Planification de la production à capacité finie dans un contexte à forte variabilité, application à l'industrie des semi-conducteurs

Emna Mhiri

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **Automatique Productique**

Arrêté ministériel : 25 mai 2016

Présentée par

Emna MHIRI

Thèse dirigée par **Mireille JACOMINO** et
coencadrée par **Fabien MANGIONE**

préparée au sein du **Laboratoire G-SCOP**
dans **l'École Doctorale EEATS**

Planification de la production à capacité finie dans un contexte à forte variabilité, application à l'industrie des semi-conducteurs

Thèse soutenue publiquement le **13 décembre 2016**,
devant le jury composé de :

Mme. Olga BATTAÏA
Professeur, Institut supérieur de l'aéronautique et de l'espace de Toulouse, Rapporteur
Mme. Nathalie SAUER
Professeur, Université de Lorraine, Rapporteur
M. André ROSSI
Professeur, Université d'Angers, Examinateur
M. Lyes BENYOUCEF
Professeur, Université d'Aix Marseille, Président
M. Philippe VIALLETELLE
Ingénieur, STMicroelectronics de Crolles, Examinateur
Mme. Mireille JACOMINO
Professeur, Grenoble INP, Directrice de thèse
M. Fabien MANGIONE
Maître de conférences, Grenoble INP, Encadrant de thèse
M. Guillaume LEPELLETIER
Ingénieur, STMicroelectronics de Crolles, Invité

Emna Mhiri : *Planification de la production à capacité finie dans un contexte à forte variabilité, application à l'industrie des semi-conducteurs.* Thèse de doctorat, 13 décembre 2016

Remerciements

Cette thèse est l'aboutissement de plusieurs années de travail et n'aurait jamais pu voir le jour sans le soutien et l'aide de nombreuses personnes qui m'ont accompagnée tout au long de cette épopée ! Je tiens donc ici à les remercier et leur témoigner de ma reconnaissance.

Mes premiers remerciements vont tout naturellement à mes encadrants qui ont suivi de près ces travaux. J'exprime mes plus profonds remerciements à ma directrice de thèse, Mireille Jacomino, pour l'intérêt qu'elle a porté à ce travail, sa disponibilité malgré les fonctions qui lui incombent, sa gentillesse et ses directives si précieuses. Je remercie également mon co-encadrant de thèse Fabien Mangione pour sa patience, ses nombreux conseils, ses idées, ses nombreuses relectures, et ses remarques qui m'ont permis d'affiner mon propos. Un grand merci pour cet encadrement de qualité intellectuellement et humainement.

Mes sincères remerciements vont également aux membres de jury qui ont accepté d'évaluer ce travail de thèse : Lyes Benyoucef, qui m'a fait l'honneur de présider le jury ; Olga Battaïa et Nathalie Sauer qui ont accepté d'être rapporteurs de ce mémoire, merci pour l'intérêt qu'elles ont porté à cette thèse. Merci à André Rossi qui a accepté d'examiner ma thèse. Merci pour ses remarques judicieuses et les améliorations qu'il a suggérées.

Cette thèse est le fruit d'une collaboration étroite avec l'entreprise STMicroelectonics, précisément le site de Crolles. Je tiens tout particulièrement à adresser mes remerciements à Philippe Vialletelle et Guillaume Lepelletier, des ingénieurs à STMicroelectronics de Crolles, pour leur disponibilité, leur suivi régulier et la richesse de nos échanges, sans lesquels je n'aurais pu collecter les données suffisantes qui ont enrichi ce travail de thèse et acquérir une vision synthétique du système de production à étudier.

Je remercie aussi tous les enseignants et les chercheurs du laboratoire G-SCOP. Ces trois années à G-SCOP m'ont donné l'occasion de passer de l'autre côté du miroir en travaillant avec ceux qui étaient auparavant mes enseignants. Je remercie particulièrement

Yannick Frein, Khaled Hadj-Hamou, Michel Tollenaere, Pierre David, Hadrien Cambazard, Maria Di Mascolo, Gülgün Alpan, Lilia Gzara et tous les autres avec qui j'ai eu l'occasion de travailler dans le cadre de mes missions d'enseignement à Grenoble INP génie industriel.

Je souhaite aussi saluer l'ensemble des ingénieurs, stagiaires et doctorants du laboratoire G-SCOP, sans qui ces quelques années auraient paru bien plus mornes. Je pense spécialement à Widad, Khadija, Wafa, Amine, Ahmed, Hassan, Hussein, Khalil, Kléber, Rami. Je remercie également mes co-bureaux Anis et Asma. Merci pour cette ambiance exceptionnelle au laboratoire et pour les bons moments passés ensemble. Merci aussi pour leur soutien et leur aide précieuse. Merci à tous qui sont passés de collègues à vrais amis.

Enfin, je remercie mes amis et ma famille qui ont largement contribué à l'aboutissement de ce projet de thèse. Merci à Afef, Maroua, Fatma d'avoir partagé aussi bien les bons que les durs moments de la thèse. Merci à mes chers parents, qui me témoignent depuis toujours une confiance sans faille. Sans leur enthousiasme, leurs encouragements infaillibles et leur soutien indéfectible, tout ceci n'aurait pas pu être possible. Merci aussi à mes deux petits frères et ma chère grand-mère.

J'adresse une pensée particulière à Riadh. Avec sa patience et son soutien, j'ai pu mener à bien la phase finale de ces travaux avec confiance et sérénité.

Une page se tourne, une autre s'ouvre, merci à tous ceux et celles qui y ont contribué et m'ont amené à faire les bons choix.

À ma famille.
À mes amis.
À mes rencontres de la vie, d'aujourd'hui et d'hier...
À tous ceux qui m'aiment !
À tous ceux que j'aime.
À la vie qui me porte.

Table des matières

Table des figures **ix**

Liste des tableaux **xiii**

Introduction générale **1**

1 Contexte industriel et problématique **5**

1.1 Introduction 7

1.2 Processus de fabrication des circuits intégrés 8

1.2.1 Description générique du processus de fabrication 8

1.2.2 Entités de base de la fabrication des semi-conducteurs 10

1.2.3 Classification du système de production des semi-conducteurs . . . 11

1.2.4 Complexité de la production dans le secteur de la microélectronique 12

1.3 Planification de la production dans l'industrie des semi-conducteurs 17

1.3.1 Niveaux de la planification de la production 18

1.3.2 Objectifs de la planification de la production 19

1.3.3 Caractéristiques des problèmes de planification de la production dans l'industrie des semi-conducteurs 20

1.3.3.1 Contraintes 20

1.3.3.2 Indicateurs de performance 21

1.3.4 Techniques de planification de la production dans l'industrie des semi-conducteurs 22

1.3.4.1 Les techniques classiques 22

1.3.4.2 Les modèles analytiques 25

1.3.4.3 Les modèles de simulation 26

1.3.4.4 Les techniques heuristiques 28

1.3.4.5 Les techniques d'intelligence artificielle 30
1.3.4.6 Autres classifications des techniques de planification . . . 31
1.3.5 Systèmes de planification de la production 31
1.4 Problématique 34
1.5 Conclusion 35

2 Planification de la fabrication des semi-conducteurs : État de l'art 37
2.1 Introduction 39
2.2 Méthodologie et classification des problèmes 39
2.3 Techniques de planification existantes et leurs limites 45
2.3.1 Planification stratégique 45
2.3.2 Planification tactique et opérationnelle 46
2.3.2.1 Techniques basées sur l'optimisation 46
2.3.2.2 Techniques basées sur la simulation 50
2.3.2.3 Techniques basées sur la théorie des files d'attente 50
2.4 Positionnement de notre problématique 51
2.5 Conclusion 52

3 Résolution analytique du problème de planification à capacité finie 55
3.1 Introduction 57
3.2 Description du problème 57
3.2.1 Contraintes 58
3.2.2 Hypothèses 59
3.2.3 Enjeux 60
3.2.4 Questions de recherche 61
3.3 Notations 61
3.4 Formulation mathématique du problème : MIP 62
3.5 Complexité 64
3.6 Résolution du problème 65
3.7 Méthodes de résolution alternatives 66
3.7.1 Procédure d'agrégation 67
3.7.2 Heuristique de décomposition 67
3.7.3 Relaxation lagrangienne 69
3.7.3.1 Principe de la relaxation lagrangienne 69
3.7.3.2 Application de la relaxation lagrangienne à notre problème 71
3.8 Résultats expérimentaux 75
3.9 Conclusion 79

4 Résolution approchée du problème de planification à capacité finie 83
4.1 Introduction . . . 85
4.2 Projection du *WIP* à capacité infinie . . . 85
4.2.1 Modèle du temps de cycle . . . 86
4.2.1.1 Modèle du temps de cycle développé . . . 88
4.2.2 Principe de la projection du *WIP* à capacité infinie . . . 93
4.3 Heuristique de décomposition à base de *MIP* . . . 94
4.4 Heuristique de décomposition à base d'algorithmes . . . 99
4.4.1 Calcul de la charge accumulée . . . 100
4.4.2 Equilibrage de la charge et la capacité . . . 102
4.5 Résultats et discussion . . . 108
4.5.1 Génération des instances aléatoires . . . 108
4.5.2 Evaluation des algorithmes heuristiques proposés en comparaison à une solution optimale . . . 109
4.5.2.1 Comparaison entre la solution optimale et l'heuristique à base de MIP . . . 109
4.5.2.2 Comparaison entre la solution optimale et l'heuristique à base d'algorithmes . . . 110
4.5.3 Comparaison entre le processus réel et les résultats de l'heuristique à base d'algorithmes pour des instances industrielles . . . 112
4.5.3.1 Analyse basée sur la mesure de la performance : nombre de *moves* par usage . . . 113
4.6 Conclusion . . . 115

5 Mise en œuvre industrielle 117
5.1 Introduction . . . 119
5.2 Présentation du cas industriel . . . 119
5.2.1 Particularités du cas industriel . . . 119
5.2.2 Description de la planification de la production chez ST Crolles . . . 120
5.3 Description de la plateforme développée . . . 123
5.3.1 Vue d'ensemble . . . 123
5.3.2 Conséquences de la mise en œuvre du système de planification à capacité finie . . . 125
5.4 Conclusion . . . 127

Conclusion générale et perspectives 129

Références bibliographiques 133

Annexe A **149**

A Le projet européen *INTEGRATE* 149

A.1 Introduction 149

A.2 WP2 : Optimisation de l'utilisation des équipements 150

A.2.1 WP2.1 : Ajustement dynamique des recettes d'équipement 151

A.2.2 WP2.2 : Gestion de l'équipement liée à l'état de l'équipement 151

A.2.3 WP2.3 : Performance de l'équipement / fabrication visuelle 151

A.2.4 WP2.4 : Mise en œuvre pilote 151

A.3 WP4 : Contrôle des flux de production 151

A.3.1 WP4.1 : Planification et méthodes de répartition 152

A.3.2 WP4.2 : Des outils de simulation et de validation 152

A.3.3 WP4.3 : Implémentations pilotes 152

A.4 WP5 : Analyse des données 153

Annexe B **155**

B Preuve de complexité 155

Table des figures

1.1 Utilisation quotidienne des circuits intégrés 7
1.2 Chiffre d'affaires annuel et taux de croissance cumulé du marché des semi-conducteurs 8
1.3 Processus de fabrication d'un circuit intégré 9
1.4 Exemple d'une route. 11
1.5 Modèle simplifié d'un flux réentrant 14
1.6 Un lot de 25 plaquettes dans un FOUP. 15
1.7 Flux d'information dans un système de fabrication 20
1.8 Classification des techniques industrielles de planification de la production. 23
1.9 Relation entre les approches de planification de la production. 23
1.10 Phases de la simulation 27

2.1 Principaux auteurs. 40
2.2 Nombre d'articles par année classés selon le niveau de décision de la planification. 41
2.3 Répartition des articles selon les niveaux de décision. 42

3.1 Paramètres et variables de décision du problème. 62
3.2 Limites de résolution du MIP. 66
3.3 Principe de décomposition. 68
3.4 Planning de production de l'instance. 75
3.5 Saturation du parc d'équipements par période. 75
3.6 Résultat de l'heuristique : Planning de production. 75
3.7 Résultat de l'heuristique : Saturation du parc d'équipements par période. 76
3.8 Limites de résolution du *MIP*, la procédure d'agrégation, l'heuristique de décomposition et la relaxation lagrangienne. 77

3.9 Qualité de la solution de la procédure d'agrégation, l'heuristique de décomposition et l'heuristique post-lagrangienne. 78

4.1 Évaluation du temps de cycle total. 88
4.2 Composants du temps de cycle total. 88
4.3 Variabilité du temps de process en fonction du type de traitement de l'équipement 89
4.4 Variabilité du temps de cycle en fonction de la taille du *batch* 90
4.5 Principe du calcul du temps de cycle des *steps*. 93
4.6 Simple instance : planning de production à capacité infinie. 94
4.7 Algorigramme de l'heuristique à base de *MIP* 96
4.8 Instance simple expliquant le principe de l'heuristique à base de *MIP* . . 98
4.9 Planning obtenu à la fin du test de l'heuristique 98
4.10 Algorigramme de l'heuristique à base d'algorithmes 99
4.11 Calcul de la charge à capacité infinie. 103
4.12 Algorigramme de l'équilibrage de la charge 105
4.13 Saturation des parcs d'équipements après équilibrage de la charge 107
4.14 Le planning de l'instance obtenu en utilisant l'heuristique à base d'algorithmes 108
4.15 Limite de la résolution de l'heuristique en comparaison avec celle du MIP 110
4.16 Comparaison entre la solution optimale et l'heuristique à base d'algorithmes 111
4.17 Saturation hebdomadaire d'un parc d'équipements de photo-lithographie à capacité infinie et finie 112
4.18 Comparaison du nombre de *moves* réel vs. estimé 114
4.19 Comparaison entre le nombre total de *moves* traités par l'usage de photolithographie réel vs. estimé 114
4.20 Comparaison entre le nombre total de *moves* traités par l'usage de gravure réel vs. estimé 115

5.1 Résultats de saturation des parcs d'équipements. 121
5.2 Écarts entre la quantité prévue à livrer et la quantité livrée réellement pour une vingtaine de produits. 122
5.3 Écarts entre la date de livraison réelle et la date de livraison prévue pour une vingtaine de produits. 122
5.4 Vue d'ensemble du logiciel développé. 124
5.5 Interface graphique du logiciel développé. 125
5.6 Entrées et sorties de chaque module du logiciel développé. 125

5.7 Évolution de l'indicateur juste à temps à ST Crolles 300 126
5.8 Saturation des parcs d'équipements 126

A.1 Les *work packages* du projet *INTEGRATE* 150

B.1 Instance du problème d'ordonnancement. 156

Liste des tableaux

1.1 Classification des systèmes de production 13
1.2 Niveaux de planification de la production selon l'horizon de planification 18
1.3 Principales approches de la méthode *MRP* 25

2.1 Type de la source pour les différents articles de la revue de littérature . . 39
2.2 Source des articles de revues 40
2.3 Classification des articles selon la méthode de résolution 44
2.4 Taxonomie des approches de planification de la production à capacité finie appliquée à l'industrie des semi-conducteurs, extraites de la littérature. 53

3.1 Notations du problème 63
3.2 Synthèse des paramètres des tests réalisés 65
3.3 Valeurs des nouveaux paramètres 68
3.4 Pourcentage des lots pour chaque sous-problème pour des instances réelles 69
3.5 Données relatives aux lots pour une instance simple 74
3.6 Pourcentages des problèmes résolus en moins de $5\,min$ (%) 76
3.7 Résultats expérimentaux : Comparaison de la qualité de la solution des trois méthodes pour résoudre le problème de planification à capacité finie 79
3.8 Résultats expérimentaux : Comparaison du temps de résolution des trois méthodes pour résoudre le problème de planification à capacité finie . . . 80

4.1 Données d'une simple instance 93
4.2 Notations pour le *MIP* mono-période 96
4.3 Paramètres de l'exemple 97
4.4 Notations pour le (PL.3) 101
4.5 Ordre des lots traités par $M2$ selon $rankingCoeff_l$ 107

4.6 Les paramètres du *WIP* au début de la seconde période 107
4.7 Paramètres des tests 109
4.8 Comparaison MIP vs. heuristique 110
4.9 Comparaison du TWT réel versus estimé 115

Introduction générale

L'industrie des semi-conducteurs, et la micro-électronique de façon générale, est l'épine dorsale des innovations qui permettent la fabrication de nombreux nouveaux produits allant des produits de consommation jusqu'aux applications industrielles, aux domaines de l'automobile, des télécommunications, de la médecine, de la bureautique et bien d'autres. Cette industrie consiste à la fabrication d'un circuit intégré. Le processus de fabrication d'un circuit intégré se décompose principalement en deux grandes phases :

— la phase de fabrication des plaques ou *wafers* en anglais, appelée *front-end*.

— la phase d'encapsulation et d'assemblage du produit, qu'on appelle *back-end*.

La phase *front-end* est la phase la plus longue et la plus coûteuse dans le processus de fabrication. En effet, le coût d'une nouvelle usine de fabrication de *wafers* peut atteindre environ 4 milliards de dollars et il faut entre 4 et 6 semaines pour obtenir une plaque avec tous ses circuits intégrés [70]. Une opération aussi critique mérite qu'on s'y attarde et c'est ce à quoi nous nous sommes intéressés dans ces travaux de thèse.

Le processus de fabrication est considéré comme l'un des processus de fabrication les plus complexes [125]. L'usine de fabrication des semi-conducteurs ou le "*wafer fab*" est caractérisée par une production de forte variabilité et faible volume (*High Mix Low Volume* ou *HMLV*) : il y a des centaines de produits et le même équipement peut être partagé par de nombreux produits de diverses technologies, c'est-à-dire nécessitant différents réglages et temps de process.

En outre, le processus de fabrication des *wafers* est composé des centaines d'étapes élémentaires appelées "*steps*". Le nombre important des *steps* est dû au fait que les processus dans les installations de fabrication de semi-conducteurs sont de type ré-entrant, *i.e.* les *wafers* sont traités par les mêmes équipements à plusieurs reprises. Pour chaque *step*, le *wafer* doit être traité par divers types d'équipements selon une recette bien définie. La recette contient les instructions détaillées à utiliser au niveau de l'équipement

afin de procéder à des transformations physiques ou des mesures prévues. L'identification des équipements candidats à utiliser est effectuée par la qualification des recettes sur les équipements. Cependant, dans les *wafer fabs HMLV*, en raison de multiples différences dans des configurations matérielles et logicielles, d'où la variété des recettes à utiliser, il est impossible de qualifier toutes les recettes sur chaque équipement. La qualification est l'une des caractéristiques de la fabrication des semi-conducteurs *HMLV*. Différents parcs d'équipements (*i.e.* ensemble d'équipements parallèles et identiques) peuvent être qualifiés pour la même recette et de multiples recettes peuvent être réalisées sur le même parc d'équipements. Alors, le processus de chaque *step* sur un parc d'équipements spécifique dépend de sa qualification. Ceci est connu par la contrainte des qualifications du processus de fabrication. En outre, chaque parc d'équipements a une capacité identifiée qui se réfère à sa charge limite.

Toutes ces caractéristiques et contraintes rencontrées dans l'environnement de fabrication des semi-conducteurs rendent la planification de la production dans cette industrie très complexe.

En effet, pour établir un planning de production réalisable sur un horizon de planification à moyen terme, il faut donc considérer en plus des caractéristiques du processus de fabrication, les contraintes de capacité et de qualifications des équipements.

En plus, comme pour d'autres industries, les installations de fabrication des semi-conducteurs doivent respecter leurs engagements de livraison aux clients et tenir compte des due dates des lots de production pour survivre dans un environnement commercial concurrentiel. En effet, ne pas répondre aux dates d'échéance peut entraîner des pénalités à cause des retards de livraison et éventuellement la perte de futurs clients. Pour mesurer la qualité d'un planning au point de vue livraison à temps, plusieurs critères ont été utilisés dans la littérature, tels que la minimisation du retard total pondéré, la minimisation de la somme des avances (pondérés) et des retards (pondérés), la minimisation du nombre de lots en retard, etc. Pour les *fabs HMLV*, le temps de cycle des *steps* est très variable en raison de nombreuses sources telles que l'hétérogénéité des modes de fonctionnement des équipements, les priorités des produits, les qualifications des *steps*, etc. Il est alors crucial de considérer également des temps de cycle variables en définissant un plan de production. Dans la pratique, les données historiques de la *fab* et diverses applications de la théorie des files d'attente sont souvent utilisées.

Ainsi, l'objectif de cette étude est de proposer des outils d'aide à la décision pour la planification de la production tout en tenant compte des contraintes de capacité et des qualifications des équipements, des priorités des lots, de la variabilité des temps de cycle et plusieurs caractéristiques du processus de fabrication des semi-conducteurs d'où l'obtention d'un plan de production réalisable.

Le projet, réalisé tout au long de cette thèse, est présenté dans ce rapport réparti en cinq chapitres.

Dans un premier chapitre introductif, nous présentons le contexte industriel de cette étude et nous posons la problématique traitée. Dans le chapitre 2, un état de l'art est réalisé, qui vise à présenter les différents problèmes et techniques de planification présents dans la littérature. L'objectif de cette partie est non seulement de retracer l'état de l'art des techniques existantes, mais aussi d'identifier les écarts entre les différents travaux et de bien positionner notre étude par rapport à la littérature existante. Des méthodes de résolution analytique sont présentées au chapitre 3 et des algorithmes approchés au chapitre 4. La mise en œuvre de ce travail à travers une plateforme de planification de la production pour une entreprise fera l'objet du chapitre 5.

Enfin, nous clôturons ce rapport par une conclusion générale contenant une synthèse du travail effectué et les perspectives éventuelles qui ouvrent les horizons sur d'autres sujets pouvant être abordés à la suite de cette étude.

1

Contexte industriel et problématique

Résumé : Cette thèse a été effectuée dans le cadre d'un projet européen impliquant des industriels des semi-conducteurs. Pour cela, ce chapitre présente ce secteur industriel et les spécificités de son processus de fabrication. À partir des caractéristiques de cet environnement industriel, nous déterminons plusieurs facteurs qui rendent la planification de la production extrêmement difficile et complexe. Nous nous intéressons, par la suite, aux différents techniques et systèmes de planification de la production existants, employés au sein de cet environnement industriel, en précisant leurs avantages et leurs limites. Enfin, la problématique traitée dans ce travail est présentée en détaillant les thèmes de réflexion qui en découlent.

Sommaire

1.1 Introduction 7
1.2 Processus de fabrication des circuits intégrés 8
1.2.1 Description générique du processus de fabrication 8
1.2.2 Entités de base de la fabrication des semi-conducteurs 10
1.2.3 Classification du système de production des semi-conducteurs 11
1.2.4 Complexité de la production dans le secteur de la microélectronique . . . 12
1.3 Planification de la production dans l'industrie des semi-conducteurs 17
1.3.1 Niveaux de la planification de la production 18
1.3.2 Objectifs de la planification de la production 19
1.3.3 Caractéristiques des problèmes de planification de la production dans l'industrie des semi-conducteurs 20
1.3.4 Techniques de planification de la production dans l'industrie des semi-conducteurs 22
1.3.5 Systèmes de planification de la production 31
1.4 Problématique 34
1.5 Conclusion 35

1.1 Introduction

Les circuits intégrés [1] sont de plus en plus utilisés dans tous les domaines de la vie quotidienne. En moyenne, en 2014, une personne utilise plus de 280 puces et 6 milliards de transistors par jour. Ces puces sont installées dans presque tous les équipements qui nous entourent allant des lave-vaisselles, fours, micro-ondes, écrans plats aux téléphones et équipements de bureau [51]. La figure 1.1 illustre l'utilisation moyenne des puces électroniques dans les différentes activités de la vie quotidienne.

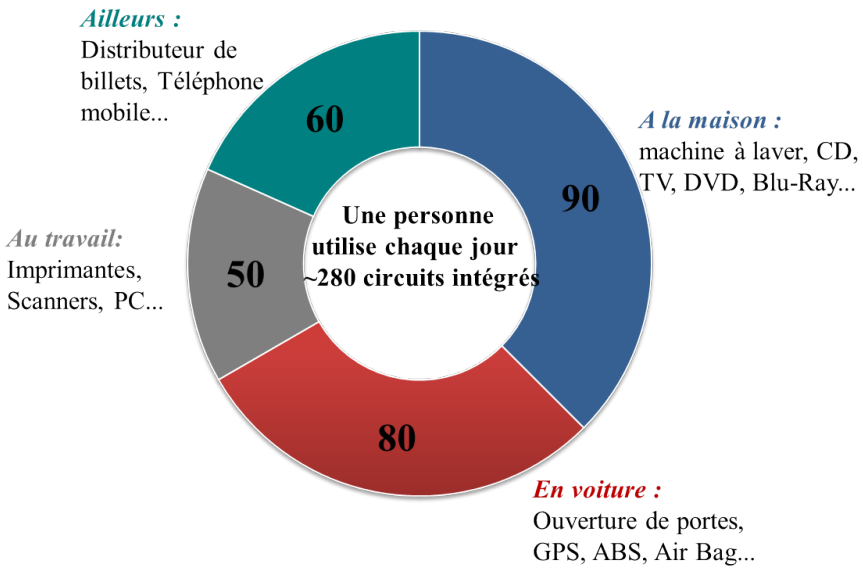

Figure 1.1 – Utilisation quotidienne des circuits intégrés

La demande en circuits intégrés est en forte croissance dans tous les domaines (automobile, communication, divertissement, multimédia, soin médical,...). Le revenu annuel de l'industrie des semi-conducteurs a atteint 335 milliards de dollars en 2015. Durant ces vingt dernières années, son taux de croissance annuel composé (TCAC) atteint 4.31% [1]. La figure 1.2 illustre l'évolution des revenus annuels et du taux de croissance cumulé du marché des semi-conducteurs dans le monde de 1995 à 2015. Cette augmentation de la demande a conduit à une concurrence accrue sur le marché. Par conséquent, les industriels des semi-conducteurs ne doivent pas limiter leur intérêt à la conception du produit mais doivent aussi accorder plus d'attention à la capacité de fabrication afin d'assurer un coût raisonnable et une livraison à temps du produit. En effet, une bonne compréhension de la capacité est essentielle pour maintenir la rentabilité au fil du temps. Dans de nombreux cas, la demande d'un produit est supérieure à la capacité de l'entreprise pour satisfaire

1. Circuit intégré (aussi appelé puce électronique) : composant électronique reproduisant une ou plusieurs fonctions électroniques plus ou moins complexes dans un volume réduit.

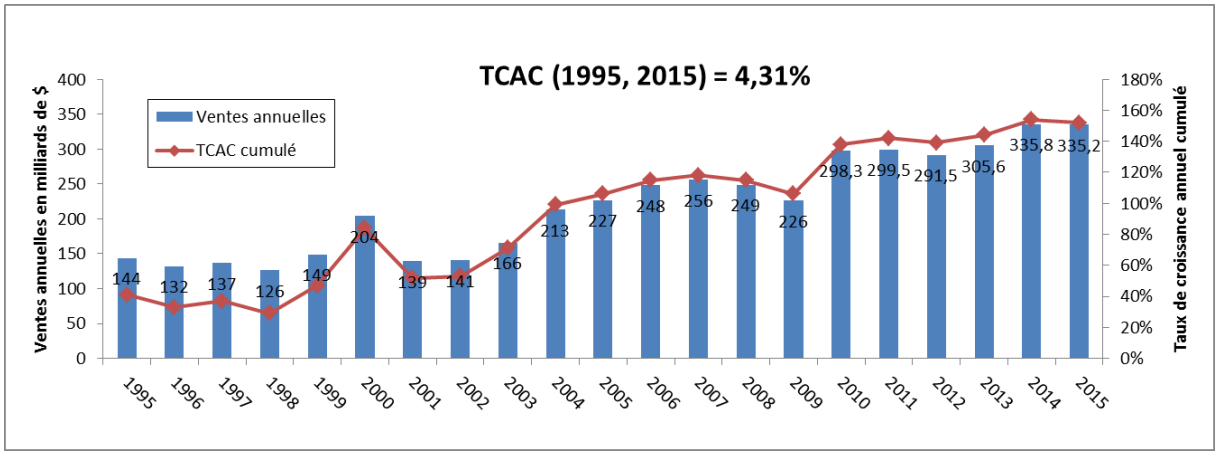

Figure 1.2 – Chiffre d'affaires annuel en milliards de dollars et taux de croissance cumulé du marché des semi-conducteurs dans le monde entier [1].

cette demande. Cela implique une pénalité importante, en termes de pertes de revenus, afin que le chargement de l'usine soit à un niveau inférieur au maximum qu'elle peut gérer. D'autre part, des conséquences négatives importantes peuvent provenir lors d'une surcharge de l'usine. Ces résultats incluent les longs temps de cycle, les dates de livraison manquées, des stocks excessifs, et les rendements éventuellement faibles. Par conséquent, il est essentiel que les fabricants de semi-conducteurs utilisent des méthodes précises pour la planification de leur capacité.

1.2 Processus de fabrication des circuits intégrés

1.2.1 Description générique du processus de fabrication

Les circuits intégrés sont constitués de deux parties : la plaquette (appelée aussi plaque ou *« wafer »* en anglais), partie active en silicium, et le boîtier qui protège la plaquette de son environnement externe et en facilite le montage dans les systèmes électroniques.

La fabrication d'un circuit intégré est effectuée par l'association de plusieurs composants électroniques élémentaires interconnectés (transistors, résistances, etc.) qui sont réalisés sur une même plaquette d'un matériau semi-conducteur (généralement du silicium). La figure 1.3 résume le processus de fabrication. Tout d'abord, les plaquettes brutes sont fabriquées en découpant des lingots de silicium monocristallin. En parallèle, les différents masques sont conçus et réalisés pour chacun des produits. Ensuite, les plaques rentrent dans la première phase de fabrication appelée *« Front End »*. Dans cette phase, on trouve une succession d'opérations d'élaboration des composants et de leurs connexions ainsi qu'un test électrique de validation. Les flux physiques (lots de plaquettes) entre les différents ateliers de l'usine sont schématisés par des flèches sur la figure 1.3. Ensuite, les

plaquettes passent à l'usine de *« Back End »* où elles sont découpées pour obtenir des circuits individuels. Ces derniers sont assemblés, mis en boitier et testés pour obtenir les circuits intégrés.

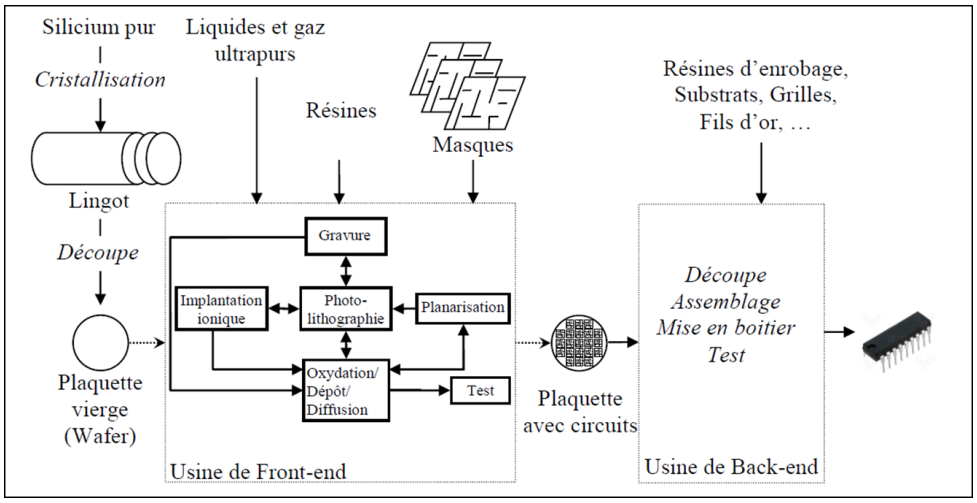

Figure 1.3 – Processus de fabrication d'un circuit intégré [16].

Cette thèse porte sur la planification de production pour la phase de *front end* et plus précisément la fabrication des plaques électroniques. Cette phase correspond à la partie la plus coûteuse (le coût d'une wafer fab peut atteindre jusqu'à 5 milliards de dollars [125]), la plus complexe et la plus longue dans le processus de fabrication par rapport aux autres phases de fabrication. Une description détaillée du processus de fabrication est présentée dans la suite de cette section.

Les plaquettes de silicium vierges subissent des centaines d'opérations de fabrication suivant une gamme spécifique à chaque produit appelée *« route »*[3]. Ces opérations sont répétées pour chaque couche du circuit sur la plaquette et elles peuvent être classées en six catégories :

L'oxydation : Il s'agit de la formation d'une couche de silice (SiO_2) à la surface du *wafer* de silicium par l'oxydation dans un four à haute température. Une couche de résine photosensible est ensuite déposée sur la couche d'oxyde.

La photolithographie : Elle consiste à imprimer des motifs d'un masque sur la couche photosensible déposée sur la plaquette par projection de lumière.

La gravure : C'est une élimination des zones générées par le masque (gravure sèche) ou suppression des éléments indésirables du matériau photosensible (gravure humide).

3. Gamme opératoire pour un produit donné

L'implantation ionique ou dopage : Après la gravure, les surfaces exposées peuvent être chargées électriquement en utilisant un faisceau d'ions de haute intensité.

La déposition chimique : Il s'agit de déposer une couche isolante sur la plaquette pour en aplanir la surface et en isoler les couches de métal.

La création des interconnexions : Cette étape permet de créer des connexions entre les différents composants de la puce par dépôt des couches conductrices sur la plaquette.

1.2.2 Entités de base de la fabrication des semi-conducteurs

Dans cette section, les entités impliquées dans le processus de fabrication des semi-conducteurs sont introduites tout en présentant les relations et les interactions entre elles.

Dans ce contexte industriel, on trouve un mélange entre les termes extraits de l'industrie et ceux utilisés dans la littérature. Par conséquent, nous introduisons des définitions spécifiques et précises afin d'avoir une analyse claire lors de cette étude.

- *Plaque*, *plaquette* ou *« wafer »* : Les plaques non traitées sont la matière première du processus de fabrication des semi-conducteurs et les plaques traitées correspondent au produit final. Ce sont des disques circulaires de silicium constituant le support des circuits intégrés.
- *Produit* : Un produit spécifie le circuit intégré fabriqué. En général, des centaines à des milliers d'unités du même produit sont fabriqués simultanément l'un près de l'autre sur une plaque.
- *Lot* : Un lot se réfère à un ensemble de plaques qui suivent une route ensemble. Toutes les plaques dans un lot correspondent à un même produit. Dans notre cas d'étude, un lot est composé de 25 plaques.
- *Route* : Une route décrit le flux de processus de fabrication d'un lot. Elle se compose d'un ensemble d'opérations depuis le lancement du lot jusqu'à sa livraison.
- *Opération* : Une opération est une étape dans une route. Elle est composée d'un ensemble de tâches élémentaires appelées *« steps »*. On trouve des opérations de process et des opérations de métrologie *i.e.* de contrôle du processus de fabrication. Dans cette étude, seules les opérations de process sont considérées.
- *Step* : Un *step* caractérise une étape élémentaire du processus de fabrication correspondant au passage d'un lot d'une machine à une autre. Il est associé à une recette et un parc d'équipements.

- *Equipement* : Les équipements ou *« tools »* en anglais correspondent aux machines de traitement des différentes opérations du processus de fabrication.
- *Parc d'équipements* : Un parc d'équipements ou *« toolset »* en anglais est un ensemble d'un ou plusieurs équipements ayant des caractéristiques similaires.
- *Usage* : Un regroupement de parcs d'équipements selon leur utilisation *i.e.* un niveau plus agrégé qu'un parc d'équipements.
- *Recette* : Une recette est l'ensemble des instructions nécessaires pour effectuer un *step* du processus de fabrication sur un parc d'équipements. On trouve plusieurs parcs d'équipements pouvant effectuer la même recette. On dit qu'ils sont qualifiés pour la recette.

La relation et l'interaction entre les différentes entités présentées sont illustrées dans la figure 1.4. Cette figure montre l'exemple d'une route composée de n opérations. Chaque opération comporte un certain nombre de *steps* par exemple l'opération 1 se compose de 4 *steps*. Chaque *step* est associé à une recette et un parc d'équipements. Chaque parc d'équipements est qualifié pour différentes recettes par exemple le parc d'équipements *Station B* est qualifié pour les recettes *Rec. B1*, *Rec. B2* et *Rec. B3*.

Opération	*Step*	Parc d'équipements	Recette
Lancement du lot			
Opération 1	*Step* 1.1	Station A	Rec. A1
	Step 1.2	Station B	Rec. B1
	Step 1.3	Station C	Rec. C1
	Step 1.4	Station D	Rec. D1
Opération 2	*Step* 2.1	Station B	Rec. B2
	Step 2.2	Station C	Rec. C1
Opération 3	*Step* 3.1	Station E	Rec. E1
	Step 3.2	Station B	Rec. B3
	Step 3.3	Station D	Rec. D2
⋮			
Opération n	*Step* n.1	Station C	Rec. C2
	Step n.2	Station A	Rec. A2
Livraison du lot			

Figure 1.4 – Exemple d'une route.

1.2.3 Classification du système de production des semi-conducteurs

Les systèmes de production peuvent être classés selon différents critères. Ils peuvent être regroupés dans des classes en fonction du type de production, volume de produc-

tion, flux de production et niveau d'automatisation. Le tableau 1.1 résume les principales classifications des systèmes de production en se basant sur la revue de la littérature (par exemple Groover [69], Zarembra *et al.* [184]). L'environnement industriel, auquel nous sommes intéressés dans cette étude, fait partie de la classe des systèmes de production caractérisée par une forte variabilité (*High mix* en anglais), un faible volume, des ateliers *job shop*, des flux ré-entrants et de très haut niveau d'automatisation. Ces caractéristiques sont mises en évidence en gras et italique dans le tableau 1.1.

1.2.4 Complexité de la production dans le secteur de la micro-électronique

Dans une unité de fabrication de semi-conducteurs (ou *« wafer fab »* [4] en anglais), il y a plusieurs facteurs qui en font un environnement particulièrement difficile à gérer (cf. [37], [70], [124], [125], [167], [168]).

Des flux ré-entrants

Dans le système de fabrication des semi-conducteurs, le produit est réalisé par des traitements sur des couches successives. La réalisation de chaque couche nécessite l'exécution de plusieurs *steps* individuels (photolithographie, gravure, etc.). De plus, plusieurs de ces *steps* sont répétés plusieurs fois. Les équipements utilisés pour les exécuter sont coûteux et ne sont donc pas multipliés dans les ateliers. Ainsi, les lots doivent passer plusieurs fois sur le même équipement au cours de son processus de fabrication. Les systèmes de fabrication dans lesquels les lots doivent visiter à plusieurs reprises les mêmes équipements à des étapes différentes de la gamme opératoire sont appelés lignes ré-entrantes [99]. Un exemple de flux ré-entrant est présenté en figure 1.4. On remarque que le parc d'équipements *Station B* est utilisé plusieurs fois. Ce dernier a traité les *steps 1.2, 2.1* et *3.2*. Un diagramme des flux illustrant une séquence typique du processus de fabrication des semi-conducteurs est schématisé dans la figure 1.5. Les chiffres sur les arcs représentent le numéro du flux parcouru par une plaque. Dans cet exemple particulier, l'oxydation est répétée six fois (flux 1, 2, 3, 6, 7 et 8) et la photolithographie quatre fois (flux 3, 4, 5 et 6).

La principale conséquence du caractère ré-entrant est que des *wafers* à différents stades du processus de fabrication sont en compétition pour le même équipement. Par conséquent, les plaques passent une grande partie de leur temps soit en attente d'un équipement, soit dans une activité de transport entre deux opérations, au lieu d'être effective-

4. Usine de fabrication des plaques électroniques.

Tableau 1.1 – Classification des systèmes de production

Critère de classification	Types	Caractéristiques
Volume de production	***Production à faible volume***	Gamme de production (1 à 100 unités/an), des ateliers de type *Job shop*
	Production à moyen volume	Gamme de production (100 à 10000 unités/an), la complexité augmente avec l'augmentation du *mix* produit
	Production à volume élevé	Gamme de production (10000 à des millions d'unités/an), production de masse
Flux de production	***Job Shop***	Différents *jobs* réalisés sur différents équipements, la séquence des *jobs* est prédéfinie
	Production par batch	La variété des produits est très importante, des *batches* pour chaque type de produit, et la plupart du temps des ordres répétés
	Production de masse / *Flow shop*	Volume élevé de produits, conception stable et demande exigée des produits, ordre de passage unique pour tous les *jobs* sur les différents équipements
	Production à forte variabilité (High mix)	Mix produit élevé, système de production flexible, production à la commande
	Flux ré-entrants	La séquence des opérations est répétée plusieurs fois durant le processus de fabrication
Niveau d'automatisation	Manuel	Toutes les tâches sont effectuées par des opérateurs
	Semi-automatisé	Combinaison entre des tâches manuelles et des tâches automatisées, par exemple les lignes d'assemblage
	Automatisé	Manutention robotisée (*Automatic Material Handling Systems*), transfert inter-cellulaire automatisé

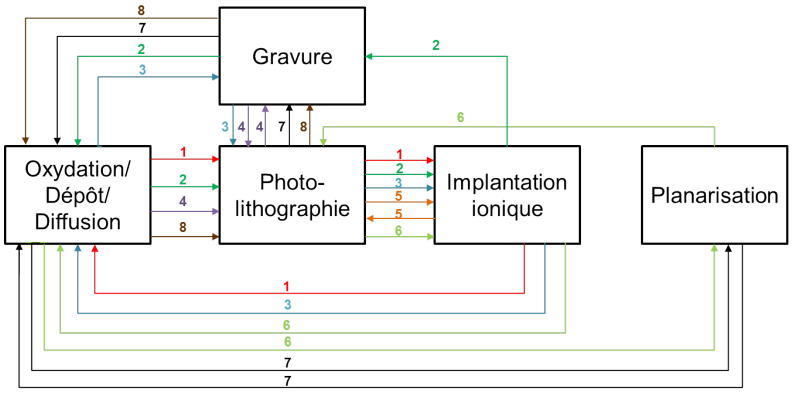

Figure 1.5 – Modèle simplifié d'un flux ré-entrant dans l'industrie des semi-conducteurs.

ment traitées sur un équipement. Cette caractéristique principale engendre des problèmes de planification difficiles à gérer dans la pratique et intraitables sur le plan théorique [99].

Diversité des produits

Les *fabs* peuvent être classées en *low-mix* ou *high-mix* selon le nombre de produits différents. Dans les *fabs low-mix*, les équipements peuvent être dédiés aux produits, tandis que dans les *fabs high-mix*, le même équipement doit être partagé par de nombreux produits de diverses technologies, c'est-à-dire nécessitant une configuration et des temps de process différents.

Dans cette étude, nous nous intéressons aux *fabs high-mix* où la planification de production est plus complexe.

Multiplicité des opérations

Le processus de fabrication des semi-conducteurs est composé d'environ 250 opérations par produit. Chaque opération comporte entre 400 et 800 *steps* (nettoyage, processus, mesure).

Hétérogénéité des opérations

Dans une usine de fabrication de semi-conducteurs typique, il y a des dizaines de flux de production pour lesquels le « *mix* »[5] produit évolue au fil du temps [125]. En outre, selon la nature des opérations, le temps de process des *steps* varie de manière significative. Certains *steps* de processus nécessitent 15 minutes ou moins pour traiter un lot tandis que d'autres demandent plus de 12 heures de temps de process. Les longs *steps*, représentant le tiers des opérations dans la *fab*, se réalisent par *batch* où plusieurs lots sont traités

simultanément.

Différents modes opératoires

Certains *steps* sont effectués sur des plaques individuelles, d'autres sur des groupes de plaques (lots), et d'autres encore sur des groupes de lots (*batches*). Un lot est composé généralement de 25 plaques, tandis qu'un *batch* typique contient jusqu'à six lots. Dans tous les cas, les plaquettes sont transportées par lot dans des caissons de protection en plastique appelés *« FOUP » (Front-Opening Unified Pod)* pouvant contenir jusqu'à 25 plaquettes (cf. figure 1.6). Le transfert des *FOUPs* entre les différents ateliers est réalisé manuellement ou avec des systèmes automatisés appelés *« AMHS » (Automatic Material Handling Systems).*

Figure 1.6 – Un lot de 25 plaquettes dans un FOUP.

Longueur du temps de cycle

Le temps de cycle est le temps écoulé entre le moment où un lot de *wafers* entre à l'usine de fabrication et le moment où il sort. En général, le temps de cycle de fabrication d'un lot est de l'ordre de 7 à 8 semaines. Il est composé d'un temps d'attente devant les équipements et d'un temps de process.

Variabilité du processus de fabrication

Le processus de fabrication des semi-conducteurs est caractérisé par une forte variabilité due à plusieurs facteurs, parmi lesquels on cite :

- Les arrêts des équipements, prévus et imprévus, conduisant à une faible disponibilité des équipements en comparaison avec les autres industries de production de masse.

5. Mot anglais utilisé dans le secteur de microélectronique exprimant la complexité de l'en-cours de production liée à la variété des produits et des technologies et la demande du marché.

- La réentrance des flux aboutissant à une grande variabilité dans les arrivées des lots aux parcs d'équipements.
- Le fonctionnement par *batch* qui résulte en une augmentation de la variabilité des temps d'inter-arrivées des lots aux parcs d'équipements traitant les opérations ultérieures.
- Le *mix* produit dû à la diversité des routes et des processus de fabrication.

Les quatre facteurs présentés contribuent ensemble à une variation importante des temps d'inter-arrivées et des temps opératoires des lots. Ainsi, le processus de fabrication des circuits intégrés est caractérisé par une variabilité du temps de cycle.

Diversité des caractéristiques des équipements

Une *wafer fab* est composée de plus d'une centaine d'équipements organisés en ateliers de fabrication. Chaque atelier est composé de groupements d'équipements identiques, parallèles et qui exécutent des opérations similaires, formant un parc d'équipements. Chaque parc d'équipements possède généralement ses propres caractéristiques selon le débit de production, le temps de processus ou de réglage, la taille de lot, la configuration, etc. Certains équipements ont d'importants temps de *setup* dépendant de la séquence telles que les implanteurs ioniques, tandis que d'autres n'en ont pas.

Selon le mode opératoire, les parcs d'équipements peuvent se diviser en deux types, ceux fonctionnant en série et d'autres fonctionnant par *batch*. Les équipements fonctionnant en série traitent les plaques une par une, alors que ceux fonctionnant par *batch*, utilisés pour les opérations de gravure et de traitement thermique, traitent plusieurs lots simultanément.

Le coût d'un équipement de fabrication des semi-conducteurs atteint 75% des coûts globaux d'investissement de l'usine [70]. Cela est dû principalement à la haute précision des équipements tels que les scanners pour la photolithographie qui valent jusqu'à 70 millions d'euros l'unité. Par conséquent, les équipements ne sont pas multipliés dans les ateliers de fabrication. Ainsi, l'utilisation des équipements présente une contrainte importante lors de la planification de production des semi-conducteurs.

Qualifications recette-équipement

La contrainte de qualification définit l'éligibilité d'un équipement à traiter un produit. La réalisation d'un process sur un équipement nécessite la qualification de la recette sur l'équipement considéré. À cause des caractéristiques matérielles et informatiques, les opérations ne peuvent pas être effectuées sur tous les équipements. Autrement dit, une

recette ne peut être appliquée que sur certains parcs d'équipements.

Multiple priorités des lots

Afin de maintenir la compétitivité et satisfaire les commandes clients urgentes, les industriels des semi-conducteurs définissent plusieurs niveaux de priorité des ordres de fabrication. Les priorités de production peuvent être divisées en trois niveaux selon l'urgence de livraison : haut, moyen et faible. Le traitement opérationnel doit respecter les ordres de priorité identifiés à priori.

Délais de livraison à respecter

Comme toute autre industrie, l'industrie des semi-conducteurs a pour objectif de répondre à l'une des attentes les plus importantes des clients : maximiser la livraison à temps *i.e.* minimiser les retards de livraison. En effet, si le fabricant ne peut pas répondre aux dates d'échéance de livraison prédéfinies, il pourrait être confronté à de fortes pénalités et éventuellement perdre ses clients. Afin d'évaluer la qualité d'un planning en termes de respect des dates d'échéance de livraison, plusieurs critères ont été utilisés tels que minimiser la somme des retards pondérés, minimiser le retard moyen, minimiser le nombre de tâches en retard, etc.

1.3 Planification de la production dans l'industrie des semi-conducteurs

La planification de la production est un processus de prise de décision qui joue un rôle important dans la plupart des industries manufacturières et de services. La fonction de planification vise à prévoir l'utilisation des ressources et l'exécution des *jobs* pour atteindre des objectifs déterminés. Les ressources peuvent être matérielles telles que les machines ou la matière première ou humaines *i.e.* le personnel du centre de travail. Les *jobs* peuvent être des opérations dans l'atelier ou des tâches (de maintenance ou de process) à effectuer. Chaque job peut avoir un niveau de priorité, une date de début au plus tôt et une date d'échéance de livraison. Les objectifs d'optimisation sont nombreux tels que la réduction du temps de cycle des *jobs*, la minimisation des retards, la maximisation de l'utilisation des ressources,... Le processus de planification prend en considération le contexte et les contraintes internes et externes connues actuellement ou prévisibles dans le futur.

Le rôle de la planification dans un environnement de fabrication est crucial. Les commandes des clients sont converties en *jobs* avec des dates d'échéance associées. Ces *jobs*

doivent souvent être réalisés sur les machines d'un centre de travail selon une séquence donnée. Le traitement des *jobs* peut être retardé si certaines machines sont occupées. Des événements imprévus, tels que les pannes de machines ou des temps de process plus longs que prévu, doivent également être pris en compte, car ils peuvent avoir un impact majeur sur le planning.

1.3.1 Niveaux de la planification de la production

Dès 1965, Anthony [2] a distingué trois niveaux de planification de la production selon la longueur de l'horizon de planification : le long terme (planification stratégique), le moyen terme (planification tactique) et le court terme (planification opérationnelle). La dimension temporelle diffère d'une industrie à une autre [170].

Le tableau 1.2 présente une description de chaque niveau de planification selon l'horizon temporel.

Tableau 1.2 – Niveaux de planification de la production selon l'horizon de planification

Niveau de planification	**Exemples**	**Horizon**
Long terme (stratégique)	— Recherche de nouveaux partenaires industriels, — Sélection des fournisseurs et sous-traitants, — Ouverture, fermeture ou délocalisation des sites de production, — Développement d'un nouveau produit, — Configuration de l'usine,...	de 1 à 5 ans
Moyen terme (tactique)	Programme directeur de production, équilibrage de la charge des ressources	de 3 à 6 mois
Court terme (opérationnel)	Ordonnancement, suivi des ateliers, règles de répartition	de 1 à 6 semaines

Au niveau stratégique, les décisions prises se traduisent par une formulation à long terme de la politique de l'entreprise (vision sur plusieurs années). Ces décisions concernent la mise au point des installations de production (par exemple, la taille et l'emplacement de nouvelles usines, l'acquisition de nouveaux équipements ...) ou du processus de fabrication (développement d'un nouveau produit ou d'une nouvelle gamme de production ...). Le niveau stratégique fixe un cadre au niveau tactique.

Les décisions tactiques concernent l'organisation des produits et des ressources en fonction des prévisions commerciales et correspondent à un ensemble de décisions à moyen terme. À ce niveau, on retrouve notamment la planification de la production qui vise à

calculer un plan de production. Dans ce plan, les quantités à produire par période sont calculées de façon à répondre aux demandes au moindre coût. Ces décisions tactiques sont des directives pour la production détaillée et la planification au niveau opérationnel.

À court terme, les décisions opérationnelles consistent à gérer le fonctionnement quotidien des ateliers de fabrication. À ce niveau, les décisions opérationnelles concernent les volumes de production et les dates de passage par produit ou lot, l'utilisation détaillée de la capacité par parc d'équipements, etc.

Dans cette étude, nous nous intéressons à la planification de la production à moyen et court terme impliquant des décisions plus tactiques sur des périodes mensuelles ou hebdomadaires. Nous considérons la planification de la capacité de production qui consiste à estimer la capacité de production nécessaire pour satisfaire les commandes clients.

1.3.2 Objectifs de la planification de la production

Planifier les activités et les ressources de production, en tenant compte de l'état des encours de production, les prévisions de la demande, et la capacité des ressources, est un objectif principal de la planification sur le niveau tactique. Les décisions, prises à ce propos, ont une incidence directe sur le processus de planification. La planification de la capacité de production est l'un des niveaux les plus critiques de la planification, car elle est le lien entre la planification stratégique et la planification opérationnelle. Elle est conduite, comme le montre la figure 1.7, par des décisions prises à moyen et court terme.

Les activités principales, à ce niveau de planification, sont la détermination des dates de début et de fin des tâches à exécuter et le calcul de la charge des ressources. La planification de la capacité de production doit interagir avec d'autres procédures de prise de décision utilisées dans l'usine. La figure 1.8 représente, de façon générale, le flux d'informations dans un système de fabrication [135].

A ce niveau de décision, il est courant de planifier dans les grandes lignes c'est-à- dire de ne pas prendre en compte certains détails qui complexifient trop la prise de décision. Ainsi, un ensemble de variables possédant des caractéristiques communes sont remplacées par une variable agrégée. Par exemple, les produits sont agrégés par familles, les ressources en grandes catégories .... L'agrégation des décisions permet de simplifier considérablement la formulation et la résolution des problèmes. Le plan de production ainsi obtenu est appelé plan agrégé. Cependant, ce plan doit être calculé de manière à être réalisable.

Généralement, parmi les objectifs des systèmes de planification de la production, on trouve la livraison à temps, la minimisation des encours de production, la réduction des temps de cycle et l'utilisation maximale des ressources. Cependant, ces objectifs peuvent être divergents, par exemple la satisfaction des délais de livraison peut être inconciliable avec la réduction des stocks ou le respect des contraintes de capacité. Le but de la pla-

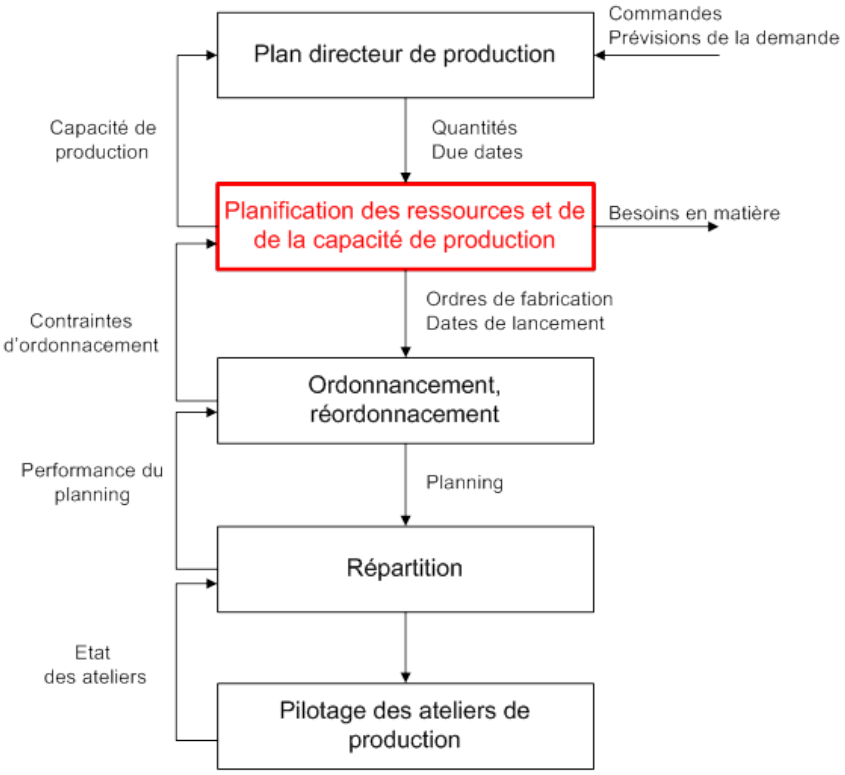

Figure 1.7 – Flux d'information dans un système de fabrication [135].

nification de la capacité est de trouver un équilibre profitable entre ces objectifs contradictoires [80]. Il s'agit d'établir un planning de production réalisable respectant simultanément les capacités disponibles et les dates d'échéance de livraison en utilisant les équipements au plus près de leur charge maximale.

Les problèmes de planification ont été et sont encore largement rencontrés dans l'industrie des semi-conducteurs. Une attention particulière est portée sur les problèmes de planification de la capacité de production qui visent à déterminer un équilibrage entre la charge et la capacité de production. Nous allons donc présenter les caractéristiques des problèmes de planification ainsi que les techniques de planification utilisées pour les résoudre dans la suite de ce chapitre.

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

1.3.3.1 Contraintes

Comme mentionné ci-dessus, il s'avère que le processus de fabrication des semi- conducteurs est très complexe impliquant une grande variété de produits, un grand nombre d'opérations, un long temps de cycle, de multiples contraintes liées aux équipements, une

forte variabilité du processus de fabrication et une pénalité élevée en cas des retards de livraison. Ces caractéristiques rendent la planification de la production, au sein de cet environnement complexe, très difficile. En plus des contraintes issues du processus de fabrication des semi-conducteurs, on trouve les contraintes de capacité et la forte variabilité des temps de cycle.

1.3.3.2 Indicateurs de performance

Un indicateur permet de mesurer de façon objective l'efficacité d'un dispositif mis en place : c'est un outil décisionnel synthétique, c'est-à-dire facile à établir et pratique à utiliser, facilitant le dialogue entre les utilisateurs ayant des cultures et des préoccupations différentes. Dans les *fabs* de semi-conducteurs, plusieurs indicateurs sont utilisés pour mesurer la performance. Le lecteur intéressé peut se référer à Montoya-Torres [126] pour plus de détails. Parmi les indicateurs de performance de la planification de la production les plus importants et les plus employés dans cette industrie, on cite :

Temps de cycle

Le temps de cycle de production, encore appelé temps de séjour ou délai de fabrication, est la durée totale nécessaire pour fabriquer un lot de *wafers*. Il mesure le temps écoulé entre le moment où le lot entre dans l'unité de fabrication et le moment où il sort.

Le temps de cycle inclut le temps opératoire, le temps de transfert entre les opérations, le temps d'attente devant les équipements et le temps d'attente pour les transferts. Dans la plupart des *wafer fabs*, un lot de *wafers* peut passer entre 50% et 90% de son temps de cycle en attente d'un équipement ou d'un transfert [126]. Le temps de cycle est une mesure de la capacité requise par la fab pour délivrer ses produits à temps. Généralement, une réduction du temps de cycle implique une augmentation de la satisfaction des clients. Les entreprises, dont la fabrication s'effectue avec des temps de cycle courts, sont capables de lancer souvent de nouveaux produits, de pénétrer plus rapidement de nouveaux marchés, de réagir plus efficacement aux changements de la demande et de délivrer à temps leurs produits.

Dans l'industrie des semi-conducteurs, le *X-factor*, représentant le rapport entre temps de cycle total et le temps total de process [114], est plus employé que le temps de cycle comme indicateur de performance. Avec cet indicateur, l'objectif recherché est d'avoir un X-factor le plus proche possible de l'unité. Cela impose une diminution de la moyenne et de la variance du temps de cycle.

Niveau d'encours de production

Le niveau d'encours (*Work-In-Progress*, WIP) est défini comme le nombre de *wafers* de production se trouvant dans la *fab* à un instant donné. Le nombre de *wafers* pris en compte est aussi bien sur une activité de production que sur une activité de non-production (transport et attente des lots). Les encours représentent du capital immobilisé dans les ateliers, donc trop d'encours diminue la trésorerie disponible.

Débit du système ou *throughput rate*

Le débit du système, appelé « *throughput rate* », est défini comme le nombre de *wafers* finis qui sortent de la fab par période. Cet indicateur indique le niveau de saturation ou d'utilisation du système *i.e.* il permet de savoir si le système est stable, c'est-à-dire s'il n'y a pas d'accumulation de produits à certaines positions dans le système.

Niveau de service

Dans le cadre de la fabrication des *wafers*, des indicateurs de performance importants sont des objectifs liés aux dates d'échéance de livraison pour mesurer le niveau de service offert au client (interne ou externe).

1.3.4 Techniques de planification de la production dans l'industrie des semi-conducteurs

Les techniques industrielles de planification de la production peuvent être réparties en : (cf. figure 1.8)

- Techniques classiques : *MRP, MRPII, JIT* et *TOC*
- Modèles analytiques
- Modèles de simulation
- Heuristiques
- Techniques basées sur l'intelligence artificielle

Souvent, ces techniques de résolution sont combinées pour fournir des solutions plus complètes, voir figure 1.9.

1.3.4.1 Les techniques classiques

Parmi les techniques industrielles de planification de la production les plus classiques et les plus répandues, on trouve la technique *Material Requirement Planning (MRP)* introduite par Orlicky en 1975 [130] qui permet de calculer les besoins en composants sans

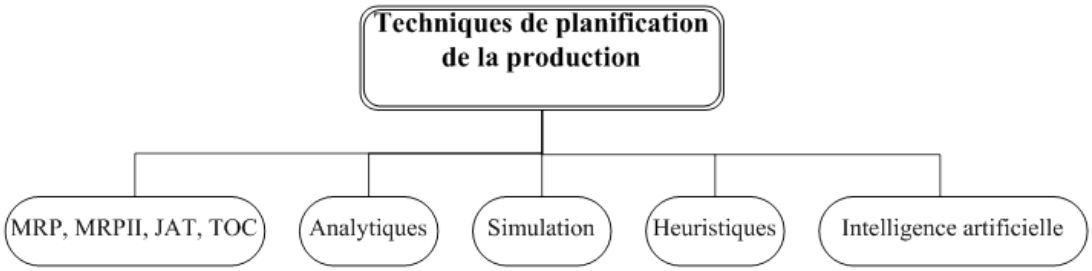

Figure 1.8 – Classification des techniques industrielles de planification de la production.

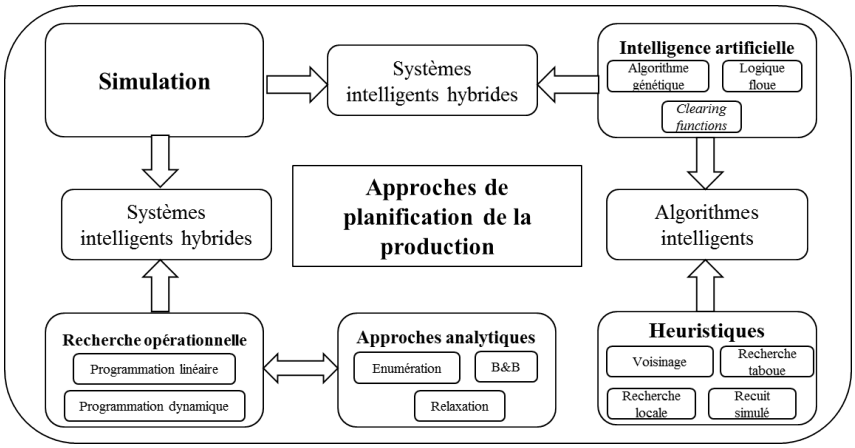

Figure 1.9 – Relation entre les approches de planification de la production.

capacité et son évolution, la technique *Manufacturing Resource Planning (MRPII)* développée par Wight en 1981 [176] qui intègre un système ajustant les niveaux de capacité. Ces techniques permettent de générer des plans de production sur la base des commandes d'entrée au système. Le concept principal de la méthode *MRP*, dans la génération de ces plans, est de considérer un temps de cycle fixe (ou *lead time* en anglais) et une capacité infinie des machines.

L'application de ces techniques traditionnelles pour la planification des capacités dans l'industrie des semi-conducteurs présente ainsi certaines limitations. En effet, il est prouvé que la méthode *MRP* est inefficace. Elle peut provoquer des problèmes de types retards de livraison ou stocks d'en-cours et produire des plannings de production irréalisables lorsqu'elle est appliquée à l'industrie des semi-conducteurs vu qu'elle ne tient pas compte des contraintes de capacité et suppose que les temps de cycle sont fixés [17, 161]. Cependant, dans les installations de semi-conducteurs, les temps de cycle dépendent de nombreux facteurs, tels que le taux d'utilisation des équipements, la taille du lot, les volumes de stock et les règles de répartition, et sont donc variables.

L'approche *MRP II* va plus loin que la précédente en cherchant à ajuster la charge souhaitée et la capacité disponible pour chaque centre de production. Les avantages offerts par l'utilisation de cette méthode sont multiples : plus grande réactivité et meilleure

maîtrise de contrôle de production, diminution des immobilisations en stock .... Cette méthode résout donc un certain nombre de problèmes évoqués pour l'approche *MRP*. Cependant, ce système prend en compte des contraintes de capacité agrégées ne permettant pas de garantir la faisabilité du plan de production au niveau opérationnel.

D'autres concepts tels que la planification des besoins de capacité (*Capacity Requirement Planning (CRP)* en anglais) et le pilotage d'atelier (*Shop Floor Control (SFC)* en anglais) sont ensuite développés. Le but de ces méthodes est de remédier au problème de la non prise en compte des contraintes de capacité après le calcul *MRP* pour obtenir un plan de production réalisable [76, 161]. Bien que les systèmes *CRP* et *SFC* puissent résoudre le problème de la capacité, ils ne résolvent pas le problème au niveau du calcul *MRP*. La façon appropriée de considérer les contraintes de capacité à ce stade est d'intégrer la méthode *MRP* avec la planification à capacité finie [7]. Ainsi, des systèmes *MRP* à capacité finie (*Finite capacity Material Requirement Planning (FCMRP)* en anglais) ont été développés pour aborder le problème de planification de la capacité au niveau du calcul *MRP* [133, 180]. Les systèmes *FCMRP* développés utilisent deux approches différentes afin d'inclure la capacité finie : pré/post analyse du calcul *MRP* et l'ordonnancement à capacité finie [127]. Mais, aucune de ces approches ne résout le problème de la capacité pendant l'exécution du calcul *MRP*. Kanet et Stößlein (2010) [93] ont développé le système « *Capacitated ERP* » (*CERP*). Ce système diffère du système *MRP* classique dans la prise en compte des contraintes de capacité avant le calcul des besoins nets. Cependant, le modèle établi est limité à la production en une seule étape, avec une nomenclature à un seul niveau et une seule machine.

Les principales approches de *MRP* et la planification de la capacité sont résumées dans le tableau 1.3. Ainsi, on constate que bien que la prise en compte de la capacité limitée dans les systèmes *MRP* présente un problème de recherche ancien, la résolution de ce problème n'a pas encore abouti jusqu'à maintenant à des résultats satisfaisants.

Les approches *MRP-CRP*, *MRP-SFC* et *FCMRP* nécessitent beaucoup de temps de calcul et tentent de résoudre le problème de capacité après le calcul *MRP*. Les contributions de recherche qui tentent d'intégrer les contraintes de capacité lors du calcul *MRP* sont souvent limitées à des processus de fabrication simples et pour une production répétitive, par exemple les lignes d'assemblage, ce qui est différent de l'environnement complexe et de forte variabilité de l'industrie des semi-conducteurs.

En plus de la technique *MRP* et ses évolutions, on trouve d'autres techniques classiques industrielles telles que la méthode *Just In Time (JIT)* [64] et la théorie des contraintes (*Theory Of Constraints (TOC)*) [63].

Bien que la technique *JIT* prouve ses forces, elle présente certaines limitations dans les systèmes de production à forte variabilité et faible volume. Elle est plus adéquate pour

Tableau 1.3 – Principales approches de la méthode *MRP*

Approche	Références	Limites de l'étude
MRP, MRPII	[130], [176]	*lead times* constants, capacité infinie
MRP-CRP	[76]	Identification des problèmes de capacité après l'exécution du calcul *MRP*, l'intervention des planificateurs est indispensable
MRP-SFC	[161]	Le problème de capacité n'est pas résolu au niveau du calcul *MRP*
FCMRP	[133, 180]	Le problème de capacité n'est pas résolu au niveau du calcul *MRP*
Approche intégrée du MRP et planification de capacité	[161]	*lead times* fixes
	[93]	Production en une seule étape, un seul niveau de nomenclature et avec une seule machine

un environnement de production répétitive avec une demande stable et une faible gamme de produits [23].

La TOC semble une technique de planification de la capacité efficace dans l'industrie des semi-conducteurs [141], mais elle ne tient compte que des postes goulots d'étranglement et elle est limitée dans le cas de variation des goulots d'étranglement.

1.3.4.2 Les modèles analytiques

L'approche analytique consiste à générer une ou plusieurs solutions par la résolution d'un modèle établi à partir du problème traité. Le modèle analytique décrit, selon un formalisme mathématique, le système étudié et les objectifs à atteindre. La formalisation mathématique repose sur :

- les variables de décision,
- les paramètres ou les variables d'état du système,
- les relations liant les variables de décision aux variables d'état,
- les contraintes du système issues des caractéristiques statiques (ex : la limitation des capacités des équipements, etc.),
- la définition du critère à optimiser (maximiser ou minimiser). Ce critère est la fonction objectif qui dépend des décisions prises.

Ainsi, le but d'une modélisation analytique est d'obtenir, par le calcul, les décisions à prendre pour optimiser le critère choisi tout en satisfaisant les contraintes imposées. Un modèle analytique peut, par exemple, être formulé en utilisant la programmation

linéaire, la programmation dynamique, la théorie des files d'attentes, etc. Les tableurs et les logiciels solveurs, commerciaux ou libres, sont abondants et offrent des environnements complets pour la programmation et la résolution sophistiquée de tels modèles, ce qui facilite la tâche du décideur.

Toutefois, la conception d'un modèle analytique est souvent limitée par la complexité du système étudié. Dans ce cas, le modèle analytique subit forcément d'importantes simplifications par rapport à la réalité, ce qui peut dégrader l'efficacité des décisions calculées.

Concernant la planification de production dans l'industrie des semi-conducteurs, plus spécifiquement les problèmes de planification à moyen et court terme de la production, parmi les modèles analytiques développés, on trouve les modèles statiques ([13], [129], [132], [177]), les modèles de files d'attente ([43], [152]), la programmation linéaire ([15], [25], [104], [102]) et la programmation stochastique ([10], [79], [158]). Dans ces études, il est montré que la résolution des modèles analytiques, pour des problèmes industriels, demande énormément de temps de calcul. En pratique, ces modèles sont sensiblement liés à la taille du problème et leur utilisation est sanctionnée par des temps d'exécution souvent inacceptables.

1.3.4.3 Les modèles de simulation

L'approche par simulation permet l'évaluation des performances du système de production. Il s'agit de l'expérimentation d'une ou plusieurs solutions envisagées, sur un modèle du système réel étudié, pour retenir la meilleure des solutions testées. Classiquement, nous distinguons quatre grandes phases pour mener une simulation, comme le montre la figure 1.10. Tout d'abord, nous commençons par la modélisation du système étudié. Ensuite, le modèle conçu est programmé via un logiciel de simulation. Après, l'expérimentation est mise en œuvre à l'aide de ce logiciel. Enfin, les résultats simulés sont interprétés et accompagnés d'actions puis resimulés et ainsi de suite, jusqu'à avoir des résultats qui satisfont le décideur compte tenu des critères visés.

En comparaison avec l'approche analytique, une approche par simulation présente un certain nombre de différences. En effet, la simulation ne construit pas une solution ; elle permet de traduire l'effet d'une solution déjà conçue par l'utilisateur sur la suite des états du système et sur les valeurs des critères retenus sur l'horizon de l'étude. En plus, la simulation n'est pas une technique d'optimisation au sens propre du mot. Elle ne peut qu'indiquer les conséquences des décisions prises par l'utilisateur sur le comportement du système. Dans ce sens, la simulation se présente comme une technique itérative qui n'établit pas une solution finale. Elle permet au décideur d'envisager des choix possibles en fonction de ce qui répond le mieux aux problèmes posés.

Les modèles de simulation sont mieux adaptés aux systèmes complexes que les modèles

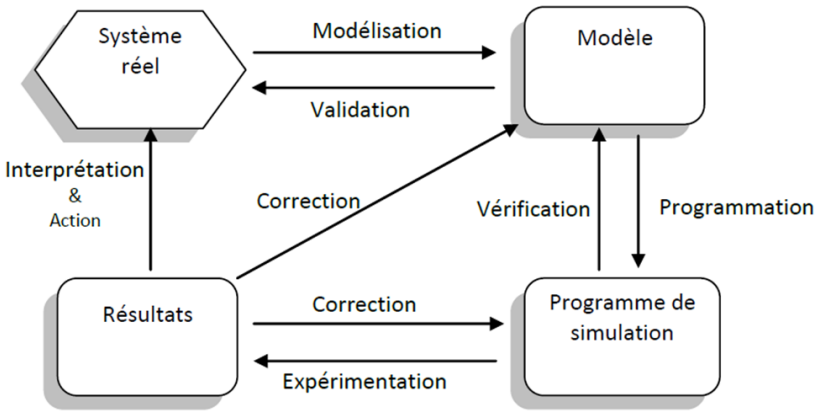

Figure 1.10 – Phases de la simulation (extrait du cours de Jean-Louis Boimond, Université d'Angers ISTIA).

analytiques car ils tiennent compte de l'aspect dynamique et des interactions entre les différents composants du système en intégrant beaucoup plus des contraintes dynamiques du processus de fabrication. C'est le cas des unités de fabrication de semi-conducteurs où, à chaque événement, plusieurs phénomènes peuvent se produire en même temps. Ainsi, plusieurs modèles de simulation pour le processus de fabrication des semi-conducteurs ont été établis ([3], [138], [140], [155], [175]). L'objectif principal de ces modèles consiste à observer une valeur réaliste du temps de cycle du processus de fabrication ([56], [163], [165]). En effet, la simulation à événements discrets est considérée comme la seule méthode pratique permettant de mesurer le temps de cycle en fonction de la disponibilité des équipements et la vitesse de production. Le modèle de simulation peut également être utilisé pour déterminer les goulots d'étranglement dans le cadre d'un mix produit donné et de prendre des décisions stratégiques concernant l'achat de matériel [137]. La simulation à évènements discrets est combinée parfois avec des approches de recherche opérationnelle. Le travail de Hung et Leachman [84] est un exemple d'une telle approche.

Cependant, les modèles de simulation présentent quelques inconvénients. En effet, ils exigent une énorme quantité de données d'entrée, y compris les détails sur les équipements, les politiques de gestion des encours de production, les informations sur les produits, etc. Ils nécessitent également beaucoup d'efforts pour la construction des modèles. En se basant sur la nature des modèles de simulation, plusieurs réplications sont nécessaires pour effectuer une analyse statistique confiante ce qui entraîne un long temps d'exécution. En particulier, pour l'industrie des semi-conducteurs, la complexité des flux de production et la forte variabilité du processus de fabrication engendrent une difficulté lors du développement des modèles de simulation et un écart entre le processus modélisé et le processus réel.

1.3.4.4 Les techniques heuristiques

Au fil de l'essor des outils de la recherche opérationnelle, les heuristiques ont été largement sollicitées afin de résoudre une panoplie de problèmes pratiques. Ce succès est dû essentiellement à leur grande capacité d'exploration et d'exploitation de l'espace des solutions.

a) **Techniques de programmation mathématique**

Deux techniques de résolution classiques pour des problèmes de programmation mathématique sont les algorithmes de séparation et évaluation (*Branch & Bound*) et la relaxation Lagrangienne. Les algorithmes de *Branch & Bound* coupent les branches de l'arbre d'énumération et réduisent donc sensiblement le nombre de nœuds générés. Une solution optimale peut être trouvée en examinant systématiquement les sous-ensembles d'une solution réalisable. Plusieurs algorithmes différents existent pour des problèmes de planification de la production appliquée à l'industrie des semi-conducteurs [10, 169]. L'un des principaux inconvénients de toutes les méthodes de *Branch & Bound* est l'absence de fortes bornes inférieures afin de couper des branches de l'arbre d'énumération le plus tôt possible [18].

Pour certains problèmes de planification de la production, il s'est avéré que la relaxation Lagrangienne est une méthode efficace permettant d'obtenir des solutions approchées. Cette technique est appliquée aux problèmes de planification stratégique [158] ou de planification opérationnelle *i.e.* ordonnancement [35, 96, 109]. L'idée générale derrière l'approche de relaxation lagrangienne est de décomposer le problème principal de planification en sous-problèmes. Ceci est possible en relâchant les contraintes de capacité grâce à l'utilisation des multiplicateurs de Lagrange. Les dates de début de chaque job sont alors déterminées en considérant les multiplicateurs de capacité et la spécificité du process des *jobs* ainsi que les exigences de priorité. Les résultats, en utilisant cette approche, ne sont pas garantis pour des problèmes complexes et de très grande taille.

b) **Techniques de voisinage**

Des techniques de voisinage sont aussi utilisées pour la résolution des problèmes de planification, essentiellement à court terme. Les méthodes de voisinage sont souvent basées sur une recherche locale, qui tente habituellement de trouver une meilleure décision que la solution courante dans son voisinage. Un outil est ajouté de manière

itérative en raison de certains critères, et la performance est évaluée jusqu'à ce qu'il n'y a pas d'amélioration dans la fonction objectif. A ce moment, la procédure est arrêtée avec des solutions quasi-optimales. Le recuit simulé [183] et les techniques de recherche tabou [58] sont les principales techniques de recherche locale qui ont été testées sur des problème de planification de la production. Dans les deux cas, la structure de voisinage est basée sur la rectification d'un planning initial. Les inconvénients majeurs de ces techniques, dans ce contexte, sont la longueur du temps d'exécution et la non robustesse de la solution obtenue.

c) **Techniques basées sur des algorithmes**

Dans le domaine des approches algorithmiques, Horiguchi *et al.* [81] ont proposé un algorithme qui estime la date de début et de fin de chaque tâche planifiée sur chaque ressource critique : leur algorithme considère le temps disponible pour toutes les combinaisons possibles de période et de ressource critique, et il réduit le temps de production disponible à chaque fois qu'une nouvelle commande est ajoutée au planning. Cette approche, en raison du niveau élevé d'agrégation de la modélisation des ressources et des relations, pourrait conduire à des ordres qui se chevauchent sur la même ressource dans la même période de planification (*i.e.* des plannings irréalisables).

Ensuite, Leachman *et al.* [142] ont introduit une méthode de calcul des objectifs de production à court terme avec des structures de produits simples, appelée *SLIM*. L'idée principale de leur algorithme est de déterminer un *WIP* cible pour chaque machine à partir de la quantité à livrer mensuellement. La méthode commence par définir le *WIP* cible pour les machines goulots d'étranglement, puis elle s'intéresse aux postes non-goulots. En outre, la méthode propose une politique de répartition afin de suivre le profil du *WIP* cible et d'atteindre l'objectif de livraison. Bien que cette méthode permet d'établir un plan de production faisable, elle est un peu complexe et difficile à utiliser et les fabricants des semi-conducteurs préfèrent souvent utiliser des méthodes de planification plus simples basées sur des règles empiriques.

Chen *et al.* (2005) [30], Chen *et al.* (2009) [31], Chen et Chen (2010) [29], Chen *et al.* (2010) [32] et Chen *et al.* (2015) [33] ont développé des systèmes de planification de la production à capacité infinie qui tiennent compte de la qualification des équipements pour une seule *fab*, double *fabs*, de multiples *wafer fabs*, des usines d'emballage des circuits intégrés et des usines de test final des circuits intégrés, respectivement.

1.3.4.5 Les techniques d'intelligence artificielle

Le besoin de solutions rapides a incité les chercheurs à utiliser des techniques d'intelligence artificielle comme les algorithmes de recherche en faisceau, la logique floue et toute combinaison de ces techniques.

La technique recherche en faisceau est un dérivé de la méthode *branch & bound*. Elle tente d'éliminer les branches d'une manière intelligente de sorte que toutes les branches ne doivent pas être examinées. Cette technique nécessite donc moins de temps de calcul par rapport à la méthode *branch & bound*, mais elle ne peut plus garantir une solution optimale. Avec la recherche en faisceau, seuls les nœuds les plus prometteurs à tous les niveaux sont sélectionnés en tant que nœuds de début de branchement. Les nœuds restants à ce niveau sont mis au rebut de façon permanente. De et Lee [46] ont utilisé la technique de recherche en faisceau pour le développement d'un système de planification des opérations de test des semi-conducteurs. Fargher *et al.* [52] et Fargher et Smith [53] ont utilisé un algorithme de recherche en faisceau en combinaison avec des étapes de retour sur trace (*backtracking steps* en anglais) pour la libération des lots et pour la détermination des plannings de façon agrégée. Habenicht et Mönch [71] ont utilisé également un algorithme de recherche en faisceau pour déterminer les dates de début et de fin prévues pour les opérations macro d'un lot. Chua *et al.* [41] ont développé un système intelligent multi-contraintes pour le lancement des lots à capacité finie. Ce système a été conçu, développé et mis en œuvre pour résoudre les problèmes de libération des lots dans un environnement de fabrication discrète avec une énorme gamme de produits et de multiples contraintes de capacité.

La logique floue peut être utile pour la modélisation et la résolution de problèmes de planification de la production avec des temps de process et des contraintes incertains. Ces incertitudes peuvent être représentées par des nombres flous qui sont décrits en utilisant le concept d'un intervalle de confiance. Ces approches sont généralement intégrées à d'autres méthodes telles que les procédures de recherche et la relaxation de contrainte.

Wang *et al.* [172] ont proposé un modèle d'allocation des ressources en utilisant la logique floue et ils l'ont appliqué à des usines de test final des semi-conducteurs.

Azzaro-Pantel *et al.* [6] ont utilisé l'approche floue pour la modélisation de la performance d'une usine de fabrication de semi-conducteurs. Un modèle de simulation à événements discrets (MELISSA) pour l'évaluation de la performance a été développé pour traiter les incertitudes modélisées par des nombres flous. Les techniques floues permettent au décideur d'avoir une large gamme de valeurs possibles pour les dates de fin, les délais moyens de stockage, et la charge de travail.

Les techniques d'intelligence artificielle ont été développées pour résoudre des problèmes de satisfaction de contraintes. Par conséquent, nous ne pouvons pas espérer de

bonnes caractéristiques d'optimisation.

1.3.4.6 Autres classifications des techniques de planification

Les techniques de planification de la capacité de production peuvent être classées aussi en techniques de planification à capacité infinie et techniques de planification à capacité finie [150]. Il s'avère que les deux types de techniques sont importants pour la planification de la capacité de production [33].

Basé sur l'hypothèse d'une capacité illimitée des équipements, il est inutile de modifier les dates d'échéance de livraison des commandes (les *due dates*) parce qu'il est autorisé de dépasser le seuil supérieur de la capacité des équipements. La planification à capacité infinie peut être utilisée pour estimer la charge future des équipements. Par conséquent, l'estimation de la charge sert à :

- déterminer les postes goulots d'étranglement,
- équilibrer la charge des équipements,
- identifier les périodes de maintenance préventive des équipements,
- aider les directeurs d'usine et / ou les planificateurs de la production à décider du niveau de l'externalisation ou de la surcharge.

Par contre, lors d'une planification à capacité finie, le taux d'utilisation des équipements ne doit pas dépasser le seuil supérieur de leur capacité, mais les dates d'échéance des commandes peuvent être changées en fonction de la limite de la capacité des équipements. L'utilité de la planification à capacité finie réside dans l'estimation des dates réalisables d'échéance de livraison des commandes donc la définition d'un planning de production faisable.

1.3.5 Systèmes de planification de la production

La fabrication des semi-conducteurs est parmi les processus de fabrication les plus complexes et à forte intensité de capital dans le monde. Elle est très compétitive [154]. Par conséquent, les installations de fabrication des semi-conducteurs affrontent une pression croissante pour réduire les coûts, augmenter la qualité et améliorer la performance de livraison. Ils ont besoin d'outils automatiques efficaces pour transformer des milliers d'opérations en un produit final complexe.

Dès l'apparition de l'industrie des semi-conducteurs en 1971 [100], il y avait un besoin d'outils d'analyse de la capacité. Ces outils ont pour rôle d'évaluer la capacité de la ligne de production afin de satisfaire les besoins clients en termes de quantité, *mix* et volume des produits.

La planification de la capacité de production, à moyen et court terme, dans l'industrie des semi-conducteurs est réalisée en utilisant les tableurs, les systèmes de planification de la production tels que les progiciels de Gestion de Production Assistée par Ordinateur (GPAO), les progiciels de gestion intégrée (*Entreprise Resource Planning - ERP*) et les systèmes de planification avancée (*Advanced Planning System - APS*).

Ces outils comportent les modèles analytiques et les modèles de simulation présentés dans la section précédente et nécessitant un effort et une puissance de calcul intensifs.

Les tableurs

Les tableurs sont souvent utilisés pour avoir des réponses rapides à des questions ou des problèmes urgents. L'objectif principal de l'utilisation des tableurs est le calcul d'une charge estimée d'un ou plusieurs équipements en connaissant la demande. Ce calcul est nécessaire pour prendre les décisions suivantes :

- Quand et de combien l'entreprise doit investir dans l'achat de nouveaux équipements ?
- Où se situent les goulets d'étranglement potentiels dans le flux de production ?
- Est-ce que le planning réalisé est faisable *i.e.* satisfait-il les dates d'échéance de livraison des commandes clients ?

Les tableurs sont des outils de planification à capacité infinie vu que la planification réalisée se déroule en deux étapes : (1) Planification des produits basée sur un temps de cycle constant ou variable extrait de l'historique et (2) Estimation de la saturation des équipements basée sur les résultats de planification. Parmi les avantages de l'utilisation des tableurs on cite :

— Rapidité du temps de calcul

— Facilité d'utilisation et d'analyse

Cependant, les modèles statiques présentent deux inconvénients majeurs :

— L'imprécision de l'estimation de la charge des équipements calculée

— La non prise en compte de l'aspect dynamique du processus de fabrication et de la variabilité du temps de cycle *i.e.* ils considèrent un temps de cycle fixe indépendant de la capacité des équipements

Les tableurs de capacité semblent plus intéressants pour les lignes de production ayant peu de produits différents et une composition de produits stable [86].

1.1.3 Planification de la production dans l'industrie des semi-conducteurs

Les outils classiques de GPAO

Les progiciels de Gestion de Production Assistée par Ordinateur (GPAO) sont nombreux et se sont sophistiqués au cours du temps. L'objet de ces progiciels est de répondre à la plupart des fonctions assurées par la gestion de la production. La planification de production est considérée comme l'une des fonctions principales de ces systèmes. Au niveau tactique, les progiciels de GPAO permettent l'élaboration du plan directeur de production avec les calculs des besoins nets et les calculs de charge associés. Au niveau opérationnel, la planification est effectuée à capacité infinie ce qui nécessite l'intervention des gestionnaires pour remédier aux problèmes d'inadéquation entre la charge et la capacité en modifiant la répartition de la charge. L'utilisation de ces outils dans les environnements de production complexes tels que celui de l'industrie des semi-conducteurs présente des limites : (i) le besoin de la gestion d'une grande masse de données techniques et (ii) le long temps de calcul des besoins nets et de la charge de production [50].

Enterprise Resource Planning (ERP)

Enterprise Resource Planning est un système d'information intégré utilisé pour gérer la production aux niveaux stratégique, tactique et opérationnel. *ERP* comporte la fonctionnalité de la planification des besoins en composants (*MRP*). Toutefois, comme il est mentionné dans la section précédente, la méthode *MRP* présente certaines limites lorsqu'elle est appliquée à l'industrie des semi-conducteurs. Elle ne tient pas compte des contraintes de capacité et suppose des temps de cycle fixes. Ainsi, l'utilisation de la fonctionnalité *MRP* des systèmes *ERP* est relativement faible dans les *wafer fabs*. La fonctionnalité de gestion des commandes proposée par les systèmes *ERP* est souvent la fonctionnalité la plus importante dans ce contexte industriel. La deuxième fonctionnalité importante est la planification et la prévision de la demande [125].

Cependant, au niveau de la planification globale, les systèmes *ERP* sont souvent complétés par les systèmes *APS*.

Advanced Planning System (APS)

Les systèmes *APS* sont des outils d'aide à la décision pour une optimisation de la planification de la chaîne logistique à long terme, moyen terme et court terme. Ils peuvent être considérés comme des extensions de systèmes *ERP* qui sont incapables de résoudre la totalité des problèmes de prise de décision associés à une chaîne logistique. Les systèmes *APS* assurent la planification de production en utilisant des méthodes de recherche opé-

rationnelle et d'intelligence artificielle en prenant en compte les contraintes de capacité.

L'application des systèmes *APS* dans l'industrie des semi-conducteurs a montré certaines défaillances. Il est constaté que les systèmes de planification avancée ont la même performance que les techniques manuelles où les industriels prennent des décisions de planification avec des supports informatiques tels que les tableurs de calcul [110]. En plus, ils posent le problème de la robustesse de la planification.

1.4 Problématique

La problématique, que nous décrivons ici, est issue du domaine de l'industrie de semi-conducteur vu le cadre du projet européen dans lequel ces travaux de recherche ont été menés (cf. annexe A). Néanmoins, la complexité des processus de fabrication et la diversité de problématiques traitées dans ce secteur pourraient aussi s'appliquer à d'autres secteurs industriels ayant des caractéristiques semblables.

Après avoir présenté le contexte général des ateliers de fabrication des semi-conducteurs, les outils et les techniques industriels utilisés afin de gérer la planification de la production au sein de cet environnement industriel très complexe, nous avons constaté que les limites essentielles de ces techniques résident dans l'élaboration d'un planning de production réalisable. En effet, le planning établi ne tient pas compte de la capacité et la disponibilité des ressources, la priorité des lots et les dates de livraison à respecter. En plus, il y a souvent une mauvaise estimation du temps de cycle des *steps*. Ainsi, une fois que le planning de production est réalisé, il est soumis à plusieurs types de perturbations imprévisibles telles que la défaillance des équipements clés, l'annulation et l'accélération des commandes clients et des problèmes de processus imprévus ce qui modifie les temps de cycle estimés. Généralement, la gestion des modifications nécessaires du planning pour agir face aux différentes perturbations est effectuée manuellement. Cela provoque une dégradation des performances dans l'usine et entraîne également une propagation des perturbations ayant des conséquences imprévues ultérieurement. Ainsi, les problèmes détectés lors de la planification de la production des usines de fabrication de semi-conducteurs sont :

- La difficulté de la gestion manuelle des modifications de planification en termes de temps et d'énergie des ressources humaines à cause des contraintes suivantes :
  - Le nombre de critères à respecter (encours de production, capacité et disponibilité des équipements, priorité des lots etc.)
  - La forte variabilité des processus et des produits
  - La réentrance des flux
  - Le nombre de *steps* très important

— La mauvaise estimation des temps d'attente

- Les erreurs effectuées lors de la planification engendrant des retards de livraison.

Ainsi, l'objectif de ce travail est de développer de nouveaux outils et techniques permettant de gérer efficacement la forte variabilité du processus de fabrication des semi-conducteurs. En outre, l'approche proposée doit répondre à l'exigence clé des industriels des semi-conducteurs, concernant le calcul rapide des plans de production réalisables (en cinq minutes au plus sur un ordinateur personnel) afin de faciliter l'analyse "*what-if*".

Pour la planification de la capacité de production, on révèle les principaux défis suivants :

- l'identification précoce des goulots d'étranglement de la ligne de production
- la validation du plan de commandes et de l'engagement aux dates d'échéance des livraisons aux clients
- la détermination du temps de cycle avec des lots circulant à des vitesses différentes
- l'élaboration des politiques de répartition et des règles de planification pour l'équilibrage de la ligne de production

Pour répondre à ces questions, trois notions et outils différents sont impliqués :

1. **Planification des capacités** : Quelle est la capacité nécessaire pour répondre à la demande ?
2. **Planification de la production** : Quel est le plan de production le plus proche de la réalité en connaissant la demande et la capacité installée ?
3. **Projection des encours de production (*Work In Progress* ou *WIP* en anglais)** : Quelle est l'activité nécessaire pour assurer la livraison du *WIP* ?

Bien que la planification à capacité infinie permette généralement de répondre à (1), pour résoudre (2) et (3), la relation entre la limitation des capacités et le temps de cycle doit être prise en compte vu que les goulots d'étranglement augmentent le temps de cycle et les produits différés créent des conflits de priorité entre les encours de production.

1.5 Conclusion

Dans ce chapitre, nous avons présenté les bases nécessaires à notre étude. La présentation du contexte dans lequel se fabriquent les puces, permet de mieux appréhender les contraintes de fabrication entraînant une complexité de la planification de la capacité de production.

Les différentes techniques de planification de la production, présentées dans ce chapitre, ont des caractéristiques qui déterminent les problèmes auxquels elles peuvent être

appliquées plus efficacement. Les limites de ces techniques sont dues à un certain nombre de raisons telles que le temps de calcul et la qualité des solutions obtenues.

Nous avons présenté aussi les outils de gestion de production utilisés par les industriels permettant la planification de la capacité de production *i.e.* les tableurs, les outils de GPAO et les progiciels *ERP* et *APS*. Ces outils comportent soit des techniques classiques *i.e.* la méthode *MRP* soit des modèles analytiques ou/et des modèles de simulation. L'utilisation de ces modèles, dans l'industrie des semi-conducteurs, a démontré certaines limites. Partant de ce constat, l'objectif de ce travail est de développer des techniques de planification de la production à capacité finie permettant d'établir un plan de production réalisable en tenant compte des dates d'échéance des livraisons aux clients et des particularités du processus de fabrication des semi-conducteurs.

Ainsi, dans ce premier chapitre, nous avons répondu aux quatre questions :

1. Quelles sont les caractéristiques du processus de fabrication des semi-conducteurs ?
2. Pourquoi la planification d'un système de fabrication des semi-conducteurs est très complexe ?
3. Quels sont les outils et les techniques de planification les plus utilisés par les industriels des semi-conducteurs ? Quels sont leurs avantages et leurs inconvénients ?
4. Quel est l'objectif principal de ce travail ?

La revue de la littérature concernant ce problème de planification de la production appliquée à l'industrie des semi-conducteurs et un positionnement du problème par rapport à cette littérature ainsi qu'une description des contributions de ce travail seront présentés dans le chapitre suivant.

2

Planification de la fabrication des semi-conducteurs : État de l'art

Résumé : Ce chapitre est dévolu à retracer l'état de l'art des principaux problèmes de planification de la production rencontrés dans l'industrie des semi-conducteurs ainsi que les techniques de planification employées pour résoudre ces problèmes. Un autre état de l'art sur les techniques de modélisation du temps de cycle (le temps prévu écoulé depuis le début jusqu'à la fin d'un processus de production) est présenté dans le chapitre 4.

Nous commençons ce chapitre par la méthodologie de recherche établie. Ensuite, nous présentons les différents problèmes et techniques de planification existants dans la littérature. Enfin, nous effectuons une synthèse des études antérieures nous permettant de positionner nos travaux de recherche. Une analyse des écarts observés permet de dégager deux axes de travail pour cette thèse, qui sont des problèmes fréquemment rencontrés dans l'industrie des semi-conducteurs mais peu abordés dans la littérature : la variabilité des temps de cycle des *steps* en fonction des dates d'échéance de livraison des lots, et la prise en compte des contraintes de capacité et de qualification des parcs d'équipements.

Sommaire

2.1 Introduction 39
2.2 Méthodologie et classification des problèmes 39
2.3 Techniques de planification existantes et leurs limites 45
2.3.1 Planification stratégique 45
2.3.2 Planification tactique et opérationnelle 46
2.4 Positionnement de notre problématique 51
2.5 Conclusion 52

2.1 Introduction

Dans le chapitre précédent, nous avons montré la complexité du processus de fabrication des semi-conducteurs ainsi que l'importance et la difficulté de la planification de la production au sein de cet environnement industriel. Nous avons présenté les limites des outils et techniques industriels utilisés pour la planification de la production, en général. L'objectif de ce chapitre est donc non seulement de donner un aperçu sur les problèmes et les techniques de planification de la fabrication des semi-conducteurs, rencontrés dans la littérature, mais aussi d'identifier les écarts entre les différents travaux et de bien positionner notre étude par rapport à la littérature existante.

Afin de comparer les problèmes étudiés dans la littérature, nous définissons, dans la section suivante 2.2, la méthodologie de recherche et les critères utilisés pour la classification des travaux antérieurs. Puis, nous présentons, dans la section 2.3, l'état de l'art sur les techniques de planification de la production appliquée à l'industrie des semi-conducteurs. Enfin, dans la section 2.4, une analyse critique de la littérature est présentée afin d'affiner la problématique traitée et de faire ressortir les contributions de notre travail.

2.2 Méthodologie et classification des problèmes

Nous avons effectué notre recherche en utilisant différentes combinaisons de mots clés suivants : *Production planning, capacity planning, WIP projection, cycle time variability, due dates, minimize tardiness, high mix low volume production, reentrant job shops, semiconductor manufacturing, wafer fabrication, decision making, finite capacity, infinite capacity, algorithms, heuristics.*

Globalement, 130 articles, provenant de différentes sources détaillées, dans le tableau 2.1, ont été consultés.

Tableau 2.1 – Type de la source pour les différents articles de la revue de littérature

Source	Nombre d'articles
Journal	83
Conférence	35
Mémoire de thèse	7
Mémoire de master	1
Rapport technique	2
Livre	2

Les titres des revues dans lesquelles les 83 articles ont été publiés sont résumés dans le tableau 2.2. La figure 2.1 donne un aperçu sur les auteurs qui ont contribué de manière significative dans ce domaine, avec leur localisation géographique.

Tableau 2.2 – Source des articles de revues

Titre du journal	Nombre d'articles
IEEE Transactions on Semiconductor Manufacturing	10
International Journal of Production Research	10
IIE Transactions	8
European Journal of Operational Research	7
Naval Research Logistics	4
Production Planning & Control	4
Computers & Industrial Engineering	3
Interfaces	3
International Journal of Advanced Manufacturing Technology	3
Journal of Manufacturing Systems	3
Journal of the Operational Research Society	3
Computers & Operations Research	2
IEEE Transactions on Automation Science and Engineering	2
Journal of Scheduling	2
Annals of Operations Research	1
AT&T Technical Journal	1
European Journal of Industrial Engineering	1
IEEE Transactions on Automatic Control	1
International Journal of Computer and Communication Engineering	1
International Journal of Computer Integrated Manufacturing	1
International Journal of Production Economics	1
International Transactions In Operational Research	1
Journal of the Chinese Institute of Industrial Engineers	1
Journal of Industrial Engineering and Management	1
Management science	1
Mathematical and Computer Modelling	1
Operations Research	1
Queueing systems	1
Robotics and Computer-Integrated Manufacturing	1
Semiconductor International	1
SEMATECH Technology Transfer	1
Semicon/West Technical Program	1
VLSI Design	1

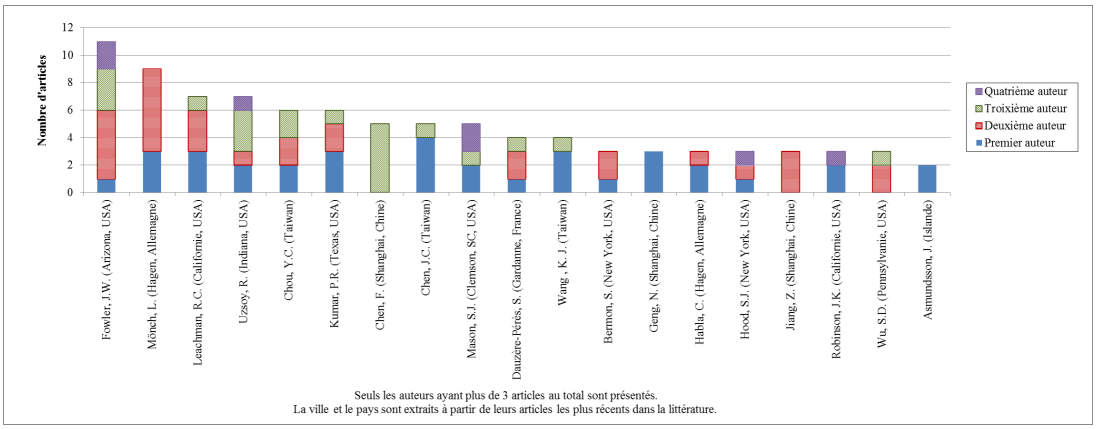

Figure 2.1 – Principaux auteurs.

Historiquement, comme le montre la figure 2.2, l'intérêt pour le problème de la planification de la production dans l'industrie de semi-conducteurs est apparu à la fin des

années 1980.

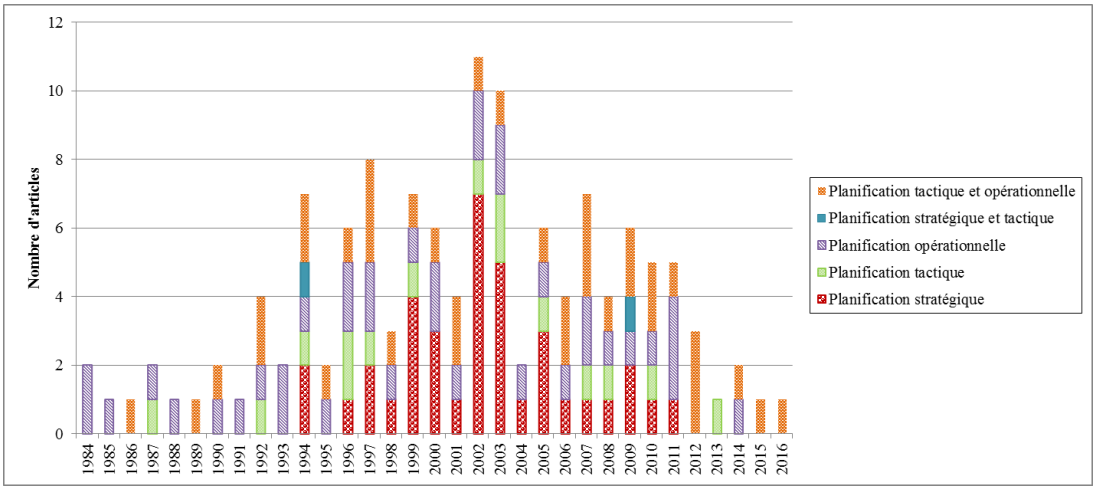

Figure 2.2 – Nombre d'articles par année classés selon le niveau de décision de la planification.

Les premiers travaux traitant ce problème sont effectués par Lohrasbpour et Sathaye (1984) [113], Dayhoff et Atherton (1984) [45], Resende (1985) [140], Burman *et al.* (1986) [20], Glassey et Resende (1988) [62] et Wein (1988) [175]. Ces auteurs ont recours à la simulation à évènements discrets pour déterminer la capacité de production et la sensibilité de cette dernière à des ressources supplémentaires, aux règles opérationnelles et aux caractéristiques du processus de fabrication. Le domaine de recherche de la planification de la production dans l'industrie de semi-conducteurs a attiré de plus en plus l'attention des chercheurs, voir les survey de Uzsoy *et al.* [167, 168] en 1992 et 1994 et Mönch *et al.* [125] en 2013. Les travaux existants peuvent être classés selon différents critères. Parmi lesquels, on cite :

— **Les niveaux de décision** : stratégique, tactique ou opérationnel (cf. figure 2.2 et figure 2.3).

  Certains travaux traitent l'intégration de deux niveaux de décision. La figure 2.3 illustre la répartition des articles selon le niveau de décision lors de la planification. Elle montre que la plupart des travaux existants s'intéressent à la planification tactique opérationnelle. L'intégration de ces deux niveaux de décisions a attiré aussi l'attention de plusieurs auteurs. Ils se sont intéressés, par exemple, aux problèmes d'affectation des ressources, libération des lots, ordonnancement, règles de répartition ([90], [99], [175], [148]). Dans notre étude, nous nous intéressons à la planification de la production au niveau tactique qui est peu étudiée dans la littérature. (cf. figure 2.3).

— **Le type d'approche** : déterministe vs. stochastique.

  Les modèles de planification de la production peuvent être classés aussi en deux catégories principales. La première catégorie correspond aux modèles stochastiques,

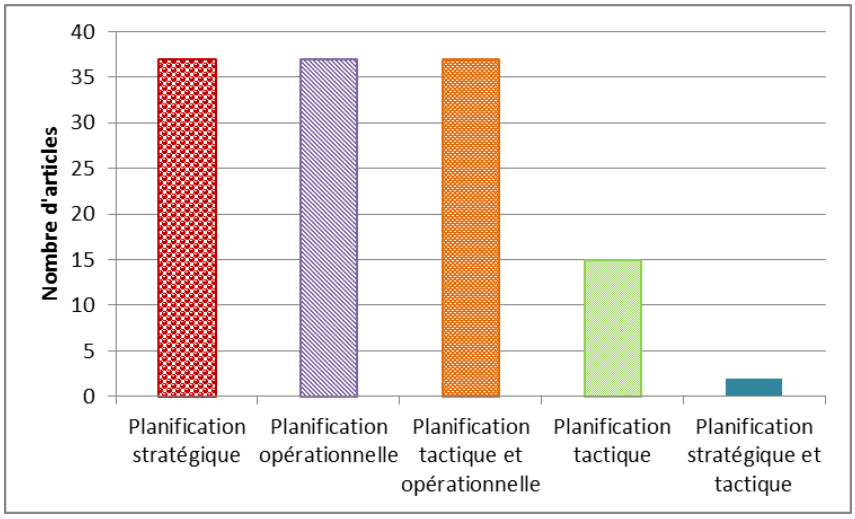

Figure 2.3 – Répartition des articles selon les niveaux de décision.

qui caractérisent la performance estimée d'un système, tout en se basant sur un modèle probabiliste de sa dynamique.

L'autre catégorie de la littérature correspond aux techniques déterministes qui divisent l'horizon de planification en périodes discrètes. Ils attribuent alors la capacité de chaque période aux produits tout en satisfaisant un ensemble de contraintes représentant la capacité et la dynamique du système à un niveau agrégé. Cependant, il peut être impossible d'exécuter une solution répondant à ces contraintes agrégées à cause de la dynamique opérationnelle, en particulier les temps de cycle qui ne sont pas modélisés explicitement.

Dans notre revue de littérature, la majorité des travaux existants ont recours à des approches déterministes (90 papiers *i.e.* environ 75%). Nous citons, par exemple, les travaux de Leachman [84, 97, 103, 104, 142], Bang et Kim [9], Bermon et Hood [15], Chen *et al.* [36], Denton *et al.* [47], Habla et Mönch [72] et Habla *et al.* [73], .

Certains auteurs ont considéré l'aspect incertain de la demande et différents scénarios potentiels de la fabrication des semi-conducteurs (pannes des équipements, urgence de livraison,...) lors du développement des méthodes de résolution pour des problèmes de planification de la production. Nous citons, par exemple, les travaux de Barahona *et al.* [10], Geng *et al.* [61], Hood *et al.* [79], Lin *et al.* [111], Swaminathan [158].

— **Le type de la demande**

Il existe plusieurs types de demandes qui peuvent être réparties selon deux groupes :

- Demandes *constantes*, les valeurs des demandes ne changent pas sur l'horizon de temps, ou demandes *dynamiques*, les valeurs varient au cours du temps.
- Demandes *certaines*, les valeurs sont connues à l'avance, ou demandes *stochas-*

tiques, les valeurs sont basées sur des probabilités.

— **La méthode de résolution** : exacte vs. approchée

Les techniques de la recherche opérationnelle sont très employées pour la planification de la production, appliquée à l'industrie des semi-conducteurs. Ces techniques peuvent être réparties en méthodes exactes et méthodes approchées.

La programmation linéaire (PL) et la programmation linéaire entière mixte (MIP), des formes typiques des méthodes exactes, sont généralement utilisées pour modéliser les problèmes de planification de la production dans l'industrie des semi-conducteurs. La PL et la MIP sont utilisées pour formuler des contraintes et une fonction objectif afin d'optimiser une prise de décision. En plus de la PL et la MIP, certains auteurs ont recours à la programmation non linéaire pour modéliser des problèmes de planification de la production dans l'industrie micro-électronique. Il s'agit de la recherche de l'optimum d'une fonction non linéaire sur un sous-ensemble convexe ou non, d'un espace donné. On trouve aussi la programmation stochastique pour formuler des problèmes de planification de la production sous incertitude de la demande ou la capacité. La décomposition de Benders est aussi utilisée comme méthode exacte pour la résolution des problèmes de planification dans ce contexte.

Cependant, ces formulations peuvent être très volumineuses pour les environnements de production complexes tels que celui de l'industrie des semi-conducteurs. Sullivan et Fordyce [157] soulignent que les PLs peuvent nécessiter un temps très long pour générer les fichiers de données d'entrée qui doivent être introduits dans le logiciel de planification mathématique, et demander aussi d'énormes volumes de mémoire et d'espace disque pour stocker ces données. Ainsi, pour simplifier les problèmes, plusieurs méthodes approchées *i.e.* des heuristiques et des méta-heuristiques sont développées. Parmi les heuristiques utilisées, nous trouvons les méthodes de décomposition telles que les algorithmes de recherche en faisceau, les algorithmes itératifs, les heuristiques gloutonnes, les heuristiques basées sur des programmes linéaires et la relaxation Lagrangienne ou autres heuristiques spécifiques. Les algorithmes génétiques et quelques techniques de recherche locale sont les méta-heuristiques les plus employées pour la résolution des problèmes de planification de la production appliquée à l'industrie des semi-conducteurs.

Le tableau 2.3 donne un aperçu sur les méthodes de résolution utilisées dans les différents articles. On peut noter qu'un article peut utiliser plusieurs méthodes de résolution, donc quelques références apparaissent à plusieurs reprises dans le tableau.

— **La prise en compte des contraintes de capacité** : planification à capacité infinie vs. planification à capacité finie.

Tableau 2.3 – Classification des articles selon la méthode de résolution

	Méthode de résolution	Articles utilisant cette méthode
Méthodes exactes	Programmation linéaire	[4], [9], [12], [15], [24], [25], [36], [31], [38], [55], [72], [73], [84], [85], [91] [97], [104], [105], [122], [181]
	Programmation non linéaire	[5], [11], [19], [173], [174]
	Programmation stochastique	[10], [40], [59], [61], [79], [83], [94], [95], [151], [153], [156], [158], [159], [166], [173], [174]
	Méthodes de décomposition (Benders)	[12]
Heuristiques	Heuristiques de décomposition	[123]
	Algorithmes de recherche en faisceau	[52], [53], [71]
	Heuristiques gloutonnes	[11], [48], [49]
	Algorithmes itératifs	[26], [44]
	Heuristiques basées sur des programmes linéaires	[10], [21], [38], [55], [73], [97], [103], [105], [122]
	Heuristiques basées sur la relaxation Lagrangienne	[12], [24], [25], [48], [73], [158]
	Autres heuristiques	[22], [30], [32], [33], [41], [66], [81], [142], [115], [116]
Méta-heuristiques	Algorithmes génétiques	[8], [82], [128], [173], [174], [179]
	Recherche locale	[11]

Dans notre revue de la littérature, la considération des contraintes de capacité, dans l'industrie micro-électronique, a attiré l'attention de plusieurs auteurs. En effet, environ 90% des études existantes traitent des problèmes de planification de la production à capacité finie. Nous citons, par exemple les travaux de Barahona *et al.* [10], Bermon et Hood [15], Habla et Mönch [72], Habla *et al.* [73], Horiguchi *et al.* [81], Leachman et Carmon [104], Thompson [163]. Quelques travaux se sont intéressés à la planification de la production à capacité infinie tels que les travaux de Chen *et al.* [30–32].

2.3 Techniques de planification existantes et leurs limites

Dans cette section, les différentes techniques de planification, développées dans des études antérieures, sont reportées en expliquant leurs avantages et leurs limites vis à vis de notre objectif en utilisant les critères de classification définis dans la section précédente.

2.3.1 Planification stratégique

La planification stratégique se réfère à la détermination de la séquence et du calendrier de l'achat ou l'élimination des équipements ce qui est un processus de prise de décision multi-critères impliquant des compromis entre le coût, le débit, le temps de cycle et le risque. Elle est aussi nommée plan d'expansion de la capacité, plan d'achat d'outils, ou plan d'un portefeuille d'outils ou de ressources. La recherche s'intéressant à la planification stratégique appliquée à l'industrie des semi-conducteurs est apparue depuis une vingtaine d'années (cf. figure 2.2) et elle est devenue un enjeu très important pour les investissements de capacité de production. Une étude détaillée des méthodes de planification stratégique de la capacité dans la fabrication des semi-conducteurs est présentée par Geng et Zhang [60]. De nombreux travaux appliqués à la fabrication de semi-conducteur appartiennent à cette catégorie de planification. Bermon et Hood [15] et Çatay *et al.* [25] ont développé un MIP déterministe pour une prise de décision stratégique concernant l'acquisition de nouveaux équipements. Afin de déterminer la configuration optimale d'un parc d'équipements, Bard *et al.* [11] ont proposé un modèle de programmation non linéaire en nombres entiers qui minimise le temps de cycle moyen soumis à des contraintes budgétaires, une exigence de débit de production, un *mix* produit et une technologie de processus de fabrication. Barahona *et al.* [10], Christie et Wu [40], Hood *et al.* [79], Ahmed [151], Swaminathan [158, 159] et Wu *et al.* [179] ont pris en compte aussi de l'incertitude de la demande dans les problèmes d'expansion de capacité.

Wang et Lin [171], Wang *et al.* [173, 174] et Yang *et al.* [182] ont étudié les problèmes de répartition et d'expansion des capacités avec des investissements en équipements pour une installation de test de semi-conducteurs.

Chou et You [39] ont proposé une méthodologie pour un plan d'expansion de la capacité, qui a trois composantes principales : un meilleur modèle statique de la capacité, un modèle de file d'attente, et une procédure d'ajustement du portefeuille. Karabuk et Wu [95] présentent une programmation stochastique fondée sur des scénarios pour le problème de la planification stratégique dans plusieurs usines de fabrication de semi-conducteurs en considérant les incertitudes de la demande et de la capacité.

Dans notre étude, nous ne nous intéressons pas à ce niveau de planification de la production.

2.3.2 Planification tactique et opérationnelle

Les techniques de planification tactique et opérationnelle sont regroupées en fonction de leur fonctionnalité principale d'optimisation, simulation et théorie de la file d'attente, tout en définissant les différences fondamentales dans les approches de modélisation. L'optimisation se réfère à la programmation linéaire, les méthodes de décomposition, les heuristiques et la programmation stochastique. La simulation se réfère à la dynamique du système, l'analyse des scénarios, l'analyse de la sensibilité et l'analyse probabiliste des risques. La théorie des files d'attente se réfère à l'analyse de la performance du système.

2.3.2.1 Techniques basées sur l'optimisation

L'optimisation se réfère à l'objectif de minimiser ou maximiser une fonction soumise à des contraintes données. L'optimisation est généralement présentée sous forme déterministe telle que la programmation linéaire, les méthodes de décomposition et les heuristiques. Dans la littérature, on trouve aussi la forme stochastique de l'optimisation mais elle ne présente pas un intérêt à notre étude vu que nous considérons une approche déterministe.

La programmation linéaire

La PL est largement appliquée à des problèmes spécifiques, rencontrés dans la planification des capacités pour l'industrie des semi-conducteurs. Elle est devenue le principal outil d'aide à la décision pour la planification des capacités [79]. Elle est très utilisée pour résoudre les problèmes de planification à capacité finie à moyen et court terme. La grande majorité des modèles de programmation linéaire dans notre centre d'intérêt, abordent le problème de la même manière. L'horizon temporel considéré est divisé en périodes de temps discrètes, généralement, de la même longueur. Les variables de décision sont associées à chaque période, la capacité est considérée comme une borne supérieure fixe représentant la quantité de produits qui peut être consommée par une ressource, pendant une période de temps, et l'objectif est généralement de réduire au minimum le coût total.

Leachman a présenté les premiers travaux de PL sur les systèmes de planification de la production appliquée à l'industrie des semi-conducteurs (Dessouky et Leachman [48], Hackman et Leachman [74], Hung et Leachman [84], Leachman et Carmon [104]).

Leachman et Carmon [104] ont proposé un modèle de planification de la capacité, en utilisant la PL qui implique un ensemble de contraintes approximatives et utilise une

hypothèse de proportionnalité pour les temps de process.

Leachman [102] a utilisé la PL pour affecter la capacité de production dans le but de maximiser le profit en tenant compte de la qualification des équipements. Il a proposé un cadre de programmation linéaire à grande échelle pour la planification de la production dans l'industrie des semi-conducteurs où les stocks de sécurité sont abordés en utilisant des classes de demande. La production nécessaire pour remplacer les stocks de sécurité est modélisée comme une classe de la demande qui a une priorité plus faible que les commandes fermes des clients. La capacité est allouée à cette classe de demande sauf si cela ne compromet pas la capacité du système à répondre aux commandes fermes.

Leachman *et al.* [103] a développé un système de planification de la production basé sur l'optimisation au niveau de l'entreprise qui comprend de multiples installations et intègre les processus de fabrication dans ces installations. Son modèle génère des calendriers réalisables, de début et de fin de production, en considérant les contraintes de capacité pour chaque usine de fabrication dans l'entreprise. Le moteur de planification de ce système intègre des techniques de formulation pour la planification des exigences de produits, pour représenter la consommation de la capacité dynamique par les flux de processus rentrants, et pour développer de multiples calculs d'optimisation qui reflètent les priorités de commercialisation.

Dessouky et Leachman [48] ont développé deux formulations de programmation en nombres entiers pour la planification de production en tenant compte des caractéristiques de la fabrication de semi-conducteurs à volume élevé telles que les flux ré-entrants, les lots semblables et les machines identiques. Leurs approches sont basées sur la limitation du domaine autorisé des événements pour le début du traitement du lot.

Chou et Hong [38] ont proposé un MIP pour les décisions du mix produit dans une *wafer fab* et ont développé une procédure fondée sur le process des postes goulots d'étranglement.

Bermon *et al.* [14] ont introduit aussi un modèle de PL pour résoudre le problème de planification de la production en considérant une demande déterministe. Ils ont développé un outil d'aide à la décision qui permet d'identifier les équipements les plus importants dans l'usine de fabrication.

Chen [34] a présenté un modèle de PL floue pour la planification de la production mensuelle dans une *wafer fab*. Basé sur un modèle mixte de programmation en nombres entiers, Hwang et Chang [85] ont proposé un système de planification et d'ordonnancement hiérarchique à deux niveaux pour une usine de fabrication de semi-conducteurs. Ce système est composé d'un module de planification à moyen terme et un ordonnanceur à court terme.

Habla *et al.* [73] ont proposé une formulation MIP pour déterminer les dates de fin des

lots dans les postes goulots d'étranglement. D'autres formulations de PL ont été proposées par Habla *et al.* [72] pour trouver des quantités optimales de production, sur un horizon de planification de la production à moyen-terme, en tenant compte de la réalisation de la demande, les restrictions de capacité, et l'état actuel des encours de production (*WIP*). Leur objectif était de maximiser les revenus.

Chen *et al.* [27] ont présenté un modèle mathématique pour aider le gestionnaire de production, dans un environnement de production basée sur la gestion à la commande (*make to order*), à sélectionner un ensemble de commandes des clients potentiels. L'objectif de ce modèle est de maximiser le profit opérationnel de telle sorte que tous les ordres sélectionnés sont achevés à leurs dates d'échéance. Le modèle proposé considère les délais habituels, les heures supplémentaires, et la sous-traitance comme sources d'extension de la capacité pour chaque type de ressource. L'expérimentation de ce modèle montre un contraste entre l'utilisation de la capacité maximale et le profit opérationnel optimal.

Chen *et al.* [36] ont proposé une stratégie de planification de la capacité de production à moyen terme, basée sur un programme linéaire, pour augmenter le nombre d'outils auxiliaires afin d'augmenter la flexibilité de la production. Les décisions considérées concernent la façon de répartir adéquatement les demandes de prévisions de produits entre plusieurs sites et comment décider sur les quantités de produits dans chaque site de production après avoir reçu des commandes confirmées par le client.

La PL est parfois combinée avec la simulation à événements discrets pour résoudre les problèmes de planification à capacité finie à moyen terme. Le travail de Hung et Leachman [84] est un exemple d'une telle approche. Étant donné des estimations initiales des *lead-times*, un modèle de PL pour la planification de la production est formulé et résolu. Le plan résultant est introduit dans un simulateur pour estimer les temps de cycle que le plan imposerait à un système réel. Si ces temps de cycle ne coïncident pas avec les *lead-times* utilisés dans le PL, le PL est résolu avec les nouvelles estimations de *lead-times*. Cependant, les propriétés de convergence de ces méthodes ne sont pas bien comprises [89], et la nécessité d'un modèle de simulation détaillé est problématique pour les grandes installations. A partir de ces études, on remarque que les limitations de la PL résident dans :

1. la difficulté de sélectionner une fonction objective appropriée dans la planification de la fabrication des semi-conducteurs [65].
2. la formulation détaillée du problème d'une manière intégrée, basée sur un modèle de programmation mathématique, qui exige des données, impossible à obtenir de manière fiable.
3. la nécessité de ressources informatiques considérables pour satisfaire le grand nombre de contraintes dans la planification des capacités.

Ainsi, en raison de l'importance du temps nécessaire pour générer des données d'entrée et du besoin de mémoire et d'espace disque pour stocker les données comme le souligne Sullivan et Fordyce [157], la PL est généralement combinée avec des méthodes de décomposition ou avec la relaxation lagrangienne pour réduire le temps d'exécution.

Les méthodes de décomposition

Ces méthodes se réfèrent à la décomposition d'un grand problème complexe en nombreux petits sous-problèmes solvables, réduisant ainsi le temps de calcul. Les solutions sont développées pour chacun des sous-problèmes individuellement, puis rassemblées pour constituer une solution du problème initial [131]. Parmi les méthodes de décomposition les plus utilisées, la décomposition de Benders. Cet algorithme génère des contraintes au fur et à mesure de sa progression vers la solution. Elle est utilisée pour la résolution du problème de planification à moyen terme considéré par Bard et *al.* [12] pour minimiser la somme des écarts entre l'objectif en termes de quantités à livrer et les stocks de produits finis. Uzsoy *et al.* [169] ont proposé des algorithmes de planification de la production pour l'ordonnancement des opérations de test de semi-conducteurs par décomposition du problème en un certain nombre de postes de travail et en utilisant une représentation en graphes disjonctifs. Ils ont développé ces algorithmes pour minimiser le retard maximum et le nombre de jobs tardifs.

En plus de la réduction du temps de calcul, la méthode de décomposition a d'autres avantages par rapport à la PL. En effet, dans cette approche, l'intégration des différents niveaux de planification est possible. Son avantage réside aussi dans son efficacité et sa capacité à gérer les non-linéarités.

Les heuristiques

En raison de la limite de résolution des méthodes exactes, les méthodes approximatives ont été largement utilisées pour développer soit des systèmes de planification à capacité infinie ou finie pour l'industrie des semi-conducteurs. Les systèmes de planification à capacité infinie sont développés pour estimer la charge future de l'équipement afin d'identifier les goulots d'étranglement et d'équilibrer la charge de chaque équipement de production sur tout l'horizon de planification [30, 32, 33]. Sachant l'importance des contraintes de capacité, de nombreux auteurs ont développé des systèmes de planification à capacité finie. Faragher *et al.* [52] ont utilisé un algorithme de recherche en faisceau en combinaison avec des étapes de retour sur trace pour la libération des lots et pour la détermination d'un planning dans un sens agrégé. Horiguchi *et al.* [81] ont proposé un algorithme de planification à capacité finie simple, très similaire à l'algorithme *capacitated MRP (MRPC)* de Tardif et Spearman [162]. L'objectif de cet algorithme est de calculer une date de

lancement pour chaque commande à chaque poste goulot d'étranglement et d'estimer sa date de fin. Le modèle considère explicitement la capacité uniquement pour des postes goulots d'étranglement spécifiés, et suppose que tous les autres postes ont une capacité infinie, ce qui est différent de l'approche *MRP* classique. Habenicht et Mönch [71] ont proposé un algorithme de recherche en faisceau pour déterminer les dates de début et d'achèvement prévues pour les opérations macro d'un lot. Chua *et al.* [41] ont développé un système intelligent multi-contrainte pour la libération des lots à capacité finie. Ce système a été conçu, développé et mis en œuvre pour résoudre les problèmes de libération des lots dans un environnement de fabrication discret avec une grande gamme de produits et de multiples contraintes de capacité.

2.3.2.2 Techniques basées sur la simulation

L'optimisation par son algorithme de recherche d'un objectif est de nature normative. Une approche plus descriptive et exploratoire est réalisée par simulation. Les outils de simulation sont mis en œuvre par un ensemble différent d'objectifs : ne pas trouver la solution optimale, mais expérimenter avec des valeurs différentes. La simulation peut être réalisée manuellement ou de manière automatique par un ordinateur.

La simulation à événements discrets est souvent utilisée pour la prise des décisions de planification des capacités dans les *wafer fabs* [140]. Elle permet de répondre aux questions « *what-if* » en créant des scénarios avec des conditions modifiées et en les comparant à la sortie de la simulation avec un scénario d'origine. L'objectif principal de cette technique consiste à calculer une valeur du temps de cycle du processus de fabrication la plus réaliste possible [56, 68, 163, 165]. En effet, la simulation à événements discrets est considérée comme la seule méthode pratique permettant de prévoir le temps de cycle *i.e* le temps prévu écoulé depuis le début jusqu'à la fin d'un processus de production en fonction de la disponibilité des ressources et la vitesse de production. Le modèle de simulation peut également être utilisé pour déterminer les goulots d'étranglement dans le cadre d'un mix produit donné [137]. Cependant, les modèles de simulation utilisés pour la planification des capacités dans l'industrie des semi-conducteurs présentent des limitations sévères. Le développement, l'exécution et l'analyse des modèles nécessitent un temps de calcul assez important à cause du volume et de la complexité des données requises pour les modèles concernés. De plus, généralement, la simulation ne considère pas une fonction « objectif » [86, 139]. Elle n'est pas un outil d'optimisation.

2.3.2.3 Techniques basées sur la théorie des files d'attente

Concernant les modèles de réseaux de files d'attente, Shanthikumar *et al.* [152] ont effectué une étude sur les différentes applications de la théorie des files d'attente pour les

systèmes de fabrication de semi-conducteurs. Ils ont admis que malgré la rapidité de leur développement et de leur exécution par rapport aux modèles de simulation, la précision des modèles de files d'attente classiques n'est pas satisfaisante due à la complexité du processus de fabrication de semi-conducteurs. En plus, Ignizio et Garrido [86] affirment que les modèles de file d'attente fournissent des prévisions seulement sur l'état d'équilibre de l'installation tandis qu'ils ignorent « les hauts et les bas » rencontrés avant d'atteindre cet état. Ils déclarent aussi que les modèles de files d'attente ont été souvent développés pour des petites usines donc leur utilisation pour des installations aussi complexes et dynamiques telles que les *wafer fabs* nécessiterait beaucoup d'effort.

2.4 Positionnement de notre problématique

Dans cette étude, un problème de planification à capacité finie à moyen terme est considéré en supposant que la demande et la capacité, à chaque période de l'horizon de planification, sont des données du problème. Dans le tableau 2.4, une synthèse des études, portant sur le même problème et utilisant la PL et les techniques heuristiques comme méthodes de résolution, est présentée. Pour chaque étude, les différents objectifs et contraintes algorithmiques et opérationnelles considérés sont répertoriés. Comme on peut le remarquer, même si certains modèles proposés considèrent simultanément les contraintes de capacité, les dates d'échéance des lots et la variabilité des temps de cycle, ils ne considèrent pas le même objectif.

Dans notre approche, l'objectif est de minimiser la somme des retards pondérés *TWT*. Les essais expérimentaux ont été réalisés sur une étude de cas réelle. Les principales questions traitées dans les études existantes sont généralement limitées aux règles de répartition et aux politiques de contrôle ce qui est une portée plus restreinte que celle de notre travail de recherche. En outre, une exigence clé des industriels des semi-conducteurs est de permettre un calcul rapide des plans de production possibles (une analyse *"what-if"*). Notre objectif est donc de proposer un outil d'aide à la décision qui répond à cette exigence.

Dans cette étude, le problème de projection des encours de production (*WIP*) est considéré. Il s'agit de prévoir l'évolution du *WIP* le long du processus de fabrication *i.e.* estimer les dates de début, les temps d'attente et les dates de fin des différents *steps* restants du *WIP* ainsi que la charge accumulée sur les parcs équipements sur chaque période de l'horizon de planification.

Dans la littérature, il existe peu d'études envisageant le problème de projection du *WIP* dans l'industrie des semi-conducteurs [66, 97, 105]. Dans ces travaux, les auteurs considèrent différents objectifs et ne prennent pas en compte toutes les contraintes citées.

Kim et Leachman [97] ont proposé une formulation en programme linéaire pour le problème de projection du WIP. Ils ont résolu le PL de grande taille obtenu par une méthode heuristique de décomposition afin de déterminer la demande et les capacités des ressources. Ils ont testé leurs approches en utilisant des données aléatoires. Lee *et al.* [105] ont employé des techniques de programmation linéaire déterministe pour le problème de projection du *WIP* dans l'usine de *wafers*, en considérant explicitement des temps de cycle variables. Govind et Fronckowiak [66] ont considéré le problème de projection du *WIP* pour mesurer la performance de la production de l'usine de fabrication des *wafers* de diamètre 300mm à IBM en calculant la productivité et des objectifs du *WIP* à capacité infinie.

Comme on peut le voir, même si certains documents abordent des problèmes de planification similaires, aucun des modèles déjà proposés ne traite explicitement notre problème spécifique (cf. section 1.4).

2.5 Conclusion

Dans ce chapitre nous présentons un état de l'art des travaux traitant le problème de planification de la capacité de production dans l'industrie des semi-conducteurs. A partir de cette revue bibliographique, nous avons noté d'une part, l'absence des travaux concernant la planification à capacité finie dans un contexte industriel à forte variabilité et faible volume de production ; d'autre part, la rareté des travaux intégrant les niveaux de décisions tactique et opérationnel en tenant compte des délais de livraison et des contraintes de capacité. En outre, notre approche doit répondre à l'exigence clé des industriels des semi-conducteurs, concernant le calcul rapide des plans de production réalisables (en cinq minutes au plus sur un ordinateur personnel) afin de faciliter l'analyse "*what-if*". Dans les chapitres suivants, nous présentons nos approches de résolution pour le problème considéré. Nous formalisons d'abord dans le chapitre 3 notre problème en un programme linéaire mixte et nous déterminons ses limites de résolution. Ensuite, nous proposons dans les chapitres 4 et 5 des heuristiques permettant d'obtenir un plan de production réalisable dans un temps d'exécution raisonnable.

Tableau 2.4 – Taxonomie des approches de planification de la production à capacité finie appliquée à l'industrie des semi-conducteurs, extraites de la littérature.

Approche	Référence	Objectifs		Contraintes et hypothèses				Test
		Algorithmique	Opérationnel	Contraintes de capacité	Contraintes de qualification	Due dates	Temps de cycle variables	
Programmation linéaire	Hung and Leachman [84], Bermon *et al.* [14]	Maximiser le profit	Déterminer les quantités de *wafers* à lancer	✓			✓	Étude de cas réelle
	Leachman [102]	Maximiser le profit	Générer des plannings réalisables	✓	✓	✓	✓	Étude de cas réelle
	Habla *et al.* [73]	Minimiser le retard total pondéré	Déterminer les dates de fin objectifs des *steps* goulots d'étranglement	✓		✓		Exemple
Algorithmes /heuristiques	Fargher *et al.* [52]	Réduire le temps de cycle et la variance du temps de cycle	Déterminer the work to release into the factory at any time	✓		✓	✓	Étude de cas réelle
	Horiguchi *et al.* [81]	Estimer les dates de début et de fin de chaque *step* affecté à une ressource critique	Améliorer la performance de livraison et la prévisibilité du système	✓		✓		Exemple
	Habenicht and Mönch [71]	Déterminer les dates de début et de fin de chaque opération de chaque lot	Établir un plan de production réalisable	✓		✓	✓	Exemple
	Chua *et al.* [41]	Calculer les dates de lancement des commandes d'assemblage pour la phase back end de la fabrication des semi-conducteurs	Résoudre le problème de libération des lots (*lot release problem*)	✓				Étude de cas réelle
	Notre étude	Minimiser le retard total pondéré	Établir un plan de production réalisable	✓	✓	✓	✓	Étude de cas réelle

3

Résolution analytique du problème de planification à capacité finie

Résumé : Ce chapitre est consacré à la description formelle du problème de planification de la capacité de production considéré dans cette étude. Tout d'abord, les contraintes, les hypothèses, les enjeux visés et les questions de recherche posées sont présentés. Ensuite, nous introduisons l'ensemble des notations mathématiques employées dans la suite du manuscrit. Puis, une formulation mathématique du problème sous forme d'un programme linéaire mixte (*MIP*) est proposée. Enfin, la complexité de calcul du problème est examinée. On montre expérimentalement la limite de résolution de la méthode exacte pour des instances industrielles. Ce résultat nous ramène à proposer trois méthodes de résolution approchées utilisant la programmation mathématique. Ce sont des heuristiques basées sur la décomposition et la relaxation dont l'objectif est de simplifier le problème et d'obtenir des solutions réalisables, dans un temps d'exécution réduit. Une analyse des résultats obtenus des tests des trois méthodes, sur des instances générées aléatoirement, est présentée à la fin de ce chapitre.

Sommaire

3.1 Introduction **57**
3.2 Description du problème **57**
3.2.1 Contraintes 58
3.2.2 Hypothèses 59
3.2.3 Enjeux 60
3.2.4 Questions de recherche 61
3.3 Notations **61**
3.4 Formulation mathématique du problème : MIP **62**
3.5 Complexité **64**
3.6 Résolution du problème **65**
3.7 Méthodes de résolution alternatives **66**
3.7.1 Procédure d'agrégation 67
3.7.2 Heuristique de décomposition 67
3.7.3 Relaxation lagrangienne 69
3.8 Résultats expérimentaux **75**
3.9 Conclusion **79**

Les travaux présentés dans ce chapitre sont aussi présentés dans l'article suivant :

[118] E. Mhiri, M. Jacomino, F. Mangione, P. Vialletelle, and G. Lepelletier. *A step toward capacity planning at finite capacity in semiconductor manufacturing.* In Proceedings of the 2014 Winter Simulation Conference, pages 2239-2250. IEEE Press (2014).

3.1 Introduction

Dans ce chapitre, le problème de planification de la production des semi-conducteurs, considéré dans cette thèse, est bien identifié et formalisé mathématiquement.

Les différentes contraintes prises en compte dans notre étude sont présentées. Pour chaque contrainte, la correspondance avec la difficulté réelle rencontrée, est précisée. Certaines contraintes additionnelles rencontrées dans la réalité sont occultées lors de la modélisation et la résolution de notre problème, mais seront reprises à la fin de notre étude comme perspectives de recherche.

Les enjeux visés sont justifiés. Les questions de recherche auxquels nous allons répondre le long de cette étude sont posées. Les notations, utilisées tout au long du document pour représenter le problème étudié, sont introduites dans la section 3.3.

Dans la section 3.4 de ce chapitre, une première approche de modélisation et de résolution du problème est présentée en utilisant la programmation linéaire mixte (*MIP*). Le recours à la programmation linéaire malgré ses limites de résolution pour les problèmes de grande taille, prouvées dans la littérature [178], est justifié par les raisons suivantes :

- Il s'agit d'une méthode de résolution exacte ;
- C'est l'un des outils les plus puissants et les plus utilisés pour l'aide à la décision ;
- Elle permet de décrire le problème mathématiquement.

La formulation proposée est d'abord testée sur des instances générées aléatoirement. Nous procédons ensuite à une étude expérimentale de la complexité. Nous testons le *MIP* sur des instances industrielles dont les résultats seront analysés pour discerner l'apport et les limites de la procédure proposée. Vu la limite de résolution du *MIP* pour des instances de grande taille, trois méthodes de résolution approchées, basées sur la programmation mathématique, sont proposées dont le but est de simplifier le problème et réduire sa taille. Il s'agit d'une procédure d'agrégation, d'une heuristique de décomposition et de la relaxation lagrangienne. Afin d'évaluer la performance de ces trois méthodes, elles sont testées sur des instances académiques et d'autres industrielles.

3.2 Description du problème

Le problème considéré consiste à proposer un planning prévisionnel qui pourrait estimer les dates de début, les temps d'attente et les dates de fin pour les différents *steps* restants des lots au début ou en cours de production ainsi que la charge accumulée sur les équipements à chaque période d'un horizon de planification à moyen terme.

3.2.1 Contraintes

Pour le problème considéré, on trouve des contraintes classiques et habituelles de planification de la production mais aussi des contraintes spécifiques à l'environnement industriel considéré dans notre étude.

Les contraintes classiques sont les suivantes :

Les contraintes de temps : Lors de la planification de la production, on doit tenir compte du calendrier de production, des délais de production et des dates d'échéance de livraison.

Les contraintes de capacité : En général, produire un lot nécessite la mobilisation d'un ou plusieurs équipements à capacité limitée. Un plan de production devrait ainsi tenir compte de cette capacité limitée.

Les contraintes découlant des spécificités de l'industrie des semi-conducteurs et donnant au problème son caractère original et complexe, sont les suivantes :

Succession des *steps* : Les *steps* sont successifs et chaque *step* est exécuté exactement une seule fois durant l'horizon de planification pour chaque lot.

Diversité des priorités des lots : Les lots n'ont pas tous le même degré d'urgence. Certains, selon leur temps d'attente, le produit final auquel ils sont affectés, ou encore le client, sont prioritaires.

Dans la littérature, il y a plusieurs règles de répartition employées pour la gestion des priorités des lots [134]. Dans notre étude, la priorité d'un lot est définie par sa date d'échéance de livraison donc on considère la règle *Earliest due date (EDD)*. Un poids est défini pour chaque lot indiquant la priorité de sa fabrication.

Qualifications des équipements : Comme il est indiqué dans le premier chapitre, on trouve dans les *wafer fabs*, certains équipements qualifiés pour un *step* alors que d'autres non.

Il y a d'autres contraintes qui doivent être considérées pour la mise en œuvre industrielle. Ces contraintes sont occultées dans nos approches de résolution proposées vu qu'elles complexifient davantage le problème considéré mais elles seront proposées comme perspectives de recherche à la fin du manuscrit.

Elles sont les suivantes :

Contraintes d'enchaînement : Les contraintes d'enchaînement entre des *steps* consécutifs présentent une autre restriction importante. Par exemple, il y a souvent une restriction de temps entre les opérations de gravure et les opérations d'oxydation / diffusion [149]. Les *steps* dans le processus de fabrication des semi-conducteurs sont caractérisés par des contraintes d'enchaînement entre eux *i.e.* le temps entre la fin d'un step n et le début d'un *step* $n + q$ doit être inférieur à une limite de temps, afin de garantir la qualité des lots. En effet, les lots dont les contraintes d'enchaînement des *steps* sont violées, doivent souvent être mis au rebut car leur reprise est généralement non autorisée.

Contraintes de *batching* : Il s'agit de tenir compte des équipements fonctionnant par *batch i.e.* plusieurs *steps* peuvent être traités simultanément. Dans l'industrie des semi-conducteurs, ces équipements existent dans l'atelier de diffusion, plus précisément dans deux types d'équipements : les équipements de nettoyage et les fours. Les raisons de *batching* sont la réduction des réglages, la facilitation de la manutention, la réduction du taux d'utilisation des équipements,...

La formation d'un *batch* sur un équipement doit tenir compte de la taille minimale et maximale du *batch* sur l'équipement considéré.

Une fois que le traitement d'un *batch* a été lancé, aucun lot ne peut être enlevé ou ajouté au *batch*. En raison de la nature chimique du procédé, il est impossible de traiter des *steps* avec des recettes différentes en même temps dans le même *batch*. Ainsi, tous les *steps* ayant la même recette ont le même temps de process.

Contraintes de *setup* : Notons qu'il y a des contraintes de *setup* dépendant de la séquence des *steps* (*Sequence-dependent setup time* en anglais). Ces temps de réglage se produisent dans certains ateliers de production et sont liés au changement de la température, la pression du gaz, la composition du métal, etc.

3.2.2 Hypothèses

Cette recherche est basée sur des hypothèses, issues de la littérature [147], telles que :

1. L'usine de plaquettes produit une grande diversité de produits.
2. Un lot est composé de 25 plaquettes.
3. Les parcs d'équipements sont composés d'une ou plusieurs machines parallèles.
4. Les machines parallèles dans un parc d'équipements sont identiques sur tous les aspects.

5. Les temps de process de tous les *steps* de fabrication sont déterministes. Ils varient entre 15 minutes et 12 heures.
6. Le temps d'installation de chaque *step* de fabrication est inclus dans le temps de process.
7. La taille de la file d'attente devant les parcs d'équipements est supposée infinie.
8. Les dates d'échéance de livraison sont données pour chaque lot de chaque commande d'un produit.

D'autres hypothèses sont spécifiques à cette étude :

1. Le modèle de temps de cycle des *steps* est extrait de l'historique des données.
2. Les temps de transfert d'un *step* à un autre, y compris les temps de *setup* dépendant de la séquence, sont considérés négligeables.
3. Les capacités des équipements sont déterministes.
4. Les demandes des clients pour chaque produit et pour chaque période sont connues. Nous n'abordons pas dans ces travaux, les aspects liés à l'incertitude des demandes.
5. Le fonctionnement par *batch* est négligé c'est-à-dire que le fonctionnement par plaque est considéré à ce niveau d'étude. Il sera proposé pour la poursuite de nos travaux.
6. Tout l'horizon de planification est divisé en périodes au cours desquelles les demandes et les capacités de production sont bien définies. La durée de chaque période de planification est constante et correspond à une semaine.

3.2.3 Enjeux

L'enjeu principal de cette étude est de proposer un planning de production réalisable pour chaque lot, durant un horizon de planification à moyen terme divisé en périodes hebdomadaires, et dans un temps d'exécution réduit (de quelques minutes) en tenant compte des contraintes citées dans la section précédente.

Rappelons qu'un plan correspond à un ensemble de quantités de produits devant être disponibles à la fin de chaque période et ces quantités correspondent à des ordres de fabrication à lancer dans l'atelier. Un plan de production est réalisable s'il existe au moins une planification des ordres de fabrication permettant effectivement de fabriquer les quantités aux périodes prévues.

Ainsi, afin d'obtenir un planning de production réalisable, les objectifs considérés dans cette étude sont :

- Minimiser les retards de livraison des lots
- Équilibrer la charge et la capacité des parcs d'équipements

Dans notre étude, l'indicateur de performance à optimiser en pratique correspond à la minimisation de la somme des retards pondérés ou "*Total Weighted Tardiness (TWT)*". En effet, il faut considérer des priorités entre les lots à cause de leur temps d'attente dans l'unité de production et parce qu'ils sont liés à des produits et des clients différents. Par conséquent, il est nécessaire de tenir compte d'un critère qui intègre cette hiérarchie entre les lots à traiter. C'est pourquoi on considère les pondérations. En plus, TWT est une mesure de performance traduisant une pénalité pour chaque lot dont la fin de son processus de fabrication dépasse sa date de livraison promise donc cet indicateur mesure la performance de l'usine en terme de respect des délais de livraison.

3.2.4 Questions de recherche

A partir des objectifs détaillés ci-dessus, nous tirons trois questions auxquelles cette thèse vise à répondre :

1. Comment peut-on satisfaire les commandes clients en termes de délais et quantités avec la capacité disponible ?
2. Quelles sont les précautions nécessaires à prendre afin d'éviter les retards de livraison ?
3. Comment peut-on équilibrer la charge et la capacité des parcs d'équipements en tenant compte de la diversité et des qualifications des équipements ?

3.3 Notations

Avant d'entamer la formulation mathématique du problème considéré, nous présentons, dans cette section, les différentes notations employées. Nous considérons un problème dans lequel nous cherchons à planifier la production de L lots de produits différents, composé chacun de Q_l *wafers*, sur un horizon de planification décomposé en T périodes discrètes de durée constante P_t.

Pour chaque lot $l \in \{1, \ldots, L\}$, il reste un nombre S_l de *steps* à exécuter. Chaque lot, de poids w_l indiquant sa priorité, a une date de lancement r_l et une *due date* d_l. Chaque *step* restant $s_l \in \{1, \ldots, S_l\}$ du lot l est traité sur un ou plusieurs parcs d'équipements $i \in \{1, \ldots, I\}$. L'exécution d'un *step* s_l sur un parc d'équipement i dépend de la qualification de la recette du *step* au parc d'équipement notée $Q_{s_l,l,i}$. Comme entrée du problème, on définit une matrice de qualifications $[Q]$ dont les éléments $Q_{s_l,l,i}$ sont binaires. $Q_{s_l,l,i}$ est égale à 1 si le parc d'équipement i peut effectuer le step s_l du lot l et 0 sinon. La quantité de *wafers* d'un lot l affectée au parc d'équipements i, pouvant traiter le *step* s_l pendant la

période t, est connue d'avance et notée $a_{s_l,l,i,t}$. Le temps de process unitaire d'un *step* s_l du lot l sur chacun des parcs d'équipements i qui lui est qualifié, est noté $p_{s_l,l,i}$. Chaque parc d'équipements dispose d'une capacité limitée prédéfinie pour chaque période t de l'horizon de planification notée $C_{i,t}$. Les capacités de production $C_{i,t}$ doivent être respectées.

Les variables de décision considérées dans ce problème sont les dates de début et les dates de fin de chaque *step* restant s_l du lot l notées respectivement $s_{s_l,l}$ et $e_{s_l,l}$, la date de fin de fabrication de chaque lot l notée C_l et la charge accumulée sur chaque parc d'équipement i, pendant une période t, notée $L_{i,t}$.

La figure 3.1 illustre les paramètres et les variables de décision relatifs à la projection d'un lot en partant de la date courante jusqu'à la fin de l'horizon de planification. Le temps de cycle d'un *step* est composé d'un temps d'attente et d'un temps de process.

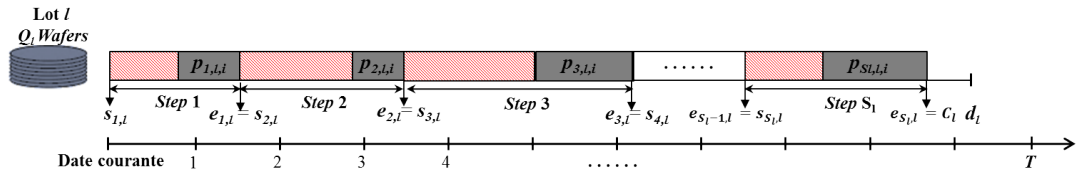

Figure 3.1 – Paramètres et variables de décision du problème.

La fonction objectif considérée est de minimiser la somme des retards pondérés (TWT). Le retard T_l d'un lot l est alors défini par $T_l = max(0, C_l - d_l)$. Le retard pondéré du lot l (WT_l) est défini par $WT_l = (w_l \times T_l)$. Le retard total pondéré calcule la somme pondérée des valeurs des retards : $TWT = \sum_l WT_l$.
Le tableau 3.1 récapitule la notation proposée.

3.4 Formulation mathématique du problème : MIP

Dans cette section, une formulation du problème considéré en programme linéaire mixte (MIP) est proposée. En utilisant la notation présentée dans la section 3.3, le modèle *MIP* se compose de la fonction objectif (3.1) et les contraintes ((3.2)...(3.14)).

Il peut alors s'écrire comme suit (MIP.1) :

Tableau 3.1 – Notations du problème

Indices	**Description**
L	Nombre de lots
$l = 1..L$	Indice du lot
S_l	Nombre de *steps* restants du lot l
$s_l = 1..S_l$	Indice du *step* du lot
I	Nombre de parcs d'équipements
$i = 1..I$	Indice du parc d'équipement
T	Nombre de périodes
$t = 1..T$	Indice de la période
Paramètres	**Description**
Q_l	Quantité de *wafers* pour chaque lot l
r_l	Date de lancement du lot l
w_l	Poids du lot l
d_l	*Due date* du lot l
P_t	Durée de la période t
$p_{s_l,l,i}$	Temps de process unitaire du *step* s_l du lot l sur le parc d'équipements i
$C_{i,t}$	Capacité du parc d'équipements i sur la période t
$Q_{s_l,l,i}$	$= 1$ si le *step* s_l du lot l est qualifié sur le parc d'équipements i, 0 sinon
$a_{s_l,l,i,t}$	Quantité de *wafers* du lot l au *step* s_l, pouvant être traité par le parc d'équipements i, pendant la période t
Variables de décision	**Description**
$s_{s_l,l}$	Date de début d'un *step* s_l du lot l
$e_{s_l,l}$	Date de fin d'un *step* s_l du lot l
C_l	Date de fin du lot l
T_l	Retard du lot l
$L_{i,t}$	Charge du parc d'équipement i pendant la période t
$y_{s_l,l,t}$	$= s_{s_l,l}$ si le *step* s_l du lot l est lancé pendant la période $[t,t+1[$, 0 sinon
$x_{s_l,l,t}$	=1 si le *step* s_l du lot l est exécuté pendant la période t, 0 sinon

Programme linéaire 1 (MIP.1).

$$min \quad \sum_l w_l T_l \tag{3.1}$$

$$s.c. \quad s_{1,l} \geq r_l \qquad l = 1, \ldots, L \tag{3.2}$$

$$s_{s_l,l} + \sum_i \sum_t p_{s_l,l,i} \times a_{s_l,l,i,t} \times Q_{s_l,l,i} \times x_{s_l,l,t} \leq e_{s_l,l} \quad s_l = 1, \ldots, S_l,\ l = 1, \ldots, L \tag{3.3}$$

$$s_{s_l,l} = e_{s_l-1,l} \quad s_l = 1, \ldots, S_l,\ l = 1, \ldots, L \tag{3.4}$$

$$\sum_t y_{s_l,l,t} = s_{s_l,l} \quad s_l = 1, \ldots, S_l,\ l = 1, \ldots, L \tag{3.5}$$

$$\sum_t x_{s_l,l,t} = 1 \quad s_l = 1, \ldots, S_l,\ l = 1, \ldots, L \tag{3.6}$$

$$C_l = e_{S_l,l} \quad l = 1, \ldots, L \tag{3.7}$$

$$T_l \geq C_l - d_l \quad l = 1, \ldots, L \tag{3.8}$$

$$T_l \geq 0 \qquad l = 1, \ldots, L \tag{3.9}$$

$$t \times P_t \times x_{s_l,l,t} \leq y_{s_l,l,t} \quad s_l = 1,\ldots,S_l,\ l = 1,\ldots,L, \quad t = 1,\ldots,T \tag{3.10}$$

$$(t+1) \times P_t \times x_{s_l,l,t} > y_{s_l,l,t} \quad s_l = 1,\ldots,S_l,\ l = 1,\ldots,L, \quad t = 1,\ldots,T \tag{3.11}$$

$$L_{i,t} = \sum_{l}\sum_{s_l} p_{s_l,l,i} \times a_{s_l,l,i,t} \times Q_{s_l,l,i} \times x_{s_l,l,t}$$

$$i = 1,\ldots,I,\ t = 1,\ldots,T \tag{3.12}$$

$$L_{i,t} \leq C_{i,t} \quad i = 1,\ldots,I,\ t = 1,\ldots,T \tag{3.13}$$

$$x_{s_l,l,t} = \{0,1\} \quad s_l = 1,\ldots,S_l,\ l = 1,\ldots,L,\ t = 1,\ldots,T \tag{3.14}$$

La fonction objectif (3.1) minimise le retard total pondéré (TWT). Les contraintes du *MIP* peuvent être classées en deux catégories : les contraintes temporelles ((3.2)...(3.11)) et les contraintes cumulatives ((3.12)-(3.13)).

Les contraintes (3.2) définissent la date de début du premier *step* restant à traiter pour chaque lot. La date de fin de chaque *step* restant de chaque lot est supérieure ou égale à la somme de la date de début et du temps de process (cf. Contraintes (3.3)). La différence entre ces deux termes est égale au temps d'attente estimé pour chaque *step*. Les contraintes (3.4) présentent les contraintes de succession des *steps*. Les contraintes (3.5) garantissent que chaque *step* restant de chaque lot est lancé une fois. Les contraintes (3.6) vérifient que chaque *step* restant de chaque lot est exécuté une fois pendant la période de planification considérée. Les contraintes (3.7) définissent les dates de fin des lots. Les contraintes (3.8) et (3.9) calculent le retard pour chaque lot. Les contraintes (3.10) et (3.11) indiquent que chaque *step* restant de chaque lot est traité dans l'intervalle $[t, t+1[$. Les contraintes (3.12) calculent la charge accumulée sur chaque parc d'équipements pendant chaque période en tenant compte de la qualification du parc d'équipements au *step* traité et du pourcentage de la quantité de *wafers* des lots affectés au parc d'équipements considéré. Les contraintes (3.13) sont les contraintes de capacité. Les contraintes (3.14) sont les contraintes d'intégrité.

3.5 Complexité

Le problème considéré est avéré NP-difficile au sens fort, même dans un cas simplifié. La NP-difficulté est prouvée par une transformation à partir du problème de sac à dos, qui est NP-difficile au sens fort [57] (cf. annexe B).

3.6 Résolution du problème

Le modèle mathématique présenté dans la section 3.4 a été résolu par le solveur `ILOG CPLEX` 12.6. Les expérimentations ont été effectuées sur un PC `Intel® Core™ i5` fonctionnant avec un processeur de 2,7 GHz et 4 Go de RAM. Des tests ont été réalisés sur 30 instances du problème générées aléatoirement de sorte à faire ressortir les principales caractéristiques des données industrielles et à garder un certain degré de généralité, et ce en vue de conserver toute la difficulté du problème. En effet, en se basant sur l'analyse de la littérature ([117], [125]), nous avons identifié six paramètres importants du problème qui pourraient affecter la performance de l'approche proposée : le nombre de lots (L), le nombre maximum de *steps* restants pour chaque lot dans le *WIP* ($\max S_l$), le nombre de parcs d'équipements (I), la longueur de l'horizon de planification (T), les temps de process unitaires ($p_{s_l,l,i}$) et les dates d'échéance de livraison des lots (d_l). Nous considérons le cas de 3, 5, 10, 20, 100 et 300 parcs d'équipements avec une capacité fixe égale à 100%. Les poids des lots w_l sont définis en utilisant la loi uniforme sur (0,1). Les dates de lancement des lots r_l et la quantité de plaquettes pour chaque lot Q_l sont supposées fixées respectivement à 0 et 25 pour tous les lots. Les dates d'échéance de livraison des lots d_l et les temps de process unitaires des steps $p_{s_l,l,i}$ sont extraits à partir des données réelles. Les valeurs de d_l varient entre 1 et 210 jours par rapport à la date de lancement et les temps de process unitaires $p_{s_l,l,i}$ sont compris entre 0.01 et 0.25 heures. L'horizon de planification est fixé à 24 périodes (semaines). Le tableau 3.2 présente les différents paramètres des tests générant 30 instances de taille différente.

Tableau 3.2 – Synthèse des paramètres des tests réalisés

Paramètre du problème	**Valeurs utilisées**
Nombre de lots (L)	2, 3, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 240, 1000, 1700, 2000
Maximum nombre de *steps* restants d'un lot l ($\max S_l$)	1, 2, 5, 6, 8, 10, 20, 30, 40, 50, 60, 100, 150, 200, 250, 680
Nombre de parcs d'équipements (I)	3, 5, 10, 20, 100, 300
Nombre de périodes (T)	24
Poids du lot (w_l)	Loi uniforme (0,1)
Dates de lancement des lots (r_l)	0
Due dates des lots (d_l)	r_l+[1..210]
Quantité de plaques dans un lot (Q_l)	25
Temps de process unitaire des *steps* ($p_{s_l,l,i}$)	[0.01..0.25]

Des résultats optimaux ont été obtenus dans un temps raisonnable pour les instances de taille réduite. En augmentant la taille des instances testées (jusqu'à environ 4000 *steps* à planifier), la résolution du *MIP* a été interrompue pour un temps de calcul très important

et faute de mémoire informatique. En analysant la longueur du temps de calcul, il s'avère qu'au delà de 4000 *steps*, le solveur n'arrive pas soit à trouver la solution optimale soit ou à prouver l'optimalité de la solutions trouvée en moins de 5 minutes. La figure **??** illustre les limites de résolution du *MIP*.

Une instance réelle correspond à un *WIP* composé de 2000 lots, chaque lot ayant un maximum de 680 *steps* restants pour être traités sur 300 parcs d'équipements pendant un horizon de planification composé de 24 périodes (semaines). Donc, un problème réel présente 70 742 400 contraintes et 69 011 200 variables. Ainsi, la taille du cas réel est évidemment trop large pour être résolue en utilisant le *MIP* proposé (cf. figure 3.2). Partant des preuves empiriques sur les difficultés de calcul afin d'obtenir un planning

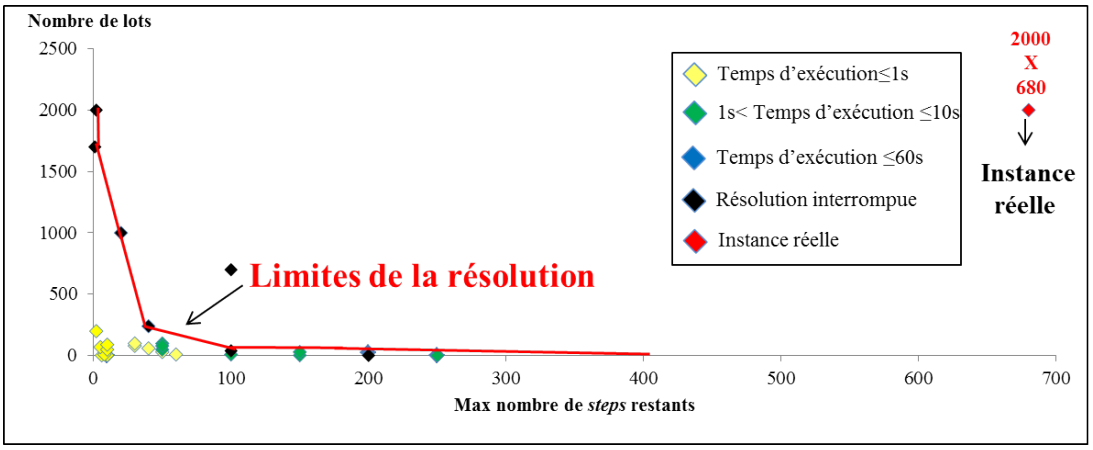

Figure 3.2 – Limites de résolution du MIP.

optimal en tenant compte des dates d'échéance de livraison des lots et des contraintes de capacité, il est évident que le problème de planification de la production considéré sera informatiquement non résolu.

3.7 Méthodes de résolution alternatives

Afin de simplifier le problème considéré, trois méthodes alternatives de résolution sont proposées. Ces techniques correspondent à une procédure d'agrégation, une heuristique de décomposition et une relaxation lagrangienne. La procédure d'agrégation et l'heuristique de décomposition sont proposées pour construire une solution réalisable et obtenir une borne supérieure. Une borne inférieure du problème est calculée par la relaxation lagrangienne. Le principe de chaque méthode est expliqué. Nous avons codé les algorithmes proposés en `JAVA` et nous les avons testés en reprenant les mêmes instances testées sur le *MIP*. Tout d'abord, nous procédons à des tests pour les instances de petite taille pour valider chacune des approches proposées. Ensuite, nous analysons expérimentalement la

complexité des différentes approches, en nous basant sur les temps de calcul pour des instances industrielles de grande taille.

3.7.1 Procédure d'agrégation

Afin de réduire le nombre de variables et de contraintes et trouver des solutions à notre problème réel, une procédure d'agrégation pour les *steps* de process et les parcs d'équipements a lieu en premier. Les opérations sont considérées au lieu des *steps* individuels. Une opération est définie comme un ensemble de *steps* consécutifs d'un seul lot. Concernant les parcs d'équipements, ils sont remplacés par des usages. Plus formellement, si un lot l a un flux de processus $s_l \in \{1, \ldots, S_l\}$ à exécuter, alors nous remplaçons s_l par la séquence $o_l \in \{1, \ldots, O_l\}$ des opérations. En général, une opération se compose de trois ou quatre *steps* consécutifs. Ces opérations sont exécutées sur des parcs d'équipements $i \in \{1, \ldots, I\}$ qui sont remplacés aussi par des usages $u \in \{1, \ldots, U\}$. Un usage est composé de un à huit parcs d'équipements. Le temps de process d'une opération sur un usage est égal à la somme des temps de process des *steps* composant l'opération sur tous les parcs d'équipements composant l'usage. Par exemple, si une opération est composée de n *steps* et est qualifiée pour un usage composé de m parcs d'équipements, le temps de process de l'opération est égal à :

$$p_{o_l,l,u} = \sum_{s_l=1}^{n} \sum_{i=1}^{m} p_{s_l,l,i} \tag{3.15}$$

Ainsi, pour notre instance réelle, au lieu de planifier au maximum 680 *steps* par lot traités sur 330 parcs d'équipements, on planifie au maximum 220 opérations par lot exécutées sur 190 usages. Par conséquent, le nombre de variables devient égal à 22 008 560 au lieu de 69 011 200 variables et le nombre de contraintes se réduit de 70 742 400 à 22 897 120 contraintes.

Pour cette méthode de résolution, nous avons repris les mêmes tests effectués en utilisant le MIP. Pour ces instances, nous avons modifié le nombre maximum de steps restants d'un lot et le nombre de parcs d'équipements par le nombre maximum d'opérations restantes d'un lot et le nombre d'usages, respectivement. Les valeurs de ces paramètres sont indiquées dans le tableau 3.3. Les résultats de ces tests sont présentés ultérieurement.

3.7.2 Heuristique de décomposition

En considérant la procédure d'agrégation, une autre approche de résolution a été proposée dont le but est de réduire davantage la taille du problème et de trouver des solutions aux problèmes de grande taille. Il s'agit d'une heuristique de décomposition

Tableau 3.3 – Valeurs des nouveaux paramètres

Paramètre du problème	**Valeurs utilisées**
Maximum nombre d'opérations restantes d'un lot l ($\max O_l$)	1, 2, 3, 5, 7, 10, 13, 18, 20, 35, 50, 67, 70, 83, 220
Nombre d'usages (U)	1, 2, 3, 7, 30, 190

dont le principe est de décomposer le problème global et complexe en de multiples sous problèmes de petite taille.

La décomposition est effectuée en 2 étapes :

1. Répartir les lots selon la règle de priorité "*Earliest Due Date*" (EDD) en trois classes (les bornes de d_l sont définies par les responsables de production du partenaire industriel) :
   - Classe 1 pour les lots de priorité élevée ("*hot lots*") $d_l \leq 10 jours$.
   - Classe 2 pour les lots de priorité normale $10 jours < d_l \leq 30 jours$.
   - Classe 3 pour les lots de faible priorité $d_l > 30 jours$.
2. Pour chaque classe des lots, répartir les lots en deux catégories (de même, les bornes de O_l sont définies par les responsables de production du partenaire industriel) :
   - Catégorie 1 pour les lots dont le nombre maximum d'opérations restantes à traiter est supérieur à 100 : $O_l > 100$
   - Catégorie 2 pour les lots dont le nombre maximum d'opérations restantes à traiter est inférieur ou égal à 100 : $O_l \leq 100$.

Ainsi, à la fin de cette décomposition, on obtient six sous problèmes. Pour chaque sous-problème, on applique le *MIP* sur $L11$, $L12$, $L21$, $L22$, $L31$ et $L32$ dans cet ordre. A la fin de l'exécution du *MIP*, pour chaque sous-problème, la capacité des usages est mise à jour pour chaque période. Elle est égale à la capacité initiale (donnée du *MIP*) moins la charge $L_{u,t}$ calculée par le *MIP*. La figure 3.3 montre le principe de décomposition.

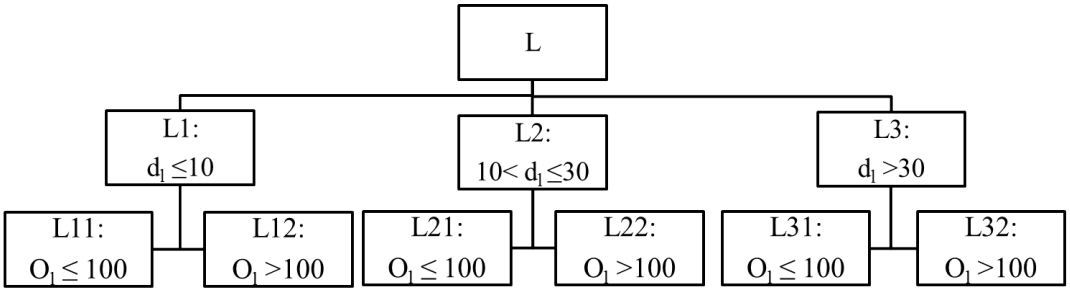

Figure 3.3 – Principe de décomposition.

Cette heuristique est développée en `JAVA`. Pour chaque sous-problème, le *MIP* a été exécuté en utilisant le logiciel `CPLEX 12.6`. Pour des instances réelles, les pourcentages de lots pour chaque sous-problème sont affichés dans le tableau 3.4.

Tableau 3.4 – Pourcentage des lots pour chaque sous-problème pour des instances réelles

Paramètre	**Valeur moyenne**
$L1$	15% de L
$L11$	100% de $L1$
$L12$	0% de $L1$
$L2$	20% de L
$L21$	90% de $L2$
$L22$	10% de $L2$
$L3$	65% de L
$L31$	45% de $L3$
$L32$	55% de $L3$

3.7.3 Relaxation lagrangienne

La relaxation lagrangienne [54, 106] est une méthode de résolution applicable à un grand ensemble de problèmes en programmation mathématique. L'idée consiste à relâcher un ensemble de contraintes difficiles (souvent couplant des variables) qui sont introduites dans la fonction objectif sous la forme d'une pénalité. La dualisation des contraintes difficiles dans la fonction objectif résulte en un problème Lagrangien plus facile à résoudre.

Avant d'appliquer cette technique à notre problème, nous allons expliquer son principe de façon générale.

3.7.3.1 Principe de la relaxation lagrangienne

Lagrangien et fonction duale

Considérons le problème d'optimisation (P) suivant :

$$\begin{aligned} min \quad & cx \\ s.c. \quad & Ax \leq b \qquad (3.16) \\ & Bx \leq d \qquad (3.17) \end{aligned}$$

Nous associons un vecteur u dit vecteur Lagrangien à l'ensemble des contraintes (3.17). La fonction lagrangienne associée au problème (P) est :

$$L(x,u) = cx + u(Bx - d) \qquad (3.18)$$

On définit par (S) l'ensemble des solutions x qui satisfont les contraintes (3.16). La fonction duale $w(u)$ est définie pour tout $u \geq 0$ par :

$$w(u) = min_{x \in S} L(x, u) \tag{3.19}$$

Pour tout $u \geq 0$, on définit le problème Lagrangien par le programme linéaire suivant :

$$min_{x \in S} L(x, u) \tag{3.20}$$

$w(u)$ définit une borne inférieure au problème (P). Ainsi, le but est de rechercher le vecteur u^* qui maximise la fonction $w(u)$. Ceci est obtenu en résolvant le problème dual Lagrangien (D) suivant :

$$max_{u \geq 0} w(u) \tag{3.21}$$

Questions clés

Deux questions se posent pour définir une approche de relaxation lagrangienne :

— **Quelles contraintes doit-on relâcher ?**

La relaxation des contraintes permet d'obtenir des problèmes plus simples à résoudre. Cependant, on ne peut pas générer des problèmes trop faciles puisque dans ce cas la borne inférieure sera de mauvaise qualité. De plus, en pratique, il est important de préserver une taille raisonnable au vecteur des multiplicateurs, le nombre de contraintes relâchées doit être limité.

— **Comment résoudre le problème du dual Lagrangien ?**

Cette question signifie : comment trouver de bonnes valeurs des multiplicateurs Lagrangiens ?

Il existe des méthodes heuristiques permettant de mettre à jour les multiplicateurs Lagrangiens : la méthode des faisceaux, la génération de colonnes et le sous-gradient. Dans ce qui suit, nous présentons seulement la méthode du sous-gradient, qui est la plus fréquemment utilisée.

Résolution du problème dual

L'idée de cette méthode pour trouver la solution optimale u^* de (D) est de commencer à un point initial u^0 et de l'améliorer itérativement dans la direction du sous-gradient avec un certain pas de déplacement. A chaque étape k, le vecteur Lagrangien est déterminé

comme suit :

$$u_{k+1} = u_k + \lambda_k \frac{s^k}{\|s^k\|} \tag{3.22}$$

Où s^k est le sous-gradient de w^u à l'itération k et $\|s^k\|$ sa norme. Le scalaire λ_k est appelé pas de déplacement.

Polyak [136] a montré que la suite u^k converge vers la solution optimale w^* sous les conditions suivantes : $\lim_{k \to +\infty} \lambda_k = 0$ et $\sum \lambda_k = +\infty$

Une formule ayant montré une grande efficacité pour calculer le pas de déplacement est la suivante :

$$\lambda_k = \delta_k \frac{(\bar{w}) - w(u^k)}{\|s^k\|} \tag{3.23}$$

Où $\bar{w}$ correspond à la meilleure solution obtenue pour (P). Il est appelé coefficient de relaxation et est choisi entre 0 et 2 $\forall k$. Souvent, δ_0 est initialisé à 2, et δ_k est réduit par un facteur de 2 toutes les x itérations (paramètre fixé à l'avance) [77]. La méthode de sous-gradient est résumée dans l'algorithme (1). L'algorithme du sous-gradient s'arrête si l'une des conditions suivantes est vérifiée :

— un nombre maximal d'itérations est atteint

— $\bar{w} - BI < \epsilon \quad (\epsilon > 0)$

— $\|u^{k+1} - u^k\| < \epsilon \quad (\epsilon > 0)$

3.7.3.2 Application de la relaxation lagrangienne à notre problème

Nous utilisons la méthode de relaxation lagrangienne afin de calculer une borne inférieure au problème de planification de la production considéré. Plusieurs auteurs ont proposé de relâcher les contraintes de capacité dans les problèmes de planification de la production ([12], [24], [73], [158]).

Comme on remarque dans le *MIP* proposé, la fonction objectif, ainsi que toutes les contraintes, à l'exception des contraintes de capacité (3.12 et 3.13) peuvent être divisées en sous parties à un seul lot. Ainsi, nous relâchons les contraintes de capacité afin d'être en mesure de diviser le problème en L sous-problèmes simples à un seul lot et sans capacité. Les autres contraintes restent non modifiées et la nouvelle fonction objectif (fonction de Lagrange L), en intégrant le multiplicateur Lagrangien $u_{i,t}$ $\forall i \in I$ et $t \in T$, est donnée

Algorithme 1 : Algorithme du sous-gradient

Initialisation des multiplicateurs $u^0 \geq 0$;
Initialisation de la meilleure borne inférieure $BI := -\infty$;
$\delta_0 := 2$;
$k := 1$;
tant que *le critère d'arrêt n'est pas atteint* **faire**
  $w(u^k) := min_{x \in S} L(x, u^k)$
  si *($Bx^k \leq d$) et ($u^k(Bx^k - d) = 0$)* **alors**
    x^j est solution optimale de (P)
    ARRETER
  fin si
  si $BI < w(u^k)$ **alors**
    $BI := w(u^k)$
  fin si
  $s^k := Bx^k - d$
  $\lambda_k = \delta_k \frac{(\bar{w}) - w(u^k)}{\|s^k\|}$
  $u^{k+1} := max(0, u_k + \lambda_k \frac{s^k}{\|s^k\|})$
  si *aucune amélioration après x itérations* **alors**
    $\delta_{k+1} := \delta_{k+1}/2$
  sinon
    $\delta_{k+1} := \delta_{k+1}$
  fin si
  $k := k + 1$
fin tant que

par l'expression suivante :

$$\begin{aligned} L(\vec{x}, \vec{u}) &= \sum_l w_l T_l + \sum_i \sum_t u_{i,t}(L_{i,t} - C_{i,t}) \\ &= \sum_l w_l T_l + (\frac{1}{Q_l} \sum_i \sum_t \sum_l \sum_{s_l} (u_{i,t} \times p_{s_l,l,i} \times a_{s_l,l,i,t} \times Q_{s_l,l,i} \times x_{s_l,l,t} - C_{i,t}) \end{aligned} \quad (3.24)$$

Nous appelons la formulation du modèle d'origine primale et la nouvelle formulation du problème duale. L'idée de base de l'approche Lagrange est qu'une maximisation de la fonction de Lagrange L correspond à la minimisation de la fonction objectif primale. Une solution optimale est destinée à être produite étape par étape, en alternant de manière itérative le vecteur $\vec{u}$. Le problème dual est divisé en deux types de problèmes :

- Un problème principal : Les multiplicateurs Lagrangiens doivent être déterminés en utilisant l'algorithme du sous-gradient.
- Des sous-problèmes à un seul lot : Un planning optimal d'un seul lot doit être déterminé en respectant les contraintes temporelles ((3.2)...(3.11)) et les contraintes

d'intégrité (3.14) et les coûts de consommation des capacités $u_{i,t}$.

Résolution du problème principal : Méthode sous-gradient

Les sous gradients peuvent être déterminés pour tous les parcs d'équipements i et toutes les périodes t. Ils sont donnés par le terme :

$$\sum_{l}\sum_{s_l}(C_{i,t} - u_{i,t} \times p_{s_l,l,i} \times a_{s_l,l,i,t} \times Q_{s_l,l,i} \times x_{s_l,l,t}) \quad \forall i = 1,\ldots,I,\ t = 1,\ldots,T \qquad (3.25)$$

La procédure de calcul des multiplicateurs Lagrangiens est ainsi la suivante :

1. Initialiser tous les multiplicateurs Lagrangiens à 0.
2. Produire une solution initiale (notée $\bar{w}$ dans l'algorithme sous-gradient présenté ci-dessous) en utilisant l'heuristique de décomposition décrite dans la section précédente.
3. Résoudre tous les sous-problèmes.
4. Produire une solution primale réalisable et calculer la valeur objective (TWT) en utilisant l'heuristique de décomposition. Si la solution obtenue est meilleure par rapport à la meilleure solution des itérations précédentes alors on l'enregistre.
5. Calculer tous les sous gradients et recalculer le vecteur $\vec{u}$.
6. Arrêter si l'un des critères d'arrêt est satisfait. Sinon, reprendre l'étape 3.

Dans les tests expérimentaux, l'algorithme de relaxation lagrangienne s'arrête si :

— Un nombre maximal d'itérations Q est atteint. On fixera Q à 10 ;

— L'écart entre la borne inférieure et la borne supérieure est inférieur à 1%. Celui-ci est mesuré par le paramètre $GAP = \bar{w} - BI < \bar{w}$. BI est la valeur de la meilleure borne inférieure.

Amélioration de la solution de la relaxation lagrangienne

Lors de la résolution du problème Lagrangien, la solution obtenue est souvent non réalisable. En effet, comme les contraintes de capacité sont relaxées, les charges des parcs d'équipements peuvent dépasser les capacités disponibles. Pour construire une solution réalisable, les contraintes violées dans la résolution du modèle relaxé sont identifiées et la solution du problème relaxé est modifiée afin de satisfaire ces contraintes. Ainsi, nous avons proposé une heuristique de lissage de la production d'avant en arrière sur l'horizon de temps. Dans l'heuristique proposée, nous considérons la notion d'un lot critique et de période initiale :

Définition 3.1. *Un lot critique est un lot dont les steps seront déplacés.*

Définition 3.2. *Une période initiale est une première période, en partant du début de l'horizon de planification où on a une surcharge d'un parc d'équipement.*

La procédure de l'heuristique est la suivante : Pour chaque parc d'équipements $i \in I$, on calcule $L_{i,t}$. Pour un parc d'équipement sélectionné, pour chaque période $t = 1 \dots T$, sélectionner les périodes de l'horizon de planification où les contraintes de capacité ne sont pas satisfaites ($L_{i,t} > C_{i,t}$). Dans ce cas :

1. Sélectionner la première période de surcharge, en partant du début de l'horizon de planification.
2. Sélectionner un lot critique qui correspond au lot le moins prioritaire et dont le step traité sur le parc d'équipements saturé est réalisé à la fin de la période sélectionnée.
3. Décaler le *step* sélectionné et ses successeurs à la période suivante.
4. Mettre à jour la charge des parcs d'équipements saturés.

Exemple 3.1. *Pour expliquer plus le principe de l'heuristique, nous l'appliquons sur un exemple simple. Nous considérons deux lots, un avec six steps et l'autre avec trois steps restants à traiter sur un parc d'équipements et pendant un horizon de temps composé de trois périodes. Les données relatives aux lots sont représentées dans le tableau 3.5.*

Tableau 3.5 – Données relatives aux lots pour une instance simple

Lot	**Maximum nombre de *steps* restants**	**Temps de process unitaire** $p_{s_l,l,i}$	**Poids** w_l	**Due date** d_l
1	6	[0.012, 0.008, 0.018, 0.014, 0.022, 0.016]	0.2	3
2	3	[0.012, 0.018, 0.01]	0.3	1

La figure 3.4 représente le planning de production établi en utilisant la relaxation lagrangienne et la figure 3.5 illustre la saturation i.e. le rapport entre la charge et la capacité du parc d'équipements par période.

On remarque que le parc d'équipements est surchargé pendant les périodes 1 et 2. En appliquant l'algorithme, on se positionne à la première période de l'horizon de temps et on sélectionne le lot critique. Il correspond au lot 1 qui est le moins prioritaire et dont le step 3.1 est réalisé à la fin de la période 1. On décale donc ce step et ses successeurs à la période 2. Le parc d'équipements est encore surchargé (130%). On décale encore le step 3.2. Ainsi, le parc d'équipement est non saturé sur la période 1 ($L_{1,1}/C_{1,1}$=95%). La nouvelle saturation du parc d'équipements sur la période 2 est 135% donc on décale le step 1.4 à la troisième période. Ainsi, le parc d'équipements n'est plus saturé sur toutes les périodes. Les résultats en termes de planning et saturation du parc d'équipements sont affichés dans les figures 3.6 et 3.7, respectivement.

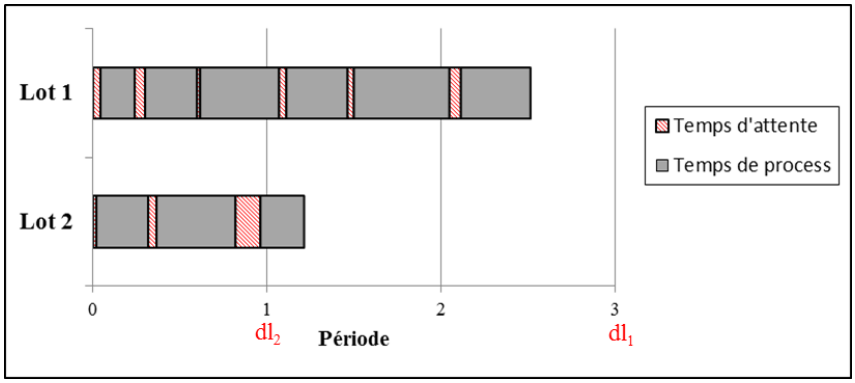

Figure 3.4 – Planning de production de l'instance.

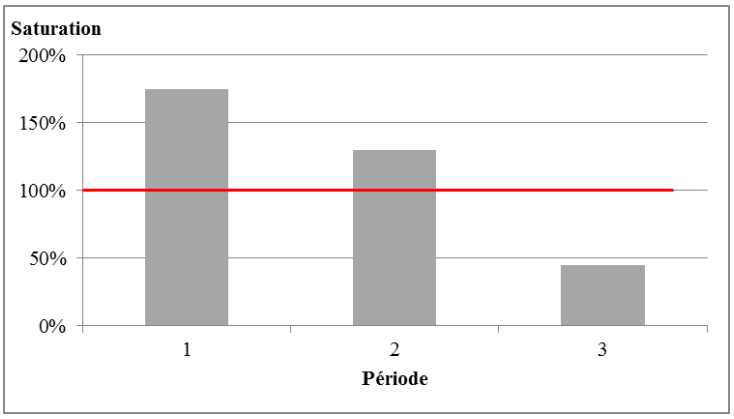

Figure 3.5 – Saturation du parc d'équipements par période.

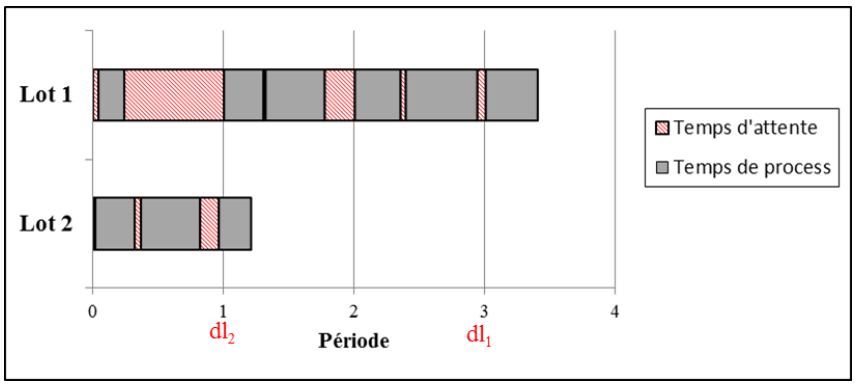

Figure 3.6 – Résultat de l'heuristique : Planning de production.

3.8 Résultats expérimentaux

Dans cette partie, nous analysons expérimentalement les différentes méthodes développées, à savoir la procédure d'agrégation, l'heuristique de décomposition et la relaxation lagrangienne en comparaison avec les résultats du *MIP*. Toutes les méthodes sont développées et exécutées, sur un ordinateur personnel avec un processeur `intel`® `Core`™ `i5-3340M CPU 2,70 GHz` et 4,00 `GO` de RAM.

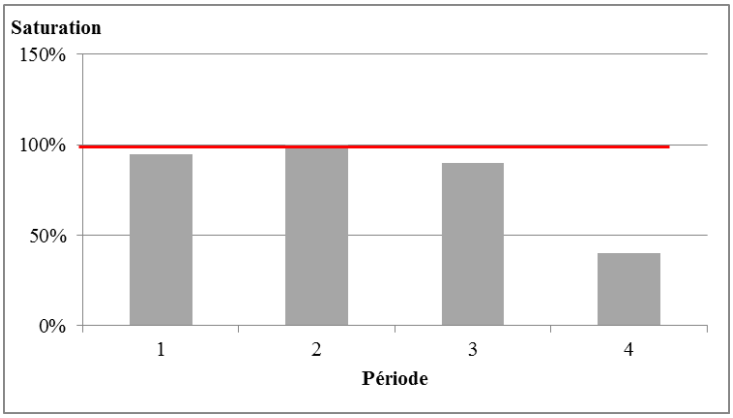

Figure 3.7 – Résultat de l'heuristique : Saturation du parc d'équipements par période.

Les trois méthodes ont été implémentées en `JAVA`. Le *MIP* dans la procédure d'aggrégation et l'heuristique de décomposition a été exécuté en utilisant le logiciel `CPLEX 12.6`.

Les tests numériques sont réalisés sur les instances générées aléatoirement, testées sur le *MIP* et présentées dans la section 3.5. Nous avons aussi exécuté chaque instance pendant un temps de calcul maximal égal à $5\,min$.

Pourcentage des instances résolues avant la condition d'arrêt

Dans un premier temps, nous présentons l'efficacité de ces méthodes en termes de pourcentage d'instances résolues avant la condition d'arrêt ($5\,min$). Sans surprise, les trois méthodes nous ont permis de résoudre des instances de petite taille. Les pourcentages des problèmes résolus en utilisant les trois méthodes en comparaison avec le *MIP* sont présentés dans le tableau 3.6.

Tableau 3.6 – Pourcentages des problèmes résolus en moins de $5\,min$ (%)

MIP	**Procédure d'agrégation**	**Heuristique de décomposition**	**Relaxation lagrangienne**
73%	93%	97%	97%

Nous remarquons que la méthode de relaxation lagrangienne et l'heuristique de décomposition ont permis la résolution de toutes les instances proposées sauf l'instance de taille réelle. Pour déterminer la limite de résolution de ces méthodes, d'autres instances de taille importante sont testées.

La figure 3.8 illustre les limites de la résolution des trois méthodes de résolution proposées en comparaison avec le *MIP*.

Nous remarquons qu'en réduisant la taille du problème, la résolution de plusieurs instances, interrompue pour le *MIP*, faute de mémoire informatique ou par ce que le temps de calcul dépasse $5\,min$, devient possible avec les méthodes de résolution proposées.

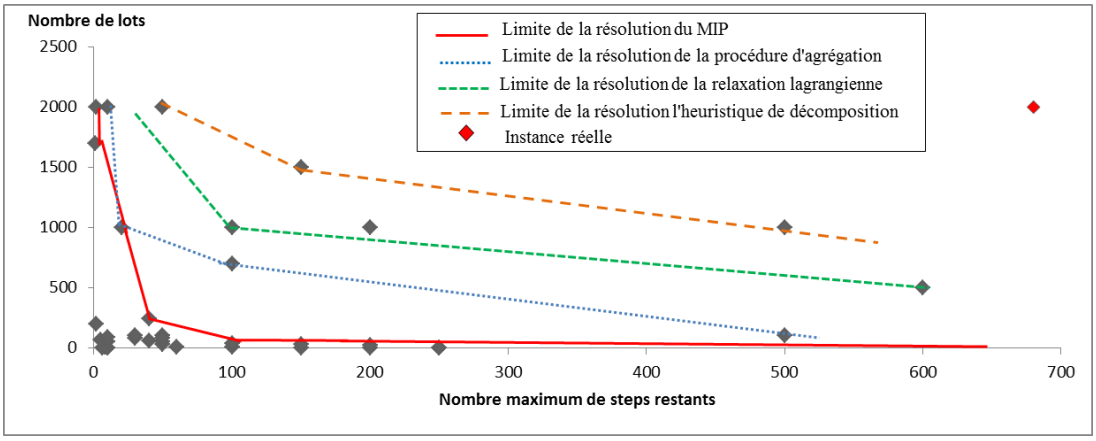

Figure 3.8 – Limites de résolution du *MIP*, la procédure d'agrégation, l'heuristique de décomposition et la relaxation lagrangienne.

Comme il est indiqué dans la section 3.5, le *MIP* permet de résoudre l'ensemble des instances dont la taille est inférieure ou égale à $L \times \max S_l = 4000$ *steps* à planifier avant 5 *min*. La méthode d'agrégation est capable de résoudre des instances de taille allant jusqu'à $L \times \max S_l = 7000$ *steps* à planifier. La limite de résolution de l'heuristique de décomposition est égale à $L \times \max S_l = 50000$ *steps* à planifier alors que celle de la relaxation lagrangienne atteint $L \times \max S_l = 10000$ *steps* à planifier. Ainsi, l'heuristique de décomposition est la plus performante en temps de calcul par rapport à la relaxation lagrangienne et la procédure d'agrégation. Par contre, à partir de la figure 3.8, nous remarquons que la résolution des instances de taille réelle est impossible en utilisant les trois méthodes proposées dans un temps de calcul raisonnable.

Temps de calcul et performances

On s'intéresse, dans ce paragraphe, à la comparaison et l'analyse de différents résultats obtenus liés au temps de calcul et les performances des trois méthodes de résolution. Nous présentons pour chaque instance :

— la valeur de la borne supérieure obtenue par la méthode d'agrégation et l'heuristique de décomposition ;

— la valeur de la borne inférieure obtenue par la relaxation lagrangienne ainsi que la valeur de la borne supérieure obtenue par l'heuristique de lissage.

Afin d'évaluer les solutions trouvées par rapport à la meilleure solution obtenue *i.e.* la solution optimale du MIP si elle est connue ou la meilleure borne supérieure (la plus faible) des trois heuristiques proposées, nous définissons le gap entre la solution trouvée

par la méthode heuristique et la meilleure solution qui se calcule comme suit :

$$GAP_{heuristique} = \frac{|\text{Solution méthode heuristique} - \text{Meilleure solution}|}{\text{Meilleure solution}}$$

La meilleure méthode de résolution correspond à celle ayant obtenue le gap le plus faible. Pour l'évaluation en terme de temps de calcul, nous avons calculé un gain en temps de calcul par rapport à la procédure de la méthode exacte. Le gain en temps de calcul s'écrit comme suit :

$$GAIN_{heuristique} = \frac{\text{Temps de calcul (méthode exacte)} - \text{Temps de calcul (méthode heuristique)}}{\text{Temps de calcul(méthode exacte)}}$$

Les résultats expérimentaux détaillés des comparaisons des trois méthodes appliquées aux différentes instances sont présentées dans les tableaux 3.7 et 3.8. Le tableau 3.7 illustre une évaluation de la qualité de la solution obtenue pour les trois méthodes en comparaison avec la solution optimale. $BS_{agrég}$, $BS_{décomp}$ et $BS_{Lagrange}$ représentent les bornes supérieures obtenues par la procédure d'agrégation, l'heuristique de décomposition et l'heuristique d'amélioration de la solution de la relaxation lagrangienne, respectivement. BI représente la valeur de la meilleure borne inférieure obtenue à la fin de la relaxation lagrangienne après 10 itérations. Le $GAP_{heuristique}$ mesure la qualité de la solution approchée. La figure 3.9 illustre les résultats de calcul de $GAP_{heuristique}$ des trois heuristiques. Le tableau 3.8 présente une comparaison des trois méthodes en termes de temps de calcul noté $Time_{heuristique}$ et calculé en millisecondes.

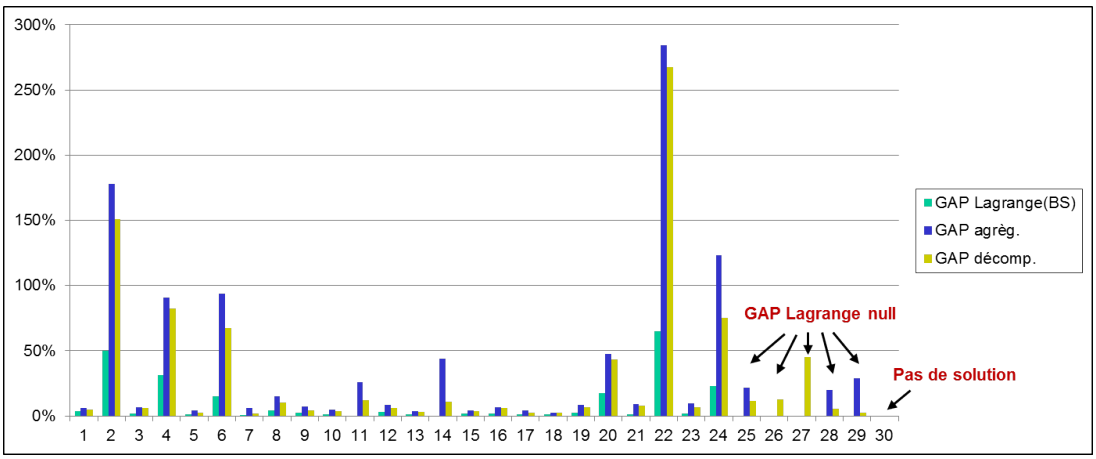

Figure 3.9 – Qualité de la solution de la procédure d'agrégation, l'heuristique de décomposition et l'heuristique post-lagrangienne.

D'après le tableau 3.7 et la figure 3.9, il est clair que la qualité de la solution de heuristique d'amélioration de la relaxation lagrangienne est meilleure que celle de la méthode d'agrégation et de décomposition. Pour les instances de taille réduite, le gap entre

Tableau 3.7 – Résultats expérimentaux : Comparaison de la qualité de la solution des trois méthodes pour résoudre le problème de planification à capacité finie

L	max S_l	I	Optimum	$BS_{agrég}$	$BS_{décomp}$	$BS_{Lagrange}$	BI	$GAP_{agrég}$	$GAP_{décomp}$	$GAP_{Lagrange}$
2	10	5	0,33	0,349	0,346	0,342	0,322	5,76%	5,01%	3,64%
2	200	20	1,55	4,31	3,88	2,32	0,645	178,06%	150,77%	49,68%
2	250	20	1,72	1,83	1,82	1,75	0,82	6,34%	5,73%	1,74%
3	6	20	0,07	0,13	0,127	0,092	0,039	90,93%	82,26%	31,43%
3	10	3	18,7	19,47	19,16	18,92	18,43	4,16%	2,51%	1,18%
3	150	5	7,98	15,47	13,34	9,2	5,39	93,85%	67,13%	15,29%
10	8	20	10,58	11,23	10,78	10,625	10,38	6,16%	1,89%	0,43%
10	60	6	0,28	0,31	0,302	0,29	0,26	14,76%	10,04%	3,94%
10	100	20	0,56	0,6	0,58	0,575	0,548	7,11%	4,39%	2,68%
20	100	20	5,46	5,74	5,64	5,52	5,39	4,71%	3,33%	1,10%
20	200	10	-	4,18	3,72	3,32	2,237	25,90%	12,05%	0%
30	50	10	1,37	1,49	1,45	1,41	1,33	8,67%	6,24%	2,92%
30	150	20	4,45	4,62	4,58	4,51	4,37	3,88%	3,06%	1,35%
40	100	20	-	9,34	7,19	6,49	6,3	43,91%	10,79%	0%
50	10	10	3,2	3,34	3,31	3,26	3,13	4,48%	3,36%	1,87%
50	50	10	5,8	6,19	6,16	5,89	5,54	6,89%	6,26%	1,73%
60	40	10	2,51	2,61	2,57	2,541	2,48	3,93%	2,29%	1,24%
70	5	10	5,8	5,95	5,94	5,86	5,68	2,53%	2,39%	1,03%
80	30	10	11,57	12,57	12,33	11,82	11,043	8,67%	6,24%	2,92%
80	50	10	8,7	12,82	12,48	10,23	6,46	47,34%	43,45%	17,59%
90	10	10	25,94	28,23	28,13	26,3	24,55	8,83%	8,08%	1,39%
100	30	10	17,75	68,22	65,26	29,3	4,94	284,34%	267,66%	65,07%
100	50	10	14,57	15,96	15,52	14,87	13,87	9,56%	6,55%	2,06%
200	2	10	7,6	16,97	13,3	9,32	5,006	123,34%	75,13%	22,63%
240	40	10	-	13,8	12,62	11,35	10,8	21,59%	11,19%	0%
700	100	20	-	-	25,7	23,9	22,3	-	12,55%	0%
1000	20	20	-	-	23,12	15,92	12,58	-	45,23%	0%
1700	1	10	-	14,37	12,65	11,98	10,68	9,95%	5,59%	0%
2000	2	20	-	7,92	6,3	6,145	5,9	28,89%	2,52%	0%
2000	680	300	-	-	-	-	-	-	-	-

la solution de l'heuristique de lissage et la solution optimale est faible. Cependant, ce dernier augmente avec le nombre de *steps* à planifier. Nous remarquons aussi que la solution de l'heuristique de décomposition est plus proche de la solution optimale par rapport à l'heuristique d'agrégation.

D'après le tableau 3.8, on peut conclure que l'heuristique de décomposition est plus performante en termes de temps de calcul par rapport à l'heuristique d'agrégation et la relaxation lagrangienne. La majorité des instances sauf l'instance industrielle ont été calculées dans un délai ne dépassant pas 1 min.

3.9 Conclusion

Dans ce chapitre, une formulation appropriée du problème en un programme linéaire mixte (*MIP*) a été proposée dont la fonction objectif est la minimisation de la somme des

Tableau 3.8 – Résultats expérimentaux : Comparaison du temps de résolution des trois méthodes pour résoudre le problème de planification à capacité finie

L	$\max S_l$	I	$Time_{MIP}(ms)$	$GAIN_{agrég}(\%)$	$GAIN_{décomp}(\%)$	$GAIN_{Lagrange}(\%)$
2	10	5	62	97,9%	98,8%	96,9%
2	200	20	12324	95,8%	97,2%	89,2%
2	250	20	13930	93,9%	96,4%	89,7%
3	6	20	147	80,2%	91,2%	78,4%
3	10	3	94	84,2%	89,4%	82,2%
3	150	5	6162	62,9%	75,3%	72,1%
10	8	20	540	65,8%	83,5%	79,6%
10	60	6	796	62%	79%	63,5%
10	100	20	8300	57,4%	69,7%	61,3%
20	100	20	interrompu	>42,3%	>52,7%	>45,5%
20	200	10	562	67,6%	81,9%	72,3%
30	50	10	2324	77,4%	84,2%	79,3%
30	150	20	288720	50,4%	69,8%	67,2%
40	100	20	interrompu	>32,9%	>45,6%	>35,7%
50	10	10	312	55%	73,6%	64%
50	50	10	858	43%	59%	48,7%
60	40	10	75	59%	69,7%	62,3%
70	5	10	47	62,8%	76,4%	63,4%
80	30	10	89	65,3%	79,9%	69,4%
80	50	10	317	63,7%	82,4%	71,9%
90	10	10	38	56,3%	69,1%	62,5%
100	30	10	168	49,3%	65,2%	57,3%
100	50	10	237	51,3%	57,8%	53,9%
200	2	10	252	96,2%	97,9%	95,8%
240	40	10	interrompu	>25,3%	>42,6%	>39,7%
700	100	20	interrompu	interrompu	>24,6%	>19,1%
1000	20	20	interrompu	interrompu	>21,3%	>11,6%
1700	1	10	interrompu	>11,6%	>29,7%	>15,6%
2000	2	20	interrompu	>13,5%	>43,2%	>22,9%
2000	680	300	interrompu	interrompu	interrompu	interrompu

retards pondérés. En testant le *MIP* sur des instances aléatoires de petites tailles, nous obtenons des plannings de production réalisables où toutes les contraintes considérées sont respectées. Toutefois, en augmentant la taille des instances testées, la résolution du problème est interrompue à cause des contraintes spatiales (manque de mémoire informatique) et temporelles (temps de calcul important).

La NP-difficulté du problème étudié implique qu'il est impossible de trouver un algorithme efficace garantissant la résolution optimale du problème. L'application des méthodes de résolution exacte n'est pas utilisable dans le cas des problèmes à grande échelle. Par conséquent, pour trouver des solutions à notre problème réel, nous avons essayé trois techniques de résolution, à savoir une procédure d'agrégation, une technique de décomposition et la relaxation lagrangienne. Des expérimentations, menées sur différentes instances générées aléatoirement, ont montré l'efficacité de la relaxation lagrangienne en termes de qualité de la solution tout en donnant des solutions proches de l'optimal pour des instances

de taille réduite. Cependant, cette méthode est moins performante que l'heuristique de décomposition en terme de temps de calcul. Malgré l'importance de ces techniques dans la réduction du temps de calcul et de l'amélioration de la limite de résolution, il s'est avéré qu'elles sont inefficaces au point de vue qualité de la solution et temps de calcul en testant des instances industrielles.

Ainsi, il est nécessaire de développer des méthodes de résolution plus adaptées à la structure du problème. Dans le chapitre suivant, des méthodes approchées sont proposées pour la résolution du problème considéré.

4

Résolution approchée du problème de planification à capacité finie

Résumé : D'après les résultats acquis au chapitre précédent, il a été montré que le problème de planification à capacité finie est NP-difficile au sens fort. Trouver une solution optimale pour les instances industrielles de grande taille demeure un challenge. Ainsi, pour résoudre le problème de projection des encours de production à capacité finie, deux approches de résolution approchée sont proposées. Il s'agit de deux algorithmes itératifs par période de l'horizon de planification. Les deux heuristiques sont composées par des modules. Le premier module commun correspond à la projection du *WIP* à capacité infinie. La première heuristique se compose d'un deuxième module se basant sur le *MIP* présenté dans le chapitre précédent. La deuxième heuristique se compose de deux autres modules : un deuxième module pour le calcul de la charge accumulée sur les parcs d'équipements et un troisième module pour l'équilibrage de la charge et la capacité des parcs d'équipements en cas de surcharge.

Les résultats apportés par ces approches ainsi qu'une évaluation de leur performance en termes de qualité de la solution et temps de calcul sont présentés en fin de ce chapitre.

Sommaire

4.1 Introduction 85
4.2 Projection du *WIP* à capacité infinie 85
4.2.1 Modèle du temps de cycle 86
4.2.2 Principe de la projection du *WIP* à capacité infinie 93
4.3 Heuristique de décomposition à base de *MIP* 94
4.4 Heuristique de décomposition à base d'algorithmes 99
4.4.1 Calcul de la charge accumulée 100
4.4.2 Equilibrage de la charge et la capacité 102
4.5 Résultats et discussion 108
4.5.1 Génération des instances aléatoires 108
4.5.2 Evaluation des algorithmes heuristiques proposés en comparaison à une solution optimale 109
4.5.3 Comparaison entre le processus réel et les résultats de l'heuristique à base d'algorithmes pour des instances industrielles 112
4.6 Conclusion 115

Les résultats développés dans ce chapitre ont été présentés dans les articles suivants :

[119] E. Mhiri, M. Jacomino, F. Mangione, P. Vialletelle, and G. Lepelletier. *Finite capacity planning algorithm for semiconductor industry considering lots priority.* IFAC-PapersOnLine (2015), 48(3), 1598-1603.

[120] E. Mhiri, M. Jacomino, F. Mangione, P. Vialletelle, and G. Lepelletier. *Prise en compte des priorités des lots pour la projection des encours de production dans l'industrie des semi-conducteurs.* In *16ème conférence ROADEF Société Française de Recherche Opérationnelle et Aide à la Décision*, Marseille, France (2015).

[121] E. Mhiri, F. Mangione, M. Jacomino, P. Vialletelle, and G. Lepelletier. *Approche heuristique pour la projection des encours de production (WIP) à capacité finie, application à l'industrie des semi-conducteurs.* In *17ème conférence ROADEF Société Française de Recherche Opérationnelle et Aide à la Décision*, Compiègne, France (2016).

4.1 Introduction

Dans le chapitre précédent, nous avons proposé des approches exactes pour résoudre le problème de projection du *WIP* à capacité finie dont le but est de minimiser la somme des retards pondérés. L'apport de ces méthodes est la modélisation du problème. Cependant, le temps d'exécution long ainsi que la mémoire requise pour une résolution exacte, rendent les algorithmes exacts inutilisables pour des problèmes réels. Afin de trouver une solution réalisable de bonne qualité au problème de planification de grande taille, les chercheurs se sont intéressés à des heuristiques de décomposition. Les approches de décomposition fonctionnent généralement sur une ou plusieurs dimensions du problème. En effet, les heuristiques de décomposition pour les problèmes de planification à moyen terme sont classées en deux catégories : les heuristiques période par période et les heuristiques référence par référence. Les heuristiques période par période fonctionnent souvent de manière récursive. La solution d'un problème à la période t est utilisée comme entrée au problème à la période $t + 1$. Une solution réalisable du problème initial est obtenue lorsque tous les sous-problèmes de chaque période sont résolus. Une telle heuristique doit s'assurer de la faisabilité de la solution des sous-problèmes. Les heuristiques référence par référence fonctionnent par ensembles de références. A chaque étape, un ensemble de références est planifié jusqu'à ce qu'un plan de production soit obtenu pour toutes les références.

Dans ce chapitre, nous continuons à explorer le problème de projection du *WIP* à capacité finie en développant des heuristiques de décomposition. En premier lieu, nous présentons une méthode basée sur une hybridation du programme linéaire mixte présenté dans le chapitre 3 et une stratégie de décomposition de l'horizon de planification afin de fournir une solution réalisable au problème traité. En second lieu, une heuristique de décomposition sur des algorithmes itérés par période de l'horizon de planification est proposée. Ces heuristiques correspondent à deux systèmes de planification considérant les contraintes de capacité et de qualifications des équipements ainsi que la priorité des lots et la variabilité du temps de cycle. Les deux systèmes se composent par des modules. Ils ont un premier module commun de projection du *WIP* à capacité infinie qui sera présenté au début de ce chapitre. Ensuite, l'algorithme de chaque heuristique sera détaillé en présentant les différents modules le composant. Ces approches sont testées sur des instances générées aléatoirement pour les valider et sur des instances industrielles afin d'évaluer leur performance en termes de qualité de la solution obtenue et temps de calcul.

4.2 Projection du *WIP* à capacité infinie

Prendre en compte des priorités des lots pour assurer une livraison des produits à temps et considérer la variabilité du processus de fabrication sont des objectifs principaux

dans cette thèse. Ainsi, le principe du premier module de planification est de proposer un modèle de temps de cycle qui tient compte de ces contraintes mais qui néglige, en premier lieu, les contraintes de capacité.

L'algorithme de projection du *WIP* prévoit la trajectoire de chaque lot par période, depuis sa position dans la route jusqu'à sa due date afin de connaître l'activité induite par période. En effet, il détermine pour chaque lot son propre modèle du temps de cycle basé sur :

— la vitesse nécessaire et suffisante des *steps* restants de chaque lot pour atteindre la due date,

— un modèle de temps de cycle objectif moyen pour chaque *step* qui tient compte de la variabilité du process, extrait de l'historique des données.

Pour le calcul du temps de cycle, une nouvelle variable notée $CTCoeff_l$ est introduite. Il s'agit d'un coefficient de temps de cycle pour chaque lot qui identifie sa vitesse nécessaire et suffisante pour atteindre la date d'échéance de livraison en tenant compte du temps de cycle objectif. Après avoir établi le modèle de temps de cycle pour chaque lot, l'algorithme de projection calcule la date de début et de fin de chacun des *steps* restants pour chaque lot ainsi que le nombre de *moves* à exécuter, estimé pour chaque période de l'horizon de planification noté M_t. Ce module est composé de trois étapes principales comme il est expliqué dans l'algorithme 2.

Dans ces projections, chaque lot a son propre modèle du temps de cycle qui représente le temps de traitement des *steps* restants en se basant sur la vitesse nécessaire et suffisante pour atteindre les dates d'échéance de livraison des lots. Dans la section suivante, on explique le modèle du temps de cycle considéré.

4.2.1 Modèle du temps de cycle

Le temps de cycle est le temps prévu écoulé depuis le début jusqu'à la fin d'un processus de production. Dans le cas où le processus de fabrication est composé d'une séquence distincte d'étapes de traitement telles que la fabrication des semi-conducteurs, le temps de cycle d'un lot est égal à la somme des temps de cycle individuels pour chaque étape de traitement avec l'hypothèse qu'une étape de traitement est indépendante de toute autre. Cette hypothèse est appliquée à l'estimation du temps de cycle dans la présente recherche, comme cela est représenté sur la figure 4.1.

Le temps de cycle de chaque *step* CTs_l est aussi décomposé en deux parties : un temps d'attente devant les équipements $wt_{s_l,l,i}$ et un temps de process $p_{s_l,l,i}$. La figure 4.2 illustre les composants du temps de cycle.

Algorithme 2 : Algorithme de projection du WIP à capacité infinie

Données :

- S_l Nombre de *steps* restants pour chaque lot l dans le *WIP*
- d_l Due date du lot l
- Q_l Quantité de *wafers* dans un lot l
- w_l Poids du lot l
- r_l Date de lancement du lot l
- $p_{s_l,l,i}$ Temps de process unitaire de chaque *step* s_l du lot l sur le parc d'équipements i
- $ObjCT_{s_l,l}$ Temps de cycle objectif moyen pour chaque *step* s_l du lot l

Résultat :

- $s_{s_l,l}$ Date de début du *step* s_l du lot l
- $e_{s_l,l}$ Date de fin du *step* s_l du lot l
- $x_{s_l,l,t}$ Variable de décision indiquant si un *step* s_l du lot l est traité pendant la période t ou non
- C_l Date de fin du lot l
- T_l Retard du lot l
- $CTCoeff_l$ Coefficient du temps de cycle du lot l
- M_t Nombre de *moves* par période t

début

pour chaque $t \in 1 \dots T$ **faire**

pour chaque $l \in 1 \dots L$ **faire**

pour chaque $s_l \in 1 \dots S_l$ **faire**

1 – Calculer un coefficient du temps de cycle $CTCOeff_l$ pour chaque lot l basé sur sa due date (voir section 4.2.1)

2 – Calculer les dates de début $s_{s_l,l}$ et les dates de fin $e_{s_l,l}$ pour tous les *steps* de tous les lots, les variables de décision des variables $x_{s_l,l,t}$, les dates de fin C_l et les retards des lots T_l

3 – Calculer le nombre de moves M_t pour chaque période t égale à $\sum_l \sum_{s_l} x_{s_l,l,t}$

fin pour

fin pour

fin pour

fin

Sur le plan opérationnel, le temps de process $p_{s_l,l,i}$ peut être divisé en :

- Un temps de traitement (PR), indiquant le temps opérationnel passé sur l'équipement,
- Un retard (DL) provoqué par le chevauchement de traitement des lots consécutifs.

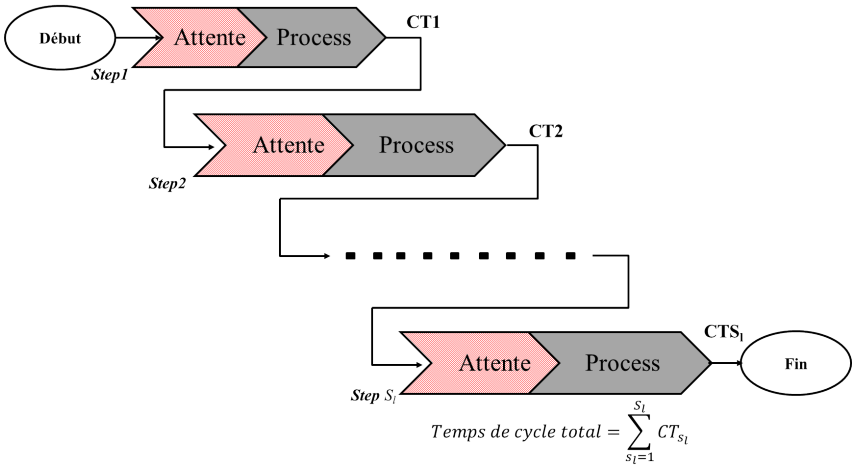

Figure 4.1 – Évaluation du temps de cycle total.

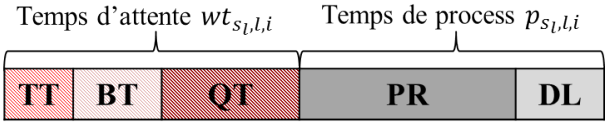

Figure 4.2 – Composants du temps de cycle total.

Le temps d'attente $wt_{s_l,l,i}$ peut aussi se répartir en trois composants :

- Le temps de transport (TT) correspondant au temps pendant lequel le lot circule entre les équipements de fabrication et les zones de stockage,
- Le temps de construction et de dissolution d'un *batch* (BT), pour les équipements fonctionnant par *batch*, indiquant à la fois le temps passé à attendre les autres lots pour former un *batch*, et le temps d'attente après le traitement du *batch*,
- Le temps d'attente restant (QT) désignant le temps passé à attendre qu'un équipement soit disponible.

Généralement, les temps de process des *steps* sont fixes et supposés être donnés, par contre, les temps d'attente sont très variables. Par conséquent, le principal défi dans la modélisation du temps de cycle réside dans la détermination des temps d'attente des *steps*, générés par la variabilité du système de production.

4.2.1.1 Modèle du temps de cycle développé

Comme il est précisé dans le chapitre 1, la méthode *MRP* présente des limites lorsqu'elle est appliquée à l'industrie des semi-conducteurs car elle considère des temps de cycle constants le long de l'horizon de planification. Par conséquent, cette technique ne tient pas compte des différences individuelles des caractéristiques des commandes [92]. En

outre, pour surmonter le problème de la capacité infinie, les temps de cycle constants sont souvent étendus pour pouvoir répondre plus facilement aux exigences de capacité. Cependant, cela aura des impacts négatifs sur la qualité du planning de la production et sur les performances industrielles, car il entraîne d'importants encours de production *(WIP)* et peut conduire à des défaillances dans le respect des délais de livraison des commandes clients [88, 144, 185].

Dans le contexte industriel considéré, les temps de cycle sont variables pour une même référence produit et dépendent des facteurs déterministes suivants :

1. L'utilisation des équipements
2. Les flux ré-entrants
3. La variabilité du temps de process en fonction du type de traitement de l'équipement (par *wafer* ou par *batch*) : Les différents types d'équipements provoquent un comportement différent de temps de process pour un lot. Dans les opérations de *batching*, le temps de process est indépendant du nombre de *wafers* traités. Dans le cas d'équipements fonctionnant par *wafer*, le temps de process peut être supposé linéaire par rapport à la taille du lot utilisé (cf. figure 4.3).
4. La taille du *batch* : La figure 4.4 illustre un aperçu qualitatif de la relation entre la taille du *batch* et le temps de cycle. Dans le cas d'un *batch* d'une grande taille, l'équipement doit attendre un long moment jusqu'à ce que le *batch* soit rempli et le processus puisse commencer. Dans le cas contraire, si la taille du *batch* est trop petite, le débit de l'opération de chargement du *batch* est très faible. La raison est la longueur du temps de process commun des équipements fonctionnant par *batch*.
5. Le *mix* produit engendrant une diversité des routes.
6. Les différentes priorités des lots : On trouve plusieurs classes de priorité des lots qui influencent la distribution des temps de cycle de *steps*. Les lots critiques *i.e.* présentant une urgence de livraison ont des priorités élevées ce qui exige un temps de cycle très court.

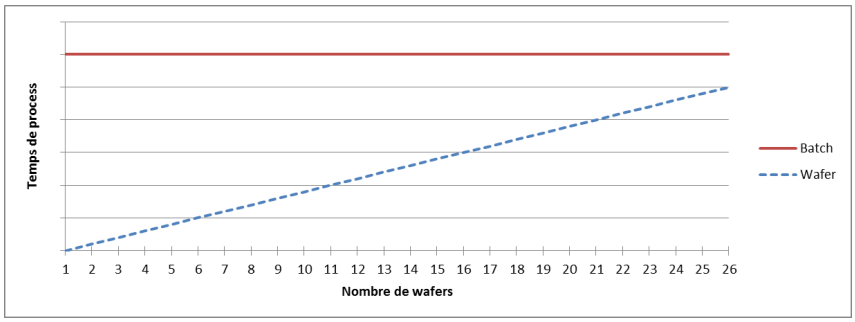

Figure 4.3 – Variabilité du temps de process en fonction du type de traitement de l'équipement

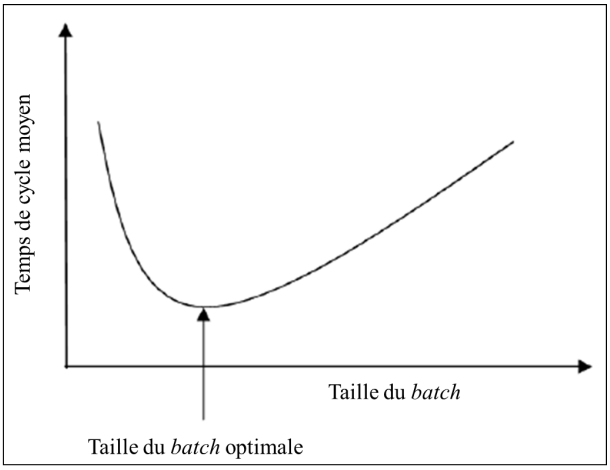

Figure 4.4 – Variabilité du temps de cycle en fonction de la taille du *batch*

Ainsi, proposer un modèle de temps de cycle qui prend en compte la variabilité du processus et les priorités des lots est essentiel pour développer un outil d'aide à la décision pour la planification de la production des *wafer fabs*.

Dans la littérature, certains modèles ont été développés pour le problème de de planification de la production avec les contraintes de temps de cycle cible [39, 43, 68]. Cependant, ces modèles sont basés sur un seul *mix* produit. Un *mix* produit désigne les ratios des demandes pour chaque famille de produits. Par exemple, le *mix* produit de trois familles de produits peut être A : B : C = 0,5 : 0,3 : 0,2. Un seul ratio du *mix* produit implique que le ratio de la demande est fixé le long de l'horizon de planification. Néanmoins, un *wafer fab* peut avoir divers *mixes* produits. La planification basée sur un *mix* produit fixe conduit souvent à une inadéquation entre l'offre et la demande lorsque le *mix* produit change considérablement. Cette incohérence peut dégrader les performances opérationnelles en engendrant un faible débit de production et un long temps de cycle. D'autres études portent sur le problème de planification pour multi-mix produit [10, 98, 158]. Cependant, leurs modèles ne se concentrent pas spécifiquement sur les *wafer fabs*, et donc le facteur de temps de cycle n'est pas considéré.

Les méthodes les plus utilisées pour déterminer le temps de cycle sont les réseaux de file d'attente [39, 43, 179] et la simulation à événements discrets [45, 68, 112, 113, 145].

La théorie des files d'attente a toujours été considérée non suffisamment précise pour estimer le temps de cycle, car elle ne prend pas en compte le facteur de variabilité de la *fab*. Le facteur de variabilité est défini par Li *et al.* [108] comme « un indicateur physique et significatif de la variabilité du système qui a une large influence sur le temps de cycle et il a deux composants qui sont un facteur de la variabilité du temps de process et un facteur de la variabilité des flux ». Li *et al.* ont défini ces deux types de variabilité comme deux ensembles de différents intervenants : dans le premier cas, la variabilité du temps de

process se manifeste dans la configuration et le *mix* produit/process, la reprise des lots, etc. Dans le second cas, la variabilité des flux se trouve dans les modes de fonctionnement, les règles de dispatching, etc.

Les méthodes de simulation sont précises pour la modélisation des *wafer fabs*. Cependant, leur développement nécessite beaucoup d'efforts et leur temps de calcul est très long.

Très souvent, les meilleures données disponibles des temps de cycle sont sous la forme des distributions observées précédemment [52]. En plus, vu la complexité de l'élaboration d'un modèle de temps de cycle qui tient compte des différentes sources de variabilité du processus de fabrication des semi-conducteurs, on s'est limité dans cette étude à considérer un modèle de temps cycle extrait de l'historique des données. En effet, le temps de cycle pour chaque *step* est calculé en se basant sur une formule semi-empirique. Cette formule multiplie le temps de process par un coefficient de variabilité appelé $Xfactor_{s_l,l}$ qui tient compte du taux d'utilisation des équipements dépendant des cinq premiers facteurs de variabilité cités. Ce facteur est extrait de l'historique des données. Le temps de cycle résultant présente le temps de cycle objectif pour un *step i.e.* la durée maximale que pourrait passer un lot à ce *step*, y compris le temps d'attente et le temps de process.

Dans le modèle de temps de cycle extrait de l'historique, la priorité des lots n'est pas considérée. Par conséquent, on propose des modifications sur le modèle du temps de cycle initial dans le but de mieux gérer les priorités des lots et réaliser l'un des objectifs principaux *i.e.* respecter les due dates des lots et minimiser les retards des livraison des lots de production.

Pour ce faire, comme il est indiqué dans l'algorithme de projection du WIP à capacité infinie, l'étape 1 consiste à calculer un coefficient de temps de cycle $CTCoeff_l$ pour chaque lot. Il est égal au rapport entre le temps de cycle estimé restant $RemExpCT_l$ et le temps de cycle objectif restant d'un lot $RemObjCT_l$. $CTCoeff_l$ augmente la priorité des lots en retard et accélère leur production et réduit la priorité des lots en avance et étend leur process.

$$CTCoeff_l = \frac{RemExpCT_l}{RemObjCT_l} \tag{4.1}$$

Le temps de cycle estimé restant d'un lot $RemExpCT_l$ est égal au maximum entre la différence entre sa date d'échéance de livraison et la date courante t et le temps de process

de ses steps restants $RemPT_l$.

$$RemExpCT_l = \max(d_l - t, RemPT_l) \tag{4.2}$$

$$RemPT_l = \sum_{s_l=1}^{S_l} \sum_{i=1}^{I} p_{s_l,l,i} \times Q_l \tag{4.3}$$

Le temps de cycle objectif restant d'un lot $RemObjCT_l$ est la somme des temps de cycle objectif de ses *steps* extraits de l'historique des données.

$$RemObjCT_l = \sum_{s_l=1}^{S_l} ObjCT_{s_l,l} \tag{4.4}$$

$$ObjCT_{s_l,l} = Xfactor_{s_l,l} \times p_{s_l,l,i} \tag{4.5}$$

Ensuite, pour chaque lot l, le $RemExpCT_l$ est divisé sur ses steps élémentaires pour calculer un temps de cycle attendu par *step* $ExpCT_{s_l,l}$ qui est égal au produit de $ObjCTs_l, l$ et $CTCoeff_l$.

$$ExpCT_{s_l,l} = CTCoeff_l \times ObjCT_{s_l,l} \tag{4.6}$$

L'équation (4.7) donne une estimation du temps d'attente pour chaque *step*. Par conséquent, le temps d'attente par *step* $wt_{s_l,l}$ peut être calculé :

$$wt_{s_l,l} = ExpCT_{s_l,l} - \sum_{i=1}^{I} p_{s_l,l,i} \tag{4.7}$$

Exemple 4.1. *Pour bien expliquer le principe de calcul du temps de cycle, la figure 4.5 illustre un exemple de 2 lots avec différentes dates d'échéance, ayant 3 steps restants chacun. Le premier a une date d'échéance plus tôt i.e. une priorité plus élevée et un $RemExpCT_l$ inférieur au second.*

En utilisant la projection classique basée sur des données historiques, les deux lots ont la même distribution des temps d'attente des steps restants durant l'horizon de planification. Cependant, le module de projection proposé répartit les temps de cycle prévus en tenant compte des priorités des lots i.e. les due dates des lots. En effet, comme il est mentionné auparavant, il existe de multiples niveaux de priorité des lots de production. Les priorités de production peuvent être divisées en deux niveaux selon l'urgence de la livraison : élevée (hot) et standard. Donc, pour respecter ces priorités, le module de projection établi rétrécit les temps d'attente des steps des lots prioritaires afin de satisfaire leurs dates d'échéance. Cependant, pour les lot standards, il étend les temps d'attente de leurs steps en respectant leurs due dates.

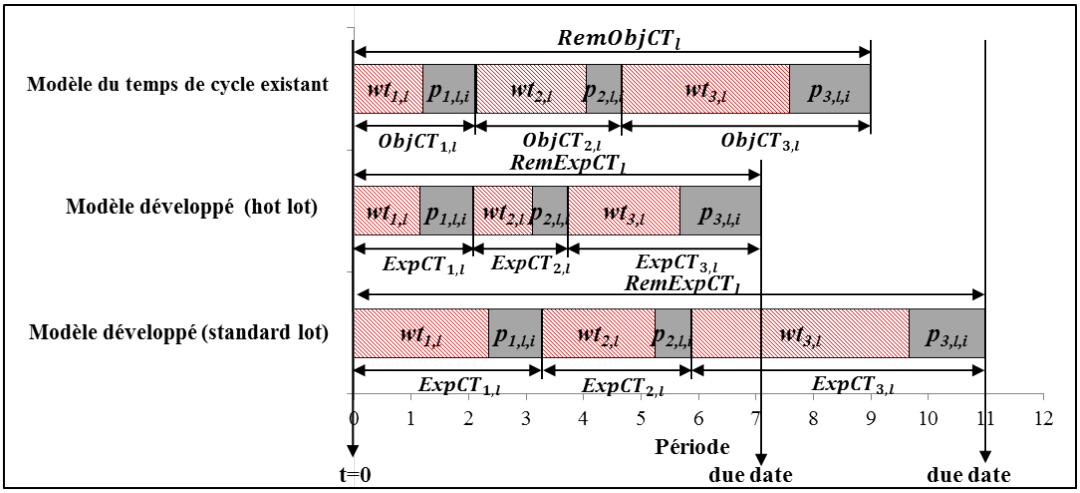

Figure 4.5 – Principe du calcul du temps de cycle des *steps*.

4.2.2 Principe de la projection du *WIP* à capacité infinie

Après avoir établi un modèle de temps de cycle des lots tenant compte de leurs niveaux de priorité, le module de projection du WIP détermine les dates de début et les dates de fin des steps de chaque lot ainsi que les dates de fin des lots et les retards de livraison et les variables de décision $x_{l,s_l,t}$ selon les contraintes temporelles ((3.2)...(3.11)), présentées dans le modèle mathématique dans le chapitre 3.

Exemple 4.2. *Pour expliquer davantage le concept de projection du WIP, une instance aléatoire simple est testée. Le WIP considéré est constitué de 10 lots de 25 wafers chacun, suivant des routes différentes, et ayant aussi de différentes due dates. Le tableau 4.1 présente, pour chaque lot, le nombre de steps restants, $RemPT_l$, $RemObjCT_l$, $RemExpCT_l$ et $CTCoeff_l$.*

Tableau 4.1 – Données d'une simple instance

Lot l	**Poids** w_l	**Nombre de *steps* restants** S_l	$RemPT_l$ **en jours**	$RemObjCT_l$ **en jours**	$RemExpCT_l$ **en jours**	$CTCoeff_l$
Lot 1	0,33	6	1,1	1,6	5	3,125
Lot 2	1	4	0,8	1,1	0,5	0,45
Lot 3	0,5	2	0,25	0,41	1,5	3,65
Lot 4	0,5	8	1,7	2,3	1,5	0,65
Lot 5	0,5	6	1	1,4	1,5	1,07
Lot 6	0,33	4	0,75	1,02	5	4,9
Lot 7	0,5	8	0,86	1,05	1,5	1,43
Lot 8	1	4	0,8	1,05	0,5	0,48
Lot 9	0,5	4	0,8	1,05	1,5	1,43
Lot 10	0,5	6	1,4	1,9	1,5	0,79

La figure 4.6 illustre les résultats de la projection des 10 lots au cours de la première période de l'horizon de planification. Pour certains lots, une séquence d'étapes est répétée

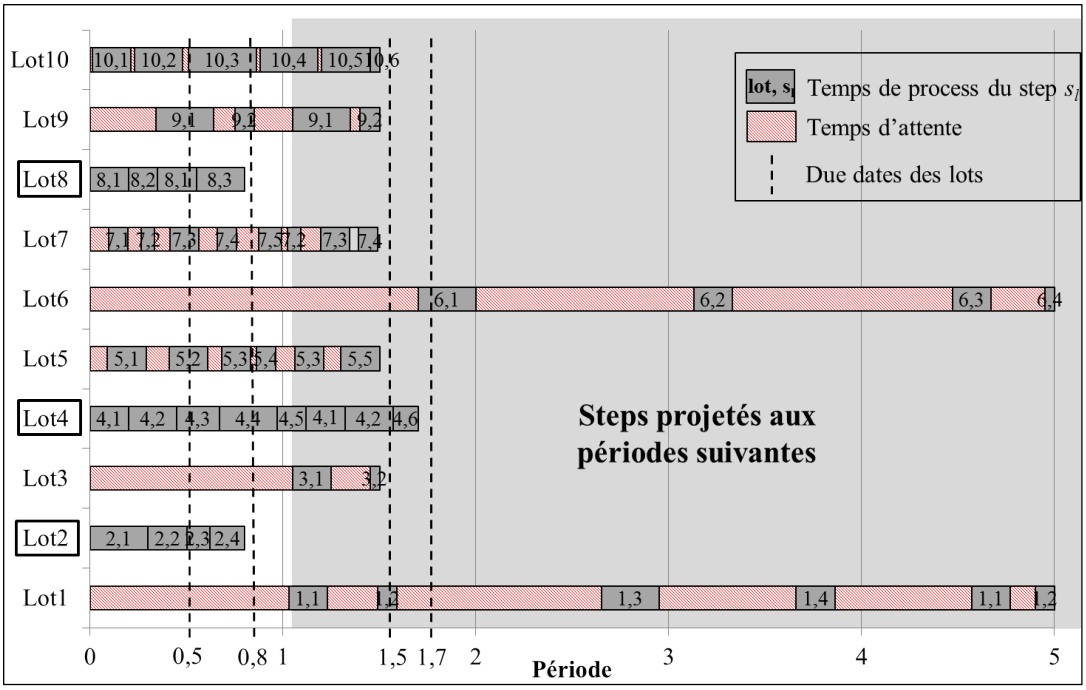

Figure 4.6 – Simple instance : planning de production à capacité infinie.

deux fois (les lots 1,4,5,7,8,9) i.e. les lots visitent le même parc d'équipements deux fois ce qui illustre la réentrance des flux. La figure 4.6 présente les dates de début et de fin, les temps d'attente et le temps de process pour chaque step restant du processus de fabrication au cours de la période considérée. Certains steps (step 4.5 et step 10.4) commencent dans la première période et finissent dans les périodes ultérieures de l'horizon de planification. Cette figure montre que le moteur de projection permet l'extension des temps d'attente des steps restants pour les lots où il y a une grande marge entre la date courante et leurs due dates ce qui est le cas des lots 1 et 6 et il rétrécit les temps de cycle des steps dans le cas où les dates d'échéance sont proches de la date courante tel est l'exemple des lots 2, 4 et 8. Les lots 2, 4 et 8 ne sont pas livrés à temps. Leurs due dates ne sont pas réalisables donc leurs dates de sortie sont égales à la somme de la date actuelle ($t = 0$) et la durée restante du temps de process $RemPT_l$.

4.3 Heuristique de décomposition à base de *MIP*

Cette approche est une méthode itérative, elle résout à chaque période de l'horizon de planification en deux étapes un problème réduit. L'objectif de cette approche est de réduire le nombre de variables et le nombre de contraintes. Cette réduction permet de diminuer le temps de calcul. Le principe de l'heuristique est de commencer par sélectionner une période t de l'horizon de planification. Ensuite, la projection du *WIP* à capacité infinie est effectuée de la période t sélectionnée jusqu'à la fin de l'horizon de planification selon

l'algorithme décrit dans la section 4.2. A partir de cette projection, on détermine le nombre de lots projetés pendant la période t noté L_t, le nombre maximum de steps projetés pour chaque lot pendant la période t noté $S_{l,t}$ et une due date périodique pour chaque lot notée $d_{l,t}$. Cette dernière est égale au minimum entre la due date du lot d_l et la date de fin du lot à la période t *i.e.* la date de fin du dernier *step* du lot projeté à la période t notée $C_{l,t}$.

$$d_{l,t} = \min d_l, C_{l,t} \tag{4.8}$$

Dans la seconde étape de l'algorithme, la projection de l'ensemble des lots L_t sur la période t est refaite en utilisant le *MIP* décrit dans la section 3.4 du chapitre 3. L'objectif de cette re-projection est de considérer les contraintes de capacité. La résolution du *MIP* est effectuée sur une seule période *i.e.* la période t sélectionnée ce qui permet de réduire le nombre de variables et de contraintes par rapport au MIP initial. La résolution du *MIP* permet de mettre à jour les variables $s_{s_l,l}$, $wt_{s_l,l}$ et $e_{s_l,l}$ des *steps* restants de l'ensemble des lots à projeter pendant la période t ($S_{l,t}$) ainsi que les dates de fin et les retards de l'ensemble des lots L_t. Elle permet aussi de déterminer la charge accumulée sur les équipements $L_{i,t}$ tout en respectant les contraintes de capacité et d'identifier le *WIP* à projeter à la prochaine période. La fonction objectif est de minimiser la somme des retards pondérés des lots pendant la période t. Toutes les contraintes présentées dans le *MIP* au chapitre 3 sont maintenues identiques mais exécutées seulement sur la fenêtre de temps $[t, t+1[$. Ces deux étapes sont répétées jusqu'à atteindre la fin de l'horizon de planification. Le retard d'un lot l calculé sur tout l'horizon est déterminé à la dernière itération de l'algorithme. Il est égal à $T_{l,T}$.

La figure 4.7 illustre les étapes de cet algorithme.

Les notations utilisées dans le *MIP* à une période t sont présentées dans le tableau 4.2.

Le programme linéaire mono-période peut alors s'écrire comme suit (PL.2) :

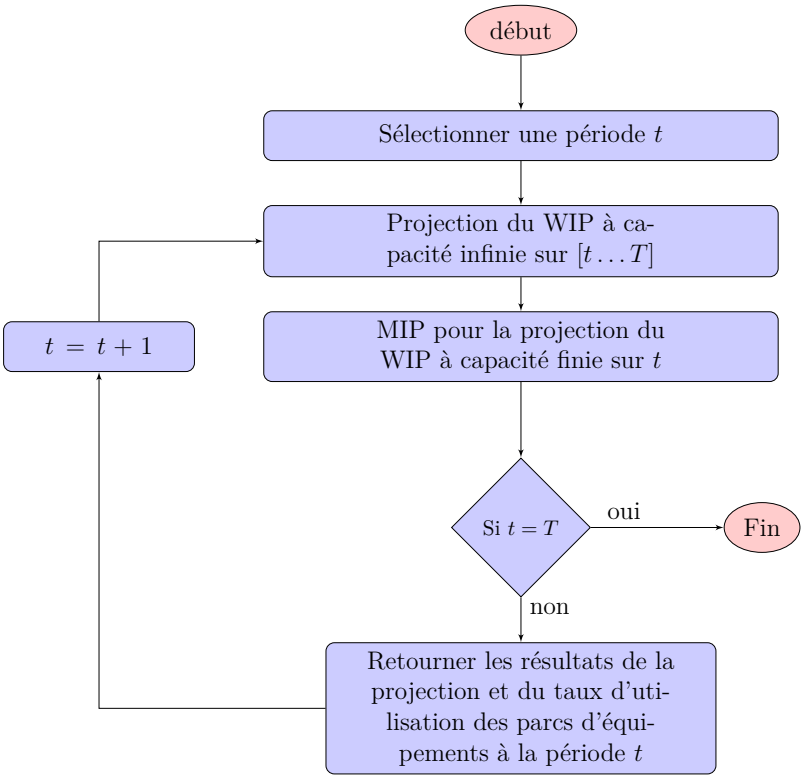

Figure 4.7 – Algorigramme de l'heuristique à base de *MIP*

Tableau 4.2 – Notations pour le *MIP* mono-période

Indices	**Description**
L_t	Nombre de lots à projeter à la période t
$l = 1..L_t$	Indice du lot à projeter à la période t
$S_{l,t}$	Maximum nombre de *steps* restants du lot l à projeter à la période t
$s_{l,t} = 1..S_{l,t}$	Indice du *step* du lot
I_t	Nombre de parcs d'équipements exécutant les steps $S_{l,t}$
$i_t = 1..I_t$	Indice du parc d'équipement à utiliser pendant la période t
Paramètres	**Description**
Q_l	Quantité de *wafers* pour chaque lot l
r_l	Date de lancement du lot l
w_l	Poids du lot l
$d_{l,t}$	*Due date* du lot l pendant la période t
$p_{s_{l,t},l,i}$	Temps de process unitaire du *step* s_l du lot l sur le parc d'équipements i
$C_{i,t}$	Capacité du parc d'équipements i sur la période t
$Q_{s_{l,t},l,i}$	= 1 si le *step* $s_{l,t}$ du lot l est qualifié sur le parc d'équipements i, 0 sinon
$a_{s_l,l,i,t}$	Quantité de *wafers* du lot l au *step* s_l, pouvant être traité par le parc d'équipements i, pendant la période t
Variables de décision	**Description**
$s_{s_{l,t},l}$	Date de début d'un *step* s_l du lot l
$e_{s_{l,t},l}$	Date de fin d'un *step* s_l du lot l
$C_{l,t}$	Date de fin du lot l pendant la période t
$T_{l,t}$	Retard du lot l pendant la période t
$L_{i,t}$	Charge du parc d'équipement i pendant la période t
$y_{s_l,l,t}$	$= s_{s_l,l}$ si le *step* s_l du lot l est lancé pendant la période $[t,t+1[$, 0 sinon
$x_{s_l,l,t}$	=1 si le *step* s_l du lot l est exécuté pendant la période t, 0 sinon

Programme linéaire 2 (PL.2).

$$min \quad \sum_{l} w_l T_{l,t} \tag{4.9}$$

$$s.c. \quad s_{1,l} \geq r_l \qquad l = 1, \ldots, L_t \tag{4.10}$$

$$s_{s_{l,t},l} + \sum_{i} p_{s_{l,t},l,i} \times Q_l \leq e_{s_{l,t},l} \qquad s_{l,t} = 1, \ldots, S_{l,t},\ l = 1, \ldots, L_t \tag{4.11}$$

$$s_{s_{l,t},l} = e_{s_{l-1,t},l} \qquad s_{l,t} = 1, \ldots, S_{l,t},\ l = 1, \ldots, L_t \tag{4.12}$$

$$\sum_{t} y_{s_{l,t},l,t} = s_{s_{l,t},l} \qquad s_{l,t} = 1, \ldots, S_{l,t},\ l = 1, \ldots, L_t \tag{4.13}$$

$$\sum_{t} x_{s_{l,t},l,t} = 1 \qquad s_{l,t} = 1, \ldots, S_{l,t},\ l = 1, \ldots, L_t \tag{4.14}$$

$$C_{l,t} = e_{S_{l,t},l} \qquad l = 1, \ldots, L_t \tag{4.15}$$

$$T_{l,t} \geq C_{l,t} - d_{l,t} \quad l = 1, \ldots, L_t \tag{4.16}$$

$$T_{l,t} \geq 0 \qquad l = 1, \ldots, L_t \tag{4.17}$$

$$t \times P_t \times x_{s_{l,t},l,t} \leq y_{s_{l,t},l,t} \qquad s_{l,t} = 1, \ldots, S_{l,t},\ l = 1, \ldots, L_t, \tag{4.18}$$

$$(t+1) \times P_t \times x_{s_{l,t},l,t} > y_{s_{l,t},l,t} \qquad s_{l,t} = 1, \ldots, S_{l,t},\ l = 1, \ldots, L_t \tag{4.19}$$

$$L_{i,t} = \frac{1}{Q_l} \times \sum_{l} \sum_{s_{l,t}} p_{s_{l,t},l,i} \times a_{s_{l,t},l,i,t} \times Q_{s_{l,t},l,i} \times x_{s_{l,t},l,t} \qquad i = 1, \ldots, I_t \tag{4.20}$$

$$L_{i,t} \leq C_{i,t} \qquad i = 1, \ldots, I_t \tag{4.21}$$

$$x_{s_l,l,t} = \{0, 1\} \qquad s_l = 1, \ldots, S_l,\ l = 1, \ldots, L_t \tag{4.22}$$

Exemple 4.3. *Afin de mieux comprendre cette approche, nous considérons l'exemple suivant : 3 lots avec un nombre maximum de steps respectifs égal à 5,4 et 3 traités sur un seul parc d'équipements. L'horizon de planification est composé de 3 périodes (jours). Les différents paramètres sont présentés dans le tableaux 4.3.*

Tableau 4.3 – Paramètres de l'exemple

L	S_l	d_l	w_l	$p_{s_l,l,i}$
1	5	3	0.1	[0.1, 0.09, 0.08, 0.14, 0.07]
2	4	2	0.3	[0.08, 0.08, 0.12, 0.06]
3	3	2	0.3	[0.04, 0.07, 0.08]

La figure 4.8 illustre les résultats de l'heuristique appliquée à l'exemple donné pour la première période.

Après avoir effectué la projection à capacité infinie sur les trois périodes de l'horizon

de temps, on obtient $L_1 = 3$, $S_{l,1} = \{2, 2, 2\}$ *et* $d_{l,1} = 0.9, 0.95, 1$. *L'application du MIP modifie le planning initial obtenu pour la période 1. Les steps 1.2 et 3.2 sont décalés à la période 2 à cause des contraintes de capacité.*

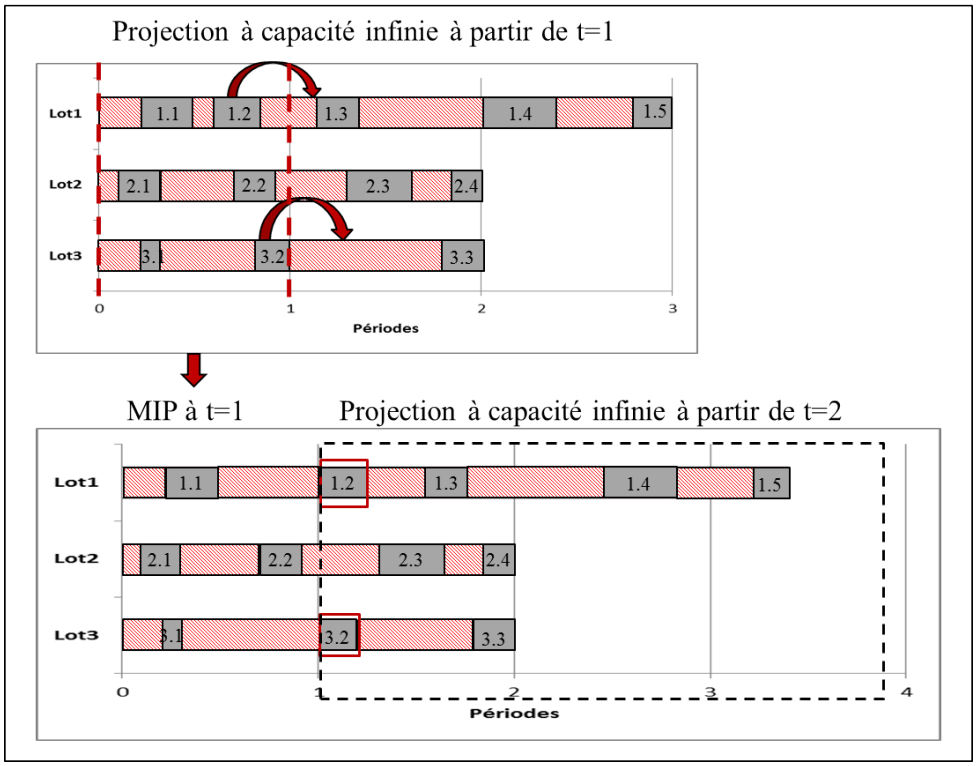

Figure 4.8 – Instance simple expliquant le principe de l'heuristique à base de *MIP*

Le planning final obtenu pour cette instance en exécutant l'heuristique sur les 3 périodes de l'horizon de planification est illustré dans la figure 4.9. On remarque que la livraison du lot 1 est retardée de 0,45 jours et la livraison du lot 2 est retardée de 0,4 jours. La solution conduit à un retard total pondéré de 0,165 jours.

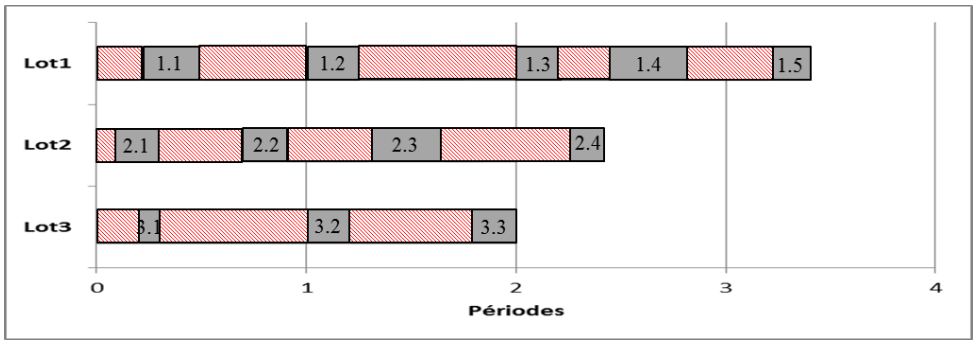

Figure 4.9 – Planning obtenu à la fin du test de l'heuristique sur l'exemple 4.3

4.4 Heuristique de décomposition à base d'algorithmes

La deuxième heuristique proposée est aussi un algorithme itératif composé de trois modules principaux : (i) la projection du *WIP* à capacité infinie, (ii) l'accumulation de la charge de travail et l'analyse des capacités et (iii) l'équilibrage de la charge et la capacité des parcs d'équipements. L'algorithme est exécuté par itérations sur les périodes de l'horizon de planification.

Pour chaque période définie, le module de projection du *WIP* estime l'évolution du *WIP*, lot par lot, en se basant sur les due dates des lots comme il est expliqué dans la section 4.2. Ensuite, le module d'accumulation de charge calcule le taux de chargement prévu pour chaque parc d'équipements. En cas de surcharge des parcs d'équipements, le module d'équilibrage est utilisé pour réduire la charge de ces derniers en décalant des *steps* qui lui sont affectés à des périodes ultérieures. Le module de projection à capacité infinie est déjà présenté dans la section 4.2 alors que les deux autres modules seront présentés en détails dans les sections suivantes. La figure 4.10 illustre l'algorigramme du système de planification développé.

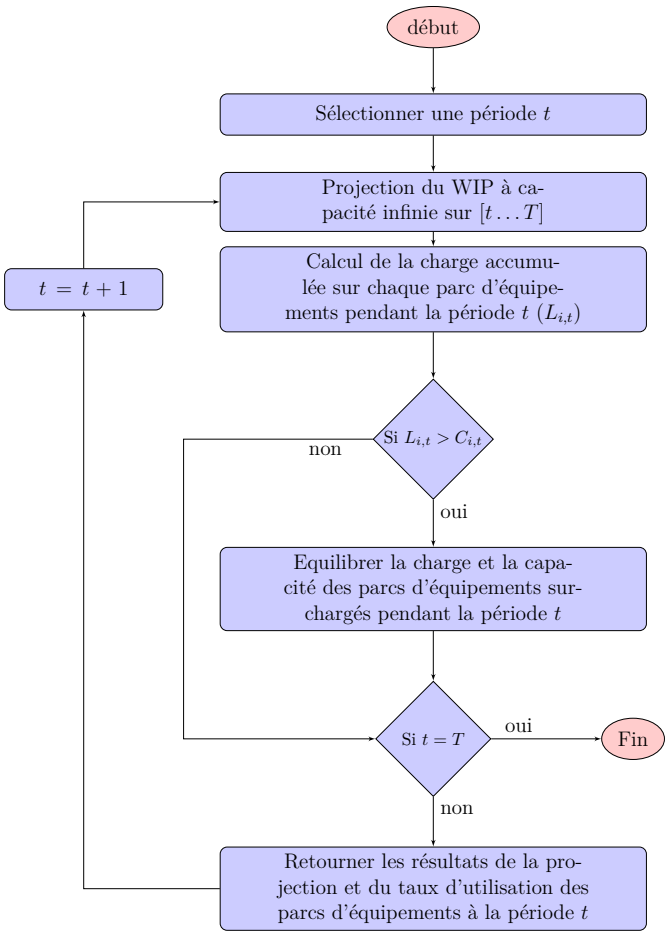

Figure 4.10 – Algorigramme de l'heuristique à base d'algorithmes

4.4.1 Calcul de la charge accumulée

Après la projection du *WIP*, la charge accumulée sur chaque parc d'équipements, au cours de la période considérée notée $L_{i,t}$, est calculée sur la base de l'hypothèse d'une capacité de production infinie et en considérant les contraintes de qualifications des équipements.

Les entrées pour ce module sont :

- Le nombre de *moves* sur la période t (résultats du module de projection),
- Le modèle des équipements indiquant le nombre d'équipements dans le parc d'équipements, la capacité de chaque parc d'équipements,
- le modèle des recettes indiquant les qualifications des recettes aux équipements avec les temps de process correspondants.

Pour optimiser le temps de calcul, les parcs d'équipements sont répartis en groupes d'équilibrage (*Balancing groups*). Un *balancing group* est un ensemble de parcs d'équipements qui ont les mêmes qualifications et partagent les mêmes recettes. Cette approche permet de décomposer le problème en petits sous-problèmes.

Pour répartir la charge de travail, sur la période sélectionnée, entre les différents parcs d'équipements, appartenant au même *balancing group* en tenant compte des contraintes de qualifications, un programme linéaire (PL.3) est proposé. Les notations utilisées dans le programme linéaire (PL.3) sont présentées dans le tableau 4.4.

En utilisant les paramètres et les variables de décision ci-dessus, nous pouvons représenter la formulation du programme linéaire de la manière suivante (PL.3) :

Programme linéaire 3 (PL.3).

$$min \quad \alpha \cdot Lmax - \beta \cdot Lmin + \gamma \cdot \sum_r Lmax_r - \delta \cdot \sum_r Lmin_r + \delta \cdot (\sum_i^I L_i - Lmin) \tag{4.23}$$

$$avec \quad \alpha = I_b^2, \beta = I_b, \gamma = 1, \delta = 1/I_b$$

$$s.c. \quad L_i = \sum_r p_{r,i} \times W_{r,i} \qquad i = 1, \ldots, I_b \tag{4.24}$$

$$\sum_{i=1}^{I_r} W_{r,i} = \sum_l \sum_{s_l} x_{s_l,l,t} \times Q_l \times a_{l,s_l,r} \times Q_{s_l,l,i} \qquad r = 1, \ldots, R \tag{4.25}$$

$$L_i \geq Lmin_r \qquad r = 1, \ldots, R, i = 1, \ldots, I_r \tag{4.26}$$

$$L_i \leq Lmax_r \qquad r = 1, \ldots, R, i = 1, \ldots, I_r \tag{4.27}$$

$$L_i \geq Lmin \qquad i = 1, \ldots, I_b \tag{4.28}$$

$$L_i \leq Lmax \qquad i = 1, \ldots, I_b \tag{4.29}$$

Tableau 4.4 – Notations pour le (PL.3)

Indices	**Description**
B	Nombre de *balancing groups*
$b = 1..B$	Indice du *balancing group*
R_b	Nombre de recettes liées au *balancing group* b
$r = 1..R_b$	Indice de la recette
I_b	Nombre de parcs d'équipements composant le *balancing group*
I_r	Nombre de parcs d'équipements qualifiés pour la recette r, $I_r \subseteq I_b$
$i = 1..I_b$	Indice du parc d'équipement
Paramètres	**Description**
$x_{l,s_l,t}$	Variables de décision, résultats du module de projection du *WIP* sur la période t
$C_{i,t}$	Capacité du parc d'équipements i sur la période t
$a_{l,s_l,r}$	=1 si la recette r correspond au step s_l du lot l, 0 sinon
$Q_{sl,l,i}$	Qualification du parc d'équipements i pour le step s_l du lot l
$p_{r,i}$	Temps de process de la recette r sur le parc d'équipements i
Q_l	Quantité de *wafers* des lots
Variables de décision	**Description**
L_i	La charge du parc d'équipements i
$W_{r,i}$	La quantité de *wafers* produite par le parc d'équipements i qualifié pour la recette r
$Lmax$	La charge du parc d'équipements le plus chargé dans le *balancing group*
$Lmin$	La charge du parc d'équipements le moins chargé dans le *balancing group*
$Lmax_r$	La charge, pour une recette donnée r, du parc d'équipements le plus chargé parmi ceux sur lesquels r est qualifiée
$Lmin_r$	La charge, pour une recette donnée r, du parc d'équipements le moins chargé parmi ceux sur lesquels r est qualifiée

Le choix de la pondération de la fonction objectif est justifié par le fait qu'on cherche à :

- Minimiser la charge du parc d'équipements le plus chargé dans le *balancing group* notée $Lmax$.
- Maximiser la charge du parc d'équipements le moins chargé dans le *balancing group* notée $Lmin$.
- Minimiser la charge de travail totale des parcs d'équipements $\sum_i^I L_i$ et maximiser la charge de travail totale du parc d'équipements le moins chargé par recette $\sum_r Lmin_r$, avec le même degré de priorité.
- Minimiser la charge de travail totale des parcs d'équipements les plus chargés par recette $\sum_r Lmax_r$.

Nous nous retrouvons donc face à une agrégation de critères. Comme l'explique Rossi [146], les fonctions objectifs agrégées présentent l'inconvénient majeur de nécessiter la détermination d'une pondération des critères acceptables pour l'utilisateur. Dans le cas du problème de répartition, il nous est au moins possible d'estimer un ordre de grandeur relatif des pondérations. En effet, nous introduisons cette nouvelle fonction objectif afin d'améliorer le critère minimisation de $Lmax$. Cela signifie que la minimisation de $Lmax$

doit être dominante devant les autres critères. Dans un deuxième temps, nous souhaitons équilibrer la répartition trouvée. Intuitivement, on cherche ici une répartition qui réduit l'écart entre la charge de l'équipement goulot d'étranglement et l'équipement le moins chargé. Le critère de maximisation de la charge de l'équipement le moins chargé est donc secondaire. Ceci constitue notre première approximation. Si m équipements sont qualifiés pour une même recette, on souhaite que l'équipement le plus chargé parmi les m, ait une charge la plus faible possible. Réciproquement mais avec une importance moindre, on souhaite que la charge de l'équipement le moins chargé parmi les m, ait la charge la plus élevée possible. Enfin, on peut souhaiter que la répartition calculée soit la plus économique possible.

Ainsi, les objectifs du module de calcul de la charge accumulée sont : (i) anticiper les principaux goulets d'étranglement pendant la période considérée et (ii) ajuster le plan de production en tenant compte des équipements goulots identifiés.

Exemple 4.4. *En reprenant l'exemple 4.2, les steps restants des 10 lots sont considérés traités par 6 parcs d'équipements* $\{M1, M2, M3, M4, M5, M6\}$*. Ces parcs d'équipements sont répartis en 4 balancing groups* $\{M1, M6\}, \{M2, M4\}, \{M3\}$ *et* $\{M5\}$*.*

La figure 4.11 illustre le pourcentage de saturation des parcs d'équipements i.e. le rapport entre la charge et la capacité disponible au cours de la première période de l'horizon de planification ($\frac{L_{i,1}}{C_{i,1}}, i = 1..6$) *lors du process des steps restants ordonnés par ordre croissant de la date de début.*
Dans cet exemple, la capacité de tous les parcs d'équipements considérés ($C_{i,1}, i = 1..6$) *est égale à 24 heures/jour. La figure 4.11 montre qu'il y a deux parcs d'équipements saturés* ($M2$ *et* $M6$) *dont la charge dépasse le seuil de la capacité.*

4.4.2 Equilibrage de la charge et la capacité

Après le calcul de la charge accumulée sur les parcs d'équipements à capacité infinie, on peut trouver des parcs d'équipements dont la charge dépasse la capacités maximale *i.e.* les contraintes 3.13 du MIP (PL.1) ne sont pas satisfaites. Dans ce cas, le parc d'équipements saturé est incapable de traiter tous les *steps* qui lui sont affectés pendant la période considérée. Par conséquent, sa charge doit être équilibrée et lissée sur des périodes ultérieures. Le problème d'équilibrage de la charge de travail a attiré l'attention de plusieurs chercheurs ces dernières années. Ham [75] a appliqué des heuristiques et l'algorithme Min-Max à une ligne de production flexible et a équilibré la charge de travail entre les opérations afin d'accroître la productivité par substitution d'une route.

Swarnkar et Tiwari [160] ont décrit un problème d'équilibrage d'une machine d'un système de fabrication flexible, dont les objectifs sont bi-critères : (i) minimiser le déséquilibre

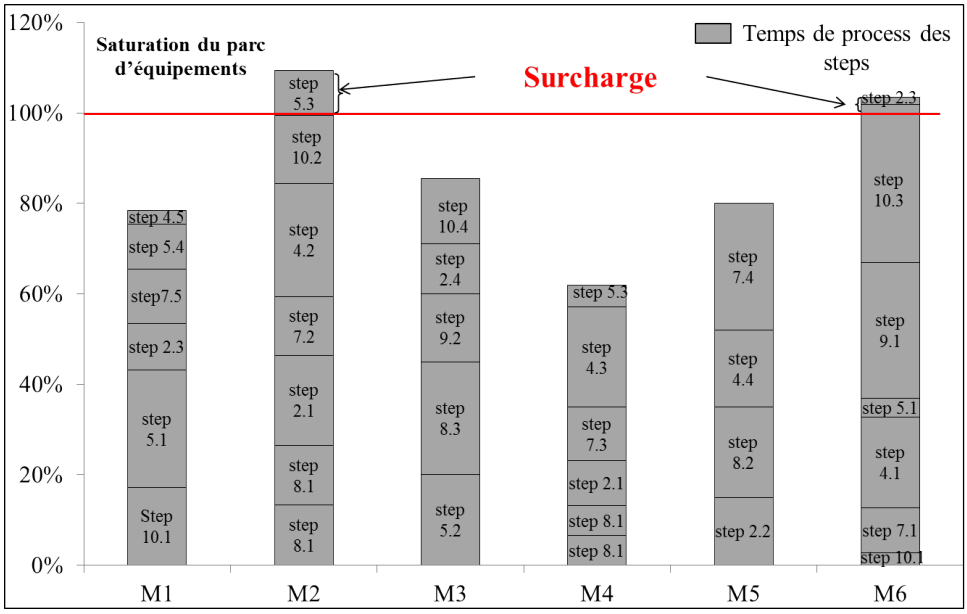

Figure 4.11 – Calcul de la charge à capacité infinie.

du système et (ii)maximiser le débit en présence de contraintes technologiques.

Toba *et al.* [164] ont proposé une méthode pour l'équilibrage de la charge pour équilibrer toutes les opérations de process des produits semi-conducteurs entre plusieurs *wafer fabs* en utilisant les résultats de la planification prédictive.

Henrich, Terre et Gaalman [78] ont indiqué que le contrôle efficace de la charge de travail exige une décision de regroupement des machines.

Chung et Jang [42] ont développé un *MIP* pour le nouveau concept d'équilibrage du *WIP* qui tient compte de la charge des équipements goulots d'étranglement pour un débit plus élevé.

Chen *et al.* [33] ont proposé un système de planification à capacité infinie pour les installations de test des circuits intégrés.Ce système comporte un module d'équilibrage de la charge par ajustement des dates de début des lots. Dans le cas de surcharge à une période $t > 0$, ce module avance le traitement des lots à des périodes antérieures.

Dans notre étude, pour l'équilibrage de la charge, nous avons proposé de décaler les *steps* des lots traités sur les parcs d'équipements saturés pendant la période t considérée à la période $t + 1$. Le décalage des *steps* est effectué tout en tenant compte des due dates des lots, des dates de passage des *steps* sur les parcs d'équipements et la répartition de la quantité produite de *wafers* sur les parcs d'équipements appartenant à un même *balancing group*. Un coefficient de priorité, noté $rankingCoeff_l$, est calculé pour chaque lot l indiquant sa priorité en termes de sa position dans la séquence de process du parc d'équipements i et l'urgence de sa livraison traduite par son coefficient de temps de cycle

$CTCoeff_l$:

$$rankingCoeff_l = \frac{1}{CTCoeff_l} + \frac{po_{l,i}}{P_t} \tag{4.30}$$

La position du lot l dans la séquence de process du parc d'équipements i, notée $po_{l,i}$, est déterminée par la date de début du *step* s_l. Elle est égale à la différence entre la date courante et la somme de la date de début du *step* s_l et son temps d'attente devant le parc d'équipement i.

$$po_{l,i} = t - (s_{s_l,l} + wt_{s_l,l}) \tag{4.31}$$

L'objectif de ce décalage est de réduire la charge des parcs d'équipements et lisser l'activité de production sur l'horizon de planification.

Les données pour cet algorithme sont :

- $L_{i,t}$ Charge du parc d'équipements i sur la période t (résultat du module d'accumulation de charge) $\forall i \in \{1 \dots I\}$
- $C_{i,t}$ Capacité du parc d'équipements i sur la période t $\forall i \in \{1 \dots I\}$
- $s_{s_l,l}$ Date de début du *step* s_l du lot l projeté pendant la période t $\forall l \in \{1 \dots L\}, \forall s_l \in \{1 \dots S_l\}$
- $wt_{s_l,l}$ Temps d'attente du *step* s_l du lot l
- $a_{s_l,l,i,t}$ Quantité de *wafers* du lot l au *step* s_l, pouvant être traité par le parc d'équipements i, pendant la période t après décalage

Les résultats de l'algorithme d'équilibrage sont :

- $L'_{i,t}$ Charge du parc d'équipements i sur la période t après équilibrage $\forall i \in \{1 \dots I\}$ égale à l'ancienne charge en enlevant le temps de process unitaire du *step* s'_l multiplié par le pourcentage du nombre de *wafers* produits par le parc d'équipement i au step s'_l $(a_{s'_l,l,i,t}/Q_l)$.
- M'_t Nombre de *moves* de la période t
- WIP de la période $t+1$

La figure 4.12 illustre l'algorigramme du module d'équilibrage de la charge.

L'algorithme pour le module d'équilibrage de la charge pour une période t est le suivant :

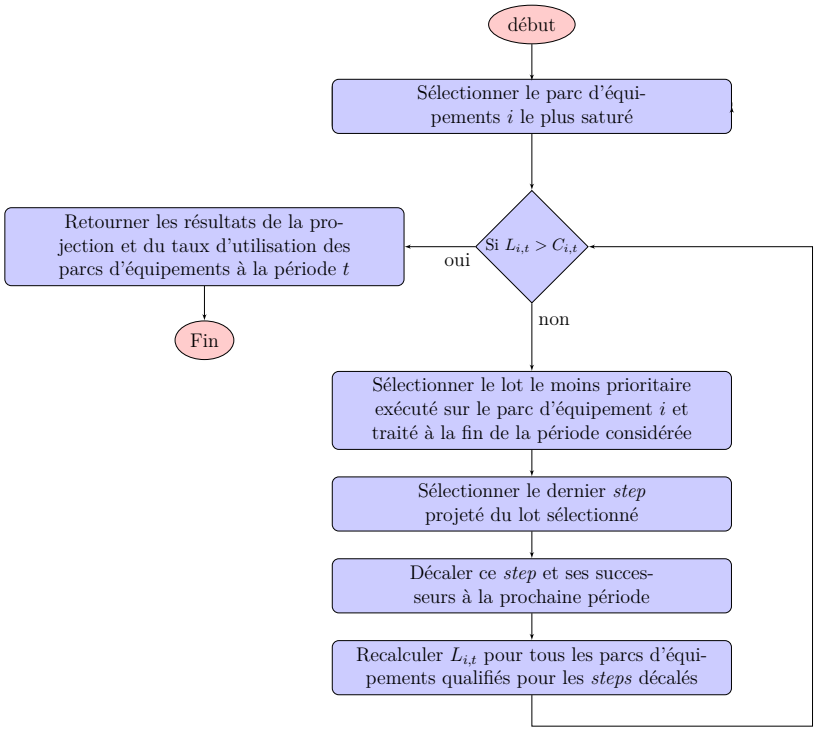

Figure 4.12 – Algorigramme de l'équilibrage de la charge

Exemple 4.5. *Pour mieux expliquer le principe de l'algorithme d'équilibrage de la charge, on reprend l'exemple 4.2. Dans cet exemple, on a trouvé après le calcul de la charge accumulée que les parcs d'équipements M2 et M6 sont saturés. En appliquant l'algorithme d'équilibrage, on sélectionne M2 comme parc d'équipements le plus saturé* ($\frac{L_{2,1}}{C_{2,1}} = 109.3\%$). *Ensuite, on sélectionne les lots 2, 4, 5, 7, 8 et 10 traités par ce parc d'équipements (4.11). Ces lots sont triés par ordre croissant de* $rankingCoeff_l$ *comme il est mentionné dans le tableau 4.5.*

Afin de diminuer la charge du parc d'équipements M2, les steps 5.3 et 7.2 et ses successeurs sont déplacés à la prochaine période de l'horizon de planification. Par conséquent, la charge de M2 devient inférieure à sa capacité maximale : $\frac{L_{6,1}}{C_{6,1}} = 96.4\%$. *M4 est également qualifié pour le step 5.3, de sorte que sa charge diminue de 5.06%. Le step 5.4 projeté sur la première période est également reporté car il est le successeur du step 5.3 décalé. Ainsi, la charge de M1 exécutant le step 5.4 devient égale à 58.83%. Le décalage des successeurs du step 7.2 (steps 7.3, 7.4 et 7.5) conduit à la diminution de la charge des parcs d'équipements M1, M4 et M5. Le même algorithme est appliqué au parc d'équipements M6 par le décalage du step 9.1 et de son successeur le step 9.2. Ainsi, sa charge diminue à 72%. La charge des parcs d'équipements obtenue après le décalage des steps est illustré à la figure 4.13.*

Le tableau 4.6 présente le WIP et les paramètres calculés ($RemPT_l$, $RemObjCT_l$, $RemExpCT_l$ *et* $CTCoeff_l$*) au début de la période suivante.*

Algorithme 3 : Algorithme d'équilibrage de la charge

– Trier les parcs d'équipements par ordre décroissant de saturation (charge/capacité). Soit I' l'ensemble des parcs d'équipements triés.
pour chaque $i \in 1 \dots I'$ **faire**
  si *(*$L_{i,t} > C_{i,t}$*)* **alors**
    pour chaque $l \in 1 \dots L$ **faire**
      pour chaque $s_l \in S_l \dots 1$ **faire**
        si *(*$x_{s_l,l,t} \times a_{s_l,l,i,t} > 0$*) (i.e. le step* s_l *est exécuté sur le parc d'équipements* i*)* **alors**
          –Identifier la position $po_{l,i}$du lot l dans la séquence de process du parc d'équipements i
          –Calculer le coefficient de priorité $rankingCoeff_l$ pour le lot l :
        fin si
      fin pour
    fin pour
    –Trier les lots par ordre croissant du $rankingCoeff_l$. Soit L' l'ensemble des lots triés
    pour chaque $l \in 1 \dots L'$ **faire**
      pour chaque $s_l \in S_l \dots 1$ **faire**
        si *(*$x_{s_l,l,t} \times a_{s_l,l,i,t} > 0$*)* **alors**
          pour chaque $s'_l \in s_l \dots S_l$ **faire**
            –Décaler le *step* s'_l à la période $t+1$ *i.e.* mettre à jour les variables de décision $s_{s'_l,l}$ au début de la prochaine période
            pour chaque $i \in 1 \dots I$ **faire**
              si *(*$x_{l,s'_l,t} \times a_{s'_l,l,i,t} > 0$*) i.e. le parc d'équipements* i *est qualifié au step* s'_l **alors**
                –Calculer la charge $L'_{i,t}$:
                $L'_{i,t} = L_{i,t} - \sum_{l'} \sum_{s'_l} (\frac{a_{s'_l,l,i,t}}{Q_l}) \times p_{s'_l,l,i} \times x_{l,s'_l,t} \times Q_{s'_l,l,i}$
              fin si
              –mettre à jour les variables de décision $x_{s'_l,l,t}$
              – Calculer le nombre de *moves* M'_t de la période t :
              $M'_t = \sum_l \sum_{s_l} x_{s_l,l,t}$
            fin pour
          fin pour
        fin si
      fin pour
    fin pour
  fin si
fin pour

L'approche proposée est testée sur un horizon de planification de cinq jours. En effet, vu la variation des données industrielles utilisées pour cette étude, il a été décidé de se

Tableau 4.5 – Ordre des lots traités par $M2$ selon $rankingCoeff_l$

Lot l	**CTCoeff$_l$**	**Step** s_l	**Date de début**	**RankingCoeff$_l$**
Lot 5	1,07	Step 5.3	0,68	1, 25
Lot 7	1,43	Step 7.2	0,265	1, 43
Lot 10	0,79	Step 10.2	0,23	2, 04
Lot 4	0,65	Step 4.2	0,2	2, 34
Lot 8	0,48	Step 8.1	0	3, 08
Lot 2	0,45	Step 2.1	0	3, 22

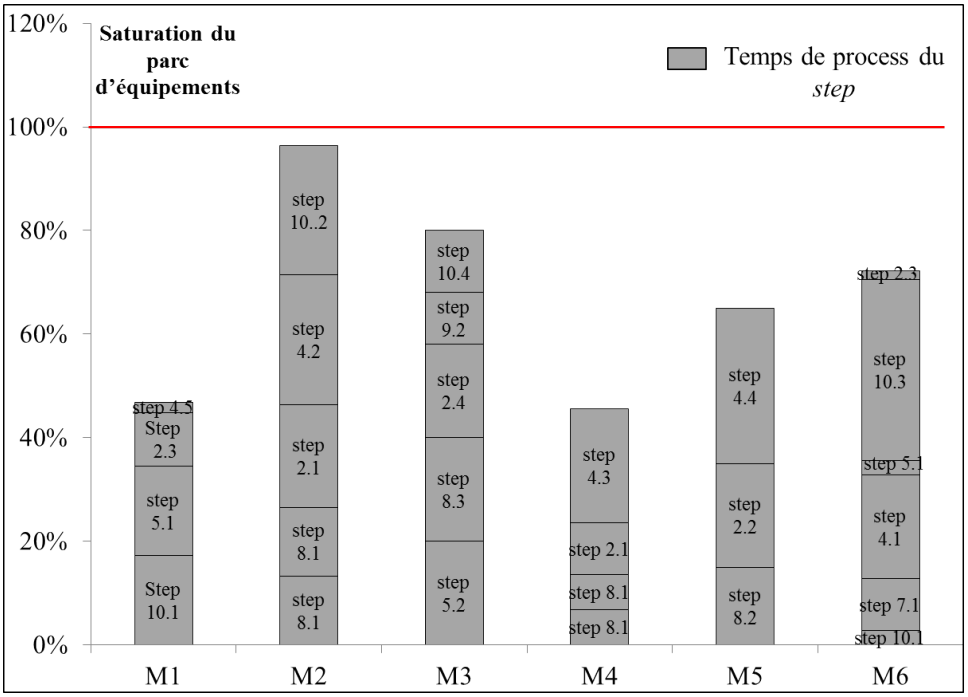

Figure 4.13 – Saturation des parcs d'équipements après équilibrage de la charge

concentrer sur un horizon de planification très court pour évaluer l'approche proposée.

Il est clair que plus la période est courte, plus les résultats de l'approche sont précis. Le planning final obtenu pour cette instance est illustré sur la figure 4.14. Pour cet exemple,

Tableau 4.6 – Les paramètres du *WIP* au début de la seconde période

Lot l	**Poids** w_l	**Nombre de *steps* restants** S_l	$RemPT_l$ **en jours**	$RemObjCT_l$ **en jours**	$RemExpCT_l$ **en jours**	$CTCoeff_l$
Lot 1	0,33	6	1,1	1,6	4	2,5
Lot 3	0,5	2	0,25	0,41	0,5	1,22
Lot 4	0,5	3	0,58	0,76	0,7	0,92
Lot 5	0,5	4	0,6	0,83	0,5	0,6
Lot 6	0,33	4	0,75	1,02	4	3,92
Lot 7	0,5	7	0,76	0,91	0,5	0,55
Lot 9	0,5	4	0,8	1,05	0,5	0,47
Lot 10	0,5	2	0,3	0,41	0,5	1,22

le TWT est égal à 1.46 jours et on a cinq lots retardés.

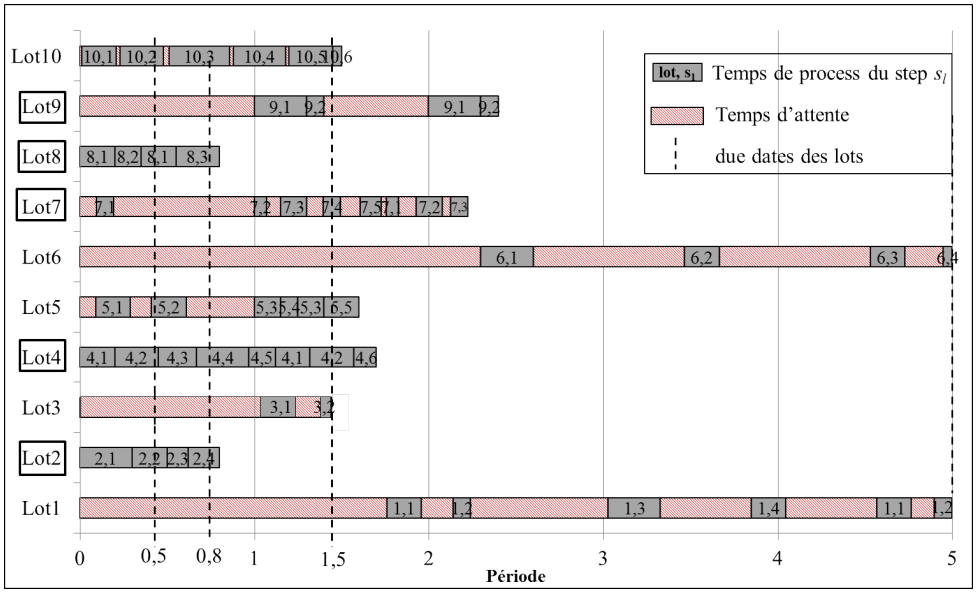

Figure 4.14 – Le planning de l'instance obtenu en utilisant l'heuristique à base d'algorithmes

4.5 Résultats et discussion

Les algorithmes proposés sont codés en `JAVA` et ils sont testés sur un ordinateur personnel avec un processeur `intel`® `Core`™ `i3-4130 CPU 2,40 GHz` et 4,00 `GO` de RAM. Nous avons effectué deux types d'expériences pour évaluer la performance des approches proposées. Le premier type correspond à une comparaison entre la méthode exacte et l'heuristique en utilisant un ensemble d'instances générées aléatoirement. Dans le second type d'expérience, on compare les résultats de l'approche proposée en utilisant des données réelles avec le processus réel dans le *wafer fab*.

4.5.1 Génération des instances aléatoires

Afin de valider et tester la qualité des méthodes de résolution proposées, nous avons généré aléatoirement de façon uniforme, des problèmes et nous avons fait varier différents paramètres. Pour le paramètre nombre de lots L, cinq valeurs sont proposées (L = 10, 15, 20, 50 et L = 100). Les paramètres générés pour les exemples proposés sont présentés dans le tableau 4.7

Alors, nous obtenons trois instances aléatoires pour chaque combinaison de paramètres fixée, donnant un total de 270 problèmes à tester.

Tableau 4.7 – Paramètres des tests

Parametère du problème	**Valeurs utilisées**	**Nombre total de valeurs**
Nombre de lots (L)	10, 15, 20, 50, 100	5
Maximum nombre de *steps* restants du lot l ($\max S_l$)	10, 20, 30, 40, 50, 100	6
Nombre de parcs d'équipements (I)	5, 10, 20	3
Nombre de périodes (T)	24	1
Poids du lot w_l	Loi uniforme (0,1)	1
Dates de lancement des lots r_l	0	1
Due dates des lots d_l	Loi uniforme (1,30)	1
Quantité de *wafers* des lots Q_l	25	1
Temps de process unitaire des *steps* $p_{s_l,l,i}$	0.0001× Loi uniforme(5,50)	1
	Nombre total de combinaisons de paramètres	90
	Nombre des instances par problème	3
	Nombre total des problèmes	270

4.5.2 Evaluation des algorithmes heuristiques proposés en comparaison à une solution optimale

Chacune des instances des 270 problèmes générés a été résolue en utilisant le solveur ILOG CPLEX et les algorithmes heuristiques proposés. Les résultats de TWT obtenus pour chaque instance en utilisant le modèle de MIP et en utilisant chacun des algorithmes itératifs proposé sont enregistrés.

En outre, pour chaque instance de taille $L \times \max S_l \times I$, nous calculons :

— L'écart absolu= |Valeur de TWT de l'algorithme heuristique - Valeur optimale de TWT|

— L'écart relatif= $\frac{\text{Ecart absolu}}{\text{Valeur optimale de TWT}}$

4.5.2.1 Comparaison entre la solution optimale et l'heuristique à base de MIP

En comparaison avec le MIP, cet algorithme permet de réduire énormément le nombre de variables et de contraintes. Le tableau 4.8 illustre une comparaison entre le MIP et cette heuristique.

La figure 4.15 illustre la limite de résolution de l'heuristique en considérant la condition d'arrêt *i.e.* un temps maximal de calcul égal à $5min$. Nous remarquons que cette heuristique permet de projeter au maximum 10000 *steps* ce qui est très réduit par rapport à la taille des instances industrielles. Bien que cet algorithme permette d'obtenir un plannnig de production réalisable en utilisant les données réelles, il ne satisfait pas l'objectif en

Tableau 4.8 – Comparaison MIP vs. heuristique

	MIP multi-périodes	**MIP mono-période**
Nombre de variables	69 011 200	84000
Nombre de contraintes	70 742 400	148000
Résolution	impossible	possible
Temps de calcul		1heure30min
Solution		≈ 20% des lots en retard TWT= 254 jours

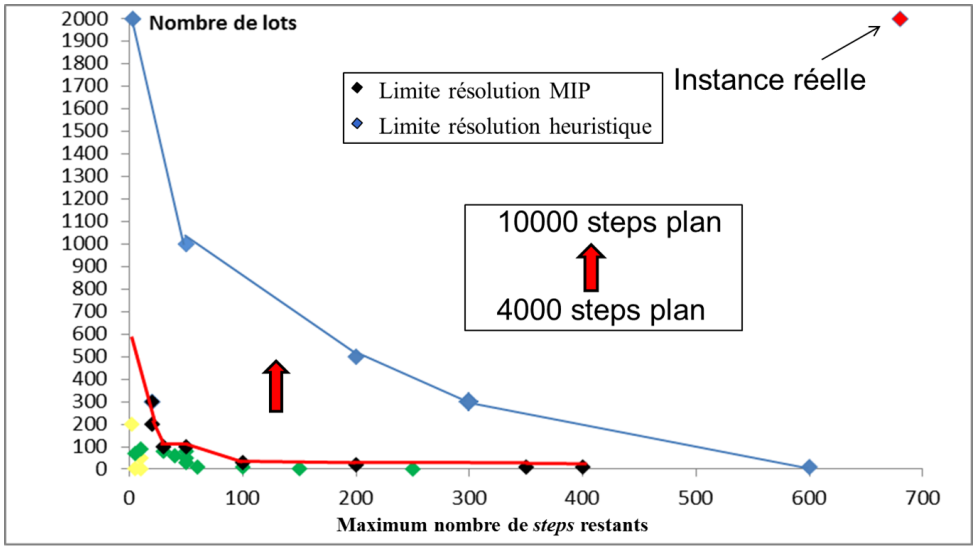

Figure 4.15 – Limite de la résolution de l'heuristique en comparaison avec celle du MIP

terme de temps de calcul (5 minutes au maximum). Le temps de résolution du MIP pour une période est long, il est égal à environ 20 minutes. Le nombre d'itérations augmente davantage le temps de résolution du MIP pour le problème entier. La solution obtenue est trouvée en environ 1heure et 30 minutes pour un horizon de 24 périodes pour l'instance industrielle considérée.

4.5.2.2 Comparaison entre la solution optimale et l'heuristique à base d'algorithmes

Sur la base des résultats de TWT obtenus pour chaque instance, la solution heuristique correspond exactement à la solution optimale dans 53 cas. La figure 4.16 représente l'écart relatif en fonction de l'écart absolu calculé pour les 270 instances aléatoires.

Dans cette figure, on peut définir quatre zones ou classes en fonction de la taille de l'instance :

— La première zone (correspondant aux valeurs de l'écart absolu $\in$ [0..30] jours et aux valeurs de l'écart relatif$\leq$ 1) : Environ 92% des instances testées sont situées

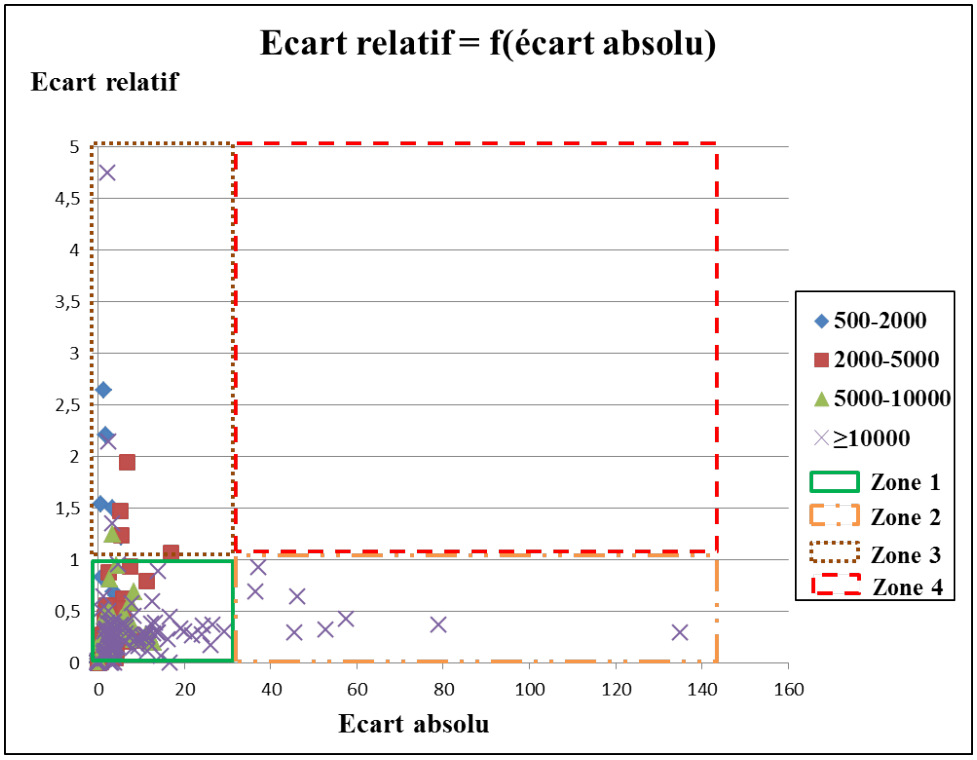

Figure 4.16 – Comparaison entre la solution optimale et l'heuristique à base d'algorithmes

dans cette zone. Par conséquent, dans la plupart des cas, la solution heuristique est proche de la solution optimale.

— La seconde zone (correspondant aux valeurs de l'écart absolu $\in$]30..140] jours et aux valeurs de l'écart relatif ≤ 1) : Les 8 instances ($\simeq 3\%$ du total des instances testées) appartenant à cette catégorie sont des instances de grande taille (≥ 10000). Par exemple, on trouve une instance de taille égale à 10000 (L=100, max S_l=10, I=10) qui a une déviation absolue égale à 79 jours et un écart relatif égal à 0.36. Cette instance a une solution TWT optimale égale à 218 jours. Ainsi, la valeur importante de l'écart absolu n'est pas significative en raison des valeurs élevées de TWT.

— La troisième zone (correspondant aux valeurs de l'écart absolu $\in$ [0..30] jours et aux valeurs de l'écart realtif $>> 1$) : 14 instances ($\simeq 5\%$ du total des instances testées) sont situées dans cette zone. On peut citer l'exemple de l'instance avec une taille égale à 15000 (L=50, max S_l=30, I=10), une faible valeur de l'écart absolu égale à 2.23 jours et une valeur élevée de l'écart relatif égale à 4.74. Pour cette instance, la solution optimale et la solution approximative présentent une faible valeur de TWT. Par conséquent, dans cette zone, l'importance de l'écart relatif n'a pas de signification.

— La quatrième zone (ce qui correspond aux valeurs de l'écart absolu > 30 jours et aux valeurs de l'écart relatif > 1) : Aucune instance ne se trouve dans cette zone,

caractérisée par des valeurs élevées des écarts absolus et relatifs.

4.5.3 Comparaison entre le processus réel et les résultats de l'heuristique à base d'algorithmes pour des instances industrielles

L'objectif de cette section est d'évaluer la capacité de l'approche proposée pour résoudre les problèmes réels. Le test de l'instance réelle ($L = 2000$ $maxS_l = 680$, $I = 300$, $T = 24$), non résolue dans un temps d'exécution raisonnable en utilisant l'approche MIP, la procédure d'agrégation, l'heuristique de décomposition et la relaxation Lagrangienne présentées dans le chapitre 3, est traité. Le temps d'exécution de cette instance avec l'algorithme proposé est d'environ 30 secondes. Dans le planning de production obtenu, 80% des lots projetés sont livrés à temps. En outre, la saturation des parcs d'équipements est maintenue au-dessous du seuil prédéfini tout en minimisant les retards des lots.

La figure 4.17 illustre la saturation hebdomadaire obtenue à capacité infinie et finie d'un parc d'équipements de photo-lithographie considéré comme un goulot d'étranglement. Comme nous avons précisé dans la section 3.2.3 du chapitre 3, pour mesurer la

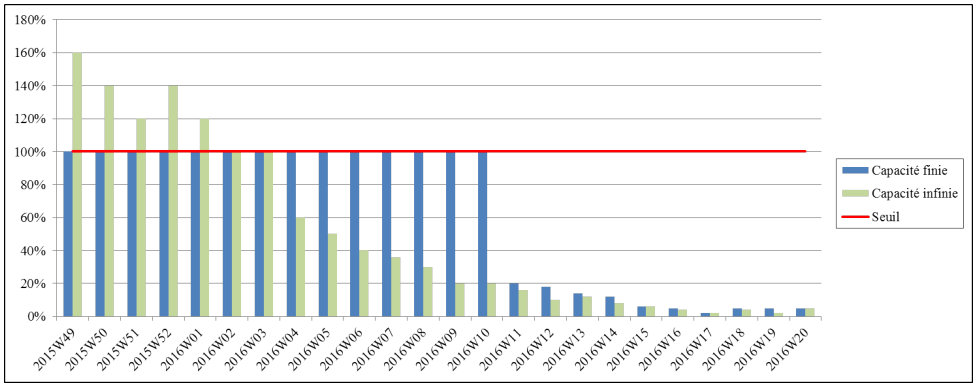

Figure 4.17 – Saturation hebdomadaire d'un parc d'équipements de photo-lithographie à capacité infinie et finie

performance des approches proposées, on a recours aux indicateurs de performance suivants :

— *Nombre de moves* : Ceci correspond au nombre de steps accomplis sur chaque période de l'horizon de planification, qui peut être comparé au nombre de *moves* réel dans la ligne de production.

— *Nombre de moves par usage* : Il s'agit du nombre de *steps* traités par un ensemble de parcs d'équipements appartenant au même atelier de production nommé « usage » sur chaque période de l'horizon de planification.

— *Retard total pondéré TWT* : Cet indicateur est utilisé pour évaluer les temps d'attente des lots pour le traitement.

Dans cette section, nous comparons les indicateurs de performance cités calculés pour la solution heuristique avec les indicateurs déterminés dans la ligne de production réelle. Pour assurer cette expérience, six essais ont été effectués sur des instances réelles extraites de quatre mois de production. La projection a été effectuée sur six périodes différentes (semaine1, semaine2, semaine3, semaine4, Semaine5 et semaine6) et les trois indicateurs ont été déterminés pour chaque projection. Après avoir effectué la projection, nous avons regardé ce qui est passé réellement dans l'usine pendant les périodes considérées et nous avons extrait les indicateurs de performance. Pour des raisons de confidentialité, nous ne sommes pas autorisés à fournir les valeurs réelles de la *fab*. Voilà pourquoi, nous calculons l'écart relatif entre la valeur estimée et la valeur réelle pour chaque période de l'horizon de planification :

$$\text{Écart relatif} = \frac{|\text{Valeur estimée} - \text{valeur réelle}|}{\text{valeur réelle}}$$

Analyse basée sur la mesure de performance : nombre de *moves*

La figure 4.18 montre des écarts relatifs du nombre de *moves* sur 15 périodes (semaines) de l'horizon de planification.

Elle montre que, dans les 6 premières périodes pour les différentes instances, l'écart relatif entre le nombre réel de *steps* de process et la valeur calculée est faible. La moyenne des écarts relatifs moyens de plus de six périodes pour les différents tests est égale à 12.7%, ce qui reflète une faible différence entre le nombre estimé de *moves* et celui réalisé. En s'éloignant du début de la projection, l'écart relatif entre la solution obtenue et le nombre réel de *moves* augmente ce qui s'explique par la forte variabilité du processus de fabrication et aussi par l'effet de l'incertitude non prise en compte dans cette étude. Par conséquent, il existe une convergence entre ce qui est estimé et ce qui est réalisé en termes d'activité périodique pour un horizon de planification à court terme.

4.5.3.1 Analyse basée sur la mesure de la performance : nombre de *moves* par usage

Pour évaluer dans quelle mesure la solution heuristique anticipe la charge du *wafer fab*, nous calculons l'écart absolu du nombre de *moves* par l'ensemble des parcs d'équipements partageant les mêmes qualifications nommées « usage » pour six instances sur chaque période de l'horizon de planification. Les figures 4.19 et 4.20 montrent la différence entre le nombre total de *steps* accomplis traités par deux types d'usages considérés comme goulots d'étranglement (photolithographie et de gravure), respectivement.

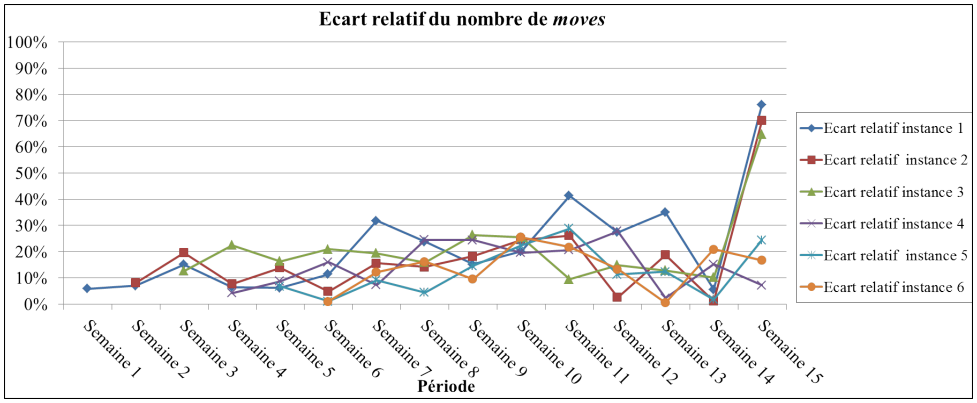

Figure 4.18 – Comparaison du nombre de *moves* réel vs. estimé

Pour cet indicateur aussi, nous observons une convergence entre la planification et le processus réel pour les 6 premières périodes avec une moyenne des écarts relatifs moyens au cours de ces périodes égales à 6.5% pour l'usage de photolithographie et 12.3% pour l'usage de gravure. Par conséquent, l'heuristique contenant l'algorithme d'équilibrage fournit de bonnes estimations sur la charge des équipements en respectant les contraintes de capacité.

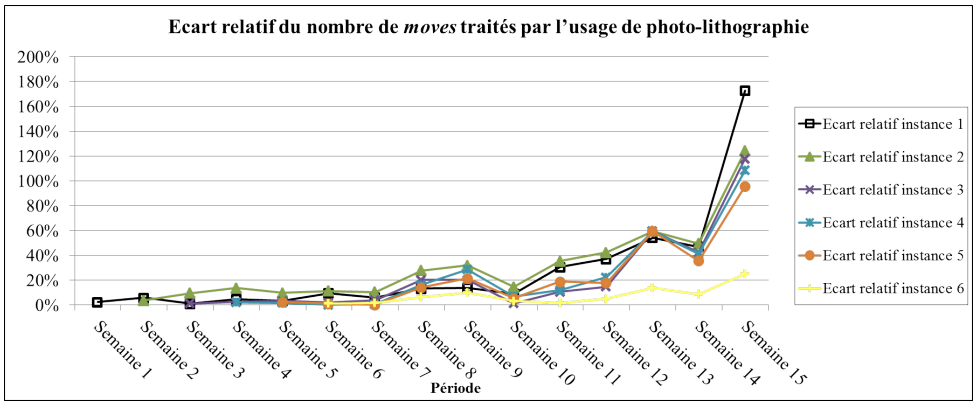

Figure 4.19 – Comparaison entre le nombre total de *moves* traités par l'usage de photolithographie réel vs. estimé

Analyse basée sur la mesure de la performance : TWT

Pour comparer le retard total pondéré réel et la valeur obtenue de cet indicateur en utilisant l'algorithme itératif pour les six essais, les écarts absolus et relatifs sont calculés et présentés dans le tableau 4.9. A partir du tableau 4.9, nous notons que la valeur estimée de TWT est proche du retard réel tout en respectant les dates d'échéance des lots. En effet, la moyenne des écarts relatifs pour les six instances est égale à 4%.

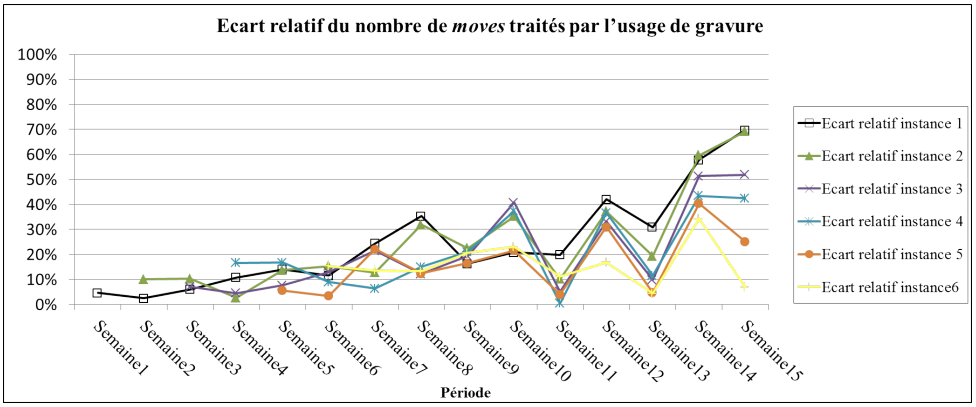

Figure 4.20 – Comparaison entre le nombre total de *moves* traités par l'usage de gravure réel vs. estimé

Tableau 4.9 – Comparaison du *TWT* réel versus estimé

Instance	Écart absolu (jours)	Écart relatif (%)
Instance 1	228.255	6
Instance 2	98.305	2.62
Instance 3	108.86	3.39
Instance 4	47.77	1.83
Instance 5	50.13	2.08
Instance 6	146.23	7.92

4.6 Conclusion

Pour résoudre le problème de planification de la production des *wafer fabs* en considérant les priorités des lots, les contraintes de capacité et les contraintes de qualification, nous avons proposé dans ce chapitre, deux heuristiques : Une heuristique de décomposition à base de *MIP* et une heuristique de décomposition à base d'algorithmes.

Pour la première heuristique, bien que le système de planification développé permette d'obtenir un plannnig de production réalisable en utilisant les données réelles, il ne satisfait pas l'objectif en terme de temps de calcul, défini par notre partenaire industriel (5 minutes au maximum). Ainsi, une deuxième heuristique est développée dont l'objectif de réduire le temps de calcul. Il s'agit d'un système de planification composé de trois modules : un module de projection du *WIP* à capacité infinie, un module de calcul de la charge accumulée sur les parcs d'équipements et un module d'équilibrage de la charge en cas de saturation des équipements. Dans le premier module, les priorités des lots en termes d'urgence de livraison sont prises en compte. Dans le deuxième module, la charge accumulée sur les parcs d'équipements est calculée en considérant les contraintes de qualifications. Dans le troisième module, le plan de production par période est ajusté de façon à respecter les contraintes de capacité.

Tout d'abord, cette heuristique est testée sur 270 instances générées aléatoirement.

Les tests de calcul ont montré un écart faible entre la solution approchée et la solution optimale pour la majorité des instances.

Ensuite, cette heuristique est testée sur des instances industrielles. Les tests de calcul, faits sur des cas réels de production, ont montré que des solutions acceptables sont obtenues rapidement. En effet, le TWT pourrait être minimisé et le taux moyen d'utilisation de l'équipement pourrait être équilibré de manière significative en utilisant le système mis au point. Par ailleurs, le temps de calcul pour des instances réelles atteint environ 30 secondes, ce qui est efficace pour les problèmes de planification avec un horizon à moyen/court terme.

Les résultats des tests de l'heuristique sur des instances industrielles sont comparées avec des données du processus réel en se basant sur trois critères : le nombre total de *moves*, le nombre de *moves* par usage et TWT. Cette comparaison a montré une nette convergence entre ce qui est prévu à l'aide de l'approche développée et ce qui est réalisé dans le processus réel sur un horizon de planification à court terme. Ces résultats montrent que la mise en œuvre du système de planification à capacité finie dans de véritables usines de fabrication semble très intéressante pour réduire au minimum les retards des lots et pour établir un planning de production réalisable sur un horizon de planification à court terme. Ainsi, cet outil est implémenté dans l'usine de fabrication des semi-conducteurs de STMicroelectronics située à Crolles. Les détails de la mise en œuvre sont présentés dans le chapitre suivant.

5

Mise en œuvre industrielle

Résumé : Nos travaux ont été réalisés dans le cadre du projet Européen *INTEGRATE*, en partenariat avec l'entreprise *STMicroelectronics*. Dans ce chapitre, nous présentons tout d'abord l'unité de fabrication considérée dans cette thèse pour la mise en œuvre de l'approche de planification de la production. Nous nous intéressons plus particulièrement au processus de fabrication et aux outils et techniques de planification de la production utilisés en signalant leurs limites et le besoin d'un nouvel outil d'aide à la décision permettant de réduire les retards de lots et d'obtenir un planning de production réalisable (cf. section5.2). L'approche développée dans cette étude est implémentée dans une plateforme de planification développée par notre partenaire industriel. La description de la plateforme et les conséquences de l'implémentation de l'algorithme proposé dans le *wafer fab* sont présentées dans la section 5.3.

Sommaire

5.1 Introduction 119
5.2 Présentation du cas industriel 119
5.2.1 Particularités du cas industriel 119
5.2.2 Description de la planification de la production chez ST Crolles 120
5.3 Description de la plateforme développée 123
5.3.1 Vue d'ensemble 123
5.3.2 Conséquences de la mise en œuvre du système de planification à capacité finie 125
5.4 Conclusion 127

5.1 Introduction

Les travaux présentés dans ce manuscrit ont reposé sur un projet industriel dont l'objectif est de développer des techniques avancées de gestion de flux qui interagissent avec des niveaux inférieurs et supérieurs de décisions. Il s'agit d'un des objectifs du projet européen *INTEGRATE* (Voir annexe **??**), visant entre autres à améliorer la connaissance et l'efficacité des procédés de fabrication dans le domaine de la fabrication des semi-conducteurs.

Nos travaux ont été réalisés en partenariat avec *STMicroelectronics*, une société internationale d'origine française et italienne qui développe, fabrique et commercialise des semi-conducteurs. L'unité de production, considérée dans cette étude, est la seule unité de fabrication 300mm de l'entreprise et est localisée à Crolles (France). Notre travail s'inscrit dans une volonté de l'entreprise de développer un outil d'aide à la décision efficace permettant de bien gérer la production dans la *fab* et d'établir des plannings de production réalisables dans un temps de calcul très réduit.

Ainsi, un diagnostic et une description détaillée de l'état de l'existant relatif à la planification du processus de fabrication au sein de ce *wafer fab* s'avèrent indispensables pour bien déterminer les limitations du système de planification utilisé et essayer de les remédier.

L'un des objectifs de ST dans le cadre du projet *INTEGRATE* est d'automatiser les décisions de planification du site de Crolles 300. L'objectif est de proposer une plateforme de planification des lots se trouvant dans le *WIP* ou des nouvelles commandes tout en considérant les priorités des lots et les contraintes de capacité et des qualifications des parcs d'équipements.

Ce chapitre a pour objet la description détaillée du problème industriel à l'origine de nos travaux de recherche. Nous présentons le système de planification utilisé par notre partenaire industriel en insistant sur ses limites et le besoin d'un nouvel outil d'aide à la décision prenant en compte les différentes contraintes liées au contexte industriel. La description de l'outil proposé et les conséquences de son intégration sur le processus de production à ST Crolles sont présentées à la fin de ce chapitre.

5.2 Présentation du cas industriel

5.2.1 Particularités du cas industriel

Cette section présente les caractéristiques qui distinguent l'unité de fabrication de STMicroelectronics de Crolles, particulièrement complexe, des autres *wafer fabs* étudiées

dans la littérature. Plus précisément, la taille de la fab *i.e.* la surface de la salle blanche est de 10000 m^2 et la complexité intrinsèque du processus de fabrication de wafers font de cet environnement de production un véritable challenge pour la recherche en génie industriel.

La *wafer fab* de ST est caractérisée par une production à forte variabilité et faible volume. Elle fabrique une gamme très diversifiée de produits, de simples transistors aux micro-contrôleurs et circuits intégrés les plus complexes regroupant des millions de composants sur la même puce. Nous trouvons plus de 1200 produits dans cette usine. Seules des opérations, de la phase *front end* du processus de fabrication des plaquettes sont effectuées. Le processus de fabrication des plaques est très complexe, il se compose de plusieurs *steps* dont le nombre peut varier entre 400 et 800 pour les technologies de production actuelles et jusqu'à 1200 *steps* pour les dernières générations, ce qui montre une forte réentrance des flux de production. Ceci représente plus du double de *steps* de fabrication par rapport aux *fabs* existantes de l'ancienne génération de *wafers* (200mm de diamètre). En particulier, dans l'usine considérée, un lot visite environ 39 fois l'atelier de photolithographie et un peu plus de 152 fois l'atelier de métrologie. On trouve environ 530 équipements regroupés en près de 300 parcs d'équipements, chacun est composé d'un ou plusieurs équipements parallèles et identiques. Cette usine est caractérisée aussi par une forte fluctuation des équipements goulots d'étranglement qui est liée à diverses sources telles que la qualification insuffisante des équipements, les pannes fréquentes, la croissance exponentielle des priorités des lots en cas de crise des capacités. Actuellement, toutes ces sources de variabilité et leur impact respectif, ne sont pas considérées suffisamment dans la gestion de production de cette *fab* et dans les outils d'aide à la décision utilisés.

5.2.2 Description de la planification de la production chez ST Crolles

Actuellement, pour la planification de la production et le calcul de la capacité de production, ST Crolles 300 utilise un outil interne fonctionnant essentiellement en flux poussé. Il s'agit d'un tableur Excel dont le calcul de la capacité de production est fait par produit et par opération. Cet outil effectue la planification de la production à capacité infinie sans tenir compte de la capacité et la disponibilité des ressources, la priorité des lots les uns par rapport aux autres et les dates d'échéance de livraison.

En utilisant la semaine comme période de référence, la capacité pour chaque parc d'équipements est fixée initialement à 168 heures ($7 jours \times 24 heures$) traduite à 100%. Cette capacité est dégradée des facteurs de perte de capacité tels que les pannes, la maintenance préventive, les reprises et les réglages. Chaque facteur de perte, généralement exprimé en pourcentage, est soustrait de 100%. La valeur obtenue est multipliée par un facteur de variabilité allant de 75% à 90% qui représente le temps d'inactivité prévu

sur les équipements. Une fois que les facteurs de perte et de variabilité sont pris en compte, la charge cumulée sur chaque parc d'équipements ainsi que leurs saturations *i.e.* le rapport entre la charge et la capacité sont déterminées en respectant les contraintes de qualifications. La figure 5.1 illustre les résultats du calcul de la saturation moyenne des parcs d'équipements des différents ateliers pendant une semaine. Pour cette instance, les parcs d'équipements de gravure sèche et humide et de photo-lithographie sont les goulets d'étranglement étant donné que leur saturation dépasse 100%.

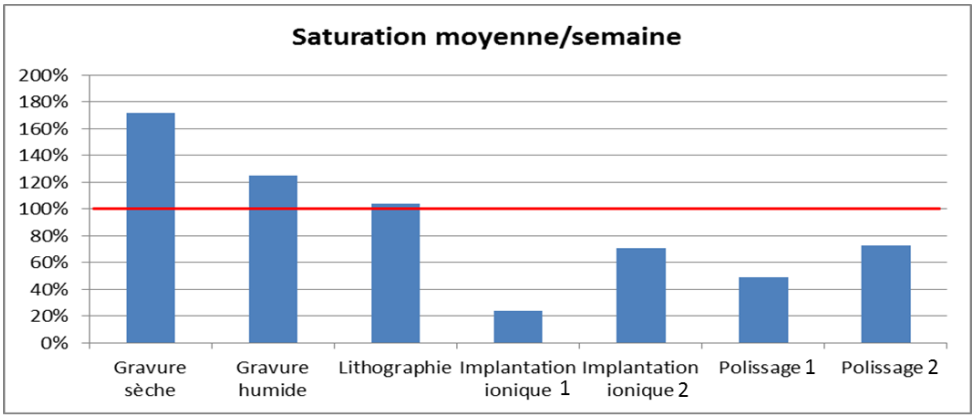

Figure 5.1 – Résultats de saturation des parcs d'équipements.

Le principe de calcul de la charge accumulée sur les parcs d'équipements dans cet outil est similaire à celui employé dans les tableurs classiques [143] mais la différence entre cet outil et les tableurs classiques réside dans la considération d'un temps de cycle variable par l'intégration d'un coefficient de variabilité de temps de cycle dans les données d'entrée. Ce coefficient est déterminé à partir des données historiques. Le modèle de temps de cycle employé est expliqué avec plus de détails dans le chapitre 4.

En plus du calcul de la charge accumulée sur les équipements, cet outil détermine aussi les dates de fin de fabrication des lots et la quantité à produire par produit et par semaine.

Le département génie industriel de ST Crolles nous a fourni des résultats réels de l'outil de planification ainsi que des données représentant ce qu'il se passe réellement dans l'usine concernant les dates de livraison et les quantités livrées. A partir de ces données, nous avons pu comparer les résultats de planification du tableur utilisé par ST avec ce qui se passe réellement dans l'usine afin de bien définir les limites de l'outil utilisé. Les figures 5.2 et 5.3 nous montrent respectivement les écarts entre le volume de production prévu à livrer et la quantité de plaquettes livrée réellement et les écarts entre les dates de livraison prévues par le système de planification et les dates de livraisons réelles pour un horizon de planification d'un mois pour une vingtaine de produits différents. Pour des raisons de confidentialité, les références des produits sont masquées.

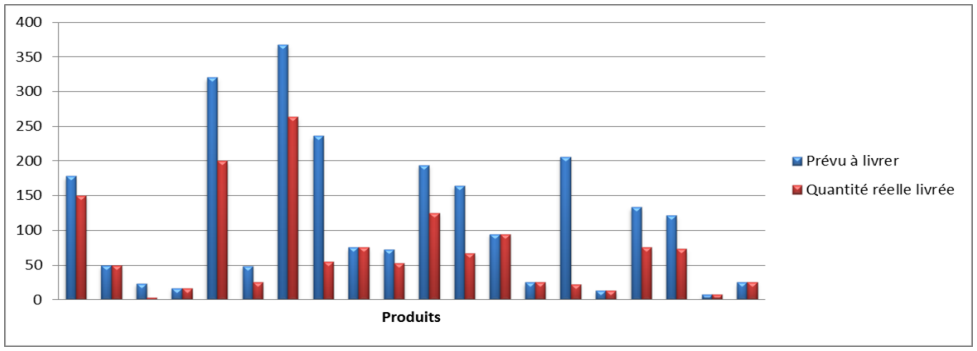

Figure 5.2 – Écarts entre la quantité prévue à livrer et la quantité livrée réellement pour une vingtaine de produits.

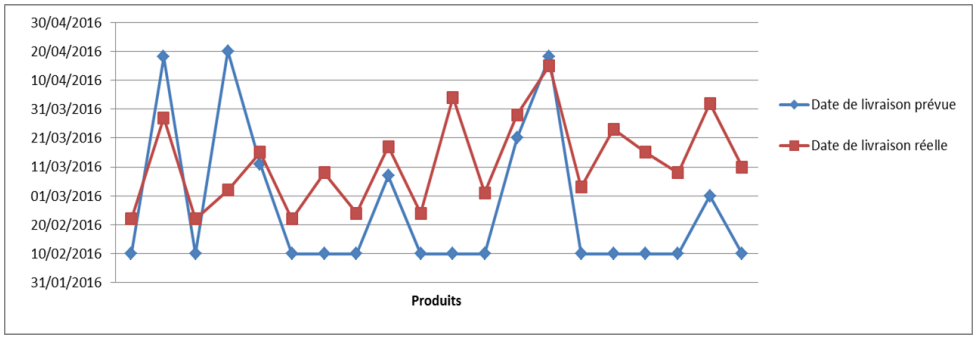

Figure 5.3 – Écarts entre la date de livraison réelle et la date de livraison prévue pour une vingtaine de produits.

La figure 5.2 montre que pour cette instance, aucun produit n'est livré à temps. 15% des lots sont livrés en avance et 85% sont livrés en retard avec un retard moyen de 22 jours.

D'après la figure 5.3, nous remarquons que la quantité à livrer n'est respectée que pour 40% des produits et elle est inférieure à ce qui est prévu pour 60% des produits avec un écart moyen égal à 80 *wafers*.

Ainsi, les quantités et les due dates planifiées sont souvent non réalisées. En effet, à ST Crolles, une fois que le planning de production est réalisé, il est soumis à plusieurs types de perturbations imprévisibles telles que la défaillance des équipements clés, l'annulation et l'accélération des commandes clients et des problèmes de processus imprévus. La gestion des modifications nécessaires du planning pour agir face aux différentes perturbations est effectuée au niveau local (en dehors du système), et souvent manuellement. Cela provoque des performances sous-optimales dans l'usine et entraîne également une propagation des perturbations ayant des conséquences imprévues ultérieurement. D'où, les problèmes détectés lors de la planification de la production de ST Crolles 300 :

1. La difficulté de la gestion manuelle des modifications de planification en termes de temps et d'énergie des ressources humaines à cause des contraintes suivantes :

- Le nombre de contraintes à considérer (*WIP*, capacité et disponibilité des machines, priorité des lots etc.),
- La forte variabilité des processus et des produits,
- La réentrance des flux,
- Le nombre d'opérations assez important,
- La mauvaise estimation des temps de cycle.

2. Les retards de livraison.
3. La difficulté de l'explication et de l'analyse des résultats.

Ainsi, l'outil utilisé présente deux limitations principales :

1. Il fonctionne à capacité infinie. Donc, la faisabilité d'un plan de production n'est pas garantie. Les valeurs de temps de cycle peuvent ne pas être réalistes.
2. Il fonctionne en mode poussé *i.e.* la livraison à temps des commandes clients n'est pas considérée. L'outil fournit une idée de « ce qui devrait se produire » si tout se passe comme dans le modèle. En plus, les modèles sont très lourds à maintenir et difficiles à maîtriser.

A partir de ces limitations, ST Crolles, dans le cadre du projet *INTEGRATE* a exprimé le besoin d'un nouvel outil d'aide à la décision permettant de répondre aux questions suivantes dans un temps de calcul très réduit ($\leq 5min$) :

— Comment gérer les priorités des lots afin de respecter les engagements de livraison ?

— Quels sont les parcs d'équipements qui seront saturés dans les x prochains jours ?

— Quelle est la date d'expédition réalisable pour chaque lot ?

Le long de cette étude, nous avons essayé de répondre à ces questions et nous avons réussi à développer l'outil demandé répondant aux différentes exigences requises. Il s'agit de l'algorithme itératif présenté dans le chapitre 4. Les résultats pertinents des tests de cet algorithme sur des instances industrielles a attiré l'attention des ingénieurs de génie industriel à ST Crolles d'où leur proposition d'implémenter cet algorithme dans un logiciel de planification.

5.3 Description de la plateforme développée

5.3.1 Vue d'ensemble

Le logiciel développé permet de faire des projections d'activité et de *WIP*, d'analyser et équilibrer la charge induite sur les parcs d'équipements. Les macro-composants du logiciel sont (cf. figure 5.4) les trois modules de l'algorithme itératif décrit dans le chapitre 4 :

— Un moteur de projection de capacité infinie (*"Projection"*),

— Un moteur de calcul de la charge accumulée sur les parcs d'équipements (*"Balancing"*)

— Un moteur d'équilibrage de la charge et la capacité (*"Finite capacity"*)

D'autres modules sont intégrés dans cette plateforme tels que le module de calcul d'un modèle de temps de cycle dynamique (*"Dynamic ST Model"*) qui tient en compte de la variabilité du processus de fabrication. Ce module est en cours d'élaboration lors d'une thèse CIFRE entre ST Crolles et le laboratoire G-SCOP. En plus, on trouve le module *"Lot priorities"* qui détermine les priorités des lots en se basant sur les due dates.

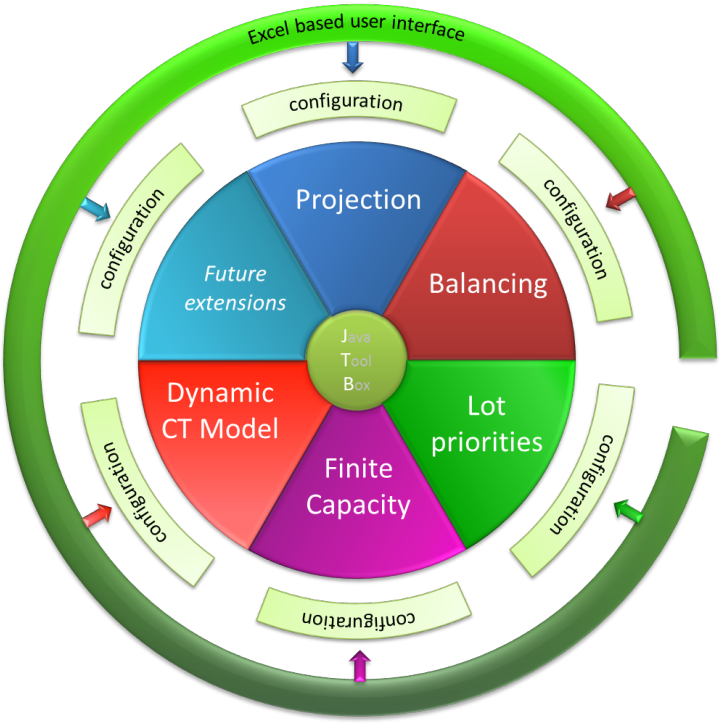

Figure 5.4 – Vue d'ensemble du logiciel développé.

Comme l'algorithme proposé, cette plateforme est aussi développée en `JAVA` et l'interface graphique est générée sur Excel (cf. figure 5.5). Pour tout utilisateur, l'exécution du moteur de projection se fait grâce au bouton "*Execute Projection Engine*". Les données résultats de la projection sont alors automatiquement mises à jour dans Excel. Pour l'exécution du moteur d'équilibrage, il faut au préalable définir le nombre de périodes d'analyse dans le volet "*Capacity Analysis*". L'exécution du moteur d'équilibrage se fait grâce au bouton "*Execute Capacity Analysis Engine*". Les données résultats de l'équilibrage sont alors automatiquement mises à jour dans Excel.

Avant l'exécution de chaque module, un appel à la base de données convenable est effectué pour charger les données nécessaires comme entrées à l'algorithme à exécuter. A

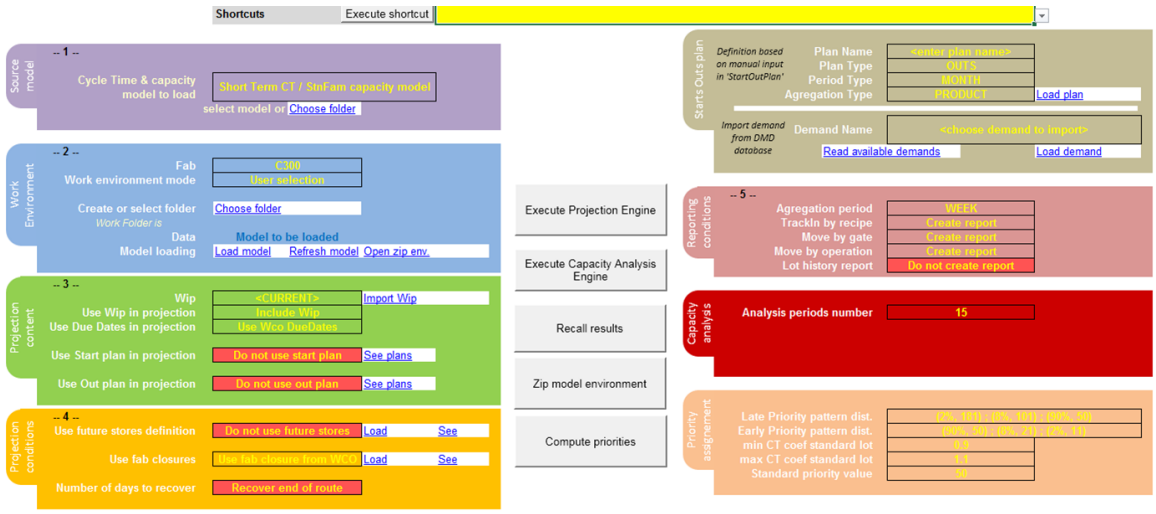

Figure 5.5 – Interface graphique du logiciel développé.

la fin de l'exécution de chaque module, des rapports résultats sont générés. La figure 5.6 illustre les entrées et les sorties des trois modules de notre algorithme implémentés dans la plateforme logicielle de planification à ST Crolles.

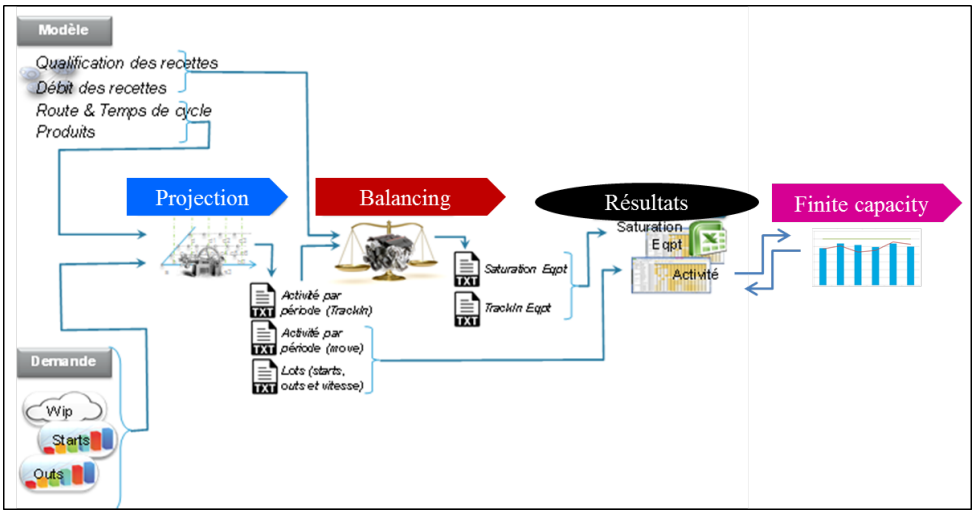

Figure 5.6 – Entrées et sorties de chaque module du logiciel développé.

5.3.2 Conséquences de la mise en œuvre du système de planification à capacité finie

La mise en œuvre industrielle de l'algorithme et le développement de la plateforme logicielle de planification a conduit à une amélioration significative des performances en termes de livraison à temps, saturation des équipements et temps d'exécution. L'intégration de cet outil a permis d'améliorer considérablement la «livraison conforme», d'où la satisfaction des clients, qui a été particulièrement difficile auparavant à cause de la forte

variabilité du processus de fabrication. L'indicateur de livraison « juste-à-temps » est amené à un niveau jusqu'ici inégalé (entre 98% et 100%, alors qu'il stagnait aux alentours des 80%. La figure 5.7 illustre l'évolution de cet indicateur sur 12 périodes. En plus, la

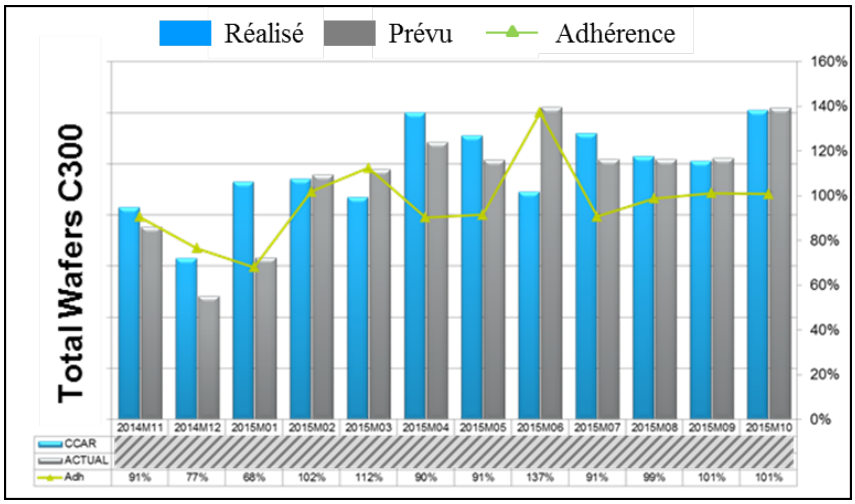

Figure 5.7 – Évolution de l'indicateur juste à temps à ST Crolles 300

saturation des équipements est maintenue en dessous des seuils de saturation prédéfinis tout en minimisant les retards de livraison des lots. La figure 5.8 présente les résultats de saturation des différents parcs d'équipements de gravure et de photolithographie, considérés comme goulets d'étranglement pour une planification à capacité infinie et finie sur différentes périodes d'un horizon de planification. Dans ces exemples, la charge est calculée en nombres de *wafers* produites.

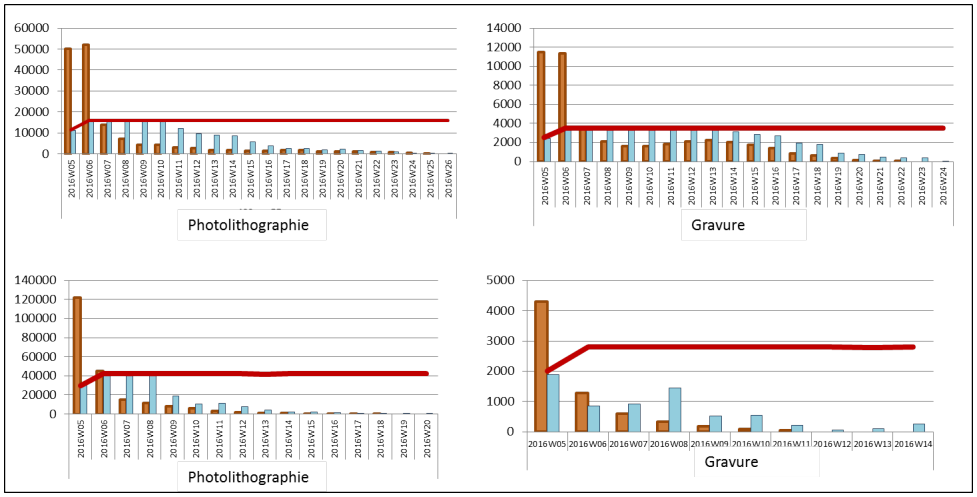

Figure 5.8 – Saturation des parcs d'équipements

Le temps d'exécution des différents modules est très réduit. Il est inférieur à 1 minute. Par exemple, pour calculer un plan de production réalisable pour un *WIP* de 7500 lots sur

une période de 6 mois, exécutés sur 360 parcs d'équipements (48 secondes pour déterminer un planning à capacité finie par rapport à 35 secondes pour un planning capacité infinie).

5.4 Conclusion

Les travaux menés ont abouti à une réalisation industrielle, ce qui est le but originel de nombreuses études en recherche opérationnelle et aide à la décision. Un projet long, difficile à mener, a été entamé et n'est à ce jour pas terminé. La décision d'industrialisation a été prise et on peut considérer que l'achèvement du projet est en bonne voie. Plusieurs enseignements se dégagent de cette expérience et permettent d'éclairer l'idée qu'on a de la mise en œuvre d'un logiciel d'optimisation et d'aide à la décision en milieu industriel.

— La validation de l'outil ne peut se faire sans la mise en place d'indicateurs de performance pertinents pour évaluer de façon concrète et objective la qualité d'une solution.

— La difficulté de l'intégration d'un nouvel outil de planification dans un système d'information existant lors de la communication avec les différentes bases de données.

— L'élaboration d'une interface graphique simple, compréhensible par tous les utilisateurs et facilitant la communication entre les différents utilisateurs .

Quelques mois séparent alors la décision d'implantation et la réalisation. Les tests effectués avec l'outil pour déterminer son potentiel gain en situation réelle de fonctionnement montrent l'efficacité du logiciel développé pour la détermination d'un plan de production réalisable à court/moyen terme sur des tranches horaires, en considérant toutes les informations actualisées de façon régulière. Plusieurs extensions du projet sont envisagées :

— L'intégration du module de calcul d'un modèle de temps de cycle dynamique permettant de modéliser les variations de temps de cycle en fonction de la saturation des équipements. Ce module doit être basé sur la théorie des files d'attente en tenant compte de la saturation des équipements et la variabilité du processus de fabrication.

— L'intégration d'un modèle de simulation à événements discrets pour le calcul de la capacité finie par période (limitée aux équipements goulots d'étranglement) ou sur tout l'horizon de planification.

— La migration de l'outil vers d'autres sites de STMicroelectronics. L'entreprise pourrait en utiliser différentes versions, et ainsi procéder à son adaptation. Il est déjà en cours de migration vers le site de Crolles 200.

Conclusion générale et perspectives

Les travaux que nous avons présentés dans ce document traitent un problème de planification de la production à moyen terme dans le contexte de l'industrie des semi-conducteurs. Ils prennent place au sein du projet Européen *INTEGRATE* et ont été développés en collaboration étroite avec notre partenaire industriel STMicroelectronics.

L'objectif de cette thèse est de développer des approches permettant de déterminer un plan de production faisable pour les lignes de fabrication des semi-conducteurs. La fabrication des semi-conducteurs est caractérisée par une forte variabilité due à la diversité des produits, des équipements et des modes de fonctionnement. Elle est caractérisée aussi par une complexité du processus de fabrication et des flux ré-entrants. En plus des contraintes de capacité et des caractéristiques du processus de fabrication, l'approche développée doit tenir compte de la variabilité des temps de cycle, des priorités des lots et des contraintes de qualifications des parcs d'équipements *i.e.* l'éligibilité d'un équipement à traiter un produit.

Le problème de planification considéré est la projection des encours de production. Il s'agit d'estimer des dates de début et de fin de chacun des *steps* restants et d'anticiper la charge accumulée sur les équipements.

Dans la revue de littérature, que nous avons présentée dans le chapitre 2, plusieurs problèmes en planification de la production sont identifiés et de différentes modélisations et méthodes de résolution issues de la littérature et les solutions logicielles pour formuler et résoudre ces problèmes sont décrites. Plusieurs classifications des approches existantes sont proposées selon le niveau de décision, le type d'approche, le type de la demande, le type de la méthode de résolution, la prise en compte ou non des contraintes de capacité, etc. A partir de la revue bibliographique, nous avons noté :

— La rareté des travaux traitant le problème de projection du *WIP* à capacité finie,

— L'absence des travaux concernant la planification à capacité finie dans un contexte industriel à forte variabilité et faible volume de production,

— La rareté des travaux intégrant les niveaux de décisions tactique et opérationnel en tenant compte des priorités des lots et des contraintes de qualification et de capacité des équipements.

Dans le chapitre 3, nous avons présenté et expérimenté la complexité du problème dans l'obtention d'une solution optimale. Nous avons formulé le problème sous forme d'un programme linéaire mixte (*MIP*) dont la fonction objectif est la minimisation de la somme des retards pondérés tout en tenant compte des contraintes temporelles et des contraintes de capacité. L'analyse de complexité numérique a permis de déterminer les limites du programme linéaire en termes de nombre de lots et nombre de steps à planifier. Pour les instances industrielles, nous n'obtenons pas de solution optimale au bout d'un temps de résolution raisonnable. Par ailleurs, des méthodes approchées ont été développées afin de corriger la limite de la méthode exacte qui est le temps de calcul important.

Ces méthodes sont une heuristique d'agrégation, une heuristique de décomposition et la relaxation lagrangienne. Pour des instances de taille réduite, nous avons montré expérimentalement la performance de la relaxation lagrangienne en termes de qualité de la solution tout en donnant des solutions proches de l'optimal. Cependant, cette méthode est moins performante que l'heuristique de décomposition en termes de temps de calcul. Malgré l'importance de ces techniques dans la réduction du temps de calcul et de l'amélioration de la limite de résolution, il s'est avéré qu'elles sont inefficaces au point de vue qualité de la solution et temps de calcul en testant des instances industrielles.

Dans le chapitre 4 de ce mémoire, deux méthodes approchées à savoir une heuristique à base de *MIP* et une heuristique à base d'algorithmes sont proposées. Pour la première heuristique, bien que le système de planification développé permette d'obtenir un plannnig de production réalisable en utilisant les données réelles, il ne satisfait pas l'objectif en terme de temps de calcul, défini par notre partenaire industriel (5 minutes au maximum). Ainsi, une deuxième heuristique est développée dont l'objectif est de réduire le temps de calcul. Il s'agit d'un système de planification composé de trois modules : un module de projection du *WIP* à capacité infinie, un module de calcul de la charge accumulée sur les parcs d'équipements et un module d'équilibrage de la charge en cas de saturation des équipements. Dans le premier module, les priorités des lots en termes d'urgence de livraison sont prises en compte. Dans le deuxième module, la charge accumulée sur les parcs d'équipements est calculée en considérant les contraintes de qualifications. Dans le troisième module, le plan de production par période est ajusté de façon à respecter les contraintes de capacité.

La mise en œuvre industrielle de l'outil proposé et ses conséquences sont présentées dans le chapitre 5. L'industrialisation de l'outil a conduit à une amélioration significative des performances en termes de livraison à temps, saturation des équipements et temps d'exécution dans le *wafer fab* de ST Crolles.

Ainsi, la contribution de nos travaux est double. Sur le plan scientifique, nous avons utilisé diverses méthodes de la recherche opérationnelle en les appliquant au problème

de planification de la production considéré. Nous avons aussi réussi à développer une méthode approchée satisfaisant plusieurs contraintes dans un temps de calcul réduit. Sur le plan industriel, un outil d'aide à la décision répondant au besoin des industriels en termes de planning réalisé et temps de calcul a été élaboré.

Ce travail a donné lieu à des publications dans des conférences internationales ([118] et [119]) et nationales ([120] et [121]).

Perspectives

Le travail réalisé dans cette thèse permet d'envisager les développements suivants :

— Prendre en compte d'autres contraintes

- Le modèle de temps de cycle objectif utilisé est extrait de l'historique des données. Chen et *al.* [28] ont montré que les données historiques ne sont pas suffisantes pour estimer les coefficients de variation. Par conséquent, il sera intéressant de développer un modèle de temps de cycle dynamique basé sur la théorie des files d'attente en tenant compte de la saturation des équipements et la variabilité du processus de fabrication.
- Dans le chapitre 3, nous avons présenté des contraintes issues du processus de fabrication des semi-conducteurs qui sont occultées dans nos approches de résolution proposées vu qu'elles complexifient davantage le problème. Il s'agit des contraintes d'enchaînement, contraintes de *batching* et contraintes de *setup*. Il sera intéressant d'intégrer ces contraintes dans le système proposé afin d'améliorer la précision du système et de s'approcher davantage à la situation réelle.

— Enlever des hypothèses

- Dans notre étude, nous avons supposé que les capacités des équipements sont déterministes. En pratique, cette hypothèse n'est pas toujours vraie. Il serait donc intéressant d'intégrer un modèle de simulation à événements discrets pour le calcul des capacités des équipements.

— Améliorer la résolution du problème

- Améliorer la résolution du programme linéaire en utilisant d'autres solveurs tel que LocalSolver, un solveur métaheuristique basé sur la recherche locale.
- Pour la relaxation lagrangienne, nous avons appliqué la méthode des sous gradient pour mettre à jour les multiplicateurs lagrangiens mais nous pouvons envisager d'améliorer le temps de calcul de cette approche en appliquant la méthode des faisceaux pour résoudre le problème dual.

- Proposer d'autres heuristiques de décomposition ou d'agrégation par exemple la combinaison des méthodes proposées.

— Diffuser les résultats

- Comme extension de la plateforme logicielle développée, la suite logique serait d'intégrer d'autres modules tels que le module de temps de cycle dynamique et le module de détermination des capacités et aussi d'étendre ces travaux à d'autres sites de STMicroelectronics.
- Une autre perspective peut être considérée, elle consiste à étendre l'application de notre approche dans d'autres secteurs industriels et d'autres domaines qui peuvent être concernés par des problématiques similaires tels que l'industrie automobile, le domaine nucléaire ou énergétique, etc.

Références bibliographiques

[1] Worldwide semiconductor trade statistics, market statistic reports. Rapport technique, 2016. 7, 8

[2] R. N. Anthony. *Planning and control systems : a framework for analysis*. Division of Research, Graduate School of Business Administration, Harvard University Boston, MA, 1965. 18

[3] A. Arisha, P. Young et M. El Baradie. A simulation model to characterize the photolithography process of a semiconductor wafer fabrication. *Journal of materials processing technology*, 155:2071–2079, 2004. 27

[4] J. Asmundsson, R. L. Rardin, C. H. Turkseven et R. Uzsoy. Production planning with resources subject to congestion. *Naval Research Logistics (NRL)*, 56(2):142–157, 2009. 44

[5] J. Asmundsson, R. L. Rardin et R. Uzsoy. Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. *IEEE Transactions on Semiconductor Manufacturing*, 19(1):95–111, 2006. 44

[6] C. Azzaro-Pantel, P. Floquet, L. Pibouleau et S. Domenech. A fuzzy approach for performance modeling in a batch plant : application to semiconductor manufacturing. *IEEE Transactions on Fuzzy Systems*, 5(3):338–357, 1997. 30

[7] N. A. Bakke et R. Hellberg. The challenges of capacity planning. *International journal of production economics*, 30:243–264, 1993. 24

[8] H. Balasubramanian, L. Mönch, J. Fowler et M. Pfund. Genetic algorithm based scheduling of parallel batch machines with incompatible job families to minimize total weighted tardiness. *International Journal of Production Research*, 42(8):1621–1638, 2004. 44

[9] J.-Y. Bang et Y.-D. Kim. 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, 2010. 42, 44

[10] F. Barahona, S. Bermon, O. Günlük et S. Hood. Robust capacity planning in semiconductor manufacturing. *Naval Research Logistics (NRL)*, 52(5):459–468, 2005. 26, 28, 42, 44, 45, 90

[11] J. F. Bard, K. Srinivasan et D. Tirupati. An optimization approach to capacity expansion in semiconductor manufacturing facilities. *International Journal of Production Research*, 37(15):3359–3382, 1999. 44, 45

[12] J.F. Bard, Y. Deng, R. Chacon et J. Stuber. Midterm planning to minimize deviations from daily target outputs in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 23(3):456–467, 2010. 44, 49, 71

[13] M. Baudin, V. Mehrotra, B. Tullis, D. Yeaman et R. A. Hughes. From spreadsheets to simulations : a comparison of analysis methods for IC manufacturing performance. *In Semiconductor Manufacturing Science Symposium, 1992. ISMSS 1992., IEEE/SEMI International*, pages 94–99, 1992. 26

[14] S. Bermon, G. Feigin et S. Hood. Capacity analysis of complex manufacturing facilities. *In Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 2, pages 1935–1940, 1995. 47, 53

[15] S. Bermon et S. Hood. Capacity optimization planning system (CAPS). *Interfaces*, 29(5):31–50, 1999. 26, 42, 44, 45

[16] B. Bettayeb. *Conception et évaluation des plans de surveillance basés sur le risque.* Thèse de doctorat, Université de Grenoble, 2012. 9

[17] P. J. Billington, J.O. McClain et L.J. Thomas. Mathematical programming approaches to capacity-constrained MRP systems : Review, formulation and problem reduction. *Management Science*, 29(10):1126–1141, 1983. 23

[18] J. Błażewicz, W. Domschke et E. Pesch. The job shop scheduling problem : Conventional and new solution techniques. *European journal of operational research*, 93(1):1–33, 1996. 28

[19] K. M. Bretthauer et M. J. Côté. Nonlinear programming for multiperiod capacity planning in a manufacturing system. *European Journal of Operational Research*, 96(1):167–179, 1997. 44

[20] D. Y. Burman, F. J. Gurrola-Gal, A. Nozari, S. Sathaye et J. P. Sitarik. Performance analysis techniques for IC manufacturing lines. *AT&T technical journal*, 65(4):46–57, 1986. 41

[21] Y. Cai. *Semiconductor manufacturing inspired integrated scheduling problems : production planning, advanced process control, and predictive maintenance.* ProQuest, 2008. 44

[22] M. Cakanyildirim et R. O. Roundy. Evaluation of capacity planning practices for the semiconductor industry. *IEEE Transactions on Semiconductor Manufacturing*, 15(3):331–340, 2002. 44

[23] J. G. Carlson et A. C. Yao. Mixed model assembly simulation. *International Journal of Production Economics*, 26(1-3):161–167, 1992. 25

[24] B. Çatay, Ş. Erengüç et A. J. Vakharia. Capacity allocation with machine duplication in semiconductor manufacturing. *Naval Research Logistics (NRL)*, 52(7):659–667, 2005. 44, 71

[25] B. Çatay, Ş.S Erengüç et A. J. Vakharia. Tool capacity planning in semiconductor manufacturing. *Computers & Operations Research*, 30(9):1349–1366, 2003. 26, 44, 45

[26] S.-C. Chang, L.-H. Lee, L.-S. Pang, TW-Y. Chen, Y.-C. Weng, H.-D. Chiang et DW-H. Dai. Iterative capacity allocation and production flow estimation for scheduling semiconductor fabrication. *In Electronics Manufacturing Technology Symposium, 1995.'Manufacturing Technologies-Present and Future', Seventeenth IEEE/CPMT International*, pages 508–512. IEEE, 1995. 44

[27] C.-S. Chen, S. Mestry, P. Damodaran et C. Wang. The capacity planning problem in make-to-order enterprises. *Mathematical and computer modelling*, 50(9): 1461–1473, 2009. 48

[28] H. Chen, J. M. Harrison, A. Mandelbaum, A. Van Ackere et L. M. Wein. Empirical evaluation of a queueing network model for semiconductor wafer fabrication. *Operations Research*, 36(2):202–215, 1988. 131

[29] J. C Chen et C.-W. Chen. Capacity planning of serial and batch machines with capability constraints for wafer fabrication plants. *International Journal of Production Research*, 48(11):3207–3223, 2010. 29

[30] J. C. Chen, C. W. Chen, C. J. Lin et H. Rau. Capacity planning with capability for multiple semiconductor manufacturing fabs. *Computers & Industrial Engineering*, 48(4):709–732, 2005. 29, 44, 49

[31] J. C. Chen, Y.-C. Fan et C.-W. Chen. Capacity requirements planning for twin fabs of wafer fabrication. *International Journal of Production Research*, 47(16): 4473–4496, 2009. 29, 44

[32] J. C. Chen, L.-H. Su, C.-J. Sun et M.-F. Hsu. Infinite capacity planning for IC packaging plants. *International Journal of Production Research*, 48(19):5729–5748, 2010. 29, 44, 49

[33] J. C. Chen, C.-J. Sun et T.-L. Chen. Capacity planning for integrated circuit final test plants. *International Journal of Computer Integrated Manufacturing*, 28(12): 1262–1274, 2015. 29, 31, 44, 49, 103

[34] T. Chen. A fuzzy mid-term single-fab production planning model. *Journal of Intelligent Manufacturing*, 14(3-4):273–285, 2003. 47

[35] T.-R. M. Chen et T. Hsia. Scheduling for ic sort and test facilities with precedence constraints via lagrangian relaxation. *Journal of Manufacturing Systems*, 16(2):117–128, 1997. 28

[36] Y.-Y. Chen, T.-L. Chen et C.-D. Liou. Medium-term multi-plant capacity planning problems considering auxiliary tools for the semiconductor foundry. *The International Journal of Advanced Manufacturing Technology*, 64(9-12):1213–1230, 2013. 42, 44, 48

[37] C.-F. Chien, S. Dauzère-Pérès, H. Ehm, J. W. Fowler, Z. Jiang, S. Krishnaswamy, T.-E. Lee, L. Mönch et R. Uzsoy. Modelling and analysis of semiconductor manufacturing in a shrinking world : challenges and successes. *European Journal of Industrial Engineering*, 5(3):254–271, 2011. 12

[38] Y.-C. Chou et L-H. Hong. A methodology for product mix planning in semiconductor foundry manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 13(3):278–285, 2000. 44, 47

[39] Y-C. Chou et R-C. You. Resource portfolio planning methodology for semiconductor wafer manufacturing. *The International Journal of Advanced Manufacturing Technology*, 18(1):12–19, 2001. 45, 90

[40] R. M. Christie et S. D. Wu. Semiconductor capacity planning : stochastic modelingand computational studies. *IIE Transactions*, 34(2):131–143, 2002. 44, 45

[41] T. J. Chua, M. W. Liu, F. Y. Wang, W. J. Yan et T. X. Cai. An intelligent multi-constraint finite capacity-based lot release system for semiconductor backend assembly environment. *Robotics and Computer-Integrated Manufacturing*, 23(3): 326–338, 2007. 30, 44, 50, 53

[42] J. Chung et J. Jang. A wip balancing procedure for throughput maximization in semiconductor fabrication. *IEEE Transactions on semiconductor manufacturing*, 22(3):381–390, 2009. 103

[43] D. P. Connors, G. E. Feigin et D. D. Yao. A queueing network model for semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 9(3):412–427, 1996. 26, 90

[44] S. Dauzère-Pérès et J.-B. Lasserre. On the importance of sequencing decisions in production planning and scheduling. *International Transactions in Operational Research*, 9(6):779–793, 2002. 44

[45] J. E. Dayhoff et R. W. Atherton. Simulation of vlsi manufacturing areas. *VLSI Design*, 4:84–92, 1984. 41, 90

[46] S. De et A. Lee. Towards a knowledge-based scheduling system for semiconductor testing. *International Journal of Production Research*, 36(4):1045–1073, 1998. 30

[47] B.T. Denton, J. Forrest et R. J. Milne. IBM solves a mixed-integer program to optimize its semiconductor supply chain. *Interfaces*, 36(5):386–399, 2006. 42

[48] M.M. Dessouky et R.C. Leachman. Dynamic models of production with multiple operations and general processing times. *Journal of the Operational Research Society*, 48(6):647–654, 1997. 44, 46, 47

[49] G. Dobson et R. S. Nambimadom. The batch loading and scheduling problem. *Operations research*, 49(1):52–65, 2001. 44

[50] G. Doumeingts, D. Breuil et L. Pun. *La gestion de production assistée par ordinateur : GPAO*. Hermes Publishing France, 1983. 33

[51] G. W. A. Dummer. *Electronic inventions and discoveries : electronics from its earliest beginnings to the present day*. Elsevier, 2013. 7

[52] H. E. Fargher, M. A. Kilgore, P. J. Kline et R. A. Smith. A planner and scheduler for semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 7(2):117–126, 1994. 30, 44, 49, 53, 91

[53] H. E. Fargher et R. A. Smith. Planning in a flexible semiconductor manufacturing environment. *Intelligent scheduling*, pages 545–580, 1994. 30, 44

[54] M. L. Fisher. The lagrangian relaxation method for solving integer programming problems. *Management science*, 27(1):1–18, 1981. 69

[55] K. Fordyce et G. Sullivan. A dynamically generated rapid response capacity planning model for semiconductor fabrication facilities. *In The Impact of Emerging Technologies on Computer Science and Operations Research*, pages 103–127. Springer, 1995. 44

[56] J.W. Fowler, S. Brown, H. Gold et A. Schoemig. Measurable improvements in cycle-time-constrained capacity. *In Proceedings of IEEE International Symposium On Semiconductor Manufacturing Conference*, pages 21–24, San Francisco, United States, 1997. 27, 50

[57] M.R. Garey et D.S. Johnson. *Computers and Intractability : A Guide to the Theory of NPCompleteness*. W. H. Freeman & Co., New York, NY, USA, 1979. 64

[58] C. D. Geiger, K. G. Kempf et R. Uzsoy. A tabu search approach to scheduling an automated wet etch station. *Journal of Manufacturing Systems*, 16(2):102–116, 1997. 29

[59] N. Geng et Z. Jiang. Capacity planning for semiconductor wafer fabrication with uncertain demand and capacity. *In 2007 IEEE International Conference on Automation Science and Engineering*, pages 100–105. IEEE, 2007. 44

[60] N. Geng et Z. Jiang. A review on strategic capacity planning for the semiconductor manufacturing industry. *International Journal of Production Research*, 47(13):3639–3655, 2009. 45

[61] N. Geng, Z. Jiang et F. Chen. Stochastic programming based capacity planning for semiconductor wafer fab with uncertain demand and capacity. *European Journal of Operational Research*, 198(3):899 – 908, 2009. 42, 44

[62] C. R. Glassey et M. G. C. Resende. Closed-loop job release control for VLSI circuit manufacturing. *IEEE Transactions on Semiconductor manufacturing*, 1(1): 36–46, 1988. 41

[63] E. M. Goldratt. *Theory of constraints : What is this thing called Theory of Constraints and how should it be implemented*. North River Press, 1990. 24

[64] D. Y. Golhar et C. L. Stamm. The just-in-time philosophy : a literature review. *International Journal of Production Research*, 29(4):657–676, 1991. 24

[65] JJ. Golovin. A total framework for semiconductor production planning and scheduling. *Solid State Technology*, 29(5):167–170, 1986. 48

[66] N. Govind et D. Fronckowiak. Setting performance targets in a 300mm wafer fabrication facility. *In Proceedings of Advanced Semiconductor Manufacturing Conference and Workshop*, pages 75–79, 2003. 44, 51, 52

[67] R. L. Graham, E. L. Lawler, J. K. Lenstra et A. R. Kan. Optimization and approximation in deterministic sequencing and scheduling : a survey. *Annals of discrete mathematics*, 5:287–326, 1979. 155

[68] N.S. Grewal, A.C. Bruska, T.M. Wulf et J.K. Robinson. Integrating targeted cycle-time reduction into the capital planning process. *In Proceedings of the 1998 Winter Simulation Conference– WSC 1998*, pages 1005–1010, Washington, United States, 1998. 50, 90

[69] M. P. Groover. *Fundamentals of modern manufacturing : materials processes, and systems*. John Wiley & Sons, 2007. 12

[70] J. N. D. Gupta, R. Ruiz, J. W. Fowler et S. J. Mason. Operational planning and control of semiconductor wafer fabrication. *Production Planning and Control*, 17(7):639–647, 2006. 1, 12, 16

[71] K. Habenicht et L. Mönch. A finite-capacity beam-search-algorithm for production scheduling in semiconductor manufacturing. *In Simulation Conference, 2002. Proceedings of the Winter*, volume 2, pages 1406–1413. IEEE, 2002. 30, 44, 50, 53

[72] C. Habla et L. Mönch. Solving volume and capacity planning problems in semiconductor manufaturing : a computational study. *In Proceedings of 2008 Winter Simulation Conference – WSC 2008*, pages 2260–2266, Texas, United States, 2008. 42, 44, 48

[73] C. Habla, L. Mönch et R. Drissel. A finite capacity production planning approach for semiconductor manufacturing. *In Proceedings of the 3rd Annual IEEE Conference on Automation Science and Engineering*, pages 82–87, Scottsdale, United States, 2007. 42, 44, 47, 53, 71

[74] S.T. Hackman et R. C. Leachman. A general framework for modeling production. *Management Science*, 35(4):478–495, 1989. 46

[75] HO-S. Ham. Job route selection model for workload balancing between workstations in flexible flow line. *Production Planning & Control*, 7(4):430–438, 1996. 102

[76] J. E. Harl. *Reducing Capacity Problems in Material Requirements Planning Systems*. New York University, Graduate School of Business Administration, 1981. 24, 25

[77] M. Held, P. Wolfe et H. P. Crowder. Validation of subgradient optimization. *Mathematical programming*, 6(1):62–88, 1974. 71

[78] P. Henrich, M. Land et G. Gaalman. Grouping machines for effective workload control. *International Journal of Production Economics*, 104(1):125–142, 2006. 103

[79] S. J. Hood, S. Bermon et F. Barahona. Capacity planning under demand uncertainty for semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 16(2):273–280, 2003. 26, 42, 44, 45, 46

[80] W. Hopp et M.L. Spearman. Factory physics, 2000. *International edition. McGraw Hill*. 20

[81] K. Horiguchi, N. Raghavan, R. Uzsoy et S. Venkateswaran. Finite-capacity production planning algorithms for a semiconductor wafer fabrication facility. *International Journal of Production Research*, 39(5):825–842, 2001. 29, 44, 49, 53

[82] Y. Hsiung, M.-C. Wu et H.-M. Hsu. Tool planning in multiple product-mix under cycle time constraints for wafer foundries using genetic algorithm. *Journal of the Chinese Institute of Industrial Engineers*, 23(2):174–183, 2006. 44

[83] W. T. Huh et R. O. Roundy. A continuous-time strategic capacity planning model. *Naval Research Logistics (NRL)*, 52(4):329–343, 2005. 44

[84] Y.-F. Hung et R. C. Leachman. A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations. *IEEE Transactions on Semiconductor Manufacturing*, 9(2):257–269, 1996. 27, 42, 44, 46, 48, 53

[85] T.-K. Hwang et S.-C. Chang. Design of a lagrangian relaxation-based hierarchical production scheduling environment for semiconductor wafer fabrication. *IEEE Transactions on Robotics and Automation*, 19(4):566–578, 2003. 44, 47

[86] J. P. Ignizio et H. Garrido. Fab simulation and variability. *Future Fab International*, 41:41–45, 2012. 32, 50, 51

[87] ENIAC IMPROVE. Official website. 2012. 150

[88] G. Ioannou et S. Dimitriou. Lead time estimation in mrp/erp for make-to-order manufacturing systems. *International Journal of Production Economics*, 139(2): 551–563, 2012. 89

[89] D. F. Irdem, N. B. Kacar et R. Uzsoy. An experimental study of an iterative simulation-optimization algorithm for production planning. *In 2008 Winter Simulation Conference*, pages 2176–2184. IEEE, 2008. 48

[90] P. K. Johri. Practical issues in scheduling and dispatching in semiconductor wafer fabrication. *Journal of Manufacturing Systems*, 12(6):474–485, 1993. 41

[91] N. B. Kacar, L. Mönch et R. Uzsoy. Modeling cycle times in production planning models for wafer fabrication. *IEEE Transactions on Semiconductor Manufacturing*, 29(2):153–167, 2016. 44

[92] J. J. Kanet. Toward a better understanding of lead times in mrp systems. *Journal of operations management*, 6(3):305–315, 1986. 88

[93] J. J. Kanet et M. Stösslein. Integrating production planning and control : towards a simple model for capacitated erp. *Production Planning & Control*, 21(3): 286–300, 2010. 24, 25

[94] S. Karabuk et S. D. Wu. Decentralizing semiconductor capacity planning via internal market coordination. *IIE transactions*, 34(9):743–759, 2002. 44

[95] S. Karabuk et S. D. Wu. Coordinating strategic capacity planning in the semiconductor industry. *Operations Research*, 51(6):839–849, 2003. 44, 45

[96] C. A. Kaskavelis et M. C. Caramanis. Efficient lagrangian relaxation algorithms for industry size job-shop scheduling problems. *IIE transactions*, 30(11):1085–1097, 1998. 28

[97] J. S. Kim et R. C. Leachman. Decomposition method application to a large scale linear programming wip projection model. *European Journal of Operational Research*, 74(1):152–160, 1994. 42, 44, 51, 52

[98] R. Kotcher et F. Chance. Capacity planning in the face of product-mix uncertainty. *In IEEE international symposium on semiconductor manufacturing conference proceedings*, volume 1, pages l–l3, 1999. 90

[99] P.R. Kumar. Scheduling semiconductor manufacturing plants. *IEEE Control Systems*, 14(6):33–40, 1994. 12, 14, 41

[100] R. Kumar. *Fabless Semiconductor Implementation*. McGraw-Hill Professional USA, 2008. 31

[101] E. L. Lawler. A "pseudopolynomial" algorithm for sequencing jobs to minimize total tardiness. *Annals of discrete Mathematics*, 1:331–342, 1977. 155

[102] R. C. Leachman. Modeling techniques for automated production planning in the semiconductor industry. *In* T.A. Ciriani et R.C. Leachman, éditeurs. *Optimisation in Industry : Mathematical Programming and Modeling*, pages 1–30, Wiley, New York, 1993. 26, 47, 53

[103] R. C. Leachman, R. F. Benson, C. Liu et D. J. Raar. IMPRESS : an automated production planning and delivery quotation system at harris corporation-semiconductor sector. *Interfaces*, 26(1):6–37, 1996. 42, 44, 47

[104] R. C. Leachman et T.F. Carmon. On capacity modeling for production planning with alternative machine types. *IIE transactions*, 24(4):62–72, 1992. 26, 42, 44, 46

[105] Y. Lee, S. Kim, S. Yea et B. Kim. Production planning in semiconductor wafer fab considering variable cycle times. *Computers & Industrial Engineering*, 33(3-4):713–716, 1997. 44, 51, 52

[106] C. Lemaréchal. Lagrangian relaxation. *In Computational combinatorial optimization*, pages 112–156. Springer, 2001. 69

[107] J. K. Lenstra, A. R. Kan et P. Brucker. Complexity of machine scheduling problems. *Annals of discrete mathematics*, 1:343–362, 1977. 155

[108] N. Li, L. Zhang, M. Zhang et L. Zheng. Applied factory physics study on semiconductor assembly and test manufacturing. *In ISSM 2005, IEEE International Symposium on Semiconductor Manufacturing, 2005.*, pages 307–310. IEEE, 2005. 90

[109] D.-Y. Liao, S.-C. Chang, K.-W. Pei et C.-M. Chang. Daily scheduling for R&D semiconductor fabrication. *IEEE transactions on semiconductor manufacturing*, 9(4):550–561, 1996. 28

[110] C. H. Lin, S. L. Hwang et M. Y. E. Wang. The mythical advanced planning systems in complex manufacturing environment. *In Information Control Problems in Manufacturing*, volume 12, pages 703–708, 2006. 34

[111] J. T. Lin, T.-L. Chen et H.-C. Chu. A stochastic dynamic programming approach for multi-site capacity planning in TFT-LCD manufacturing under demand uncertainty. *International Journal of Production Economics*, 148:21–36, 2014. 42

[112] J. Liu, F. Yang, H. Wan et J. W. Fowler. Capacity planning through queueing analysis and simulation-based statistical methods : a case study for semiconductor wafer fabs. *International Journal of Production Research*, 49(15):4573–4591, 2011. 90

[113] E. Lohrasbpour et S. Sathaye. Simulation modeling of IC wafer fabrication lines. *Semicon/West Technical Program*, pages 93–99, 1984. 41, 90

[114] D. P. Martin. Total operational efficiency (TOE) : the determination of two capacity and cycle time components and their relationship to productivity improvements in a semiconductor manufacturing line. *In Advanced Semiconductor Manufacturing Conference and Workshop, 1999 IEEE/SEMI*, pages 37–41. IEEE, 1999. 21

[115] S. J. Mason et J. W. Fowler. Maximizing delivery performance in semiconductor wafer fabrication facilities. *In Simulation Conference, 2000. Proceedings. Winter*, volume 2, pages 1458–1463. IEEE, 2000. 44

[116] S. J. Mason, J. W. Fowler et C.W. Matthew. A modified shifting bottleneck heuristic for minimizing total weighted tardiness in complex job shops. *Journal of Scheduling*, 5(3):247–262, 2002. 44

[117] M. Mathirajan, V. Bhargav et V. Ramachandran. Minimizing total weighted tardiness on a batch-processing machine with non-agreeable release times and due dates. *The International Journal of Advanced Manufacturing Technology*, 48(9-12):1133–1148, 2010. 65

[118] E. Mhiri, M. Jacomino, F. Mangione, P. Vialletelle et G. Lepelletier. A step toward capacity planning at finite capacity in semiconductor manufacturing. *In Proceedings of the 2014 Winter Simulation Conference*, pages 2239–2250. IEEE Press, 2014. 56, 131

[119] E. Mhiri, M. Jacomino, F. Mangione, P. Vialletelle et G. Lepelletier. Finite capacity planning algorithm for semiconductor industry considering lots priority. *IFAC-PapersOnLine*, 48(3):1598–1603, 2015. 84, 131

[120] E. Mhiri, M. Jacomino, F. Mangione, P. Vialletelle et G. Lepelletier. Prise en compte des priorités des lots pour la projection des encours de production dans l'industrie des semi-conducteurs. *In 16ème conférence ROADEF Société Française de Recherche Opérationnelle et Aide à la Décision*, 2015. 84, 131

[121] E. Mhiri, M. Jacomino, F. Mangione, P. Vialletelle et G. Lepelletier. Approche heuristique pour la projection des encours de production (WIP) à capacité finie, application à l'industrie des semi-conducteurs. *In 17ème conférence ROADEF Société Française de Recherche Opérationnelle et Aide à la Décision*, 2016. 84, 131

[122] R. J. Milne, C.-T. Wang, CK A Yen et K. Fordyce. Optimized material requirements planning for semiconductor manufacturing. *Journal of the Operational Research Society*, 63(11):1566–1577, 2012. 44

[123] L. Mönch, H. Balasubramanian, J. W. Fowler et M. E. Pfund. Heuristic scheduling of jobs on parallel batch machines with incompatible job families and unequal ready times. *Computers & Operations Research*, 32(11):2731–2750, 2005. 44

[124] L. Mönch, J. W. Fowler, S. Dauzère-Pérès, S. J. Mason et O. Rose. A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations. *Journal of Scheduling*, 14(6):583–599, 2011. 12

[125] L. Mönch, J. W. Fowler et S. J. Mason. *Production planning and control for semiconductor wafer fabrication facilities*. Springer New York, 2013. 1, 9, 12, 14, 33, 41, 65

[126] J.R. Montoya-Torres. Manufacturing performance evaluation in wafer semiconductor factories. *International Journal of Productivity and Performance Management*, 55(3/4):300–310, 2006. 21

[127] P.B. Nagendra et S.K. Das. Finite capacity scheduling method for MRP with lot size restrictions. *International Journal of Production Research*, 39(8):1603–1623, 2001. 24

[128] K. E. Nee, J. F. Chin, W. P. Loh et M. C.-L. Tan. A constraint programming-based genetic algorithm for capacity output optimization. *Journal of Industrial Engineering and Management*, 7(5):1222, 2014. 44

[129] T. J. Occhino. Capacity planning model : the important inputs, formulas, and benefits. *In Advanced Semiconductor Manufacturing Conference and Workshop, 2000 IEEE/SEMI*, pages 455–458, 2000. 26

[130] J.A. Orlicky. *Material requirements planning*. McGraw-Hill Professional, 1975. 22, 25

[131] I. M. Ovacik et R. Uzsoy. *Decomposition methods for complex factory scheduling problems*. Springer Science & Business Media, 2012. 49

[132] O. Ozturk, M. B. Coburn et S. Kitterman. Conceptualization, design and implementation of a static capacity model. *In 2003 Winter Simulation Conference*, pages 1373–1376. IEEE, 2003. 26

[133] P.C. Pandey, P. Yenradee et S. Archariyapruek. A finite capacity material requirements planning system. *Production planning & control*, 11(2):113–121, 2000. 24, 25

[134] M. Pfund, S. Mason et J.W. Fowler. Dispatching and scheduling in semiconductor manufacturing. *Handbook of production scheduling*, pages 213–241, 2006. 58

[135] M. Pinedo et X. Chao. Operations scheduling with applications in manufacturing and services, 1998. 19, 20

[136] B. T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969. 71

[137] K. Potti et S. J. Mason. Using simulation to improve semiconductor manufacturing. *Semiconductor International*, 20(8):289–292, 1997. 27, 50

[138] K. Prasad. A generic computer simulation model to characterize photolithography manufacturing area in an IC FAB facility. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, 14(3):483–487, 1991. 27

[139] A. A. B. Pritsker et K. Snyder. Production scheduling using FACTOR. *In The Planning and Scheduling of Production Systems*, pages 337–358. Springer US, 1997. 50

[140] M.G.C. Resende. A program for simulation of semiconductor wafer fabrication. Rapport technique, University of California, Berkeley, Operations Research Center, 1985. 27, 41, 50

[141] C. Rippenhagen et S. Krishnaswamy. Implementing the theory of constraints philosophy in highly reentrant systems. *In Proceedings of the 1998 Winter Simulation Conference*, pages 993–996, Piscataway, New Jersey, 1998. 25

[142] C. L. Robert, K. Jeenyoung et L. Vincent. SLIM : Short cycle time and Low Inventory in Manufacturing at samsung electronics. *Interfaces*, 32(1):61–77, 2002. 29, 42, 44

[143] J. K. Robinson. *Capacity planning in a semiconductor wafer fabrication facility with time constraints between process steps*. Thèse de doctorat, Citeseer, 1998. 121

[144] W. O. Rom, O. I. Tukel et J. R. Muscatello. Mrp in a job shop environment using a resource constrained project scheduling model. *Omega*, 30(4):275–286, 2002. 89

[145] O. Rose. Improved simple simulation models for semiconductor wafer factories. *In Winter Simulation Conference*, pages 1708–1712. IEEE, 2007. 90

[146] A. Rossi. *Ordonnancement en milieu incertain, mise en oeuvre d'une démarche robuste*. Thèse de doctorat, Institut National Polytechnique de Grenoble-INPG, 2003. 101

[147] S. C. Sarin, V. D. Shenai et L. Wang. Releasing and scheduling of lots in a wafer fab. *Lecture Notes in Computer Science*, 4508:108, 2007. 59

[148] S. C Sarin, A. Varadarajan et L. Wang. A survey of dispatching rules for operational control in wafer fabrication. *Production Planning and Control*, 22(1):4–24, 2011. 41

[149] W. Scholl et J. Domaschke. Implementation of modeling and simulation in semiconductor wafer fabrication with time constraints between wet etch and furnace operations. *IEEE Transactions on Semiconductor Manufacturing*, 13(3):273–277, 2000. 59

[150] P. Schönsleben. *Integral logistics management : planning & control of comprehensive business processes*. St. Lucie Press, Boca Raton, London, 2003. 31

[151] A. Shabbir. Semiconductor tool planning via multi-stage stochastic programming. *In Proceedings of the International Conference on Modeling and Analysis in Semiconductor Manufacturing*, pages 153–157, 2002. 44, 45

[152] J. G. Shanthikumar, S. Ding et M. T. Zhang. Queueing theory for semiconductor manufacturing systems : A survey and open problems. *IEEE Transactions on Automation Science and Engineering*, 4(4):513–522, 2007. 26, 50

[153] Y. Shen. Robust capacity modeling in semiconductor manufacturing. *Semiconduct. Intl*, 2002. 44

[154] T. W. Sloan. Shop-floor scheduling of semiconductor wafer fabs : Exploring the influence of technology, market, and performance objectives. *IEEE Transactions on semiconductor manufacturing*, 16(2):281–289, 2003. 31

[155] A. Spence et D. Welter. Capacity planning of a photolithography work cell in a wafer manufacturing line. *In IEEE International Conference on Robotics and Automation*, pages 702–708. IEEE, 1987. 27

[156] M. Stafford. A product-mix capacity planning model. Rapport technique, Technical Report ORP97-03, Graduate Program in Operations Research, The University of Texas at Austin, 1997. 44

[157] G. Sullivan et K. Fordyce. IBM Burlington's Logistics Management System. *Interfaces*, 20(1):43–64, 1990. 43, 49

[158] J. M. Swaminathan. Tool capacity planning for semiconductor fabrication facilities under demand uncertainty. *European Journal of Operational Research*, 120(3):545–558, 2000. 26, 28, 42, 44, 45, 71, 90

[159] J. M. Swaminathan. Tool procurement planning for wafer fabrication facilities : a scenario-based approach. *IIE Transactions*, 34(2):145–155, 2002. 44, 45

[160] R. Swarnkar et M.K. Tiwari. Modeling machine loading problem of fmss and its solution methodology using a hybrid tabu search and simulated annealing-based heuristic approach. *Robotics and Computer-Integrated Manufacturing*, 20(3):199–209, 2004. 102

[161] M. Taal et J. C. Wortmann. Integrating MRP and finite capacity planning. *Production Planning & Control*, 8(3):245–254, 1997. 23, 24, 25

[162] V. Tardif et M. L. Spearman. Diagnostic scheduling in finite-capacity production environments. *Computers & Industrial Engineering*, 32(4):867–878, 1997. 49

[163] M. Thompson. Using simulation-based finite capacity planning and scheduling software to improve cycle time in front end operations. *In Proceedings of 1995 IEEE/SEMI Advanced Semiconductor Manufacturing Conference Workshop*, pages 131–135, 1995. 27, 44, 50

[164] H. Toba, H. Izumi, H. Hatada et T. Chikushima. Dynamic load balancing among multiple fabrication lines through estimation of minimum inter-operation time. *IEEE Transactions on Semiconductor Manufacturing*, 18(1):202–213, 2005. 103

[165] B. Tullis, V. Mehrotra et D. Zuanich. Successful modeling of a semiconductor R & D facility. *In Proceedings of the 1990 IEEE/SEMI International Semiconductor Manufacturing Science Symposium*, pages 26–32, Burlingame, California, United States, 1990. 27, 50

[166] A. M. Uribe, J. K. Cochran et D. L. Shunk. Two-stage simulation optimization for agile manufacturing capacity planning. *International Journal of Production Research*, 41(6):1181–1197, 2003. 44

[167] R. Uzsoy, C.-Y. Lee et L.A. Martin-Vega. 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, 1992. 12, 41

[168] R. Uzsoy, C.-Y. Lee et L.A. Martin-Vega. A review of production planning and scheduling models in the semiconductor industry part II : shop-floor control. *IIE Transactions*, 26(5):44–55, 1994. 12, 41

[169] R. Uzsoy, L. A. Martin-Vega, C-Y. Lee et P. A. Leonard. Production scheduling algorithms for a semiconductor test facility. *IEEE Transactions on Semiconductor Manufacturing*, 4(4):270–280, 1991. 28, 49

[170] T.E. Vollmann, W.L. Berry, D.C. Whybark et F.R. Jacobs. *Manufacturing planning and control for supply chain management*. McGraw-Hill/Irwin New York, 2005. 18

[171] K.-J. Wang et S.-H. Lin. Capacity expansion and allocation for a semiconductor testing facility under constrained budget. *Production Planning & Control*, 13(5): 429–437, 2002. 45

[172] K.-J. Wang, Y-S. Lin, C.-F. Chien et J.C. Chen. A fuzzy-knowledge resource-allocation model of the semiconductor final test industry. *Robotics and Computer-Integrated Manufacturing*, 25(1):32–41, 2009. 30

[173] K.-J. Wang, S.-M. Wang et J.-C. Chen. A resource portfolio planning model using sampling-based stochastic programming and genetic algorithm. *European Journal of Operational Research*, 184(1):327–340, 2008. 44, 45

[174] K.-J. Wang, S.-M. Wang et S.-J. Yang. A resource portfolio model for equipment investment and allocation of semiconductor testing industry. *European Journal of Operational Research*, 179(2):390–403, 2007. 44, 45

[175] L. M. Wein. Scheduling semiconductor wafer fabrication. *IEEE Transactions on Semiconductor Manufacturing*, 1(3):115–130, 1988. 27, 41

[176] O. W. Wight. *MRP II : Unlocking America's productivity potential*. Omneo, 1981. 23, 25

[177] J. D. Witte. Using static capacity modeling techniques in semiconductor manufacturing. *In Advanced Semiconductor Manufacturing Conference and Workshop, 1996. ASMC 96 Proceedings. IEEE/SEMI 1996*, pages 31–35, 1996. 26

[178] L. A. Wolsey et G. L. Nemhauser. Integer and combinatorial optimization. *John Willey & Sons*, 1999. 57

[179] S. D. Wu, M. Erkoc et S. Karabuk. Managing capacity in the high-tech industry : A review of literature. *The Engineering Economist*, 50(2):125–158, 2005. 44, 45, 90

[180] T. Wuttipornpun et P. Yenradee. Development of finite capacity material requirement planning system for assembly operations. *Production Planning & Control*, 15(5):534–549, 2004. 24, 25

[181] J.-l. Yang. An approach to determine appropriate fab development plans by taking space constraints and cost-effectiveness into consideration. *In Semiconductor Manufacturing, 2000. Proceedings of ISSM 2000. The Ninth International Symposium on*, pages 217–220. IEEE, 2000. 44

[182] S.-J. S. Yang, F.-C. Yang, K.-J. Wang et Y. Chandra. Optimising resource portfolio planning for capital-intensive industries under process-technology progress. *International Journal of Production Research*, 47(10):2625–2648, 2009. 45

[183] S. J. Yim et D. Y. Lee. Scheduling cluster tools in wafer fabrication using candidate list and simulated annealing. *Journal of Intelligent Manufacturing*, 10(6):531–540, 1999. 29

[184] M. B. Zaremba et B. Prasad. *Modern manufacturing : information control and technology*. Springer Science & Business Media, 2012. 12

[185] W.H.M. Zijm et R. Buitenhek. Capacity planning and lead time management. *International Journal of Production Economics*, 46:165–179, 1996. 89

Annexe A

A Le projet européen *INTEGRATE*

A.1 Introduction

L'industrie européenne des composants micro-électroniques est confrontée à une forte concurrence de la part des États-Unis et de l'Asie qui attirent de plus en plus d'investissements que ce soit dans la fabrication ou la R&D (Recherche et Développement). Comme dans toute activité appliquant le principe de l'ouverture du marché, une meilleure compétitivité est donc nécessaire pour faire face à cette concurrence. Ce qui implique de proposer des produits de haute qualité à des coûts raisonnables. Le projet européen *INTEGRATE* a été lancé dans cette perspective. Il a duré 3 ans, du 01 janvier 2013 au 31 décembre 2015. Il regroupe près de 26 partenaires :

- 6 grands fabricants Européens de semi-conducteurs : STMicroelectronics France (site de Crolles et Rousset), STMicroelectronics Italie (Site d'Agrate et de Catania), LFoundry Italie, INTEL Irlande, NXP Pays-Bas.
- 5 PME fournisseurs de solutions : Probayes, Technofittings, XENIA Progetti, AVISI, Camline.
- 2 fournisseurs d'équipement : Adixen, ASM International.
- 12 partenaires académiques : Ecole des Mines de Saint Etienne - Centre Microélectronique de Provence, G-SCOP Grenoble INP, Laboratoire Gama UCBL, LSIS, TIMC-IMAG, National University of Ireland - Maynooth, Politecnico di Milano, Università Cattolica del Sacro Cuore, Università degli studi di Milano Bicocca, University of Pavia, CNR-IMM, TU-Delft.
- 1 institut : l'organisation néerlandaise pour la recherche scientifique appliquée (TNO).

A G-SCOP, deux thèses et un post-doc ont été financés par le projet *INTEGRATE*.

Le projet se compose de 6 volets ou groupes de tâches (*work packages* en anglais) dont trois (WP2, WP4 et WP5) représentent les axes de recherche majeurs visés par le projet (voir figure A.1). Le projet vise à conduire des études relatives au développement

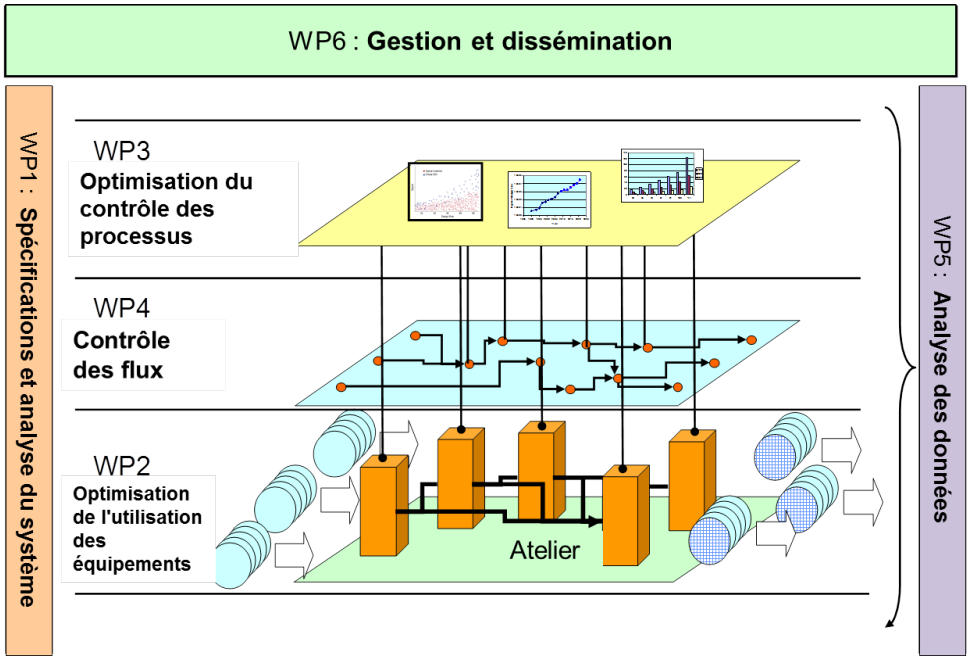

Figure A.1 – Les *work packages* du projet *INTEGRATE*

de nouveaux procédés de fabrication, de nouvelles organisations de systèmes de production, et de nouveaux outils de contrôle et d'information. L'ensemble de ces éléments devra permettre la mise en place de lignes de production capables de gérer de manière optimale la fabrication de produits associant différentes technologies, de lots hétérogènes de taille et priorité différentes pour le développement, l'ingénierie et le prototypage. Le précédent projet *IMPROVE* [87] a permis de développer les algorithmes permettant de passer d'opérations « réactives » à « prédictives ». *INTEGRATE* va plus loin en répondant aux besoins d'amélioration des environnements de production de forte variabilité.

Durant ce projet, des outils d'optimisation intégrés de contrôle de processus et de contrôle de l'équipement, ainsi que des techniques de contrôle de flux des lots avancées qui interagissent avec des niveaux de décisions inférieurs et supérieurs, mais aussi tenir compte de divers éléments de la *fab* (état de l'équipement, les ressources auxiliaires, les qualifications etc.).

A.2 WP2 : Optimisation de l'utilisation des équipements

Le premier objectif du projet est de développer des solutions qui peuvent être mises en œuvre dans les usines existantes pour accroître l'efficacité et réduire le temps perdu dans la mise en place des recettes, le changement des conditions de production, le remplacement des *batches* ou wafers et le temps d'attente d'inactivité de l'équipement.

Le deuxième objectif sera d'optimiser l'utilisation de l'équipement de production en profitant de l'état des équipements en temps réel, quantifié par l'indicateur de santé de

l'équipement (*Equipment Health Factor* ou EHF développé dans le projet *IMPROVE*) et son statut opérationnel.

A.2.1 WP2.1 : Ajustement dynamique des recettes d'équipement

La tâche 2.1 se concentre sur l'amélioration de l'utilisation de l'équipement par l'optimisation de la gestion des recettes de l'équipement et par l'ajustement dynamique des recettes pour couvrir les principaux aspects de lignes de fabrication flexibles.

A.2.2 WP2.2 : Gestion de l'équipement liée à l'état de l'équipement

La tâche 2.2 met l'accent sur l'identification des pertes d'efficacité avec un enregistrement des événements plus détaillé (les événements SECS/GEM de l'outil). L'élimination de ces pertes par l'EHF et l'amélioration du dispatching sont les objectifs principaux de ce WP.

A.2.3 WP2.3 : Performance de l'équipement / fabrication visuelle

La tâche 2.3 consiste à créer des outils qui permettent de visualiser les indicateurs et les informations des tâches 2.1 et 2.2 dans une interface graphique visuelle. Ainsi, les indicateurs de performance en temps réel deviennent directement disponibles pour conduire des réactions rapides et optimisées sur les anomalies et l'évolution des besoins de production.

Dans ces moyens visuels, nous avons l'intention d'inclure les résultats des simulations et/ou des prédictions afin d'évaluer l'impact des options d'optimisation. Pour cela, nous avons besoin de mettre en place des modèles mathématiques qui peuvent décrire les processus logistiques dans les outils complexes multi-composants.

Le résultat de ces modèles sera utilisé comme entrée pour WP4 pour optimiser l'ordonnancement des lots.

A.2.4 WP2.4 : Mise en œuvre pilote

Dans cette tâche, les implémentations pilotes des solutions (outils) décrites dans les tâches 2.1, 2.2 et 2.3 seront mises en place. Les constatations et les évaluations seront rassemblées dans un rapport d'évaluation.

A.3 WP4 : Contrôle des flux de production

L'objectif de ce WP est le développement des méthodes et des outils de planification avancée, d'ordonnancement et de répartition des lots afin de maintenir l'utilisation des

équipements à un niveau comparable, tout en réduisant la taille du lot et en augmentant le ratio des lots "très prioritaires" dans un environnement de forte variabilité.

A.3.1 WP4.1 : Planification et méthodes de répartition

La tâche 4.1 développe des algorithmes avancés pour mieux gérer la variabilité des flux dans les *wafer Fabs*. Elle comporte :

— L'intégration des indexes de santé et de performance des équipements dans l'ordonnancement et la répartition des lots

— La gestion automatique des contraintes d'enchaînement et l'optimisation multi-steps

— L'ordonnancement actif du lot en avance (optimiser à la fois l'efficacité de l'équipement et le rendement du lot ou le temps de cycle, la réduction des temps d'attente de l'équipement pour les lots prioritaires)

A.3.2 WP4.2 : Des outils de simulation et de validation

La tâche 4.2 propose des bancs d'essai et des données pour valider et alimenter les moteurs de WP4.1. Elle consiste à :

— La validation du moteur d'ordonnancement local ou des règles de répartition par la simulation.

— La validation du plan de production et l'anticipation des goulets d'étranglement dynamiques.

— Le diagnostic en temps réel des contraintes d'enchainement.

— L'évaluation de l'impact de la taille du lot et de la politique de répartition sur le temps de cycle et la capacité.

A.3.3 WP4.3 : Implémentations pilotes

Dans ce package, les implémentations pilotes concernent :

— L'intégration des EHI dans les décisions de répartition ou la qualification des recettes

— Les schedulers pour la photolithographie, la diffusion et les ateliers d'implantation ionique

— La gestion automatique des contraintes d'enchaînement

— La projection du *WIP* à capacité finie

A.4 WP5 : Analyse des données

WP5 vise à :

— manipuler des ensembles de données hétérogènes avec des caractéristiques différentes, y compris la disponibilité de l'accès aux données en temps réel .

— développer des outils statistiques dédiés de corrélation de données optimisée pour la construction de modèles de rendement pour identifier rapidement le rendement détracteur.

— évaluer et comparer dans le contexte industriel des outils et des modèles optimisés pour la construction du facteur santé de l'équipement et VM.

Annexe B

B Preuve de complexité

Lors de l'étude mathématique du problème de planification à capacité finie (cf. section 3.4), ce dernier peut être ramené à un problème d'ordonnancement classique et bien connu, le problème à une seule machine avec le critère d'optimisation est la minimisation de la somme des retards pondérés. En utilisant la notation de Graham [67], ce problème est noté $1||\sum_l w_l T_l$. La simplification effectuée est comme suit :

- $S_l = 1 \quad l = 1, \ldots, L$
- $Q_l = 1 \quad l = 1, \ldots, L$
- $r_l = 0 \quad l = 1, \ldots, L$
- $I = 1$
- $Q_{s_l,l,i} = 1 \quad l = 1, \ldots, L,\ s_l = 1, \ldots, S_l,\ i = 1, \ldots, I$
- $C_{i,t} = 1 \quad l = 1, \ldots, L,\ t = 1, \ldots, T$

Le problème $1||\sum_l w_l T_l$ est connu NP-difficile au sens fort en le réduisant à un problème de sac à dos ([101],[107]).

Problème de sac à dos Soit $a_1, \ldots, a_l, b$ des entiers positifs. Existe t-il un sous-ensemble $K \subset L = \{1 \ldots l\}$ tel que $\sum_{i \in S} a_i = b$?
La transformation du problème de planification à capacité finie est comme suit :
En utilisant la même notation utilisée lors de la présentation du problème de sac à dos, nous définissons l'instance du problème avec :

- $L = l + 1$;
- $A = \sum_{i \in L} a_i$;
- $p_{1,l,1} = w_l = a_i,\ d_l = 0 (l \in L)$;
- $p_{1,L,1} = 1, w_L = 2, d_L = b + 1$;
- $y = \sum_{1 \leq i \leq j \leq l} a_i a_j + A - b$

On a pour une séquence de process, $(\{S_l|l \in K\}, S_L, \{S_l|l \in L-K\})$ tel que $\sum_{i\in K} a_i - b = C_L - d_L$ (cf. figure B.1).

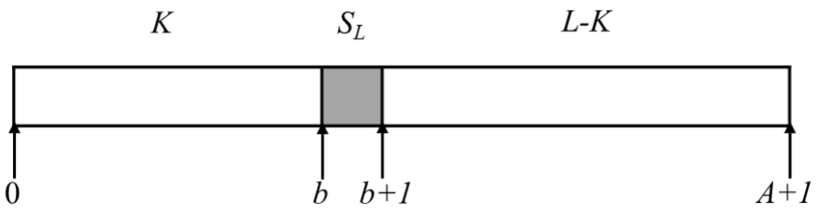

Figure B.1 – Instance du problème d'ordonnancement.

Puisque $p_{1,l,1} = w_l$ et $d_l = 0\ \forall l \in L$, la valeur de $\sum_{l\in L} w_l T_l$ n'est pas influencée par l'ordre de K et $L - K$ et nous avons :

$$\begin{aligned}\sum w_l T_l &= \sum_{l\in L} a_l C_l + 2T_L \\ &= \sum_{1\le i\le j\le l} a_i a_j + \sum_{i\in L-K} a_i + 2\max\{0, C_L - d_L\} \\ &= y + |C_L - d_L| \ge y\end{aligned}$$

Emna Mhiri : *Planification de la production à capacité finie dans un contexte à forte variabilité, application à l'industrie des semi-conducteurs.* Thèse de doctorat, 13 décembre 2016

Résumé

L'industrie des semi-conducteurs est caractérisée par une production de forte variabilité et de faible volume, des flux de production réentrants ainsi que d'un processus de fabrication complexe.

Au sein de ce contexte industriel complexe, a été considéré un problème de planification à capacité finie. C'est le problème de projection des encours de production et des commandes clients à capacité finie. Il s'agit d'estimer les dates de début, les temps d'attente et les dates de fin de chacun des steps des différents lots ainsi que la charge accumulée sur les équipements. Cette projection doit tenir compte des contraintes de capacité et qualifications des équipements et des dates d'échéance de livraison des lots. La contrainte de qualification définit l'éligibilité d'un équipement à traiter un produit. Ainsi, l'objectif de cette étude consiste à établir un plan de production réalisable à moyen terme. Afin de réaliser cet objectif, des méthodes exactes et approchées sont proposées. Des résultats en termes de complexité, et d'algorithmes de résolution, ont permis une application industrielle, dans la mesure où un logiciel de planification de la production à capacité finie a été développé.

Mots clés : Industrie des semi-conducteurs, planification de la capacité, projection du *WIP*, contraintes des qualifications, variabilité, équilibrage de charge, capacité infinie, capacité finie, programme linéaire mixte, algorithme itératif, méthodes exactes, heuristique.

Abstract

In this study, we consider the problem of production planning in the semiconductor industry characterized by high mix low volume production, reentrant flows and complex manufacturing process.

The aim of this work is to establish a feasible production schedule that takes into account the limited capacity of the manufacturing system, equipment qualifications constraints and delivery due dates. In this context, we have formulated the objective and constraints in a mixed linear program (MIP). The objective of the MIP is to minimize delivery delays to guarantee on-time delivery. While executing different tests of the MIP, we have reached a limit of resolution in a reasonable time. Thus, we use an approximate method to solve the problem. The results show the effectiveness of the heuristic established as solution quality and time resolution. The obtained results led to an industrial application and a software that provides feasible schedules in reduced execution time in a specific fab.

Keywords : Semiconductor manufacturing, capacity planning, WIP projection, qualification constraints, high mix, workload balancing, infinite capacity, finite capacity, mixed integer programming, iterative algorithm, exact methods, heuristics.