaire* permet d'améliorer la solution courante en tirant profit de certaines machines sous-chargées. Une étude expérimentale menée sur des instances industrielles a montré l'intérêt d'inclure la procédure de lissage des charges par anticipation dans le module de *step-shifting*, après la procédure de lissage vers l'avant. Cette procédure complémentaire permet d'améliorer la qualité des solutions fournies selon différents indicateurs, tout en n'augmentant que légèrement le temps d'exécution global de l'approche TSH. Parmi les règles de lissage par anticipation les plus performantes, nous pouvons citer les règles CR_{post} et MI, dont l'utilisation permet d'optimiser chacun des critères, parfois jusqu'à près de 3%. Face à ces résultats prometteurs, les prochaine étapes sont l'intégration de la procédure de lissage par anticipation dans l'outil d'aide à la planification opérationnelle, ainsi que la poursuite du développement de la procédure en augmentant le nombre d'étapes de process pouvant être anticipées, et ainsi accroître le potentiel d'amélioration de la qualité des solutions.

Chapitre 6

Développement logiciel et mise en oeuvre industrielle

.

L'un des principaux intérêts d'un projet de thèse en entreprise est de pouvoir à la fois avancer sur les aspects scientifiques de développement de nouvelles méthodes, mais également de réfléchir à l'intégration de ces méthodes au sein même du système industriel. Dans le cadre de nos travaux, cela s'est traduit par deux réalisations significatives: d'une part la création d'un cadre facilitant le développement, la validation et l'intégration de nouvelles versions du moteur de calcul (l'heuristique) dans l'outil de planification de production, et d'autre part le développement d'une interface permettant aux utilisateurs de facilement trouver les informations pertinentes et nécessaires à la prise de décision.



6.1 Introduction

Nous avons vu dans les chapitres précédents, et notamment dans le chapitre 2 sur le contexte scientifique, que les problématiques d'optimisation de la planification dans les systèmes de production ont fait l'objet de nombreuses recherches. Cependant, force est de constater que la très grande majorité de ces travaux ne porte que sur le versant technique de la problématique, à savoir le développement de méthodes toujours plus perfectionnées pour optimiser, dans notre cas, les plans de production fournis. Ainsi, il existe peu de travaux traitant de l'intégration de méthodes de recherche opérationnelle pour la planification de production au sein de systèmes industriels. Ce type d'intégration n'est pourtant pas trivial et soulève des questions importantes, parmi lesquelles on peut notamment discuter de l'utilisation des solutions proposées par l'outil développé. En effet, un plan de production peut fournir une multitude de données en sortie, mais toutes n'ont pas forcément besoin d'être montrées. La question est donc de savoir *quelles sont les données des solutions à présenter aux utilisateurs, afin de leur donner l'information nécessaire à une prise de décision efficace?*

Bien que peu d'articles traitent du développement d'outils d'aide à la planification, nous pouvons citer certains travaux faisant état de tels outils dans l'industrie des semi-conducteurs. Par exemple, le système IMPReSS mis en œuvre chez Harris Semiconductor par Leachman et al. (1996) est un outil d'aide à la décision remplissant la fonction de *Master Planning* en définissant, pour l'ensemble des usines *front end* de l'entreprise, les plans de livraison et donc les plans de charge de ces usines. L'outil est également utilisé afin d'orienter les achats de ressources de production (machines), remplissant donc également la fonction de *Capacity Planning*. Bermon and Hood (1999) introduisent *CApacity Optimization Planning System* (CAPS), un outil d'aide à la décision dont le but est de déterminer le mix produit idéal afin de maximiser les profits pour une capacité donnée, ou à l'inverse de déterminer la capacité requise afin de réaliser un mix produit donné. L'article traite de l'approche de résolution mais aussi du fonctionnement de l'outil au sein du système de production des usines d'IBM. Dans Mönch et al. (2006), les auteurs étendent le système FABMAS, un système de contrôle de la production introduit dans Mönch et al. (2003) et utilisant la simulation multi-agents, en réfléchissant à l'intégration d'un tel outil dans le système de production.

Ce nombre limité d'articles discutant de l'intégration de méthodes en entreprise, ne traduit pas nécessairement un manque d'intérêt des chercheurs, mais plus généralement une difficulté de collaboration entre entreprises et laboratoires de recherche. L'intérêt d'une thèse en entreprise est qu'elle facilite la collaboration entre ces deux entités, et permet à la fois de développer de nouvelles approches d'optimisation mais également de les expérimenter et intégrer directement dans des environnements industriels. C'est pour cette raison que nous avons jugé intéressant de présenter dans ce dernier chapitre le travail effectué durant cette thèse autour de l'intégration et l'utilisation de l'heuristique développée au sein de l'entreprise. Notamment, nous traitons dans une première partie de problématiques associées aux développement, à l'évaluation et à l'intégration de nouvelles versions de l'heuristique au sein de l'environnement industriel. Puis, dans un second temps nous discutons, de la présentation des résultats aux utilisateurs à l'aide d'interfaces pour plusieurs cas d'utilisation.

6.2 Cadre pour le développement et le test de nouvelles versions de l'outil

Comme nous l'avions mentionné dans le premier chapitre, l'outil OPERA de planification de la production s'inspire d'un autre outil de planification, à savoir l'outil CAPACE de planification de la capacité à long terme. Au fur et à mesure des années, OPERA s'est peu à peu éloigné de son grand frère pour finalement devenir un outil aux fonctions et utilisations très différentes. Cette évolution s'est faite lentement par l'introduction d'améliorations en fonction de nouveaux besoins en termes d'optimisation ou de fonctionnalités. Ainsi, depuis la création d'OPERA, près de 80 versions se sont succédées, dont plus des deux tiers durant la thèse. Pour chacune de ces versions, il est nécessaire de vérifier que des erreurs ne soient pas introduites lors de la phase de développement, ce qui devient de plus en plus difficile au fur et à mesure que l'outil de planification se complexifie. De plus, il est dans un second temps, nécessaire d'évaluer l'impact des modifications apportées sur les plans de production fournis. C'est notamment le cas pour l'influence de ces modifications sur les différents indicateurs étudiés durant la thèse tels que le retard (total, maximum,..), le temps de cycle moyen ou le taux d'utilisation des machines. Étant donné le développement régulier de nouvelles versions, ces activités nécessaires peuvent s'avérer répétitives et laborieuses. C'est pour ces raisons que nous avons développé plusieurs outils afin de faciliter ces processus d'évaluation tels qu'un outil exhaustif de détection de régressions, ainsi qu'un environnement d'analyse des indicateurs clés des plans de production fournis.

6.2.1 Validation de non-régression du moteur de calcul

Répétitivité des tests de non-régression

Le moteur de calcul, de par la nature très détaillée de l'heuristique, peut fournir beaucoup de données. En effet, il est possible de définir pour chaque opération de chaque lot dans quelle période de temps il est prévu de l'exécuter ou, par exemple, pour chaque machine la charge prévisionnelle (et les proportions de chaque produit) attendue pour chaque période. Ces informations détaillées peuvent être agrégées de différentes façons. On peut par exemple chercher à analyser l'état du WIP à la fin de la projection, ou bien analyser l'impact du plan de production sur la charge des machines, ou différentes agrégations de machines (atelier, balancing group, ...). On peut également agréger les données par lot, et analyser leur état d'avancement, l'évaluation de leur retard potentiel ou au contraire analyser globalement l'activité prévisionnelle de l'usine sur les futures semaines. Il existe plus de 30 fichiers pouvant être générés pour l'utilisateur, chacun avec leur spécificités, et le premier élément à vérifier est la consistance de ces fichiers.

Qu'entendons-nous par consistance des fichiers? Prenons par exemple le fichier LOTS-PROJECTION, qui contient pour chaque étape de process de chaque lot la date à laquelle il est prévu, On peut comprendre qu'entre deux versions de l'outil d'OPERA, disons une version A et une version B, il y aura des différences entre le fichier LOTSPROJECTION généré par chacune des versions. Certaines opérations seront réalisées plus tard ou plus tôt dans la version A par rapport à la version B, ou bien sur d'autre machines, ce qui peut par exemple changer le nombre total d'étapes de process réalisées pour une période donnée. Cependant, il y a certains indicateurs qui sont supposés être invariants. Par exemple, on comprend que si on lance l'outil de planification sur un horizon assez long afin qu'il puisse traiter l'ensemble des étapes de process de chaque lot, on doit retrouver au total le même nombre d'étapes

réalisées pour les deux versions. Sinon, en partant du principe que la version d'origine est celle validée, cela signifie qu'une régression est apparue lors du développement de la version B. Ce type d'analyse peut être faite sur chacun des fichiers selon différents indicateurs, mais la taille et le nombre de fichiers rendraient la procédure trop longue pour être faite de façon exhaustive.

Par conséquent, à l'origine, seuls quelques fichiers et indicateurs invariants étaient analysés, souvent spécifiquement aux modifications apportées dans la version nouvellement développée afin de vérifier le comportement attendu. Pourtant, cette procédure bien qu'allégée se montrait assez répétitive, obligeant à exécuter l'outil pour chacune des versions, récupérer le fichier à comparer, l'introduire dans une feuille Excel, puis analyser les fichiers en agrégeant les indicateurs voulus. Si des différences anormales étaient constatées, il fallait alors modifier la nouvelle version et relancer un processus de comparaison jusqu'à ce que les différences soient corrigées. Ce type d'analyse pouvait prendre quelques dizaines de minutes si aucune régression n'était relevée, à plusieurs heures (voire jours) dans les cas où des erreurs étaient détectées.

L'outil ENTRACTE de tests automatiques de non-régression

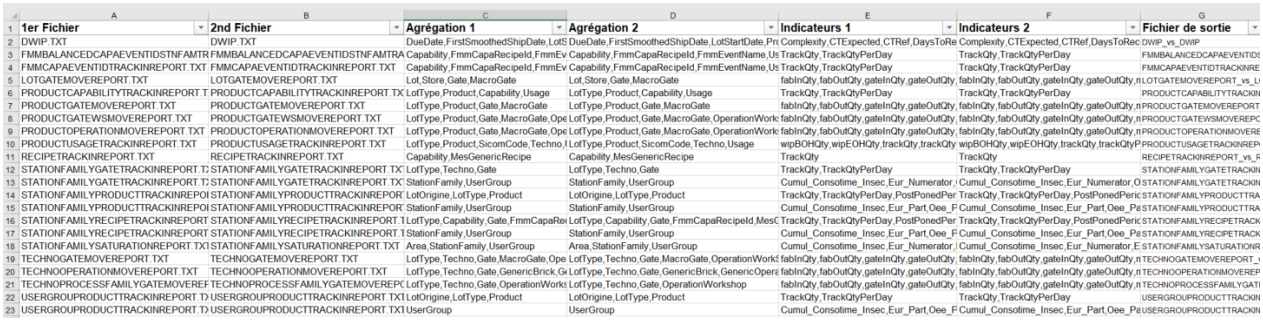
Afin de gagner du temps et de l'énergie sur ce type de processus de vérification, nous avons développé durant la thèse un outil de tests automatiques, permettant d'exécuter de façon rapide et autonome un grand nombre de vérifications. Cet outil, nommé ENTRACTE, est composé comme OPERA d'une interface Excel et d'un moteur réalisé sous Java. Ce moteur prend en entrée un ensemble de tests, le tableau 6.1 présente un exemple de comparatif pouvant être demandé à l'outil.

Dans ce tableau, les colonnes **Fichier 1** et **Fichier 2** spécifient les fichiers devant être comparés, chaque colonne correspondant à une version de l'outil. À noter qu'il est possible de comparer des fichiers différents entre deux versions car certains indicateurs sont présents dans plusieurs fichiers différents. À titre d'exemple, il est possible de comparer le nombre total d'opérations effectuées durant une planification complète, entre un fichier agrégeant les résultats au niveau de la machine et l'autre au niveau du lot. Ensuite, les colonnes **Agrégation 1** et **Agrégation 2** permettent de définir comment seront agrégés les différents indicateurs à comparer. Dans le cas de l'exemple du tableau 6.1, cela signifie que les différentes lignes de chaque fichier seront agrégées par lot. Si un fichier possède plusieurs lignes correspondant à un même lot, alors les indicateurs à comparer seront sommés (agrégation par défaut). Les colonnes **Indicateurs 1** et **Indicateurs 2** sont celles permettant de définir les valeurs selon lesquelles chaque agrégation va être comparée. Dans le cas de l'exemple, l'indicateur considéré est le TrackQty, c'est à dire le nombre de plaquettes ayant passé une étape de process. Ainsi, le test défini dans le tableau 6.1 consiste à comparer entre deux versions les fichiers LOTSPROJECTION.txt en comparant pour chaque lot le nombre total d'étapes de process qui ont été réalisées. Si, pour certains lots, des différences sont relevées, ces dernières sont renseignées dans un fichier de sortie, nommé "LotsProj_Compare" dans le cas de l'exemple.

TABLE 6.1 – Exemple de ligne pouvant être fournie en entrée à l'outil ENTRACTE d'exécution de tests en automatique

Fichier 1	Fichier 2	Agrégation 1	Agrégation 2	Indicateurs 1	Indicateurs 2	Fichier de sortie
LOTSPROJECTION	LOTSPROJECTION	Lot	Lot	TrackQty	TrackQty	LotsProj_Compare

Il est possible de fournir en entrée de l'outil une liste de tests à effectuer, sous la forme d'un ensemble de lignes décrivant les comparatifs à effectuer. Ces tests peuvent être construits directement par l'utilisateur, ou ce dernier peut également générer un ensemble de tests pré-établis. La figure 6.1 montre un exemple de tests pré-configurés permettant de détecter des cas de régression d'une nouvelle version affectant l'aspect lissage de charge de l'outil.



	1er Fichier	2nd Fichier	Agrégation 1	Agrégation 2	Indicateurs 1	Indicateurs 2	Fichier de sortie
2	DWIP.TXT	DWIP.TXT	DueDate,FirstSmoothedShipDate,LotS	DueDate,FirstSmoothedShipDate,LotStartDate,Pr	Complexity,CTExpected,CTRef,DaysToRe	Complexity,CTExpected,CTRef,DaysToRec	DWIP_vs_DWIP
3	FMMBALANCEDCAPAEVENTIDSTNFAMTR	FMMBALANCEDCAPAEVENTIDSTNFAMTRA	Capability,FmmCapaRecipeId,FmmEv	Capability,FmmCapaRecipeId,FmmEventName,Us	TrackQty,TrackQtyPerDay	TrackQty,TrackQtyPerDay	FMMBALANCEDCAPAEVENTIDS
4	FMMCAPAEVENTIDTRACKINREPORT.TXT	FMMCAPAEVENTIDTRACKINREPORT.TXT	Capability,FmmCapaRecipeId,FmmEv	Capability,FmmCapaRecipeId,FmmEventName,Us	TrackQty,TrackQtyPerDay	TrackQty,TrackQtyPerDay	FMMCAPAEVENTIDTRACKINRE
5	LOTGATEMOVEREPORT.TXT	LOTGATEMOVEREPORT.TXT	Lot,Store,Gate,MacroGate	Lot,Store,Gate,MacroGate	fabInQty,fabOutQty,gateInQty,gateOutQty,	fabInQty,fabOutQty,gateInQty,gateOutQty,n	LOTGATEMOVEREPORT_vs_L
6	PRODUCTCAPABILITYTRACKINREPORT.T	PRODUCTCAPABILITYTRACKINREPORT.TX	LotType,Product,Capability,Usage	LotType,Product,Capability,Usage	TrackQty,TrackQtyPerDay	TrackQty,TrackQtyPerDay	PRODUCTCAPABILITYTRACKIN
7	PRODUCTGATEMOVEREPORT.TXT	PRODUCTGATEMOVEREPORT.TXT	LotType,Product,Gate,MacroGate	LotType,Product,Gate,MacroGate	fabInQty,fabOutQty,gateInQty,gateOutQty,	fabInQty,fabOutQty,gateInQty,gateOutQty,n	PRODUCTGATEMOVEREPORT
8	PRODUCTGATEWSMOVEREPORT.TXT	PRODUCTGATEWSMOVEREPORT.TXT	LotType,Product,Gate,MacroGate,Ope	LotType,Product,Gate,MacroGate,OperationWork	fabInQty,fabOutQty,gateInQty,gateOutQty,	fabInQty,fabOutQty,gateInQty,gateOutQty,n	PRODUCTGATEWSMOVEREPO
9	PRODUCTOPERATIONMOVEREPORT.TXT	PRODUCTOPERATIONMOVEREPORT.TXT	LotType,Product,Gate,MacroGate,Ope	LotType,Product,Gate,MacroGate,OperationWork	fabInQty,fabOutQty,gateInQty,gateOutQty,	fabInQty,fabOutQty,gateInQty,gateOutQty,n	PRODUCTOPERATIONMOVERE
10	PRODUCTUSAGETRACKINREPORT.TXT	PRODUCTUSAGETRACKINREPORT.TXT	LotType,Product,SicomCode,Techno,I	LotType,Product,SicomCode,Techno,Usage	wipBOHQty,wipEOHQty,trackQty,trackQty	wipBOHQty,wipEOHQty,trackQty,trackQtyP	PRODUCTUSAGETRACKINREP
11	RECIPETRACKINREPORT.TXT	RECIPETRACKINREPORT.TXT	Capability,MesGenericRecipe	Capability,MesGenericRecipe	TrackQty	TrackQty	RECIPETRACKINREPORT_vs_R
12	STATIONFAMILYGATETRACKINREPORT.T	STATIONFAMILYGATETRACKINREPORT.TX	LotType,Techno,Gate	LotType,Techno,Gate	TrackQty,TrackQtyPerDay	TrackQty,TrackQtyPerDay	STATIONFAMILYGATETRACKIN
13	STATIONFAMILYGATETRACKINREPORT.T	STATIONFAMILYGATETRACKINREPORT.TX	StationFamily,UserGroup	StationFamily,UserGroup	Cumul_Consotime_Insec,Eur_Numerator,	Cumul_Consotime_Insec,Eur_Numerator,C	STATIONFAMILYGATETRACKIN
14	STATIONFAMILYPRODUCTTRACKINREPOI	STATIONFAMILYPRODUCTTRACKINREPORT	LotOrigine,LotType,Product	LotOrigine,LotType,Product	TrackQty,TrackQtyPerDay,PostPonedPer	TrackQty,TrackQtyPerDay,PostPonedPeric	STATIONFAMILYPRODUCTTRA
15	STATIONFAMILYPRODUCTTRACKINREPOI	STATIONFAMILYPRODUCTTRACKINREPORT	StationFamily,UserGroup	StationFamily,UserGroup	Cumul_Consotime_Insec,Eur_Part,Oee_F	Cumul_Consotime_Insec,Eur_Part,Oee_Pe	STATIONFAMILYPRODUCTTRA
16	STATIONFAMILYRECIPETRACKINREPORT	STATIONFAMILYRECIPETRACKINREPORT.T	LotType,Capability,Gate,FmmCapaRe	LotType,Capability,Gate,FmmCapaRecipeId,MesG	TrackQty,TrackQtyPerDay,PostPonedPer	TrackQty,TrackQtyPerDay,PostPonedPeric	STATIONFAMILYRECIPETRACK
17	STATIONFAMILYRECIPETRACKINREPORT	STATIONFAMILYRECIPETRACKINREPORT.T	StationFamily,UserGroup	StationFamily,UserGroup	Cumul_Consotime_Insec,Eur_Part,Oee_F	Cumul_Consotime_Insec,Eur_Part,Oee_Pe	STATIONFAMILYRECIPETRACK
18	STATIONFAMILYSATURATIONREPORT.TXT	STATIONFAMILYSATURATIONREPORT.TXT	Area,StationFamily,UserGroup	Area,StationFamily,UserGroup	Cumul_Consotime_Insec,Eur_Numerator,	Cumul_Consotime_Insec,Eur_Numerator,E	STATIONFAMILYSATURATIONR
19	TECHNOGATEMOVEREPORT.TXT	TECHNOGATEMOVEREPORT.TXT	LotType,Techno,Gate,MacroGate,Ope	LotType,Techno,Gate,MacroGate,OperationWork	fabInQty,fabOutQty,gateInQty,gateOutQty,	fabInQty,fabOutQty,gateInQty,gateOutQty,n	TECHNOGATEMOVEREPORT_v
20	TECHNOOPERATIONMOVEREPORT.TXT	TECHNOOPERATIONMOVEREPORT.TXT	LotType,Techno,Gate,GenericBrick,Ga	LotType,Techno,Gate,GenericBrick,GenericOpera	fabInQty,fabOutQty,gateInQty,gateOutQty,	fabInQty,fabOutQty,gateInQty,gateOutQty,n	TECHNOOPERATIONMOVEREP
21	TECHNOPROCESSFAMILYGATEMOVEREF	TECHNOPROCESSFAMILYGATEMOVEREPO	LotType,Techno,Gate,OperationWorkst	LotType,Techno,Gate,OperationWorkshop	fabInQty,fabOutQty,gateInQty,gateOutQty,	fabInQty,fabOutQty,gateInQty,gateOutQty,n	TECHNOPROCESSFAMILYGATI
22	USERGROUPRODUCTTRACKINREPORT.TX	USERGROUPRODUCTTRACKINREPORT.TXT	LotOrigine,LotType,Product	LotOrigine,LotType,Product	TrackQty,TrackQtyPerDay	TrackQty,TrackQtyPerDay	USERGROUPRODUCTTRACKIN
23	USERGROUPRODUCTTRACKINREPORT.TX	USERGROUPRODUCTTRACKINREPORT.TXT	UserGroup	UserGroup	Cumul_Consotime_Insec,Eur_Part,Oee_F	Cumul_Consotime_Insec,Eur_Part,Oee_Pe	USERGROUPRODUCTTRACKIN

FIGURE 6.1 – Exemple de liste pré-configurée de tests de non-régression

Lorsque l'ensemble des tests ont été réalisés, les résultats sont consignés dans des fichiers directement consultables via l'interface Excel. La figure 6.2 illustre le résultat obtenu une fois que l'outil est exécuté pour un ensemble de tests. On peut voir pour chaque ligne si la comparaison n'a montré aucune erreur (ligne de couleur verte) ou si des différences ont été détectées (couleur rouge avec nombre de différences affichées). Dans le cas de l'exemple, on note que trois tests ont révélé des différences entre les fichiers, 54 différences pour le fichier "DWIP", puis 24 et 11 différences pour les fichiers "STATIONFAMILYPRODUCT-TRACKINREPORT". On note deux lignes pour ces derniers fichiers car deux tests ont été effectués selon des agrégations différentes.

	A	B
1	FICHIER	TAILLE
2	DWIP_vs_DWIP(54Lines).DIFF	13053
3	FMMBALANCEDCAPAEVENTIDSTNFAMTRACKINREPORT_vs_FMMBALANCEDCAPAEVENTIDSTNFAMTRACKINREPORT(NoDifference).DIFF	507
4	FMMCAPAEVENTIDTRACKINREPORT_vs_FMMCAPAEVENTIDTRACKINREPORT(NoDifference).DIFF	479
5	LOTGATEMOVEREPORT_vs_LOTGATEMOVEREPORT(NoDifference).DIFF	849
6	PRODUCTCAPABILITYTRACKINREPORT_vs_PRODUCTCAPABILITYTRACKINREPORT(NoDifference).DIFF	459
7	PRODUCTGATEMOVEREPORT_vs_PRODUCTGATEMOVEREPORT(NoDifference).DIFF	633
8	PRODUCTGATEWSMOVEREPORT_vs_PRODUCTGATEWSMOVEREPORT(NoDifference).DIFF	673
9	PRODUCTOPERATIONMOVEREPORT_vs_PRODUCTOPERATIONMOVEREPORT(NoDifference).DIFF	739
10	PRODUCTUSAGETRACKINREPORT_vs_PRODUCTUSAGETRACKINREPORT(NoDifference).DIFF	551
11	RECIPETRACKINREPORT_vs_RECIPETRACKINREPORT(NoDifference).DIFF	395
12	STATIONFAMILYGATETRACKINREPORT_vs_STATIONFAMILYGATETRACKINREPORT_eqpt(NoDifference).DIFF	553
13	STATIONFAMILYGATETRACKINREPORT_vs_STATIONFAMILYGATETRACKINREPORT_recipe(NoDifference).DIFF	433
14	STATIONFAMILYPRODUCTTRACKINREPORT_vs_STATIONFAMILYPRODUCTTRACKINREPORT_eqpt(24Lines).DIFF	7985
15	STATIONFAMILYPRODUCTTRACKINREPORT_vs_STATIONFAMILYPRODUCTTRACKINREPORT_recipe(11Lines).DIFF	3537
16	STATIONFAMILYRECIPETRACKINREPORT_vs_STATIONFAMILYRECIPETRACKINREPORT_eqpt(NoDifference).DIFF	513
17	STATIONFAMILYRECIPETRACKINREPORT_vs_STATIONFAMILYRECIPETRACKINREPORT_recipe(NoDifference).DIFF	557
18	STATIONFAMILYSATURATIONREPORT_vs_STATIONFAMILYSATURATIONREPORT(NoDifference).DIFF	637
19	TECHNOGATEMOVEREPORT_vs_TECHNOGATEMOVEREPORT(NoDifference).DIFF	779
20	TECHNOOPERATIONMOVEREPORT_vs_TECHNOOPERATIONMOVEREPORT(NoDifference).DIFF	715
21	TECHNOPROCESSFAMILYGATEMOVEREPORT_vs_TECHNOPROCESSFAMILYGATEMOVEREPORT(NoDifference).DIFF	671
22	USERGROUPRODUCTTRACKINREPORT_vs_USERGROUPRODUCTTRACKINREPORT_eqpt(NoDifference).DIFF	535
23	USERGROUPRODUCTTRACKINREPORT_vs_USERGROUPRODUCTTRACKINREPORT_recipe(NoDifference).DIFF	443

FIGURE 6.2 – Visuel montrant le résultat d'un ensemble de tests de non-régression

Il est possible ensuite de zoomer dans les différentes lignes afin de voir les agrégations présentant des différences. En prenant le cas de la figure 6.2, on peut prendre à titre d'exemple le cas du fichier DWIP, qui représente l'état prévisionnel de chacun des lots à la fin de l'horizon de planification calculé par OPERA. On peut voir une illustration d'un tel visuel sur la figure 6.3.

	A	B	C	D	E
1	First_File	Second_File	Error_Type	Which_File	Lot
2	Version 1	Version 2	Missing_Line	Second	Q615973Q39
3	Version 1	Version 2	Missing_Line	Second	Q552932Q10
4	Version 1	Version 2	Missing_Line	Second	Q638106Q13
5	Version 1	Version 2	Missing_Line	Second	Q702223
6	Version 1	Version 2	Missing_Line	Second	Q612138Q03
7	Version 1	Version 2	Missing_Line	Second	Q650165Q08
8	Version 1	Version 2	Missing_Line	Second	Q641032Q11
9	Version 1	Version 2	Missing_Line	Second	Q614073Q03
10	Version 1	Version 2	Missing_Line	Second	Q633241
11	Version 1	Version 2	Missing_Line	Second	Q613856Q02
12	Version 1	Version 2	Missing_Line	Second	Q647092Q30
13	Version 1	Version 2	Missing_Line	Second	Q632196Q02
14	Version 1	Version 2	Missing_Line	Second	Q609214Q26
15	Version 1	Version 2	Missing_Line	Second	Q612138Q06
16	Version 1	Version 2	Missing_Line	Second	Q647995Q03
17	Version 1	Version 2	Missing_Line	Second	Q632068Q19
18	Version 1	Version 2	Missing_Line	Second	Q626106Q27
19	Version 1	Version 2	Missing_Line	Second	Q651262Q07
20	Version 1	Version 2	Missing_Line	Second	Q641032Q15
21	Version 1	Version 2	Missing_Line	Second	Q651621
22	Version 1	Version 2	Missing_Line	Second	Q614073Q06
23	Version 1	Version 2	Missing_Line	Second	Q651621Q01
24	Version 1	Version 2	Missing_Line	Second	Q616733Q07
25	Version 1	Version 2	Missing_Line	Second	Q606497Q06
26	Version 1	Version 2	Missing_Line	Second	Q548027Q09
27	Version 1	Version 2	Missing_Line	Second	Q609214Q29
28	Version 1	Version 2	Missing_Line	Second	Q652777
29	Version 1	Version 2	Missing_Line	Second	Q610991Q04
30	Version 1	Version 2	Missing_Line	Second	Q638827
31	Version 1	Version 2	Missing_Line	Second	Q641231
32	Version 1	Version 2	Missing_Line	Second	Q607520
33	Version 1	Version 2	Missing_Line	Second	Q650023Q01
34	Version 1	Version 2	Missing_Line	Second	Q616836

FIGURE 6.3 – Visuel présentant les différences anormales relevées entre des fichiers de deux versions de l'outil de planification

Au regard des résultats présentés dans la figure 6.3, on constate qu'un certain nombre de lots sont manquants dans le fichier issu de la deuxième version testée, ce qui est une différence anormale et traduit une erreur dans l'implémentation de la seconde version de l'outil. Il est alors nécessaire de comprendre la source de cette régression, et il est possible de zoomer encore sur les différences afin de trouver dans les fichiers les lignes concernées par ces différences. Les tests de non-régression ne font pas que détecter l'absence de certaines agrégations, il permettent également, comme dit précédemment, la détection de différences de valeurs associées à ces régressions. En effet, si l'on prend le cas du fichier "STATION-FAMILYPRODUCTTRACKINREPORT" qui présente lui aussi des différences, et dont le détail est présenté en figure 6.4, on constate des différences pour trois machines (nommées "*StationFamily*" dans l'exemple). Les différences portent sur le nombre de "*TrackIn*", c'est à dire le nombre de plaquettes passées par une étape de process sur la machine. On constate que pour les trois machines citées, les quantités de plaquettes traitées sont différentes selon les versions. Face à ce constat, il y a deux possibilités. Soit il est normal qu'il y ait des différences, par exemple si la nouvelle version intègre des changements dans l'équilibrage ou les politiques de lissage, et dans ce cas cela ne traduit pas de non-régression. Soit le changement est anormal, par exemple la nouvelle version n'intègre que l'ajout de nouveaux indicateurs, et il est alors nécessaire d'investiguer les raisons de cette régression.

	A	B	C	D	E	F	G	H
1	First_File	Second_File	Error_Type	Which_File	Value_Type	StationFamily	Value_FirstFile	Value_SecondFile
2	Version 1	Version 2	Different Value	both	TrackQty/TrackQty	MENDC_TAN_CUMN	20711,48	20311,48
3	Version 1	Version 2	Different Value	both	TrackQty/TrackQty	TCENT_ISSG_RTNH3	2749,10	2700,57
4	Version 1	Version 2	Different Value	both	TrackQty/TrackQty	TFORM_ANN_CU_04	9281,30	9098,44

FIGURE 6.4 – Visuel présentant les différences relevées entre des fichiers de deux versions de l'outil de planification

En conclusion, du fait des développements réguliers de nouvelles versions de l'outil OPERA, des tests sont régulièrement effectués afin de vérifier la non-régression du moteur

du calcul. Cette phase de test peut s'avérer rapidement longue et répétitive, et pourtant ne permet de vérifier qu'un nombre restreint d'indicateurs en des temps raisonnables. L'outil ENTRACTE que nous avons développé a pour but d'automatiser ce processus de validation de non-régression. Grâce à une structure à la fois simple et modulable, il permet de créer un grand nombre de tests, et de les assembler afin de constituer des ensembles pré-configurés d'analyse à réaliser selon les besoins de l'utilisateur. Ainsi, le développement de cet outil a permis un gain à la fois en temps (quelques minutes sont nécessaires pour exécuter un ensemble complet de tests) mais également en confiance dans les validations faites, grâce l'exhaustivité des tests effectués.

6.2.2 Cadre d'analyse des performances

Lors du développement d'une nouvelle version du moteur de planification, après le contrôle de non-régression vient l'évaluation des performances. Dans cette partie, on souhaite comparer la qualité des solutions fournies par différentes versions de l'outil. Cela peut être entre une version nouvellement développée et une plus ancienne, ou bien entre plusieurs modes d'une même version du moteur de calcul (comparaison de différentes politiques de lissage par exemple).

Nous avons vu dans les sections précédentes un certain nombre d'analyses numériques permettant de comparer plusieurs versions ou modes de l'outil de planification. Par exemple, il est possible d'analyser l'influence des politiques de lissage dans le module de *step-shifting* ou bien l'impact de la nouvelle méthode d'équilibrage dans le module d'*équilibrage*. Pour la plupart de ces études, les comparatifs ont été menés sur des instances directement tirées des données issues de l'usine. Or les usines *front end* (pour rappel, dédiées à la fabrication des circuits-intégrés sur des plaquettes de silicium) sont des systèmes très dynamiques et variables, notamment dans un contexte *High-Mix Low-Volume*, que ce soit par exemple au niveau de la qualification des machines, de la position du WIP ou de sa composition. Il est alors important d'avoir suffisamment d'instances réparties dans le temps afin de couvrir une grande diversité des états possibles. C'est pour cette raison que, tout au long de la thèse, un jeu d'instances s'est peu à peu développé pour au final constituer un ensemble de 50 instances dont les principales caractéristiques sont résumées dans le tableau 6.2.

Parmi les informations qui sont présentées, le WIP indique le nombre de lots présents dans l'usine au début de l'horizon de planification, avec une moyenne à 4193 lots (sachant que chaque lot est généralement composé de 25 plaquettes). Ensuite, un autre aspect pouvant impacter la planification est ce que l'on appelle le déséquilibre de la ligne de production. Les routes des produits sont généralement décomposées en 10 jalons. Ces jalons découpent la route en portions grossièrement équivalentes en nombre d'étapes de process. La mesure de la dispersion du WIP sur ces différents jalons est un bon indicateur du déséquilibre de la ligne de production. Ainsi, plus la dispersion est importante (mesurée suivant l'écart type), plus il y aura des risques de surcharge pour certaines ressources tandis que d'autres ressources seront sous utilisées. Étant donné que les volumes peuvent varier de façon importante entre les instances, on donne l'écart type du WIP divisé par la moyenne (aussi appelé Coefficient de Variation) qui est noté $CV(WIP)$ dans le tableau. On constate alors une bonne diversité dans la dispersion du WIP, allant de 18% pour les plus faibles jusqu'à plus de 50% pour les dernières instances, montrant des flux de production très déséquilibrés.

L'un des critères d'optimisation récurrents eu sein de la thèse, et plus généralement en planification, est le respect des dates de livraison. De ce point de vue, les instances peuvent être plus ou moins difficiles en fonction de la situation des lots (en termes d'avance ou de

retard) au début de l'horizon de planification (les lots qui constituent le WIP initial). Cette situation des lots peut se mesurer simplement en faisant la différence entre le temps théorique restant pour que le lot l complète toutes ses étapes de process restantes CT_l^{th} et le temps restant jusqu'à sa date de livraison t_l^d. Ainsi, on définit :

$$T_l^{init} = max\{0, CT_l^{th} - t_l^d\}$$

comme étant le retard initial du lot l au début de la planification. Basés sur cette information, nous avons constitué trois indicateurs pour évaluer la difficulté d'une instance. Le premier, nommé $\sum \frac{T^{init}}{n}$, est le retard initial moyen de chaque lot. Celui-ci varie entre 0,46 (valeur assez faible considérant des temps de cycle moyens de plusieurs mois) jusqu'à des cas plus difficiles avec 3,63 jours de retard en moyenne pour chaque lot de l'usine. Enfin, l'indicateur $\sum \frac{U^{init}}{n}$ correspond à la proportion de lots ayant un retard significatif (supérieur à 1 jour) au commencement de la planification.

Les trois dernières colonnes du tableau 6.2 témoignent de la volumétrie des problèmes traités, avec respectivement le nombre de machines considérées, le nombre de produits différents du WIP initial, ainsi que le nombre moyen d'étapes de process restant pour chacun des lots.

À chacune des instances est associé un grand nombre de fichiers, parfois de taille importante. Il est donc possible d'utiliser ces instances dans le cadre d'autres travaux. À titre d'exemple, chaque instance est constituée d'informations détaillées sur le WIP, les lots qui sont prévus d'être lancés en production, ou bien les caractéristiques des machines (capacité, qualifications,...). Il est alors possible d'utiliser ces instances (en complément d'autres données ou non) dans le cadre de travaux sur d'autres outils, par exemple pour de l'ordonnancement d'atelier cours terme ou pour de la planification plus agrégée.

Une fois les instances constituées, il est ensuite nécessaire d'exécuter les différentes versions de l'outil pour chacun des problèmes et récolter les résultats. Or, le nombre de tests à effectuer peut très vite croître en fonction des éléments que l'on souhaite comparer. À titre d'exemple, imaginons que l'on souhaite comparer l'impact de l'intégration de la capacité des machines (comparaison capacité infinie/finie) sur les retards de livraison des lots. Cela fait deux versions, chacune lancées sur les 50 instances, soit 100 dossiers de résultats (constitués de plusieurs fichiers) à analyser. Et cela reste un comparatif "réduit", si l'on compare par exemple à l'étude de règles de lissage qui requiert d'exécuter sur chaque instance près d'une dizaine de politiques différentes! Récupérer et mettre en forme ces résultats serait une tache fastidieuse et répétitive avec une forte probabilité d'erreurs de manipulation.

Il était donc nécessaire d'automatiser le processus d'exécution et de mise en forme des résultats. Pour ce faire, nous avons développé un package composé:

— D'un jeu d'instances sur lequel exécuter les différentes versions,

— D'un script configurable pour la définition des différentes versions des outils à exécuter,

— D'un fichier de mise en forme et de visualisation des résultats.

Le jeu d'instances est celui décrit plus haut. Le script quant à lui est un algorithme simple mais très configurable permettant de facilement définir les programmes (en l'occurrence l'outil OPERA) à exécuter sur les différentes instances, puis de stocker les résultats qui seront lus par le fichier de mise en forme et d'analyse.

Ce fichier a été développé en utilisant le logiciel *Power BI* qui est un outil de Business Intelligence présentant plusieurs avantages. Premièrement, il permet de facilement transformer et mettre en forme les données afin d'analyser les différentes solutions pour les indicateurs

TABLE 6.2 – Caractéristiques des instances de tests tirées de cas réels

Instance	WIP	StdDev Wip	CV (WIP)	Avg(T)	$\%U_i$	Mchns	Prdts	Avg (étapes)
1	3092	69	22%	1,00	29%	322	325	148
2	3123	71	23%	0,71	23%	322	320	149
3	3142	66	21%	0,67	22%	323	325	146
4	3148	80	25%	0,70	22%	324	324	146
5	3168	66	21%	1,50	29%	323	328	145
6	2877	92	32%	2,69	26%	327	268	134
7	2845	83	29%	1,65	31%	327	259	132
8	2757	77	28%	0,69	12%	326	243	138
9	2822	51	18%	0,86	16%	327	250	133
10	2785	60	21%	0,66	13%	326	244	135
11	2831	56	20%	0,46	11%	326	238	135
12	2883	60	21%	0,47	12%	326	239	134
13	2854	64	23%	0,50	10%	326	234	132
14	2644	67	25%	0,65	23%	326	257	140
15	2748	133	49%	2,91	53%	327	267	149
16	2799	144	51%	3,26	59%	327	277	155
17	2777	120	43%	3,63	50%	327	262	164
18	2797	107	38%	2,41	33%	327	270	163
19	2985	101	34%	1,94	33%	327	272	163
20	2866	60	21%	2,03	44%	327	265	150
21	2945	71	24%	1,36	32%	327	257	141
22	2958	79	27%	0,96	28%	327	258	139
23	2962	64	22%	1,04	26%	326	265	134
24	4011	136	34%	1,48	36%	395	433	597
25	4002	132	33%	1,50	27%	396	423	594
26	4182	139	33%	1,52	28%	397	436	594
27	4261	156	37%	1,39	28%	397	428	589
28	4199	153	36%	0,81	16%	397	439	599
29	4239	162	38%	1,44	26%	400	447	588
30	4190	154	37%	1,15	23%	400	413	593
31	4265	160	37%	1,71	26%	403	439	607
32	4242	173	41%	1,28	25%	401	439	590
33	5725	300	52%	2,41	37%	417	575	579
34	5675	299	53%	2,84	37%	415	579	584
35	5542	259	47%	2,02	30%	426	574	556
36	5608	254	45%	2,35	26%	426	577	554
37	5531	251	45%	2,06	24%	429	574	559
38	5652	256	45%	2,24	28%	429	589	559
39	5652	256	45%	2,20	27%	429	589	559
40	5735	276	48%	2,98	30%	430	584	537
41	5770	288	50%	2,04	24%	439	629	519
42	5937	258	43%	1,69	28%	445	679	511
43	6044	265	44%	1,58	28%	445	697	515
44	6092	257	42%	1,64	30%	445	706	517
45	6137	271	44%	1,70	33%	448	701	512
46	6090	291	48%	1,58	31%	448	703	518
47	5961	275	46%	1,93	31%	454	710	526
48	5968	325	54%	1,94	32%	455	708	521
49	6056	326	54%	3,13	25%	456	714	530
50	6087	340	56%	1,53	28%	457	719	539
Avg	4193	165	37%	1,66	28%	379	435	367
Max	6137	340	56%	3,63	59%	457	719	607
Min	2644	51	18%	0,46	10%	322	234	132

qui nous intéressent. Ensuite, il est facile de créer en amont des fichiers canevas qui seront ensuite utilisés pour analyser les données brutes. Il est par exemple possible de créer un fichier modèle permettant d'analyser pour un plan de production donné un certains nombre d'indicateurs clés tels que le retard des lots, l'activité de l'usine ou le temps de cycle de certains produits. La figure 6.5 illustre un exemple de canevas permettant de comparer facilement différentes versions de l'outil OPERA. Dans l'histogramme situé dans la moitié supérieure de la figure, chaque bloc de l'abscisse représente une instance de test, tandis que chaque couleur de barre représente les solutions fournies par une version de l'outil OPERA. Il est alors possible de rapidement détecter les versions les plus (ou les moins) performantes pour l'indicateur donné, en l'occurrence ici la proportion de lots livrés en retard. Ces résultats sont également résumés dans le tableau situé dans la partie basse du visuel, avec des résultats selon l'indicateur considéré pour chaque version d'OPERA et chaque instance de test. Ces valeurs sont ensuite exportables sous forme de fichiers Excel pour d'autres utilisations. De tels visuels sont disponibles pour d'autres indicateurs classiques tels que le temps de cycle global de l'usine ou de certains produits, le taux d'utilisation des machines, l'activité de l'usine ou bien le retard moyen des lots livrés.

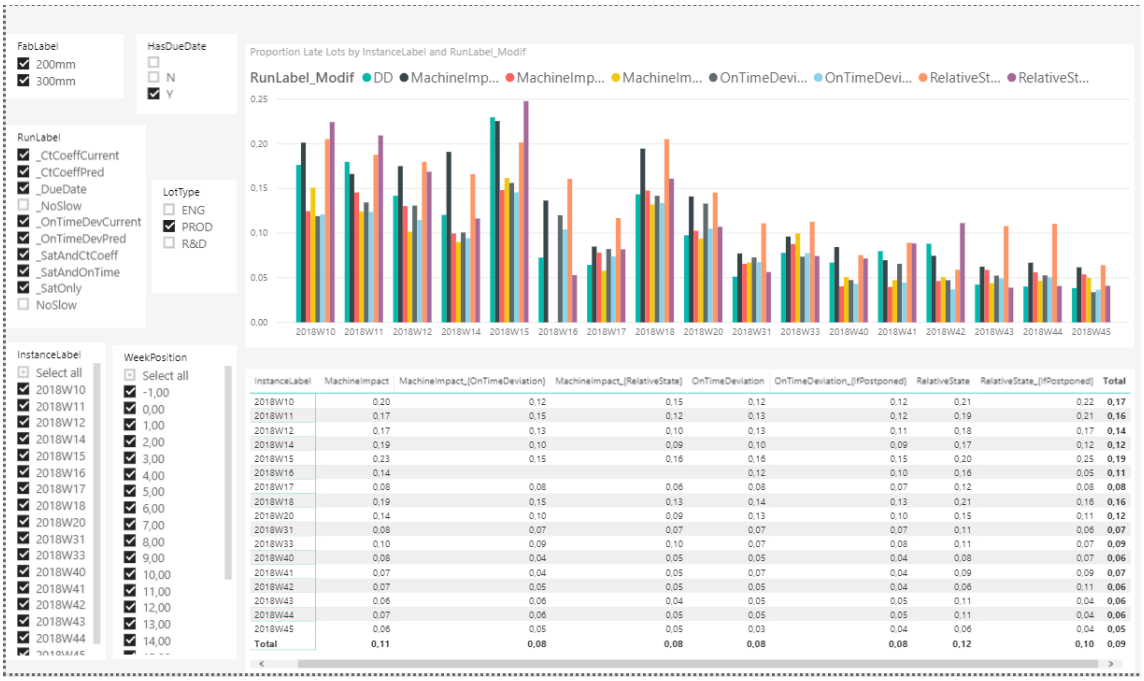


FIGURE 6.5 – Exemple de visuel pour la comparaison de la proportion de lots livrés en retard pour des solutions issues de différentes version de l'outil d'aide à la décision

De plus, les fichiers d'analyse sous ce format permettent de facilement ajouter de nouvelles versions à comparer au sein du fichier d'origine. Ainsi, il est par exemple possible de reprendre le fichier résultats concernant la comparaison de différentes règles de lissage pour la capacité finie comme vu dans le chapitre 5, d'y ajouter de nouveaux fichiers résultats issus du test d'une nouvelle règle développée, et de rapidement mettre en forme et intégrer les indicateurs des autres versions.

En résumé, nous avons développé dans cette thèse un cadre facilitant la création d'études comparatives entre différentes versions ou modes OPERA sur la base d'un important jeu d'instances, riches en information et diversifiées, issues de situations réelles de l'usine. De plus, du fait de la richesse des données contenues dans ces fichiers d'instances, il est envi-

sageable de les utiliser dans le cadre de tests pour d'autres outils de l'entreprise. D'autre part, de la même façon que pour l'automatisation des tests de non-régression, nous avons crée un package permettant de configurer aisément différentes versions à exécuter de l'outil OPERA, puis d'automatiquement agréger les résultats afin de pouvoir rapidement comparer et analyser les performances des différentes versions selon plusieurs indicateurs clés. Par ailleurs, c'est au travers de cet outil qu'ont pu être exécutés les nombreux tests comparatifs présentés dans les différents chapitres de ce manuscrit.

6.3 Interfaces pour l'analyse des données et l'aide à la décision

Dans le développement d'outils d'aide à la décision pour la planification (production, capacité, ...), la recherche de la solution optimale dans un temps minimum ne correspond souvent qu'à une partie du travail. Alors que les problématiques de planification en milieu industriel ont attiré beaucoup de chercheurs, avec un très grand nombre de travaux, force est de constater qu'un nombre limité d'articles se sont intéressés à l'intégration de leur solution au sein du système considéré. Cela n'est pas forcément le témoin d'un désintéressement du monde académique vis-à-vis de ces sujets, mais est probablement plutôt le marqueur d'une difficulté de collaboration avec les entreprises privées.

Dans le cadre d'une thèse CIFRE, le développement d'un outil d'aide à la décision à intégrer dans le système est facilité, et nous présentons dans la suite de ce chapitre les interfaces développées pour les différents processus métiers, avec la question centrale: "Quelles données montrer à l'utilisateur pour l'aider à prendre les meilleures décisions possibles?" Car si l'on reprend le modèle mathématique présenté en chapitre 3, donner pour résultat les valeurs pour l'ensemble des variables serait tout à fait contre-productif, car l'utilisateur serait noyé sous un trop grand nombre de données. De même, bien que l'heuristique proposée permette de donner pour chaque étape de process de chaque lot une date exacte de passage, on se rend compte qu'à l'échelle de l'usine, cette information est trop précise et de toute façon trop sensible à la variabilité intrinsèque d'une usine, comme souligné dans Horiguchi et al. (2001). On comprend donc qu'il est inutile d'être trop exhaustif et précis dans l'information fournie.

À l'inverse, présenter des données trop agrégées, telles que des informations globales d'activité (tout produits confondus) attendue par atelier et par période, serait certes plus proche de la réalité, mais cacherait également un grand nombre d'informations qui pourraient être utiles à l'utilisateur.

Les deux sections suivantes traitent de processus métiers intégrant OPERA dans leur fonctionnement. Le premier vise à définir consignes de production pour la semaine à suivre (aide à la décision tactique), et l'autre au contrôle quotidien de l'évolution des flux de production (aide à la décision opérationnelle). Dans les deux cas, la solution d'analyse a été développée sous l'outil *Power BI* afin de faciliter la lecture et le croisement des données. À noter que dans ces sections, les graphiques et visuels présentés dans les diverses figures ont été générés via des données fictives, pour des raisons de confidentialité. De plus, les noms des machines et produits seront floutés, de même que les chiffres, le but étant d'avoir une idée des informations pouvant être obtenues via les différents fichiers.

6.3.1 Aide à la décision tactique

Au sein de l'usine, on trouve un grand nombre de ressources aidant à l'optimisation de la production. Que ce soit des outils d'aide à la planification, tels que des outils d'ordonnancement ou de dispatching des lots sur les machines, ou bien tout simplement de ressources humaines dont le but est d'optimiser localement l'utilisation d'un ensemble de ressources. La somme des optima locaux ne correspondant généralement pas à un optimum global, il est important de pouvoir piloter ces différents outils d'optimisation locale au travers d'autres outils d'optimisation plus globaux. C'est dans cet objectif que, chaque semaine, OPERA est utilisé pour la définition d'un plan de production à suivre. Ce plan de production, nommé PRI pour *Plan de Référence Interne*, prend en entrées des consignes à suivre et notamment le plan de lancement de nouveaux lots ainsi que les livraisons prévues pour chaque semaine, mais aussi d'autres informations sur l'usine telles que le positionnement du WIP ou les caractéristiques des machines. Si l'on prend pour exemple la figure 6.6 déjà présentée dans le chapitre 1, le PRI est une des fonctions de l'*Operational Production Planning* dont le but est de fournir des consignes de production telles que des WIP cibles pour chaque jour de la semaine pour certains groupes de machines ou des produits et opérations prioritaires durant la semaine à venir.

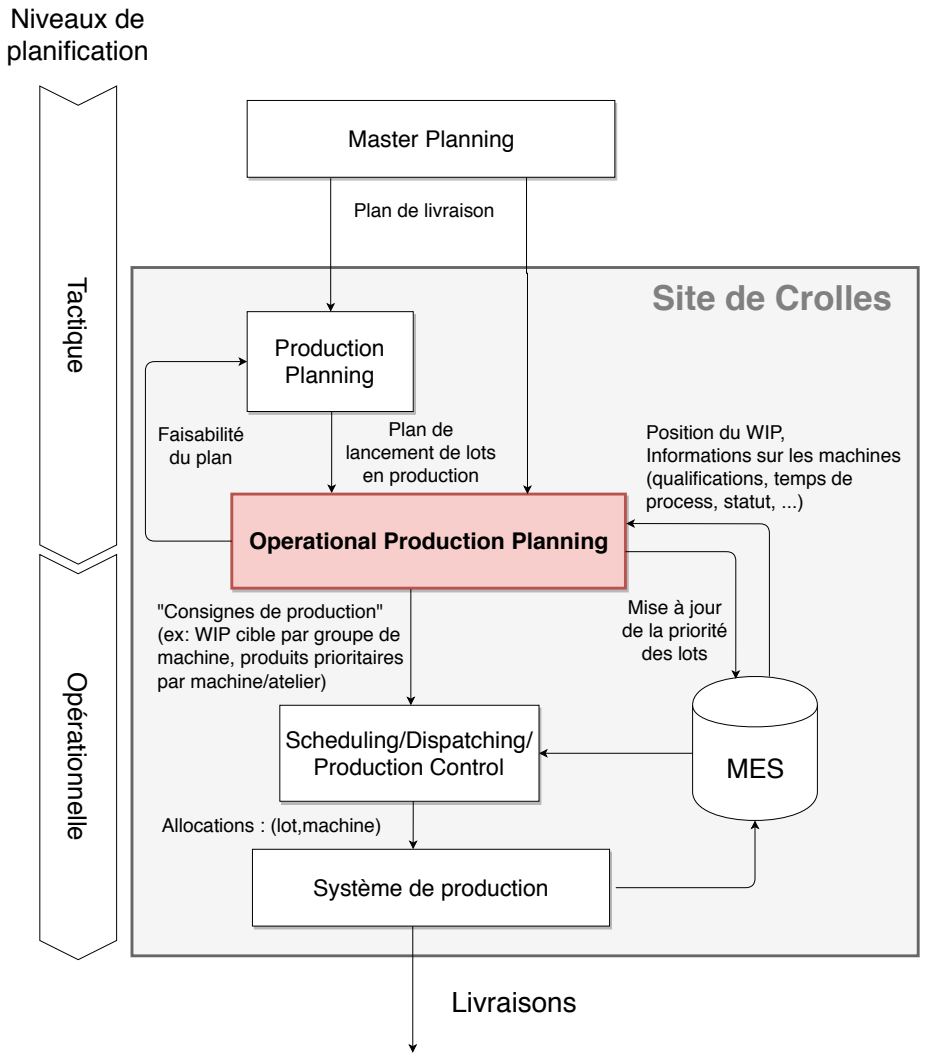


FIGURE 6.6 – Différentes étapes de planification de la production à ST Crolles

Cette solution est présentée une fois par semaine aux principaux responsables de production afin de donner les lignes directrices qui permettront en théorie de répondre aux engagements de livraison. La question est alors de savoir ce qui doit être présenté à ces différents acteurs car, comme nous l'avons vu précédemment, il n'est pas possible de donner simplement l'ensemble des fichiers de sortie à chacun d'entre eux. Les données seraient là, mais pas l'information.

Le lecteur trouvera donc ci-dessous une présentation du fichier d'analyse, et de ses différentes vues. Ce fichier est composé de plusieurs onglets présentant des informations sous plusieurs agrégations selon les questions auxquelles on cherche à répondre. À l'origine, l'analyse et la présentation des données étaient faites à l'aide d'Excel, et le comparatif sera donc également fait avec les visuels d'origine.

Du point de vue de l'activité globale de l'usine

Quelle activité globale doit-on réaliser pour atteindre nos objectifs, et principalement sur quelles parties de la ligne de production et pour quels produits?

Pour répondre à cette question, le fichier Excel fournissait le visuel de la figure 6.7. On y voit des informations globales de l'usine pour les différentes semaines, telles que l'activité, le turn moyen (nombre moyen d'opérations réalisées par lot et par jour) ou le WIP prévu. Il est également possible de décomposer ces informations selon les différents jalons, ou de filtrer selon les différents produits afin de déterminer l'activité prévisionnelle requise permettant de suivre les engagements, notamment pour les principaux d'entre eux.

	A	C	D	E	F	G	H	I
1	LotOrigine					C300 - C:\OPERA\C300\PRI_2019W20\		
2	DetailledLotType	PROD FE	90160		ENG	6150		
3	Counted	PROD BE	62770					
4	Techno	PROD ALL	152930		R&D	8690		
5								
6								
7		2019W21	2019W22	2019W23	2019W24	2019W25	2019W26	
8	Row Labels	7,0	7,0	7,0	7,0	7,0	7,0	
11	1-FE	89420	86112	85841	89344	86092	87033	
12	2-BE	62244	64574	61561	57771	54875	55044	
13	Avg Wip							
14	1-FE	41 119	40 627	40 557	41 029	41 367	41 421	
15	2-BE	23 095	22 489	21 426	20 265	19 173	18 720	
16	Turn							
17	1-FE	2,17	2,12	2,12	2,18	2,08	2,10	
18	2-BE	2,69	2,87	2,87	2,85	2,86	2,94	
19	Gate Out							
20	1-FE	29 774	27 151	27 364	28 799	27 413	29 151	
21	2-BE	23 502	21 675	21 473	20 209	17 801	18 796	
22	**PROD MovesOutPerDay**	**151664**	**150685**	**147402**	**147115**	**140967**	**142077**	
23	**PROD Avg Wip**	**64 214**	**63 116**	**61 983**	**61 294**	**60 540**	**60 141**	
24	**PROD Turn**	**2,36**	**2,39**	**2,38**	**2,40**	**2,33**	**2,36**	
25	**PROD Gate Out**	**53 276**	**48 826**	**48 837**	**49 008**	**45 214**	**47 947**	
26	**ENG**							
27	MovesOutPerDay	5927	6279	5800	6603	5850	5727	
28	Avg Wip	4 961	4 768	4 678	4 618	4 571	4 546	
29	Turn	1,19	1,31	1,24	1,43	1,28	1,26	
30	Gate Out	1 879	2 051	1 790	2 110	1 838	1 742	
31	**R&D**							
32	MovesOutPerDay	8470	7630	7952	7854	8222	9409	
33	Avg Wip	5 193	5 446	5 509	5 558	5 706	5 809	
34	Turn	1,64	1,40	1,44	1,41	1,44	1,62	
35	Gate Out	2 773	2 063	2 450	2 708	2 391	3 195	
36	**Total MovesOutPerDay**	**166061**	**164595**	**161155**	**161572**	**155038**	**157212**	
37	**Total Avg Wip**	**74 368**	**73 330**	**72 170**	**71 470**	**70 817**	**70 496**	
38	**Total Turn**	**2,23**	**2,24**	**2,23**	**2,26**	**2,19**	**2,23**	
39	**Total Gate Out**	**57 928**	**52 940**	**53 077**	**53 826**	**49 443**	**52 884**	

FIGURE 6.7 – Visuel d'origine sous Excel pour la visualisation des résultats de planification fournis par l'outil OPERA

Ces informations sont également disponibles dans le nouveau fichier proposé avec l'onglet

"Activity Analyis" (voir figure 6.8), avec cependant des informations supplémentaires sur les quantités de lots démarrés chaque semaine, ainsi que ceux devant être livrés. Il est ainsi possible de répondre à d'autres questions telles que: "Parmi l'activité prévue, quelle est celle requise pour effectuer les livraisons à court terme?" La figure 6.9 montre un exemple où, en sélectionnant la première semaine dans le graphique C, on obtient grâce au visuel A cette fois des informations sur l'activité requise pour effectuer les livraisons de la semaine courante. Cette activité requise vaut pour tous les produits, mais il est également possible de croiser davantage l'information en sélectionnant par exemple un produit dans le visuel B, ce qui nous informe sur l'activité nécessaire afin d'effectuer les livraisons prévues pour le produit considéré. Ces informations sont particulièrement intéressantes pour des produits sensibles dont on estime le retard comme très pénalisant.

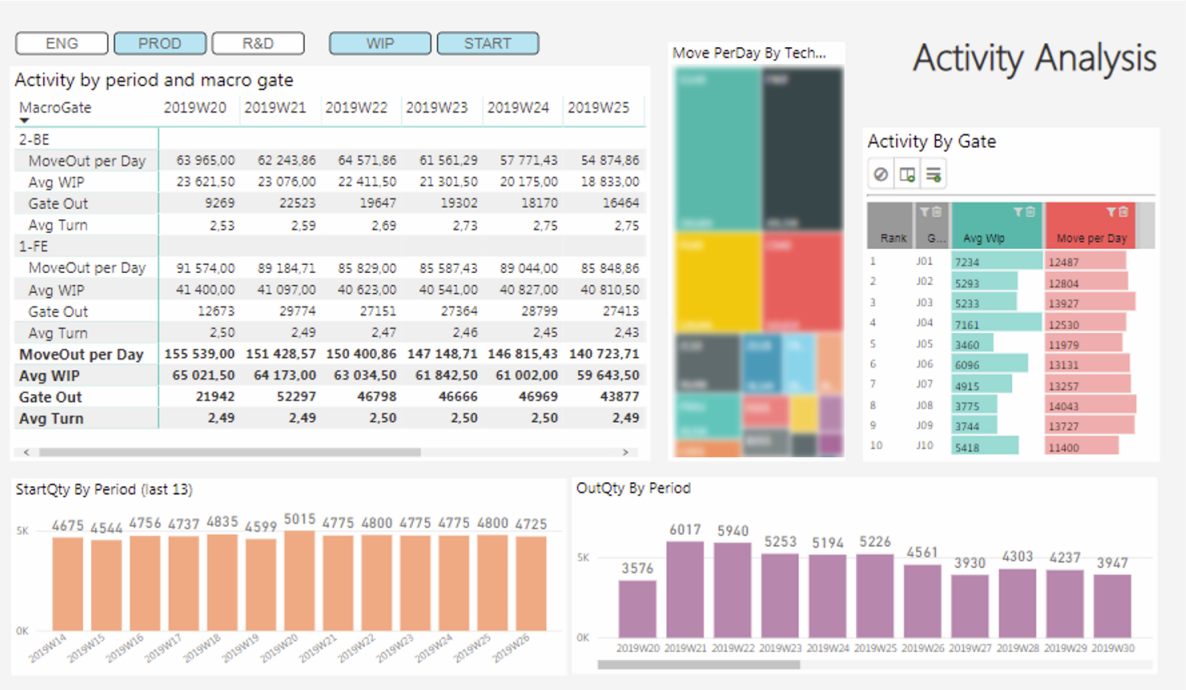


FIGURE 6.8 – Nouveau visuel d'analyse des résultats de planification fournis par l'outil OPERA

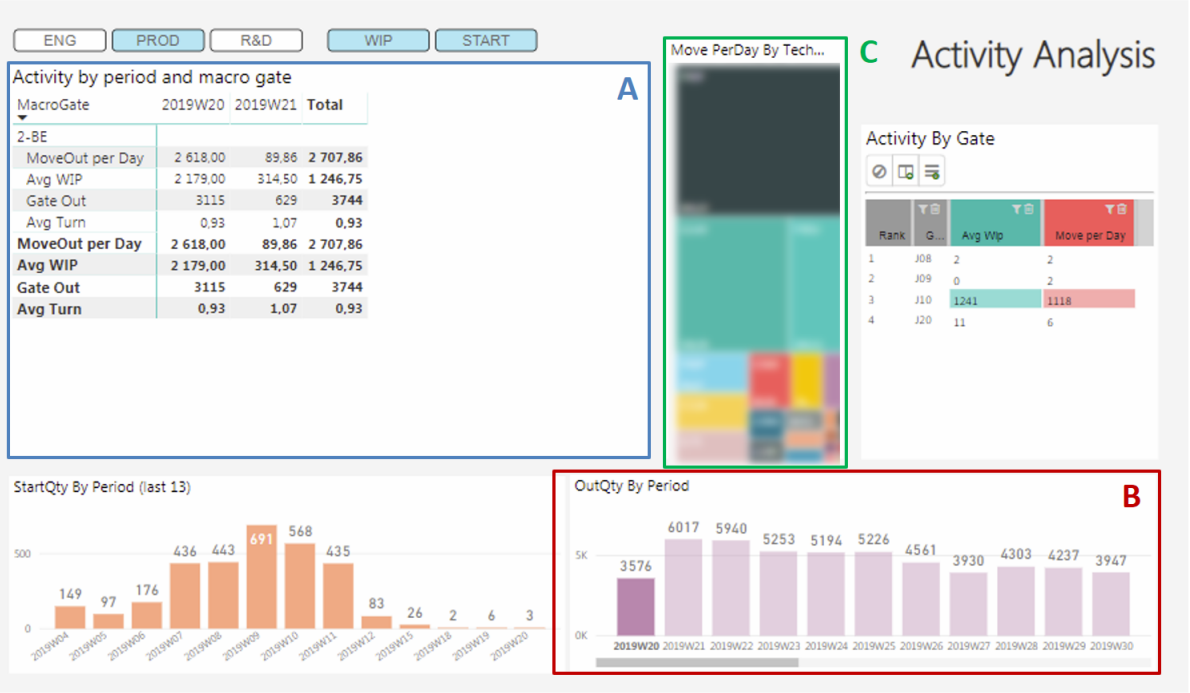


FIGURE 6.9 – Exemple de croisement de données pour déterminer l'activité théorique requise pour permettre les livraisons de la semaine courante

Grâce à ce type de croisements, il est également possible d'avoir des informations sur l'activité générée par le plan de lancement (planning, pour chaque produit, des volumes à lancer en production) des lots dans l'usine. En effet, il arrive parfois que le plan de production donne une activité globale que l'on qualifiera d'agressive, dans le sens où elle est supérieure à celle généralement observée d'un point de vue historique. Ce plan est théoriquement faisable (car généré par l'outil de planification à capacité finie), mais également difficile à tenir. Dans ce cas, il peut être intéressant d'évaluer l'impact du plan de lancement sur l'activité qu'elle génère. Car il est important de rappeler que ces nouveaux lots peuvent entrer en concurrence sur certaines machines avec des lots qui eux doivent bientôt être livrés, du fait du système de production avec flux ré-entrants. Par conséquent, un plan de lancement trop agressif peut compromettre le respect des livraisons à court terme. Ces informations peuvent être obtenues grâce au nouveau fichier, simplement en sélectionnant la période souhaitée dans le visuel B, les résultats étant toujours affichés dans le visuel A. Un cas est présenté en figure 6.10, où est donnée l'activité générée par des lots lancés en production les semaines 21 et 22, la plupart étant associés au démarrage de produits de type X (bleu), Y (noir) et Z (rouge). Ainsi, si le plan de lancement devait être revu à la baisse du fait d'une trop forte activité, les principaux leviers seraient déjà identifiés.

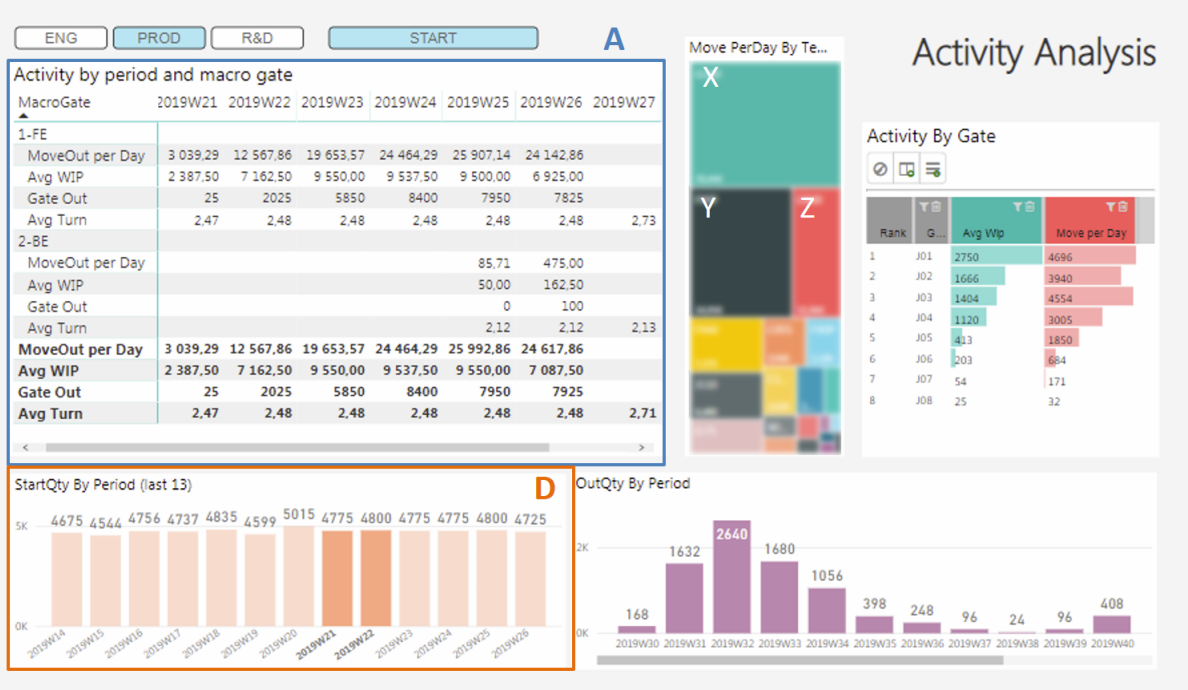


FIGURE 6.10 – Exemple de croisement de données pour déterminer l'activité théorique générée par le plan de lancement des lots

Du point de vue des machines

Nous savons donc quelle activité prévisionnelle est théoriquement requise pour suivre au mieux le plan de livraison, mais quel est l'impact sur les machines? Dans la diversité des équipements qui composent l'usine, quels sont ceux qui seront les plus impactés, les plus "tendus"? Et quels seront les principaux produits à affecter à ces équipements? Ce sont certaines questions régulièrement posées par les différents acteurs avec pour but de gérer la production sur un atelier ou un groupement de machines.

Une partie des ces questions trouve sa réponse dans le fichier Excel. La figure 6.11 est un exemple des informations mises à disposition. On y retrouve pour chaque machine et chaque semaine l'activité prévue, c'est à dire le nombre de plaquettes devant effectuer une étape de process sur la machine considérée pour chaque semaine. D'autres informations sont disponibles, telles que, pour chaque machine, le WIP moyen attendu ou bien le taux d'utilisation (noté EUR pour *Equipment Utilization Rate*). Un code couleur permet d'identifier les machines pour lesquelles la charge prévisionnelle est la plus importante par rapport à leur capacité.

	A	B	C	D	E	F	G	H
1	Area	(Multiple Items)						
2								
3		Column Labels						
4		2019W20	2019W21	2019W22	2019W23	2019W24	2019W25	2019W26
5	Row Labels	3,0	7,0	7,0	7,0	7,0	7,0	7,0
6	Machine1							
7	EUR	52%	57%	63%	60%	76%	73%	67%
8	Vulnerability Factor	83%	90%	100%	96%	120%	115%	106%
9	OEE	37%	40%	44%	43%	54%	52%	48%
10	Track Qty Per Day	367	383	403	399	469	459	441
11	nbEqpt	1	1	1	1	1	1	1
12	Required NbEqpt	0,8	0,9	1,0	1,0	1,2	1,2	1,1
13	Extra NbEqpt	0,2	0,1	0,0	0,0	-0,2	-0,2	-0,1
14	WipBoh	50	25	25	0	100	186	38
15	WipEoh	25	25	0	100	186	38	0
16	Das (hrs)	3,64	4,13	4,55	4,49	4,87	4,97	4,47
17	PostPoned Qty	0	0	0	0	0	0	0
18								
19	EUR	94%	67%	85%	63%	104%	72%	78%
20	Vulnerability Factor	122%	87%	110%	82%	135%	93%	101%
21	OEE	69%	49%	62%	47%	77%	53%	57%
22	Track Qty Per Day	270	196	259	185	321	204	211
23	nbEqpt	2	2	2	2	2	2	2
24	Required NbEqpt	2,4	1,7	2,2	1,6	2,7	1,9	2,0
25	Extra NbEqpt	-0,4	0,3	-0,2	0,4	-0,7	0,1	0,0
26	WipBoh	125	25	53	25	103	25	88
27	WipEoh	25	53	25	103	25	88	50
28	Das (hrs)	9,51	10,04	9,50	10,08	9,85	11,66	13,63
29	PostPoned Qty	0	0	0	0	0	0	0
30								
31	EUR	98%	98%	98%	98%	98%	98%	98%
32	Vulnerability Factor	100%	100%	100%	100%	100%	100%	100%
33	OEE	74%	74%	74%	74%	74%	74%	74%
34	Track Qty Per Day	1 714	1 768	1 832	1 791	1 803	1 802	1 803
35	nbEqpt	1	1	1	1	1	1	1
36	Required NbEqpt	1,0	1,0	1,0	1,0	1,0	1,0	1,0
37	Extra NbEqpt	0,0	0,0	0,0	0,0	0,0	0,0	0,0
38	WipBoh	1 939	2 018	1 379	1 648	2 705	3 170	3 646
39	WipEoh	2 018	1 379	1 648	2 705	3 170	3 646	30
40	Das (hrs)	17,41	20,10	25,36	32,30	33,94	34,04	28,47
41	PostPoned Qty	0	45	15	43	142	120	172
42								

FIGURE 6.11 – Page Excel d'analyse de l'impact du plan de production sur l'activité prévisionnelle des machines

Ainsi, les différents utilisateurs obtiennent des informations sur l'activité prévisionnelle des machines, mais il manque la dimension lot avec les produits et les opérations à considérer. En effet, on note par exemple que la "Machine1" doit réaliser une activité prévisionnelle de 383 plaquettes chaque jour pour la semaine à venir. Mais face à plusieurs lots concurrents, lesquels sont à privilégier? Ne pas donner d'information à ce sujet revient à laisser les différents ateliers optimiser localement, sans gestion globale, ce que nous souhaitons éviter. Ainsi, il est nécessaire d'apporter davantage d'informations, tous d'abord sur les lots à privilégier selon le type de produits, mais également selon leur position sur leur route. En effet, du fait des flux ré-entrants, se pose pour une même machine la question de la priorité pour des lots d'un même produit mais à des positions différentes sur leur route. Il semblerait assez naturel de penser qu'il suffit de toujours faire passer le lot le plus proche d'être fini, ou dont la date de livraison est la plus proche. Mais peut-être que ce lot n'est finalement pas en retard, voire même plutôt en avance sur sa date de livraison, et qu'un autre lot plus en amont est quant à lui plus en retard, ou bien pourrait permettre d'alimenter par la suite d'autres machines actuellement sous utilisées. La question de la position des lots (les opérations à privilégier) est donc tout aussi importante que celle des types de produit.

Le nouveau fichier développé permet de visualiser facilement ces différentes informations à travers l'onglet "Tool Saturation" dont un aperçu est visible dans la figure 6.12. Le visuel A permet de sélectionner l'atelier désiré, tandis que la matrice B présente les données qui étaient visibles dans le fichier Excel. S'ajoute à cela la possibilité de connaître la répartition de l'activité selon les différents produits et jalons grâce aux visuels C et D. Contrairement

au fichier Excel, dans ce nouveau format, il est facile de connaître les principaux types de lots sur lesquels axer la production. À titre d'exemple, le responsable de l'atelier TT (Traitement Thermique) peut rapidement constater (illustration dans la figure 6.13) l'activité attendue sur les différentes machines de l'atelier, avec une répartition assez homogène sur les produits et jalons mais cependant un accent particulier sur le jalon numéro 8. En croisant d'autres données, comme le montre la figure 6.14, le responsable peut également noter que 5 machines seront particulièrement chargées durant les deux semaines à venir, et que leur activité devra notamment se concentrer sur le jalon 6. Cette information est importante car la simple information de l'activité prévisionnelle pourrait amener l'acteur à privilégier l'activité sur d'autres jalons, ce qui aurait pour conséquence probable un déséquilibre des flux de production et un risque accru de ne pas respecter les dates de livraison à plus ou moins long terme.

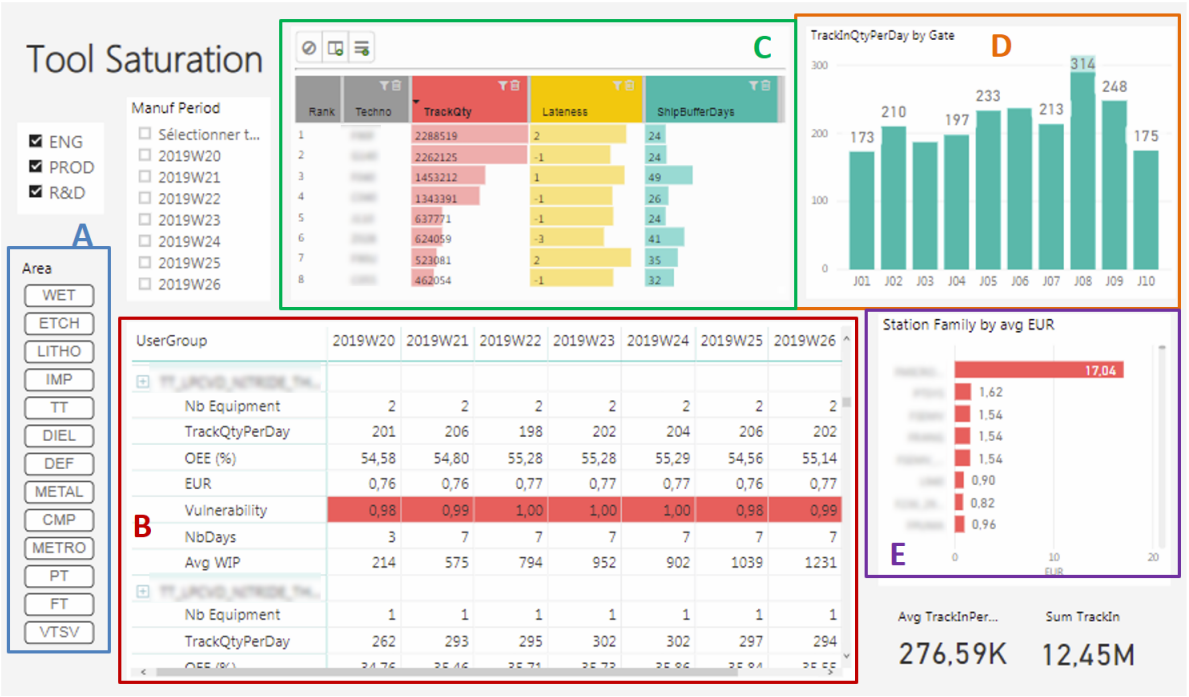


FIGURE 6.12 – Nouvelle version d'analyse de l'impact du plan de production sur l'activité prévisionnelle des machines

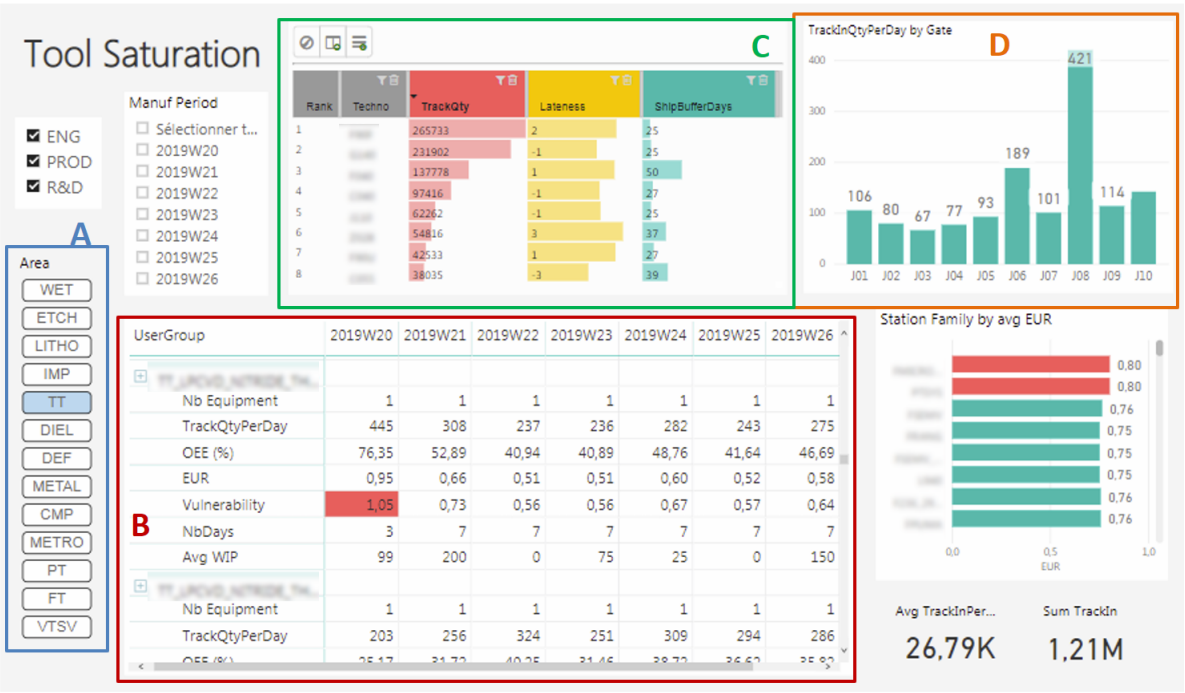


FIGURE 6.13 – Exemple de croisement de données permettant de connaître pour un atelier l'activité attendue et l'évolution de différents indicateurs

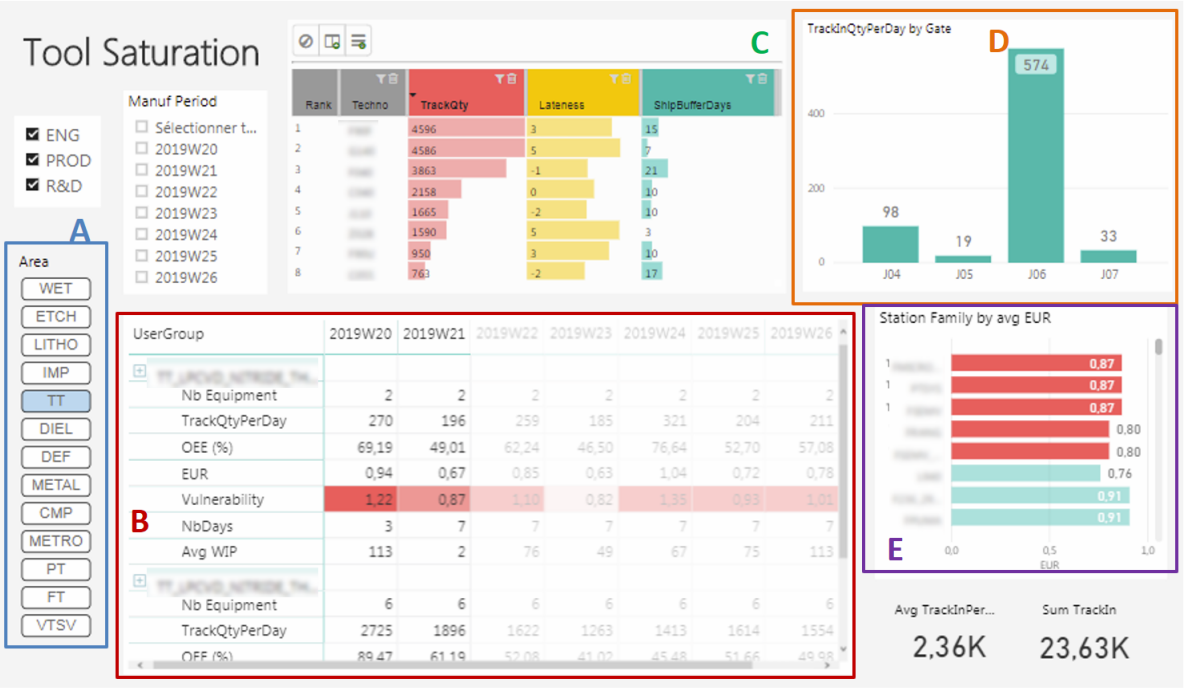


FIGURE 6.14 – Croisement poussé des données permettant de définir les équipements les plus sensibles de l'atelier TT ainsi que le type précis de lots à traiter

Du point de vue des livraisons

En plus d'informations sur les différentes machines ou sur les performances globales de l'usine, le fichier de visualisation permet également d'informer sur le planning de livraison des différents lots. Ces éléments ont plusieurs intérêts. Pour ceux qui développent les plans de production (à l'aide de l'outil), les visuels donnent des clés de compréhension permettant d'expliquer certains choix que fait l'outil. Pourquoi privilégier ces produits dans certains ateliers? Pourquoi constate-t-on peu d'activité pour tel autre produit sur tel jalon? Les réponses à ces questions se trouvent parfois dans l'analyse du planning de livraison des lots. Ensuite, ces visuels permettent de communiquer avec d'autres acteurs, notamment les responsables du planning des livraisons et de lancement des lots, afin de détecter de potentiels retards et d'éventuellement réajuster le planning des livraisons.

Ainsi, différentes informations sont fournies dans le fichier via deux onglets "Outs By Techno" et "Outs By Lateness", illustrés respectivement dans les figures 6.15 et 6.16. Ces deux onglets présentent des informations sur le plan de livraison des lots mais sous deux angles différents. Le premier donne une vue synthétique des différentes quantités de produits devant être livrées chaque semaine, tandis que le second insiste sur la position des lots vis-à-vis de leur date de livraison. Pour cela, on procède de la même manière que pour le calcul du retard/avance présenté dans le chapitre 5 pour la définition de la règle de lissage OD. Pour rappel, cette évaluation de l'état d'avancement d'un lot se fait en comparant sa date de livraison à la date où le lot sortirait s'il avançait selon son temps de cycle théorique (évalué à partir d'analyses historiques, voir chapitre 3.4.2 pour plus d'informations).

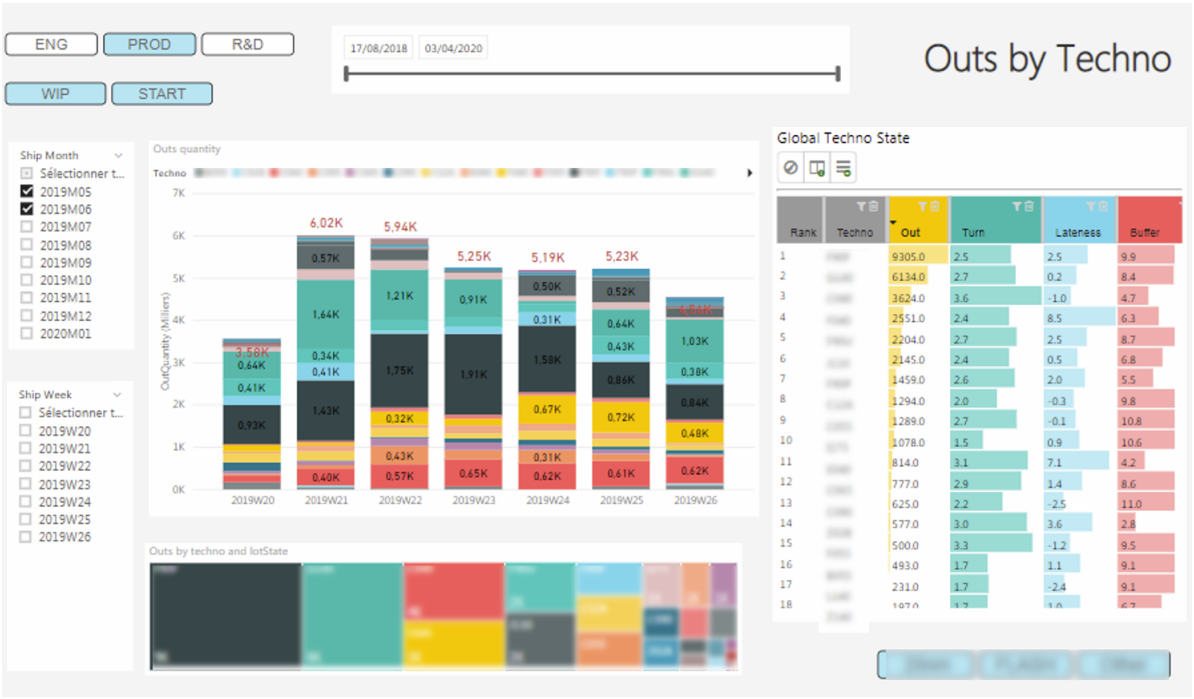


FIGURE 6.15 – Onglet de visualisation du plan de livraison considéré par l'outil de planification OPERA avec détail par type de produits

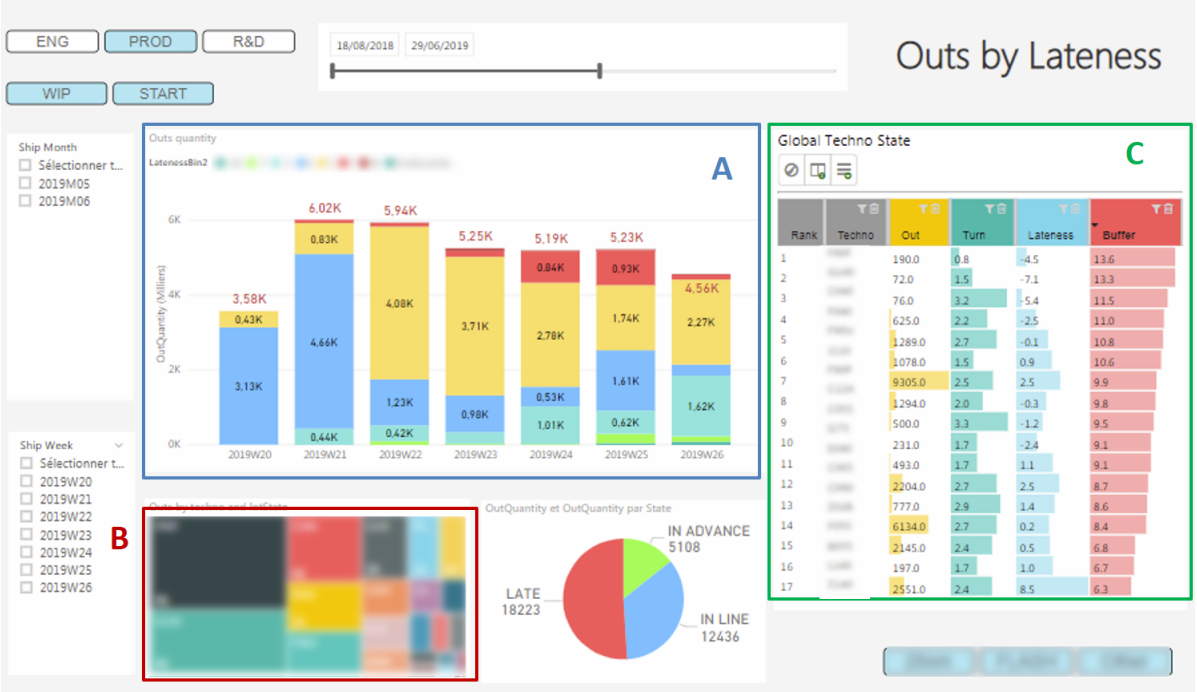


FIGURE 6.16 – Onglet de visualisation du plan de livraison considéré par l'outil de planification OPERA avec détail par situation de retard/avance des lots

Ainsi, pour chaque semaine de l'horizon, le visuel A informe sur les quantités de plaquettes à livrer et leur état d'avancement. On note par exemple que pour les plaquettes devant être livrées la semaine 21, 830 sont considérées comme ayant plus d'un jour de retard (catégorie jaune "-1") mais tout de même moins de 7 jours de retard (car sinon seraient dans la catégorie orange "-7"). Il est donc nécessaire d'être vigilant afin de garantir les livraisons, mais celles-ci semblent néanmoins être tenables. Il est ensuite possible d'avoir davantage d'informations sur les principaux produits pour ce retard, en les sélectionnant dans le visuel A et en regardant les volumes de produits dans les visuels B ou C.

Le visuel A permet également de prévenir d'éventuels risques. On note notamment que pour les semaines 24 et 25, il y a respectivement 840 et 930 plaquettes qui accusent un retard potentiel de plus de 7 jours. Cette situation n'est pas irrémédiable car il est possible, en donnant des consignes de production, de rattraper ce retard d'ici les semaines de livraison. Mais il est cependant important de prendre ces décisions suffisamment en amont avant que le temps disponible, et donc la marge de manoeuvre, soit insuffisant et qu'il soit nécessaire de revoir le plan de livraison. Ce cas pratique illustre encore l'intérêt et l'importance des informations fournies par le fichier.

6.3.2 Aide à la décision opérationnelle

Le fichier présenté dans la section précédente permet de visualiser différentes informations concernant le plan de production proposé par l'outil OPERA et qui servira de ligne directrice pour la semaine à venir. Mais l'outil est également utilisé dans un cadre encore plus opérationnel en gérant quotidiennement la priorité des lots, comme cela a été mentionné dans le chapitre 1. Les résultats issus des exécutions journalières sont stockés dans une base de données de telle sorte que sont continuellement sauvegardés les résultats des 70 derniers jours. L'intérêt de stocker ces résultats est qu'ils permettent notamment de visualiser l'évo-

lution de certains indicateurs de l'usine dans le temps, alors que l'analyse présentée dans la section précédente ne donne les informations que pour une planification exécutée à une date donnée.

Ces informations intéressent particulièrement les services responsables de la gestion globale des flux de production au sein de l'entreprise. Ces derniers, travaillant étroitement avec le service en charge des plannings de livraison et de mise en production, ont pour objectif d'assurer la bonne évolution des flux de production afin d'assurer le respect des dates de livraison. Leurs principaux leviers sont la mise en place de consignes d'activité ponctuelles (par exemple produire X plaquettes d'un produit A à l'étape Z dans les prochaines 24 heures), ou plus souvent des actions sur la priorité des lots (ralentissement manuel des lots les plus en avance, accélération des lots les plus en retard). Ces actions se font en complément des mises à jour journalières et automatiques réalisées par l'outil OPERA.

Le fichier présenté ci-dessous est un outil quotidien d'aide à décision pour plusieurs utilisateurs. Il est composé de nombreux onglets, et nous nous cantonnons à seulement trois d'entre eux pour illustrer quelques questions auxquelles ils tentent de répondre.

Quel est l'état de l'usine aujourd'hui? La situation est-elle en cours d'amélioration?

Pour répondre à cette question, l'utilisateur peut regarder le premier onglet intitulé "Global View', dont un aperçu est présenté dans la figure 6.17. Le visuel A donne la proportion de lots étant actuellement soit en retard, en avance, ou en ligne par rapport à leur date de livraison prévue. À noter qu'un lot est réellement considéré comme en retard si l'indicateur "On Time Deviation", définit dans le chapitre 5, présente un retard supérieur à 1 jour. Il en est de même, symétriquement, pour les lots en avance.

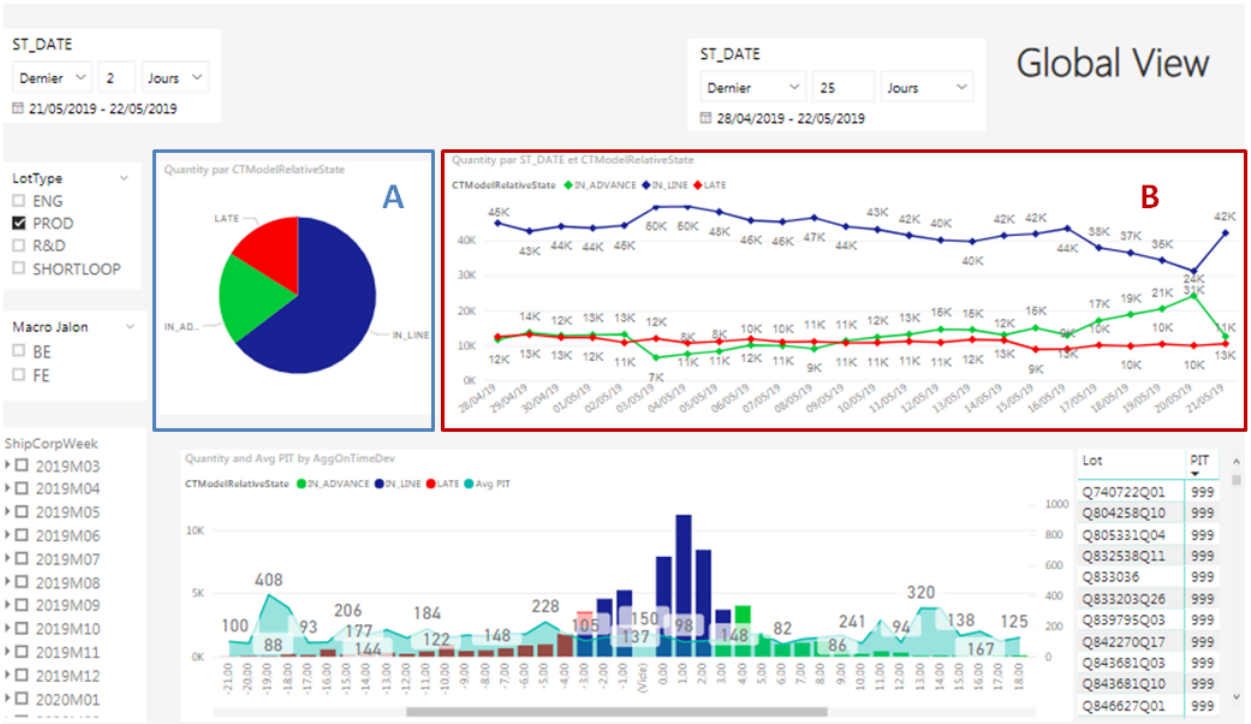


FIGURE 6.17 – Onglet permettant de visualiser l'état de l'usine en termes d'avance/retard des lots ainsi que son évolution dans le temps

Ce visuel donne l'état instantané de l'usine mais ne donne aucune information sur la répartition de ces retards ou sur leur tendance. Pour compléter l'information, le visuel B retrace l'ensemble des états de l'usine, pris chaque matin, jusqu'à un horizon souhaité dans le passé. Il permet de voir non seulement qu'actuellement l'usine présente un grand nombre de lots considérés comme "en ligne" par rapport à leur date de livraison, ainsi qu'un nombre équivalent de lots en avance et en retard, mais aussi que cet état a brusquement changé dernièrement avec un nombre de lots trop en avance augmentant continuellement. Cette cassure des courbes au niveau du 20/21 Mai montre un changement brutal du nombre de lots considérés en ligne (augmentation) ou en avance (diminution), et laisse entendre qu'il y a eu une revue des plans de livraison et que les dates de livraison des lots ont évolué (en l'occurrence, avancé) en conséquence.

Où se situent les principaux points de blocage actuellement?

Après avoir constaté la situation globale de l'usine ainsi que les tendances d'évolution, l'onglet "OnTime State By Gate" (illustré dans la figure 6.18) permet de visualiser les principales zones où le retard semble s'accumuler. L'idée est de comprendre grâce au visuel A dans quels jalons sont situés les lots en retards, ce qui permet d'évaluer les lots sur lesquels l'action doit être particulièrement portée. Par exemple, on constate que l'on trouve des lots en situation de retard dans la plupart des jalons. On note toutefois un grand volume de lots en retard dans le jalon 6, lui même déjà très chargé globalement (plus de 9000 plaquettes actuellement). Il semble donc important d'agir dans un premier temps sur les lots récents dans ce jalon, afin de réduire le WIP dans celui-ci et surtout celui des lots en situation de retard (facilement identifiables grâces aux autres visuels). De plus, il serait inutile de chercher avant tout à faire passer les lots en retard du jalon 5, étant donné que ceux-ci ne feraient qu'alimenter un prochain jalon déjà en manque de ressources. Les visuels B et C permettent de préciser l'information en déterminant les principaux produits composants ces lots en retard. Il est même possible pour l'utilisateur d'aller chercher dans le visuel D des informations sur les lots en retard, et pour lesquels il pourrait être nécessaire d'effectuer une gestion spécifique de leur évolution.

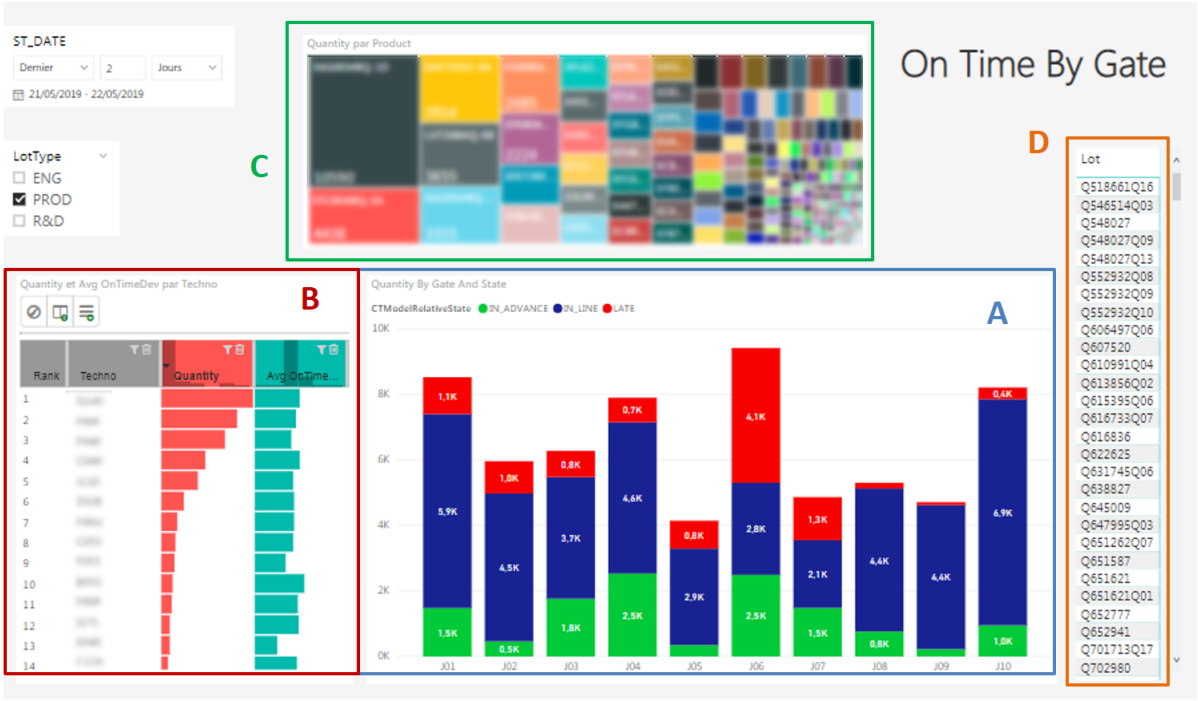


FIGURE 6.18 – Visualisation des états d'avance et de retard des lots selon leur type et leur positionnement dans la ligne de production

Existe-t-il des lots en situation anormale?

Nous avons pu le constater au fil des chapitres, la fabrication de semi-conducteurs (et le site de Crolles n'y échappe pas) implique généralement de travailler avec des volumes importants et des flux complexes. Cette complexité génère une quantité conséquente de données à tel point qu'il peut parfois devenir difficile de contrôler des situations anormales mais de faible ampleur. À titre d'exemple, on peut prendre le cas de l'onglet "Techno View" présenté en figure 6.19. Le visuel principal permet de fournir en une figure un grand nombre d'informations. Chaque colonne de l'axe des abscisses représente une semaine de livraison, tandis que chaque couleur représente une technologie (agrégation de produits). La taille des bulles informe sur le volume à livrer pour chaque technologie, et enfin la position sur l'axe des ordonnées indique le retard/avance moyen pour chacune de ces technologies. Ainsi, on constate par exemple qu'il y a quelques lots (en zoomant dessus on constaterait qu'il y en a deux) d'une technologie (cercle orange dans la deuxième colonne) devant être livrées la semaine 22 et qui ont un retard moyen de 12 jours. Ce point indique donc une situation soit anormale, car il y a une erreur dans les dates de livraison de ces deux lots, soit un risque important de livrer ces lots en retard. Cet onglet permet donc de régulièrement (et facilement) détecter des cas étranges difficilement détectables du fait du volume des données traitées, mais qui pourtant peuvent amener à perturber le bon fonctionnement du système.

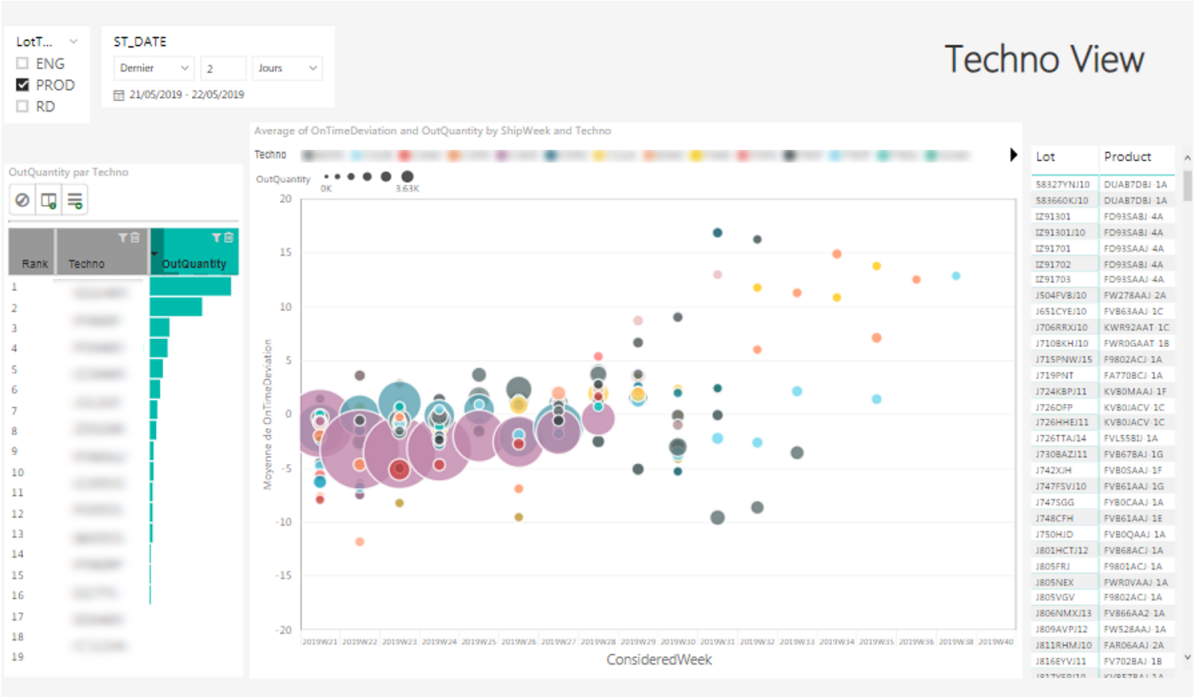


FIGURE 6.19 – Positionnement d'agrégations de produits selon les différentes semaines de livraison et leur état d'avancement

6.4 Conclusion et Perspectives

Dans ce chapitre, nous avons eu l'occasion de nous éloigner un peu de la dimension scientifique de résolution du problème de planification de production opérationnelle, pour nous intéresser à un aspect ayant bénéficié d'un nombre limité de travaux, à savoir l'intégration d'un outil d'aide à la décision au sein du système industriel. Dans le cadre de cette thèse, cela s'est traduit par deux réalisations majeures.

La première, concerne la création d'un environnement favorisant le développement et la validation de nouvelles versions de l'outil d'aide à la décision (nommé OPERA). D'une part, nous avons proposé un outil permettant l'exécution de tests automatiques et exhaustifs pour la détection de potentielles régressions au sein de la très grande quantité de données générées par l'outil. D'autre part, la constitution d'un ensemble de 50 instances tirées de situations réelles et de fichiers pré-configurés a permis de faciliter l'étape d'analyse et de comparaison de différentes versions de l'outil.

Le deuxième grande réalisation concerne le développement de fichiers de visualisation à partir d'un logiciel de *Business Intelligence*, rendant possible pour l'utilisateur de croiser les données de façon dynamique et permettant de tirer profit des solutions très détaillées proposées par l'outil d'aide à la décision. Ces fichiers ont été développés dans le cadre de deux processus métiers. Pour le premier processus, lié à la création d'un plan de production directeur pour la semaine à venir, la nouvelle interface d'analyse permet d'identifier des informations globales, telles que des consignes d'activité ou de WIP pour chaque jalon, jusqu'à des informations sur les prévisions de charges de machines et des principaux produits et étapes à réaliser. Le second processus, quant à lui, est dédié au suivi quotidien des lots dans l'usine. Grâce aux nouveaux visuels développés, l'utilisateur peut suivre d'un simple coup d'œil la tendance générale d'état d'avancement des lots en production, savoir si le retard

a tendance à s'accumuler ou au contraire à se résorber. L'utilisateur peut également évaluer plus précisément les principaux points de blocage de l'usine, en zoomant par produit, ou bien détecter des lots en situation anormale et qui pourraient être noyés dans la masse des autres lots de l'usine.

Chapitre 7

Conclusions et perspectives

7.1 Conclusions

Cette thèse traite du problème de planification de production opérationnelle en fabrication de semi-conducteurs, et plus précisément dans les usines dites *front end*, dont les systèmes de production sont connus pour être parmi les plus complexes au monde. Du fait de la complexité des flux de production et de la taille des problèmes traités, un écart est constaté entre la planification de production tactique (*Master Planning* et *Production Planning*) en charge de déterminer des plans de livraison et de lancement des lots en production, et le *Production Scheduling* chargé d'assigner et d'ordonnancer les lots sur les différentes machines. Le problème de planification de production opérationnelle a pour objectif de faire le lien entre les fonctions de *Production Planning* et *Master Planning*, et la fonction *Production Scheduling*, en fournissant des objectifs de production à la troisième fonction afin de suivre les plans fournis par les deux premières fonctions. Ce problème frontière, essentiel dans les usines *front end* aux flux complexes, n'a pourtant fait l'objet que de peu de travaux de recherche.

Cette thèse fait suite à celle de Mhiri (2016), dans laquelle une modélisation du problème a été proposée ainsi qu'une approche heuristique en trois étapes, qui est à la base d'un outil d'aide à la planification de production opérationnelle utilisé dans deux usines *front end* de production et de R&D. Le problème de planification de production opérationnelle que nous étudions dans cette thèse est cependant différent de celui présenté dans Mhiri et al. (2018), à la fois dans la modélisation de certains aspects critiques du système de production tels que la très grande hétérogénéité des machines, et dans l'approche heuristique dont l'objectif est de fournir le meilleur plan de production plutôt que d'être un outil de simulation du système de fabrication.

Après avoir formalisé le problème étudié et montré que celui-ci est NP-Difficile, une étude expérimentale est menée afin de montrer l'impossibilité de résoudre le problème de manière exacte pour des instances industrielles, justifiant l'utilisation de méthodes approchées. Nous introduisons l'heuristique en trois étapes initialement en place dans le système de production et nous montrons à travers une étude expérimentale sur de très petites instances les bonnes performances de l'approche heuristique en termes de temps et de qualité des solutions.

L'approche initiale a cependant certains défauts importants auxquels nous avons répondus. Parmi ces défauts, il y a celui de l'équilibrage des charges sur les machines dont l'objectif consiste à assigner la charge liée aux étapes de process à réaliser sur les différentes machines. Bien qu'il ne soit pas question d'ordonnancement, cette tâche n'est pas simple du fait de la grande hétérogénéité des machines, et revêt une importance capitale dans les décisions de ma-

nagement de la capacité. La version initiale de l'outil d'aide à la décision avait des difficultés dans l'accomplissement de ces tâches, notamment car les solutions d'équilibrage présentaient régulièrement des situations d'équilibrage sous-optimaux amenant les utilisateurs à douter de la validité des plans de production. Considérant un type de problèmes classiques de partage équitables de ressources, nous montrons que son application au problème d'équilibrage des charges sur des machines non-identiques permet de fournir des solutions aux propriétés intéressantes. Nous proposons ensuite une méthode permettant d'obtenir ce type de solutions, fournissant des plans de production avec un équilibrage des charges fiable et riche de sens pour les utilisateurs.

Les plans de production devant tenir compte de la capacité limitée de l'usine afin de fournir des objectifs réalistes aux outils d'ordonnancement, l'approche heuristique contient un module dont le but est de garantir que la charge des machines respecte leur capacité. Ce module utilise une approche de lissage des charges, dont la composante centrale est la règle de sélection des lots à retarder. Dans la version initiale de l'approche, la règle utilisée a pour but de minimiser les retards de livraison. Or, aucune évaluation n'a été entreprise afin de valider les performances de cette règle de lissage afin d'optimiser cet indicateur orienté client. De plus, les critères d'optimisation dans l'industrie des semi-conducteurs étant variés et parfois antinomiques, allant du respect des commandes clients à la maximisation du nombre de plaquettes lancées en production chaque semaine, nous introduisons et évaluons différentes règles de lissage afin d'optimiser la qualité des plans de production selon différents critères de performance. Nous mettons également en évidence les limites de l'approche initiale qui, considérant individuellement la charge de chaque machine dans le processus de lissage, ne permet pas de tenir compte de la possibilité de réduire la charge d'une machine en transférant un peu de sa charge sur une autre machine moins chargée. Afin de corriger ce défaut amenant à des cas sous-optimaux de machines sous-chargées, nous introduisons une nouvelle approche de lissage basée sur le regroupement des machines selon leur partage des mêmes files d'attente. Des études expérimentales montrent les performances prometteuses de cette nouvelle approche, que ce soit en termes de qualité des solutions ou de temps de calcul. Ces résultats sont cependant à nuancer, car l'approche paraît fournir des plans de production ne respectant pas toujours la capacité de l'ensemble des machines, et des travaux complémentaires (détaillés dans la section 7.2) doivent être menés. Ensuite, l'approche initiale de lissage (ainsi que celles que nous avons jusqu'alors présentées) a le défaut de garantir le respect des contraintes de capacité des machines via l'utilisation d'un processus ne pouvant que dégrader continuellement la solution courante. Nous introduisons donc une nouvelle approche dont le but est de permettre à des étapes de process, dont la réalisation est prévue à la période suivante, d'avancer leur réalisation dans la période courante. À la différence des approches par lissage des charges présentées précédemment, cette procédure *complémentaire* permet d'améliorer la solution courante en tirant profit de certaines machines sous-chargées. Une étude expérimentale menée sur des instances industrielles souligne l'intérêt d'inclure cette nouvelle approche dans la procédure de lissage existante, permettant d'améliorer la qualité des solutions fournies selon différents indicateurs. Ces résultats justifient également la poursuite du développement de la procédure, ce dont nous parlons dans la section 7.2.

Enfin, la thèse se déroulant dans un contexte industriel, nous présentons plusieurs réalisations autour de l'intégration de l'approche heuristique au sein de l'entreprise. La première concerne la création d'un environnement favorisant le développement et la validation de nouvelles versions de l'outil d'aide à la décision, grâce notamment au développement d'un autre outil permettant l'exécution de tests automatiques et exhaustifs pour la détection de potentielles régressions Le deuxième grande réalisation concerne le développement de fichiers de

visualisation à partir d'un logiciel de *Business Intelligence* selon différents processus métier, rendant possible pour l'utilisateur de croiser les données de façon dynamique et permettant de tirer profit des solutions très détaillées proposées par l'outil d'aide à la décision.

7.2 Perspectives

Cette thèse a donné lieu a plusieurs réalisations scientifiques et industrielles qui ont été résumées dans la section 7.1. Ces résultats ouvrent la porte à différentes perspectives, que nous présentons dans cette section selon deux catégories : l'amélioration de l'approche heuristique et son utilisation dans l'environnement industriel.

7.2.1 Pistes d'amélioration pour l'approche heuristique

Améliorer l'approche par *Balanced Group*

Dans la section 5.3 du chapitre 5, nous avons montré que l'approche par *Balanced Group* semblait prometteuse car permet d'améliorer les performances du module de lissage par rapport à l'approche par machine. Cependant, cette approche par *Balanced Group* donne parfois lieu à des plans de production ne respectant pas la capacité de certaines machines. Ainsi, il n'est pas garanti que les gains de performance ne soient pas uniquement liés à l'exploitation des excès de capacité. Face à ce problème, plusieurs solutions à relativement court terme sont envisageable afin d'éviter les cas de dépassement de capacité :

— Une première solution pourrait être d'effectuer, après le processus de lissage par *Balanced Group*, un nouvel équilibrage des charges sur les machines (second module de l'approche TSH), puis d'effectuer une passe de lissage des charges *par machine*, afin d'éliminer tous les cas de machines encore surchargées. Cette approche est plus longue que l'approche seulement par *Balanced Group*, car nécessite deux exécutions du module d'*équilibrage* et deux processus de lissage des charges pour chaque période, mais a l'avantage de garantir le respect des contraintes de capacité.

— Une autre possibilité serait de diminuer le seuil de taux de charge maximum à ne pas dépasser. Par exemple, en obligeant de lisser la charge afin de réduire le taux de charge des *Balanced Group* en dessous de 0,95 plutôt que 1. Ces 5% constitueraient une marge permettant de prévenir des petits dépassements. Mais il faudrait alors évaluer si cela n'annulerait pas les gains de productivité initialement gagnés par rapport à l'approche par machine.

— Enfin, une dernière solution serait de dégager des conditions suffisantes permettant de déterminer si les machines d'un même *Balanced Group* ne peuvent plus s'équilibrer du fait de l'évolution des quantités de produit. Ce type de vérification doit cependant être réalisable en un temps très court, contrairement à l'utilisation d'un Programme Linéaire qui permet de vérifier la capabilité des machines à s'équilibrer, mais en un temps de calcul trop long. Un cas simple est celui d'un *Balanced Group* constitué de deux machines m_1 et m_2. Une condition suffisante de non équilibrage est que la somme des charges ne pouvant aller que sur une machine (par exemple m_1), soit supérieure à la somme des charges restantes (ne pouvant aller que sur m_2 ainsi que sur les deux machines). De telles règles sont cependant plus complexes à mettre en oeuvre avec plus de deux machines. Une solution serait peut-être de s'orienter vers une approche de type *machine learning*, dont une phase d'apprentissage sur de nombreux cas d'équilibrage

(et de déséquilibre) pourrait permettre de dégager des règles permettant de déduire les situations où les machines ne peuvent probablement plus s'équilibrer.

Poursuite du développement de l'approche par anticipation

Ensuite, une seconde perspective à court terme concerne la poursuite du développement de l'approche de lissage par anticipation. Le principe de cette approche est de sélectionner des étapes de process prévues à des étapes ultérieures et de les réaliser plus tôt (les anticiper) dans la période courante. Pour des raisons de temps et de simplicité d'implémentation, les étapes de process potentiellement anticipables ne sont que les premières de chaque lot dans la période suivante. Le but était d'évaluer le gain de performance potentiel de l'approche en comparaison du temps supplémentaire qui serait requis. Compte tenu des résultats prometteurs présentés dans la section 5.4 du chapitre 5, le développement d'une version plus étendue de la procédure de lissage par anticipation devrait être envisagé. Mais ce développement amènera son lot de questions. Contrairement au lissage des charges vers l'avant, le lissage par anticipation est obligé de tenir compte de la position de l'étape de process précédent celle décalée afin de respecter les contraintes de temps de process minimums. Toute étape n'est donc pas anticipable, contrairement au cas du lissage vers l'avant où toute étape de process peut être repoussée (avec pour risque "seulement" de livrer le lot concerné en retard au client). De plus, anticiper une étape de process (ce qui signifie donc réduire son temps d'attente par rapport à la fin de l'étape de process précédente) pose la question de jusqu'où le délais d'attente doit être réduit ? Doit-t-on réduire ce temps d'attente au minimum ? Le risque est d'obtenir des plans de production trop agressifs, où beaucoup d'étapes de process sont prévues d'être réalisées en fin de période. Peut-être est-il alors préférable de laisser une marge plus grande entre chaque étape de process, tel que le temps de cycle théorique moyen, mais au risque de fournir des plans de moins bonne qualité ? Ceci est une question parmi d'autres qu'amènerait le développement d'une version plus étendue de la procédure de lissage par anticipation.

Nouvelles règles de lissage

Une autre possibilité d'amélioration concerne la règle de lissage utilisée dans le module de *step-shifting* avec le lissage des charges vers l'avant. Plusieurs règles ont été développées afin de sélectionner quel lot moins prioritaire doit être mis en attente devant une machine jusqu'à une période ultérieure, afin d'optimiser des critères tels que la minimisation des retards de livraison ou bien la maximisation de la productivité de l'usine. Il y a cependant une faiblesse qu'aucune des règles de lissage précédemment présentées permet d'éviter. Elle concerne le cas des machines limitantes (ou goulots) successives. Prenons comme exemple le cas de deux lots, un lot l_1 de produit A, et un lot l_2 de produit B. Le lot l_1 doit passer (dans une période donnée) sur deux machines limitantes consécutives m_1 et m_2 (la machine m_2 étant dédiée au produit A), tandis que l_2 ne doit passer que sur la machine m_1. Par ailleurs, beaucoup de produits A sont attendus par les clients, et le lot l_1 est prioritaire par rapport au lot l_2. Compte tenu de cette situation, toute règle de lissage tenant compte de la date de livraison des lots aura tendance à plutôt décaler le lot l_2 (le faire attendre jusqu'à la période d'après), ce qui aura pour conséquence de décharger la machine m_1. Cependant, étant donné la surcharge de la machine m_2, il est nécessaire de décaler une certaine quantité de produit A, dans notre exemple le lot l_1. Par conséquent, toute règle basée (uniquement ou partiellement) sur les dates de livraison, amènerait au décalage des lots l_1 et l_2, alors que le décalage du lot l_1 seul permettrait de réduire la charge des deux machines m_1 et m_2. Pour éviter ce type de situations, une solution serait de développer une nouvelle règle de lissage, dont la priorité des

lots serait basée sur l'impact de leur potentiel décalage sur la charge des machines limitantes, en favorisant le décalage des lots pour lesquels cela permettrait de réduire fortement la charge des machines surchargées. Dans ce cas, une règle de ce type pourrait privilégier de repousser à la période suivante d'abord le lot l_1, sans nécessité de décaler le second lot. Cependant, ce type de règles aurait pour conséquence probable de décharger fortement l'ensemble des machines, et pas seulement celles surchargées. Une solution pourrait être de considérer une règle composite maximisant la réduction de la charge sur les machines limitantes, tout en minimisant cette même réduction sur les machines sous-chargées. Enfin, ce type de règles risque également de décaler des étapes de process prévues initialement en début de période, car ce décalage amènerait au décalage des étapes de process suivantes (et prévues également dans la période considérée), ce qui aura tendance à engendrer une forte réduction des charges. Cependant, décaler des lots à partir d'étapes de process prévues en début de période risque de grandement influer sur la capacité de livrer ces lots dans les temps. Une solution serait alors de créer une règle composite considérant également les dates de livraison des lots, mais dans ce cas une pondération serait nécessaire entre ces deux termes aux dimensions assez différentes.

Parallélisation du processus de lissage suivant différentes règles

Nous avions montré dans la section 5.2 du chapitre 5 que pour la plupart des critères de performances étudiées, aucune règle de lissage ne fournissait toujours la meilleure solution. Une solution, étant donné la diversité des règles développées, serait d'exécuter la procédure de lissage avec chacune de ces règles et pour chaque période, de garder la solution pour laquelle le critère considéré aurait la meilleure valeur. En l'état, cette approche augmenterait le temps d'exécution du module de *step-shifting* de façon rédhibitoire, mais une solution serait d'exploiter le calcul parallèle, de la même façon que cela est fait pour l'équilibrage des *Isolated Group* dans le module d'*équilibrage*.

Repenser le processus itératif de décalage des lots

Une autre possibilité d'amélioration serait de repenser la procédure de lissage dans son approche de sélection itérative qui est gloutonne (*greedy* en anglais), en intégrant une méthode de décision plus globale. Par exemple, une solution envisageable serait de pré-calculer la charge qui serait retirée de chaque machine suite au décalage d'une étape de process d'un lot. Puis, à l'aide d'une méthode exacte (telle qu'un PLNE), de déterminer le meilleur ensemble d'étapes de process à décaler afin de minimiser un certain objectif. Il faudrait cependant vérifier que la résolution exacte d'un problème à plusieurs milliers de décisions ne dégrade pas trop fortement le temps d'exécution global de l'approche TSH. Une solution serait alors de rechercher une méthode approchée, par exemple via l'utilisation d'une méta-heuristique de recherche locale, afin de déterminer le meilleur ensemble d'étapes de process à décaler.

Gestion des multiples solutions d'allocation des quantités de recettes dans le module d'*équilibrage des charges*

Les perspectives présentées jusqu'ici concernent le module de *step-shifting*, mais des possibilités d'amélioration sont également envisageables pour les autres modules. Concernant le module d'*équilibrage*, nous avons montré que la méthode IMM permet de déterminer une solution *min-max fair* pour notre problème d'équilibrage, et que cette solution est unique.

L'unicité concerne le taux de charge des machines, c'est à dire qu'il n'existe pas deux solutions *min-max fair* où une machine n'aurait pas le même taux de charge dans les deux cas. En revanche, pour un même taux de charge, il existe plusieurs répartitions possibles des différentes recettes à réaliser sur les différentes machines. Cette répartition des quantités n'est pas contrôlée dans l'approche IMM et peut pourtant avoir une influence sur plusieurs aspects. Par exemple, la répartition des recettes sur les machines a une influence sur le module de *step-shifting* car, pour deux solutions avec des répartitions différentes (même avec des taux de charge identiques), le décalage des lots n'impactera pas de la même façon le taux de charge des différentes machines. De plus, étant donné que l'outil de planification permet de définir des objectifs de production (e.g. par quantité de recette par machine et par période) pour les outils d'ordonnancement et de répartition des lots sur les machines, le choix de la répartition des quantités à une influence sur l'optimisation locale. Par exemple, sur des machines fonctionnant par *batch* ou avec des temps de *setup* importants, il serait judicieux que la répartition des quantités minimise la diversité des recettes à exécuter sur chaque machine. Ce fonctionnement est d'ailleurs probablement le comportement par défaut de l'IMM, étant donné l'utilisation d'un programme linéaire en nombres réels dont le parcours sur les sommets du polyèdre des solutions amène généralement à des solutions de type "tout ou rien". À l'inverse, il peut être intéressant de diversifier les types de recettes réalisées sur les machines, par exemple afin de réduire l'impact de possibles pannes sur la production de certains flux de produits. Afin de traiter ce problème de répartition des recettes sur les machines, une solution pourrait être de réaliser dans un premier temps un équilibrage des charges à l'aide de la méthode IMM, puis de fixer le taux de charge de chaque machine, et ensuite d'effectuer une nouvelle répartition des quantités sur les machines. Cette répartition pourrait être faite selon le groupe d'équipements considéré, soit en cherchant à diversifier les recettes réalisées par chaque machine, ou au contraire en faisant en sorte que chacune soit spécialisée sur un petit groupe de recettes. L'ajout de ce deuxième équilibrage des recettes ne ferait qu'ajouter une étape de résolution d'un programme linéaire dans la procédure de l'IMM et ne devrait donc pas augmenter significativement le temps d'exécution du module d'*équilibrage*.

Choix des lots pour le module d'*équilibrage des charges*

Une autre piste d'amélioration concerne l'interaction entre le module de *projection* et celui d'*équilibrage*. Le second module reçoit du premier une prévision des étapes de process devant passer dans la période courante. La séparation entre la période courante est la suivante est binaire : si l'étape de process commence dans la période courante, la charge induite est à considérer dans le module d'équilibrage et cette charge n'est pas à considérer si l'étape commence après la fin de la période courante. Cette décision ne peut être remise en cause par le module d'*équilibrage*, même si l'étape commence 1 minute avant ou après la fin de la période courante. Pourtant, selon le taux de charge des machines, il pourrait être intéressant pour le module d'*équilibrage* de choisir si une étape devrait être considérée dans une période car elle peut être réalisée en respectant la capacité des machines, ou au contraire devrait être réalisée dans la période suivante afin de réduire la charge de certaines machines surchargées dans la période courante. Pour ce faire, une approche serait de considérer un ensemble d'étapes de process à la frontière autour de la période courante et la suivante (de façon analogue à l'ensemble des étapes anticipables présentées dans la section 5.4 du chapitre 5). Le module d'*équilibrage* aurait le choix de réaliser ou non ces étapes afin d'optimiser la répartition des charges sur les machines. Il sera cependant nécessaire de modifier la méthode d'équilibrage

car l'IMM, minimisant successivement la charge des machines, ne ferait que refuser de réaliser les étapes de process même si ceci n'amènerait pas à des cas de machines surchargées.

Perspectives d'amélioration face au mécanisme glouton de l'approche *TSH*

Cette thèse a permis de montrer que l'approche TSH était une solution viable et performante pour répondre au problème de planification de production opérationnelle, car elle fournit rapidement des plans de production détaillés sur des horizons de temps de plusieurs mois. Cette approche a l'avantage de considérer l'ensemble des ateliers de l'usine, contrairement aux outils locaux d'ordonnancement et de répartition des lots qui s'intéressent à un groupe de machines. L'approche TSH possède cependant certaines faiblesses inhérentes à son mécanisme itératif, que ce soit dans le traitement glouton de chaque période, la décomposition de la résolution en trois phases (projection, équilibrage, lissage) ou bien le processus itératif de décalage des lots dans le module de *step-shifting*. Le processus glouton de l'approche TSH amène donc nécessairement à des solutions sous-optimales par rapport à des approches traitant le problème de façon plus globale, par exemple à l'aide d'outils de programmation linéaire. Face à ce problème, deux perspectives nous paraissent intéressantes à étudier :

— Une première idée serait de développer une approche multi-périodes pour différents modules de l'approche TSH. Par exemple, il serait possible d'effectuer un équilibrage des charges sur la période courante mais également sur la période suivante. Ainsi, lors du processus de lissage des charges vers l'avant, il serait possible de choisir de décaler des lots augmentant la charge sur des machines sous-chargées dans la deuxième période, ou bien permettant de réduire la charge sur des machines surchargées à la fois dans les deux périodes considérées. Des considérations similaires pourraient être faites du point de vue du lissage par anticipation, qu'il serait alors cohérent de renommer lissage vers l'arrière.

— Une autre solution serait de compléter l'approche TSH avec une méthode d'optimisation globale comme celle présentée par exemple dans Barhebwa-Mushamuka et al. (2019). L'idée serait de d'abord définir des paramètres globaux de pilotage de l'usine tels que des objectifs de production par famille de produits (ou par produits) par période et par groupe de machines (ou par machines). Puis, l'approche TSH, tenant compte de ces objectifs obtenus via une approche globale, aurait pour tâche de définir un plan de production respectant la capacité et donnant des objectifs détaillés sur l'évolution des lots le long de leur gamme de fabrication. Cette approche suppose cependant que l'approche TSH soit adaptée pour suivre des objectifs de production, ce dont nous allons parler dans le paragraphe suivant.

Alignement du module de *projection* à des objectifs de production

Concernant le module de *projection*, ce dernier ne fait pas l'objet d'un chapitre particulier dans cette thèse, même si celui-ci a fait l'objet de travaux. Le principal travail concerne le fait que le module de *projection* ne possède que deux modes afin de définir la façon dont les lots évoluent dans leur gamme de fabrication. Soit les lots parcourent leur gamme de fabrication selon le temps de cycle théorique calculé grâce aux données historiques, soit la vitesse de progression est ajustée (en adaptant les temps d'attente) en fonction de la date de livraison de chacun des lots. L'approche utilisée par défaut est la seconde, permettant à la projection de tenir compte des engagements envers le client dans la définition du premier plan

de production. Il arrive cependant que les utilisateurs aient besoin de fournir des consignes spécifiques de production, par exemple une quantité à réaliser pour une opération d'un produit donné dans une période donnée, souvent du fait de l'expérience des utilisateurs face à des situations prévues ne pouvant être prises en compte par l'approche TSH. De plus, la conjonction de l'approche TSH avec une approche d'optimisation globale (comme présentée dans le paragraphe précédent) nécessite que l'approche TSH puisse suivre des objectifs de production par période, opération et produit donnés. Pour répondre à ce besoin, nous avons développé une méthode gardant la structure sous forme de modules de l'approche TSH et dont le principe est le suivant :

1. Une première étape consiste à évaluer l'ensemble des étapes de process pouvant être réalisée dans la période courante tout en respectant un temps de cycle considéré comme n'étant pas trop aggressif (i.e. trop court, ceci étant un paramètre pouvant être modifié).
2. Considérant un ensemble d'objectifs de production (par opération et produit), le but est de définir l'ensemble des étapes de process devant être réalisées afin de répondre à ces objectifs. Ce problème d'optimisation peut être assimilé à un problème de *sac à dos en arbre*, ou *Tree Knapsack Problem* (TKP, voir Cho and Shaw (1997) ou Bertsimas and Demir (2002)), où chaque étape de process est un objet et chaque objectif (opération/produit) est un sac à dos avec une capacité donnée.
3. Une fois que sont définies les étapes de process à réaliser dans la période, le module de *projection* est exécuté en adaptant la vitesse de progression de chaque lot afin que seules les étapes souhaitées soient réalisées.

Une première version de ce nouveau module de *projection* a été implémentée et a montré des performances prometteuses dans le suivi des objectifs de production. Le temps de calcul est cependant un point critique car les premiers tests montraient que l'approche TSH, avec le nouveau module de *projection*, nécessitait un temps d'exécution plus de deux fois supérieur à l'approche avec le module de *projection* initial. Une solution a été de réduire la taille du problème en obligeant certaines étapes de process (celles prévues les plus tôt dans la période considérée) à être réalisées dans la période courante, étant donnée que celles-ci seront probablement réalisées de toute façon. Des premiers tests ont montré que cette réduction de la taille du problème permet à l'approche TSH d'être exécutée en des temps acceptables, mais des analyses complémentaires doivent être menées afin d'évaluer l'influence de ce paramètre de fixation sur la capacité du module de *projection* à respecter les objectifs de production.

Utiliser la simulation afin d'évaluer les performances de l'approche TSH

Dans cette thèse, nous avons à plusieurs reprises comparé entre elles différentes versions de l'approche TSH, selon différents critères de performance. Si l'on prend par exemple le critère du nombre de lots livrés en retard, une version A de l'approche TSH était considérée plus performante qu'une version B si le plan de production fourni par la version A présentait moins de lots livrés en retard que dans le plan de production fourni par la version B de l'approche. Le critère de qualité porte donc sur la solution fournie par l'approche, et non pas sur l'impact de ces plans de production sur le pilotage de l'usine. Une perspective très intéressante serait d'évaluer l'impact de ces plans directeurs sur les performances du système de production. Dans l'idéal, il faudrait pouvoir dupliquer l'usine à un instant donné et réaliser un pilotage selon différents plans de production générés selon différentes versions de l'approche TSH. La version la plus performante de l'approche TSH serait celle dont le

pilotage amène à la meilleure performance de l'usine. Une telle duplication de l'usine n'est pas possible, mais l'utilisation d'un outil de simulation comme dans Horiguchi et al. (2001) ou Barhebwa-Mushamuka et al. (2019) pourrait permettre de comparer différentes versions de l'approche TSH selon leur capacité à effectivement bien piloter les flux de production dans l'usine.

L'utilisation d'un outil de simulation permettrait également de comparer un pilotage de l'usine à l'aide de consignes définies par la fonction d'*Operational Production Planning* avec un fonctionnement des outils d'ordonnancement de chaque atelier sans pilotage global. Une telle étude permettrait d'évaluer le gain d'intégrer une fonction intermédiaire entre le *Production Planning* et le *Production Scheduling*, comme cela est par exemple présenté dans Horiguchi et al. (2001).

7.2.2 Perspectives d'utilisation de l'outil d'aide à la décision

Exploiter les plans de production à des fins de prévisions

Dans cette thèse, nous avons introduit l'approche TSH afin de traiter le problème de planification de production opérationnelle et l'intégrant dans un outil d'aide à la décision. Le plan de production détaillé fourni par l'outil est utilisé comme un plan directeur, un chemin à suivre afin de piloter au mieux les flux de produits dans l'usine afin d'optimiser différents indicateurs de performance. Mais des travaux sont également en cours afin d'utiliser les plans fournis par l'approche TSH comme des prévisions de l'évolution des flux de produits. Le paradigme est donc très différent, au lieu de définir le chemin qui *doit* être suivi, les solutions fournies par l'outil indiquent le chemin qui *sera* suivi si la production se déroule selon les paramètres de l'approche TSH utilisés. On parle généralement d'utilisation en mode "*what if*". Par exemple, si l'on utilise l'approche TSH sans considérer les dates de livraison des lots dans le module de *projection* et que le module de *step-shifting* utilise la règle de lissage uniquement basée sur le taux de charge des machines, alors le plan de production fourni est une prévision de l'évolution des flux de produits dans le cas où le pilotage de l'usine a pour seul but de maximiser le productivité de cette dernière. Plusieurs outils utilisés en salle se servent actuellement des résultats de l'approche TSH afin de prévoir l'évolution des lots dans l'usine, par exemple afin de placer au mieux les maintenances préventives des machines. Une telle utilisation nécessite d'évaluer la qualité des prévisions effectuées, et plus précisément d'évaluer la version (ou plutôt le paramétrage) de l'approche TSH permettant de prédire au mieux les grandeurs suivies. Car suite aux différents développement autour de l'approche TSH et de l'outil d'aide à la décision lié, ce dernier est devenu très modulable. Ainsi, il est possible de choisir le type de projection à effectuer (alignée aux dates de livraisons ou simplement au temps de cycle théorique), si l'on souhaite considérer ou non les contraintes de capacité, la règle de lissage utilisée, l'utilisation ou non du lissage par anticipation, du lissage par *balanced broups*, la taille des périodes de planification, etc. De plus, quel indicateur souhaitons-nous prévoir ? Les quantités de plaques réalisées dans la journée ? Le nombre de lots sortant de l'usine durant la semaine ? Les machines limitantes demain ? Selon les grandeurs (ainsi que l'agrégation et l'échelle de temps) considérées, certaines versions de l'approche seront plus performantes que d'autre dans la prévision de leur évolution. Une analyse formelle est donc nécessaire afin de déterminer pour chaque application le paramétrage de l'approche TSH requis afin de répondre au mieux aux besoins de prévisions. Dans un second temps, une perspective pourrait être de combiner différentes versions de l'approche TSH sous forme d'un "super modèle", où chaque version possède un

poids particulier dans la réalisation de la prévision, ces poids pouvant être déterminés grâce à l'utilisation d'approches de *machine learning* et en tirant profit des grandes quantités de données historiques disponibles. Ce principe de combiner plusieurs modèles de prévision afin de créer un nouveau modèle plus performant est par exemple utilisé dans le domaine des prévisions météorologiques (voir par exemple Chaves et al. (2005) ou Flato et al. (2014)).

Extension de l'approche *TSH* à d'autres fonctions de planification de production

Une dernière perspective de développement concerne l'extension de l'approche TSH à d'autres fonctions de la planification de production.

Tout d'abord, le plan de lancement des lots en production est une donnée d'entrée pour le problème d'*Operational Production Planning* traité par l'approche TSH, ce plan de lancement étant déterminé par la fonction *Production Planning* (voir figure 6.6 dans les chapitres 1 et 6). Or, ce plan de lancement pourrait être une donnée de sortie de l'approche TSH. Une méthode simple est de considérer le plan de livraison pour chaque période (semaine), et de déterminer les quantités pouvant être livrées grâce aux en-cours (le WIP). Pour les livraisons ne pouvant être complétées par les lots actuellement en cours de production, il suffit alors de déterminer la quantité de lots à lancer en production avec pour date de lancement, la date de livraison moins le temps de cycle théorique nécessaire pour que le produit passe sur l'ensemble de sa gamme de fabrication (le *lead time*). Cette fonction est en réalité déjà implémentée dans l'outil d'aide à la décision, mais n'est actuellement pas utilisée au quotidien. Cette méthode de génération du plan de lancement suit le même principe que les méthodes à *lead time* fixe présentées dans la section 2.5.1 du chapitre 2, mais d'autres approches plus sophistiquées pourraient être envisagées, par exemple en tenants compte de l'influence entre temps de cycle et le niveau de charge de l'usine.

Enfin, une autre fonction que pourrait remplir l'approche TSH, est celle de *Capacity Planning*. Pour rappel, l'objectif consiste à estimer le nombre d'équipements nécessaires pour répondre à une demande donnée ou, de façon symétrique, la quantité de produits qu'un ensemble donné d'équipements peut fabriquer tout en maintenant un rendement acceptable (par exemple un temps de cycle compétitif). Dans la fabrication de semi-conducteurs, il s'agit généralement d'une décision à moyen/long terme sur un horizon d'un à trois ans en raison du long délai d'obtention de nouveaux équipements (section 2.4.1 du chapitre 2). Étant donné cet horizon de planification particulièrement long, les données sont agrégées avec des raisonnements par atelier (voire usine) et grandes familles de produits. Ces agrégations sont possibles dans l'approche TSH, et étant donné sa grande rapidité, son extension à des horizons de planification de plusieurs années est possible tout en gardant un temps d'exécution acceptable dans le cadre de l'utilisation prévue. Nous avions d'ailleurs présenté les qualités de la nouvelle méthode d'équilibrage (chapitre 4) pour aider à l'analyse des machines critiques. L'utilisation de l'approche TSH dans le cadre du *Capacity Planning* est donc une perspective prometteuse, et un projet est en cours au sein de STMicroelectronics afin d'évaluer la possibilité d'étendre l'outil d'aide à la décision à l'ensemble des sites *front end* de l'entreprise.

Table des figures

Liste des tableaux

Liste des Algorithmes

Bibliographie

Akçalı, E., Üngör, A. and Uzsoy, R. (2005). Short-term capacity allocation problem with tool and setup constraints, *Naval Research Logistics (NRL)* **52**(8): 754–764. 33

Albey, E., Bilge, Ü. and Uzsoy, R. (2014). An exploratory study of disaggregated clearing functions for production systems with multiple products, *International Journal of Production Research* **52**(18): 5301–5322. 34, 35, 37

Albey, E., Bilge, Ü. and Uzsoy, R. (2017). Multi-dimensional clearing functions for aggregate capacity modelling in multi-stage production systems, *International Journal of Production Research* **55**(14): 4164–4179. 37

Albey, E. and Uzsoy, R. (2015). Lead time modeling in production planning, *2015 Winter Simulation Conference (WSC)*, IEEE, pp. 1996–2007. 33

Allahverdi, A. (2015). The third comprehensive survey on scheduling problems with setup times/costs, *European Journal of Operational Research* **246**(2): 345–378. 40

Allahverdi, A., Ng, C., Cheng, T. E. and Kovalyov, M. Y. (2008). A survey of scheduling problems with setup times or costs, *European journal of operational research* **187**(3): 985–1032. 40

Ankenman, B. E., Bekki, J. M., Fowler, J., Mackulak, G. T., Nelson, B. L. and Yang, F. (2011). Simulation in production planning: an overview with emphasis on recent developments in cycle time estimation, *Planning Production and Inventories in the Extended Enterprise*, Springer, pp. 565–591. 34

Anthony, R. N. (1965). *Planning and Control Systems: A Framework for Analysis [by]*, Division of Research, Graduate School of Business Administration, Harvard University. 11, 24

Arrow, K. J., Karlin, S., Scarf, H. E. et al. (1958). Studies in the mathematical theory of inventory and production. 26

Asmundsson, J., Rardin, R. L., Turkseven, C. H. and Uzsoy, R. (2009). Production planning with resources subject to congestion, *Naval Research Logistics (NRL)* **56**(2): 142–157. 37

Atherton, R. and Dayhoff, J. (1986). Signature analysis: Simulation of inventory, cycle time, and throughput trade-offs in wafer fabrication, *IEEE transactions on components, hybrids, and manufacturing technology* **9**(4): 498–507. 29

Baker, K. R. (1974). *Introduction to sequencing and scheduling*, John Wiley & Sons. 94, 95

Bang, J.-Y. and Kim, Y.-D. (2010). Hierarchical production planning for semiconductor wafer fabrication based on linear programming and discrete-event simulation, *IEEE Transactions on Automation Science and Engineering* **7**(2): 326–336. 34

Bard, J. F., Deng, Y., Chacon, R. and Stuber, J. (2010). Midterm planning to minimize deviations from daily target outputs in semiconductor manufacturing, *IEEE transactions on semiconductor manufacturing* **23**(3): 456–467. 44, 47, 49

Barhebwa-Mushamuka, F., Dauzère-Pérès, S. and Yugma, C. (2019). Work-in-progress balancing control in global fab scheduling for semiconductor manufacturing, *2019 Winter Simulation Conference (WSC)*, IEEE. 43, 44, 45, 49, 155, 157

Behringer, F. A. (1981). A simplex based algorithm for the lexicographically extended linear maxmin problem, *European Journal of Operational Research* **7**(3): 274–283. 83

Benders, J. F. (1962). Partitioning procedures for solving mixed-variables programming problems, *Numerische mathematik* **4**(1): 238–252. 44

Beraudy, S., Absi, N. and Dauzère-Pérès, S. (2018). Production planning models with productivity and financial objective functions in semiconductor manufacturing, *Proceedings of the 2018 Winter Simulation Conference*, IEEE Press, pp. 3397–3407. 34

Bermon, S., Feigin, G. and Hood, S. (1995). Capacity analysis of complex manufacturing facilities, *Proceedings of 1995 34th IEEE Conference on Decision and Control*, Vol. 2, IEEE, pp. 1935–1940. 33

Bermon, S. and Hood, S. J. (1999). Capacity optimization planning system (caps), *Interfaces* **29**(5): 31–50. 124

Bertsekas, D. P., Gallager, R. G. and Humblet, P. (1987). *Data networks*, Vol. 2, Prentice-hall Englewood Cliffs, NJ. 77, 80, 83

Bertsimas, D. and Demir, R. (2002). An approximate dynamic programming approach to multidimensional knapsack problems, *Management Science* **48**(4): 550–565. 156

Bettayeb, B. (2012). *Conception et évaluation des plans de surveillance basés sur le risque. Limitation des incertitudes qualité avec des ressources limitées de maîtrise*, PhD thesis, Grenoble. 5, vii

Billington, P. J., McClain, J. O. and Thomas, L. J. (1983). Mathematical programming approaches to capacity-constrained mrp systems: Review, formulation and problem reduction, *Management Science* **29**(10): 1126–1141. 15, 33

Błażewicz, J., Ecker, K. H., Pesch, E., Schmidt, G. and Weglarz, J. (2007). *Handbook on scheduling: from theory to applications*, Springer Science & Business Media. 39

Bowman, E. H. (1959). The schedule-sequencing problem, *Operations Research* **7**(5): 621–624. 39

Brahimi, N., Dauzère-Pérès, S. and Najid, N. M. (2006). Capacitated multi-item lot-sizing problems with time windows, *Operations Research* **54**(5): 951–967. 66

Brucker, P. (2007). *Scheduling Algorithms*, Springer. 39

Brucker, P., Gladky, A., Hoogeveen, H., Kovalyov, M. Y., Potts, C. N., Tautenhahn, T. and Van De Velde, S. L. (1998). Scheduling a batching machine, *Journal of scheduling* **1**(1): 31–54. 40

Bureau, M., Dauzère-Pérès, S. and Mati, Y. (2006). Scheduling challenges and approaches in semiconductor manufacturing, *IFAC Proceedings Volumes* **39**(3): 739–744. 40

Bureau, M., Dauzère-Pérès, S., Yugma, C., Vermariën, L. and Maria, J.-B. (2007). Simulation results and formalism for global-local scheduling in semiconductor manufacturing facilities, *2007 Winter Simulation Conference*, IEEE, pp. 1768–1773. 43, 44, 45, 49

Burman, D. Y., Gurrola-Gal, F. J., Nozari, A., Sathaye, S. and Sitarik, J. P. (1986). Performance analysis techniques for ic manufacturing lines, *AT&T technical journal* **65**(4): 46–57. 29

Byrne, M. and Bakir, M. A. (1999). Production planning using a hybrid simulation–analytical approach, *International Journal of Production Economics* **59**(1-3): 305–311. 34

Byrne, M. and Hossain, M. (2005). Production planning: An improved hybrid approach, *International Journal of Production Economics* **93**: 225–229. 34

Cakanyildirim, M. and Roundy, R. (1999). Demand forecasting and demand planning in the semiconductor industry. 11, 30

Carlson, J. G. and Yao, A. C. (1992). Mixed model assembly simulation, *International Journal of Production Economics* **26**(1-3): 161–167. 16

Chaves, R. R., Ross, R. S. and Krishnamurti, T. (2005). Weather and seasonal climate prediction for south america using a multi-model superensemble, *International Journal of Climatology: A Journal of the Royal Meteorological Society* **25**(14): 1881–1914. 158

Cho, G. and Shaw, D. X. (1997). A depth-first dynamic programming algorithm for the tree knapsack problem, *INFORMS Journal on Computing* **9**(4): 431–438. 156

Christ, Q., Dauzère-Pérès, S. and Lepelletier, G. (2019). An iterated min-max procedure for practical workload balancing on non-identical parallel machines in manufacturing systems, *European Journal of Operational Research* **279**(2): 419–428. 66, 72

Chung, J. and Jang, J. (2009). A wip balancing procedure for throughput maximization in semiconductor fabrication, *IEEE Transactions on semiconductor manufacturing* **22**(3): 381–390. 44, 49

Cook, W., Lovász, L., Seymour, P. D. et al. (1995). *Combinatorial optimization: papers from the DIMACS Special Year*, Vol. 20, American Mathematical Soc. 84

Copil, K., Wörbelauer, M., Meyr, H. and Tempelmeier, H. (2017). Simultaneous lotsizing and scheduling problems: a classification and review of models, *OR spectrum* **39**(1): 1–64. 42

Curry, G. L. and Feldman, R. M. (2010). *Manufacturing systems modeling and analysis*, Springer Science & Business Media. 34

Dauzère-Pérès, S. and Lasserre, J.-B. (2002). On the importance of sequencing decisions in production planning and scheduling, *International transactions in operational research* **9**(6): 779–793. 13, 41, 92

Dayhoff, J. and Atherton, R. (1984a). Signature analysis of dispatch systems in wafer fabrication: Human factors and analytical comparison, *IEEE Transactions* . 29

Dayhoff, J. E. and Atherton, R. W. (1984b). Simulation of vlsi manufacturing areas, *VLSI Design* **4**(84-92): 41. 29

Dequeant, K. (2017). *Workflow variability modeling in microelectronic manufacturing*, Theses, Université Grenoble Alpes.
URL: *https://hal.archives-ouvertes.fr/tel-01652884* 14

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V. et al. (2014). Evaluation of climate models, *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, pp. 741–866. 158

Fordyce, K., Dalton, D., Gerard, B., Jesse, R. R., Sell, R. and Sullivan, G. G. (1992). Daily output planning: Integrating operations research, artificial intelligence, and real-time decision support with apl2, *Expert Systems with Applications* **5**(3-4): 245–256. 43, 49

Fordyce, K., Fournier, J., Milne, R. J. and Singh, H. (2012). Tutorial: illusion of capacity-challenge of incorporating the complexity of fab capacity (tool deployment & operating curve) into central planning for firms with substantial non-fab complexity, *Proceedings of the Winter Simulation Conference*, Winter Simulation Conference, p. 206. 12

Fox, M. S. (1983). Constraint-directed search: A case study of job-shop scheduling., *Technical report*, CARNEGIE-MELLON UNIV PITTSBURGH PA ROBOTICS INST. 46

Framinan, J. M., Leisten, R. and García, R. R. (2014). Multi-objective scheduling, *Manufacturing Scheduling Systems*, Springer, pp. 261–288. 39

Freund, R. M. and Mizuno, S. (2000). Interior point methods: current status and future directions, *High performance optimization*, Springer, pp. 441–466. 84

Gantt, H. L. (1919). *Organizing for work*, Harcourt, Brace and Howe. 26

Gao, J., Gen, M. and Sun, L. (2006). Scheduling jobs and maintenances in flexible job shop with a hybrid genetic algorithm, *Journal of Intelligent Manufacturing* **17**(4): 493–507. 40

Garey, M. R. (1979). A guide to the theory of np-completeness, *Computers and intractability* . 39

Geng, N. and Jiang, Z. (2009). A review on strategic capacity planning for the semiconductor manufacturing industry, *International journal of production research* **47**(13): 3639–3655. 30

Ghasemi, A., Heavey, C. and Laipple, G. (2018). A review of simulation-optimization methods with applications to semiconductor operational problems, *Proceedings of the 2018 Winter Simulation Conference*, IEEE Press, pp. 3672–3683. 35

Glassey, C. R. and Resende, M. G. (1988). Closed-loop job release control for vlsi circuit manufacturing, *IEEE Transactions on Semiconductor manufacturing* **1**(1): 36–46. 29

Goldratt, E. M. (1990). *Theory of constraints*, North River Croton-on-Hudson. 16, 27

Golhar, D. Y. and Stamm, C. L. (1991). The just-in-time philosophy: a literature review, *The International Journal of Production Research* **29**(4): 657–676. 16

González, M. A., Vela, C. R., González-Rodríguez, I. and Varela, R. (2013). Lateness minimization with tabu search for job shop scheduling problem with sequence dependent setup times, *Journal of Intelligent Manufacturing* **24**(4): 741–754. 40

Gopalswamy, K. and Uzsoy, R. (2017). An iterative refinement approach to fitting clearing functions to data from simulation models of production systems, *2017 Winter Simulation Conference (WSC)*, IEEE, pp. 3254–3265. 38

Gopalswamy, K. and Uzsoy, R. (2018). An exploratory comparison of clearing function and data-driven production planning models, *Proceedings of the 2018 Winter Simulation Conference*, IEEE Press, pp. 3482–3493. 38

Govind, N., Bullock, E. W., He, L., Iyer, B., Krishna, M. and Lockwood, C. S. (2008). Operations management in automated semiconductor manufacturing with integrated targeting, near real-time scheduling, and dispatching, *IEEE Transactions on Semiconductor Manufacturing* **21**(3): 363–370. 31

Graves, S. C. (1986). A tactical planning model for a job shop, *Operations Research* **34**(4): 522–533. 36

Guhlich, H., Fleischmann, M., Mönch, L. and Stolletz, R. (2018). A clearing function based bid-price approach to integrated order acceptance and release decisions, *European Journal of Operational Research* **268**(1): 243–254. 37

Habenicht, K. and Mönch, L. (2002). A finite-capacity beam-search-algorithm for production scheduling in semiconductor manufacturing, *Proceedings of the Winter Simulation Conference*, Vol. 2, IEEE, pp. 1406–1413. 46, 47, 48, 49

Habla, C., Mönch, L. and Driebel, R. (2007). A finite capacity production planning approach for semiconductor manufacturing, *2007 IEEE International Conference on Automation Science and Engineering*, IEEE, pp. 82–87. 46, 47, 48, 49

Hackman, S. T. (2007). *Production economics: integrating the microeconomic and engineering perspectives*, Springer Science & Business Media. 33

Hackman, S. T. and Leachman, R. C. (1989). A general framework for modeling production, *Management Science* **35**(4): 478–495. 33

Harris, F. W. (1913). How many parts to make at once. 26

Held, M. and Karp, R. M. (1970). The traveling-salesman problem and minimum spanning trees, *Operations Research* **18**(6): 1138–1162. 46

Held, M. and Karp, R. M. (1971). The traveling-salesman problem and minimum spanning trees: Part ii, *Mathematical programming* **1**(1): 6–25. 46

Hopp, W. J. and Spearman, M. L. (2011). *Factory physics*, Waveland Press. 15, 37

Horiguchi, K., Raghavan, N., Uzsoy, R. and Venkateswaran, S. (2001). Finite-capacity production planning algorithms for a semiconductor wafer fabrication facility, *International Journal of Production Research* **39**(5): 825–842. 45, 47, 48, 49, 105, 133, 157

Hung, Y.-F. and Cheng, G.-J. (2002). Hybrid capacity modeling for alternative machine types in linear programming production planning, *Iie Transactions* **34**(2): 157–165. 33

Hung, Y.-F. and Hou, M.-C. (2001). A production planning approach based on iterations of linear programming optimization and flow time prediction, *Journal of the Chinese Institute of Industrial Engineers* **18**(3): 55–67. 34, 35

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. 34, 35

Ignizio, J. P. (2009). Cycle time reduction via machine-to-operation qualification, *International Journal of Production Research* **47**(24): 6899–6906. 33

Irdem, D. F., Kacar, N. B. and Uzsoy, R. (2010). An exploratory analysis of two iterative linear programming—simulation approaches for production planning, *IEEE Transactions on Semiconductor Manufacturing* **23**(3): 442–455. 35

Johnson, L., Montgomery, D. and Montgomery, D. (1974). *Operations Research in Production Planning, Scheduling, and Inventory Control*, Wiley.
URL: *https://books.google.fr/books?id=EOFTAAAAMAAJ* 33

Johnson, S. M. (1954). Optimal two-and three-stage production schedules with setup times included, *Naval research logistics quarterly* **1**(1): 61–68. 26

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, 72

Johri, P. K. (1993). Practical issues in scheduling and dispatching in semiconductor wafer fabrication, *Journal of Manufacturing Systems* **12**(6): 474–485. 29

Johri, P. K. (1994). Overlapping machine groups in semiconductor wafer fabrication, *European journal of operational research* **74**(3): 509–518. 33

Kacar, N. B., Irdem, D. F. and Uzsoy, R. (2012). An experimental comparison of production planning using clearing functions and iterative linear programming-simulation algorithms, *IEEE Transactions on Semiconductor Manufacturing* **25**(1): 104–117. 37

Kacar, N. B., Mönch, L. and Uzsoy, R. (2013). Planning wafer starts using nonlinear clearing functions: A large-scale experiment, *IEEE Transactions on Semiconductor Manufacturing* **26**(4): 602–612. 37

Kacar, N. B., Mönch, L. and Uzsoy, R. (2016). Modeling cycle times in production planning models for wafer fabrication, *IEEE Transactions on Semiconductor Manufacturing* **29**(2): 153–167. 37, 38

Kacar, N. B., Mönch, L. and Uzsoy, R. (2018). Problem reduction approaches for production planning using clearing functions, *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, IEEE, pp. 931–938. 37, 46, 47

Kacar, N. B. and Uzsoy, R. (2014). A comparison of multiple linear regression approaches for fitting clearing functions to empirical data, *International Journal of Production Research* **52**(11): 3164–3184. 37

Kacar, N. B. and Uzsoy, R. (2015). Estimating clearing functions for production resources using simulation optimization, *IEEE Transactions on Automation Science and Engineering* **12**(2): 539–552. 37

Kalir, A. A. and Rozen, K. (2018). Optimizing starts for capacity, velocity, and output during the ramp-up period of a semiconductor fab, *2018 Winter Simulation Conference (WSC)*, IEEE, pp. 3494–3501. 37

Kallrath, J. and Maindl, T. I. (2006). Planning in semiconductor manufacturing, *Real Optimization with SAP® APO* pp. 105–118. 16

Karabuk, S. and Wu, S. D. (2003). Coordinating strategic capacity planning in the semiconductor industry, *Operations Research* **51**(6): 839–849. 29

Karmarkar, U. S. (1989). Capacity loading and release planning with work-in-progress (wip) and leadtimes, *Journal of Manufacturing and Operations Management* **2**(105-123). 37

Katerattanakul, P., Hong, S. and Lee, J. (2006). Enterprise resource planning survey of korean manufacturing firms, *Management Research News* **29**(12): 820–837. 27

Kim, B. and Kim, S. (2001). Extended model for a hybrid production planning approach, *International Journal of Production Economics* **73**(2): 165–173. 34, 35

Kim, S. H., Kim, J. W. and Lee, Y. H. (2014). Simulation-based optimal production planning model using dynamic lead time estimation, *The International Journal of Advanced Manufacturing Technology* **75**(9-12): 1381–1391. 34

Kim, S. H. and Lee, Y. H. (2016). Synchronized production planning and scheduling in semiconductor fabrication, *Computers & Industrial Engineering* **96**: 72–85. 35, vii

King, J. H. (1989). Allocation of scarce resources in manufacturing facilities, *Bell Labs Technical Journal* **68**(3): 103–113. 78

Klemmt, A. and Mönch, L. (2012). Scheduling jobs with time constraints between consecutive process steps in semiconductor manufacturing, *Proceedings of the 2012 Winter Simulation Conference (WSC)*, IEEE, pp. 1–10. 40

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. 28, 40

Korte, B., Vygen, J., Korte, B. and Vygen, J. (2012). *Combinatorial optimization*, Vol. 2, Springer. 57

Krafcik, J. F. (1988). Triumph of the lean production system, *MIT Sloan Management Review* **30**(1): 41. 27

Kriett, P. O., Eirich, S. and Grunow, M. (2017). Cycle time-oriented mid-term production planning for semiconductor wafer fabrication, *International Journal of Production Research* **55**(16): 4662–4679. 33

Leachman, R. C. (1993). Modeling techniques for automated production planning in the semiconductor industry, *Optimization in industry* **1**: 1. 38

Leachman, R. C., Benson, R. F., Liu, C. and Raar, D. J. (1996). Impress: An automated production-planning and delivery-quotation system at harris corporation—semiconductor sector, *Interfaces* **26**(1): 6–37. 124

Leachman, R. C. and Carmon, T. F. (1992). On capacity modeling for production planning with alternative machine types, *IIE transactions* **24**(4): 62–72. 33

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. 43, 49

Lee, B., Lee, Y. H., Yang, T. and Ignisio, J. (2008). A due-date based production control policy using wip balance for implementation in semiconductor fabrications, *International Journal of Production Research* **46**(20): 5515–5529. 43, 49

Lee, T.-E. (2008). A review of scheduling theory and methods for semiconductor manufacturing cluster tools, *2008 Winter Simulation Conference*, IEEE, pp. 2127–2135. 41

Lee, T. N. and Plenert, G. (1993). Optimizing theory of constraints when new product alternatives exist, *Production and inventory management journal* **34**(3): 51. 17

Lee, Y. H. and Lee, B. (2003). Push-pull production planning of the re-entrant process, *The International Journal of Advanced Manufacturing Technology* **22**(11-12): 922–931. 42, 43, 49

Lee, Y. H., Park, J. and Kim, S. (2002). Experimental study on input and bottleneck scheduling for a semiconductor fabrication line, *IIE transactions* **34**(2): 179–190. 43

Li, M., Yang, F., Uzsoy, R. and Xu, J. (2016). A metamodel-based monte carlo simulation approach for responsive production planning of manufacturing systems, *Journal of Manufacturing Systems* **38**: 114–133. 35, 38

Liberopoulos, G. (2002). Production capacity modeling of alternative, nonidentical, flexible machines, *International Journal of Flexible Manufacturing Systems* **14**(4): 345–359. 33

Lima, A., Borodin, V., Dauzere-Peres, S. and Vialletetelle, P. (2019). Sampling-based release control of multiple lots in time constraint tunnels, *Computers in Industry* **110**: 3–11. 40

Lin, C.-H., Hwang, S.-L. and Wang, M.-Y. E. (2006). The mythical advanced planning systems in complex manufacturing environment, *IFAC Proceedings Volumes* **39**(3): 703–708. 16, 27, 28

Liu, S. and Papageorgiou, L. G. (2013). Multiobjective optimisation of production, distribution and capacity planning of global supply chains in the process industry, *Omega* **41**(2): 369–382. 78

Liu, S. and Papageorgiou, L. G. (2018). Fair profit distribution in multi-echelon supply chains via transfer prices, *Omega* **80**: 77–94. 78

Lu, Z., Zhang, Y. and Han, X. (2013). Integrating run-based preventive maintenance into the capacitated lot sizing problem with reliability constraint, *International Journal of Production Research* **51**(5): 1379–1391. 66

Luss, H. (1999). On equitable resource allocation problems: A lexicographic minimax approach, *Operations Research* **47**(3): 361–378. 78

Luss, H. (2012). *Equitable Resource Allocation: Models, Algorithms and Applications*, Vol. 101, John Wiley & Sons. 83, 84

Luss, H. and Smith, D. R. (1986). Resource allocation among competing activities: A lexicographic minimax approach, *Operations Research Letters* **5**(5): 227–231. 78, 83

Ma, Y., Chu, C. and Zuo, C. (2010). A survey of scheduling with deterministic machine availability constraints, *Computers & Industrial Engineering* **58**(2): 199–211. 40

Mabert, V. A., Soni, A. and Venkataramanan, M. (2000). Enterprise resource planning survey of us manufacturing firms, *Production and Inventory Management Journal* **41**(2): 52. 27

Manda, A. B. and Uzsoy, R. (2018). Simulation optimization for planning product transitions in semiconductor manufacturing facilities, *2018 Winter Simulation Conference (WSC)*, IEEE, pp. 3470–3481. 34

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. 28, 31, 39

Mathirajan, M. and Sivakumar, A. I. (2006). A literature review, classification and simple meta-analysis on scheduling of batch processors in semiconductor, *The International Journal of Advanced Manufacturing Technology* **29**(9-10): 990–1001. 40

Mati, Y. (2010). Minimizing the makespan in the non-preemptive job-shop scheduling with limited machine availability, *Computers & Industrial Engineering* **59**(4): 537–543. 40

Maxwell, W. L. and Miller, L. W. (1967). *Theory of scheduling*, Reading, Mass.: Addison-Wesley Publishing Company. 26

Maynard, H. B., Stegemerten, G. J. and Schwab, J. L. (1948). Methods-time measurement. 26

Meyr, H., Wagner, M. and Rohde, J. (2015). Structure of advanced planning systems, *Supply chain management and advanced planning*, Springer, pp. 99–106. 25, 30, vii

Mhiri, E. (2016). *Capacity planning in the context of high mix, application in the semiconductor industry*, Theses, Université Grenoble Alpes.
URL: *https://tel.archives-ouvertes.fr/tel-01485148* 9, 32, 46, 48, 56, 57, 59, 61, 72, 149

Mhiri, E., Mangione, F., Jacomino, M., Vialletelle, P. and Lepelletier, G. (2018). Heuristic algorithm for a wip projection problem at finite capacity in semiconductor manufacturing, *IEEE Transactions on Semiconductor Manufacturing* **31**(1): 62–75. 46, 47, 48, 49, 64, 66, 69, 76, 77, 92, 95, 100, 121, 149

Miller, D. J. (1989). Implementing the results of a manufacturing simulation in a semiconductor line, *Proceedings of the 21st conference on Winter simulation*, ACM, pp. 922–929. 29

Milne, R. J., Mahapatra, S. and Wang, C.-T. (2015). Optimizing planned lead times for enhancing performance of mrp systems, *International Journal of Production Economics* **167**: 220–231. 33, 37

Missbauer, H. and Uzsoy, R. (2011). Optimization models of production planning problems, *Planning production and inventories in the extended enterprise*, Springer, pp. 437–507. 36, 37

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. 29, 39, 40, 41

Mönch, L., Fowler, J. W. and Mason, S. J. (2012). *Production planning and control for semiconductor wafer fabrication facilities: modeling, analysis, and systems*, Vol. 52, Springer Science & Business Media. 5, 6, 11, 15, 16, 27, 28, 29, 32, 38

Mönch, L., Stehli, M. and Zimmermann, J. (2003). Fabmas: An agent-based system for production control of semiconductor manufacturing processes, *International Conference on Industrial Applications of Holonic and Multi-Agent Systems*, Springer, pp. 258–267. 124

Mönch, L., Stehli, M., Zimmermann, J. and Habenicht, I. (2006). The fabmas multi-agent-system prototype for production control of water fabs: design, implementation and performance assessment, *Production Planning & Control* **17**(7): 701–716. 124

Mönch, L., Uzsoy, R. and Fowler, J. W. (2018a). A survey of semiconductor supply chain models part i: semiconductor supply chains, strategic network design, and supply chain simulation, *International Journal of Production Research* **56**(13): 4524–4545. 9, 10, 11, 25, 29, 45, vii

Mönch, L., Uzsoy, R. and Fowler, J. W. (2018b). A survey of semiconductor supply chain models part iii: master planning, production planning, and demand fulfilment, *International Journal of Production Research* **56**(13): 4565–4584. 29, 30, 32, 33, 34, 36, vii

Murça, M. C. R. (2018). Collaborative air traffic flow management: Incorporating airline preferences in rerouting decisions, *Journal of Air Transport Management* **71**: 97–107. 77

Nace, D. and Orlin, J. B. (2007). Lexicographically minimum and maximum load linear programming problems, *Operations research* **55**(1): 182–187. 83, 84, 90

Nace, D. and Pióro, M. (2008). Max-min fairness and its applications to routing and load-balancing in communication networks: a tutorial, *IEEE Communications Surveys & Tutorials* **10**(4). 77, 83

Naderi, B., Ghomi, S. F. and Aminnayeri, M. (2010). A high performing metaheuristic for job shop scheduling with sequence-dependent setup times, *Applied Soft Computing* **10**(3): 703–710. 40

Ogryczak, W., Luss, H., Pióro, M., Nace, D. and Tomaszewski, A. (2014). Fair optimization and networks: A survey, *Journal of Applied Mathematics* **2014**. 77

Olhager, J. (2013). Evolution of operations planning and control: from production to supply chains, *International journal of production research* **51**(23-24): 6836–6843. 26

Olhager, J. and Selldin, E. (2003). Enterprise resource planning survey of swedish manufacturing firms, *European Journal of Operational Research* **146**(2): 365–373. 27

Omar, R. S. M., Venkatadri, U., Diallo, C. and Mrishih, S. (2017). A data-driven approach to multi-product production network planning, *International Journal of Production Research* **55**(23): 7110–7134. 38

Orcun, S., Uzsoy, R. and Kempf, K. (2006). Using system dynamics simulations to compare capacity models for production planning, *Proceedings of the 38th conference on Winter simulation*, Winter Simulation Conference, pp. 1855–1862. 36

Orlicki, J. A. (1975). *Material requirements planning: the new way of life in production and inventory management*, McGraw-Hill. 15, 26

Ovacik, I. M. and Uzsoy, R. (2012). *Decomposition methods for complex factory scheduling problems*, Springer Science & Business Media. 39, 40

Ovacik, I. and Uzsoy, R. (1997). Decomposition methods for complex factory scheduling problems. 28

Pahl, J., Voß, S. and Woodruff, D. L. (2005a). Load dependent lead times—from empirical evidence to mathematical modeling, *Research methodologies in supply chain management*, Springer, pp. 539–554. 34

Pahl, J., Voß, S. and Woodruff, D. L. (2005b). Production planning with load dependent lead times, *4OR* **3**(4): 257–302. 34, 37

Pahl, J., Voß, S. and Woodruff, D. L. (2007). Production planning with load dependent lead times: an update of research, *Annals of Operations Research* **153**(1): 297–345. 34

Pfund, M. E., Mason, S. J. and Fowler, J. W. (2006). Semiconductor manufacturing scheduling and dispatching, *Handbook of Production Scheduling*, Springer, pp. 213–241. 14, 92

Pinedo, M. (2016). Scheduling: Theory, algorithms, and systems, fifth, *Cham: Springer International Publishing* . 39, 41

Pochet, Y. and Wolsey, L. A. (2006). *Production planning by mixed integer programming*, Springer Science & Business Media. 33

Potts, C. N. and Kovalyov, M. Y. (2000). Scheduling with batching: A review, *European journal of operational research* **120**(2): 228–249. 40

Qi, J. (2016). Mitigating delays and unfairness in appointment systems, *Management Science* **63**(2): 566–583. 77

Quadt, D. and Kuhn, H. (2005). Conceptual framework for lot-sizing and scheduling of flexible flow lines, *International Journal of Production Research* **43**(11): 2291–2308. 42

Quadt, D. and Kuhn, H. (2009). Capacitated lot-sizing and scheduling with parallel machines, back-orders, and setup carry-over, *Naval Research Logistics (NRL)* **56**(4): 366–384. 42

Quirk, M. and Serda, J. (2001). *Semiconductor manufacturing technology*, Vol. 1, Prentice Hall Upper Saddle River, NJ. 4

Radunovic, B. and Le Boudec, J.-Y. (2007). A unified framework for max-min and min-max fairness with applications, *IEEE/ACM Transactions on networking* **15**(5): 1073–1083. 77, 78, 80, 83

Rippenhagen, C. and Krishnaswamy, S. (1998). Implementing the theory of constraints philosophy in highly reentrant systems, *1998 Winter Simulation Conference. Proceedings (Cat. No. 98CH36274)*, Vol. 2, IEEE, pp. 993–996. 16

Rowshannahad, M., Dauzere-Peres, S. and Cassini, B. (2015). Capacitated qualification management in semiconductor manufacturing, *Omega* **54**: 50–59. 72

Roy, B. and Sussmann, B. (1964). Les problemes d'ordonnancement avec contraintes disjonctives, *Note ds* **9**. 39

Sadeghi, M., Björnson, E., Larsson, E. G., Yuen, C. and Marzetta, T. L. (2018). Max–min fair transmit precoding for multi-group multicasting in massive mimo, *IEEE Transactions on Wireless Communications* **17**(2): 1358–1373. 77

Schmidt, G. (2000). Scheduling with limited machine availability, *European Journal of Operational Research* **121**(1): 1–15. 40

Shen, L. (2014). A tabu search algorithm for the job shop problem with sequence dependent setup times, *Computers & Industrial Engineering* **78**: 95–106. 40

Shilov, A. (2018). Tsmc starts to build fab 18: 5 nm, volume production in early 2020. **URL:** *https://www.anandtech.com/show/12377/tsmc-starts-to-build-fab-18-5nm-in-early-2020* 5

SIA (2015). Global billings report history (3-month moving average). **URL:** *http://www.semiconductors.org/industry_statistics/global_sales_report/* 4, vii

Silver, E., Pyke, D. and Peterson, R. (1998). *Inventory Management and Production Planning and Scheduling*, third edn, John Wiley and Sons, New York. **URL:** *http://books.google.com.tw/books?id=GI5jQgAACAAJ* 11

Smith, S. B. (1965). Planning transistor production by linear programming, *Operations Research* **13**(1): 132–139. 33

Spence, A. and Welter, D. (1987). Capacity planning of a photolithography work cell in a wafer manufacturing line, *Proceedings. 1987 IEEE International Conference on Robotics and Automation*, Vol. 4, IEEE, pp. 702–708. 29

Srinivasan, A., Carey, M., Morton, T. E. et al. (1988). Resource pricing and aggregate scheduling in manufacturing systems, *Technical report*, Carnegie Mellon University, Tepper School of Business. 37

Stadtler, H. and Kilger, C. (2015). Supply chain management and advanced planning, *Concepts, Models, Software and Case Studies* **4**. 11, 24, 25

Sugimori, Y., Kusunoki, K., Cho, F. and UCHIKAWA, S. (1977). Toyota production system and kanban system materialization of just-in-time and respect-for-human system, *The International Journal of Production Research* **15**(6): 553–564. 27

Taal, M. and Wortmann, J. C. (1997). Integrating mrp and finite capacity planning, *Production Planning & Control* **8**(3): 245–254. 15

Tamssaouet, K., Dauzère-Pérès, S. and Yugma, C. (2018). Metaheuristics for the job-shop scheduling problem with machine availability constraints, *Computers & Industrial Engineering* **125**: 1–8. 40

Tang, C. S. (1988). A max-min allocation problem: its solutions and applications, *Operations Research* **36**(2): 359–367. 78, 83

Taylor, F. W. (1911). The principles of scientific management, *New York* **202**. 26

Toktay, L. B. and Uzsoy, R. (1994). *Capacity Allocation with Integer Side Constraints: An Application in Semiconductor Manufacturing*, School of Industrial Engineering, Purdue University. 33

Trigeiro, W. W., Thomas, L. J. and McClain, J. O. (1989). Capacitated lot sizing with setup times, *Management Science* **35**(3): 353–366. 66

Uzsoy, R., Fowler, J. W. and Mönch, L. (2018). A survey of semiconductor supply chain models part iii: demand planning, inventory management, and capacity planning, *International Journal of Production Research* **56**(13): 4546–4564. 29, 30

Uzsoy, R., Lee, C.-Y. and Martin-Vega, L. A. (1992). A review of production planning and scheduling models in the semiconductor industry part i: system characteristics, performance evaluation and production planning, *IIE transactions* **24**(4): 47–60. 5, 29, 32

Uzsoy, R., Lee, C.-Y. and Martin-Vega, L. A. (1994). A review of production planning and scheduling models in the semiconductor industry part ii: Shop-floor control, *IIE transactions* **26**(5): 44–55. 29

Vollmann, T. E. (2005). *Manufacturing planning and control for supply chain management*, McGraw-Hill. 11

Voß, S. and Woodruff, D. L. (2006). *Introduction to computational optimization models for production planning in a supply chain*, Vol. 240, Springer Science & Business Media. 33

Wang, J., Yang, J., Zhang, J., Wang, X. and Zhang, W. (2018). Big data driven cycle time parallel prediction for production planning in wafer manufacturing, *Enterprise information systems* **12**(6): 714–732. 38

Wang, L., Fang, L. and Hipel, K. W. (2008). Basin-wide cooperative water resources allocation, *European Journal of Operational Research* **190**(3): 798–817. 77

Wein, L. M. (1988). Scheduling semiconductor wafer fabrication, *IEEE Transactions on semiconductor manufacturing* **1**(3): 115–130. 29

Wight, O. (1995). *Manufacturing resource planning: MRP II: unlocking America's productivity potential*, John Wiley & Sons. 15, 27

Wilson, R. (1934). *A scientific routine for stock control*, Harvard Univ. 26

Womack, J. P., Womack, J. P., Jones, D. T. and Roos, D. (1990). *Machine that changed the world*, Simon and Schuster. 27

Wu, S. D., Erkoc, M. and Karabuk, S. (2005). Managing capacity in the high-tech industry: A review of literature, *The engineering economist* **50**(2): 125–158. 30

Wylie, L. (1990). A vision of next generation mrp ii, *Gartner Group* **40**. 27

Xiao, J., Yang, H., Zhang, C., Zheng, L. and Gupta, J. N. (2015). A hybrid lagrangian-simulated annealing-based heuristic for the parallel-machine capacitated lot-sizing and scheduling problem with sequence-dependent setup times, *Computers & Operations Research* **63**: 72–82. 42

Xiao, J., Zhang, C., Zheng, L. and Gupta, J. N. (2013). Mip-based fix-and-optimise algorithms for the parallel machine capacitated lot-sizing and scheduling problem, *International Journal of Production Research* **51**(16): 5011–5028. 42

Yaakob, N. and Khalil, I. (2016). A novel congestion avoidance technique for simultaneous real-time medical data transmission, *IEEE journal of biomedical and health informatics* **20**(2): 669–681. 77

Yugma, C., Blue, J., Dauzère-Pérès, S. and Obeid, A. (2015). Integration of scheduling and advanced process control in semiconductor manufacturing: review and outlook, *Journal of Scheduling* **18**(2): 195–205. 39

Zhu, X., Jiang, C., Yin, L., Kuang, L., Ge, N. and Lu, J. (2018). Cooperative multigroup multicast transmission in integrated terrestrial-satellite networks, *IEEE Journal on Selected Areas in Communications* . 77

Ziarnetzky, T., Kacar, N. B., Mönch, L. and Uzsoy, R. (2015). Simulation-based performance assessment of production planning formulations for semiconductor wafer fabrication, *2015 Winter Simulation Conference (WSC)*, IEEE, pp. 2884–2895. 37

Ziarnetzky, T., Mönch, L. and Uzsoy, R. (2018). Rolling horizon, multi-product production planning with chance constraints and forecast evolution for wafer fabs, *International Journal of Production Research* **56**(18): 6112–6134. 37

Zoghby, J., Barnes, J. W. and Hasenbein, J. J. (2005). Modeling the reentrant job shop scheduling problem with setups for metaheuristic searches, *European Journal of Operational Research* **167**(2): 336–348. 39

Annexe A

Tableaux détaillés des résultats comparatifs des performances des différentes règles de lissage de la charge

TABLE A.1 – Comparaison des règles de lissages par rapport au retard total

Rules / Instances	EDD	OD	OD_{Post}	RS	RS_{Post}	MI	MI_{OD}	MI_{RS}	RSMP
1	18	133	19	151	18	30	19	19	62
2	0	439	91	157	0	60	37	34	201
3	37	53	41	59	37	57	41	37	266
4	33	73	51	144	36	73	45	42	329
5	6	134	30	215	6	119	29	7	47
6	174	892	28	793	7	98	29	13	520
7	119	596	62	414	2	79	60	5	263
8	48	1001	249	791	63	321	258	67	361
9	73	1020	265	814	76	400	265	81	430
10	2086	2279	2122	3444	1353	2355	1859	1752	2307
11	1607	2489	2309	2346	1154	2586	2170	1838	2254
12	1330	2630	2169	1982	1028	2352	2078	1461	2436
13	1493	2104	2073	2191	884	2427	1878	1553	2563
14	2528	3188	2981	2842	1625	2768	2885	1759	1720
15	1254	3054	2618	3017	878	1586	2550	1540	1671
16	1610	3353	2894	2979	1061	2693	2897	2140	1237
17	1637	2827	2567	2504	1280	2734	2416	2132	1340
18	802	2460	1964	2505	628	1522	1362	1335	1370
19	1004	2331	2351	2531	700	1662	2075	2453	1397
20	973	1737	1383	2182	881	1730	1114	1349	2059
21	907	1289	1301	1188	1105	1162	931	1025	2202
22	1106	761	780	700	1363	1186	862	823	2476
23	690	683	679	1169	482	700	774	813	1961
24	678	708	702	1297	496	697	723	763	1988
25	507	658	658	996	439	743	784	631	2711
Avg	829	1476	1216	1496	**624**	1206	1126	947	1367
Max	2528	3353	2981	3444	**1625**	2768	2897	2453	2711
Min	0	53	19	59	**0**	30	19	5	47

TABLE A.2 – Comparaison des règles de lissages par rapport au retard maximum

Instances \ Rules	EDD	OD	OD_{Post}	RS	RS_{Post}	MI	MI_{OD}	MI_{RS}	RSMP
1	2	6	2	4	2	3	2	2	4
2	0	4	1	3	0	3	1	1	4
3	3	3	3	3	3	3	3	3	6
4	1	3	3	4	1	5	2	2	6
5	3	5	5	4	3	7	5	3	5
6	5	7	5	5	5	5	5	5	9
7	2	6	4	4	1	6	4	1	5
8	2	7	5	4	2	7	5	4	6
9	3	7	5	4	2	7	5	3	6
10	20	17	17	17	9	20	17	10	24
11	16	17	17	16	7	17	17	16	18
12	12	17	17	17	9	18	17	14	16
13	14	18	15	18	9	18	15	14	16
14	15	17	15	15	8	20	14	12	15
15	12	17	17	16	8	19	17	13	12
16	17	17	16	17	8	19	15	14	
17	22	18	18	18	8	21	15	15	17
18	15	14	15	14	14	20	13	13	23
19	14	11	11	9	8	14	10	10	15
20	11	11	11	11	7	11	11	11	14
21	11	12	12	10	7	9	11	11	12
22	11	13	13	11	8	12	12	12	11
23	10	12	12	10	8	11	11	12	11
24	10	12	12	10	8	11	12	12	10
25	13	13	14	12	6	14	14	14	13
Avg	9,76	11,36	10,60	10,24	**6,04**	12,00	10,12	9,08	11,52
Max	22	18	18	18	**14**	21	17	16	24
Min	0	3	1	3	**0**	3	1	1	4

TABLE A.3 – Comparaison des règles de lissages par rapport au nombre de lots en retard

Rules / Instances	EDD	OD	OD_{Post}	RS	RS_{Post}	MI	MI_{OD}	MI_{RS}	RSMP
1	17	73	18	85	17	23	18	18	55
2	0	241	91	113	0	47	37	34	139
3	35	51	39	56	35	49	39	35	144
4	33	56	45	93	36	56	42	39	210
5	2	50	15	102	2	62	14	3	41
6	169	252	12	378	2	46	13	6	363
7	82	168	39	219	2	46	37	5	180
8	47	242	197	363	62	155	209	58	238
9	71	244	202	343	75	166	202	78	291
10	698	471	478	812	888	797	493	597	526
11	712	532	490	743	829	658	576	492	634
12	587	541	474	744	698	724	539	420	687
13	508	425	398	701	491	806	420	380	433
14	955	649	605	838	1030	937	616	672	678
15	450	523	518	763	577	593	549	402	538
16	605	598	564	866	679	821	623	557	613
17	718	738	682	957	788	991	718	657	768
18	388	550	518	860	431	589	503	519	532
19	603	545	588	828	571	729	675	775	488
20	537	332	333	505	575	618	311	407	733
21	482	283	351	349	720	522	311	380	769
22	630	214	259	257	897	557	337	405	767
23	299	186	205	519	256	355	272	301	608
24	259	189	214	567	260	384	259	304	684
25	208	183	198	347	222	334	291	270	865
Avg	363,8	333,44	**301,32**	496,32	405,72	442,6	324,16	312,56	479
Max	955	738	**682**	957	1030	991	718	775	865
Min	**0**	50	12	56	**0**	23	13	3	41

TABLE A.4 – Comparaison des règles de lissages par rapport au temps de cycle moyen

Instances \ Rules	EDD	OD	OD_{Post}	RS	RS_{Post}	MI	MI_{OD}	MI_{RS}	RSMP
1	33,68	34,01	33,69	34,04	33,68	33,7	33,68	33,68	32,82
2	32,95	34,13	33,18	33,27	32,95	33,06	33,04	33,03	32,18
3	33,29	33,33	33,3	33,35	33,29	33,33	33,3	33,29	31,76
4	31,94	32,04	31,97	32,23	31,94	32,01	31,95	31,95	31,00
5	30,54	30,9	30,6	31,16	30,54	30,77	30,6	30,54	30,95
6	30,2	32,45	30,13	32,21	30,03	30,22	30,13	30,06	31,46
7	30,67	31,98	30,48	31,5	30,31	30,47	30,48	30,32	30,93
8	29,47	32,09	29,75	31,38	29,44	30,07	29,74	29,48	31,49
9	29,29	31,91	29,58	31,25	29,25	30,07	29,58	29,26	31,28
10	51,16	51,76	51,36	54,71	49,11	51,46	50,7	50,31	53,24
11	47,57	50,4	49,99	49,9	46,41	50,22	49,54	48,68	54,16
12	48,73	52,02	50,85	50,41	47,92	50,88	50,52	49,02	56,82
13	48,53	50,2	50,14	50,32	47,09	50,91	49,62	48,8	52,78
14	53,39	56,52	55,78	55,12	51,26	54,78	55,41	51,77	55,07
15	46,89	51,72	50,63	51,56	45,87	47,76	50,51	47,69	56,59
16	51,4	55,63	54,55	54,67	50,14	54,19	54,6	52,71	56,02
17	49,1	52,22	51,58	51,42	48,15	51,84	51,26	50,44	52,95
18	49,3	53,29	51,94	53,17	47,98	49,89	50,49	50,33	46,49
19	50,46	54,19	54,26	53,91	49,19	51,6	53,26	54,06	46,23
20	50,12	53,74	53,52	53,57	49,21	52,13	54,7	51,63	48,52
21	52,71	53,94	53,14	53,36	51,16	52,63	53,52	51,28	49,63
22	53,9	55,56	54,31	54,44	51,85	53,43	54,64	52,07	50,82
23	50,37	54,61	54,09	54,62	49,57	53,22	54,51	52,4	50,08
24	50,65	55,25	54,67	55,35	49,9	53,53	54,76	52,86	50,08
25	48,27	52,3	51,17	51,53	48,09	51,13	50,84	49,56	49,75
Avg	43,38	45,85	44,99	45,54	**42,57**	44,53	44,86	43,81	44,52
Max	53,90	56,52	55,78	55,35	**51,85**	54,78	55,41	54,06	56,82
Min	29,29	30,90	29,58	31,16	**29,25**	30,07	29,58	29,26	30,93

TABLE A.5 – Comparaison des règles de lissages par rapport à la productivité (*throughput*)

Rules / Instances	EDD	OD	OD_{Post}	RS	RS_{Post}	MI	MI_{OD}	MI_{RS}	RSMP
1	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
4	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
8	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10	0,00%	1,90%	1,90%	1,60%	1,20%	2,00%	2,00%	1,40%	0,3%
11	0,00%	2,00%	2,00%	1,70%	1,10%	2,40%	2,20%	1,60%	-0,6%
12	0,00%	1,80%	1,80%	1,50%	0,90%	2,00%	1,70%	1,60%	-0,2%
13	0,00%	1,60%	1,60%	1,10%	1,00%	2,00%	1,80%	1,50%	-0,6%
14	0,00%	2,80%	2,80%	2,60%	1,50%	2,90%	2,80%	2,50%	1,9%
15	0,00%	1,20%	1,30%	0,70%	0,70%	1,60%	1,40%	1,30%	1,2%
16	0,00%	1,50%	1,50%	0,80%	1,00%	1,50%	1,50%	1,50%	1,4%
17	0,00%	0,80%	1,00%	0,60%	0,60%	1,10%	0,90%	1,00%	2,2%
18	0,00%	1,30%	1,60%	0,40%	1,90%	2,10%	1,80%	1,80%	-0,3%
19	0,00%	1,10%	1,20%	1,50%	1,40%	1,90%	1,60%	1,40%	1,1%
20	0,00%	1,00%	1,20%	0,40%	1,00%	1,40%	1,10%	1,10%	-0,7%
21	0,00%	0,90%	0,90%	0,40%	1,10%	1,40%	1,40%	1,20%	1,0%
22	0,00%	0,40%	0,50%	0,40%	0,80%	1,00%	0,90%	0,70%	-0,5%
23	0,00%	0,50%	0,50%	0,50%	0,60%	0,70%	0,60%	0,60%	0,8%
24	0,00%	0,60%	0,60%	0,50%	0,60%	0,80%	0,70%	0,70%	0,6%
25	0,00%	0,10%	0,10%	-0,10%	0,20%	0,40%	0,40%	0,40%	-1,8%
Avg	0,00%	0,78%	0,82%	0,58%	0,62%	**1,01**%	0,91%	0,81%	0,232%
Max	0,00%	2,80%	2,80%	2,60%	1,90%	**2,90**%	2,80%	2,50%	2,25%
Min	0,00%	0,00%	0,00%	-0,10%	0,00%	0,00%	0,00%	0,00%	-1,85%

TABLE A.6 – Comparaison des règles de lissages par rapport à la charge globale (*Workload*)

Instances \ Rules	EDD	OD	OD_{Post}	RS	RS_{Post}	MI	MI_{OD}	MI_{RS}	RSMP
1	0,00%	2,86%	2,86%	2,86%	2,86%	2,86%	2,86%	2,86%	2,1%
2	0,00%	3,03%	3,03%	3,03%	3,03%	3,03%	3,03%	3,03%	3,03%
3	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	-1,1%
4	0,00%	0,00%	0,00%	0,00%	0,00%	3,33%	0,00%	3,33%	0,8%
5	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	-1,5%
6	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,3%
7	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	-1,2%
8	0,00%	-3,45%	0,00%	-3,45%	0,00%	0,00%	0,00%	0,00%	0,1%
9	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10	0,00%	3,13%	3,13%	3,13%	3,13%	3,13%	3,13%	3,13%	0,5%
11	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	-0,6%
12	0,00%	0,00%	3,03%	0,00%	0,00%	3,03%	3,03%	3,03%	-1,5%
13	0,00%	0,00%	2,94%	2,94%	0,00%	2,94%	2,94%	2,94%	-0,5%
14	0,00%	3,03%	3,03%	3,03%	3,03%	3,03%	3,03%	3,03%	2,6%
15	0,00%	3,03%	3,03%	0,00%	3,03%	3,03%	3,03%	3,03%	2,1%
16	0,00%	3,03%	3,03%	3,03%	3,03%	3,03%	3,03%	3,03%	2,0%
17	0,00%	3,13%	3,13%	0,00%	0,00%	3,13%	3,13%	0,00%	3,3%
18	0,00%	2,56%	2,56%	2,56%	2,56%	5,13%	2,56%	2,56%	-1,1%
19	0,00%	0,00%	0,00%	0,00%	0,00%	2,50%	2,50%	0,00%	0,8%
20	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	-1,5%
21	0,00%	0,00%	0,00%	0,00%	0,00%	2,86%	2,86%	2,86%	0,3%
22	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	-1,2%
23	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,70%	2,70%	0,1%
24	0,00%	0,00%	0,00%	0,00%	2,70%	2,70%	2,70%	0,00%	0,00%
25	0,00%	0,00%	0,00%	-2,78%	0,00%	0,00%	0,00%	0,00%	-2,6%
Avg	0,00%	0,81%	1,19%	0,57%	0,93%	**1,75**%	1,62%	1,42%	0,21%
Max	0,00%	3,13%	3,13%	3,13%	3,13%	**5,13**%	3,13%	3,33%	3,32%
Min	0,00%	-3,45%	0,00%	-3,45%	0,00%	0,00%	0,00%	0,00%	-2,59%

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2020LYSEM002

Quentin CHRIST

OPTIMIZATION AND DECISION SUPPORT FOR OPERATIONNAL PRODUCTION PLANNING IN SEMICONDUCTOR MANUFACTURING

Speciality: Industrial Engineering

Keywords: Production Planning, Scheduling, Semiconductor Manufacturing, Heuristic

Abstract:

In today's connected societies, computers, sensors, data centers, automotive electronics and portable devices have become ubiquitous. At the heart of these products are semiconductors, whose current production systems are among the most complex in the world. In this thesis, we address the problem of operational production planning, whose objective is to link tactical production planning, which defines the delivery and launch plans for products in the plant, to detailed production scheduling. Due to the complexity of production systems in semiconductor manufacturing, this essential frontier problem remains under-researched. After having formalized the studied problem and shown that it is NP-Difficile, an experimental study is carried out in order to show the impossibility of solving the problem in an exact way for industrial instances, justifying the use of approached methods. We present a three-step heuristic initially in place in the production system and highlight some of its limitations. We introduce new methods aimed at improving the quality of the solutions provided by the decision support tool according to the users' needs. As this thesis was carried out in a company, the objective was both to make progress on the scientific aspects, but also to think about the integration of these methods in the industrial system.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2020LYSEM002

Quentin CHRIST

OPTIMISATION ET AIDE A LA DECISION POUR LA PLANIFICATION DE PRODUCTION OPERATIONNELLE EN FABRICATION DE SEMICONDUCTEURS

Spécialité: Génie Industriel

Mots clefs : Planification de Production, Ordonnancement, Fabrication de Semiconducteurs, Heuristique

Résumé:

Dans les sociétés connectées d'aujourd'hui, ordinateurs, capteurs, centres de données, électronique automobile et dispositifs portables sont devenus omniprésents. Au cœur de ces produits, se trouvent les semi-conducteurs, dont les systèmes de production actuels font partie des plus complexes au monde. Dans cette thèse, nous traitons du problème de planification de production opérationnelle, dont l'objectif est de faire le lien entre la planification de production tactique, définissant les plans de livraisons et de lancements des produits dans l'usine, et l'ordonnancement détaillé de la production. Du fait de la complexité des systèmes de production en fabrication de semi-conducteurs, considérer ce problème frontière est indispensable, mais a fait l'objet de peu de travaux de recherches. Après avoir formalisé le problème étudié et montré que celui-ci est NP-Difficile, une étude expérimentale est menée afin de montrer l'impossibilité de résoudre le problème de manière exacte pour des instances industrielles, justifiant l'utilisation de méthodes approchées. Nous présentons une heuristique en trois étapes initialement en place dans le système de production et en soulignons certaines limites. Nous introduisons de nouvelles méthodes ayant pour but d'améliorer la qualité des solutions fournies par l'outil d'aide à la décision en fonction des besoins des utilisateurs. Cette thèse ayant été réalisée en entreprise, l'objectif fût à la fois d'avancer sur les aspects scientifiques, mais également de réfléchir à l'intégration de ces méthodes dans le système industriel.