Qualification Management and Closed-Loop Production Planning in Semiconductor Manufacturing

Mehdi Rowshannahad

▶ **To cite this version:**

Mehdi Rowshannahad. Qualification Management and Closed-Loop Production Planning in Semiconductor Manufacturing. Other. Ecole Nationale Supérieure des Mines de Saint-Etienne, 2015. English. NNT : 2015EMSE0784 . tel-01231124

HAL Id: tel-01231124

https://theses.hal.science/tel-01231124

Submitted on 19 Nov 2015

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

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

NNT : **2015 EMSE 0784**

THÈSE

présentée par

Mehdi ROWSHANNAHAD

pour obtenir le grade de
Docteur de l'École Nationale Supérieure des Mines de Saint-Étienne

Spécialité : **Génie Industriel**

QUALIFICATION MANAGEMENT AND CLOSED-LOOP PRODUCTION PLANNING IN SEMICONDUCTOR MANUFACTURING

soutenue publiquement à Gardanne, le 26 mai 2015.

Membres du jury

Président	Marc SEVAUX	Professeur, Université de Bretagne-Sud
Rapporteurs	Bernard PENZ	Professeur, INP Grenoble
	Farouk YALAOUI	Professeur, Université de Technologie de Troyes
Examinateur	Jan C. FRANSOO	Professeur, Eindhoven University of Technology
Directeur de thèse	Stéphane DAUZÈRE-PÉRÈS	Professeur, École des Mines de Saint-Étienne
Co-Directeur	Nabil ABSI	Maître de Recherche, École des Mines de Saint-Étienne
Encadrant industriel	Bernard CASSINI	Ingénieur, Soitec

©2015 Mehdi ROWSHANNAHAD
Réservés tous les droits.
All Rights Reserved.

Spécialités doctorales	Responsables :	Spécialités doctorales	Responsables
SCIENCES ET GENIE DES MATERIAUX	K. Wolski Directeur de recherche	MATHEMATIQUES APPLIQUEES	O. Roustant, Maître-assistant
MECANIQUE ET INGENIERIE	S. Drapier, professeur	INFORMATIQUE	O. Boissier, Professeur
GENIE DES PROCEDES	F. Gruy, Maître de recherche	IMAGE, VISION, SIGNAL	JC. Pinoli, Professeur
SCIENCES DE LA TERRE	B. Guy, Directeur de recherche	GENIE INDUSTRIEL	A. Dolgui, Professeur
SCIENCES ET GENIE DE L'ENVIRONNEMENT	D. Graillot, Directeur de recherche	MICROELECTRONIQUE	S. Dauzere Peres, Professeur

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

ABSI	Nabil	CR	Génie industriel	CMP
AVRIL	Stéphane	PR2	Mécanique et ingénierie	CIS
BALBO	Flavien	PR2	Informatique	FAYOL
BASSEREAU	Jean-François	PR	Sciences et génie des matériaux	SMS
BATTON-HUBERT	Mireille	PR2	Sciences et génie de l'environnement	FAYOL
BERGER DOUCE	Sandrine	PR2	Sciences de gestion	FAYOL
BERNACHE-ASSOLLANT	Didier	PR0	Génie des Procédés	CIS
BIGOT	Jean Pierre	MR(DR2)	Génie des Procédés	SPIN
BILAL	Essaid	DR	Sciences de la Terre	SPIN
BLAYAC	Sylvain	MA(MDC)	Microélectronique	CMP
BOISSIER	Olivier	PR1	Informatique	FAYOL
BORBELY	Andras	MR(DR2)	Sciences et génie des matériaux	SMS
BOUCHER	Xavier	PR2	Génie Industriel	FAYOL
BRODHAG	Christian	DR	Sciences et génie de l'environnement	FAYOL
BRUCHON	Julien	MA(MDC)	Mécanique et ingénierie	SMS
BURLAT	Patrick	PR1	Génie Industriel	FAYOL
COURNIL	Michel	PR0	Génie des Procédés	DIR
DARRIEULAT	Michel	IGM	Sciences et génie des matériaux	SMS
DAUZERE-PERES	Stéphane	PR1	Génie Industriel	CMP
DEBAYLE	Johan	CR	Image Vision Signal	CIS
DELAFOSSE	David	PR0	Sciences et génie des matériaux	SMS
DESRAYAUD	Christophe	PR1	Mécanique et ingénierie	SMS
DOLGUI	Alexandre	PR0	Génie Industriel	FAYOL
DRAPIER	Sylvain	PR1	Mécanique et ingénierie	SMS
FEILLET	Dominique	PR1	Génie Industriel	CMP
FEVOTTE	Gilles	PR1	Génie des Procédés	SPIN
FRACZKIEWICZ	Anna	DR	Sciences et génie des matériaux	SMS
GARCIA	Daniel	MR(DR2)	Génie des Procédés	SPIN
GERINGER	Jean	MA(MDC)	Sciences et génie des matériaux	CIS
GOEURIOT	Dominique	DR	Sciences et génie des matériaux	SMS
GRAILLOT	Didier	DR	Sciences et génie de l'environnement	SPIN
GROSSEAU	Philippe	DR	Génie des Procédés	SPIN
GRUY	Frédéric	PR1	Génie des Procédés	SPIN
GUY	Bernard	DR	Sciences de la Terre	SPIN
HAN	Woo-Suck	MR	Mécanique et ingénierie	SMS
HERRI	Jean Michel	PR1	Génie des Procédés	SPIN
KERMOUCHE	Guillaume	PR2	Mécanique et Ingénierie	SMS
KLOCKER	Helmut	DR	Sciences et génie des matériaux	SMS
LAFOREST	Valérie	MR(DR2)	Sciences et génie de l'environnement	FAYOL
LERICHE	Rodolphe	CR	Mécanique et ingénierie	FAYOL
LI	Jean-Michel		Microélectronique	CMP
MALLIARAS	Georges	PR1	Microélectronique	CMP
MAURINE	Philippe			CMP
MOLIMARD	Jérôme	PR2	Mécanique et ingénierie	CIS
MONTHEILLET	Frank	DR	Sciences et génie des matériaux	SMS
MOUTTE	Jacques	CR	Génie des Procédés	SPIN
NEUBERT	Gilles			FAYOL
NIKOLOVSKI	Jean-Pierre	Ingénieur de recherche		CMP
NORTIER	Patrice	PR1		SPIN
PIJOLAT	Christophe	PR0	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR1	Génie des Procédés	SPIN
PINOLI	Jean Charles	PR0	Image Vision Signal	CIS
POURCHEZ	Jérémy	MR	Génie des Procédés	CIS
ROBISSON	Bruno	Ingénieur de recherche		CMP
ROUSSY	Agnès	MA(MDC)	Génie industriel	CMP
ROUSTANT	Olivier	MA(MDC)	Mathématiques appliquées	FAYOL
ROUX	Christian	PR	Image Vision Signal	CIS
STOLARZ	Jacques	CR	Sciences et génie des matériaux	SMS
TRIA	Assia	Ingénieur de recherche	Microélectronique	CMP
VALDIVIESO	François	PR2	Sciences et génie des matériaux	SMS
VIRICELLE	Jean Paul	DR	Génie des Procédés	SPIN
WOLSKI	Krzystof	DR	Sciences et génie des matériaux	SMS
XIE	Xiaolan	PR1	Génie industriel	CIS
YUGMA	Gallian	CR	Génie industriel	CMP

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

BERGHEAU	Jean-Michel	PU	Mécanique et Ingénierie	ENISE
BERTRAND	Philippe	MCF	Génie des procédés	ENISE
DUBUJET	Philippe	PU	Mécanique et Ingénierie	ENISE
FEULVARCH	Eric	MCF	Mécanique et Ingénierie	ENISE
FORTUNIER	Roland	PR	Sciences et Génie des matériaux	ENISE
GUSSAROV	Andrey	Enseignant contractuel	Génie des procédés	ENISE
HAMDI	Hédi	MCF	Mécanique et Ingénierie	ENISE
LYONNET	Patrick	PU	Mécanique et Ingénierie	ENISE
RECH	Joël	PU	Mécanique et Ingénierie	ENISE
SMUROV	Igor	PU	Mécanique et Ingénierie	ENISE
TOSCANO	Rosario	PU	Mécanique et Ingénierie	ENISE
ZAHOUANI	Hassan	PU	Mécanique et Ingénierie	ENISE

Mise à jour : 28/10/2014

To my family

for their unconditional love and support for lifelong development ...

ACKNOWLEDGEMENTS

At the end of this project, I would like to thank those who supported me throughout the course of this thesis and made it an enjoyable and extremely rewarding experience.

In the first place, I would like to thank the ANRT (Association Nationale de la Recherche et de la Technologie) and the Soitec company who financed this thesis.

I would like to express my appreciation to my academic thesis advisor, Mr Stéphane Dauzère-Pérès for his in-depth guidelines, analytical insights and conveying continuously a spirit of adventure. His presence at all stages of the thesis guaranteed its success.

I had the pleasure to have Mr Bernard Cassini as the industrial supervisor at Soitec to whom I am indebted. His professional experience, willingness to explore new ways and supportive attitude provided the foundation of a successful industrial implementation.

My sincere gratitude goes to Mr Nabil Absi who co-advised the second part of this thesis that deals with production planning. His initiative, insightful comments and deep knowledge in lot-sizing made this challenge exciting and valuable.

Warm words of thanks go to the committee members: Mr Bernard Penz, Mr Farouk Yalaoui, Mr Jan Fransoo and Mr Marc Sevaux for their constructive comments and suggestions. I appreciate their extensive and extremely detailed examination.

I owe a great deal of thanks to my University, École des Mines de Saint-Étienne, and its personnel for their invaluable support; especially to the members of my department SFL (Sciences de la Fabrication et Logistique) at Centre Microélectronique de Provence.

Many thanks go to the Soitec company, where this thesis was conducted. I especially benefited from the valuable support of my colleagues in Industrial Engineering, Supply Chain Management, ex-PDM and Production Control departments.

Mehdi ROWSHANNAHAD
Grenoble, May 2015

God bless that man whose warp is wisdom and whose weft is justice.

Shahnameh of Ferdowsî

Contents

List of Tables

List of Figures

Chapter 1

Résumé (Summary in French)

Ce chapitre présente un court extrait en français de toute la thèse[1]*. Ces travaux ont été menés dans l'industrie des semi-conducteurs. La thèse est en deux parties. La première partie concerne la gestion des qualifications menées dans le but d'augmenter l'utilisation de la capacité de production dans des îlots de production. La seconde partie étudie la planification de production de deux lignes de productions couplées.*

[1] According to French universities regulation, in case a thesis is written in English, a French summary must be provided.

1.1 Introduction

L'électronique basée sur des matériaux semi-conducteurs a révolutionnée notre quotidien. L'industrie des semi-conducteurs vit dans un constant souci de miniaturisation, de performance et d'efficacité, de baisse de consommation d'énergie et finalement de rapidité de production et d'industrialisation tout en réduisant les coûts. L'évolution constante de marché, la diversité accrue des gammes de produits, la baisse de cycle de vie des produits et le coût élevé de fabrication exigent des lignes de production flexibles et agiles. Ces travaux de thèse porte sur l'optimisation de la capacité et la planification de production. Ils ont été menés au sein de l'entreprise Soitec, le leader des produits semi-conducteurs de hautes-performances, nommés SOI (Silicium sur Isolant). Bien que ces concepts soient déployés chez Soitec, ils peuvent être adaptés à d'autres lignes de production, dans l'industrie des semi-conducteurs ou autres. Dans la Section 1.2, le cadre des travaux est posé en définissant le contexte industriel. Ensuite, différents aspects concernant la gestion des qualifications (Partie I) sont explorés sous la Section 1.3. La Partie II traite de la planification d'un système de production à boucle fermée. Ce sujet est résumé sous la Section 1.4. On tire des conclusions relatives aux deux parties tout en indiquant des perspectives dans la Section 1.5.

1.2 Contexte Industriel

Des composants à base de matériaux semi-conducteurs sont utilisés en abondance dans les appareils électroniques qui forment notre vie de tous les jours. Ces composants sont en grande partie à base de silicium. Le dioxyde de silicium se trouve en abondance sous la forme de sable. Après un procédé spécifique, des tranches ou des plaques de silicium ("wafer" en anglais) sont fabriquées à partir de silicium d'une extrême pureté. Des composants électroniques, tels que des diodes, transistors, circuits intégrés et des puces sont ensuite gravés sur ces plaques de silicium. Vu la taille miniature des composants, la moindre impureté dans une étape de production les endommagerait, d'où la fabrication dans des salles extrêmement propres dites *salle blanche* ("clean room" en anglais). On appelle aussi des unités de fabrication

de semi-conducteurs, des *fabs*.

Les appareils ont d'autant plus de problèmes de batterie. En même temps, les exigences en terme de performances augmentent. Par conséquent, des matériaux semi-conducteurs de hautes performances ont été développés. Des tranches de SOI (Silicium-sur-Isolant et "Silicon-on-Insulator" en anglais) répondent à ces besoins. Une tranche de SOI est composée d'une mince couche de silicium active. Une couche de dioxyde de silicium et un support à base de silicium (voir Figure 1.1). Des

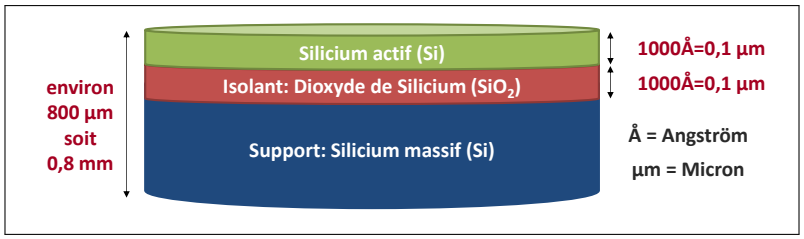


Figure 1.1: SOI ou Silicium-sur-Isolant

plaques SOI peuvent être fabriquées grâce à différents procédés. Une technologie brevetée qui permet une fabrication économique, industrielle et de qualité s'appelle la Technologie Smart-Cut™. Le procédé Smart-Cut™ est illustré dans la Figure 1.2. Les étapes de fabrication des plaques de SOI sont [66]:

- La plaque de silicium A (aussi appelée "Top") est oxydée thermiquement. L'idée est de créer une couche d'oxyde à la bonne épaisseur et avec la bonne uniformité;

- À l'étape d'implantation, le rôle est de faire pénétrer des ions d'hydrogène $H+$ en profondeur dans le silicium oxydé. La profondeur est déterminée en fonction de l'épaisseur de Si actif souhaitée. Des ions implantés créent une zone de fracture;

- La plaque A est nettoyée avant de passer à l'étape de collage. Les plaques sont traitées dans des bains chimiques afin d'ôter les contaminations métalliques et particulaires. Le nettoyage avant le collage rend les plaques hydrophiles pour faciliter le collage. Une fois le Top nettoyé, il est collé

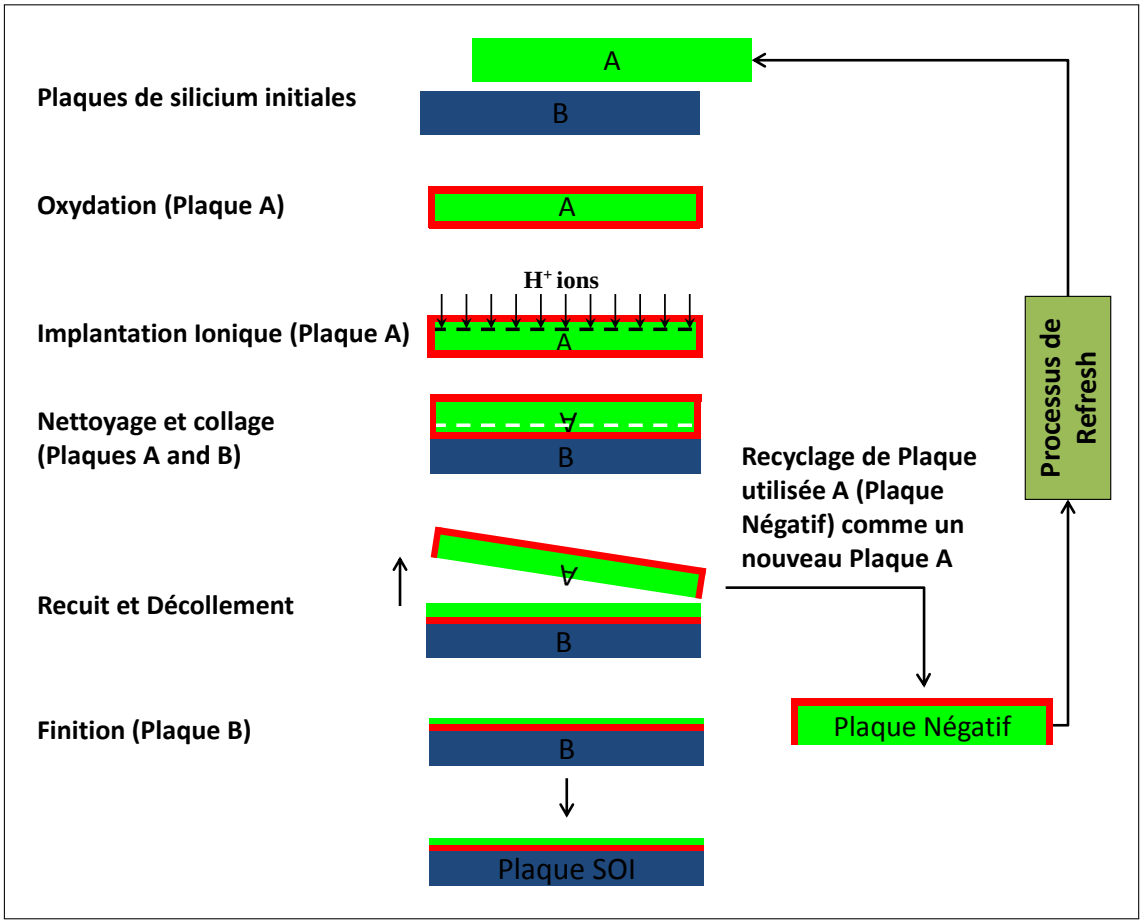


Figure 1.2: La Technologie Smart-Cut™ Utilisée pour Produire les Plaques SOI

avec la plaque B (appelée le Base). Le collage est dit “hydrophile par adhésion moléculaire”. Il est basé sur l’attraction de couches d’eau présentes à la surface des plaques;

- Ensuite, les plaques sont repassées dans des fours à l’étape *recuit*. Lors de cette étape, les ions implantés se recombinent et se dilatent de façon à générer la fracture nécessaire à l’obtention de la plaque SOI. Au niveau de l’interface de collage, les liaisons hydrogène sont remplacées par des liaisons atomiques plus solides. Le décollement se fait après de manière mécanique;

- Les trois dernières étapes de finition sont *le polissage*, *le RTA* et *la stabilisation et l'amincissement*. Le polissage a pour objectif de diminuer la rugosité de la plaque. C'est un procédé mécano-chimique qui combine l'action du frottement sur un pad avec l'action chimique du slurry. Le RTA ("Rapid Thermal Annealing" en anglais) est un recuit rapide qui permet de lisser la surface de la plaque après décollement. La stabilisation est une étape d'oxydation superficielle qui a pour but d'éliminer une couche de silicium abîmée lors du décollement ainsi que de stabiliser l'interface de collage créée lors du recuit. L'amincissement est également une oxydation superficielle qui permet d'amener le SOI à l'épaisseur finale demandée par le client.

La plaque de silicium qui reste après le décollement peut être retravaillée afin d'être réutilisée comme un nouveau Top. Une plaque Top déjà utilisée est appelée *un négatif*. Sachant que seulement une couche très mince de plaque Top reste sur la plaque SOI, la plaque négatif générée peut être retravaillée afin de revenir dans la ligne de fabrication de SOI pour en faire d'autres. Le processus de refabrication d'une plaque négatif est appelé *le processus de refresh* ou tout court *le refresh*. Cette spécificité représente un des grands avantages économiques de la Technologie Smart-Cut™.

1.3 La Gestion des Qualifications

1.3.1 La Gestion des Qualifications dans la Fabrication Semi-conducteur

Dans l'industrie des semi-conducteurs, de multiples opérations sont exécutées dans différentes étapes de production pour fabriquer un produit. Chaque étape est réalisée dans un poste de travail ("workcenter" ou "workstation" en anglais). Chaque poste de travail est un ensemble composé d'opérateurs, d'appareils de manutention et d'un parc d'équipements ("toolset" en anglais). Ce dernier est constitué de machines parallèles et souvent non-identiques qui exécutent des opérations similaires.

Par exemple, un ensemble de fours font le parc d'équipements *Traitement Thermiques* (tout court: TTH). Les machines dans un parc d'équipements peuvent avoir différentes caractéristiques en ce qui concerne, le débit de production ("throughput" en anglais) ou le temps de processus ("process time" en anglais), la taille de lot, la configurabilité, le logiciel, etc. Une *recette* est associée à chaque opération. Elle définit les instructions nécessaires pour obtenir le processus souhaité. Par exemple, une recette de TTH définit les différentes phases de la montée ou descente en température, les paliers de températures ainsi que la pression de gaz à chaque phase aussi bien que la vitesse de refroidissement.

Afin de pouvoir exécuter une opération sur une machine, la recette correspondant à l'opération doit déjà être qualifiée sur celle-ci. À cause des restrictions matérielles et informatiques, il n'est pas toujours possible de faire tourner toutes les opérations sur toutes les machines. Autrement dit, toutes les recettes ne sont pas qualifiables sur toutes les machines d'un parc d'équipements. L'exécution d'une nouvelle qualification peut être rapide, par exemple l'équivalent d'une *séquence de production* ("production run" en anglais). Toutefois, elle peut s'avérer longue, par exemple aussi longue qu'un cycle de production. Quelle que soit la durée, une qualification coûte en temps et en énergie. En règle générale, une configuration de qualification recette-machine pourrait prendre un des trois statuts suivants. Quand une recette n'est pas autorisée sur une machine, elle est *non-qualifiable*. Si la qualification est autorisée, elle est *qualifiable*. Et finalement, une recette qualifiable peut aussi être (déjà) *qualifiée*.

L'idéal serait que toutes les recettes soient qualifiées sur tout le parc d'équipements. Ainsi on pourrait allouer le volume de production (la charge ou "workload" en anglais) de chaque recette sur n'importe quelle machine sans aucune restriction. Pourtant, comme cela a déjà été dit, toutes les recettes ne sont pas qualifiables sur toutes les machines. En outre, une qualification reste coûteuse. Outre cela, l'état des qualifications est très dynamique, par exemple de nouvelles recettes sont souvent créées avec l'apparition de nouveaux produits; les recettes évoluent dans le temps créant de nouvelles recettes; la perte des qualifications (souvent dite "disqualification") surviendra suite aux maintenances, pannes machines ou absence de tournage d'une recette pendant un temps excessif sur une machine, etc. Toutes ces contraintes

démontrent l'importance de la gestion des qualifications [55].

Un des défis du management est d'ajuster les contraintes de capacité avec le plan de production afin de répondre à la demande. La capacité devrait être optimisée en équilibrant au mieux la charge suivant le temps productif des équipements et la main-d'œuvre disponible. La configuration des qualifications a un impact direct sur l'utilisation de la capacité d'un parc d'équipements. Il est impossible d'allouer une charge de production associée à une recette à un équipement non-qualifié. Donc, une configuration des qualifications adéquate permet l'optimisation de l'utilisation de la capacité disponible et garantit le bon fonctionnement de la fab. Une telle configuration ajoute de la flexibilité au système manufacturier en donnant plus de choix pour l'allocation des charges.

Par conséquent, une stratégie de gestion des qualifications est nécessaire pour garantir la satisfaction de la demande client et l'optimisation de l'utilisation de la capacité des équipements.

1.3.2 La Gestion des Qualifications sous Contrainte de Capacité

Les machines ne sont pas toujours disponibles pour la production. Le temps total disponible des machines est découpé en différentes catégories (voir Figure 1.3). Seul le créneau nommé "productive time" est vraiment utilisé pour la production. Cette donnée doit être considérée lors de l'équilibrage de charge et donc dans la gestion des qualifications. Des mesures de flexibilité ont été définies dans [54] pour la gestion des qualifications. Ces indicateurs qui varient entre 0 et 1 (0% et 100%) mesurent le lissage de l'équilibrage optimal des charges. On les appelle *mesures de flexibilité à capacité infinie* ("uncapacitated flexibility measures" en anglais) car elles ne tiennent pas compte de la capacité des machines. L'idée dans la gestion des qualifications est de mesurer le gain de flexibilité potentiel associé à la qualification d'un couple recette-machine qualifiable. Pour calculer ce gain, la flexibilité de la configuration actuelle des qualifications dans un parc d'équipements est mesurée. Ensuite à chaque fois, un couple recette-machine qualifiable est virtuellement qualifié, et la flexibilité associée à cette nouvelle configuration est calculée. Ces calculs constituent la matrice

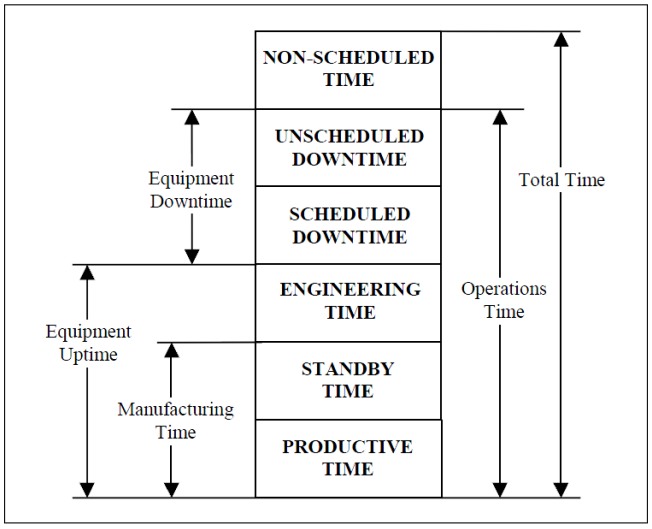


Figure 1.3: États d'un Équipement ([86])

de flexibilité après les qualifications potentielles. Pour calculer le gain correspondant à chaque nouvelle qualification, la flexibilité de la configuration initiale est déduite de chaque élément de cette matrice. Ainsi les gains de flexibilité de chaque couple qualifiable recette-machine est calculé.

Par la suite, les *mesures de flexibilité à capacité finie* (F_{Capa}) sont définies.

Paramètres

R	Le nombre total des recettes à faire,
M	Le nombre total des machines dans un parc d'équipements,
WIP_r	Le volume de production total associé à chaque recette r,
$TP_{r,m}$	Le débit de production de la recette r sur la machine m (nombre de plaques traité par heure),
$Capa_m$	La capacité de la machine m (en heures),
$Q_{r,m}$	$\begin{cases} 1 \text{ si recette } r \text{ est qualifiée sur machine } m, \\ 0 \text{ si recette } r \text{ n'est pas qualifiée sur machine } m. \end{cases}$
γ	L'exposant de l'équilibrage de charge ($\gamma \geq 1$).

Variables

$WIP_{r,m}$	Le volume de production de la recette r alloué à la machine m,
$C_{r,m}$	Le temps de production de la recette r alloué à la machine m,
WIP_m	Le volume de production total alloué à la machine m ($WIP_m = \sum_{r=1}^{R} WIP_{r,m}$),
C_m	Le temps de production total alloué à la machine m ($C_m = \sum_{r=1}^{R} \frac{WIP_{r,m}}{TP_{r,m}}$).

L'équilibrage de charge peut se faire en terme de *volume de production* ou *temps de production*. Afin de calculer la flexibilité associée à une configuration de qualification d'un parc d'équipements, un modèle d'optimisation (1.1) doit être résolu. La mesure de flexibilité souhaitée constitue la fonction objectif. La seule contrainte assure que tout le volume de production de chaque recette est fabriqué et cela seulement sur les machines qualifiées. En maximisant la fonction objectif, l'équilibrage de charge optimal et la valeur de flexibilité sont obtenus.

$$\begin{aligned} \max \quad & F_{Capa} \\ \text{subject to} \quad & \\ & \sum_{m=1|Q_{r,m}=1}^{M} WIP_{r,m} = WIP_r \qquad \forall r \\ & WIP_{r,m} \geq 0 \qquad \forall r, m \end{aligned} \tag{1.1}$$

1.3.2.1 Gestion des qualifications à capacité finie en terme de volume de production

F_{Capa}^{WIP} (1.2) mesure la flexibilité en terme de *volume de production* ("WIP" en anglais pour en-cours).

$$F_{Capa}^{WIP} = \frac{\left(\frac{\sum_{m=1}^{M} (WIP_m / Capa_m)}{M}\right)^{\gamma}}{\frac{\sum_{m=1}^{M} (WIP_m / Capa_m)^{\gamma}}{M}} \in (0, 1] \tag{1.2}$$

Même avec un équilibrage imparfait, la mesure de flexibilité F_{Capa}^{WIP} pourrait atteindre son maximum. Cette information peut induire les décideurs en erreur lors de la prise de décisions pour faire de nouvelles qualifications. Pour contrer cette conséquence non-désirable l'exposant γ est prévu. En l'augmentant, la valeur de flexibilité baisse sans aucun impact sur l'équilibrage de charge. Donc, à l'aide de γ, il faut ajuster la valeur de flexibilité de manière qu'elle représente au mieux l'équilibrage réel de la charge dans la fab.

En considérant la capacité finie des équipements, la mesure de flexibilité ne peut seule conduire le décideur vers le meilleur choix de qualification. En effet, elle nous informe seulement sur la qualité de l'équilibrage. À titre d'exemple, elle atteint son maximum si la charge est parfaitement équilibrée même si chaque machine est disons deux fois plus (ou moins) chargée que sa capacité. Donc, des indicateurs complémentaires, nommés *ratio de déviation de la capacité* DR_{Capa}, sont introduites pour assurer une meilleure prise de décision. Le Ratio de déviation de la capacité pour le volume de production DR_{Capa}^{WIP} est défini dans (1.3).

$$DR_{Capa}^{WIP} = \frac{\sum_{m=1}^{M} |WIP_m - Capa_m|}{\sum_{m=1}^{M} Capa_m} \tag{1.3}$$

1.3.2.2 Gestion des qualifications à capacité finie en terme de temps de production

La mesure de flexibilité à capacité finie en terme de temps de production F_{Capa}^{Time} est définie dans (1.4).

$$F_{Capa}^{Time} = \frac{Ideal\ Ratio_{Capa}}{\sum_{m=1}^{M} \left(\sum_{r=1}^{R} \frac{WIP_{r,m}}{TP_{r,m} \cdot Capa_m} \right)^{\gamma}} \in (0, 1]\,. \tag{1.4}$$

Le calcul de F_{Capa}^{Time} se fait en deux phases. Dans un premier temps, tous les couples recette-machine qualifiables sont virtuellement qualifiés pour calculer la constante $Ideal\ Ratio_{Capa}$ (1.5). Il représente le meilleur équilibrage qu'on pourrait atteindre

en qualifiant tous les couples recette-machine qualifiables. Pour cela il suffit de considérer (1.5) comme la fonction objectif du Modèle (1.1) sachant que cette fois, il faut la minimiser.

$$Ideal\ Ratio_{Capa} = \min \sum_{m=1}^{M} \left(\sum_{r=1}^{R} \frac{WIP_{r,m}}{TP_{r,m} \cdot Capa_m} \right)^{\gamma} \quad \text{with } Q_{r,m} = 1 \quad \forall r, m. \tag{1.5}$$

F_{Capa}^{Time} est une fonction convexe sous $\gamma \geq 1$. L'augmentation de γ non seulement baisse la valeur nominative de flexibilité mais elle a aussi un impact sur l'équilibrage de charge. En l'augmentant, on diminue l'écart entre la capacité de chaque machine et le temps de production total qu'on y alloue. Ceci au détriment de l'allègement de charge des machines rapides au profit des machines lentes.

Sur le même principe que pour le volume de production, le ratio de déviation de la capacité pour le temps de production est défini dans (1.6).

$$DR_{Capa}^{Time} = \frac{\sum_{m=1}^{M} \left| \sum_{r=1}^{R} \frac{WIP_{r,m}}{TP_{r,m}} - Capa_m \right|}{\sum_{m=1}^{M} Capa_m} \tag{1.6}$$

Des exemples et extensions concernant la gestion des qualifications à capacité finie sont décrits dans le Chapitre 4. Ce chapitre comporte aussi des méthodes de résolution dédiées à l'équilibrage de charge en terme de volume de production et temps de production.

En résumé, la capacité des machines doit être prise en compte lors de l'équilibrage de charge ainsi que la gestion des qualifications.

1.3.3 La Gestion des Qualifications sous Contrainte de Taille de Batch

Le regroupement des produits lors de la production ou *batching* est une pratique commune en industrie. En industrie des semi-conducteurs, les plaques de silicium parcourent l'ensemble des étapes de production par lots de 25. Suivant l'opération ou l'équipement, les lots sont regroupés (ou parfois même dégroupés) pour faire des

batches. Les lots dans un batch reçoivent la même opération. Le batching réduit la variabilité de fabrication tout autant que le coût de production. Donc, il est préférable de faire tourner les processus longs ou avec un coût de lancement élevé en batch.

La taille de batch peut varier dans le même parc d'équipements d'une recette à l'autre et pour une même recette d'une machine à l'autre. La variation de taille de batch d'un parc d'équipements TTH peut aller de 4 à 10 lots. Cette contrainte constitue une source importante de la variabilité et a un impact direct sur l'allocation dans l'équilibrage de charges et par conséquent sur la gestion des qualifications.

Le modèle d'optimisation tenant compte des tailles de batch est présenté au Chapitre 5. L'analyse de complexité montre que le problème est NP-difficile. Plusieurs heuristiques sont développés pour l'équilibrage de charges en terme de volume ainsi que de temps de production. Grâce aux expérimentations sur les instances industrielles de Soitec, les performances des heuristiques sont analysées notamment en comparaison avec le logiciel commercial Cplex.

En conclusion, la contrainte de la taille de batch doit être considérée lors de l'équilibrage de charge et de la gestion des qualifications pour assurer le meilleur choix de qualification.

1.3.4 La Gestion des Qualifications et la Variabilité des Charges

La variabilité dans un système manufacturier comprend les variations dans les flux de production. Elle peut avoir des sources *stochastiques* ou *déterministes*. Les sources stochastiques restent non contrôlables, par exemple: la demande, la panne des équipements, etc. En revanche, les sources déterministes sont maîtrisables, par exemple: les contraintes liées aux processabilités des produits sur les équipements, le batching, les contraintes liées au lancement, les flux ré-entrants, etc.

La variabilité est l'ennemie de la production. Elle a un impact négatif sur la planification de production, l'ordonnancement et toute gestion liée à la production. Cette influence négative est plus grave en ce qui concerne les goulots d'étranglement. Il faut donc maîtriser la variabilité, surtout dans les étapes critiques, afin d'éviter sa propagation dans toute la ligne de production. La variabilité conduit aussi à une

perte de capacité. Vu le coût élevé de production en industrie des semi-conducteurs, une utilisation optimale de la capacité disponible est souhaitée.

La variabilité diminue avec l'augmentation de la flexibilité. La flexibilité dans un parc d'équipements est la conséquence des possibilités liées à l'allocation des produits aux machines. Plus de flexibilité permettrait une meilleure allocation des charges et par conséquent l'optimisation de la capacité.

L'impact de la gestion des qualifications sur la variabilité est étudié au Chapitre 6. Des mesures sont introduites pour évaluer la variabilité des charges dans un parc d'équipements. Elles sont utilisées pour proposer de nouvelles qualifications afin de réduire la variabilité et augmenter l'utilisation de la capacité du parc d'équipements.

La mesure de la variabilité (1.7) est introduite pour évaluer la variabilité des charges dans un parc d'équipements. Pour chaque nouvelle qualification possible, la réduction en terme de variabilité est calculée. La meilleure qualification est choisie suivant la réduction de variabilité.

$$Var_{Capa}^{Time} = \sum_{m=1}^{M} (C_m - Capa_m)^{\gamma} \tag{1.7}$$

Le rôle de γ est le même que dans la mesure de flexibilité pour les temps de production (1.4). La mesure de variabilité (1.7) est similaire au ratio de déviation de capacité (1.6) non-normalisée et avec l'exposant de l'équilibrage de charge. Elle tente de diminuer l'écart de la charge allouée à chaque machine et sa capacité disponible. D'autres variantes des mesures de variabilité sont définies au Chapitre 6.

$$\begin{aligned} \min \quad & Var: && && (1.8a) \\ \text{Subject to} \quad & \sum_{m=1|Q_{r,m}=1}^{M} WIP_{r,m} = WIP_r && \forall r && (1.8b) \\ & WIP_{r,m} \geq 0 && \forall r, m \end{aligned}$$

Les mesures de variabilité doivent être minimisées (1.8a) tout en assurant que tout le volume de production de chaque recette est fabriqué et cela seulement sur les machines qualifiées (1.8b). On explique ensuite comment adapter les méthodes de résolution au Chapitre 4 pour résoudre le modèle (1.8).

Des expérimentations sont menées sur le parc d'équipements TTH de Soitec pour étudier l'impact d'une nouvelle qualification sur la variabilité des charges. En premier lieu, on considère une instance de 22 machines et 37 recettes pour une seule période. Le diagramme 1.4 montre l'équilibrage de charge avec la configuration de qualification actuelle. Les lignes horizontales définissent la capacité des machines. Chaque barre verticale montre la charge associée à chaque machine sachant que chaque couleur représente le volume d'une recette. Le dépassement des barres des lignes horizontales indique une surcharge et à l'inverse, une sous-charge. En calcu-

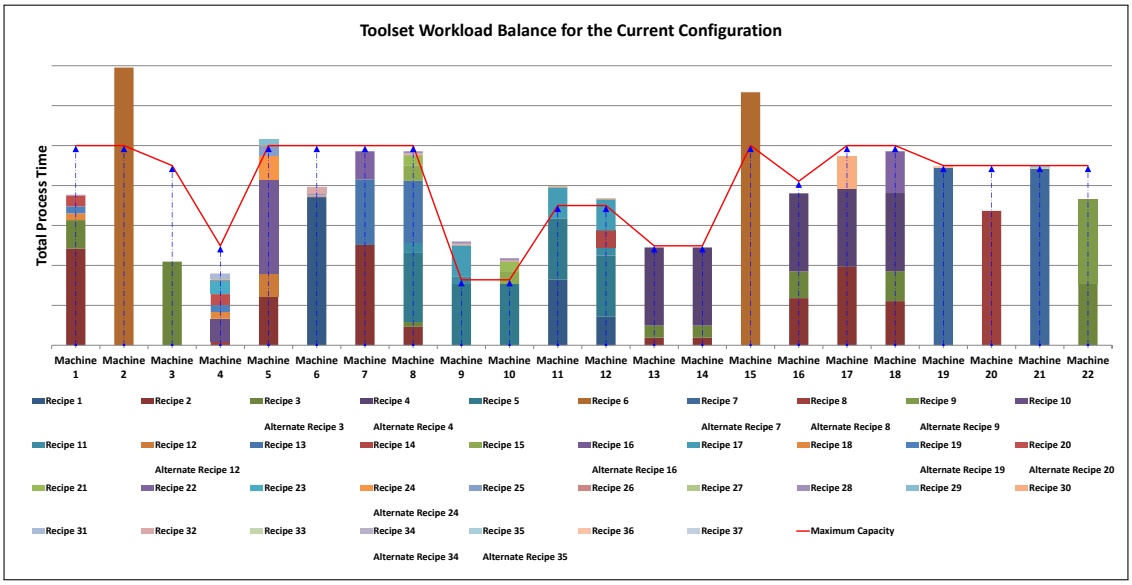


Figure 1.4: L'Équilibrage de Charge pour la Configuration Actuelle des Qualifications

lant la réduction de variabilité associée à chaque nouvelle qualification (ce qui signifie ajouter de la flexibilité), la qualification qui réduit le plus possible la variabilité est choisie. Le diagramme de l'équilibrage de charge (Figure 1.5) illustre comment la variabilité des charges est réduite après une nouvelle qualification, autrement dit après l'augmentation de la flexibilité. En continuant à faire de nouvelles qualifications qui diminuent le plus la variabilité, l'équilibrage de charge s'améliore. Au lieu de montrer les diagrammes de l'équilibrage de charge pour chaque nouvelle qualification, quelques indicateurs de performance sont utilisés dans le Tableau 1.1. On constate qu'en faisant de plus en plus de nouvelles qualifications diminuant la vari-

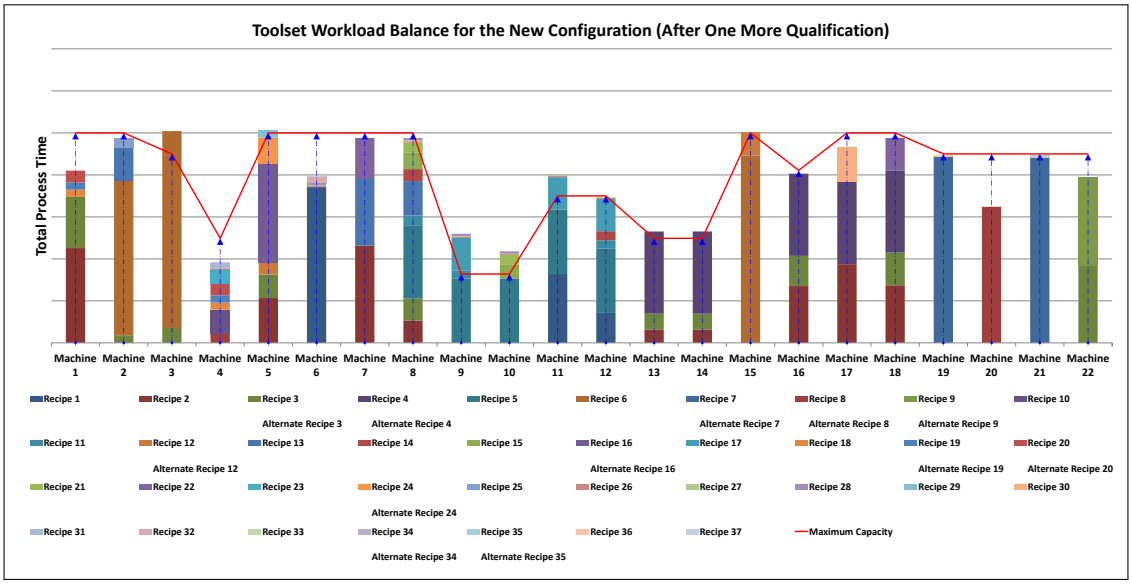


Figure 1.5: L'Équilibrage de Charge pour la Configuration après une Qualification Meilleure

abilité, on diminue la surcharge et cela en utilisant la capacité non-utilisée et en répartissant mieux la charge totale. En même temps, on remarque que l'impact de nouvelles qualifications baisse graduellement. Donc, à un certain moment un compromis doit être fait entre la diminution de la variabilité souhaitée et l'augmentation de la flexibilité. La Figure 1.6 présente les résultats de Tableau 6.1. Pour dix instances industrielles, le Tableau 1.2 montre l'impact d'une nouvelle qualification sur la réduction de la variabilité. En général, la meilleure qualification réduisant le plus la variabilité, réduit la surcharge et la capacité non-utilisée des machines considérablement, tout en augmentant légèrement la charge totale sur l'ensemble de toolset. À titre d'exemple, dans la première instance, la surcharge est entièrement éliminée (-100%), l'utilisation de la capacité non-utilisée est considérablement augmentée $(20,20\%)$ tandis que la charge totale est légèrement augmentée de $3,40\%$. Les expérimentations montrent aussi que les variations ne suivent pas un profil linéaire. Suivant le mix des produits, les caractéristiques et la configuration du parc d'équipements, le même montant de réduction de la variabilité aura plus ou moins d'impact sur les indicateurs de performances. Les Instances 1 et 10 ou 4 et 9 en sont l'exemple. En ce qui concerne les Instances 1 et 10, la variabilité est ré-

Table 1.1: Nombre de nouvelles qualifications versus la réduction de la variabilité et les variations des indicateurs de performance

	Variation			
Nouvelle(s) Qualification(s)	Variabilité	Charge	Surcharge	Capacité Non-Utilisée
1	-77,44%	3,33%	-37,33%	-32,72%
2	-86,32%	4,08%	-47,11%	-40,02%
3	-89,94%	4,55%	-57,58%	-44,66%
4	-94,41%	4,94%	-59,48%	-48,48%
5	-95,58%	5,22%	-62,22%	-51,20%
6	-96,08%	5,28%	-61,99%	-51,83%
7	-96,12%	5,23%	-65,82%	-51,31%
8	-96,60%	5,48%	-66,75%	-53,78%
9	-97,55%	5,84%	-72,56%	-57,33%
10	-97,55%	5,84%	-72,56%	-57,33%

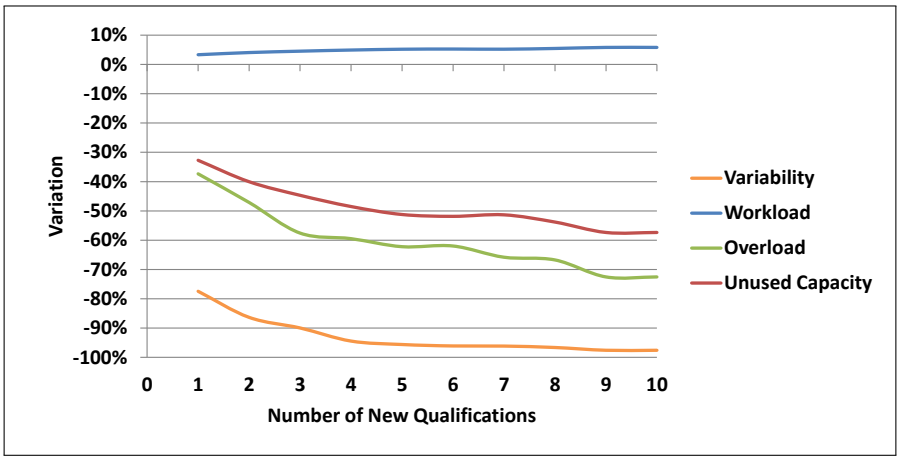


Figure 1.6: Les Variations de la Variabilité des Charges, des Surcharges, et de la Capacité Non-Utilisée versus le Nombre de Nouvelles Qualifications

duite d'environ 4% ($-4,13\%$ et $-4,71\%$) tandis que la variation des indicateurs de performances est plus violente: c'est le cas pour la variation des charges ($3,40\%$ et $1,20\%$), de la surcharge (-100% et $-1,41\%$) et de la capacité non-utilisée ($-20,20\%$ et $-5,49\%$). Une nouvelle qualification est équivalente à la création d'un lien de plus entre l'ensemble des recettes et celui des machines. Un nouveau lien ajouterait

Table 1.2: La réduction de la variabilité en faisant une nouvelle qualification

	Variation			
Numéro d'Instance	Variabilité	Charge	Surcharge	Capacité Non-Utilisée
1	-4,13%	3,40%	-100,00%	-20,20%
2	-34,52%	2,16%	-46,76%	-7,36%
3	-76,11%	4,43%	-89,42%	-22,41%
4	-43,39%	3,99%	-28,97%	-16,76%
5	-3,29%	2,20%	-21,47%	-18,06%
6	-21,90%	2,29%	-20,05%	-15,80%
7	-62,82%	1,06%	-5,80%	-5,03%
8	-99,76%	4,57%	-30,10%	-87,41%
9	-49,65%	1,75%	-18,86%	-18,23%
10	-4,71%	1,20%	-1,41%	-5,49%

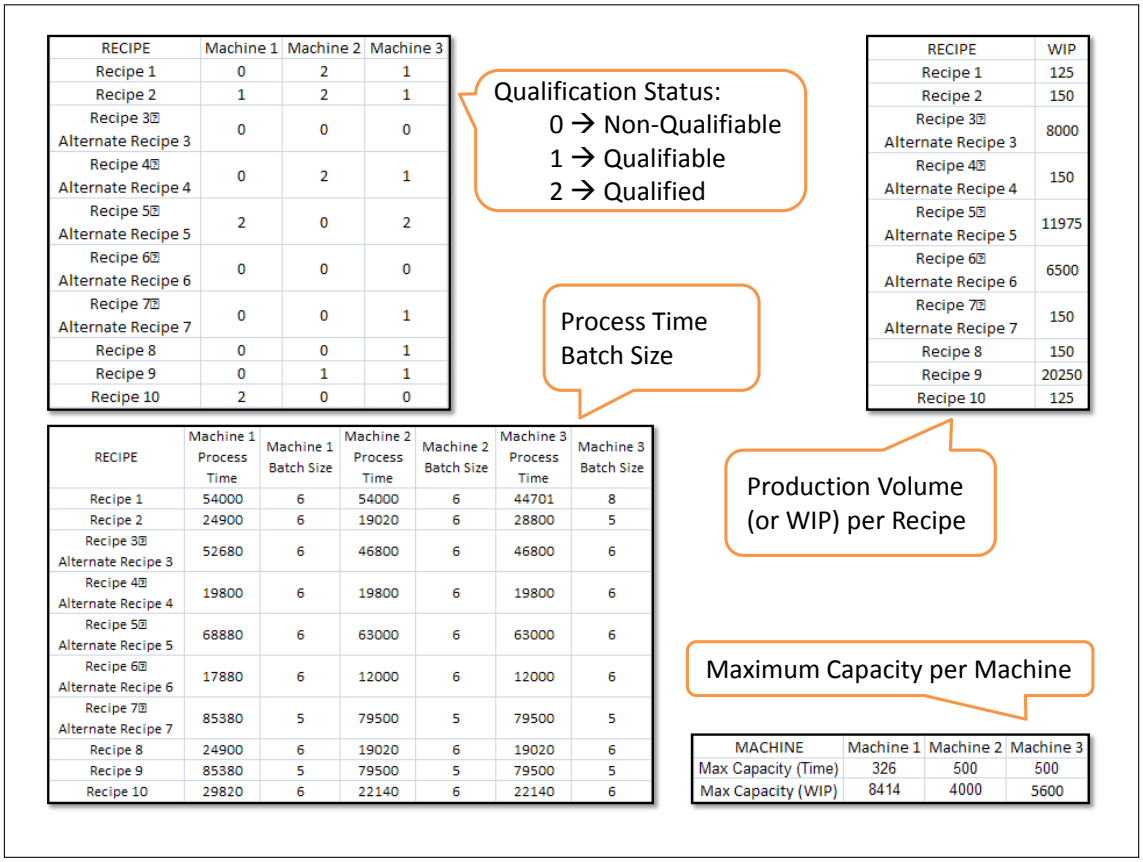
de la flexibilité au parc d'équipements et diminuerait la variabilité.

En résumé, plus de flexibilité absorbe la variabilité. Donc, une plus grande flexibilité est requise là où la variation des charges est la plus élevée mais pas forcément là où la charge est la plus élevée. En d'autres mots, si la variabilité est basse mais qu'en parallèle il subsiste des surcharges, de nouvelles qualifications n'apporteraient pas de flexibilité supplémentaire et ne résoudraient pas le problème. Dans ce cas, d'autres scénarios comme l'achat de nouveaux équipements doivent être considérés.

1.3.5 Industrialisation de la Gestion des Qualifications

Les concepts expliqués dans cette partie de thèse ont été industrialisés au sein de l'entreprise Soitec. Un nouveau processus de prise de décision a été introduit et mis en place. Ce processus implique le recueil des données qui consiste en *états des qualifications*, *débits de production*, *tailles de batch*, *volumes de production* et *capacités des machines* (voir Figure 1.7). La matrice des états de qualification comporte trois codes pour désigner l'état des qualifications des couples recette-machine: 0 pour *non-qualifiable*, 1 pour *qualifiable* et 2 pour (déjà) *qualifié*.

Les données d'entrée sont chargées dans l'application de gestion des qualifica-



RECIPE	Machine 1	Machine 2	Machine 3
Recipe 1	0	2	1
Recipe 2	1	2	1
Recipe 3 Alternate Recipe 3	0	0	0
Recipe 4 Alternate Recipe 4	0	2	1
Recipe 5 Alternate Recipe 5	2	0	2
Recipe 6 Alternate Recipe 6	0	0	0
Recipe 7 Alternate Recipe 7	0	0	1
Recipe 8	0	0	1
Recipe 9	0	1	1
Recipe 10	2	0	0

RECIPE	WIP
Recipe 1	125
Recipe 2	150
Recipe 3 Alternate Recipe 3	8000
Recipe 4 Alternate Recipe 4	150
Recipe 5 Alternate Recipe 5	11975
Recipe 6 Alternate Recipe 6	6500
Recipe 7 Alternate Recipe 7	150
Recipe 8	150
Recipe 9	20250
Recipe 10	125

RECIPE	Machine 1 Process Time	Machine 1 Batch Size	Machine 2 Process Time	Machine 2 Batch Size	Machine 3 Process Time	Machine 3 Batch Size
Recipe 1	54000	6	54000	6	44701	8
Recipe 2	24900	6	19020	6	28800	5
Recipe 3 Alternate Recipe 3	52680	6	46800	6	46800	6
Recipe 4 Alternate Recipe 4	19800	6	19800	6	19800	6
Recipe 5 Alternate Recipe 5	68880	6	63000	6	63000	6
Recipe 6 Alternate Recipe 6	17880	6	12000	6	12000	6
Recipe 7 Alternate Recipe 7	85380	5	79500	5	79500	5
Recipe 8	24900	6	19020	6	19020	6
Recipe 9	85380	5	79500	5	79500	5
Recipe 10	29820	6	22140	6	22140	6

MACHINE	Machine 1	Machine 2	Machine 3
Max Capacity (Time)	326	500	500
Max Capacity (WIP)	8414	4000	5600

Figure 1.7: Le Format des Données d'Entrée

tions. En sortie, les propositions des qualifications ainsi que l'équilibrage optimal de charges sont affichés (voir Figure 1.8).

La Figure 1.9 illustre les différentes interfaces pour la récupération des données. Les volumes par recette sont extraits en croisant les volumes de production par produit et les routings associés à chacun des produits. La capacité des machines est déterminée en considérant l'historique de l'uptime des machines, la cible de l'uptime et les maintenances en vue. Les états des qualifications, les débits de production et les tailles de batch sont fournis par le département "process".

Une application nommée "Input File Creator" rassemble toutes ces données en une feuille Excel dans un format prédéfini. Ensuite, l'utilisateur charge cette feuille dans l'application de gestion des qualifications (appelée "Nexus" chez Soitec). Les

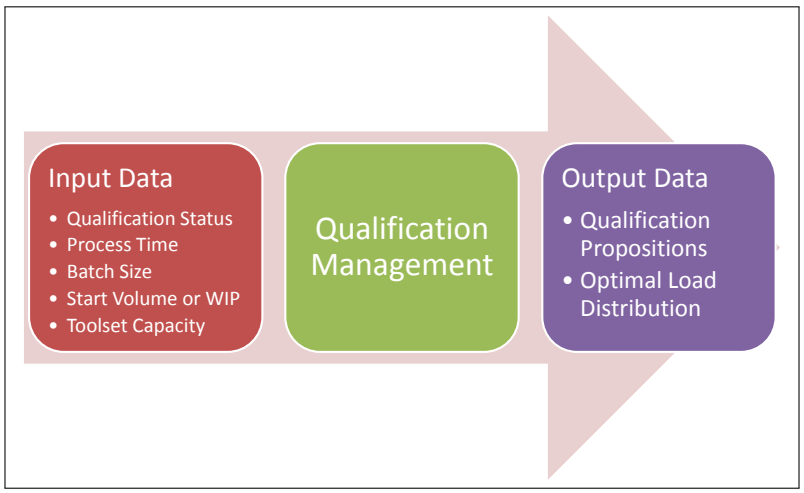


Figure 1.8: Les Flux de Données et d'Information

résultats de l'optimisation ont différentes utilisations. Les propositions des qualifications servent à choisir les meilleures qualifications pour assurer une utilisation optimale de la capacité des machines. En considérant le futur mix de production, l'équilibrage optimal des charges sert pour la planification de capacité. L'équilibrage optimal des charges en terme de volume de production sert comme guideline pour l'ordonnancement.

1.3.6 Conclusions et Perspectives

Dans la première partie de la thèse, nous avons étudié l'impact de la gestion des qualifications sur l'optimisation de l'utilisation de la capacité des équipements. Cette gestion apporte aussi de la flexibilité dans le système manufacturier. L'impact de la capacité limitée des équipements sur l'équilibrage de charge et par conséquent sur la gestion des qualifications a été étudié dans le Chapitre 4. Le Batching est une contrainte importante dans certains parcs d'équipement en industrie des semi-conducteurs. Lors de l'allocation des charges, cette contrainte doit être respectée. Plusieurs algorithmes d'équilibrage de charge sous contrainte de batch ont été introduits dans le Chapitre 5. De nouvelles qualifications apportent potentiellement plus de flexibilité au parc d'équipements. Plus de flexibilité réduit la variabilité des charges. Ce sujet est discuté dans le Chapitre 6. Les concepts de cette partie de

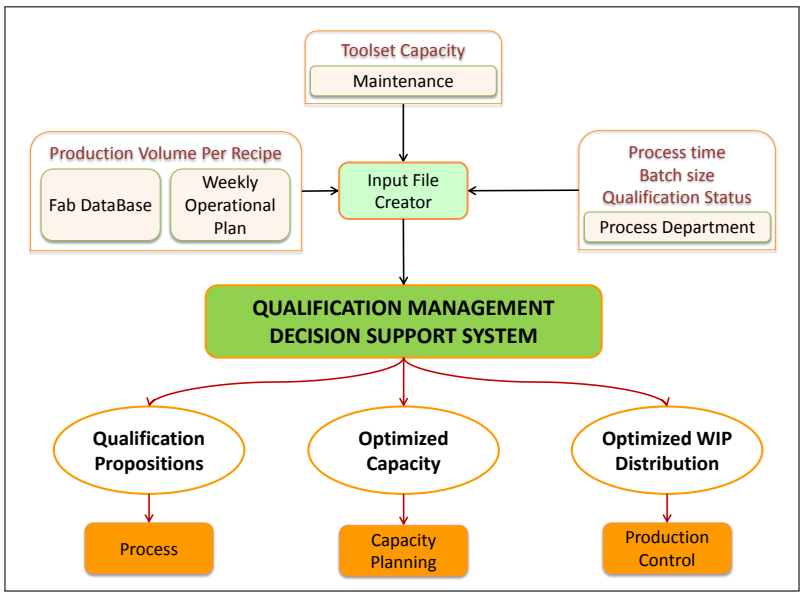


Figure 1.9: La Vue Globale du Processus de Prise de Décision, l'Application de la Gestion des Qualifications et ses Interfaces

thèse ont été mis en place au sein de l'entreprise Soitec. Le processus de prise de décision et d'industrialisation est le sujet du Chapitre 7. Plusieurs perspectives sont imaginables comme suite de ces travaux. Les coûts de qualifications ne sont pas considérés lors de la proposition de nouvelles qualifications. Aujourd'hui, ce sont les utilisateurs qui choisissent les meilleures qualifications selon les gains en terme de flexibilité tout en considérant les coûts et les difficultés associés. Donc, la considération du coût de qualification (en fonction du temps, du coût monétaire ou des difficultés), apporte une dimension de plus à la gestion des qualifications. Chaque parc d'équipements pourrait imposer des contraintes industrielles qui ne sont pas prises en compte dans ces travaux. Entre autres, on peut citer des ressources auxiliaires, des consommables (périssables ou non), la séquence de passage des lots sur les machines, le système de manutention, etc. Ces contraintes ont un impact sur l'allocation des charges et la gestion des qualifications.

1.4 Planification de Production en Boucle Fermée

1.4.1 Motivations et État de l'Art

Dans le Chapitre 9, on défini le cadre scientifique et industriel de la seconde partie de la thèse. Un des éléments de la gestion de la chaîne logistique concerne la *planification de production*. Elle fait référence aux décisions prises au niveau tactique dans l'objectif de déterminer les périodes et les quantités de production. La planification de production (autrement dit *dimensionnement des lots*) s'exprime par un modèle d'optimisation souvent avec l'objectif de minimisation des coûts (ou maximisation des profits) sujet aux contraintes. Les contraintes concernent la satisfaction de la demande à un taux de service souhaité, l'équilibrage des flux de production et les stocks, la capacité de production, la nomenclature, etc.

Le problème de dimensionnement des lots discuté dans cette seconde partie modélise le processus décrit dans la Section 1.2. Ce processus donne naissance à un système manufacturier avec deux lignes de production avec flux ré-entrants d'où l'expression *planification de production en boucle fermée*. Le dimensionnement des lots est fait sur la base d'un horizon discret avec de longues périodes et la nomenclature bi-niveau, un niveau concerne les matières premières et l'autre le produit fini.

1.4.2 Dimensionnement des Lots Multi-Produits Bi-Niveau à Capacité Finie avec Multiples Refabrications des Produits Dérivés Réutilisables

Dans le Chapitre 10, nous avons modélisé le processus de fabrication de plaques SOI en utilisant la Technologie Smart-Cut™ ainsi que le processus de "refresh" associé. Dans la suite, le problème est décrit rapidement sans rentrer dans le détail de la modélisation.

La Technologie Smart-Cut™ est décrite en Section 1.2. Les matières premières achetées sont les plaques de Fresh et de Base. En utilisant ces deux plaques, une plaque SOI est fabriquée. La plaque Fresh est aussi nommée le *Top*. La fabrication

de cette plaque SOI génère une plaque "négatif" (notée *Négatif 0* ou *Neg0*) qui est considérée comme un produit dérivé. Le Neg0 peut être recyclé (communément appelé "refreshé") pour revenir dans le cycle de fabrication de SOI. La plaque Neg0 recyclée est appelée *Refresh 1* ou *R1*. Le R1 peut être utilisé avec une autre plaque de type *Base* pour la production d'une autre plaques SOI. Ce processus de réutilisation des plaques de Top peut continuer jusqu'à une limite maximum qui dépend des produits considérés. Ce processus est illustré pour 7 niveaux de refresh dans la Figure 1.10. Pourtant le processus de refresh peut être réalisé en interne (Soitec) ou

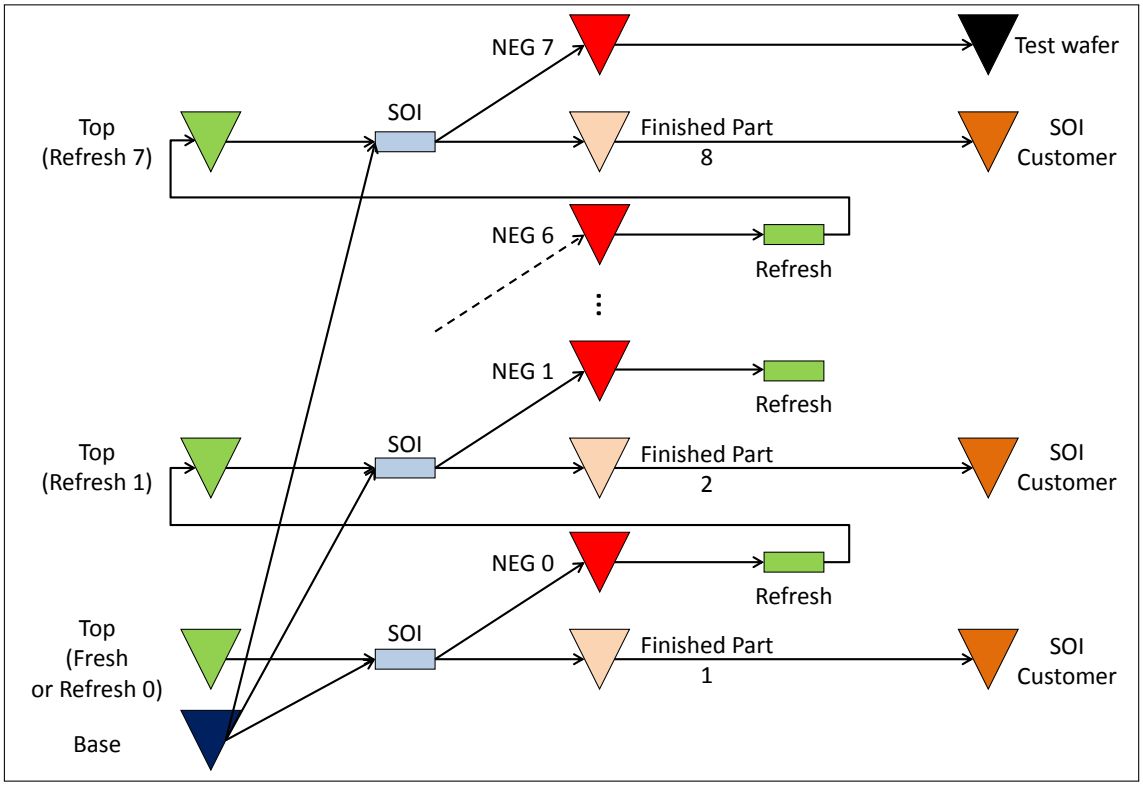


Figure 1.10: Fabrication de SOI avec un Exemple de Processus de Refresh jusqu'à 7 Niveaux ($l^{max} = 7$)

sous-traité en externe. La Figure 1.11 illustre le cas d'une ligne de refresh externe. Un modèle de planification de production est défini en Section 10.3 pour représenter le problème. Comme d'autres problèmes de dimensionnement des lots, l'objectif est de retrouver le meilleur compromis entre les coûts de stockage et ceux de lancement

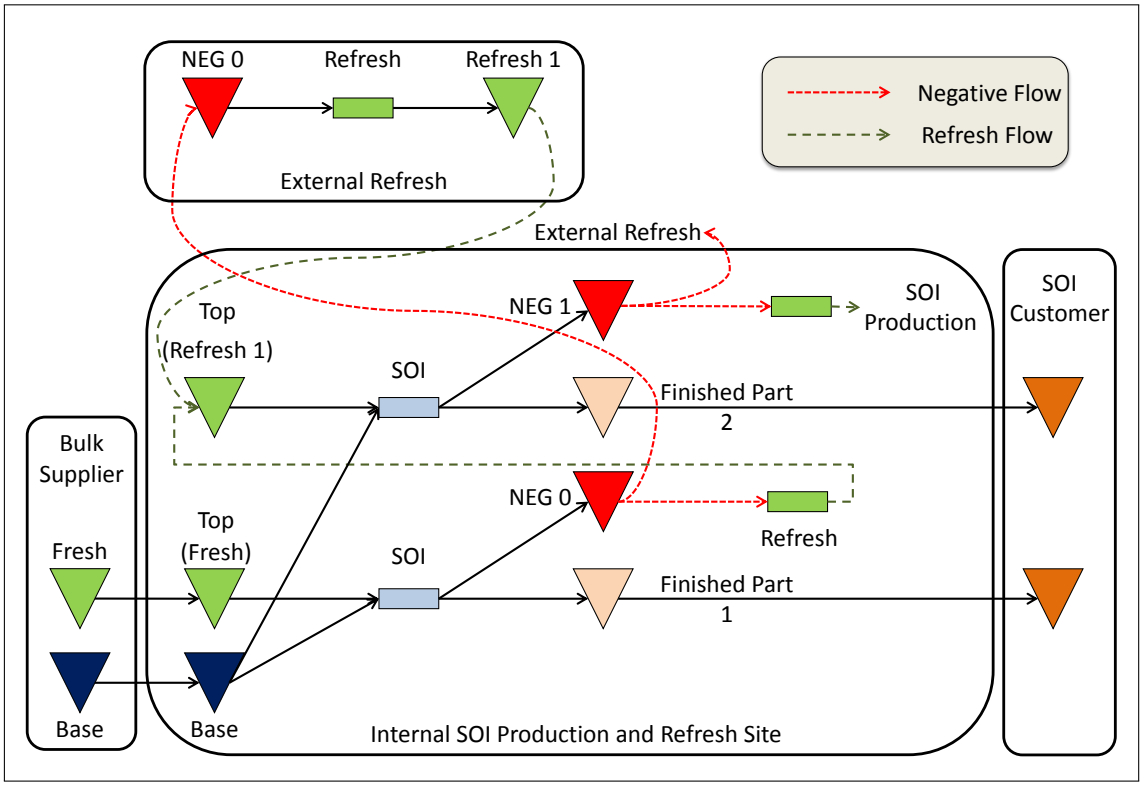


Figure 1.11: Une Simple Chaîne Logistique de Fabrication de SOI et de Processus de Refresh

de production et de commande. Les coûts de stockage concernent les matières premières achetées (les plaques de Fresh et de Base), les produits dérivés (les plaques négatifs), les plaques de Refresh et les produits finis (les plaques SOI). Les coûts de lancement concernent l'achat des matières premières (les plaques de Fresh et de Base), la production des produits finis (les plaques de SOI) et le processus de refresh. Les contraintes industrielles sont aussi considérées, par exemple: le rendement de production, l'aspect refresh hors-site, la capacité, etc.

Le modèle mathématique introduit est NP-difficile. Des jeux de données ont été construits à partir de données industrielles. Le logiciel CPLEX a été utilisé pour résoudre ce problème. Il se montre très performant et résout les plus grandes instances en moins de 10 minutes. En augmentant l'horizon de planification, l'importance du processus de refresh devient plus claire. En passant de 6 à 46 périodes, le coût

d'achat des plaques de Fresh diminue considérablement. Ce phénomène est dû au fait que le coût d'achat des plaques de Fresh est nettement inférieur à celui de refresh. Pour 6 périodes, le coût d'achat des plaques de Fresh et celui de processus de refresh occupent respectivement $18,63\%$ et $5,08\%$ du coût total du plan de production. En passant à 46 périodes, le pourcentage du coût d'achat des plaques de Fresh diminue fortement, passant à $2,72\%$. En contrepartie, le coût de refresh augmente légèrement, à savoir: $8,49\%$. Les expérimentations démontrent l'intérêt de la planification simultanée des deux processus de production de SOI et de refresh. Un autre point observé est que la production de plaques SOI ne sert pas seulement à satisfaire la demande mais aussi à générer des plaques de type négatif afin d'assurer la disponibilité des plaques de refresh pour des productions de SOI dans les périodes futures. Par ailleurs, la capacité de la ligne de refresh n'est pas seulement contrainte par elle-même mais aussi par celle de la ligne de SOI. Car c'est bien cette dernière qui assure la génération des plaques de type négatif utilisées ensuite dans la ligne refresh.

1.4.3 Dimensionnement de Lots Mono-Produit Bi-Niveaux à Capacité Infinie avec Multiples Refabrications des Produits Dérivés Réutilisables

En se basant sur le problème de planification de production du système manufacturier de la Technologie Smart-Cut™, un modèle de dimensionnement de lots mono-produit et à capacité infinie est défini dans le Chapitre 11. Ce modèle permet une étude académique plus approfondie sur des caractéristiques et des propriétés du problème. La maîtrise d'un tel modèle ouvre la voie pour proposer des heuristiques qui utilisent la décomposition et la relaxation Lagrangienne pour le cas multi-produits et à capacité finie.

La Figure 11.1 illustre le système manufacturier composé des lignes de fabrication SOI et refresh. Dans ce système, seule la gestion des plaques de Top et SOI est considérée. Le niveau de refresh est défini par l et il peut atteindre l^{max}. Le modèle mathématique comporte trois contraintes de lancement, à savoir celle de l'achat des plaques de Fresh, de processus de fabrication de produits finis (les plaques SOI) et

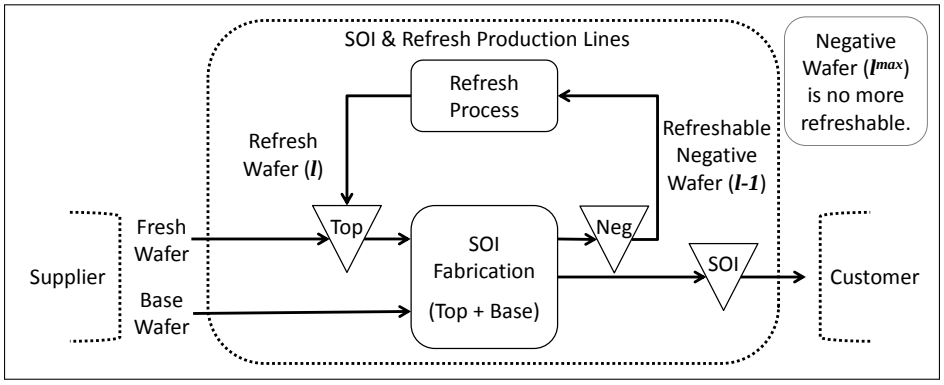


Figure 1.12: Le Système Manufacturier Simplifié de la Fabrication de SOI et du Processus de Refresh en Utilisant la Technologie Smart-Cut™

de processus de refresh. Nous avons démontré que le modèle est NP-difficile.

Afin de proposer une méthode de résolution, on a seulement considéré la contrainte de lancement d'achat de matières premières. Ensuite, on a défini la structure des coûts basée sur les réalités industrielles. Enfin, on a introduit des hypothèses de "Full Push" pour simplifier le modèle. Ces hypothèses assurent que les matières premières seront systématiquement transformées en produits finis. Ainsi, les produits dérivés sont recyclés immédiatement et ne seront disponibles qu'au début de la période suivante. Ces hypothèses entraînent la suppression de plusieurs stocks sauf celui des produits finis (voir Figure 1.13). On propose un algorithme de ré-

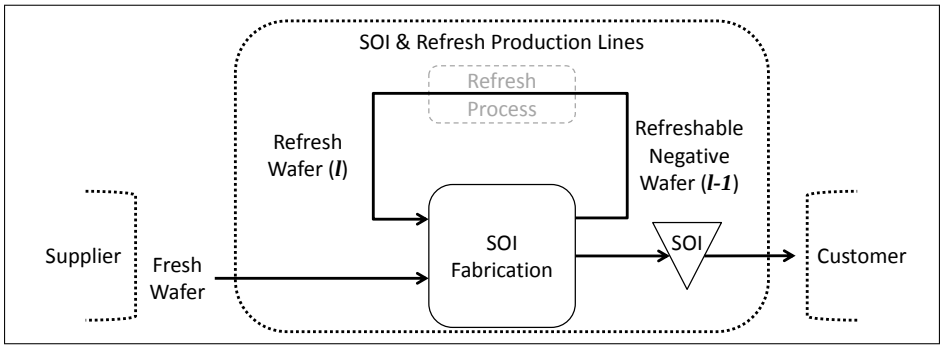


Figure 1.13: Le Schéma du Système Manufacturier SOI-Refresh Simplifié après les Hypothèses "Full Push"

solution basé sur la programmation dynamique. Malgré toutes ces simplifications,

l'algorithme reste exponentiel.

1.5 Conclusions et Perspectives Générales

Les deux parties de cette thèse sont complémentaires. Dans la première, nous avons considéré une contrainte particulièrement forte dans l'industrie des semi-conducteurs comme un levier de l'optimisation de la capacité. Dans la seconde partie, la planification de production de deux lignes de production en boucle fermée a été étudiée. La capacité est une des entrées d'un modèle de planification de production. Donc, comme perspective, ces deux idées peuvent être combinées dans un seul modèle. Dans un tel modèle, les nouvelles qualifications assurent la capacité nécessaire pour satisfaire la demande client lors de la planification de production. Un modèle a été proposé dans la Section 12.2.

General Introduction

This PhD project grew out of the collaboration between industry and academia. Soitec, a pioneer in building revolutionary semiconductor materials, is the world leader in producing high performance Silicon-On-Insulator (SOI) wafers. Soitec invests in research and development to make breakthroughs in advanced semiconductor materials. The company has specific production lines to satisfy the fabrication needs of the developed products. That is why, Soitec constantly seeks to improve its industrial practices to cope with market changes and satisfy demand. The industrial engineering department of Soitec decided to start the first PhD project of Soitec in industry by beginning a collaboration with Manufacturing Sciences & Logistics (SFL) department of École Nationale Supérieure des Mines de Saint-Étienne. The SFL department in CMP (Center of Microelectronics in Provence) has extensive experience in collaborating with industry, in particular with semiconductor manufacturing companies, such as STMicroelectronics. The French national research and technology agency (ANRT) has set up a program entitled *CIFRE* (standing for "Conventions Industrielles de Formation par la REcherche") to promote high quality research and development projects in the framework of a PhD project. This PhD project has received financial aid of ANRT. The main subject treated is *capacity and production planning*. Each subject is developed in a separate part.

The first part deals with capacity planning and optimization in the context of semiconductor manufacturing. We take a binding restriction, called *qualification*, present in semiconductor manufacturing as a lever for increasing and optimizing capacity utilization. The unstable business environment, and complex and extremely dynamic production environment require to adapt a *flexible* approach for *qualification management*. Interest in flexibility has grown rapidly in the last years both among practitioners and academicians. A large number of definitions has been given for the concept of flexibility. In Part I, we clarify what we mean by flexibility while showing its importance in capacity planning.

In each chapter of Part I, we discover different facets of qualification management and capacity planning by considering new restrictions. The successful industrial

implementation of this concept in Soitec was recognized internationally by receiving the Best Applied Paper Award during the Winter Simulation Conference 2013.

In the second part of the thesis (Part II), we are interested in production planning of two related production lines in Soitec. We tackle the industrial problem by proposing an innovative mathematical model. By taking the core challenge of the industrial case, we define and analyze an original academic problem.

The PhD project was launched with the subject of qualification management and capacity utilization. As the objectives of the industry were met, we started the second part of the thesis which is dedicated to a specific closed-loop manufacturing system.

Thesis Outline and Reading Plan

The manuscript is designed to be flexible. It is divided into an introductory chapter and two parts which are organized in a way to make each of them self-contained and independent of the other (see Figure 1.14). The introductory Chapter 2 describes the general industrial context of the whole study. Its purpose is to give an initial understanding of this topic to the readers who may not be familiar with semiconductor manufacturing (in particular SOI manufacturing using the Smart-Cut™ Technology).

Part I (Chapters 3 to 8) discusses capacity optimization and planning through flexible qualification management. In Chapter 3, the qualification management and its impact on capacity planning are discussed. Chapter 4 studies qualification management while considering the limited capacity of the machines. Batching is an important characteristic for some machines in the industry in general and in semiconductor manufacturing in particular. The impact of batching on qualification management is discussed in Chapter 5. Chapter 6 discusses the relationship of qualification management and production variability. The introduced concepts are industrialized. The industrialization is explained in Chapter 7. Extended conclusions and perspectives of Part I are presented in Chapter 8.

In Part II, the focus is on the production planning for SOI fabrication and refresh process lines in Soitec. The described production process in Section 2.3 is

exclusive to Soitec. Chapter 9 sets the industrial and scientific context of this part. Chapter 10 presents the production planning model with all industrial constraints. Chapter 11 presents and analyzes several closed-loop production problems in which raw materials may be reused several times after remanufacturing. Both parts of the dissertation are related. We do not study this link in this thesis. However, some suggestions are presented at the end of the thesis in Chapter 12.

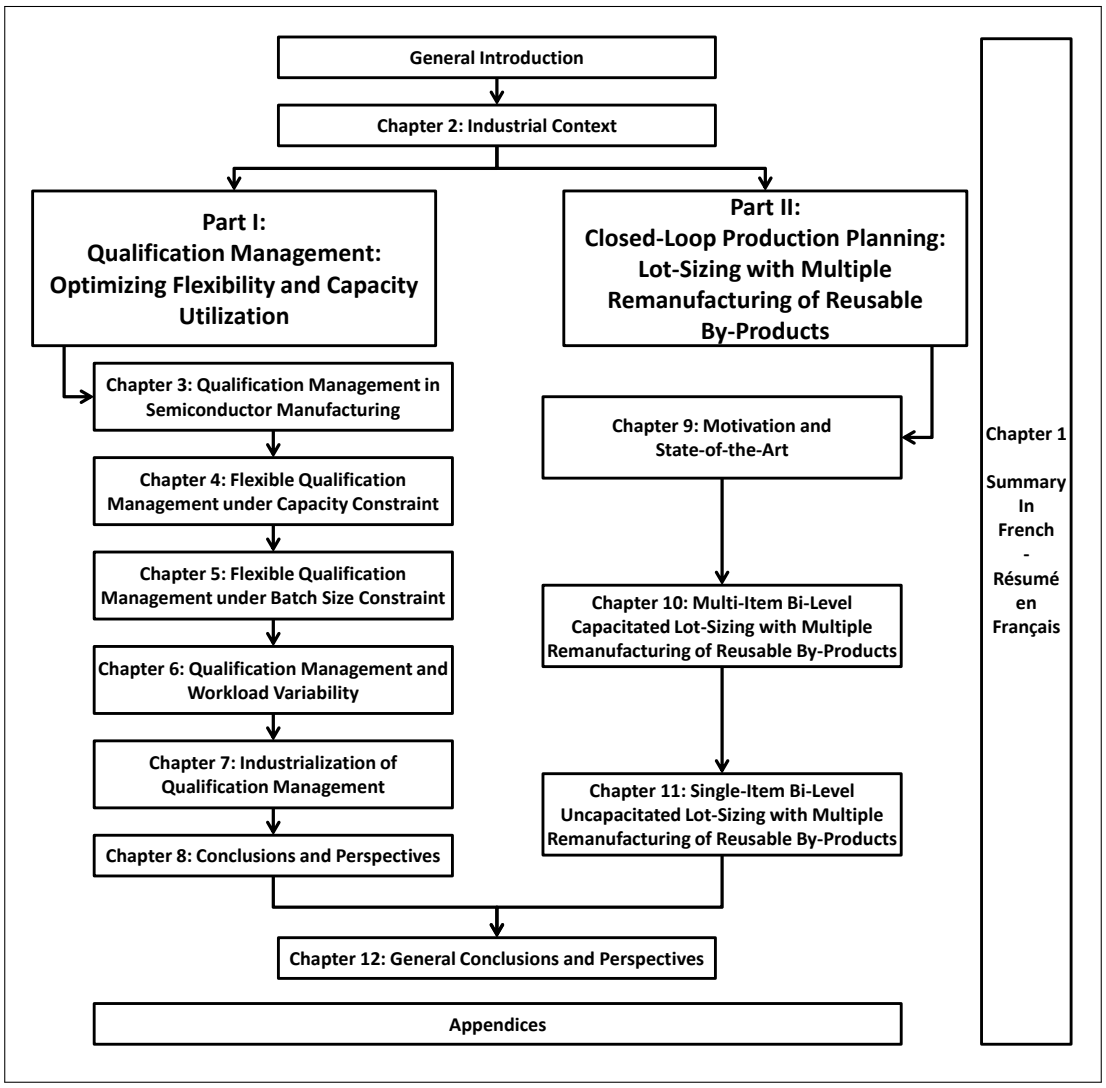


Figure 1.14: Structure of the Thesis and the Precedence Relationship between Chapters

Chapter 2

Industrial Context

This chapter introduces the industrial context of both parts of the thesis. Almost all aspects of the first part concerning capacity planning through flexible qualification management can be applied to most semiconductor fabrication facilities. The focus of the second part is on production planning of a Silicon-On-Insulator (SOI) wafer fabrication supply chain using the specific Smart-Cut™ Technology. The technology described briefly in this chapter is exclusive to the company Soitec.

2.1 Introduction

The history of semiconductor manufacturing dates back to the invention of field-effect transistor (FET) by the physicist Julius Edgar Lilienfeld in 1925 in Canada. At that time, it did not find practical use because of the lack of high-quality semiconductor material. However, the electronics industry was born with the invention of vacuum tubes in the early 19s and devices such as radios. Transistors rapidly replaced vacuum tubes as the latter consume a lot of power while producing heat. They are fragile, relatively big and also wear over time. Transistors are made up of semiconductor material. Semiconductor material is a substance which is conductor under some conditions else an insulator. The electrical conductivity of a semiconductor changes with the variation of voltage or current of the control electrode. Several elements and compounds show semiconductor characteristics. However, most of the semiconductor materials are built on silicon. The small size of the transistors has revolutionized the world. The electronics revolution has been following the Moore's law which states that the number of components in an integrated circuit doubles every 18 to 24 months. The goal of the semiconductor industry has been to keep up with this pace of evolution with the creed "smaller, faster and cheaper". To follow this slogan, constant research is conducted on semiconductor components material, design and production process. While the manufacturing processes become more complex, the cycle time and production cost must decrease, which call for a better capacity utilization of equipment and labor. Besides, the life cycle of the products shortens while their diversity increases. Therefore, the production process must be efficient, cost competitive, flexible and agile. In the remainder of this chapter, we paint a global picture of semiconductor manufacturing with an emphasis on SOI manufacturing in Section 2.2. Section 2.3 focuses on the main theme of this dissertation, i.e. production and capacity planning.

2.2 Semiconductor Manufacturing

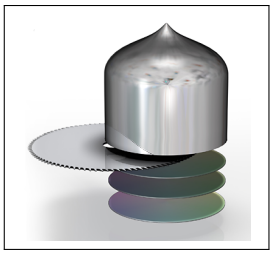
Semiconductor components are made of semiconductor elements or compounds; the most used semiconductor substance is *silicon*. Silicon (Si) can be found by mass

in dusts, sands, planetoids and planets under the form of silicon dioxide (silica) or silicates (Figure 2.1a). However, very clean and good form sand is used for making silicon. The sand (SiO_2) is heated just above its melting point to obtain silicon (Si). After several chemical processes, blocs of polycrystalline silicon are obtained (2.1b). Monocrystalline Silicon (single-crystal silicon) ingots (Figure 2.1c), produced with high purity, are used in the fabrications of semiconductor components. They are grown mostly using the Czochralski process. Silicon wafers are obtained using mechanical-chemical procedure from the silicon ingot. Firstly, the silicon ingot is ground to the desired diameter. Then a notch or a flat edge is given to the ingot serving as the future wafer orienting guide. The ingot is sliced into silicon wafers (Figure 2.1d). Several mechanical and chemical steps guarantee the flatness, thickness and smoothness of the wafers. The silicon wafers can then be used for making

(a) Sand made up of about 25% silicon

(b) Rod and lumpy polycrystalline silicon

(c) Silicon ingot



(d) Wafer slicing

Figure 2.1: From Sand to Silicon Wafers

semiconductor components such as chips and ICs (integrated circuits). However, some devices require higher performance and specific characteristics which cannot be obtained by traditional silicon wafers. That is where Silicon-on-Insulator wafers are used. In the next section, we discuss more about SOI wafers and specially one of the most interesting techniques available to make them, the Smart-Cut™ Technology.

Semiconductor components made of silicon are used in most electronic devices. The semiconductor components are etched on silicon or SOI wafers. As even a tiny bit of dust can damage an integrated circuit or a chip, the whole semiconductor manufacturing process is done in so-called clean rooms. A clean room is upto $10,000$ times cleaner than an operating theater. The air inside the clean room is continu-

ously filtered and cleaned. Operators, technicians and anyone who enters the clean room must wear special clothing to keep the area free from particles. The clean room garments include face mask, hood, coverall, shoe cover, glasses and gloves.

The whole process of semiconductor manufacturing can be divided into two sequential sub-processes: *front-end* and *back-end* (see Figure 2.2). The front-end concerns all fabrication steps for the creation of circuits on a blank wafer. The back-end begins when the chips and all circuits are created on the wafer. It concerns testing, dicing, wiring, assembly and packaging.

The front-end or wafer fabrication produces silicon chips on a blank wafer (typically silicon). A chip is a set of electronic circuits manufactured in layers. A microchip may contain over a billion transistors. *Photomasks* are used to optically transfer patterns to wafers. The *photolithography process* transfers the geometric pattern of the photomask on the wafer. After a series of chemical treatments, the desired patterns are engraved on the wafer and the undesired exposed material is etched away. The photolithography is a very precise process which contains several re-entrant steps. At the end of the front-end process, the proper functioning of the chip is electronically tested.

The first step of the back-end operation is *die preparation* in which the wafer is sawed into individual chips or dices. Then each *dice* is *attached* to the metallic support structure of the package (e.g. the leadframe) by means of an alloy or an adhesive. Then the leads on the leadframe are *wire bonded* to the input/output electrical terminals of the package. Packaging or encapsulation is the final step of the semiconductor manufacturing process in which the chip is encapsulated in a plastic or ceramic mold. The package protects the chip from corrosion and physical damage. Before shipping the integrated circuits to the customers, they are tested electronically.

2.3 Silicon-On-Insulator (SOI) Wafer Fabrication

Semiconductor components have revolutionized our daily life. Silicon wafers are extensively used in semiconductor manufacturing to produce microelectronic components such as integrated circuits and chips. However, some devices require higher

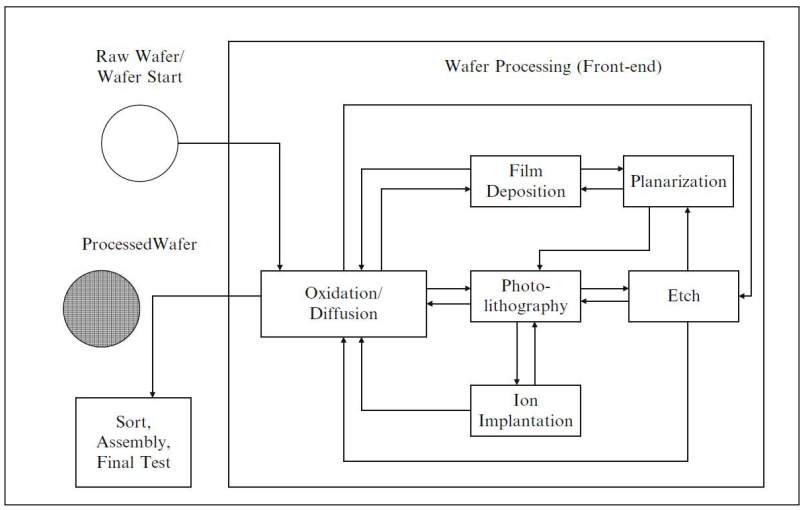


Figure 2.2: Front-End and Back-End Process ([70])

performance which cannot be delivered by traditional silicon-only wafers. Components built on *Silicon-On-Insulator* (*SOI*) wafers offer much more performance while consuming less energy compared to components on silicon-only wafers. Semiconductor components fabricated on SOI wafers are found in RF devices, imaging systems, microprocessors, automobiles, etc. (see Figure 2.3). SOI refers to a layered

Figure 2.3: Semiconductor Manufacturing and the Daily Life

silicon on insulator substrate, in which the insulator is mostly silicon dioxide (SiO_2). SOI is widely used in the microelectronics industry such as radiofrequency devices, microelectro-mechanical systems (MEMS), photonics and biotechnological chips.

An SOI wafer is composed of three layers (see Figure 2.4):

- The active silicon layer called **Top**;
- The buried Silicon Oxide layer called **BOX**;
- The monocrystalline silicon which serves as mechanical support called **Base**.

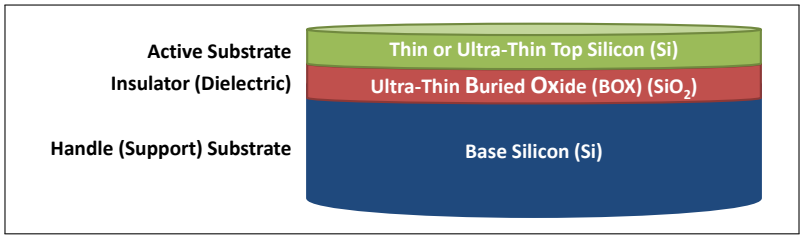


Figure 2.4: Silicon-On-Insulator

The thickness of the active silicon layer (Top) varies depending upon the SOI application between $0.01\mu m$ and $1.5\mu m$. The thickness of the dielectric insulator (BOX) varies between $0.05\mu m$ and $3\mu m$. The thickness of the silicon substrate (Base) varies between $375\mu m$ and $775\mu m$ based on the wafer diameter.

The main advantage of SOI wafers compared to silicon bulk wafers lies in the buried oxide (BOX). In a MOSFET, only the superficial layer of the silicon (0.006 to $0.1\mu m$) is used for the transport of carriers. The rest of the substrate which represents about 99.99% of its thickness, causes current leakage and undesirable parasite effects. SOI not only eliminates the parasite effects but also decreases the current consumption [25].

Circuits based on SOI technology are more compact, faster, more power-efficient and more heat-resistant compared to traditional silicon wafers. SOI technology is challenging more and more the traditional bulk silicon wafers. For instance, more than 60% of mobile devices and 80% of game consoles produced in 2012 use SOI chips [66].

The main advantages of the SOI technology are [66]:

- High performance, both in speed and power in comparison to traditional silicon bulk;
- Smaller chip area due to better scaling; and
- CMOS process simplification.

A group of important companies in the microelectronics industry have formed the SOI Industry Consortium to promote the benefits of the SOI technology in the market and enhance its usage.

SOI wafers may be produced using different technologies such as *SIMOX™* (Separation by IMplantation of OXygen), *wafer bonding* or *Seed methods*. The production process studied in this thesis considers a type of wafer bonding technology called the *Smart-Cut™ Technology*.

Smart-Cut™ Technology

Smart-Cut™ Technology has been developed since 1993 to obtain SOI materials [15]. The Smart-Cut™ Technology is based on direct bonding of two wafers. One of the wafers is implanted by light gas ions such as hydrogen. At the mean depth of the ion penetration, a weakened zone is formed. Once the implanted wafer is bonded with the second wafer, the splitting occurs at the implanted zone. Therefore, a thin silicon layer is transferred from the implanted wafer to the second wafer. The second wafer acts only as a mechanical support. The semiconductor components are built on the transferred thin silicon layer.

The Smart-Cut™ Technology steps used to produce SOI Wafers are as follows [66] (see Figure 2.5):

- The silicon Wafer A is oxidized thermally. The ***B**uried **Ox**ide* (BOX) is now formed;
- The oxidized Wafer A is implanted by ions of Hydrogen creating micro-cavities. The micro-cavities constitute the weakened zone. The weakening is defined by the energy and the implantation dose. The implantation energy determines the hydrogen ion penetration depth through the BOX;

- The implanted Wafer A is cleaned and bonded to Wafer B at room temperature. The cleaning eliminates the particles from the surface of the wafers. Moreover, it covers the wafers with radicals of OH which facilitate the bonding. The very plane surface of the wafers allow molecular adhesion of two wafers;

- The bonded Wafers A and B are split, leaving a thin layer of Wafer A on Wafer B. The splitting takes place in an oven. By increasing the temperature, the pressure of the Hydrogen (H_2) in the micro cavities cause horizontal fracture at the weakened zone;

- Finally, using thermal annealing, the bonding surface is consolidated. The roughness left on the SOI Wafer is removed using mechanical-chemical process.

The silicon bulk which remains after the splitting can be reworked again to be used as another Wafer A. The process of reworking the used Wafer A is called the *refresh process*. The second Part of this dissertation (Chapters 9 to 11) is dedicated to production planning of the SOI Production and Refresh Process. One of the most economical advantages of the Smart-Cut™ Technology is the possibility of reusing Wafer A. The recycling (or "refresh" process) contains several steps. First the oxide layer of the *used Top Wafer* is removed. Then the edge is polished. Using Double Side Polishing (DSP), the wafer is polished roughly. After cleaning, the Chemical-Mechanical Planarization (CMP) process smooths the wafer which is ready for the final cleaning. Once cleaned, the wafer is graded. If the wafer fits the specifications, it is used to make SOI wafers. If it fails the inspection, depending on the status, it returns either to the refresh line for a second try or it is discarded or used as a test wafer. A detailed modeling of the refresh process line can be found in Appendix B.

As Wafer A is recycled several times, its quality must be excellent. While the quality of Wafer B is not critical. It serves only as a mechanical support (from which the name *handle wafer* is derived). Because of the low cost of Wafer B and that Wafer A can be recycled multiple times, the Smart-Cut™ Technology is cost competitive. It is almost similar to a mono-wafer technology.

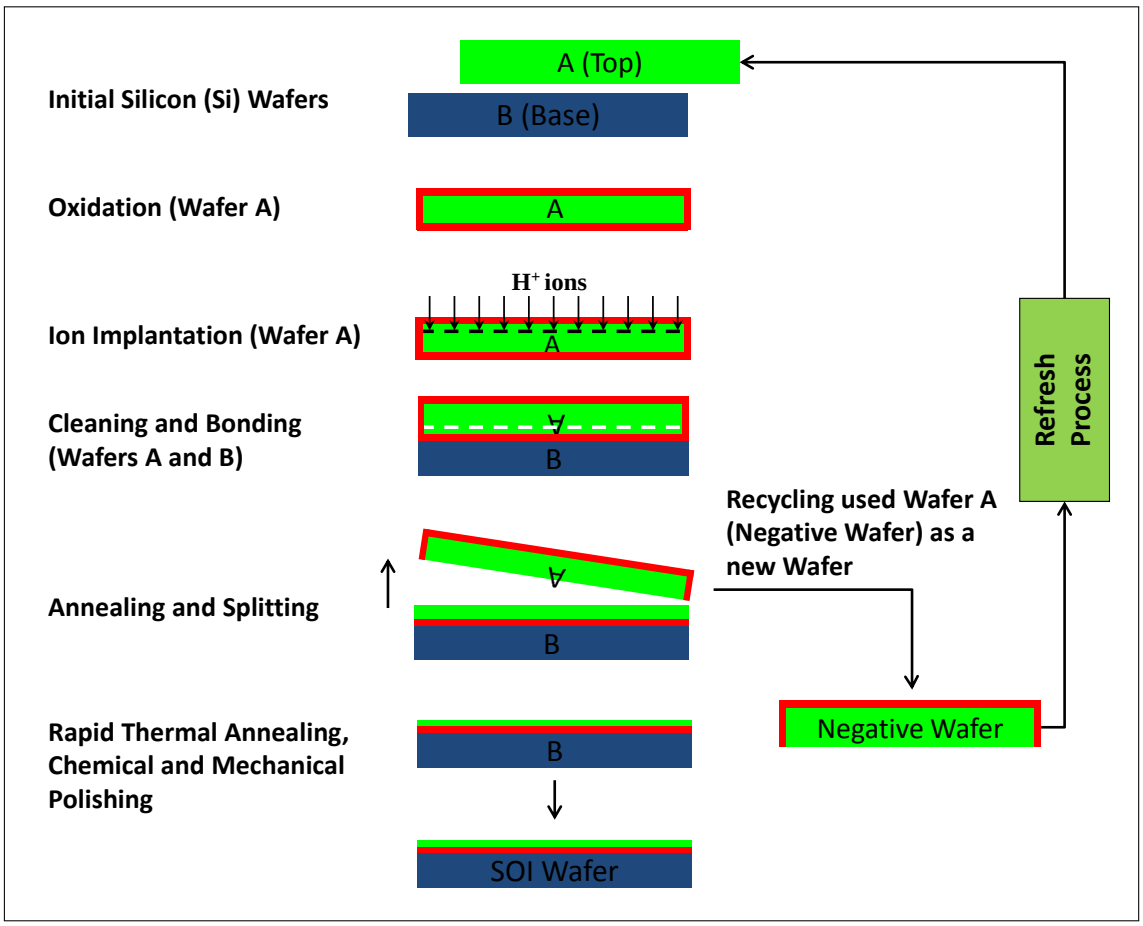


Figure 2.5: Smart-Cut™ Technology Used to Produce SOI Wafers

Among other Smart-Cut™ Technology advantages, we can mention the high quality and the thickness homogeneity of the transferred active layer. Furthermore, the Smart-Cut™ Technology can be adapted to transfer single-crystal substrates on different supports [66]. SOI wafers are produced in *clean rooms* like other semiconductor components.

2.4 Production and Capacity Planning in Semiconductor Manufacturing

The goal of any production system is to transform raw materials into finished products by adding more value to the initial raw material at each production step. This is also the case in semiconductor manufacturing with the difference that its production and business environment are very complex. The production complexity comes from the sophisticated nature of the products and associated manufacturing systems. The relatively high product and production cost and the competitive business environment call for an efficient and effective production planning system.

The objective of a production system is to satisfy the customer demand over a time horizon in the most efficient and effective way while considering multiple constraints and optimizing different production indicators. Some production measures which are to be optimized are cycle time, (work-in-process) WIP levels and machine capacity utilization.

The second part of this study deals with the production planning of the SOI and refresh production lines. Special constraints associated with the SOI fabrication using the Smart-Cut™ Technology are analyzed and considered in the problem modeling. In particular, the refresh process, which is one of the exclusive features of SOI manufacturing using the Smart-Cut™ process is detailed.

The capacity planning is often categorized into long-, medium- and short-term horizons. The time dimension differs from industry to industry [97].

The focus in the first part (Chapters 3 to 7) of this thesis is on medium- to short-term horizon capacity planning which involve more operational decisions on a monthly to weekly basis.

The short-term capacity planning determines the expected machine utilization of (a single machine or a toolset) for some determined production plan or loading [70]. This machine utilization is used to see if the capacity meets the production plan. If not, either the capacity must be increased or the production plan must be revised. The interested reader can refer to [70] for more literature on subjects related to production planning in semiconductor manufacturing.

2.5 Conclusion

In this chapter, we have set the stage for the remainder of the dissertation. The importance and essential role of semiconductor manufacturing has been pointed out. This study is conducted in Soitec, a company which produces high performance wafers used for high demanding applications. Rapid expansion of the semiconductor industry and introduction of new technologies have created high-mix low-volume production lines. In order to cope with rapid changes, to lower the cycle time and production costs, efficient and flexible capacity and production planning is required.

Part I

Qualification Management: Optimizing Flexibility and Capacity Utilization

Chapter 3

Qualification Management in Semiconductor Manufacturing

Semiconductor manufacturing is one of the most complex industries in the world. The increased diversity of products and short product life cycles call for re-configurable machines. Qualification is a kind of setup which gives a piece of equipment new capability. Nevertheless, qualifications consume time and energy. All of these constraints in the dynamic fab environment suggest an efficient and effective qualification management. This chapter opens up the discussion about flexible qualification management and capacity planning in semiconductor manufacturing.

3.1 Introduction

The semiconductor industry is one of the most complex and modern industries in the world. Expensive manufacturing equipment and the ever-growing competition call for an optimal capacity utilization of the fabrication facilities (called "fabs"). In order to cope with the fast-changing business environment, the production system must be flexible enough to produce a wide range of products. As the life cycle time of products shortens more and more, production systems must be agile to rapidly manufacture new products or adapt to product mix changes.

In semiconductor manufacturing, wafers undergo operations at workstations called "toolset". Each toolset is a collection of nonidentical multi-purpose parallel machines that are reconfigurable. In order to perform an operation, a *recipe* must be executed on the product. A recipe is the machine instructions to obtain the desired process. In order to perform an operation on a product, its corresponding recipe must be *qualified* on the machine. However, due to multiple hardware and software restrictions, maintenance or retrofit costs, it is not possible to *qualify* all recipes on every machine.

Qualification is one of the characteristics of the semiconductor manufacturing. It is a kind of setup with the difference is that a qualification is performed once and not before each production run. And the qualified machine remains qualified for that recipe until a disqualification occurs. Based on the toolset, several reasons may cause recipe disqualification on a machine. Machine breakdown or maintenance may require the re-qualification of all previously qualified recipes. Not running a recipe for a long time on a machine may also lead to *automatic disqualification*.

Qualifications have a direct impact on production capacity. If a recipe is not qualified on a machine, it is not possible to allocate the production volume of this recipe to the machine. The impact of the recipe-to-machine qualification configuration in a toolset is discussed in 3.3.

Besides, qualifying a recipe on a machine can be very time- and energy-consuming. Test products must be used for test runs. During test runs, the machines are under scheduled downtime status, therefore in a non-productive status. Metrology and defect inspection resources must also be extensively used. Hence, it is not economi-

cally wise to perform a great number of qualifications. If a poor recipe-to-machine configuration results in loss of capacity, too many qualifications (and maintaining them) are very costly and also cause scheduled downtimes. These two contradictory constraints call for an efficient *qualification management* policy. Qualifications are also one of the sources of variability. Variability often lead to the increase of the buffer stocks in the fab. Variability and hence stock and cycle times are reduced by a better qualification management. This leads to a smoother product flow in the production line and cycle time reduction.

3.2 Qualification in Semiconductor Manufacturing

In the course of production, each wafer undergoes *operations* at each *workstation* (named as "toolset"). A *toolset* is a collection of parallel and mostly nonidentical machines (or "tools") performing similar operations. For instance, a collection of furnaces, with eventually different specifications (throughput, batch size, hardware and software) constitutes the "Thermal Treatment Toolset" or a collection of "furnaces" constitutes the "Thermal Treatment Toolset". Each *operation* (which corresponds to a production step) is associated with a *recipe* which corresponds to the machine instructions to obtain the desired process. A *thermal treatment recipe* may specify the ramp up and ramp down temperature values and duration, cool-down rate and gas pressure at each phase. Or each "implantation recipe" defines the implantation energy and duration besides other technical specifications.

In order to perform an operation on a machine, its recipe must be *qualified* beforehand on the machine. Due to machine hardware and software restrictions, all operations cannot be performed on all machines. In other words, all recipes cannot be qualified on every machine of the toolset. Therefore, in order to qualify a recipe on a machine, several setups (both software and hardware) and tests may be needed. A qualification can be relatively quick to perform (e.g. equivalent to a production run), but also very time-consuming (e.g. equivalent to a production cycle time) and hence costly.

The ideal configuration would be that all recipes are qualified on all machines, so that we can freely allocate the production volume of any recipe to any machine. However, due to process restrictions, some of the recipe-to-machine qualifications are not authorized. Qualifying "qualifiable recipes" on machines is costly and time consuming, causing scheduled downtimes. Therefore, in practice only a few number of qualifications can be performed.

In order to reduce qualification costs and to avoid performing unnecessary qualifications, an efficient qualification strategy must be set up. In [55], the importance of qualification management in wafer fabs is pointed out. As already stated, a poor qualification configuration may result in an unbalanced toolset making production planning and also scheduling difficult. Although, if this were the case (i.e. all recipes were authorized to be qualified on all machines), it would be too costly to qualify all recipes on all machines. The recipes are sorted into three categories: *Unauthorized* (or *unqualifiable* or *non-qualifiable*), *authorized* (or *qualifiable*) and *qualified*. This qualification classification links the recipes and the machines. In this thesis, without loss of generality, the recipes are considered to be either "(already) qualified" or "non-qualifiable" on machines. In Chapter 7, we explain the industrialization of all cases. The recipes however may also be sorted into other categories: *Normal recipes* which are qualified at least on one machine and *Not previously Qualified Recipes* (termed hereafter *NQR*) which are mostly new recipes corresponding to new products which have not been previously qualified on any machine. The case of *not previously qualified recipes* may occur when a new product is launched or when a recipe has been disqualified from all machines. For some toolsets, such as "thermal treatment", disqualifications may occur automatically due to not processing a recipe during a certain time window on the machine. However, disqualifications may be caused after changing the hardware, corrective maintenance operation, (rarely) uninstalling some software, etc.

3.3 Qualification Management and Capacity Planning

One of the main management challenges is to match capacity constraints to forecast plans either by increasing capacity (or capacity utilization) or by adapting the plans to the capacity constraints by delaying the delivery date or canceling the orders. Other possible courses of action to match capacity and forecast plans are subcontracting, purchasing manufactured parts instead of raw materials and choosing possible alternate routing ([97] and [89]).

In a production line or a workcenter, the capacity is the output rate, i.e. the the processing work amount per time. The capacity must be optimized through workload balancing according to available workforce and equipment productive time and capabilities [64].

The recipe-to-machine qualification configuration has a direct impact on the capacity utilization of the toolset. If the recipe of a product is not qualified on a machine, it is not possible to allocate quantities of the product to that machine. Therefore, a poor qualification configuration causes loss of capacity. The effect is intensified for a bottleneck toolset. Qualification constraints are like machine eligibility restrictions (also called machine dedications). However, as the machines can be qualified or disqualified, they can be considered as *re-configurable*. Due to product mix, dynamic fab environment and toolset limited capacity, this may lead to backlog or unsatisfied demand. Therefore, an adequate recipe-to-machine qualification configuration is necessary for the smooth running of the fab [55].

Generally speaking, capacity planning (or capacity allocation) is seen as critical in semiconductor manufacturing and is still investigated in the literature (see for instance [63], [18] and [105]). Equipment qualification is considered in capacity planning in semiconductor manufacturing. For instance, in [78], the authors calculate the required inspection capacity in a dynamic sampling strategy. However, as the inspection machines are only qualified for some inspection process operations (inspection recipes), the machine qualification is considered in the sampling strategy to optimize the capacity allocation.

3.4 Flexibility in Semiconductor Manufacturing

Today's economy is marked by globalization and as a result increased competition. The technological breakthrough has shortened the product life cycle requiring the companies very short manufacturing response time. Therefore, flexible production systems are required to quickly respond to market fluctuation, increase the production agility and answer the diverse market demand.

Since 1980s flexibility in manufacturing decision making has become more into focus. *Dedicated machinery* and *non integrated general purpose machine tools* are two traditional forms of manufacturing systems [33]. The first system allows mass production of very limited products (or mostly a single product) with almost no flexibility. The second system permits very small batch production of a variety of products. The flexibility of this type of production system is very high but it is usually not suitable for large production volumes. One other disadvantage of such a system is the high production cost per unit and required highly qualified operators.

Nowadays, most production systems are computer-driven, making them flexible and more capable of producing large volumes of multiple products at a low unit cost. Flexible Manufacturing Systems (FMS), Computer Aided Design (CAD) or Computer Aided Manufacturing (CAM) systems are supporting the development of this third type of manufacturing systems [45].

According to [17], flexibility is defined as the ability of a manufacturing system to cope with changing circumstances. It is one of the main objectives and critical measures of any manufacturing system [45].

Manufacturing flexibility (MF) can be categorized from different points of view, such as the time horizon, software or hardware flexibility, etc. Flexibility is very important in semiconductor manufacturing due to high investment, labor, facilities and equipment costs.

The investigations conducted in this part of the thesis, is an effort, in continuation of [53], to quantitatively measure manufacturing flexibility in a workcenter (toolset) in semiconductor industry. Several types of manufacturing flexibilities are defined in [14]: Machine Flexibility, Process Flexibility, Product Flexibility, Routing Flexibility, Volume Flexibility, Expansion Flexibility, Process Sequence Flexibility

and Production Flexibility.

The discussed flexibility measures in [53] do not fit exactly one single category. The ultimate objective of each flexibility measure is defined in its place.

A comprehensive literature synthesis on flexibility is available in [45]. Flexible manufacturing system features have also been taken into account while production planning and scheduling [34].

In this thesis, flexibility is to gain very much with very little (or even no) effort. The criterion that we consider, is the levelness of the workload in a toolset. By reaching full flexibility, we mean that the toolset is well-balanced, i.e. all machines are equally loaded, according to their capacity of course.

3.5 Literature Review

In the following chapters, flexible qualification management is considered from different aspects. We try to gather the related literature in this section.

Flexibility

For many years, manufacturing flexibility has drawn interest from researchers and practitioners as a key factor of competitivity in dynamic and uncertain environments. Operations flexibility is defined in [85] as the assignment of production tasks to workcenters, assuming that the number of tasks are larger than the number of workcenters and that the workcenters are able to perform all tasks. Benefits of the defined operations flexibility for a flowshop environment with the objective of minimizing the completion time of all jobs and also maximizing workcenter utilization is studied further. The solution approach is based on a two-phase heuristic: Job to workcenter allocation and job sequencing.

Early studies consider manufacturing flexibility between plants and products, with various capacity limitations and demands [10]. However, in this study, we focus on manufacturing flexibility and capacity utilization at workcenter level, i.e. to which extent the flexibility of recipe-to-machine assignments affect toolset capacity utilization.

Qualification Management

Recipe-to-machine qualification management has come into attention in recent years due to its importance in the semiconductor industry. Here, we discuss the main studies on the subject. Recipe-to-machine qualification configuration is studied as a configuration problem for a parallel multi-purpose machines workshop in [3]. The study takes two features into account: Demand uncertainty and qualification cost. To hedge against demand uncertainty, the recipe-to-machine qualification configuration must be robust while at the same time, the qualification cost must be minimized. A bi-objective optimization model is proposed which maximizes the robustness level with a certain fixed qualification cost, i.e. a fixed number of new qualifications. Then the model is solved with a branch-and-bound method.

Finding the qualification configuration at minimum cost in order to balance the workload on the toolset while meeting demand, termed as *load-balanced production plan* is presented in [5]. They present a mixed integer linear program and they show that the problem is NP-hard in the strong sense. Modeled as a transportation problem, lower and upper bounds on the setup costs for uniform parallel machines with identical setup times are estimated. The recipe-to-machine qualification configuration of a parallel multi-purpose machine is studied in [4]. Models and resolution approaches are proposed to find the best compromise between the number of qualifications (cost) and workload balance robustness. It is argued that as the demand variability is high in forecasts, the deviation from the forecast plan (robustness) must be maximized while maintaining a load-balanced production plan without exceeding a given total setup.

New indicators, called *Flexibility Measures* are proposed in [54] and [53] to estimate the flexibility of recipe-to-machine qualification configuration of the whole toolset depending upon two different objectives. Several extensions of the measures and optimization approaches for multiple qualifications are also proposed and validated in [53]. We continue these studies. Therefore, the measures are recalled in Section 4.2.1 in detail.

An optimization model with binary variables is proposed for a photolithography toolset in [51] with the objective of balancing the workload on all the machines. To do this, for each new qualification, the qualification binary variables are set to 1. In

the model, it is possible to define a minimum and/or a maximum number of allowed qualifications.

Discrete-event simulation has been used in [58] to investigate the impact of recipe-to-machine qualification configuration and production start volume on the workload of each machine in a toolset. Different simulations are performed with different production start volumes and recipe-to-machine qualification configurations. Using the results of these simulations, the workload associated with recipes with only one qualification is compared to the workload of recipes qualified on several machine. For instance, they show that the overall workload is higher when each recipe has only one qualified machine. It is also stated that the immature products are more likely to have only one or few machines qualified compared to high-volume mature products.

Batching

Batching is an important industrial constraint which has significantly be studied in the scientific literature, especially in the context of semiconductor manufacturing. In Chapter 5, in order to evaluate the recipe-to-machine qualification configuration, we solve a problem which is close to a scheduling problem on nonidentical unrelated parallel batch machines where splitting recipes is allowed. However, our objectives are quite different from the ones considered in the literature. Some relevant papers on batch machine scheduling are cited below.

In the semiconductor industry context, [44] propose an algorithm to minimize three due-date objectives (average tardiness, maximum tardiness, and number of tardy jobs in batch processes) simultaneously in scheduling a *single* batch processing machine. [21] propose an efficient genetic algorithm to minimize the makespan for a *single* batch-processing machine with nonidentical job sizes. [104] solve the flexible job-shop scheduling problem on nonindentical batching machines with a metaheuristic based on a disjunctive graph representation and for different objectives functions. [23] study the scheduling problem of minimizing the weighted and unweighted number of tardy jobs on a *single* batch processing machine with incompatible job families. The scheduling problem on unrelated parallel batch ma-

chines with capacity restrictions is treated in [67]. Several heuristics for minimizing the makespan are proposed. [7] propose and compare metaheuristics to solve the scheduling problem of jobs with ready times on identical parallel machines and when the total weighted tardiness is minimized. Again in the semiconductor industry, [52] consider the scheduling problem of a set of n jobs with different job sizes on a set of m *identical* and parallel batch machines with the objective of minimizing the makespan. A metaheuristic in two phases is proposed: In the first phase, the jobs are grouped into batches and, in the second phase, the batches are scheduled on the machines.

Chapter 5 deals with the impact of batch size constraint on qualification management. In all cited qualification management studies, batch sizes are ignored.

Variability

Stochastic modeling has been widely used to measure the production variability of production lines [47]. Many studies have been done on the measurement of the variability of the production flows in a fluid modeling network [19]. In our study, we only consider one workcenter in one period and evaluate the variability using an optimization model.

The benefits of process flexibility in capacity utilization and sales increase in supply chains is extensively studied in [56]. [38] define a flexibility measure for supply chain systems. At plant level, manufacturing flexibility between plants and products, with various capacity limitations and demands is studied in [10]. At workcenter level, [54] propose flexibility measures to evaluate the manufacturing flexibility of a given workcenter configuration according to the production volume or production time. The industrial implementation and consideration of special cases for new and alternate recipes and their impact on toolset capacity is further studied in [82].

The relationship between manufacturing flexibility and production variability is studied in [71]. The study is conducted on a multi-plant multi-product make-to-order manufacturing supply chain. Based on an optimization-based simulation model, it is shown that, by increasing the manufacturing flexibility, the production

variability is reduced.

Production throughput variability is calculated for a single workstation with deterministic process times and random downtimes in [62]. Probability density function and variance of time to produce are developed for a fixed lot size. Finally, it is shown how the proposed probability density function can be used in discrete event simulation to generate a cycle time distribution of a lot size of one.

In Chapter 6, we study the impact of *qualification management* on *production variability*.

3.6 Conclusions

In this chapter, we defined the recipe-to-machine qualification as a characteristic of semiconductor manufacturing. Qualifications provide the possibility of workload allocation of a recipe to a machine. An appropriate recipe-to-machine qualification configuration allows an efficient capacity utilization of the toolset. Flexibility is seen as an important manufacturing element in semiconductor industry. With this background, we attack the core of the subject. In the remainder of this part of the thesis, we discuss flexible qualification management and capacity planning in semiconductor industry from different angles. In the next chapter, we will see how limited machine capacity affects qualification management.

Chapter 4

Flexible Qualification Management under Capacity Constraint

In Semiconductor Manufacturing, machines are usually qualified to process a limited number of recipes related to products. It is possible to qualify recipes on machines to better balance the workload on machines in a given toolset. However, all machines of a toolset do not have equal uptimes, e.g. they may have different scheduled and unscheduled downtimes. This may heavily impact an efficient recipe-to-machine qualification configuration. In this chapter, we propose indicators for recipe-to-machine qualification management based on the overall toolset workload balance under capacity constraints. The models, deployed in industry, demonstrate the importance of considering toolset capacity while managing qualifications. Industrial experiments are presented and discussed.[1]

[1]Part of this chapter has been published in OMEGA [84].

4.1 Introduction and Motivation

A machine is not always *productive*. The *Total Time* of a machine is usually categorized under these superstates [86] (Figure 4.1): *Non-Scheduled Time*, *Unscheduled Downtime*, *Scheduled Downtime*, *Engineering Time*, *Standby Time* and *Productive Time*. The only state which is really used for production is the *Productive Time*. Each machine, depending upon its condition, usage and importance in the production system, has a target productive time. This productive time is referred hereafter to as the *maximum capacity* (or shortly, the *capacity*) of the machine. The capacity of machines in the same toolset may be different, and is dynamic over time due to for instance maintenance plans. At each time interval and planning phase, the available capacity of each machine must be considered. Ignoring this important factor while planning may lead to infeasible or inefficient plans. In this chapter, we take into consideration the capacity of each machine in a toolset for qualification management (QM). We extend the WIP and Time Flexibility measures introduced in [54] and, after analyzing these extended measures, we show that additional measures are required. According to Figure 4.2, the only time slot in which production takes

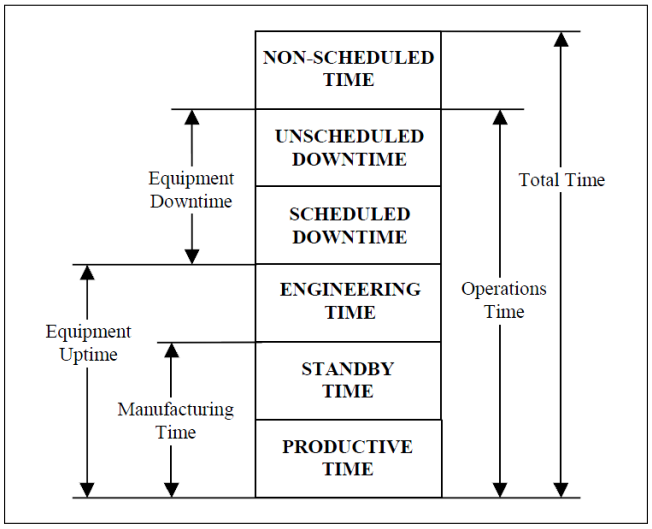


Figure 4.1: Equipment States Stack Chart ([86])

place is the "Productive Time". This status is a percentage of the total time which we refer to as the *maximum available time* or the *maximum capacity*. *Maximum*

capacity may also be limited when the machine is shared with another production line.

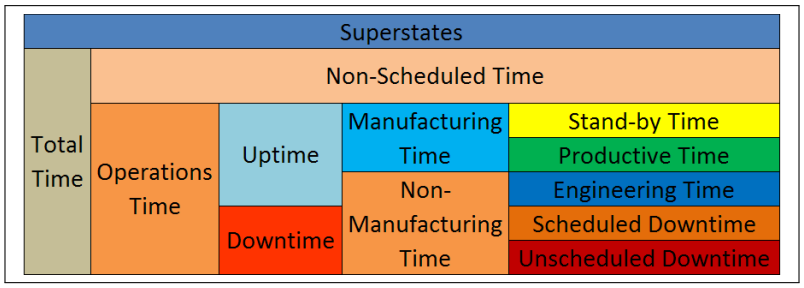


Figure 4.2: Total Time Components

4.2 Capacitated Flexibility Measures

Flexibility Measures based on two different criteria (*recipe-to-machine configuration robustness* and *toolset workload balance*) are defined in [54] for QM. The *Toolset Flexibility Measure* evaluates the *robustness* of a recipe-to-machine qualification configuration. Taking into account capacity in this flexibility measure is not critical and will not be discussed in this thesis. The two other flexibility measures are *WIP (Work-In-Process) Flexibility* and *Time Flexibility*. They aim at *balancing the workload* on the machines in a toolset. The *WIP Flexibility Measure* evaluates the recipe-to-machine qualification configuration with regard to the workload balance in terms of production volumes (or WIP). The *Time Flexibility Measure* evaluates the qualification configuration with regard to the workload balance in terms of production times. Flexibility Measures vary between 0 and 1. Higher flexibility values indicate a more effective qualification configuration. In order to evaluate the impact of each new qualification, the flexibility value of the current qualification configuration is calculated and stored. Then each qualifiable recipe-to-machine couple is virtually qualified, and the resulting qualification configuration is recalculated and stored. By subtracting the flexibility values for each new configuration from the current flexibility value, the *flexibility gain* of each new qualification is computed.

In Section 4.2.1, we recall the WIP and Time Flexibility Measures proposed

in [54]. These measures assume that all machines have (unlimited) equal capacity. Hereafter, we refer to these Flexibility Measures as *Uncapacitated Flexibility Measures*. By modifying these measures, we define in Section 4.2.2 new flexibility measures which consider the capacity of each machine. These new flexibility measures are referred to as *Capacitated Flexibility Measures*. While taking into account capacity constraints, we show that complementary measures, called *Capacity Deviation Ratio*, are required to appropriately evaluate the qualification configuration of a toolset. These measures are introduced in Section 4.2.3. In Section 4.4, we discuss how *Capacitated Flexibility* and *Capacity Deviation Ratio* measures may be used to interpret the workload balancing diagram used for capacity planning.

Below, the *parameters* and *variables* used throughout the chapter are defined.

Parameters

R	Total number of recipes to be processed,
M	Total number of machines in the toolset,
WIP_r	Total production volume of recipe r,
$TP_{r,m}$	Throughput rate of recipe r on machine m (number of wafers per hour),
$Capa_m$	Capacity of each machine m (in hours),
$Q_{r,m}$	$\begin{cases} 1 \text{ if recipe } r \text{ is qualified on machine } m, \\ 0 \text{ if recipe } r \text{ is not qualified on machine } m. \end{cases}$
γ	Workload balancing exponent ($\gamma \geq 1$).

Variables

$WIP_{r,m}$	The production volume of recipe r assigned to machine m,
$C_{r,m}$	The production time of recipe r on machine m,
WIP_m	The total production volume assigned to machine m ($WIP_m = \sum_{r=1}^{R} WIP_{r,m}$),
C_m	The total production time assigned to machine m ($C_m = \sum_{r=1}^{R} \frac{WIP_{r,m}}{TP_{r,m}}$).

4.2.1 Uncapacitated Flexibility Measures

Two flexibility measures are defined in [54] to evaluate how well the workload can be balanced on a toolset with its qualification configuration. These measures do not take into account the capacities of machines in a toolset. Following, the *Uncapacitated WIP* and *Time Flexibility Measures* are recalled and briefly discussed.

4.2.1.1 Uncapacitated WIP Flexibility Measure

By balancing the workload in terms of WIP in the toolset, the Uncapacitated WIP FM F^{WIP}_{Uncapa} evaluates the recipe-to-machine qualification configuration of a toolset [54]:

$$F^{WIP}_{Uncapa} = \frac{\left(\frac{\sum_{m=1}^{M} WIP_m}{M}\right)^{\gamma}}{\frac{\sum_{m=1}^{M} (WIP_m)^{\gamma}}{M}} \in (0,1]\,. \tag{4.1}$$

For any γ, F^{WIP}_{Uncapa} attains its maximum value, i.e. 1, when the numerator and denominator are equal. The term $(\sum_{m=1}^{M} WIP_m)$ in the numerator represents the overall workload on all machines, and is equal to the total production volume of all recipes $(\sum_{r=1}^{R} WIP_r)$. This is because all of the production volume of each recipe must be produced. Hence, the numerator in (4.1) represents an equal distribution of the total production volume over all machines $(\sum_{r=1}^{R} WIP_r/m)$. This is the case of a perfectly-balanced toolset. The denominator represents the same value from a machine perspective. It corresponds to an equal distribution of the total workload over all machines. If the numerator and denominator are equal, it means that the average production volume distribution and the average workload distribution are the same. Therefore, by maximizing F^{WIP}_{Uncapa} under a given recipe-to-machine qualification configuration, the optimal workload balance of the toolset is obtained.

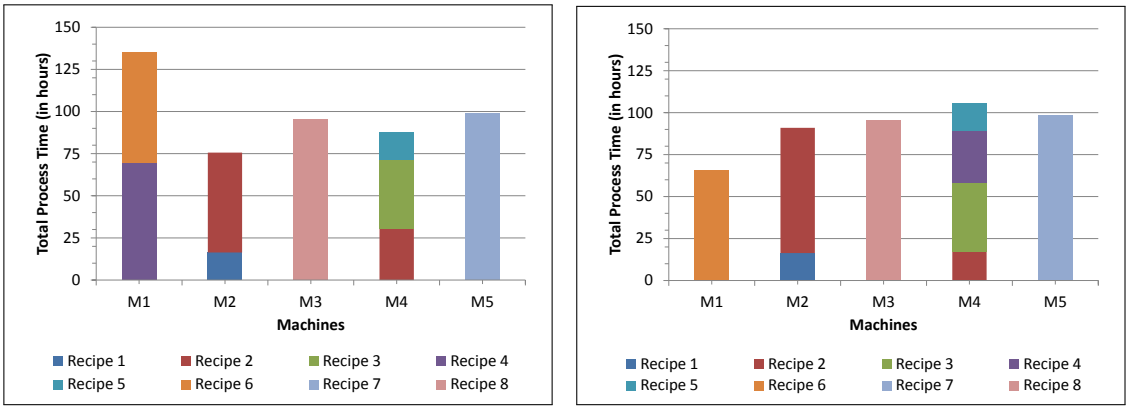
4.2.1.2 Uncapacitated Time Flexibility Measure

The *Uncapacitated Time FM* F^{Time}_{Uncapa} evaluates the recipe-to-machine qualification configuration with respect to the workload balance on the toolset in terms of production times. The definition from [54] is as follows:

$$F^{Time}_{Uncapa} = \frac{Ideal\ Ratio_{Uncapa}}{\sum_{m=1}^{M}\left(\sum_{r=1}^{R}\frac{WIP_{r,m}}{TP_{r,m}}\right)^{\gamma}} \in (0,1]\,. \tag{4.2}$$

As with F^{WIP}_{Uncapa}, F^{Time}_{Uncapa} varies between 0 and 1. The numerator called $IdealRatio_{Uncapa}$ is the minimum value of the total production times $\sum_{m=1}^{M}(\sum_{r=1}^{R}\frac{WIP_{r,m}}{TP_{r,m}})^{\gamma}$ when all of the qualifiable recipe-to-machine couples are "virtually" qualified.

Contrary to F^{WIP}_{Uncapa}, γ not only decreases the flexibility value but also affects the workload balancing in F^{Time}_{Uncapa}. A reduced example based on industrial data is used throughout the article to demonstrate **(a)** the behavior of Capacitated and Uncapacitated FMs, namely capacitated versus uncapacitated workload balancing, **(b)** the impact of γ while workload balancing using Time FMs and **(c)** the impact of a qualification on the toolset workload balance. The dataset is composed of 8 recipes and a toolset of 5 machines. Figure 4.3a shows the optimal workload balance diagram for the current toolset qualification configuration using F^{Time}_{Uncapa}, and with $\gamma = 2$. After calculating the flexibility gains for all qualifiable recipe-to-machine couples, the best qualification gain is evaluated at 15.4%. This brings F^{Time}_{Uncapa} to 97.7%. The optimal workload balance for the toolset qualification configuration after performing the best qualification is shown in Figure 4.3b. Now, the total workload is better distributed and more evenly balanced in the toolset. Note that, as the capacity of all machines are *implicitly* considered to be equal and unlimited, the total process time of all machines is centered around the average (about 100 hours). Let us evaluate again the Uncapacitated Time FM using a larger γ ($\gamma = 10$). Figure 4.4 shows the optimal workload balance on the toolset before and after the best qualification. It is worth to note that the flexibility value for the current qualification configuration has decreased from 82.3% (for $\gamma = 2$) to 11.3% (for $\gamma = 10$). The most important change is in the optimal workload balance. Figures 4.3 and 4.4 illustrate that increasing γ shifts more workload from "high speed" machines ($M3$, $M4$ and $M5$) to machines

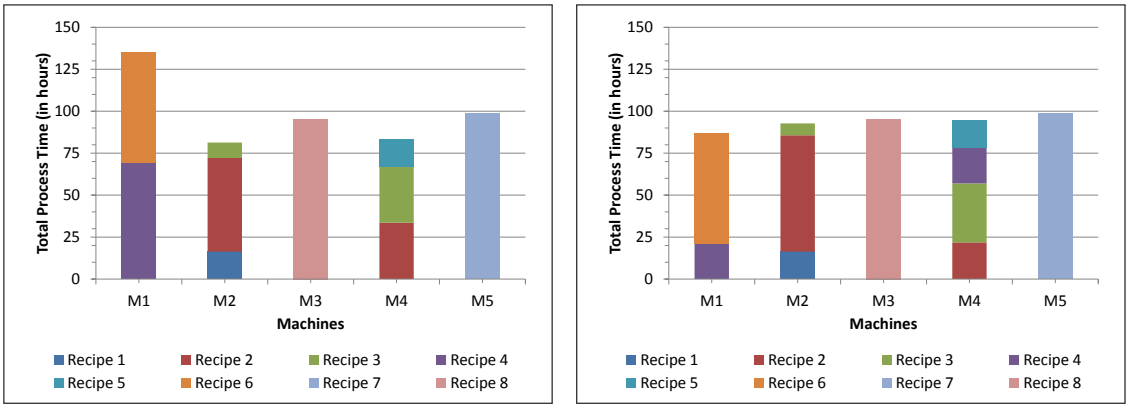


(a) Current Process Time Distribution ($F^{Time}_{Uncapa} = 82.3\%$)

(b) Process Time Distribution after Best Qualification ($F^{Time}_{Uncapa} = 97.7\%$)

Figure 4.3: Toolset Workload Balance Using Uncapacitated Time FM ($\gamma = 2$)

with lower throughput ($M1$ and $M2$). Therefore, increasing γ leads to an increase of the total process time and at the same time to a decrease of the maximum process time. These remarks call for a careful choice of γ via observation of the toolset workload in the fab and the priority of managers.



(a) Current Process Time Distribution ($F^{Time}_{Uncapa} = 11.3\%$)

(b) Process Time Distribution after Best Qualification ($F^{Time}_{Uncapa} = 89.3\%$)

Figure 4.4: Toolset Workload Balance Using Uncapacitated Time FM ($\gamma = 10$)

4.2.2 Capacitated Flexibility Measures

In this section, by modifying the Uncapacitated FMs, we introduce new FMs which take into account the *limited* and *unequal capacity* of each machine of a toolset. As Uncapacitated FMs, Capacitated FMs vary between 0% and 100%.

4.2.2.1 Capacitated WIP Flexibility Measure

The *Capacitated WIP FM* may be used for toolsets with homogeneous machines or machines with similar process times for all recipes. The capacity of each machine is restricted. The capacity restriction may be due to several reasons. For example, the *Preventive Maintenance (PM)* counter may be limited to a certain level and it is not desired to trigger the PM before the next planning period. In this case, the maximum capacity of the machine(s) in question must be limited to the desired level. In the same category, we can mention the case of machines using consumable materials such as *Chemical-Mechanical Polishing (CMP)* machines. The pads used for polishing are worn out over time. These pads must be changed after polishing a given number of wafers. In order not to interrupt the production in a planning period, the maximum capacity of the machines may be restricted to a fixed number of wafers. F_{Capa}^{WIP} evaluates the flexibility of a recipe-to-machine qualification configuration of a toolset from the standpoint of WIP workload balance under capacity constraint.

$$F_{Capa}^{WIP} = \frac{\left(\frac{\sum_{m=1}^{M}(WIP_m/Capa_m)}{M}\right)^{\gamma}}{\frac{\sum_{m=1}^{M}(WIP_m/Capa_m)^{\gamma}}{M}} \in (0,1] \tag{4.3}$$

Since $WIP_m = \sum_{r=1}^{R} WIP_{r,m}$, (4.3) can be reformulated as in (4.4):

$$F_{Capa}^{WIP} = \frac{M \cdot \left(\frac{\sum_{m=1}^{M} \sum_{r=1}^{R} (WIP_{r,m}/Capa_m)}{M} \right)^{\gamma}}{\sum_{m=1}^{M} \left(\sum_{r=1}^{R} WIP_{r,m}/Capa_m \right)^{\gamma}} \in (0,1]\,. \tag{4.4}$$

When $\gamma \geq 1$, F_{Capa}^{WIP} is a convex function. For any value of $\gamma = 1$, F_{Capa}^{WIP} attains its maximum value of 1 when the numerator and the denominator are equal. This means that the workload of each machine against its maximum capacity is equal. As with F_{Uncapa}^{WIP}, γ has no impact on the toolset workload balancing. The increase of γ just decreases the value of F_{Uncapa}^{WIP}. γ is an interesting parameter to adjust the WIP FMs (both Uncapacitated and Capacitated) according to the real workload distribution on the shop floor. With a low γ, for instance 1, the WIP FM may exceed 90% even with a poor workload balance. This high flexibility value may be misleading. Therefore γ must be adjusted using historical data and careful observation of the actual workload balance on the shop floor. In the numerical experiments presented in this paper, γ is set to 6, which according to our experience, leads to relevant WIP Flexibility values.

4.2.2.2 Capacitated Time Flexibility Measure

Many toolsets are composed of heterogeneous machines with different throughputs for each recipe. Therefore, it does not make sense to evaluate the qualification configuration of such toolsets based on their workload balance using *WIP FMs*. Therefore, production times must be considered. Below, we introduce the *Capacitated Time FM* (F_{Capa}^{Time}):

$$F_{Capa}^{Time} = \frac{Ideal\ Ratio_{Capa}}{\sum_{m=1}^{M} \left(\sum_{r=1}^{R} \frac{WIP_{r,m}}{TP_{r,m} \cdot Capa_m} \right)^{\gamma}} \in (0,1]\,. \tag{4.5}$$

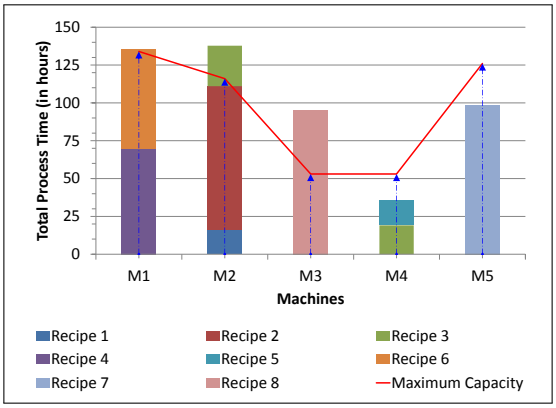
The constant $Ideal\ Ratio_{Capa}$ is the minimum value of relative production times when all of the qualifiable recipe-to-machine couples in the toolset are qualified. This is calculated by "virtually" setting all $Q_{r,m}$ to 1:

$$Ideal\ Ratio_{Capa} = \min \sum_{m=1}^{M} \left(\sum_{r=1}^{R} \frac{WIP_{r,m}}{TP_{r,m} \cdot Capa_m} \right)^{\gamma} \quad \text{with } Q_{r,m} = 1 \quad \forall r, m. \quad (4.6)$$

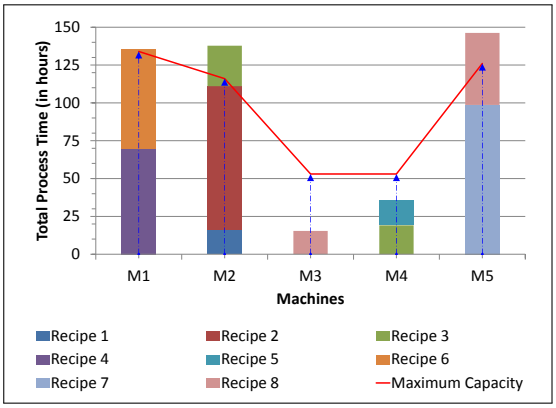
F_{Capa}^{Time} is a convex function when $\gamma \geq 1$. $Ideal\ Ratio_{Capa}$ being a constant, by maximizing F_{Capa}^{Time}, we minimize the denominator. The denominator represents the sum of relative production times which can at best be equal to $Ideal\,Ratio_{Capa}$ when the best qualification configuration is obtained.

The way capacity is modeled in (4.5) is straightforward. The term $(TP_{r,m} \cdot Capa_m)$ states that the machine throughput is relevant to its capacity. For instance, a machine with two times more capacity is the same as a machine with two times larger throughputs.

Let us reconsider the same example under capacity constraint. Figure 4.5 shows the capacitated workload balance diagram when $\gamma = 2$ for the current qualification configuration. The line depicts the maximum capacity of each machine. In this example, the Capacitated FM value is lower than the Uncapacitated FM value: $F_{Uncapa}^{Time} = 82.3\%$ and $F_{Capa}^{Time} = 61.1\%$ for the current configuration, and $F_{Uncapa}^{Time} = 95.5\%$ and $F_{Capa}^{Time} = 97.7\%$ for the toolset configuration after the best qualification. With capacity constraint, the best qualification proposition may be different than in the uncapacitated case. This is the case in our example: "Recipe 8" on "Machine 5" for *capacitated QM* instead of "Recipe 4" on "Machine 4" for *uncapacitated QM*. The impact of γ on capacitated QM using Time Flexibility is illustrated in Figures 4.5 and 4.6. By increasing γ, the total process time is increased due to a workload shift from machines with high throughput rates to machines with lower throughput rates (as in F_{Uncapa}^{Time}). However in the capacitated case, the increase of γ decreases the difference between the *process time on each machine* and the *capacity of the machine*; meaning that the capacity is better respected. By increasing the *Load Balancing Exponent* (γ), the value of the total production time (in F_{Uncapa}^{Time}) and the total relative production time (in F_{Capa}^{Time}) increase exponentially, which leads to a better leveling between each individual production time (in F_{Uncapa}^{Time}) or each



(a) Current Process Time Distribution ($F_{Capa}^{Time} = 61.1\%$ and $DR_{Capa}^{Time} = 22.9\%$)

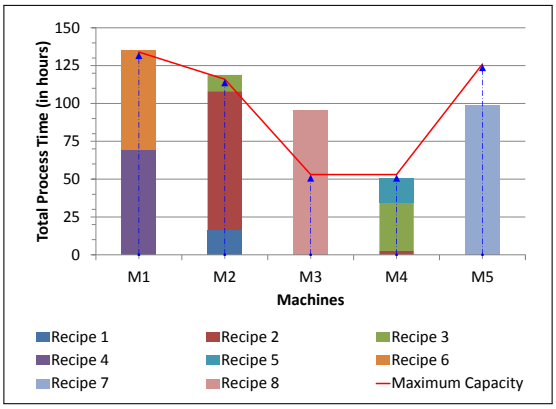


(b) Process Time Distribution after Best Qualification ($F_{Capa}^{Time} = 95.5\%$ and $DR_{Capa}^{Time} = 20.4\%$)

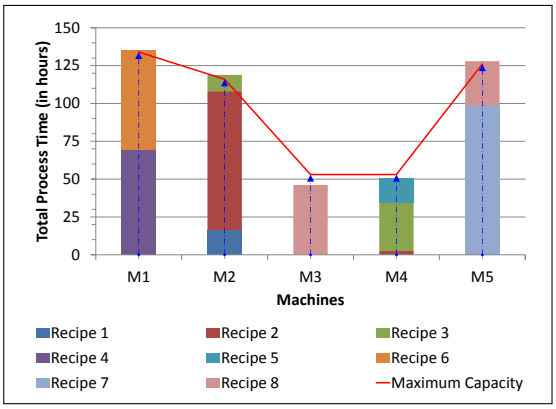
Figure 4.5: Toolset Workload Balance Using Capacitated Time FM ($\gamma = 2$)

individual relative production time (in F_{Capa}^{Time}).

As the workload balance changes when modifying γ, any change in γ influences the qualification propositions. These multiple aspects show the importance of γ in F_{Uncapa}^{Time} and F_{Capa}^{Time}. While considering capacity, we want to increase the toolset



(a) Current Process Time Distribution ($F_{Capa}^{Time} = 0.9\%$ and $DR_{Capa}^{Time} = 15.7\%$)



(b) Process Time Distribution after Best Qualification ($F_{Capa}^{Time} = 73.4\%$ and $DR_{Capa}^{Time} = 3.1\%$)

Figure 4.6: Toolset Workload Balance Using Capacitated Time FM ($\gamma = 10$)

capacity utilization while decreasing the overload. In the next section, we show that

it is not only interesting but also decisive to measure these objectives in QM under capacity constraint.

4.2.3 Capacity Deviation Measurement

Let us define new indicators called *Capacity Deviation Ratios*, which are required to complement the *Capacitated FMs*. When balancing workload under capacity constraint, some machines may be over- or underloaded. Overloaded machines slow down the production flow and increase the variability while underloaded machines lead to loss of capacity and productivity. These elements must be carefully considered in QM.

4.2.3.1 WIP Capacity Deviation Ratio (DR^{WIP}_{Capa})

As the name suggests, the *WIP Capacity Deviation Ratio* measures the workload deviation of each machine from its maximum capacity in terms of production volume. It computes, at the same time, machine under- and over-utilization.

Consider the *Capacitated WIP FM* defined in (4.3). When the ($\frac{WIP_m}{Capa_m}$) ratio is the same for all machines in the toolset, F^{WIP}_{Capa} reaches its maximum value, regardless of the possible capacity over- or under-utilization of each machine. This shows that only considering F^{WIP}_{Capa} may be misleading for QM under capacity constraint.

Below, we define the *WIP Capacity Deviation Ratio* DR^{WIP}_{Capa} which evaluates the absolute mean deviation from the maximum capacity of each machine. Note that $DR^{WIP}_{Capa} \in \mathbb{R}^+$ and that lower DR^{WIP}_{Capa} values indicate a better machine capacity utilization.

$$DR^{WIP}_{Capa} = \frac{\sum_{m=1}^{M} |WIP_m - Capa_m|}{\sum_{m=1}^{M} Capa_m} \tag{4.7}$$

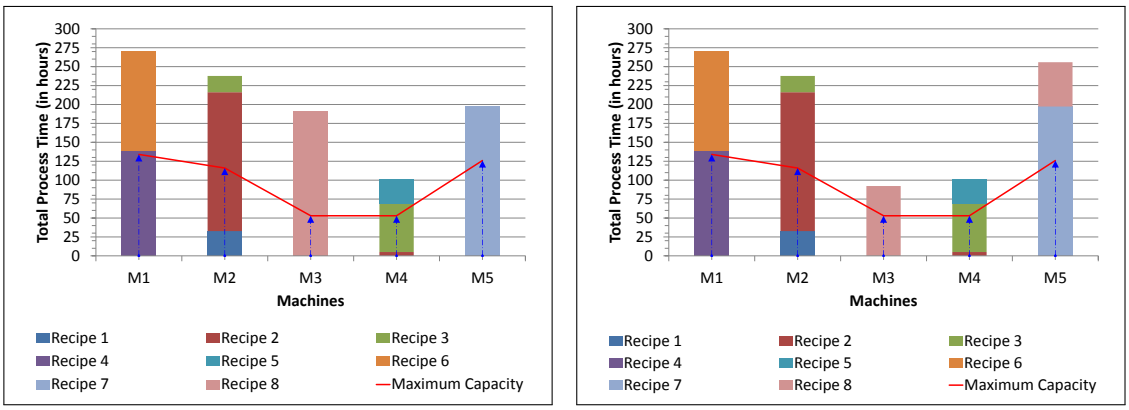
4.2.3.2 Time Capacity Deviation Ratio (DR^{Time}_{Capa})

As for F^{WIP}_{Capa}, F^{Time}_{Capa} may reach 100% although the toolset is more (or less) loaded than its maximum capacity. Hence, F^{Time}_{Capa} must be combined with the *Time Capacity*

Deviation Ratio DR^{Time}_{Capa}.

$$DR^{Time}_{Capa} = \frac{\sum_{m=1}^{M} \left| \sum_{r=1}^{R} \frac{WIP_{r,m}}{TP_{r,m}} - Capa_m \right|}{\sum_{m=1}^{M} Capa_m} \tag{4.8}$$

This case is illustrated in Figure 4.7 where we have artificially doubled the production volume of all recipes in our example. While the flexibility values remain the same ($F^{Time}_{Capa} = 0.9\%$ for the current toolset qualification configuration and $F^{Time}_{Capa} = 73.4\%$ for the toolset qualification configuration after the best qualification), the deviation ratio has shot up ($DR^{Time}_{Capa} = 15.7\%$ versus $DR^{Time}_{Capa} = 107\%$ for the current configuration and $DR^{Time}_{Capa} = 3.1\%$ versus $DR^{Time}_{Capa} = 98.8\%$ for the configuration after the best qualification). The best case is with 100% flexibility and 0% deviation ratio. This means that the workload is perfectly balanced while using the overall capacity of the toolset without over- or under-utilizing any machine. Using 27 industrial instances, the impact of γ on the *average toolset capacity*



(a) Current Process Time Distribution ($F^{Time}_{Capa} = 0.9\%$ and $DR^{Time}_{Capa} = 107\%$)

(b) Process Time Distribution after Best Qualification ($F^{Time}_{Capa} = 73.4\%$ and $DR^{Time}_{Capa} = 98.8\%$)

Figure 4.7: Toolset Workload Balance Using Capacitated Time FM ($\gamma = 10$) - Twice More Load

utilization, the *Time Capacity Deviation Ratio* and the *Capacitated Time Flexibility*

is illustrated in Figure 4.8. These industrial instances include on average 40 recipes and 26 machines with throughput rates which vary with a factor of 8.

Because F_{Capa}^{Time} depends on γ, it may not be a good measure to evaluate the impact of γ. However, DR_{Capa}^{Time} is independent of γ. Figure 4.8 depicts how increasing γ leads to increasing the toolset capacity utilization by shifting the workload from high throughput machines to machines with lower throughput rates. This is visualized with a decrease of DR_{Capa}^{Time}. However, DR_{Capa}^{Time} and the average toolset capacity utilization variations stabilize for large values of γ. *Capacity Deviation*

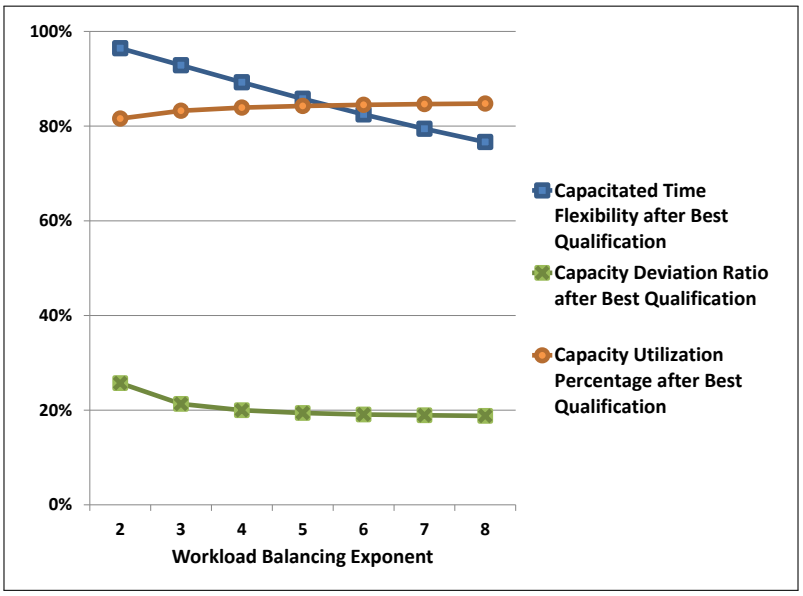


Figure 4.8: Workload balancing exponent (γ) variation versus Capacitated Time FM (F_{Capa}^{Time}), Time Capacity Deviation Datio (DR_{Capa}^{Time}) and toolset capacity utilization percentage

Ratio Indicators are essential in QM. While evaluating the qualification flexibility gain under capacity constraints, the selected qualification should increase the most the desired FM and at the same time reduce as much as possible the corresponding capacity deviation ratio indicator. This requires a bi-objective optimization of these measures which goes beyond the scope of this study.

4.2.4 Weighted Capacitated Flexibility Measures

Capacitated flexibility measures are proposed because machines do not have the same capacity. In the capacitated flexibility measures by increasing the *workload balancing exponent* (γ), the machine capacity utilization is increased, i.e. machine overload and underload are penalized at the same level although their impact is different. There is a loss of money related to non fully utilized machines (labor cost, overhead cost, ...) while, if a machine is overloaded, production is delayed and cycle times are consequently increased. Moreover, an overloaded machine tends to increase the production variability. Hence, it is relevant to weigh differently overloaded and underloaded machines.

The flexibility measures allow to identify qualifications which tend to level the workload. It may happen that the (first) new qualification(s) level more the workload of the underloaded machine while removing the overload is more important. That is why we propose weighted capacitated flexibility measures in which over- or underloaded machines are penalized linearly (using weights o_m and u_m) and nonlinearly using different workload balancing exponents (γ_o and γ_u).

The additional notations used in this section are listed below.

Parameters

γ_o	Overload Balancing Exponent,
γ_u	Underload Balancing Exponent,
o_m	Overload weight of machine m,
u_m	Underload weight of machine m.

The weighted capacitated Time Flexibility measure is defined:

$$F^{Time}_{Capa\neq} = \frac{IdealRatio_{Capa\neq}}{\sum\limits_{m=1;C_m>Capa_m}^{M} o_m\left(\frac{C_m}{Capa_m}\right)^{\gamma_o} + \sum\limits_{m=1;C_m\leq Capa_m}^{M} u_m\left(\frac{C_m}{Capa_m}\right)^{\gamma_u}}. \tag{4.9}$$

Since $C_m = \sum_{r=1}^{R}(WIP_{r,m}/TP_{r,m})$, $F_{Capa\neq}^{Time}$ can be written:

$$F_{Capa\neq}^{Time} = \frac{IdealRatio_{Capa\neq}}{\sum\limits_{m=1|C_m>Capa_m}^{M} o_m(\frac{\sum_{r=1}^{R}\frac{WIP_{r,m}}{TP_{r,m}}}{Capa_m})^{\gamma_o} + \sum\limits_{m=1|C_m\leq Capa_m}^{M} u_m(\frac{\sum_{r=1}^{R}\frac{WIP_{r,m}}{TP_{r,m}}}{Capa_m})^{\gamma_u}} \tag{4.10}$$

where

$$IdealRatio_{Capa\neq} = \min \sum\limits_{m=1|C_m>Capa_m}^{M} o_m(\frac{C_m}{Capa_m})^{\gamma_o} + \sum\limits_{m=1|C_m\leq Capa_m}^{M} u_m(\frac{C_m}{Capa_m})^{\gamma_u}$$
$$\text{with } Q_{r,m} = 1 \quad \forall r, m. \tag{4.11}$$

Overload and underload weights (o_m and u_m) can be chosen empirically for example such that $o_m + u_m = 1$. However, they can also be determined based on the cost or profit of being over- or underloaded on each machine m, e.g. by considering the cycle time decrease or increase; or the opportunity cost for not producing a product (loss of revenue).

In the same way, we define also the **Weighted Deviation Measures**:

$$DR_{Capa\neq}^{Time} = \frac{IdealDeviation_{Capa\neq}}{\sum\limits_{m=1|C_m>Capa_m}^{M} o_m|C_m - Capa_m|^{\gamma_o} + \sum\limits_{m=1|C_m\leq Capa_m}^{M} u_m|C_m - Capa_m|^{\gamma_u}} \tag{4.12}$$

$$IdealDeviation_{Capa\neq} =$$
$$\min \sum\limits_{m=1|C_m>Capa_m}^{M} o_m|C_m - Capa_m|^{\gamma_o} + \sum\limits_{m=1|C_m\leq Capa_m}^{M} u_m|C_m - Capa_m|^{\gamma_u} \tag{4.13}$$
$$\text{with } Q_{r,m} = 1 \quad \forall r, m.$$

$$C_m > Capa_m \Rightarrow \qquad o_m = 1, u_m = 0 \tag{4.14a}$$
$$C_m \leq Capa_m \Rightarrow \qquad o_m = 0, u_m = 1 \tag{4.14b}$$

$$DR_{Capa\neq}^{Time} = \frac{IdealDeviation_{Capa\neq}}{\sum\limits_{m=1}^{M}(o_m + u_m)|C_m - Capa_m|^{\gamma_o+\gamma_u}} \tag{4.15}$$

$$IdealDeviation_{Capa\neq} = \min \sum_{m=1}^{M} (o_m + u_m)|C_m - Capa_m|^{\gamma_o+\gamma_u} \quad \text{with } Q_{r,m} = 1 \quad \forall r, m. \tag{4.16}$$

4.3 Solution Approaches: Capacitated Workload Balancing

The introduced Capacitated Flexibility Measures are based on the optimal workload balance of the toolset. However, there is a noticeable difference between workload balancing in terms of production volumes (or WIP) and that of production times. Concerning the workload balance in terms of WIP, the sum of production volume ($\sum_{m=1}^{M} WIP_m$) is always the same regardless of the way the load is distributed on the toolset. Whereas, when balancing the workload in terms of production times, the total production time ($\sum_{m=1}^{M} \frac{WIP_{r,m}}{TP_{r,m}}$) depends on how the production volumes are distributed on the toolset. This makes the problem resolution considerably more difficult.

4.3.1 Capacitated Workload Balancing in terms of Production Volumes (F_{Capa}^{WIP})

The production volume workload balancing with capacity constraints is done in two phases. In the first phase, an initial (and most often non-optimal) solution is constructed. The second phase consists of improving the initial solution obtained in the first phase until finding an optimal solution. Let us first present the algorithm used for the construction of the initial workload balance and then, the algorithm used for improving this initial workload balance.

Constructing Initial WIP Workload Balance In order to construct a real-valued initial solution, the production volume of each recipe is divided among all of the qualified machines for that recipe. Following is a formalization of the algorithm:

Step 0. Set $r = 1$.

Step 1. For each machine in the toolset, allocate the production volume of recipe r to its qualified machines by equally distributing the production volume of recipe r (WIP_r) among all of the qualified machines for r ($WIP_{r,m} = WIP_r / \sum_{m=1}^{M} Q_{r,m}$).

Step 2. If $r < R$, set $r = r + 1$ and go to **Step 1**.

WIP Workload Balancing Algorithm with Capacity Constraints The *initial WIP workload balance* (initial solution) must be improved in order to find the optimal (or precise-enough near-optimal) workload balance. The main idea of this algorithm has been presented in [54] for *uncapacitated production volume workload balancing*. In this paper, we formalize and explain in detail the adapted algorithm for *capacitated production volume workload balancing*.

The algorithm begins with an initial solution x_0, generated by the initial WIP distribution algorithm.

Before describing the algorithm, let us define the set of *loading machines* for recipe r ($\mathscr{LM}_r$) to be the set of qualified machine(s) for recipe r, with the smallest $WIP_m/Capa_m$ ratio among all of the qualified machines for r, i.e.:

$$m^* \in \mathscr{LM}_r \quad \text{if} \quad \frac{WIP_{m^*}}{Capa_{m^*}} = \min_{\forall m | Q_{r,m}=1} \left\{ \frac{WIP_m}{Capa_m} \right\} \tag{4.17}$$

Step 0. Start with initial solution x_0. Set $r = 1$ and $k = 1$.

Step 1. For all of the qualified machines for recipe r, remove the WIP of recipe r initially allocated to machine(s) m:

$$\text{If } WIP_{r,m} > 0 \text{ then } WIP_m = WIP_m - WIP_{r,m} \quad \forall m \tag{4.18}$$

Sort the machines according to their $WIP_m/Capa_m$ ratio in an ascending order.

Step 2. Allocate (equally) WIP quantities to loading machine(s) until the $WIP_{m^*}/Capa_{m^*}$ (or $\sum_{r=1}^{R} WIP_{r,m^*}/Capa_{m^*}$) ratio on one (or more) loading machine(s)

m^* in $\mathscr{LM}_r$ is equal to the $WIP_{m'}/Capa_{m'}$ ratio of a machine m' qualified for recipe r but not in $\mathscr{LM}_r$:

$$\frac{WIP_{m^*}}{Capa_{m^*}} = \frac{WIP_{m'}}{Capa_{m'}} \tag{4.19}$$

Then, m' is added to the set of loading machines, i.e. $\mathscr{LM}_r \equiv \mathscr{LM}_r \cup \{m'\}$.
If $r < R$, set $r = r + 1$ and go to **Step 1**.

Step 3. If $F^{WIP}_{Capa\,k} - F^{WIP}_{Capa\,k-1} > 0$, then set $k = k + 1$, $r = 1$ and go to **Step 1**.

4.3.2 Capacitated Workload Balancing in terms of Production Times (F^{Time}_{Capa})

The resolution method for the Time Flexibility Measure is more complex than the WIP Flexibility Measure. F^{Time}_{Capa} can be solved using the same approach proposed in [53] for F^{Time}_{Uncapa}.

In order to find the optimal workload balance for the Capacitated Time Flexibility Measure (F^{Time}_{Capa}), the optimization problem (4.20) must be solved when F^{Time}_{Capa} is maximized. The only constraint of the model ensures that the production volume of each recipe r is entirely allocated to qualified machines ($Q_{r,m} = 1$).

$$\begin{aligned} \max \quad & F^{Time}_{Capa} = \frac{Ideal\ Ratio_{Capa}}{\sum_{m=1}^{M}(\sum_{r=1}^{R}\frac{WIP_{r,m}}{TP_{r,m}\cdot Capa_m})^{\gamma}} \\ \text{subject to} \quad & \\ & \sum_{m=1|Q_{r,m}=1}^{M} WIP_{r,m} = WIP_r \qquad \forall r \\ & WIP_{r,m} \geq 0 \qquad \forall r, m \end{aligned} \tag{4.20}$$

Since $Ideal\ Ratio_{capa}$ is constant, by maximizing F^{Time}_{Capa}, the sum of the relative

production times (4.21) must be minimized.

$$\min \qquad f = \sum_{m=1}^{M} (\sum_{r=1}^{R} \frac{WIP_{r,m}}{TP_{r,m} \cdot Capa_m})^{\gamma} \tag{4.21}$$

Since (4.21) is a continuous function, the problem is reduced to an unconstrained minimization problem. All algorithms that minimize unconstrained continuous functions start from an initial solution (x_0) and improve the solution at each iteration until either no more improvement is made or the solution is good enough [72]. The algorithm described hereafter is a line search method called *active-set method*. In line search methods, at each iteration k (at solution (x_k)), the algorithm finds a search direction called p_k. Then the algorithm searches a new solution (x_{k+1}) by moving along the direction (p_k) at a length (α_k) with a lower function value [72]. In the active-set method, we assume that a feasible solution to the problem exists.

The idea behind the active-set method is to partition the set of constraints into active (binding) and inactive (non-binding) constraints, where the inactive constraints are essentially ignored. At each iteration (solution point k, x_k), the active-set method determines a set of constraints called the *working set* ($\mathscr{W}_k$) to be active. The surface defined by the working set is called the *working surface*. At each iteration, by moving on the working surface, the solution is improved until the optimal solution is found. Moving on the working surface consists of finding a nonzero *search direction* (p_k) and a nonnegative *step length* (α_k) over which the objective value is decreased. There exist several methods and approaches for finding a search direction and a feasible step length. While moving on the working surface, the inactive constraints must not be violated to avoid getting out of the feasible region and the solution to become infeasible. While moving on the working surface and defining the step length, one or several constraints may limit the step length. This or these constraints are called *blocking constraint(s)*. If one (or several) blocking constraint is encountered, it is added to the working set making a surface of lower dimension than before [68]. At each iteration, the Lagrangian multipliers (λ) of the working set are evaluated. If all are nonnegative, the optimal solution is obtained else we drop one or more of the constraints having a negative Lagrangian multiplier. At each iteration of the algorithm, either the solution x_k or the working set $\mathscr{W}_k$ is modified,

where either $x_k \neq x_{k+1}$ or $\mathscr{W}_k \neq \mathscr{W}_{k+1}$. Therefore for a non-degenerate case, the algorithm terminates after a finite number of iterations [103].

An investigation of the active-set method is out of the scope of this chapter, therefore we refer the interested reader to more specialized textbooks on the subject ([72], [35] and [39]).

The workload balancing in terms of production times is done in two phases (like workload balancing in terms of production volume). In the first phase, an initial (and most often nonoptimal) solution is constructed. In the second phase, the initial solution is improved until an optimal (or precise-enough near-optimal) solution is reached.

Note that the structure of the algorithm constructing the initial solution is the same as the one used for WIP initial workload balancing with the difference that production times are considered.

Step 0. Set $r = 1$.

Step 1. For each machine in the toolset, allocate the production volume of recipe r to its qualified machines by equally distributing the production volume of recipe r (WIP_r) among all of the qualified machines for r ($WIP_{r,m} = WIP_r / \sum_{m=1}^{M} Q_{r,m}$). Calculate and allocate the corresponding production time for each load allocation ($WIP_{r,m}/TP_{r,m}$).

Step 2. If $r < R$, set $r = r + 1$ and go to **Step 1**.

The initial solution is then improved in an iterative procedure. In each iteration, only one recipe (r^*) is considered. Therefore (4.21) is decomposed into the relative production time of recipe r^* on machine m ($\frac{WIP_{r^*,m}}{TP_{r^*,m} \cdot Capa_m}$) plus the relative production time of all other recipes on machine m ($\bar{f}_{r^*,m}$) (4.22) ([53] and [39]).

$$f = \left(\frac{WIP_{r^*,m}}{TP_{r^*,m} \cdot Capa_m} + \bar{f}_{r^*,m} \right)^{\gamma} \tag{4.22}$$

Step 0. Set $r^* = 1$ and $k = 1$.

Step 1. Remove the initially allocated load $WIP_{r^*,m}$ of recipe r^* from all machines m in the toolset (4.23).

$$\text{If } WIP_{r^*,m} > 0 \text{ then } WIP_m = WIP_m - WIP_{r^*,m} \quad \forall m \tag{4.23}$$

Step 2. Decompose (4.21) for recipe r^* to obtain (4.22). Re-distribute WIP_{r^*} using the *active-set method*. If $r^* < R$, set $r^* = r^* + 1$ and go to **Step 1**.

Step 3. If $f_k - f_{k-1} > 0$ (4.21), then set $k = k + 1$ and $r^* = 1$ and go to **Step 1**.

The active-set method is the core of the algorithm discussed above for the re-distribution of recipe r^*. Before starting the description of a simple active-set method, we introduce the notations and some necessary elements of the algorithm.

Notations

$\mathscr{W}_k$	Working set of active constraints at solution k,
$\bar{A}$	Constraint matrix for active constraints,
$\bar{A}_r$	Right inverse for $\bar{A}$,
$\bar{Z}$	Null-space matrix for $\bar{A}$,
∇f	Gradient vector of f,
$\nabla^2 f$	Diagonal elements of the Hessian matrix of f.

In order to calculate the Lagrangian multipliers and step length, we need the gradient vector (4.24) and diagonal elements of the Hessian matrix (4.25) of (4.22).

$$\nabla f = \frac{\gamma}{TP_{r^*,m} \cdot Capa_m} \left(\frac{WIP_{r^*,m}}{TP_{r^*,m} \cdot Capa_m} + \bar{f}_{r^*,m} \right)^{\gamma-1} \tag{4.24}$$

$$\nabla^2 f = \frac{\gamma\,(\gamma-1)}{(TP_{r^*,m} \cdot Capa_m)^2} \left(\frac{WIP_{r^*,m}}{TP_{r^*,m} \cdot Capa_m} + \bar{f}_{r^*,m} \right)^{\gamma-2} \tag{4.25}$$

An algorithm of a simple active-set method is provided below ([39], [53], [72], [28]).

Step 0. Start with an initial feasible solution x_0 and working set of active constraints at x_0 ($\mathscr{W}_0$). Set $k = 1$.

Step 1. The Optimality Test

If $\bar{Z}^T \nabla f(x_k) = 0$

Step 1a. If there are no active constraints ($WIP_{r^*,m} = 0$), STOP - Local Stationary Point.

Step 1b. Else, compute Lagrangian Multipliers:

$$\bar{\lambda} = \bar{A}_r^T \nabla f(x_k). \tag{4.26}$$

Step 1c. If $\bar{\lambda} \geq 0$, STOP - Local Stationary Point. Else delete the constraint corresponding to the most negative Lagrangian multiplier from the working set $\mathscr{W}_k$.
Update $\mathscr{W}_k$, $\bar{Z}$, $\bar{A}$ and $\bar{A}_r$.

Step 2. The Search Direction

Compute a descent feasible search direction p_k with respect to the active constraints in $\mathscr{W}_k$. The *Reduced Newton Search Direction* (5.7) [50] is an efficient search direction.

$$p_k = -\bar{Z}\left(\bar{Z}^T \nabla^2 f(x_k)\, \bar{Z}\right)^{-1} \bar{Z}^T \nabla f(x_k) \tag{4.27}$$

Step 3. The Step Length

Compute a step length α_k such that $f(x_k + \alpha_k p_k) < f(x_k)$ subject to retaining feasibility with respect to all constraints. For that $\alpha_k \leq \bar{\alpha}_k$, where $\bar{\alpha}$ is the maximum feasible step length along p_k:

$$\alpha_k = \min_{i \notin \mathscr{W}_k, p_k < 0} \frac{WIP_k}{-p_k} \tag{4.28}$$

Step 3a. If $\alpha_k < \bar{\alpha}_k$, α_k is an unconstrained step, i.e. no blocking constraint is encountered and $x_{k+1} = x_k + \alpha_k p_k$ remains feasible.

Step 3b. If $\alpha_k \geq \bar{\alpha}_k$, then the step along p_k is blocked by one (or more) (inactive) constraint(s) not in $\mathscr{W}_k$ but which are active in x_{k+1}. In this case, $\mathscr{W}_k$ is modified to include the blocking constraint to the

working set ($\mathscr{W}_{k+1} \leftarrow \mathscr{W}_k \cup i$).
Update $\bar{Z}$, $\bar{A}$ and $\bar{A}_r$. Set $k = k + 1$ and go to **Step 1**.
If more than one blocking constraint exists (i.e. more than one constraint boundary is reached), the solution is degenerate and only one of the blocking constraints is added to the working set.

4.3.3 Capacitated Workload Balancing in terms of Production Volumes (F^{WIP}_{Capa}) - Alternative Approach

An alternative approach to the algorithm proposed in Section 4.3.1 is to substitute $TP_{r,m} \cdot Capa_m$ with the capacity of each machine ($Capa_m$) in the Capacitated Time Flexibility Measure (F^{Time}_{Capa}).

Note that $F^{Time}_{\bullet}$ reduces to $F^{WIP}_{\bullet}$ when the throughput of all recipes are equal on all machines.

4.4 Managerial Insights

The proposed methodology is being applied in industry. Based on the industrial experiments, some of the managerial insights are pointed out.

The Flexibility Measures described in this chapter, were implemented with the programming language *VBA* in Microsoft Excel®. The application solves the optimization problem for the current qualification configuration and saves the *current flexibility level* according to the selected flexibility measure. Then it virtually qualifies each of the qualifiable recipes and saves it in a matrix. Finally the flexibility gain is calculated by subtracting each element of the matrix from the current flexibility level while highlighting the best qualification.

The application has been used for more than one year for the fab's main bottleneck toolset as a decision support system for qualification management and capacity planning.

The results show that, on average, by performing the best qualification with $\gamma = 6$, the Capacitated Time Flexibility Measure (F^{Time}_{Capa}) increased by 23.7%, the Time

Capacity Deviation Ratio (DR_{Capa}^{Time}) decreased by 9.6% and the toolset maximum workload decreased also by 4%. The indicators show the considerable improvements in Toolset Flexibility and capacity utilization with only one qualification.

The capacitated flexibility measures proposed in this study are based on toolset workload balance. Several outputs of the model can be used and interpreted by different users and from different aspects. Thanks to the flexibility gain table, the process department can perform a limited number of qualifications bringing more flexibility. At the same time, useless and costly qualifications with poor flexibility gains are avoided. Used as a decision support system, among several qualification alternatives, the best qualification can be chosen by making a trade-off between the flexibility gain and the required qualification effort. In a more long-term vision, necessary qualifications could be anticipated by taking into account future production volumes. This reduces the obligation of performing the so-called *rush qualifications*. By identifying in advance the set of new qualifications to be performed, the incurred downtime is minimized and the impact on the production line can be better managed.

Using the *Capacitated Time Flexibility Measure*, the optimal workload balance in terms of process time is calculated. It is possible to observe the workload balance improvement after performing a new qualification. Therefore, the workload balance diagram is used by capacity planners, which serves also as a guideline for capacity improvement actions.

Thanks to Capacitated Flexibility Measures, we are able to study the impact of preventive maintenance or machine breakdown on the toolset qualification configuration and workload balance. The proposed model can also be used in the case of production ramp-up and new machine acquisition. By creating several scenarios consisting of dummy machine(s) and running the application, the impact of each scenario is calculated on flexibility level and toolset workload. By plotting the graph of the flexibility gain and/or deviation ratio against the *Return On Investment (ROI)* of each scenario, the best solution can be chosen.

In the model, the volumes are allocated without considering scheduling constraints, in other words as if preemption and splitting of recipes on machines were allowed. Therefore the effect of scheduling has not been taken into account. How-

ever the detailed WIP allocation on the machines can be used as a rough guideline for load allocation. It is also possible to use this output as the input of a scheduler and dispatcher. In addition, note that a new qualification with high flexibility gain helps to better balance the workload on the machines and gives more alternatives for processing a recipe. This gives more flexibility for dispatching lots on the machines and levels the workforce of the toolset.

The toolset qualification configuration changes over time. This is due to new qualifications, disqualifications, machine breakdowns, machine acquisitions) but also the change of the production volume and mix. The flexibility level of a recipe-to-machine qualification configuration, when monitored continuously, can serve as a warning. Preventive measures (such as (re-)qualification and at a strategic level, new machine acquisition) can be triggered when the Flexibility Measures falls below a certain level.

4.5 Conclusions and Perspectives

In this chapter, we investigated an approach for qualification management in semiconductor manufacturing. Indicators called Capacitated Flexibility Measures were proposed to evaluate the workload balance of a toolset under equipment capacity restriction for a given qualification configuration. It was shown that Flexibility Measures alone are not enough to select new qualifications. So, complementary indicators called Deviation Ratios were introduced. In the development of solution approaches, two main problems have been considered. In the first one, the toolset consists of homogeneous machines where all process times are equal. The second problem considers heterogeneous machines with different recipe-to-machine process times. For both optimization models (WIP and Time), dedicated solution approaches were presented. Using real fab data, we illustrated the relevance of dedicated Capacitated Flexibility Measures and Capacity Deviation Ratios for qualification management and capacity planning.

Due to lack of space, some other practical aspects of qualification management are not discussed in this chapter. As stated in the introduction, some recipes are *non-qualifiable*. This makes qualification management a bit more complicated. Besides,

due to new product launch or disqualifications, some recipes may need to be qualified first on the machines. A study of these recipe-types called *not previously qualified recipes* for *Uncapacitated Flexibility Measures* can be found in [82]. Another aspect of qualification management is *recipe priority*. Some recipes may be more important than others because of deadlines, customer needs, etc.

As a perspective for future investigations concerning this chapter, it is interesting to study bi-objective optimization of the both flexibility and deviation ratio measures.

In order to reduce setup and production costs, and obtain a uniform result on the products, batch processing is used in some toolsets in semiconductor manufacturing. Batch size therefore impacts the toolset workload balance and must be considered for qualification management. The next chapter is centered around this industrial issue.

Chapter 5

Flexible Qualification Management under Batch Size Constraint

Qualification and batch size constraints are two important characteristics of many toolsets in semiconductor manufacturing. In this chapter, we consider the impact of considering batch sizes on two flexibility measures used for qualification management. We propose several solution approaches for balancing the workload on machines in terms of production volume or production time. Using numerical experiments conducted on real fab data, the solution approaches are compared and the impact of batch sizes on qualification management is discussed.[1]

[1]Part of this chapter has been published in [81] and submitted to IJPE. The article presented in Winter Simulation Conference 2013 has won the Best Applied Paper Award.

5.1 Introduction and Motivation

Until now, we have demonstrated that qualifications have a direct impact on production capacity. In this chapter, we study the impact of batching on qualification management. Wafers of the same type travel together in units called *lots*. Each lot consists of normally 25 wafers. Depending upon the machine, lots are grouped to form *batches*. The lots of a batch undergo the same process. Batching is used for the reduction of manufacturing process variability which can result in defective parts. It is also used to reduce the manufacturing costs. Hence, it is mostly appropriate for processes with high setup or changeover costs, for instance long-run processes such as those in furnaces. Besides, batching assures a uniform processing condition for all lots. Small batch size increases manufacturing cost (setup, labor, tool replacement, etc.) while a batch size being *too large*, *may* lead to increased tolerance and product variability. Therefore an optimal batch size must be determined according to machine specifications and manufacturing process variability [87].

Batch processing is widely used in semiconductor manufacturing as well as other industries such as the metallurgical industry. Some typical batch processing steps in semiconductor manufacturing are thermal treatment, cleaning, implantation, etc. Batching may be either serial (called s-batching) or parallel (called p-batching). In semiconductor manufacturing, p-batching is more frequent than s-batching [69]. In this chapter, we consider p-batching.

In some cases, lots are ungrouped when a machine processes less than 25 wafers at each *production run*. In the toolset for which the numerical experiments in Section 5.4 are conducted, the batch size varies from 4 to 10 lots at each production run (equivalent to 100 to 250 wafers per run)! The wide range of the batch size motivates the study of its impact on qualification management and hence capacity planning. As the machine specifications are different, the batch size of the same recipe may vary depending upon the machine. Since the throughput of each machine differs from recipe to recipe, an efficient qualification management considering these constraints optimizes the capacity utilization of the toolset while reducing the production variability and the cycle time.

Both recipe-to-machine qualification configuration and batches are two impor-

tant deterministic sources of variability ([83] and see Chapter 6). To avoid variability propagation through the production line, mastering variability is more crucial for bottleneck toolsets.

In this chapter, we investigate the impact of the batch size on the recipe-to-machine allocation used in qualification management and toolset capacity. In Section 5.2, the optimization model for batch size constraints is presented and the complexity is analyzed. In Section 5.3, several solution approaches for the flexibility measures considering batch size constraint are proposed. In Section 5.4, numerical experiments on real fab data for a "Thermal Treatment" toolset are presented and discussed. Finally, we draw conclusions and mention some research perspectives in Section 5.5.

5.2 Optimization Model and Complexity Analysis

In the following sections, we discuss the related mathematical model and complexity analysis.

5.2.1 Optimization Model

In this chapter, we propose solution approaches for workload balancing for both the WIP and time Flexibility measures. In addition, recipe-to-machine WIP allocations must be an integer multiple of the batch size of that recipe-to-machine, except for at most one recipe-to-machine allocation for each recipe. The latter is because it is not always possible to find batch sizes for recipe r so that their sum is exactly equal to WIP_r. Let us denote by $Y_{r,m}$ a binary variable which is equal to 1 if the quantity of recipe r assigned to machine m is *not* an integer multiple of the batch size and 0 otherwise. Let us also denote by $WIPC_{r,m}$ the WIP allocation of recipe r to machine m which can be continuous, i.e. not an integer multiple of the batch size.

The total WIP allocated to each machine can constitute full batch(es) (if any) and/or a non full-batch allocation. Therefore, in the flexibility measures (F^{WIP} and

F^{Time}), $WIP_{r,m}$ is to be substituted with $Run_{r,m}BS_{r,m} + WIPC_{r,m}$. $Run_{r,m}BS_{r,m}$ states that the WIP allocation of recipe r on machine m is an integer multiple (i.e. number of full-batch runs) of the batch size $BS_{r,m}$ of recipe r on machine m.

$$\begin{aligned}
\max \quad & F && && (5.1a)\\
\text{Subject to} \quad & \sum_{m=1}^{M} (Run_{r,m}BS_{r,m} + WIPC_{r,m}) = WIP_r && \forall r && (5.1b)\\
& Run_{r,m}BS_{r,m} + WIPC_{r,m} \leq Q_{r,m}WIP_r && \forall r, m && (5.1c)\\
& WIPC_{r,m} \leq (BS_{r,m} - 1)Q_{r,m} && \forall r, m && (5.1d)\\
& WIPC_{r,m} \leq Q_{r,m}WIP_r Y_{r,m} && \forall r, m && (5.1e)\\
& \sum_{m=1}^{M} Y_{r,m} \leq 1 && \forall r && (5.1f)\\
& Y_{r,m} \in \{0, 1\} && \forall r, m &&\\
& Run_{r,m} \in \mathbb{N}^+ && \forall r, m &&\\
& WIPC_{r,m} \geq 0 && \forall r, m &&
\end{aligned}$$

The objective function F (5.1a) is either F^{WIP} or F^{Time}. Constraint (5.1b) ensures that all of the production volume of each recipe is allocated to the toolset. Constraint (5.1c) guarantees that recipe r is only allocated to machines m that are qualified for r, i.e. such that $Q_{r,m} = 1$. Constraint (5.1d) states that the non full-batch allocation must be smaller than or equal to the batch size (else a full batch can be made). Constraint (5.1e) and Constraint (5.1f) ensure that at most only one variable $WIPC_{r,m}$ can be strictly positive for each recipe r, i.e. at most one non full-batch run is allowed for each recipe r.

Note that γ is even more important in F^{Time} when considering batch sizes. By varying γ, the WIP allocation of recipes to machines may change. Since the batch size of each recipe on each machine may be different, if the workload balancing is done with batch size constraints, significant changes in γ may significantly impact the recipe-to-machine qualification configuration. This issue is discussed in Section 5.4.

5.2.2 Complexity Analysis

It is possible to show that the problem with F^{WIP} or F^{Time} as a criterion is NP-hard in the strong sense. This is because, even if all recipes are qualified on all machines, i.e. $Q_{r,m} = 1 \; \forall r, m$, then the problem can be simplified to the scheduling problem on parallel machines $P||C_{max}$, which is known to be NP-hard in the strong sense [31]. The transformation is done by noting that $F^{WIP} = 1$ or $F^{Time} = 1$ means that $\sum_{r=1}^{R} WIP_{r,m}$ is the same for each machine m, i.e. $\sum_{r=1}^{R} WIP_{r,m} = \frac{\sum_{r=1}^{R} WIP_{r,m}}{M}$. The decision problem "Is there a solution to the qualification problem such that $F^{WIP} = 1$ or $F^{Time} = 1$" can be transformed to solving a related scheduling problem $P||C_{max}$ and checking that $C_{max} = \frac{\sum_{r=1}^{R} WIP_{r,m}}{M}$.

5.3 Workload Balancing with Batch Size Restrictions

The algorithms for *real-valued* workload balancing for the WIP and Time Flexibility measures are discussed in [54] and in [53], respectively. Here, we discuss two types of algorithms for workload balancing for the WIP and Time Flexibility measures *with batch size constraints*.

The first type algorithms construct a full-batch solution directly. While, the second type algorithms, called *Batch-Feasibility Algorithms* takes the real-valued optimal solution and makes it feasible according to batch constraints.

In the next section, we compare the results of proposed algorithms with the standard commercial solver, IBM ILOG CPLEX 12.5.1.

5.3.1 Workload Balancing in terms of Production Volumes

In order to find the workload balance with batch size constraint on the toolset in terms of the production volume (WIP), we consider two main approaches. The first approach is a to solve the problem while considering batch sizes. The second approach consists of a heuristic which takes the optimal real-valued solution as

an initial solution and then, according to the batch size of each recipe-to-machine couple, generates a solution considering batch size.

For balancing the production volume on the toolset, we maximize (F^{WIP}) while distributing the production quantity of each recipe among the qualified machines for that recipe. The problem is formalized in (5.1).

The full-batch resolution is composed of two algorithms. The first algorithm constructs a real-valued or full-batch initial solution and the second algorithm progressively improves the initial solution to obtain either the optimal or a precise-enough near-optimal full-batch solution.

The algorithm used to construct the real-valued initial solution equally distributes the WIP of each recipe to their corresponding qualified machines. This yields a feasible and (almost always) non-optimal solution. The initial solution can be constructed according to the batch sizes or not.

The second (and the main) algorithm takes this initial solution as an input and improves it to find the optimal solution. The results in [81] show that the second algorithm gives the same final full-batch solution regardless of the initial solution, i.e. whether constructed according to batch size or not.

5.3.1.1 Full-Batch Algorithm

Here, we modify the algorithm described in [54] to consider batch sizes. Several strategies are possible which are described along the section. We may consider batch sizes from the beginning, i.e. while constructing the initial feasible solution or only when improving the initial feasible solution, constructed considering or ignoring batch sizes.

Constructing Full-Batch Initial Solution For calculating a real-valued initial feasible solution, we equally divide the WIP of each recipe among all of the qualified machines for that recipe. To consider batch sizes, in each loop, we start allocating the WIP by making as many batches as possible, i.e. $\lfloor \frac{WIP_{r,m}}{BS_{r,m}} \rfloor$. The new WIP allocation is composed of one or more batches which represent one or more production runs of recipe r on machine m.

Step 0. Set recipe index r to 1 ($r = 1$).

Step 1. For each machine in the toolset, determine the production volume allocation of recipe r to machines by equally distributing the production volume of recipe r (WIP_r) among all of the qualified machines for r. If for each of the qualified machines for recipe r, i.e. where $Q_{r,m} = 1$, the calculated $WIP_{r,m}$ is more than a full batch ($WIP_{r,m} \geq BS_{r,m}$) and forms an integer multiple of the batch size ($BS_{r,m}$), then attribute the WIP, else attribute only the integer part of the calculated $WIP_{r,m}$, i.e.$\lfloor \frac{WIP_{r,m}}{BS_{r,m}} \rfloor$. However, if the calculated $WIP_{r,m}$ is less than a full batch ($WIP_{r,m} \leq BS_{r,m}$), then allocate $WIP_{r,m}$ to the less loaded machine among the machines qualified for recipe r.

Step 2. Set $r = r + 1$, if $r \leq R$, then go to step 1.

Full-Batch Optimization Algorithm The algorithm begins with an initial solution S_0, generated by the initial WIP distribution algorithm. The initial solution is improved iteratively until the optimal or precise-enough near-optimal solution is obtained.

One case that often happens is that the overall production volume of each recipe does not correspond to an integer number of full batches, which is also difficult to consider from the beginning (when production planning is done). This is because the batch size of each recipe depends on the machine and we do not know exactly in advance how much production volume of which recipe is allocated to which machine considering that the machine statuses and qualification configuration change over time. In this case, even by attributing $WIP_{r,m}$ considering the batch size of recipe r on machine m ($BS_{r,m}$), the rest of the production volume for each recipe r is not equal to a production run, i.e. one full batch equivalent to $BS_{r,m}$. The question is to which of the qualified machines of recipe r, the remaining production volume must be assigned. We will allocate this volume to the less loaded machine among all qualified machines for recipe r. However it is better, if possible, to determine the production volumes from the beginning to form full batches. This is because processing only one wafer or a full batch of for instance 250 wafers will take the same

process time. Besides, if full batches cannot be made, in some toolsets, including Thermal Treatment toolset, *filler wafers* must be used to fill the vacant places in the carrying boats. By making full batches, we eliminate the cost of filler wafers and reduce the non productivity in the fab. Another case that must be considered is when the production quantity of a recipe is smaller than the smallest batch size of any machine qualified for this recipe. In this case, the production quantity of this recipe is produced only on the machine with the smallest total workload (WIP_m).

In summary, while allocating the workload, three cases may happen: (**a**) One (or several) full batch(es) and one real-valued WIP allocation; (**b**) One real-valued WIP allocation and no full batch and finally (**c**) One (or several) full batch(es).

Let us define by $\mathcal{LM}_r$ the set of *loading machines* for recipe r as the set of qualified machine(s) for r, with the smallest WIP quantity among all of the qualified machines for r, i.e.

$$m^* \in \mathcal{LM}_r \Rightarrow WIP_{m^*} = \min_{\forall m | Q_{r,m}=1} \{WIP_m\}.$$

Step 0. Start with initial solution S_0. The recipe index r and the iteration k are set to 1 ($r = 1$ and $k = 1$).

Step 1. Remove the WIP of recipe r initially allocated to machine m ($WIP_{r,m}$) from machine m, ($WIP_m = WIP_m - WIP_{r,m}$). Then, the total WIP quantities on the machines are sorted in a decreasing order.

Step 2. Calculate the WIP allocation quantities (WIP_{r,m^*}) if we had equally distributed the WIP of r on the loading machine(s) m^* in $\mathcal{LM}_r$ until the WIP quantity on one (or more) loading machine(s) m^* in $\mathcal{LM}_r$, (i.e. WIP_{r,m^*}) is equal to the WIP quantity of a machine m' qualified for recipe r and not in $\mathcal{LM}_r$, i.e. $WIP_{m^*} = WIP_{m'}$ with $m^* \in \mathcal{LM}_r$ and $m' \notin \mathcal{LM}_r$.

Step 3. WIP Re-Allocation

Step 3a. If $WIP_{r,m^*} > \lfloor \frac{WIP_{r,m^*}}{BS_{r,m^*}} \rfloor BS_{r,m^*}$, then there are one (or several) full-batch(es) and one real-valued WIP allocation. The full-batch

WIP allocation(s) is (are) equal to $WIP_{r,m^*} = \lfloor \frac{WIP_{r,m^*}}{BS_{r,m^*}} \rfloor BS_{r,m^*}$. The remainder ($WIP_r - \sum_r WIP_{r,m^*}$ where $WIP_{r,m^*} = \lfloor \frac{WIP_{r,m^*}}{BS_{r,m^*}} \rfloor BS_{r,m^*}$) is added in the last loop to the least-loaded loading machine in the loading machine set.

Step 3b. If $WIP_{r,m^*} < BS_{r,m^*}$, then there is no full batch. The real-valued WIP allocation is equal to WIP_{r,m^*}.

Step 3c. If $WIP_{r,m^*} = \lfloor \frac{WIP_{r,m^*}}{BS_{r,m^*}} \rfloor BS_{r,m^*}$, then there is (are) one (or several) full batch(es) which is (are) equal to WIP_{r,m^*} ($\lfloor \frac{WIP_{r,m^*}}{BS_{r,m^*}} \rfloor BS_{r,m^*}$).

Then, m' is added to the set of loading machines, i.e. $\mathcal{LM}_r \equiv \mathcal{LM}_r \cup \{m'\}$.
If $r < R$, then go to **Step 1**.

Step 4. If $F_k^{WIP} - F_{k-1}^{WIP} > 0$, then set $k = k + 1$, $r = 1$ and go to **Step 1**.

As the proposed algorithms are based on progressive improvement of an initial solution, it is not possible to implement the condition $F_k^{WIP} - F_{k-1}^{WIP} > 0$ of Step 3 in a strict manner. Therefore, the flexibility value of the previous solution S_{k-1} is compared with the flexibility value of the new solution S_k by defining an acceptable gap called ϵ. ϵ is a very small value, for example 1×10^{-20}. The algorithm iterates until precise-enough near-optimal solution is reached.

5.3.1.2 Batch-Feasibility Algorithm

When taking into account batch sizes, the problem becomes NP-hard in the strong sense for both flexibility measures as discussed in Section 5.2.2. This is why we propose a heuristic which takes as input the optimal real-valued workload balance, and makes it feasible by considering the batch size of each recipe on each machine. The resolution approach is different from the previous algorithms in the sense that we first choose the machine and then the recipe, whereas in the previous methods, a recipe is first chosen from which WIP values are assigned to its qualified machines. The advantages of our approach are that we already know the optimal solution of the real-valued problem and we have more flexibility in the problem

resolution. Furthermore, the real-valued solution serves as an upper bound of the full-batch problem solution.

Step 0. The optimal solution generated by the real-valued WIP distribution algorithm is the initial solution (S_0).

Step 1. In the whole toolset, find the most loaded machine ($m^\star$). From the set of all qualified recipes for this machine to which WIP values have been allocated ($Q_{r,m^\star} = 1$ and $WIP_{r,m^\star} > 0$), remove the load from the WIP allocation which is nearest (or in a second version: "farthest") to form a full batch. The removed value is equal to $\{\frac{WIP_{r,m^\star}}{BS_{r,m^\star}}\}$. Attribute this value to the less loaded machine which has already a *non-integer* WIP allocation for the same recipe (i.e. $\{\frac{WIP_{r,m^\star}}{BS_{r,m^\star}}\} > 0$). Note that this last condition (i.e. non-integer WIP allocation) has been added to avoid cycling.

Step 2. Once all the WIP allocations are redistributed (if necessary), the stop criterion is checked. If $F_k^{WIP} - F_{k-1}^{WIP} > 0$, then $k = k + 1$, $r = 1$ and go to step 1.

Two versions of the algorithm have been implemented. In the first version, at the beginning of Step 1, the preliminary workload on the toolset is ordered decreasingly and then all WIP attributions are done using this order. In the second version, the ordering is redone each time a WIP allocation is performed. The numerical results in this chapter are based on the second version which yields better results on average.

5.3.2 Workload Balancing in terms of Production Times

As with the production volumes, two approaches for workload balancing in terms of process times are presented. In the first approach, the batch constraint is considered from the beginning. The second approach is based on the same heuristic used for WIP workload balancing. *Batch-Feasibility Algorithm* takes the optimal real-valued load balancing solution of the Time Flexibility and makes it feasible with respect to the batch size constraint (see Model 5.1).

The *active-set method* has been used in [53] to find the optimal *real-valued* workload balance for the *Time Flexibility measure* (F^{Time}). The method is adapted below to consider the batch size constraint by limiting the step length.

5.3.2.1 Full-Batch Resolution: Adapted Active Set Method

In order to find the optimal workload balance for the Time Flexibility measure (F^{Time}), the optimization problem (5.1) must be solved when F^{Time} is maximized. Since C_{ideal} is constant, by maximizing F^{Time}, the sum of the production times (5.2) must be minimized.

$$\min \qquad f = \sum_{m=1}^{M} (\sum_{r=1}^{R} \frac{WIP_{r,m}}{TP_{r,m}})^{\gamma} \tag{5.2}$$

The essential part of the resolution algorithm is based on a line search method called *active-set method* described in Section 4.3.2.

In the active-set method, the solution is improved at each iteration k (corresponding to solution (x^k)). The solution improvement consists of finding a nonzero *search direction* (p^k) and a nonnegative *step length* α^k on the working surface to reach a new solution new solution (x^{k+1}). A solution in our problem corresponds to a WIP reallocation ($WIP^{k+1}_{r^*,m}$). The working surface is defined by the *working set* ($\mathscr{W}^k$) which corresponds to a subset of constraints which are binding (active). Here, we will restrict the step length (α^k) in a way not to violate the batch size constraint (BS_{r,m^*}). While moving on the working surface and defining the step length, one or several *blocking constraint(s)* may limit the step length. If one (or several) blocking constraint is encountered, it is added to the working set making a surface of lower dimension than before [68]. In our problem, the variables ($WIP_{r,m}$) are upper-bounded by their batch size ($BS_{r,m}$). Therefore, the variable becomes active, each time a bound constraint is encountered, i.e. $WIP^k_{r^*,m} = Run_{r,m}BS_{r^*,m}$ where $Run_{r,m} \in \mathbb{N}^+ \quad \forall r, m$ [35]. At each iteration, the Lagrangian multipliers (λ) of the working set are evaluated. If all are nonnegative, the optimal solution is obtained else we drop one or more of the constraints having a negative Lagrangian multiplier. Step length definition makes the algorithm not to violate the batch size constraint. As the method is already presented in Section 4.3.2, we only present the adapted active-set method, the decomposition, the gradient and the Hessian functions.

Instead of Equations (4.22), (4.24) and (4.25), Equations (5.3), (5.4) and (5.5) must be used, respectively.

$$f = \left(\frac{WIP_{r^*,m}}{TP_{r^*,m}} + \bar{f}_{r^*,m} \right)^{\gamma} \tag{5.3}$$

$$\nabla f = \frac{\gamma}{TP_{r^*,m}} \left(\frac{WIP_{r^*,m}}{TP_{r^*,m}} + \bar{f}_{r^*,m} \right)^{\gamma-1} \tag{5.4}$$

$$\nabla^2 f = \frac{\gamma\left(\gamma-1\right)}{\left(TP_{r^*,m}\right)^2} \left(\frac{WIP_{r^*,m}}{TP_{r^*,m}} + \bar{f}_{r^*,m} \right)^{\gamma-2} \tag{5.5}$$

Step 0. Start with an initial feasible solution x_0 and working set of active constraints at x^0 ($\mathscr{W}^0$). Set $k = 1$.

Step 1. The Optimality Test If $\bar{Z}^T \nabla f \left(WIP^k_{r^*,m} \right) = 0$

Step 1a. If there are no active constraints ($WIP^k_{r^*,m} = 0$), *Local Stationary Point* - **STOP**.

Step 1b. Else, compute Lagrangian Multipliers:

$$\bar{\lambda} = \bar{A}_r^T \nabla f \left(WIP^k_{r^*,m} \right). \tag{5.6}$$

Step 1c. If $\bar{\lambda} \geq 0$, *Local Stationary Point* - **STOP**. Else delete the constraint corresponding to the most negative Lagrangian multiplier from the working set $\mathscr{W}^k$. Update $\mathscr{W}^k$, $\bar{Z}$, $\bar{A}$ and $\bar{A}_r$.

Step 2. The Search Direction Compute a descent feasible search direction p^k with respect to the active constraints in $\mathscr{W}^k$. The *Reduced Newton Search Direction* (5.7) [50] is an efficient search direction.

$$p^k = -\bar{Z} \left(\bar{Z}^T \nabla^2 f \left(WIP^k_{r^*,m} \right) \bar{Z} \right)^{-1} \bar{Z}^T \nabla f \left(WIP^k_{r^*,m} \right) \tag{5.7}$$

Step 3. The Step Length Compute a step length α^k such that $f(WIP^k_{r^*,m} + \alpha^k p^k) < f(WIP^k_{r^*,m})$ subject to retaining feasibility with respect to all constraints, i.e. $\alpha^k \leq \bar{\alpha}^k$, where $\bar{\alpha}^k$ is the maximum feasible step length

along p^k.

The objective of this Step is to redistribute $WIP^k_{r^*,m}$ according to its corresponding batch size, i.e. $BS_{r^*,m}$. Three cases may happen:

Step a. If $WIP^k_{r^*,m} > \lfloor \frac{WIP^k_{r^*,m}}{BS_{r,m^*}} \rfloor BS_{r,m^*}$, then one (or several) full-batch allocation(s) and just one real-valued allocation is done. The step length α^k is defined in 5.8.

$$\alpha^k = \min_{i \notin \mathscr{W}^k, p^k < 0} \frac{\lfloor \frac{WIP^k_{r^*,m}}{BS_{r,m^*}} \rfloor BS_{r,m^*} - WIP^k_{r^*,m}}{p^k}. \tag{5.8}$$

Since the variable lies on its upper bound, the bound constraint is active. Therefore, $\mathscr{W}^k$ is modified to include the bound constraint to the working set ($\mathscr{W}^{k+1} \leftarrow \mathscr{W}^k \cup i$). Update $\bar{Z}$, $\bar{A}$ and $\bar{A}_r$.

Step b. If $WIP^k_{r^*,m} < BS_{r,m^*}$, then no full-batch allocation can be made and only a real-valued allocation is done. The step length α^k is defined in 5.9.

$$\alpha^k = \min_{i \notin \mathscr{W}^k, p^k < 0} \frac{-WIP^k_{r^*,m}}{p^k}. \tag{5.9}$$

[Step 3b-I.] If $\alpha^k < \bar{\alpha}^k$, α^k is an unconstrained step, i.e. no blocking constraint is encountered and x^{k+1} remains feasible ($x^{k+1} = x^k + \alpha^k p^k$).

[Step 3b-II.] If $\alpha^k \geq \bar{\alpha}^k$, then the step along p^k is blocked by one (or more) (inactive) constraint(s) not in $\mathscr{W}^k$ but which are active in x^{k+1}. In this case, $\mathscr{W}^k$ is modified to include the blocking constraint to the working set ($\mathscr{W}^{k+1} \leftarrow \mathscr{W}^k \cup i$). Update $\bar{Z}$, $\bar{A}$ and $\bar{A}_r$.

Step c. If $WIP^k_{r^*,m} = \lfloor \frac{WIP^k_{r^*,m}}{BS_{r,m^*}} \rfloor BS_{r,m^*}$, then one (or several) full-batch allocation(s) are done. The step length α^k is zero. This can be verified by considering 5.8. Since the variable lies on its upper bound, the bound constraint is active. Therefore, $\mathscr{W}^k$ is modified to include the bound constraint to the working set ($\mathscr{W}^{k+1} \leftarrow \mathscr{W}^k \cup i$). Update $\bar{Z}$, $\bar{A}$ and $\bar{A}_r$.

Set $k = k + 1$ and go to **Step 1**.

5.3.2.2 Batch-Feasibility Algorithm

When taking into account batch sizes, the problem becomes NP-hard in the strong sense for both flexibility measures as discussed in Section 5.2.2. That is why we propose a heuristic which takes as input the optimal real-valued workload balance, and makes it feasible by considering the batch size of each recipe on each machine. The resolution approach is different from the previous algorithms in the sense that we first choose the machine and then the recipe, whereas in the previous methods, a recipe is first chosen from which WIP values are assigned to its qualified machines. The advantages of this approach are that as the real-valued optimal solution of the problem is known, more flexibility exists in the problem resolution. Furthermore, the real-valued solution serves as an upper bound of the full-batch problem solution.

Step 0. Start with the optimal real-valued solution (S_0). Set the solution index to one ($k = 1$).

Step 1. Order the machines in the toolset according to their workload in terms of production times in a decreasing order. Choose the most loaded machine ($m^\star$) (potential contributor machine). From the set of all qualified recipes for this machine to which WIP values have been allocated ($Q_{r,m^\star} = 1$ and $WIP_{r,m^\star} > 0$), remove the load from the WIP allocation which is nearest to form a full batch (if any). The removed value is equal to $\{\frac{WIP_{r,m^\star}}{BS_{r,m^\star}}\}$.

Step 2. Attribute this value to the less loaded machine which has already a *not full batch* WIP allocation for the same recipe (i.e. $\{\frac{WIP_{r,m^\star}}{BS_{r,m^\star}}\} > 0$). Note that this last condition (i.e. non-integer WIP allocation) is added to avoid cycling.

Re-order the machines in the toolset according to their workload in a decreasing order.

Repeat **Step 1** and **Step 2** until all WIP allocations are redistributed (if necessary).

Step 3. If $F_k^{WIP} - F_{k-1}^{WIP} > 0$, then update the solution index ($k = k + 1$) and go to **Step 1**.

5.4 Numerical Experiments

The quality of the solution provided by our approaches and the impact of the batch size constraint are studied in this section. As the concept is industrialized, we use 25 industrial instances of Soitec for the experimentation.

5.4.1 Comparing with Exact Solutions

The commercial solver IBM ILOG CPLEX 12.5.1 has been used to evaluate the proposed algorithms. However, IBM ILOG CPLEX can only solve quadratic programs and not other types of non-linear programming models. Therefore, γ is set to 2 in this section. For WIP flexibility, the optimal solution is obtained by minimizing the denominator $(\sum_{m=1}^{M}(\sum_{r=1}^{R} WIP_{r,m})^{\gamma})$. Then, F^{WIP} is calculated. The same approach is used for F^{Time}. The best flexibility values obtained by our methods are bold-faced.

WIP Flexibility Measure

Table 5.1 presents the WIP flexibility values for each instance using our resolution approaches and IBM ILOG CPLEX for $\gamma = 2$. Note that, because of the small value of γ, the WIP flexibility measures are relatively big and very close for each instance. For all instances (except one), the full-batch algorithm (Section 5.3.1.1) outperforms the batch-feasibility heuristic. Moreover, the WIP flexibility values of the full-batch algorithm are very close to the optimal solutions.

Time Flexibility Measure

Table 5.2 shows that the adapted active-set method (Section 5.3.2.1) yields very good solutions that are close to the optimal ones. The batch-feasibility heuristic

Table 5.1: WIP flexibility with $\gamma = 2$ for the current toolset recipe-to-machine qualification configuration

Instance	Real-value	Full batch	Batch feasibility	CPLEX
1	95.10%	**95.05%**	94.92%	95.10%
2	94.89%	**90.55%**	90.35%	90.58%
3	96.99%	**96.97%**	96.83%	96.99%
4	95.53%	**94.26%**	94.06%	95.53%
5	87.94%	**82.47%**	82.42%	83.94%
6	99.41%	**94.85%**	94.66%	94.89%
7	76.59%	**76.34%**	76.32%	76.59%
8	88.58%	**67.23%**	67.08%	67.49%
9	96.26%	68.50%	**68.55%**	68.75%
10	99.81%	**99.03%**	96.00%	99.81%
11	96.60%	**96.02%**	94.49%	96.60%
12	88.72%	**88.58%**	86.69%	88.72%
13	96.31%	**95.12%**	91.33%	96.31%
14	90.80%	**86.14%**	85.22%	86.48%
15	94.80%	**94.25%**	92.83%	94.80%
16	92.60%	**87.27%**	84.92%	88.19%
17	95.66%	**90.29%**	87.55%	91.10%
18	90.92%	**85.94%**	83.39%	86.59%
19	93.84%	**84.65%**	83.12%	84.88%
20	84.42%	**76.38%**	75.29%	76.74%
21	86.31%	**81.75%**	79.89%	82.39%
22	83.67%	**82.64%**	81.75%	83.67%
23	85.43%	**85.17%**	83.53%	85.43%
24	90.08%	**89.73%**	88.48%	90.08%
25	89.57%	**89.08%**	85.69%	89.57%

(Section 5.3.2.2) also yields good results but not as good as the adapted active-set method.

Table 5.2: Time flexibility with $\gamma = 2$ for the current toolset recipe-to-machine qualification configuration

Instance	Real-value	Batch feasibility	Adapted Active-Set	CPLEX
1	84.79%	84.51%	**84.63%**	84.67%
2	83.02%	82.74%	**82.85%**	83.01%
3	98.06%	**97.92%**	96.54%	97.96%
4	99.56%	**98.42%**	92.91%	98.92%
5	89.87%	**89.70%**	89.62%	89.70%
6	97.24%	**96.97%**	94.94%	97.03%
7	98.58%	**97.82%**	**97.82%**	97.95%
8	82.90%	78.35%	**82.45%**	82.45%
9	76.67%	75.64%	**76.44%**	76.44%
10	97.57%	91.41%	**96.34%**	97.14%
11	99.86%	99.39%	**99.54%**	99.79%
12	96.68%	95.50%	**96.31%**	96.31%
13	93.60%	**92.67%**	90.74%	92.90%
14	95.92%	**92.36%**	91.46%	92.36%
15	99.65%	95.63%	**98.88%**	99.09%
16	96.83%	91.83%	**96.14%**	96.38%
17	95.21%	92.57%	**94.80%**	95.01%
18	96.77%	94.26%	**96.57%**	96.74%
19	94.82%	92.89%	**93.54%**	93.85%
20	96.50%	91.76%	**94.66%**	96.31%
21	96.16%	**92.96%**	92.06%	92.96%
22	99.86%	97.27%	**97.85%**	97.85%
23	99.14%	94.18%	**98.18%**	98.86%
24	99.20%	96.55%	**98.80%**	99.11%
25	98.31%	95.22%	**98.23%**	98.25%

5.4.2 Impact of Batch Size Constraint on Qualification Decisions

In the industrial decision support system, our methods are implemented in VBA using Microsoft Excel. For the experiments of this section, γ is set to 6. The choice of γ is based on historical experiments which corresponds the best to the actual toolset workload in the fab. The WIP and time flexibility measures for the

current recipe-to-machine qualification configuration and the qualification configuration after the best qualification are presented in Tables 5.3 and 5.5. As the optimal real-valued solutions for both WIP and time flexibility measures are available, they are used as upper bounds to evaluate the quality of our algorithms with batch size constraints. Tables 5.4 and 5.6 present the qualification choice correspondence using the full-batch algorithms compared to the real-valued solutions for the WIP and time flexibility measures.

WIP Flexibility Measure

Table 5.3 shows the WIP flexibility value for the current recipe-to-machine configuration and the WIP flexibility for the new configuration after the best new qualification. The second and third columns show the optimal real-valued WIP flexibility values for all instances. The fourth and fifth columns show the WIP flexibility values taking batch sizes into account using the full-batch algorithm. The sixth and seventh columns show the flexibility measures taking batch sizes into account using the batch-feasibility heuristic. When looking at Table 5.3 and although, in most cases, the WIP flexibility is very close with or without batch size constraints, the difference is quite large in Instance 12 (from 79.24% down to 64.54% for the current configuration and from 93.88% down to 75.30% for the configuration after the best qualification). This result could be due to the fact that our solution approaches considering batch sizes do not necessarily find the optimal solution. Table 5.3 also shows that the full-batch algorithm strongly dominates the batch-feasibility heuristic.

An important question is whether the best qualification proposed when solving the problem with real-valued variables is also relevant if the problem is solved with batch size constraints. Table 5.4 compares the WIP flexibility measure for the best qualification proposed by our batch solution approaches with the best qualification proposition obtained using the real-valued resolution but solved with our batch solution approaches. This shows the flexibility evaluation deviation if we had applied the qualification decision without considering batch sizes.The columns “Best” contain the WIP flexibility measure for the recipe-to-machine configuration after the best (new) qualification. The columns “Best (RV)” show the WIP flexibility measure for the recipe-to-machine configuration after the best (new) qualification proposed

	Real value		Full batch		Batch feasibility	
Instance	Current	Best	Current	Best	Current	Best
1	74.43%	87.69%	**74.22%**	**86.53%**	72.20%	85.46%
2	53.88%	67.55%	**53.70%**	**67.22%**	50.47%	65.86%
3	68.83%	78.13%	**68.54%**	**77.61%**	66.60%	76.37%
4	53.49%	70.88%	**53.29%**	**70.66%**	52.86%	69.81%
5	5.08%	33.01%	**5.07%**	**31.46%**	**5.07%**	5.09%
6	72.73%	98.33%	**72.28%**	**96.96%**	67.99%	93.69%
7	7.25%	14.78%	**7.04%**	**14.23%**	6.81%	14.16%
8	11.01%	20.98%	**10.56%**	**20.41%**	10.32%	19.55%
9	97.17%	97.17%	**90.33%**	**91.72%**	45.97%	64.39%
10	61.76%	62.18%	56.58%	**60.88%**	**57.44%**	59.87%
11	29.09%	33.83%	**28.88%**	**32.87%**	27.09%	31.63%
12	79.24%	93.88%	**64.54%**	**75.30%**	40.24%	75.11%
13	25.30%	53.03%	**24.41%**	**49.36%**	22.96%	45.80%
14	49.96%	49.96%	**47.69%**	**49.01%**	45.87%	47.28%
15	32.99%	36.86%	**32.08%**	33.05%	20.22%	**33.61%**
16	44.81%	69.53%	**39.87%**	**63.63%**	32.87%	46.71%
17	28.19%	47.20%	**24.73%**	**44.62%**	22.11%	29.28%
18	31.11%	38.50%	**27.01%**	**37.32%**	15.84%	26.75%
19	15.01%	24.72%	**14.77%**	**23.75%**	13.41%	16.56%
20	20.17%	37.41%	**19.22%**	**35.13%**	17.46%	32.51%
21	19.62%	19.78%	16.95%	18.76%	**18.33%**	**19.14%**
22	17.83%	21.86%	**17.37%**	21.34%	16.57%	**21.35%**
23	32.80%	33.16%	**31.55%**	**32.23%**	25.69%	28.69%
24	30.24%	32.63%	26.05%	**31.22%**	**27.16%**	30.05%
25	46.75%	68.86%	**45.56%**	**67.67%**	43.73%	66.40%

Table 5.3: Comparing solution approaches for WIP flexibility measure ($\gamma = 6$).

when using the (optimal) real-valued solution approach. The columns "Match" indicate whether the next (new) qualification of the corresponding approach is the same as the (new) qualification proposition using the (optimal) real-valued resolution approach. It is interesting to note that there are multiple instances for which the proposed qualifications differ and with substantial differences in terms of WIP flexibility, such as Instances 5 (31.46% and 0.70%), 9 (91.72% and 83.59%) and

12 (75.30% and 69.25%). This illustrates that not considering batch sizes may induce less effective qualification decisions. Note that the full-batch algorithm is quite efficient, considering the (relatively small) gap with the results of the real-valued resolution. As a conclusion, among the workload balancing algorithms and when considering batch sizes for the WIP flexibility measure, the full-batch algorithm is the most efficient.

	Full batch			Batch feasibility		
Instance	Best	Best (RV)	Match	Best	Best (RV)	Match
1	86.53%	86.53%	*	85.46%	85.46%	*
2	67.22%	67.22%	*	65.86%	65.86%	*
3	77.61%	77.61%	*	76.37%	74.56%	
4	70.66%	70.66%	*	69.81%	69.81%	*
5	31.46%	0.70%		5.09%	3.41%	
6	96.96%	96.96%	*	93.69%	93.69%	*
7	14.23%	14.07%		14.16%	14.09%	
8	20.41%	20.41%	*	19.55%	19.55%	*
9	91.72%	83.59%		64.39%	46.92%	
10	60.88%	60.88%	*	59.87%	59.72%	
11	32.87%	32.87%	*	31.63%	31.63%	*
12	75.30%	69.25%		75.11%	66.39%	
13	49.36%	49.36%	*	45.80%	45.80%	*
14	49.01%	47.31%		47.28%	46.83%	
15	33.05%	33.05%	*	33.61%	33.61%	*
16	63.63%	63.63%	*	46.71%	46.71%	
17	44.62%	44.62%	*	29.28%	29.28%	*
18	37.32%	37.32%	*	26.75%	21.62%	*
19	23.75%	23.75%	*	16.56%	13.80%	
20	35.13%	35.13%	*	32.51%	29.29%	
21	18.76%	18.13%		19.14%	18.42%	
22	21.34%	21.34%	*	21.35%	21.07%	
23	32.23%	32.23%	*	28.69%	26.79%	
24	31.22%	31.22%	*	30.05%	24.75%	
25	67.67%	67.54%		66.40%	64.56%	

*: Proposed qualifications are identical.

Table 5.4: Comparing best qualifications for WIP flexibility measure.

Time Flexibility Measure

Looking at Table 5.5, with $\gamma = 6$, the adapted active-set method still yields considerably better results than the batch-feasibility heuristic. In most cases where the batch-feasibility heuristic gives better results, the difference is negligible. For example, in the first instance, the current flexibility using the batch-feasibility heuristic is 40.43% while the result obtained by the adapted active-set method is 40.40%. To get better results, it would be interesting to calculate the time flexibility considering the batch sizes using both approaches and choose the best solution. However, this would be very time consuming.

The batch-feasibility heuristic is less efficient with higher values of γ. In the real-valued resolution, by increasing gamma, we increase the number of recipe splitting while allocating workload. As the batch-feasibility heuristic "blindly" re-allocates the production volumes to make full batches (without considering the workload of the machines based on their process times), the efficiency of the algorithm decreases. Therefore, for high values of γ, it is preferable to solve the problem in one shot using the adapted active-set method.

The results of Table 5.5 show that considering batch sizes may lead to large difference in the time flexibility value of the toolset. This is the case for many instances (Instances 7 to 12 and 14 to 25). Table 5.6 shows the qualification matches using the real-valued resolution and the full-batch resolution. As it can be seen, in most cases, the qualification propositions do not match. But what would have happened if we had chosen the real-valued qualification proposition? According to Table 5.6, it could have led to a large misevaluation of the time flexibility. Some instances (7 to 9, 19, 22, 24 and 25) illustrate this phenomenon. Our numerical experiments on industrial data shows that, by ignoring batch sizes, wrong qualification decisions may be taken, leading to potential capacity loss.

5.5 Conclusions and Perspectives

Batching is an important characteristic of various workshops in the semiconductor manufacturing industry. In this chapter, we have proposed different workload

Table 5.5: Comparing solution approaches for the Time flexibility measure.

Instance	Real value		Batch feasibility		Adapted Active-Set	
number	Current	Best	Current	Best	Current	Best
1	40.91%	72.68%	**40.43%**	66.89%	40.40%	**69.78%**
2	34.29%	49.75%	32.83%	45.60%	**33.79%**	**48.08%**
3	85.53%	93.43%	**82.92%**	88.27%	80.54%	**90.93%**
4	33.89%	38.44%	32.94%	36.84%	**33.10%**	**37.45%**
5	34.42%	63.71%	33.44%	61.90%	**33.48%**	**61.94%**
6	82.58%	94.82%	**80.03%**	**92.72%**	79.29%	92.49%
7	98.00%	99.99%	62.11%	81.03%	**66.91%**	**85.87%**
8	29.34%	41.40%	13.37%	**38.72%**	**25.68%**	37.24%
9	28.02%	49.44%	**17.36%**	**34.27%**	17.16%	31.52%
10	67.82%	85.11%	**20.89%**	**49.74%**	18.36%	49.12%
11	99.22%	99.55%	43.64%	49.50%	**47.00%**	**66.12%**
12	81.14%	96.45%	49.58%	57.28%	**67.04%**	**69.57%**
13	48.91%	68.81%	**45.06%**	49.68%	45.01%	**53.91%**
14	70.41%	97.50%	29.66%	47.30%	**38.26%**	**48.66%**
15	99.64%	99.96%	37.27%	**58.74%**	**55.10%**	58.68%
16	74.42%	88.54%	**27.43%**	**27.72%**	26.28%	27.43%
17	61.70%	76.55%	**13.68%**	32.44%	13.09%	**57.54%**
18	75.33%	92.15%	38.60%	52.36%	**44.72%**	**58.34%**
19	74.70%	83.46%	**26.95%**	51.60%	**26.95%**	**64.95%**
20	69.94%	84.04%	54.98%	61.85%	**59.28%**	**63.86%**
21	71.49%	85.33%	13.32%	41.37%	**33.40%**	**46.46%**
22	95.20%	98.17%	45.70%	59.68%	**47.06%**	**71.61%**
23	97.70%	99.13%	42.96%	**78.09%**	**68.30%**	72.30%
24	96.03%	98.01%	**58.23%**	68.45%	58.06%	**73.84%**
25	92.74%	95.00%	26.52%	55.44%	**29.82%**	**61.93%**

balancing algorithms (in terms of production quantities and production times) considering batch sizes. The workload allocations are used in qualification management. Using real fab data, we have shown that ignoring batch size constraints may lead to inappropriate qualification decisions which in turn leads to a non-optimal toolset capacity utilization.

Various perspectives are possible. Having near-optimal recipe-to-machine WIP

Instance	Batch feasibility			Adapted Active-Set		
number	Best	Best (RV)	Match	Best	Best (RV)	Match
1	66.89%	66.89%	*	69.78%	69.78%	*
2	45.60%	45.60%	*	48.08%	48.08%	*
3	88.27%	88.27%	*	90.93%	90.93%	*
4	36.84%	36.84%	*	37.45%	37.45%	*
5	61.90%	61.90%	*	61.94%	61.94%	*
6	92.72%	92.72%	*	92.49%	92.49%	*
7	81.03%	78.64%		85.87%	68.68%	
8	38.72%	25.56%		37.24%	29.21%	
9	34.27%	19.42%		31.52%	18.19%	
10	49.74%	49.74%	*	49.12%	49.12%	*
11	49.50%	49.19%		66.12%	60.73%	
12	57.28%	52.51%		69.57%	68.57%	
13	49.68%	45.28%		53.91%	47.21%	
14	47.30%	47.30%	*	48.66%	48.66%	*
15	58.74%	58.74%	*	58.68%	57.70%	
16	27.72%	27.46%		27.43%	26.60%	
17	32.44%	25.99%		57.54%	57.54%	*
18	52.36%	52.36%	*	58.34%	58.34%	*
19	51.60%	27.04%		64.95%	37.47%	
20	61.85%	57.83%		63.86%	60.58%	
21	41.37%	26.21%		46.46%	46.46%	*
22	59.68%	48.69%		71.61%	49.62%	
23	78.09%	67.44%		72.30%	71.57%	
24	68.45%	63.22%		73.84%	64.23%	
25	55.44%	36.00%		61.93%	30.93%	

*: Proposed qualifications are identical.

Table 5.6: Comparing best qualifications for Time flexibility measure.

allocations considering batch sizes, it may be possible to use the workload balancing allocations as an input for a scheduler or a dispatcher of lots. The batch-feasibility heuristic might be improved by taking the flexibility measure value (objective function value) into account. This is achievable by adding another condition while workload re-allocating. This means that we calculate the flexibility value if we have

had distributed the (removed) WIP for the machines of the same recipe having a non-integer WIP value. And if several alternatives exist, the WIP allocation is chosen which yields the largest flexibility measure.

A general perspective for qualification management is to propose flexibility measures which consider the yield for each recipe on each machine and propose relevant qualifications taking this parameter into account. Another aspect which may be considered directly or indirectly in flexibility measures is recipe priorities, i.e. more priority could be given to the recipes related to more important products or "hot lots". Auxiliary resource management (for example masks in the lithography toolset) is also a subject related to qualification management which can enrich future studies.

Chapter 6

Qualification Management and Workload Variability

Variability is an inherent component of all production systems. To prevent variability propagation through the whole production line, variability must be constantly monitored, especially for bottleneck toolsets. In this chapter, we propose measures to evaluate workload variability for a toolset configuration. Using industrial data, we show how making the toolset configuration more flexible by qualifying products on machines decreases variability. By quantifying the toolset workload variability, our variability measures makes it possible to estimate the variability reduction associated to each new qualification. The industrial results show significant workload variability reduction and capacity improvement.[1]

[1]Part of this chapter has been presented in Winter Simulation Conference 2014 [83].

6.1 Introduction and Motivation

The dynamic business environment in semiconductor manufacturing industry calls for an increasing product-mix flexibility. Besides, production systems are restricted by machine capabilities. The ultimate goal of every manufacturing system is to satisfy customer demand while making the best possible use of the production facilities.

Variability is the uncertainties and variations in the production flow. Variability has *stochastic* and *deterministic* sources. The stochastic variability sources are uncontrollable, and the most well-known are demand, machine breakdown, rework, operator delay, etc. On the other hand, deterministic variability sources are controllable. They include machine eligibility, batches, setups, re-entrant flows, etc. The controllable deterministic variability source considered in this chapter is *product-to-machine eligibility*, called *qualification*.

Production variability influences production planning and scheduling, capacity planning, inventory management, equipment and labor cost, etc. [62]. Even little variability in bottleneck workcenters can cause high variability in the whole production line. Therefore, production variability reduction at bottleneck workcenters is crucial to prevent variability propagation to the whole production line. Production variability leads to loss of capacity. Therefore, due to expensive equipment cost, mastering variability is critical in semiconductor manufacturing. Production variability decreases as manufacturing systems become more flexible. The flexibility of a manufacturing system is determined based on to which extent a product can be allocated to a machine. Hence, capacity allocation determines the flexibility of a production system [71].

This chapter is a continuation of our investigations about qualification management and production flexibility. In the previous chapters, variability is not explicitly considered. This is why, in this chapter, we aim at reducing production variability by defining variability measures. The workload variations in a single workcenter in one period are considered. We want to show how variability is reduced when the flexibility of the toolset is increased by performing new qualifications. The variability measures proposed can be used to measure the impact of batch or single

wafer processing. However additional constraints must be added to the optimization model. Chapter 5 specifically address the problem of batch processing. Through experiments with industrial data, we illustrate that production variability and manufacturing flexibility are two sides of the same coin.

This chapter is organized as follows. Section 6.2 details the framework of the study. The measures that we propose to evaluate the workload variability are presented in Section 6.3. Some extensions and resolution approaches to the problem are presented in 6.4. In Section 6.5, numerical experiments on industrial data are discussed. Finally, Section 6.6 provides conclusions and proposes future investigation paths.

6.2 Toolset Workload Variability and Manufacturing Flexibility

Recipe-to-machine qualification restrictions are comparable to product-to-machine eligibility constraints. The qualification concept is also alike the configuration concept. Reconfigurable manufacturing systems have more flexibility. By qualifying a *qualifiable* recipe, the toolset manufacturing flexibility increases. Figure 6.1 shows three configurations of a toolset. Figure 6.1a shows dedicated machines to recipes. This dedicated strategy is sometimes referred to as no flexibility. In Figure 6.1b, each recipe is qualified on at least two machines. The setting is called partial or sparse flexibility. Figure 6.1c depicts a totally flexible toolset where all recipes are qualified on all machines. For us, we reach full-flexibility when all machines are equally loaded based on its capacity. So, it is not absolutely necessary to qualify all recipes on all machines to attain full flexibility. Fortunately this is not the case, because qualifying all recipes on all machines is firstly not always possible and then not economically feasible. This is the miracle of flexibility that we gain a lot by no or little effort. Note that "theoretically", we may reach flexibility even in case where one recipe is qualified only on one machine, i.e. the dedicated strategy. However this case is scarce in industry. In the next section, we define variability measures to evaluate to which extent the increase of the manufacturing flexibility contributes to

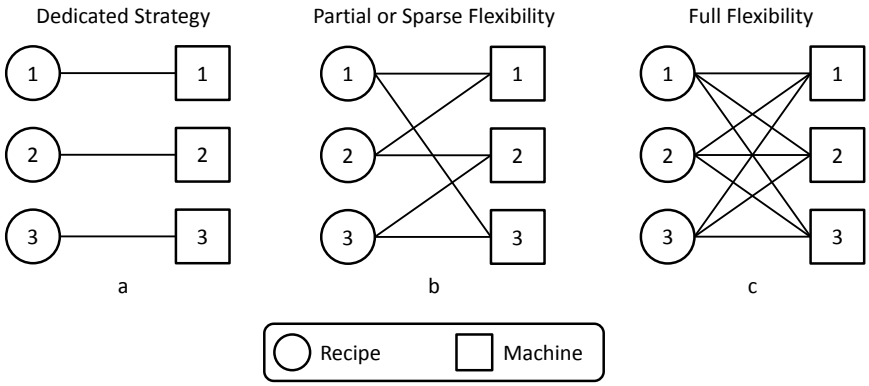


Figure 6.1: Dedicated Strategy (a), Partial or Sparse Flexibility (b) and Full Flexibility (c) Recipe-to-Machine Configurations (adapted from [38] and [71])

workload variability reduction.

Increasing manufacturing flexibility implies at least one new qualification, enabling the qualified machine to process another recipe by creating a new link between the qualified recipe-machine couple. As already specified in the introduction, the notion of *manufacturing flexibility* in this study refers to [54], where *flexibility measures* are defined to evaluate the flexibility gain associated with performing qualifications. One of the main objectives of their measures is to estimate the impact of qualifications on workload balancing. We have further discussed these measures in previous chapter and in the remainder of this part. They are used to evaluate new qualifications in the daily fab operations. However, by only considering these flexibility measures, the impact of a qualification on the toolset workload variability is not clear to decision makers. This study aims at throwing light on the other side of qualification management via variability measurement. Moreover, compared to [54], we also consider machine capacity restrictions.

6.3 Variability Measures

In order to obtain a unique variability value, the optimal workload balance of the toolset according to its configuration must be calculated. By minimizing the proposed measures, which are inspired from statistical moments, the optimal workload

balance of a toolset must be calculated.

The variability measures introduced in the next sections (Var^{Time}_{Uncapa} in Section 6.3.1, Var^{Time}_{Capa} in Section 6.3.2, and $Var^{Time}_{Capa\neq}$ in Section 6.3.3) are used as the objective function ($Var^{\bullet}_{\bullet}$) of the mathematical model below. The only set of constraints (6.1b) of the model guarantees that the production volume of recipe r is only allocated to machines that are qualified for r, i.e. machines m such that $Q_{r,m} = 1$. By minimizing the selected variability measure subject to the set of constraints, the optimal toolset workload balance is obtained.

$$\min \quad Var^{\bullet}_{\bullet} \tag{6.1a}$$

$$\text{Subject to} \quad \sum_{m=1|Q_{r,m}=1}^{M} WIP_{r,m} = WIP_r \qquad \forall r \tag{6.1b}$$

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

This model can be solved by adapting the Active Set method described in [53] and Chapter 4. In Section 6.4, the resolution is briefly addressed.

6.3.1 Uncapacitated Time Variability Measure (Var^{Time}_{Uncapa})

For the current toolset qualification configuration, by minimizing the sum of the total process time of each machine (6.2), the workload variability is calculated. Then, for each qualifiable recipe-to-machine couple, we re-calculate the variability (Var^{Time}_{Uncapa}) by virtually qualifying the qualifiable couple. In order to evaluate the variability reduction associated with each new qualification, the variability of each new configuration is subtracted from the variability of the initial configuration.

$$Var^{Time}_{Uncapa} = \sum_{m=1}^{M} (C_m)^{\gamma} \tag{6.2}$$

By increasing the workload balancing exponent (γ), the load of high speed machines is shifted to slower machines where qualification is allowed. Increasing γ creates a smoother workload distribution on the toolset.

6.3.2 Capacitated Time Variability Measure (Var^{Time}_{Capa})

Machine failures, operator unavailability, scheduled and unscheduled maintenance are sources of variability which affect the uptime of machines. The uptimes of each machine in the same toolset can be different and are considered to be deterministic in this paper. Var^{Time}_{Capa} (6.3) evaluates the workload variability of a toolset while considering the capacity of machines.

$$Var^{Time}_{Capa} = \sum_{m=1}^{M} (C_m - Capa_m)^{\gamma} \tag{6.3}$$

In order to calculate the variability reduction of each new qualification, as explained in Section 6.3.1, the variability of the current qualification configuration and the configuration after each new qualification must be calculated.

By increasing the workload balancing exponent (γ), the model tries to fit better the toolset workload to the available capacity by shifting workload from overloaded machines to less loaded machines.

6.3.3 Weighted Capacitated Time Variability Measure ($Var^{Time}_{Capa\neq}$)

As in Section 4.2.4 for flexibility measures, we propose a *Weighted Capacitated Time Variability Measure ($Var^{Time}_{Capa\neq}$)* in which overloaded machines and underloaded machines are linearly and non-linearly penalized. The linear penalizers of a machine m are o_m and u_m and nonlinear penalizers correspond to different workload balancing exponents γ_o and γ_u.

$$Var^{Time}_{Capa\neq} = \sum_{m=1|C_m>Capa_m}^{M} o_m(C_m - Capa_m)^{\gamma_o} + \sum_{m=1|C_m\leq Capa_m}^{M} u_m(C_m - Capa_m)^{\gamma_u}$$

The qualification that reduces the most the production variability depends on the values chosen for the penalizers (o_m, u_m, γ_o and γ_u).

6.4 Resolution Approaches

Model 6.1 can be solved by adapting the resolution approach based on [53] and [39] and described in Section 4.3.2. Only a summary of the functions and the necessary changes in the resolution approach are mentioned below.

6.4.1 Capacitated Time Variability Measure (Var^{Time}_{Capa})

$$Var^{Time}_{Capa} = \sum_{m=1}^{M} (C_m - Capa_m)^{\gamma} \tag{6.5}$$

Equations (4.22), (4.24) and (4.25) are respectively replaced by (6.6), (6.7) and (6.8).

$$f = (C_m - Capa_m)^{\gamma} + \bar{f}_m \tag{6.6a}$$

$$f = \left(\frac{WIP_{r^*,m}}{TP_{r^*,m}} - Capa_m\right)^{\gamma} + \bar{f}_m \tag{6.6b}$$

$$\nabla f = \frac{\gamma}{TP_{r^*,m}} (C_m - Capa_m)^{\gamma-1} \tag{6.7}$$

$$\nabla^2 f = \frac{\gamma(\gamma-1)}{(TP_{r^*,m})^2} (C_m - Capa_m)^{\gamma-2} \tag{6.8}$$

6.4.2 Weighted Capacitated Time Variability Measure ($Var^{Time}_{Capa\neq}$)

$$Var^{Time}_{Capa\neq} = \sum_{m=1|C_m>Capa_m}^{M} o_m (C_m - Capa_m)^{\gamma_o} + \sum_{m=1|C_m\leq Capa_m}^{M} u_m (C_m - Capa_m)^{\gamma_u} \tag{6.9}$$

Equations (4.22), (4.24) and (4.25) are respectively replaced by:

$$f = \begin{cases} o_m (C_m - Capa_m)^{\gamma_o} + \bar{f}_m & \text{if } C_m > Capa_m \\ u_m (C_m - Capa_m)^{\gamma_u} + \bar{f}_m & \text{if } C_m \leq Capa_m \end{cases}$$

$$\nabla f = \begin{cases} \frac{o_m \cdot \gamma_o}{TP_{r^*,m}} (C_m - Capa_m)^{\gamma_o - 1} & \text{if } C_m > Capa_m \\ \frac{u_m \cdot \gamma_u}{TP_{r^*,m}} (C_m - Capa_m)^{\gamma_u - 1} & \text{if } C_m \leq Capa_m \end{cases}$$

$$\nabla^2 f = \begin{cases} \frac{o_m \cdot \gamma_o (\gamma_o - 1)}{(TP_{r^*,m})^2} (C_m - Capa_m)^{\gamma_o - 2} & \text{if } C_m > Capa_m \\ \frac{u_m \cdot \gamma_u (\gamma_u - 1)}{(TP_{r^*,m})^2} (C_m - Capa_m)^{\gamma_u - 2} & \text{if } C_m \leq Capa_m \end{cases}$$

6.5 Numerical Experiments

In this section, we study the concept of workload variability by conducting experiments on industrial data of a Thermal Treatment Toolset in an "SOI" (Silicon-On-Insulator) production line. The Thermal Treatment Toolset, which consists of non-homogeneous furnaces, is a bottleneck toolset. Each standard SOI product must at least visit three times this toolset. For the industrial experiments, the value of γ is set to 4. Careful observations of the shop floor workload allocation shows that any value between 4 to 6 suits the model for practical purposes. As the capacity of the machines are different, we use the *Capacitated Time Variability Measure* (Var^{Time}_{Capa}) (6.3) as the objective function of the mathematical model.

First, we consider a data set for one period and study the impact of performing a single qualification on the percentage of the workload variability reduction. Other independent performance indicators used to interpret the workload balance are: The overall toolset workload variation percentage (6.10), overload (6.11) and unused capacity (underload) (6.12) variation percentages. The variation comparison for each performance indicator is simply calculated as shown in (6.13) for overload (OL) variation comparison. Using the static workload balance diagram, we show the impact of one new qualification on the workload variability. Finally, we discuss the impact of new qualifications on the reduction of the production variability for some industrial instances taken from the daily fab operations.

$$Workload\ Sum = \sum_{m=1}^{M} C_m \tag{6.10}$$

$$Overload\ Sum = \sum_{m=1|C_m \geq Capa_m}^{M} (C_m - Capa_m) \tag{6.11}$$

$$Unused\ Capacity\ Sum = \sum_{m=1|C_m \leq Capa_m}^{M} (C_m - Capa_m) \tag{6.12}$$

$$OL\ Comparison = \frac{(OL_{New\ Config.} - OL_{Current\ Config.})}{OL_{Current\ Config.}} \times 100 \tag{6.13}$$

First, we consider an industrial instance of a toolset consisting of 22 machines and 37 recipes for a single period. Diagram 6.2 depicts the current toolset workload balance. The vertical lines correspond to machine capacities. Each horizontal bar represents the workload of a machine. Bars above the capacity lines show the overloaded machines, and the opposite for underloaded machines. By calculating the

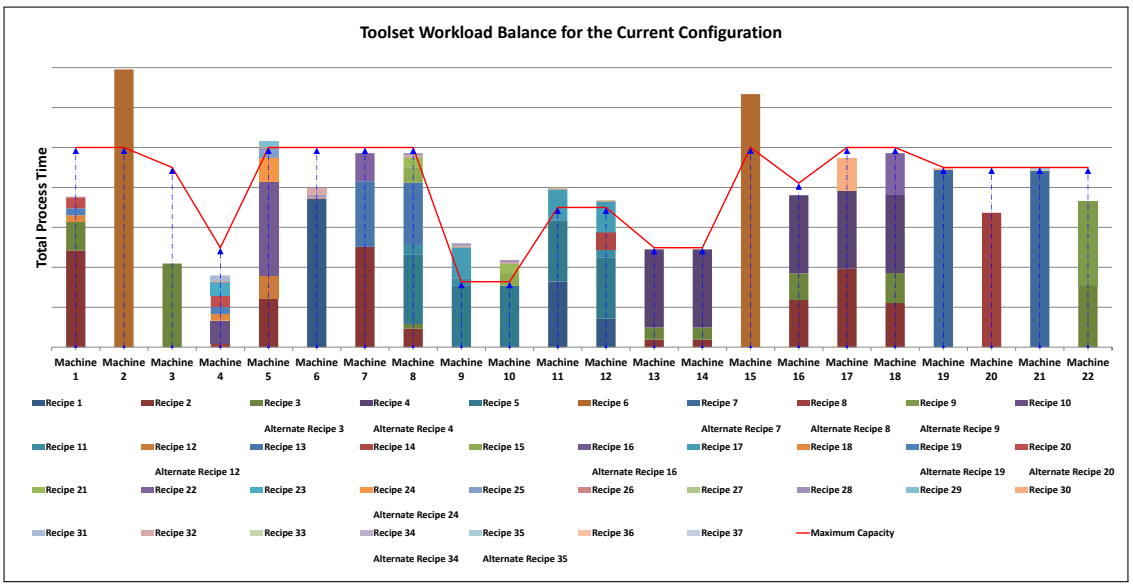


Figure 6.2: Toolset Workload Balance for the Current Recipe-to-Machine Qualification Configuration

toolset variability associated with each new qualification (creating additional manufacturing flexibility), the qualification which reduces the most workload variability, is chosen. The workload balancing diagram for the new toolset qualification configuration (Figure 6.3) illustrates how the workload variability is reduced after performing one new qualification. Note that both diagrams have the same scale. By continuing

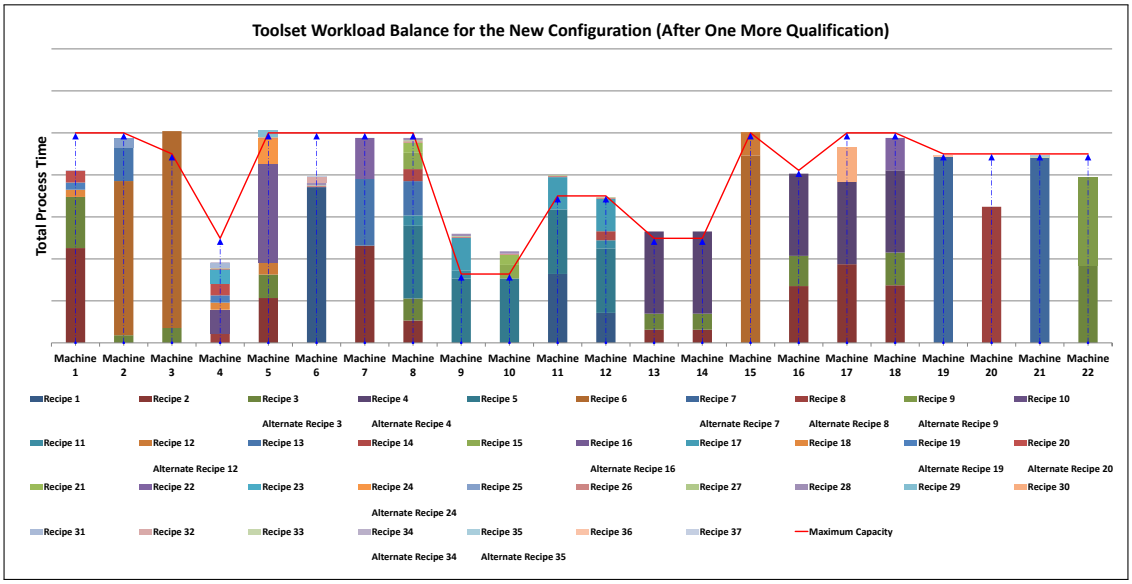


Figure 6.3: Toolset Workload Balance for the Configuration After Performing One New Qualification

to perform new qualifications, the toolset capacity allocation improves. However, instead of showing the workload diagram, variability and performance indicators are presented in Table 6.1. It shows the workload variability, overload and unused capacity reduction and used capacity increase percentages after new qualifications, i.e. creating new links between recipe set and machine set. It can be observed that performing more and more new qualifications reduces less and less the production variability. A trade-off must be made between the cost of performing new qualifications and the benefit of reducing workload variability. Figure 6.4 depicts the results of Table 6.1. It is worth to note that workload variations are not linear as the number of new qualifications increases linearly. Some qualifications decrease variability more than others. However, too many new qualifications do not decrease variability very much. Table 6.2 presents how one new recipe-to-machine qualification affects production variability for ten industrial instances. In general, one new qualification reduces variability, overload and unused capacity drastically while only slightly increasing the total workload. Note that, in the first instance, the overload is completely eliminated (-100%), capacity utilization is highly increased (20.20%) while the total workload only increases by 3.40%. The experiments illustrate again

Table 6.1: Number of new qualifications versus variability reduction and performance indicators variations

	Variation			
New Qualification(s)	Variability	Workload	Overload	Unused Capacity
1	-77.44%	3.33%	-37.33%	-32.72%
2	-86.32%	4.08%	-47.11%	-40.02%
3	-89.94%	4.55%	-57.58%	-44.66%
4	-94.41%	4.94%	-59.48%	-48.48%
5	-95.58%	5.22%	-62.22%	-51.20%
6	-96.08%	5.28%	-61.99%	-51.83%
7	-96.12%	5.23%	-65.82%	-51.31%
8	-96.60%	5.48%	-66.75%	-53.78%
9	-97.55%	5.84%	-72.56%	-57.33%
10	-97.55%	5.84%	-72.56%	-57.33%

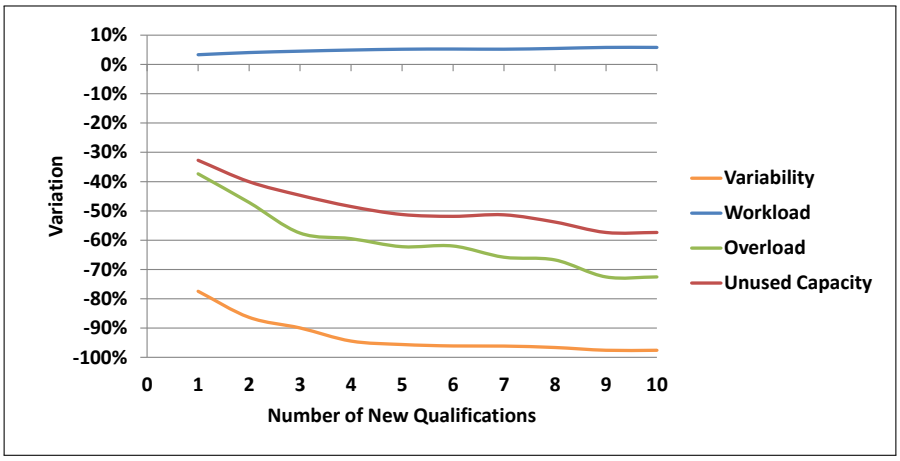


Figure 6.4: Toolset Variability, Workload, Overload and Unused Capacity Variations versus Number of New Qualifications

that the variations do not follow a linear pattern. Depending upon the data set and toolset configuration, nearly the same amount of variability reduction can lead to more or less impact on the performance measures. This is illustrated in Instances 1 and 10. While both instances record a variability decrease of about four percent 4% (-4.13% and -4.71%), the variations of workload (3.40% and 1.20%), overload

(−100% and −1.41%) and unused capacity (−20.20% and −5.49%) are very different. Tables 6.1 and 6.2 show that the toolset workload increase percentage is not

Table 6.2: Variability reduction by performing one new qualification

	Variation			
Instance Number	Variability	Workload	Overload	Unused Capacity
1	-4.13%	3.40%	-100.00%	-20.20%
2	-34.52%	2.16%	-46.76%	-7.36%
3	-76.11%	4.43%	-89.42%	-22.41%
4	-43.39%	3.99%	-28.97%	-16.76%
5	-3.29%	2.20%	-21.47%	-18.06%
6	-21.90%	2.29%	-20.05%	-15.80%
7	-62.82%	1.06%	-5.80%	-5.03%
8	-99.76%	4.57%	-30.10%	-87.41%
9	-49.65%	1.75%	-18.86%	-18.23%
10	-4.71%	1.20%	-1.41%	-5.49%

equal to the variability, overload and unused capacity decrease. This implies that, by performing one new qualification, only a small increase of workload leads to a high decrease of variability. The reason is that the process times of recipes are different from machine to machine and, by creating a new link between a recipe-machine couple via qualification, a better capacity allocation becomes possible.

6.6 Conclusions and Perspectives

In this chapter, we have studied the relationship between toolset qualification configuration (toolset manufacturing flexibility) and workload variability. A product can only be processed in a toolset when its associated recipe is qualified on at least one machine. Increasing the number of qualifications adds flexibility to the toolset and allows better capacity allocation.

Variability measures are presented to evaluate the workload variability of a toolset qualification configuration. Based on industrial data, we showed that, by

performing the best qualification according to the proposed variability measures, toolset overload, unused capacity and also variability are reduced.

It was shown that that without incurring the high qualification cost to achieve a full flexible system (Figure 6.1c), we can possibly reach the desired flexibility at a very lower cost by having a partial flexibility configuration (Figure 6.1b).

In conclusion, more manufacturing flexibility is required where the workload variability is the largest and not simply where the workload is the largest. In other words, more manufacturing flexibility absorbs workload variability. If the workload variability is low, meaning that (almost) all machines of a toolset are loaded equally according to their capacity, more manufacturing flexibility does not reduce variability. In this case, acquiring new machines might be necessary.

Several perspectives are possible for this study. An important source of variability is batching. The same variability measures can be used to evaluate how batches affect workload variability. One concept closely related to variability is *robustness*. It is interesting to study the relative quality of the solution for specific recipe-to-machine qualification configuration or data variability. The proposed approach leads to local variability reduction. Using stochastic modeling, it would be interesting to integrate the present toolset variability measurement to decrease the global production variability. In this case, the impact of manufacturing flexibility on buffer stock requirements between workcenters can be studied. In the same context, it is interesting to answer the following questions: how does the variability propagates in the production line? how does the variability pattern changes after crossing other workcenters? Can variability create a Bullwhip Effect? If yes, how to quantify the effect? Machine failure is often a major element of toolset variability. Although they reflect machine breakdowns, the machine capacities used in our measures are deterministic. It could be interesting to explicitly consider machine breakdown probabilities.

Finally, it could also be relevant to formalize the trade-off between the costs associated with performing and maintaining new qualifications and the gains related to the improved flexibility and variability quantified with our measures. This could lead to an interesting bi-criteria optimization problem. It remains to be seen if this will bring enough added value to the decision makers that are currently using our decision support system, since they will have to provide more information.

Chapter 7

Industrialization of Qualification Management

In this chapter, we present and discuss to which extent and how the approaches proposed in the previous chapters have been industrialized. Special industrial restrictions, workflow procedure and industrial applications are presented.[1]

[1] Part of this chapter has been published in [82].

7.1 Introduction

In previous chapters, we discussed qualification management and its impact on capacity planning from different aspects. When industrializing our approaches, we had to deal with some additional industrial constraints which are detailed in this chapter.

Not previously qualified recipes are discussed in Section 7.2. The industrial implementation for two different toolsets and the decision making process are presented and discussed in Section 7.3. Some results obtained with the application are shown in Section 7.4. We close the chapter with some conclusions and perspectives.

7.2 Not Previously Qualified Recipes

When dealing with real data, it is necessary to consider the case in which at least one recipe is initially not (or no longer) qualified on any machine. A typical case, is a recipe associated with the introduction of a new product. *Not previously qualified recipes* are denoted *NQRs*. NQRs may also result from (automatic) disqualifications when a recipe is not processed over a certain time window on the machine.

Below, we modify uncapacitated flexibility measures to consider NQRs. We define one additional parameter P as the penalty coefficient for *not previously qualified recipes (NQRs)*.

7.2.1 WIP Flexibility Measure

The following WIP flexibility measure first proposes qualifications for the *not previously qualified recipes* (*NQRs*) and, once all the *NQR*s are qualified, helps to suggest further qualifications for other recipes.

Using the *penalty coefficient* (P) in F_{NQR}^{WIP} as defined below, we artificially put a large load on the machines until all of the *NQR*s are qualified. When all *NQR*s are

qualified, the measure proposes qualifications for other recipes.

$$F_{NQR}^{WIP} = \frac{M \times \sum_{m=1}^{M} (WIP_m/M)}{\sum_{m=1}^{M} (WIP_m + P \sum_{r=1|NQR}^{R} WIP_r)} \in (0,1] \tag{7.1}$$

P must be large enough to make the maximum WIP of the set of *non previously qualified recipes* larger than the maximum value of all WIPs. Therefore, in each iteration, as NQRs are qualified, the value of P is updated.

$$MaxWIP = \max_{\forall r}\{WIP_r\} \tag{7.2a}$$

$$MaxWIP_r^{NQR} = \max_{\forall r|NQR}\{WIP_r\} \tag{7.2b}$$

$$P = \frac{MaxWIP}{MaxWIP_r^{NQR}} \tag{7.2c}$$

7.2.2 Time Flexibility Measure

The same idea as with the F_{NQR}^{WIP} is used to derive F_{NQR}^{Time}. The *penalty coefficient* (P) artificially puts a large load (in terms of process time) on the machines. The term $P \times C_{NQR}$ disappears when all NQRs are qualified and F_{NQR}^{Time} reduces to F^{Time}. P must be large enough to make the maximum process time of the set of NQRs larger than the maximum process time of any normal recipe. In other words, the *penalty coefficient* (P) makes the maximum of the sum of the process times on each tool ($\sum_{r=1}^{R} \frac{WIP_r}{TP_{r,m}}$) of NQRs (C_{NQR}) larger than the maximum of all C_ms. Therefore, in each iteration, as NQRs are qualified, the value of P is updated.

$$F_{NQR}^{Time} = \frac{C_{ideal}}{\sum_{m=1}^{M} (C_m + P \times C_{NQR})^{\gamma}} \in (0,1] \tag{7.3}$$

Since $C_m = \sum_{r=1}^{R}(WIP_{r,m}/TP_{r,m})$, (7.3) can be reformulated as in (7.4).

$$F_{NQR}^{Time} = \frac{C_{ideal}}{\sum_{m=1}^{M}\left(\sum_{r=1}^{R}\frac{WIP_{r,m}}{TP_{r,m}} + P\sum_{r=1|NQR}^{R}\frac{WIP_{r,m}}{TP_{r,m}}\right)^{\gamma}} \tag{7.4}$$

Contrary to the *Toolset Flexibility*, in order to calculate both the *WIP* and *time flexibility measures*, the optimal workload balance must be calculated. Note that capacitated flexibility measures as well as variability measures can be modified in the same way to consider NQRs. The usage of the modified flexibility measures are very important when there are NQRs in the recipe set. Using normal flexibility measures, the flexibility gain obtained for NQRs is negative! Therefore, these recipes do not appear among qualification alternatives although they should be the first for qualification.

7.3 Implementation and Industrialization

In this section, we describe how the concept has been deployed in industry.

7.3.1 Decision Making Process

In this thesis, we have succeeded to put in place a new workflow process in Soitec for qualification management.

Performing a qualification can be a time-consuming and costly task (based on the toolset). It may disturb production and several departments might have to be mobilized. As a result, it is preferable to know in advance the required new qualifications and of course the best ones. The required data (see Section 7.3.3) are collected for the next three weeks on a weekly basis. The simulation for the next three weeks is used to avoid rush qualifications (mostly due to *not previously qualified recipes*) and to provide a rough vision of the toolset capacity and flexibility. When using the application on a one-week data, the scheduled downtimes are identified and the production plan is more stable. The results of the analysis give guidelines for workload allocation and scheduling at the operational level.

Figure 7.1 shows where the data is collected and which departments use the results. The *process department* uses the flexibility gain values to determine the best qualification(s) to perform. This ensures a more robust and balanced toolset leading to capacity optimization. The *capacity planning* service must provide (if possible) the required capacity to realize the production plan determined in collaboration with the supply chain department. Once the desired qualification configuration is determined, the optimal workload balance is calculated using the application. This optimal workload distribution serves as guidelines to allocate the recipe volumes to machines by the *production control* department.

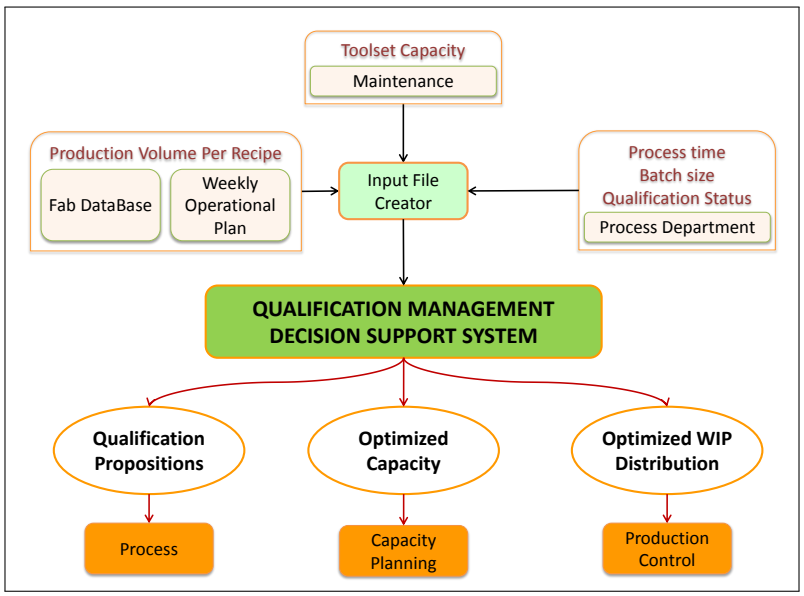


Figure 7.1: Overall Decision Making Process

7.3.2 Industrialization

The flexibility measures described in [54] were implemented with the programming language VBA™ (Visual Basic for Applications) under Microsoft Excel™ originally in [53]. The application has been completed to cover all of the subjects discussed in this thesis. Figure 7.2 is a snapshot of the control sheet of the application. Using the control panel, it is possible to choose the necessary functions (flexibil-

ity measures or variability measures, capacitated or uncapacitated, considering or ignoring not previously qualified recipes, different resolution approaches, etc.)

In a separate sheet the required parameters of the model are entered. These include: *Load Balancing Exponent* (γ), *Overload Balancing Exponent* (γ_o), *Underload Balancing Exponent* (γ_u), *Overload Weight* (o), *Underload Weight* (u) and *System Flexibility Weight* (a).

The application solves the optimization model for the current qualification configuration and saves the *current flexibility level* according to the selected flexibility measure. Then it virtually qualifies each of the qualifiable recipes and saves it in a matrix. Finally the flexibility gain is calculated by subtracting each element of the matrix from the current flexibility level while highlighting the best qualification. The flexibility gain is shown in both a tabular and listed manner. In the latter, the possible qualifications are listed in a decreasing order of their flexibility measure. In a separate worksheet, the optimal workload balance according to process

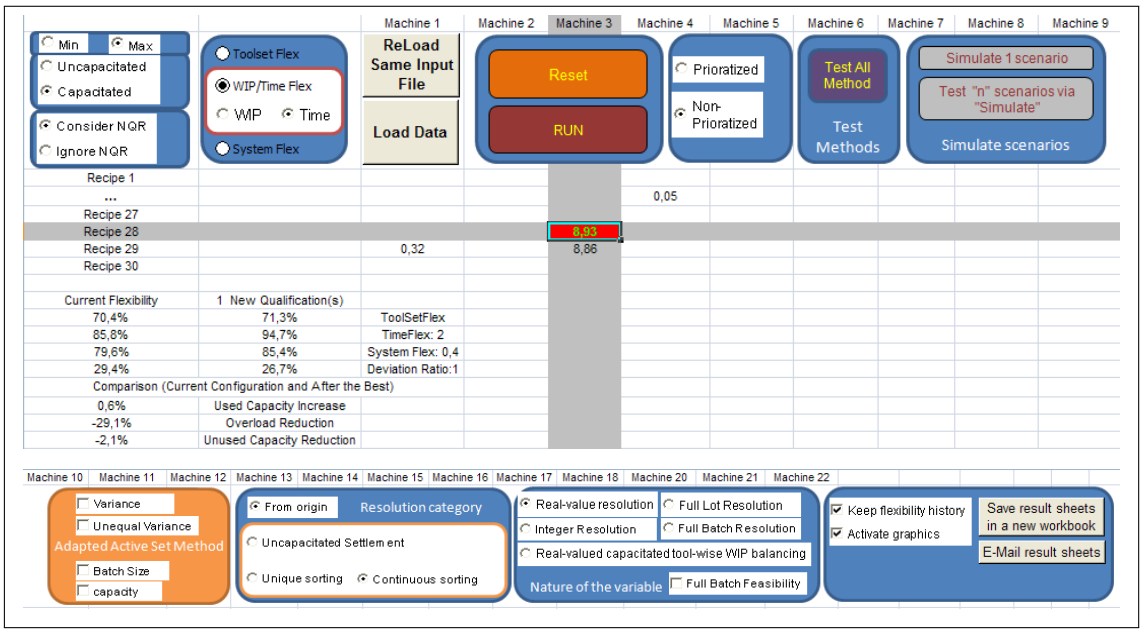


Figure 7.2: Control Panel of the Application: *Qualification Management Optimization Application*

time or WIP is calculated for the current qualification configuration and after the

best qualification. This enables the user to analyze the load on the toolset and the possible improvements that could be achieved if the best qualification is performed. Some examples of the workload balance diagram are illustrated in Figures 6.2, 6.3, 7.7 and 7.12.

7.3.3 Input Data

One of the important aspects of industrialization is input data preparation. In general, four data categories are used: Toolset qualification status, production volume (or WIP) per recipe, throughput and batch size for each qualified and qualifiable recipe for each machine and the maximum capacity per machine. *Toolset qualification status* defines the qualification status of each recipe-to-machine couple. It can be *non qualifiable* (or *unqualifiable*), *qualifiable* or *(already) qualified*. So, the application calculates the flexibility gain if each *qualifiable recipe* was *qualified*.

7.3.4 Input Data Extraction and Treatment

Figure 7.2 shows a snapshot of the Microsoft Excel™-based application, developed using the programming language VBA™. The data are extracted from an image of the MES (Manufacturing Execution System) database. Using SQL (Structured Query Language), the production route of each part is identified. The steps and the recipes are extracted. Then the production volume per product reference is cross-referenced with the recipes to give the production volume (or WIP) per recipe. Figure 7.3 shows the control sheets of the data extractor. Throughputs, batch sizes and qualification statuses are extracted automatically via another Excel-based application (called *Input file Creator*, Figure 7.4) to separate worksheets. Finally all data are exported via one Excel workbook containing four worksheets to the *Qualification Management Optimization Application* (Figure 7.5). Figure 7.6 shows an example of a data extraction.

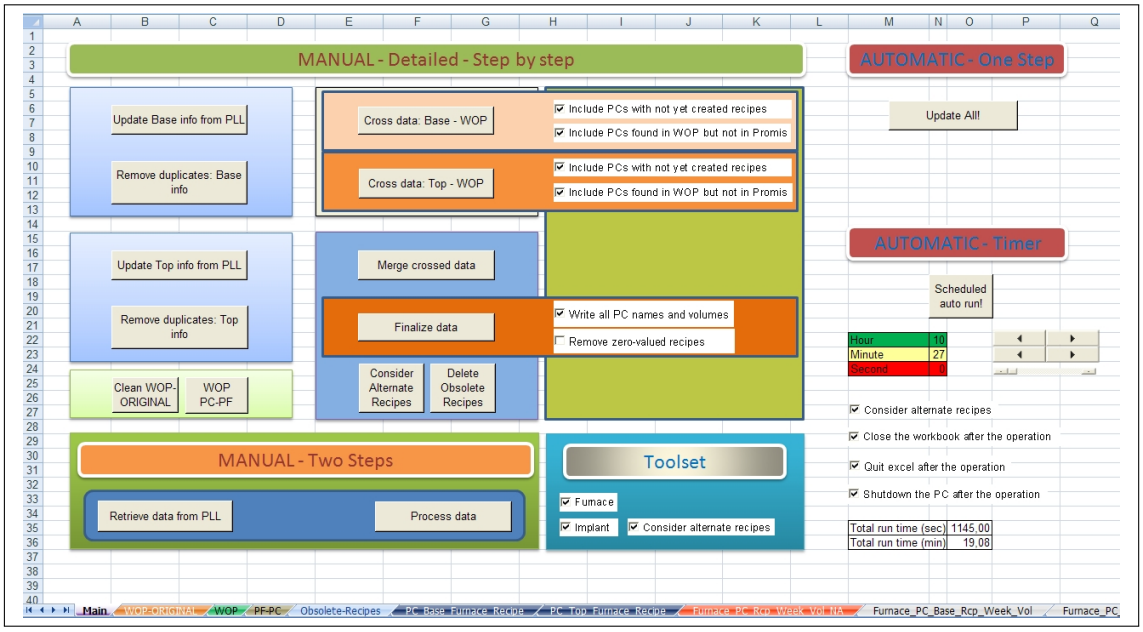


Figure 7.3: Data Extractor (WIP per Recipe) Application

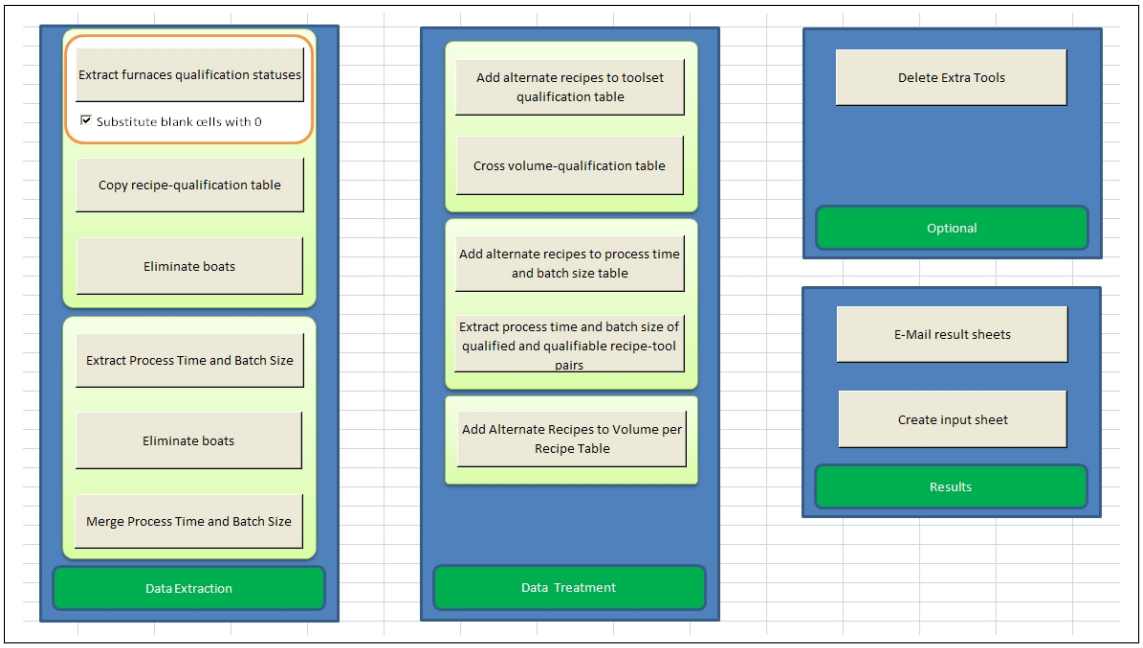


Figure 7.4: Input File Creator

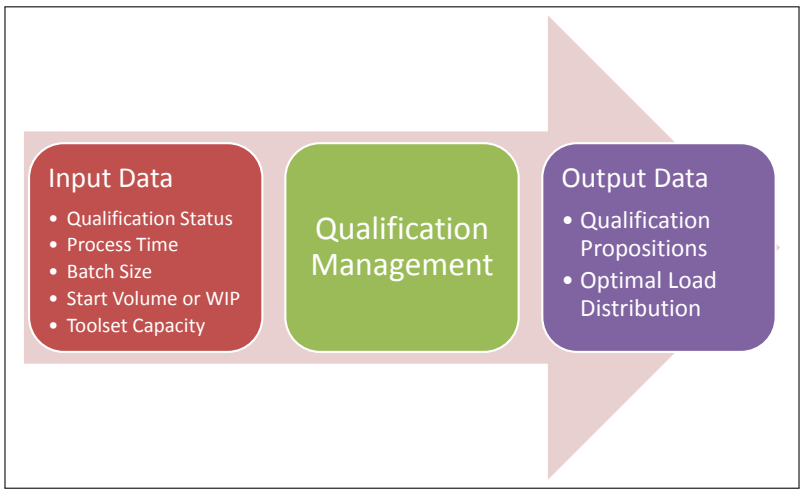


Figure 7.5: Data and Information Flow

7.4 Results

The application is being used for two bottleneck toolsets as a decision support system for qualification management and capacity planning.

7.4.1 Thermal Treatment Toolset

The *thermal treatment toolset* consists of 22 machines belonging to four different machines types designated respectively from types A to D. Each machine type has similar hardware and software characteristics. Some of the recipes have a corresponding alternate recipe which can be used to better balance the workload. As the alternate recipes are performed on different machine types, they may be faster or slower than the original recipe. With low values of γ, the optimization balancing algorithm tends to allocate more load to faster machines. When increasing γ, the algorithm tend to allocate more load to slower machines to better balance the workload among all machines.

As an example, the results of the simulation of one week are used. The *System flexibility measure* is chosen as the flexibility measure. As workload balancing is more important for the company, we chose 70% of *Time Flexibility* and 30% of *Toolset Flexibility*. After extracting and preparing data, the application is run in order

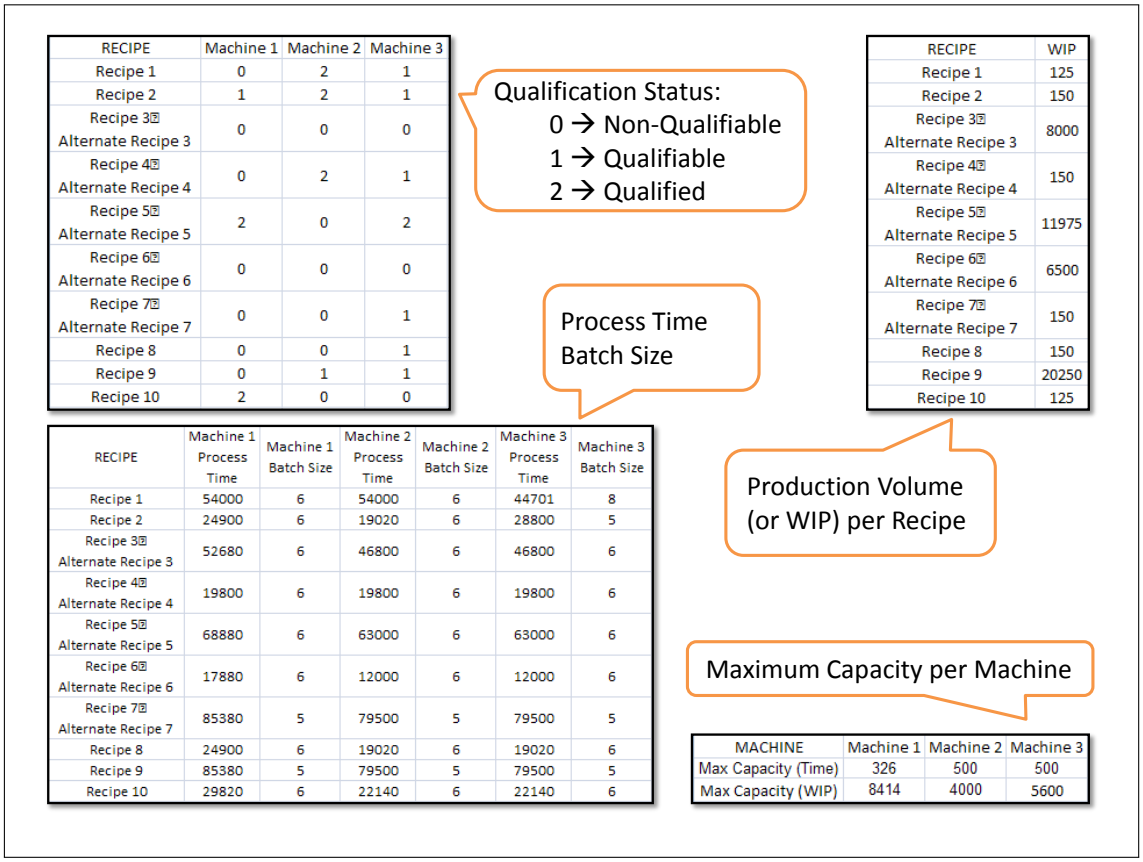


Figure 7.6: Input Data Format

to find the best qualification and the capacity of the toolset. Figure 7.7 shows the current optimal load balance on the toolset. The machines are aligned on the x-axis, the y-axis shows the production time in hours while the line shows the maximum available time (maximum capacity) of each machine. Some of the machines are overloaded ($M1, M4, M6, M8 - M12$) while others have unused capacity ($M13 - M18, M20 - M22$) and one machine is empty ($M19$). Looking at the *flexibility gain table* (Figure 7.8), one of the recipes is marked with red. This recipe has actually no qualified tool (it is a *not previously qualified recipe*). The *Time Flexibility measure* is calculated as in (7.4). It forces the model to propose to qualify first this new recipe. According to the flexibility gain table for qualifiable recipes, the toolset composition could be guessed. The four machine types are Type A ($M1$ to $M12$), Type B ($M13$ to $M17$), Type C ($M18$ to $M21$) and Type D ($M22$). As the machines belonging

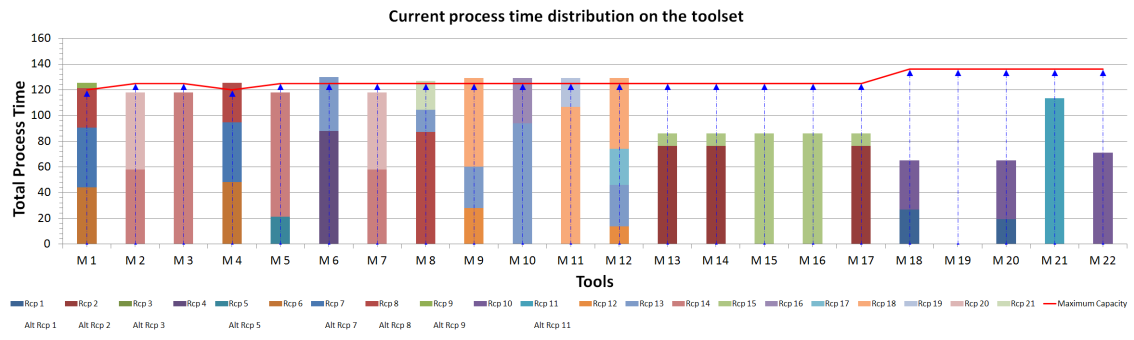


Figure 7.7: Current Workload Balancing (in Hours for the Production Volume of one Week)

to the same machine category have similar characteristics, their flexibility gain is also more or less the same. Another specificity of this recipe as that it is actually composed of two recipes, which are both equivalent but processable on different machine types. "Recipe 3" is processable on machines types A, C and D while its equivalent recipe ("Alternate recipe 3") is executable on machine Type B. By referring to the load balance (Figure 7.7), the machines of type A are the most loaded. Several reasons could explain this load allocation. Due to their hardware or software configuration, they are already qualified for a significant number of recipes. Therefore more load could be allocated to them. Another reason could be their throughput. However, as already mentioned, the behavior of the algorithm can be controlled using the *balancing exponent* (γ).

The best qualification is specified in green (i.e. to qualify "Recipe 3-Alternate Recipe 3" on "Machine 16"). It is important to note that the application is used as a *Decision Support System (DSS)*. This implies that the *decision maker* can choose *his/her best qualification* among different alternatives, according to his/her qualitative criteria. The role of the decision maker becomes more important when the differences between flexibility gains are not significant. By qualifying the new recipe, the same machines are still over-loaded despite the unused capacity on the majority of the machines (Figure 7.9). The reason is that by qualifying a new recipe, we add more load on the toolset and due to current qualification settings and restrictions, we are unable to allocate production volumes to less loaded machines.

The flexibility gain table after the first qualifications is shown in Figure 7.10. The

	M 1	M 2	M 3	M 4	M 5	M 6	M 7	...	M 13	M 14	M 15	M 16	M 17	...	M 21	M 22
Rcp 1 Alt Rcp 1								...						...		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Rcp 3 Alt Rcp 3		53,33	53,33		53,33		53,33	...	56,84	56,84	57,23	57,23	56,84	...	48,46	
Rcp 4				0,12				...						...		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Rcp 21				0,39		0,39		...						...		

Figure 7.8: Flexibility Gain Table Containing one *Not Previously Qualified Recipe (NQR)*, i.e. "Recipe 3-Alternate Recipe 3"

best qualification increases the *System flexibility* (F^{Sys}) by 12.90%. By decomposing F^{Sys}, F^{Time} increases by 17.43% (i.e. from 79.4% to 96.8%) and F^{TS} increases by 2.4% (i.e. from 68.1% to 70.5%). The workload balance is depicted in Figure 7.11. It is worth noting that only one new qualification can increase the machine utilization and eliminate overload. Finally, one of the other outputs of the application is the *optimal WIP distribution diagram* corresponding to each optimal workload balance. In Figure 7.12, the optimal WIP distribution diagram corresponding to the optimal workload balance of Figure 7.11 is shown. Each color represents a recipe, the x-axis shows the machines of the toolset and the y-axis represents the production volume (for confidentiality reasons, the figures are removed).

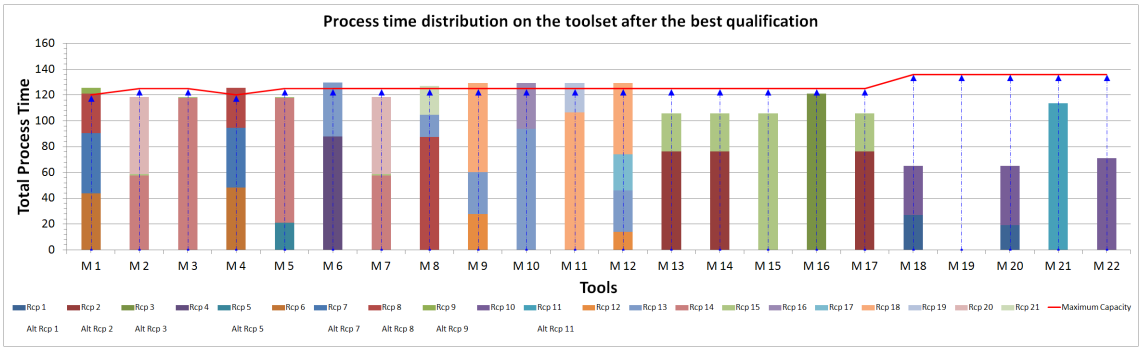


Figure 7.9: Workload Balancing after Qualifying "Recipe 3-Alternate Recipe 3" on "Machine 16"

	M 1	M 2	M 3	M 4	M 5	M 6	...	M 13	M 14	M 15	M 16	M 17	M 18	M 19	M 20	M 21	M 22
Rcp 1⊐ Alt Rcp 1							...							0,45			
Rcp 2⊐ Alt Rcp 2							...										
Rcp 3⊐ Alt Rcp 3		2,07	2,07		2,07		...	1,91	1,91	1,91		1,91				1,56	
Rcp 4				0,14			...										
Rcp 5⊐ Alt Rcp 5			0,47				...										
Rcp 6							...										
Rcp 7⊐ Alt Rcp 7							...							12,90		1,00	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Rcp 21				0,72		1,37	...										

Figure 7.10: Flexibility Gain Table after the 1st Qualification

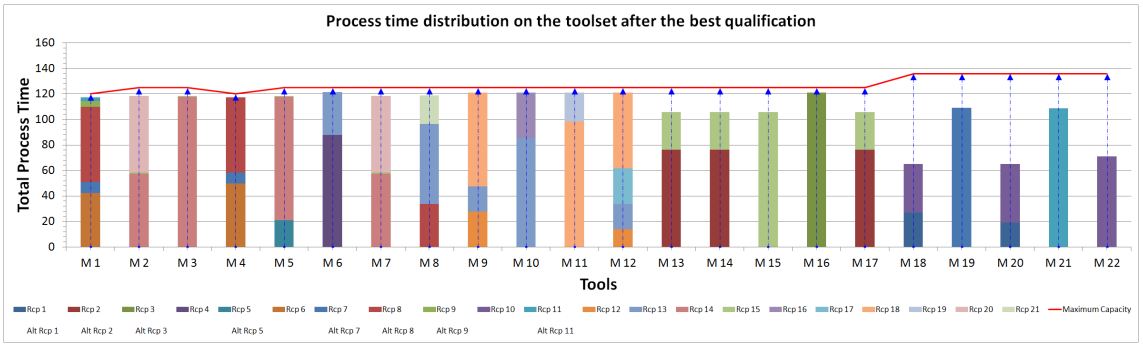


Figure 7.11: Workload Balancing after Performing the Best Qualification, i.e. “Recipe 7” on “Machine 19”

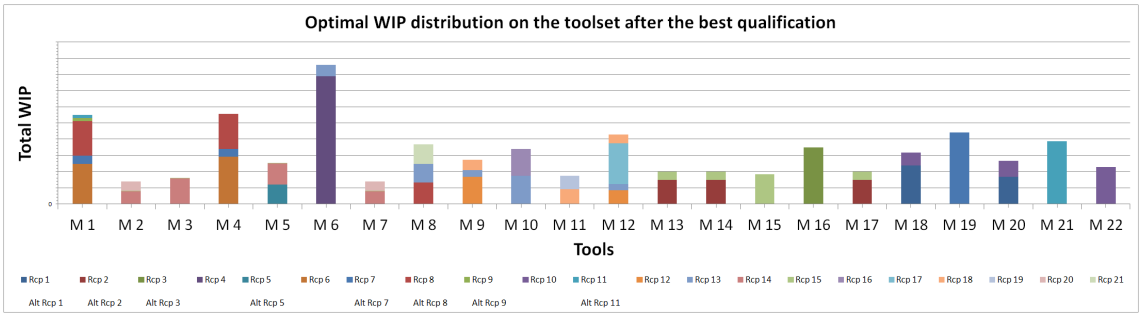


Figure 7.12: Optimal WIP Distribution after Performing the Best Qualification, i.e. “Recipe 7” on “Machine 19”

7.4.2 Implantation Toolset

The application has also been used for the *implantation toolset*. The number of recipes and the constraints in this toolset are far less compared to the *thermal treatment* toolset. Besides, the implantation toolset consists of only ten machines of two different types. The implantation recipes are sorted under categories (named *families*). By creating new recipes or introducing new products, the new recipe is considered to be qualified on all machines already qualified for the recipe family to which it belongs. Therefore, the qualification proposals are useful when we intent to qualify a new *recipe family* on a machine type. The workload balance diagram is used to visualize the workload balance on different machines.

7.5 Conclusion and Perspectives

In this chapter, we showed how the research conducted in this part of the thesis is applied in industry. In particular, we discussed an important industrial constraint, namely *not previously qualified recipes*. Using modified flexibility measures, the impact of new products can be estimated on the variability and the configuration flexibility of the toolset. By omitting a machine from the toolset list and calculating the configuration flexibility, we can estimate to which extent the machine is important for the production of the current product-mix. This *machine criticality* can be calculated before planning preventive maintenance to study the impact of stopping a machine. Each toolset may have specific restrictions. Some of the most current in semiconductor manufacturing are auxiliary resource management, yield issues of running some recipes on some machine, consumable usage reduction, etc.

Chapter 8

Conclusions and Perspectives of Qualification Management

8.1 Conclusions

In the first part of the thesis, we studied various facets of flexible qualification management in semiconductor industry and its relationship with capacity optimization and planning. Here we give a summary of the main achievements together with some conclusions.

Qualification Management

In Chapter 3, we discussed the importance of qualification management in semiconductor manufacturing industry. It enables a machine to process a new recipe. Qualifications in a toolset are like machine eligibility restrictions. Qualifications allow workload allocation of recipes to machines. Thus, qualifications are important in production management. Capacity planning tries to match the production plan with the available labor and equipment. Qualification management is one of the possibilities of increasing equipment capacity utilization and capacity planning. Our aim is to find the most appropriate recipe-to-machine qualification configuration. The qualification configuration changes over time, with recipe changes, disqualifications, machine breakdown, etc. Besides, a qualification configuration may be suitable for a certain product-mix and volume while with product-mix changes, the configuration may turn out to be poor. Performing each new qualification adds flexibility to the toolset qualification configuration. However, some new qualifications may not improve capacity utilization due to existing qualifications, low workload on the

newly qualified machines, low production volume of the newly qualified recipes, etc. Therefore qualification configuration flexibility must be increased intelligently.

Capacitated Qualification Management

The flexibility measures introduced in [54] do not consider the unequal limited capacity of the equipment while qualification management. QM considering the toolset maximum capacity was discussed in Chapter 4. Limited equipment available capacity affects the optimal workload balance and hence QM. We have modified the uncapacitated flexibility measures to consider unequal and limited equipment capacity. However, in order to make sound judgments, we need other indicators which we have called *Capacity Deviation Ratios*. These two complementary measures are used for capacitated QM. Flexibility and deviation ratio measures increase the flexibility to fight against unbalanced workload, regardless if equipment over- or under-utilization. The equipment capacity under-utilization or over-utilization have different impacts on the production line. While underloaded machines cause loss of productivity, overloaded machine increase the production line variability and endanger demand satisfaction. Therefore, we have introduced unequally penalized flexibility and capacity deviation measures to emphasize either equipment under- or over-utilization. Numerical experiments demonstrate that the equipment capacity is to be considered in QM to find the best qualification configuration.

Qualification Management with Batch Size Constraint

Batching is one of the characteristics of many processes. Similar products are grouped to form a batch. In batch machines, a batch is processed at a production run instead of a single product. Batching helps to reduce production and setup costs while yielding products of the same quality. Batching is preferable where long and expensive runs are needed. In Chapter 5, we studied the impact of batch sizes on qualification decisions and hence capacity utilization. In order to calculate the configuration flexibility of a toolset, we have solved a problem similar to the scheduling of unrelated nonidentical parallel machines with eligibility constraints. The problem is known to be NP-hard in strong sense. Several heuristics are proposed

to resolve the problems. Numerical experiments show that batch sizes must be considered in QM to optimize the toolset capacity utilization.

Qualification Management and Production Variability

Variability is known to be the enemy of production. It should be constantly monitored and reduced to get a smooth production flow. Equipment qualification, and capacity and batching are among sources of variability. In Chapter 6, we developed variability measures to show to which extent an efficient equipment qualification management strategy can reduce the variability. It was shown that too many qualifications tend to reduce less and less the workload variability. Qualification management helps to increase toolset capacity utilization where the variability is the highest and not necessarily where the equipment workload is the highest. Variability control helps to decrease inventory and in-line WIP while increasing production performance.

Industrialization

The concept is used in Soitec as a decision support system for qualification management and capacity planning. Some snapshots of the developed applications and the decision making process are presented in Chapter 7.

Flexibility and Cost

In Chapter 6, we showed that we may achieve full flexibility at very little cost by performing a minimum number of qualifications. This means that we can achieve "full flexibility", by not qualifying all recipes on all machines as illustrated in Figure 6.1c. In summary, by using the developed models in Soitec, flexibility has increased the capacity utilization without any capital investment in machine purchase. Meaning that small flexibility increase can bring about a huge amount of improvements in KPIs such as workload, overload and unused capacity (see Table 6.1).

8.2 Perspectives

At the end of each chapter, we have listed various perspectives. Here, we present some indications for future investigations.

Capacitated Time Flexibility Measure (F_{Capa}^{Time}) Considering the Required Qualification Duration

Setup times and costs are not always negligible. The required qualification time may consist of a considerable time portion relative to the period under study. Consider for instance a new qualification which takes about 35 hours in a study scope of one week i.e. 168 hours per week. For a toolset of 10 tools and assuming an equal uptime of 80% for each machine, this qualification represents about 2.6% of the total available time. Now consider another qualification which requires only about 10 hours, i.e. about 0.7% of the total time available in this toolset. The qualification time should be taken into account if the differences are significant in order to calculate the correct configuration flexibility, variability and capacity. Besides the qualification time, several other costs such as product inspection and incurred downtimes may also be considered.

Let us define the following parameters and variables:

Parameter

$cq_{r,m}$ qualification duration of recipe r on machine m.

Variable

$$OQ_{r,m} \quad \begin{cases} 1 \text{ if recipe } r \text{ is to be qualified on machine } m, \\ 0 \text{ otherwise.} \end{cases}$$

In the following, the Capacitated Time Flexibility measure is slightly modified to take this feature into account:

$$F_{Capa}^{Time} = \frac{Ideal\ Ratio_{Capa}}{\sum_{m=1}^{M} \left(\frac{\sum_{r=1}^{R} \left(\frac{WIP_{r,m}}{TP_{r,m}} + cq_{r,m} OQ_{r,m} \right)}{Capa_m} \right)^{\gamma}} \in (0, 1]\,. \tag{8.1}$$

The constant $Ideal\,Ratio_{Capa}$ is the minimum value of the total production time C_m plus the required qualification duration ($\sum_{r=1}^{R} cq_{r,m}OQ_{r,m}$) divided by $Capa_m$ when all qualifiable machines are "virtually" qualified :

$$Ideal\ Ratio_{Capa} = \min \sum_{m=1}^{M} \left(\frac{\sum_{r=1}^{R}\left(\frac{WIP_{r,m}}{TP_{r,m}} + cq_{r,m}OQ_{r,m}\right)}{Capa_m}\right)^{\gamma} \quad \text{with } Q_{r,m} = 1 \quad \forall r, m. \tag{8.2}$$

Disqualification

The opposite action of qualification is called disqualification. When a recipe (previously qualified) is disqualified from a machine, it is no longer possible to allocate production volume of that recipe to a machine. A disqualification may occur automatically after hardware change, maintenance, etc. Maintaining qualifications may be costly. Therefore, in some cases, we may need to disqualify recipes such as production ramp-down. By disqualifying a recipe, we can evaluate the *importance* or *criticality* of each qualification. *Qualification criticality* is calculated the same way as *qualification flexibility*. With the difference that we now virtually disqualify recipes instead of qualifying them. First, the flexibility value of the current configuration is calculated. Then each (already) qualified recipe is virtually *disqualified*. After each virtual disqualification, the flexibility of the toolset configuration is saved in a matrix. The initial configuration flexibility is subtracted from each flexibility element of the matrix. The values will be nonpositive contrary to flexibility values which are nonnegative.

Maintaining Qualifications

In order not to perform a re-qualification, it would be interesting to minimize the number of machine disqualifications. The cost of re-qualification and maintaining a qualification must be considered over time. The problem becomes dynamic and more sophisticated. This subject is partially treated in [73].

Further Industrial Deployment

Other industrial restrictions affect qualification management. Among these, we can name auxiliary resources, (perishable) consumables, material handling system restrictions, sequencing of the tasks, etc.

As an example of an interesting industrial and academic challenge, we describe some of the cleaning process constraints in SOI manufacturing. One of the recurrent process steps in semiconductor manufacturing is *cleaning*. The toolset is also very used in SOI fabrication process. The toolset is composed of several machines. Each machine has 10 tanks which are filled with liquid chemical materials, such as HF (Hydrofluoric acid). The robot can at most handle one lot (of 25 wafers) at a time. A stage corresponds to a production step in the SOI fabrication process. Depending on the product, a stage consists of different recipes. A recipe defines the different chemicals which must be used and the time the wafers are exposed to the chemical(s). During the execution of the recipes, wafers in the cassettes are dipped in the tanks for some minutes. Not all tanks must be visited for all recipes. The execution of a recipe in a stage may cause contamination in the consumable chemicals. Contamination levels are identified for each stage. Higher contamination levels add additional constraints to the problem. Chemicals in the tanks are consumables. Tank chemical must be changed after a certain amount of time a chemical has been used on wafers. Sometimes, it can be expressed based on the number of wafers (or lots) cleaned in the chemical(s). The batch constraint does not exist in a strong sense. However, wafers move from one tank to another in lots of 25 wafers.

Part II

Closed-Loop Production Planning: Lot-Sizing with Multiple Remanufacturing of Reusable By-Products

Chapter 9

Motivations and State-of-the-Art

In this introductory chapter, we provide the general framework of the second part of the thesis which addresses an original production planning problem. We begin the study by a brief description of the activities in a supply chain with a focus on production planning. The Smart-Cut Technology used for SOI manufacturing is a patented and exclusive production process. This unique manufacturing process motivates our study.

9.1 Introduction

The open market and globalization require a global strategy of the whole enterprise instead of local separated decisions. The context is more challenging in the semiconductor industry. One of the main elements of the supply chain of every manufacturing company is *production management*. A sound production management system assures the continuity of the enterprise and a stable and healthy turnover. Due to the hierarchical structure of a supply chain, production management is declined in several concepts. At the tactical level, production management concerns in particular capacity and production planning while at the operational level, it mainly concerns scheduling and dispatching. The concept of capacity planning was discussed in the Part I of this thesis. We also approached scheduling problems, especially when considering batches in Chapter 5.

This part of the thesis is dedicated to production planning. The topic is also called *lot sizing* in the academic literature. The focus is on production planning modeling dedicated to SOI manufacturing based on Smart-Cut™ Technology. We discuss a lot sizing model considering SOI fabrication and refresh process lines.

First, the industrial context of the study is set by pointing out the importance of an efficient production planning in Section 9.2. Then, a picture of the concept of supply chain management is panted in Section 9.3 with a particular focus on production planning in Section 9.4. It is shown that the addressed problem has not been treated in the scientific literature by mentioning some related papers in Section 9.5. We conclude this introductory chapter in Section 9.6.

9.2 Motivations

The global SOI production process is explained in Section 2.3. In this section, this process is recalled while underlining the motivation of the study. The exclusivity of the production process is based on a patented thin layer deposition method called *Smart-Cut™* Technology. In this method, two wafers are bonded together while a thin layer of oxide separates them. The bonded wafers are then split, leaving a thin silicon layer on the oxide. The remaining wafer can be recycled several

times for the production of more SOI wafers. The recycling (called refreshing in the industrial jargon) needs another remanufacturing line. Due to confidentiality reasons, we cannot detail the refresh process steps. Just we note that less steps are required than in the SOI fabrication. However, constant monitoring of the yield and cycle times of the refresg line are necessary for smooth running of the SOI production line.

In the following, the extreme importance of the refresh process is pointed out.

The purchased Fresh Wafer is quite expensive. This is why the goal is to recycle it more and more. The high price of the Fresh Wafer is due to its quality. The quality of Wafer A (Top Wafer) must be excellent for two reasons: First, it constitutes the active layer of the SOI Wafer and second, it must keep its characteristics after several recycling processes. On the other hand, the quality of Wafer B (support, handle or Base Wafer) is not critical. On the other hand, the recycling (refresh) process is far less expensive than the purchase of new Top Wafers (Fresh Wafers). So, the refresh process helps reducing the costs.

The refresh process (especially the intern refresh process) gives the possibility of negotiating the purchase of the bulk wafers from the suppliers and better absorb the bulk market fluctuations. The purchase negotiation concerns not only the purchase price but also the purchase commitment term. The refresh process creates more independence from suppliers.

The internal refresh process is also important because the bulk suppliers prefer to use their capacity to produce Fresh Wafers (prime wafers) rather than for the refresh process. The Smart-Cut™ Technology allows Soitec to be the only company which uses Refresh Wafers as well as Fresh Wafers.

Besides, the worldwide wafer market is out of control but the internal refresh process remains under control of Soitec. Namely, Soitec can use its expertise to create more alternatives in the BOM and manage better its material requirements. By better mastering the refresh and SOI fabrication processes, more refresh levels can be qualified for the fabrication of SOI Wafers.

Using the internal refresh process, we increase the internal capability. As the refresh can also be subcontracted, the refresh process line can be used as a buffer zone for labor planning. Meaning that if the activity of the SOI production line

increases, it is possible to decrease the activity of the refresh line for a short while by subcontracting the refresh process (or even buying more Fresh Wafers).

When purchasing Fresh Wafers, the purchase lead time may not be completely under control due to negotiations, ordering, processing and shipment procedures. When using the refresh process, not only the cycle time is decreased but we also better plan the supply chain. Securing the Top Wafer supply leads to a more reliable supply chain and a decrease of the supply chain management risk.

9.3 Supply Chain Management

A *supply chain* is a collection of at least two legally distinct organizations with the aim of satisfying the final customer demand by producing products or providing services [94]. The present organizations in a supply chain are linked together through material, information and financial flows. Usually a supply chain consists of suppliers, manufacturers, distributors, retailers and customers (Figure 9.1). The suppliers provide raw material and components for the manufacturer. Note that, the suppliers may be broken down to several levels or "tiers" which collaborate between them and make another supply chain. However, most manufacturers procure raw materials and components from "Tier 1 Suppliers". In the same way, the other building blocks of a supply chain may contain other components.

Logistics is an integrating function which assures the forward and reverse product or service and information flow from the suppliers to the final customers. It is mostly divided into *inbound* and *outbound logistics*. While *inbound logistics* deals with the inbound movements (materials, components, information, etc.) to the business, *outbound logistics* refers to all flows between the business and the customers [20].

Figure 9.1 shows an example of the information, material (or service) and financial flows in a supply chain. The information flow starts from the customers by the demand for a product (or service), the delivery date, the quantity, etc. Customers procure their needs from retailers who keep a small stock of products. Distributors (national, regional, etc.) supply retailers. In some cases, notably via internet sales, distributors directly supply the end customer. Distributors transfer the requirements information to the manufacturers. The requirements, translated in terms of

raw materials and component, are transferred to the suppliers. The material flow begin from suppliers and continue the value chain to reach the end customers. At each level, the financial flow exists from the end customer to the suppliers.

However, these flows may be reverse for some other types of supply chain. For example, in a supply chain where product return and repair is also possible, the material flow is reversed from the customer towards the manufacturer. Or in case of reimbursement, the financial flow is also reversed towards the customer. *Supply*

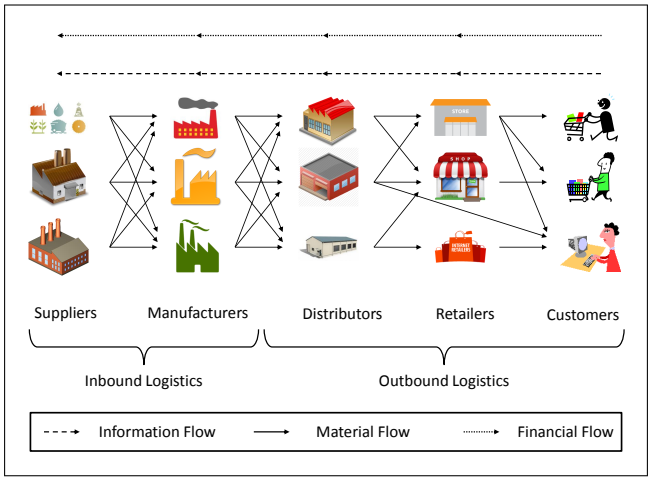


Figure 9.1: Supply Chain (An example)

Chain Management (SCM) is a set of approaches used to efficiently integrate organizational units across a supply chain while coordinating materials, information and financial flows in order to minimize system wide costs while meeting service level requirements [91]. The aim of supply chain management is to improve global competitiveness by reducing costs (improving efficiency) and increasing customer service (improving effectiveness) [20]. While Manufacturing Planning and Control (MPC) tends to optimize the performance while coordinating information and materials flows inside the company (intrafirm), SCM focuses on coordinating and optimizing the interactions across companies (interfirm) [97].

The efficiency and effectiveness of a supply chain are measured using *Key Performance Indicators (KPIs)*. Using SCM, we expect to improve KPIs. A list of KPIs and the expected improving trend can be found in [65], [30] and [94].

9.3.1 Decision Levels in Supply Chain Management

Decisions in Supply Chain Management have different importance and are updated periodically. Therefore, they are usually classified into three levels according to their importance and time horizon [2] and [89].

Strategic Level The strategic decisions, made at the highest management level, define long-term policies of the supply chain, namely for several years. They influence all decisions of lower levels. Typical examples of strategic decisions concern the design of the supply chain, new products, number and location of production facilities and warehouses and research and development projects.

Tactical Level Tactical decisions concern mid-term planning. The planning horizon is usually between one month and two years. It mostly determines rough and aggregates resource utilization such as budgeting, demand, production, distribution and sales planning.

Operational Level Operational decisions have short term effects, ranging from hours to weeks. They specify the most detailed and precise instructions for execution of tasks, such as scheduling, sales plan, inventory planning, distribution scheduling.

The decisions at different levels are linked and have mutual impact. Therefore, in order to obtain feasible decisions, the coherence between the decisions of different levels should be ensured. The production planning decisions discussed in Part II fall in the tactical level while the qualification decisions discussed in Part I have a more operational nature.

9.3.2 Closed-Loop Supply Chain

The problem tackled in this Part of the thesis has similarities to what is called a *Closed-Loop Supply Chain* (*CLSC*). A CLSC considers the backward product flow (reverse supply chain) as well as the traditional forward product flow. The backward product flow focuses on the return of the product to the company from the customers or the market. A Closed-Loop Supply Chain aims at recovering the entire product

or at least some of its modules or parts. Key additional elements in CLSCs are ([43]):

Product acquisition is the task of retrieving the right quantity of used products at the right time and with a certain quality level directly from the customers or from the market.

Reverse logistics is the planning framework for the reverse flow of the collected products (and their related information) for product recovery (such as remanufacturing, recycling, etc.), see [24] for an interesting discussion about reverse logistics.

Product disposition consists of sorting, testing and grading the returned products. Based on the result of this step, the returned product may be considered for some sort of recovery, resell or landfill.

Product recovery is the process of adding value to the selected used products. The recovery may be at the product, module or component level. Based on the recovery level and the required effort, the product recovery may be called remanufacturing, repair, part or module recovery, reconditioning, refurbishing or material recycling [42].

Remarketing is the process of reintroducing the recovered products into the market. It concerns selling and redistribution.

CLSCs are important due to enormous potential savings, environmental regulations and strategic business issues related to procurement of raw materials.

9.3.3 Production System

A production or manufacturing system is a collection of technologies (patent, know-how, etc.), equipment and labor with the aim of satisfying the customer need by the end product. The production (or manufacturing) of goods is the process of adding value to raw materials or components through transformation or assembly.

The end products are composed of raw materials, components or subassemblies. A **Bill of Materials** or **BOM** is a hierarchical list of the needed quantities to manufacture the end product. A *Single-Level BOM* displays the end product with only one level of children (the so-called *immediate* raw materials, subassemblies or parts). However, if there are more parent-children relationship between the sub-assemblies,

it concerns a *Multi-Level BOM*. Using the BOM, based on the demand for the end products (*independent demand*), we make a projection of the requirements, in terms of components (*dependent demand*). In a production system, the raw materials or components are processed or assembled in **workcenters**. A production (or manufacturing) line consists of a series of workcenters. A workcenter consists of people and equipment at shop floor level and performs one or several production operations. Each operation adds value to semi-finished products. The output of a workcenter is picked up by the subsequent workcenter. The sequence of production operations and the associated workcenters, production process (or recipe), equipment, required labor, setup and process time are defined in the production **routing**. The manufacturing time, setup time and waiting times constitutes the **lead times** of each operation. All the described elements of a production system are necessary for production and capacity planning.

Closed-Loop Manufacturing Systems In order to classify our problem, we define a *Closed-Loop Manufacturing System* as a system in which by-products, co-products and/or components can be returned to the manufacturing system after a process. The return process may concern a recovery activity such as remanufacturing, refurbishment, recycling, etc.

9.3.3.1 Production Environment

A production (or manufacturing) environment must be chosen based on the product or service which is offered by the firm and the company's strategy in how to respond to the market. The production environment type is important in supply chain planning, especially in demand and requirements planning. The production environment (or strategy) can be categorized as follows (Figure 9.2) ([100], [97] and [49]).
The **Make-to-Stock (MTS) production environment** offers standard products or services based on anticipation of demand. The demand of the customer is satisfied directly from products in stock. Therefore the level of service to the customer depends directly on the stock level. The crucial question is the replenishment policy of the stock. Higher customer service level requires high on-hand inventory or a

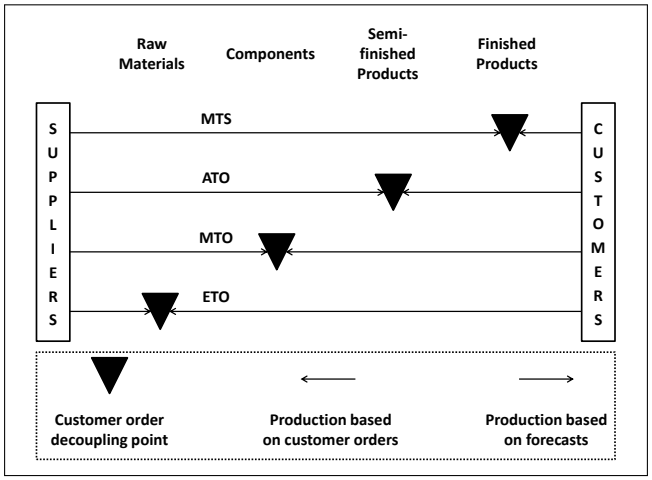


Figure 9.2: Different Production Environments and Customer Order Decoupling Point ([49])

good knowledge of future demand. Hence, a trade-off between the inventory level and customer service level must be determined. Accuracy in demand forecasting is essential for the efficiency of a Make-to-Stock production strategy. The number of products is limited with fairly long and predictable life cycles [49].

In the **Make-to-Order (MTO) production environment**, the production starts once a firm order of the customer is placed. The strategy is mostly suitable for customized products with relatively high inventory cost produced in relatively low volume. Using this policy, the excessive inventory holding expenses (common in Make-to-Stock strategy) decrease at the cost of increased waiting time for the customer to receive the product. Lead time for the final product can be reduced by procuring to forecast long lead time, highly used items with low degree of obsolescence [49].

The **Engineer-to-Order (ETO) production environment** is similar to MTO with the difference that partial or the whole product design is according to the customer specifications. Therefore, a high level of customer participation in both design and manufacturing phases is required. This production policy is suitable for low volume complex or one-off (one-of-a-kind) products. The engineered-to-order products are expensive and time-consuming to manufacture. The raw materials or basic components are often difficult to source or to design. On the other hand,

very low inventory is carried. The production time and the required manufacturing capacity are difficult to estimate. High manufacturing flexibility is also required.

In the **Assemble-to-Order (ATO) production environment**, the product is manufactured according to the customer specifications from standard modules on-hand. ATO is a hybrid strategy between Make-to-Stock and Make-to-Order strategies. The modules (or components) are kept in inventory (MTS) while the final assembly occurs after the customer order placement (MTO). Therefore, ATO requires a modular structure of the final product. The lead time between the order receipt and product delivery is lower in comparison to the MTO strategy. Besides, the products are more customizable compared to the MTS strategy and the on-hand module inventory is lower.

The production environment depends upon the customer requirements, final product design, product life cycle, demand behavior and market behavior, production facilities, production cost, inventory management and production cycle time. Different production policies may be applied for different products in the same company. A combination of these policies is also possible. A vital concept in planning is the *customer order decoupling point* which defines which material is forecast-driven or customer order-driven. Figure 9.2 shows that as we move more towards ETO policy, the production is *sales-driven* and on the inverse direction, *manufacturing-driven*.

The production environment of Soitec is located in the range of Engineer-to-Order or Make-to-Order production environments.

9.3.3.2 Remanufacturing

Until now, our discussion was centered on manufacturing. However, an emerging and promising field in both research and industry considers *product recovery*. Product recovery aims at restoring products while eliminating waste to a large degree. Some of the product recovery options are depicted in Figure 9.3. However, the options may go beyond this framework. Several recovery options may be applied for different components of the same product. More and more, product recovery is considered in the design phase of the products.

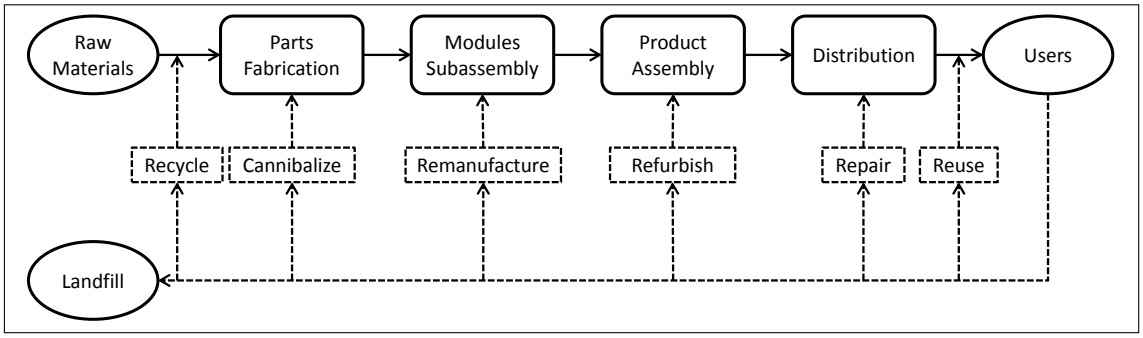


Figure 9.3: Product Recovery Options ([89])

The following concepts are defined by [41]: Remanufacture-to-Stock (RMTS), Reassemble-to-Order (RATO), and Remanufacture-to-Order (RMTO) based on the classification proposed in [101].

In our research, we consider a specific type of *by-product recovery*. The industrial jargon of Soitec for this type of recovery is *refreshing*. However, to avoid confusion and due to similarities of this type of recovery to remanufacturing, the latter term is used.

9.3.3.3 Types of Production Systems

Production is the main element of the value chain in a supply chain. In order to determine an adapted production management policy, the production systems are classified into two main categories: *Continuous* and *Intermittent Production Systems* (Figure 9.4).

Continuous Production System In a continuous production system, the inputs, outputs, processes and sequences are standard. It follows a *Make-to-Stock* policy based on anticipation of demand. The factory size is normally large with highly specialized automated machinery and material handling systems. The material requirement is very predictable. The production is without interruption and at a high rate. The production process generates usually by-products. The bottlenecks are stationary and known [89]. Therefore, production planning and control is easy. It can be categorized into *Mass Production Flows* and *Process Production Flows*.

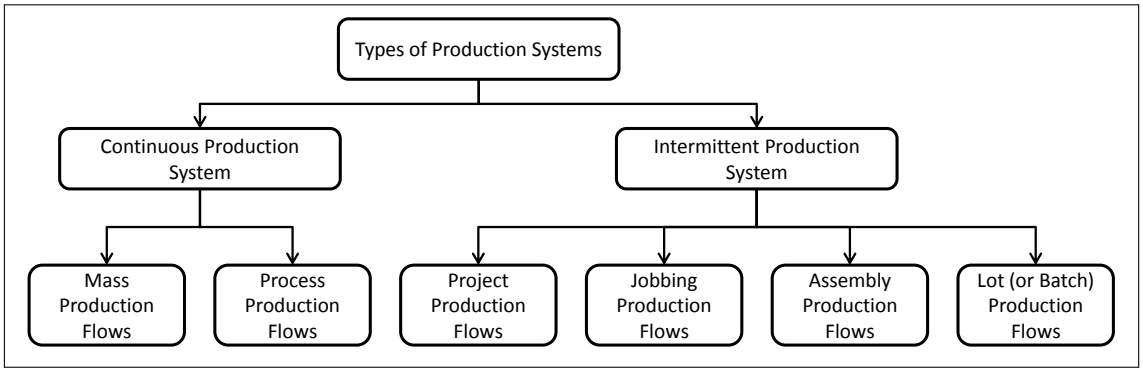


Figure 9.4: Types of Production Systems (Adapted from [32])

In **Mass Production Flows**, several products are produced or assembled on a very large scale based on demand forecast, e.g. chemicals production, final car assembly, etc. In **Process Production Flows**, the production system is designed for the production of normally a single product, e.g. glass, paper or steel production.

Intermittent Production System An Intermittent Production System is designed flexible enough to handle a variety of different products. The production flow is intermittent. The material handling system must be adapted for a variety of raw materials, components, semi-finished products and final products. The products are more customized and the production volume is relatively lower than in a Continuous Production System. The production operations, routing and process times differ from product to product. Usually, few by-products are generated using this type of production system. It is classified into *Project Production Flows*, *Jobbing Production Flows*, *Lot (Batch) Production Flows* and *Assembly Production Flows* [89].

In **Project Production Flows**, one time complex projects are realized by the company at an estimated cost and delay. The product volume is very low and the system is generally very flexible. The production strategy is *Engineer-to-Order*. Examples are construction projects, ship building, etc. The **Jobbing Production Flows** is designed to manufacture very few products according to the specifications of the customer. The workstations are flexible and multifunctional. The followed production strategies are *Engineer-to-Order* or *Make-to-Order*. Examples are spe-

cialized machinery production or prototype building. In **Assembly Production Flows**, interchangeable parts and modules are assembled together sequentially to make the final product. The products are customizable to some extent as the parts or modules are standardized. **Lot (or Batch) Production Flows** allows the production of similar products in production campaigns, in other words in lots (or batches). A lot (or batch) is a group of identical products which follow the same routing and production operations. Organizing big production campaigns (lots) generates a lot of stock while small size production campaigns must be more frequent and require more setups. The production policy is mostly Make-to-Stock or Make-to-Order depending on the firmness of the orders or demand predictions. A production system can be a combination of these production systems. *Lot Production Flows* and *Assembly Production Flows* can be combined together as a single production system.

Production planning of a Lot Production System is referred to as a lot-sizing problem. Generally speaking, a lot-sizing problem determines the size of the production campaigns of each product over a planning horizon (defined in days, weeks or months) in order to satisfy the customer demand. The objective is mostly expressed as the minimization of production setup and inventory costs while meeting demand and capacity restrictions. Of course, additional constraints may be considered in a lot-sizing optimization problem.

The SOI fabrication and Refresh process, considered in this thesis, are classified under intermittent production system with lot production flows. We discuss this lot-sizing problem in more details in Section 9.4.

9.3.3.4 Manufacturing Configuration

A manufacturing system is composed of human and machine resources. Until leaving the production line, the product visits different machines according to operations routing. One or several machines may be grouped in a workcenter. A workcenter (Aka workstation or toolset) adds value to semi-final products by performing operations. The sequence of visiting the machines inside a workcenter or between the workcenters defines the manufacturing routing. The routing between the workcenters (or the machines) can be seen as the workshop configuration. The

scheduling and capacity planning depends upon the workshop and workcenter configurations. In the following classifications, we distinguish *resource characteristics* and *routing characteristics*. None of these classifications are necessarily reserved to a workcenter nor to a workshop.

Resource Characteristics A workcenter (or a workshop) may consist of one or several machines with different characteristics. A **single machine** workcenter (workshop) is the most basic resource configuration. In a **parallel machines** configuration, several similar machines are gathered together in a workcenter (or workshop). A configuration with parallel machines may be further classified into *Identical Machines*, *Uniform Machines* and *Unrelated Machines*. ***Identical Machines*** have the same throughput. ***Uniform Machines*** have proportional throughputs. ***Unrelated Machines*** have different throughputs.

The operations routing (sequence) in a workshop (or a workcenter with parallel machines) may be classified to *Flow Shop*, *Job Shop* or *Open Shop*. In the **open shop** configuration, no special routing and sequence are defined for the operations. This type of workshop is rarely found in industry. In **flow shop** configuration, all products visit the machines (or workcenters) in the same order. The flow shop configuration may exist in both continuous and intermittent manufacturing environment. Examples are cement industry (continuous manufacturing environment) and automobile industry (intermittent manufacturing environment). In the **job shop** configuration, each product (or product family) has its own routing and sequencing on the machines or workcenters. The job shop configuration can be found in semiconductor manufacturing or pharmaceutical industry.

The workstations in SOI fabrication line consist of unrelated parallel machines with a job shop configuration. The refresh process line has a flow shop configuration with several workcenters with identical machines. The machines in the workcenters are multi-purpose and mostly configurable. *Multi-purpose* or *flexible* machines (in comparison to *single-purpose* machines) are able to perform several jobs. A *multi-purpose* machine may be *configurable* or *non-configurable*. As the name suggests, configurable multi-purpose machines can be configured to perform jobs that they could not perform before configuration. However, the configuration degree and

possibilities are not the same.

9.3.4 Production Planning

In supply chain planning, production planning concerns the tactical decision answering the questions *when to produce which product* and *how much* (or *how many*). The production planning problem is expressed as an optimization problem with the objective of minimizing the cost (or maximizing the profit) subject to several constraints. With a cost minimization objective, the challenge is to find the best trade-off between opposing costs such as inventory cost -which depends upon production quantity- and setup costs. However, depending on the problem, several other cost items may be considered, e.g. material procurement costs, production cost, maintenance cost, labor cost, backlog or lost sales costs, overtime costs, etc. The main constraint of a production planning problem is customer demand satisfaction. Additional constraints concern safety stock, capacity constraints (in terms of human or production resources), market share, etc.

In the following, we discuss some concepts of supply chain planning, with more focus on production planning.

9.3.4.1 Enterprise Resource Planning (ERP)

An enterprise can be divided into major functional areas: *Finance*, *manufacturing and logistics*, *sales and marketing* and *human resources*. *Enterprise Resource Planning* or *ERP* systems provide a cross-functional integration bed for the business activities of an organization. In other words, ERPs are comprehensive information systems which support integrated planning and execution among all functions of an organization. *Manufacturing, inventory management and logistics* are among the major modules in ERP systems. Production planning and scheduling is the core of manufacturing and logistics function. One of the basic tools used in production planning in manufacturing systems with lot production flows is *Material Requirements Planning* or simply *MRP*.

Material Requirements Planning (MRP) Intermittent production systems with lot production flows have been evolved with basic material planning tools such as *Material Requirements Planning* (or *MRP*). MRP has the objective of determining the *right quantities of the right part at the right time* [97]. The major drawback of the MRP approach is the lack of capacity checks from other departments, making the plans hardly feasible operationally. To rectify the production plans of MRP, *Manufacturing Resource Planning* (or *MRP-II*) has been created. Closed-loop MRP is an enhancement of MRP systems in which capacity checks are carried out in an iterative process.

Manufacturing Resource Planning (MRP-II) MRP II is a planning method for medium-term manufacturing resource planning and control. MRP II explicitly considers the capacities of resources while interacting more with other departments by converting some outputs of production planning and control. MRP II can be seen as a closed-loop MRP while considering financial flows in the enterprise. The decision and data consistency are more precise in MRP II. MRP II has given birth to ERP systems.

9.3.4.2 Advanced Planning System (APS)

Advanced Planning Systems have been created to optimize and synchronize the whole set of activities in a supply chain in a collaborative and global way. Based on the service level and financial objectives, the activities are optimized from demand planning until final delivery at all decision levels and in line with all interfaces.

APS solutions consist of several software modules, in which each module covers a planning scope in the supply chain. The supply chain planning matrix depicted in Figure 9.5 shows the possible modules in an APS. However, the cutting and the scope of each module may be different from one software provider to another.

APS take advantage of the integrated and consistent ERP bed across the enterprise. Therefore APS supplement the ERP system used as a transaction and execution system but do not substitute it [94]. In this thesis, we consider a supply chain in which there is also a reverse material flow from the manufacturer to the supplier or between two distinct production lines of the same manufacturer. The

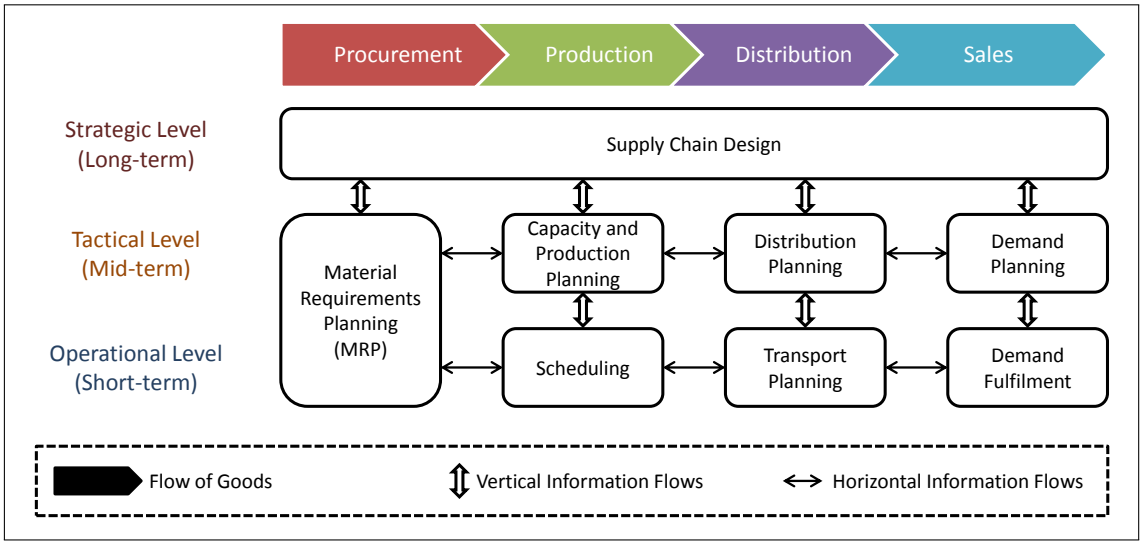


Figure 9.5: Supply Chain Planning Matrix (Adapted from [98])

concept is similar to closed-loop supply chain in which there is a possible return material flow from the customer to the manufacturer for remanufacturing. Nevertheless, it is not the final product which is returned to the remanufacturing process but the used raw material (called by-product).

9.4 Production Planning

In this part, we are interested in production planning. The term mostly used in industry and academia is lot-sizing.

9.4.1 Optimization Problems

Models are used to represent a given real-world problem. [11] defines four classes of models: *Operational exercise*, *gaming*, *simulation* and *analytical model*. In this thesis, we consider *analytical models* in which the model is entirely represented in mathematical terms bringing a high degree of simplification. Analytical models give birth to optimization problems in which a function (called *objective function*) is typically minimized or maximized over a set of points, called *feasible region*. Production

planning problems involve a sequence of decisions over multiple-periods. Industrial requirements generate a great number of variables and constraints complicating more the resolution approach.

9.4.2 Lot-Sizing Problems

We mentioned in Section 9.3.3.3 that the lot (or batch) production flow is a strategy used in intermittent production systems. This strategy consists of organizing production (or purchasing) campaigns (called "lots") per period (normally on a daily or weekly basis) over a planning horizon. By organizing production by lot, we may anticipate the demand satisfaction of several periods in order to incur less (fixed) production setup costs. However, by anticipating production, we pay more inventory costs. That is why a lot-sizing problem usually consists of finding the best trade-off between setup and inventory costs. However, the trade-off may lie among setup, production or ordering costs and inventory costs.

A detailed examination of lot-sizing problems is out of the scope of this dissertation. Moreover, in many scientific documents and textbooks, detailed state-of-the-art discussions about lot-sizing problems are presented (see [74]). So, we just look over some main ideas to determine where we stand.

A special category of lot-sizing problems deals with manufacturing processes with a potential return of material flow. In order to compare our case with cases studied in literature, we open the discussion about *Closed-Loop Lot-Sizing* later in Section 10.6.3.

9.4.2.1 Classification of Lot-Sizing Problems

The history of lot-sizing dates back to 1913 ([26]), by the Economic Order Quantity (EOQ) (or Economic Lot Size Model) introduced in [46]. EOQ and all its extensions consider a constant and deterministic demand rate over an infinite period. However, in most practical cases, the average demand rate varies over time and the planning horizon is limited. While the EOQ model recommends the same replenishment quantities for all periods, this would not be optimal with *time-varying demand*. The problems addressing time-varying demand are termed as *lot-sizing* problems.

The first basic lot-sizing model was analyzed in [99]. The proposed dynamic programming algorithm is named after the authors, Wagner and Whitin (hereafter WW). With very low demand rate variations, it would make sense to use the EOQ model. If not, exact solution methods can be used to obtain the optimal solution. However, lot-sizing problems are mostly difficult to solve, therefore heuristics are developed to get approximate solutions. Lot-sizing problems can be classified based on different characteristics, such as *number of levels in BOM*, *length of the periods*, *demand behavior*, *production capacity* and so on.

For a classification of lot-sizing problems for single-level and multi-level BOM (Figure 9.6), see [102]. In single-level lot-sizing problems, the demand of each product is independent of other products. While in multi-level lot-sizing, the problem becomes more complicated because of the parent-children relationship between products at different levels. In lot-sizing, the time scale may be *discrete* or *continuous*.

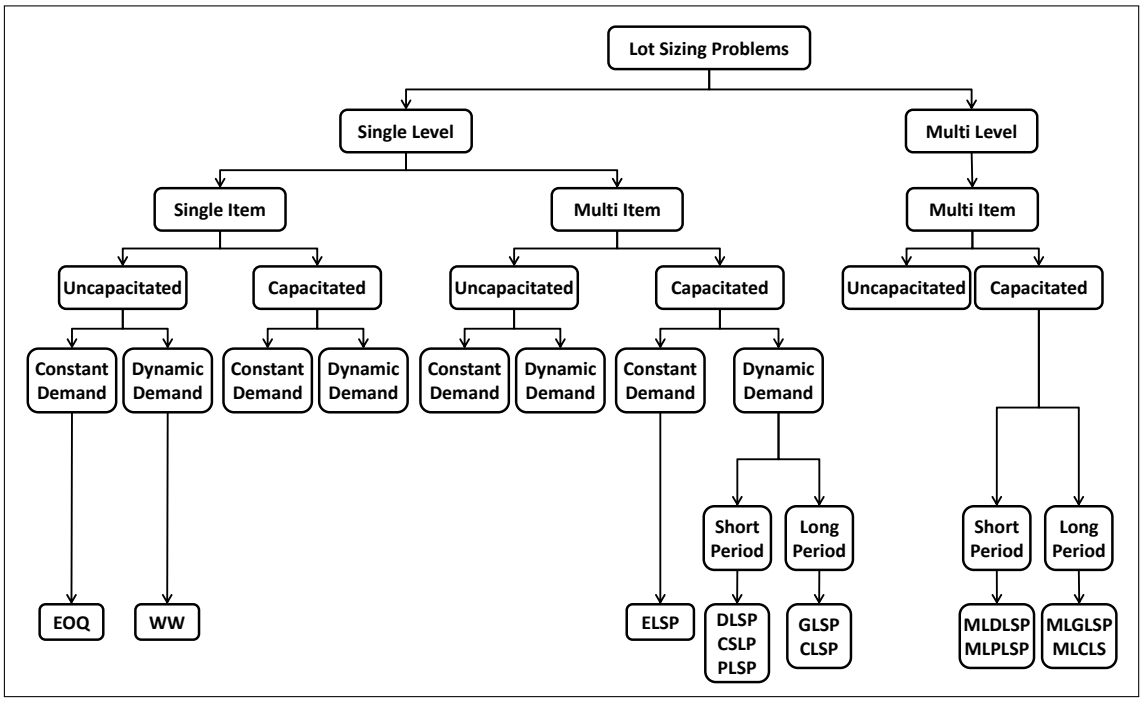


Figure 9.6: Classification of Lot-Sizing Problems for single- and multi-level BOM ([102])

The planning horizon is generally considered to be either *finite* or *infinite*. Mostly,

problems with discrete time periods have a finite planning horizon while continuous time problems have an infinite planning horizon. The EOQ model is namely a continuous time model with infinite planning horizon. From the resource point of view, a problem may be *capacitated* or *uncapacitated*. In real-world problems, resources have limited capacity. However, as capacitated lot-sizing problems are considerably more difficult to solve, for simplicity, the resource capacity restrictions may be ignored. A literature review of the models and the algorithms for uncapacitated and capacitated single level lot sizing problems can be found in [59]. Capacitated lot sizing problems with deterministic demand considering at each time one of the following extensions: Back-order, setup carryover, sequencing and parallel machines are discussed in [75]. For a classification and review of solution algorithms for capacitated lot-sizing problems, see [16].

Based on the number of items, lot-sizing problems might be *single-item* or *multi-item*. In single-item models, the items (products) do not compete for resource capacity utilization. In multi-item models, because of the interdependency between products, they compete to use (scarce) capacity. Both EOQ and WW are single-item models. For a literature review of single-item lot-sizing problems, see [13].

The discrete time lot-sizing problems are classified into *extra big time bucket*, *big time bucket* and *small time bucket* depending on the *period length*. The planning periods length of extra big time bucket are of the order of months, for big time buckets are about a few days or weeks and for small time bucket problems are about a few hours. Longer time buckets can be considered as aggregations of smaller time buckets. In this case, we talk about hierarchical lot-sizing (see [8]).

Interested readers can find a concise discussion and state-of-the-art about supply chain management with a focus on production planning in [36] (in French) and [12] (in English).

In this part of the thesis, we treat big time bucket discrete lot-sizing problems. The industrial problem considered in Chapter 10 is multi-item and capacitated. In Chapter 11, we consider a single-item uncapacitated problem. The problems are bi-level; with one level of raw materials (level 1) and one level of end product (level 0).

9.5 Literature Review

The problem studied in this part is related to different domains of lot-sizing. Some related studies are recalled to distinguish our research from previous investigations.

Our study concerns a multi-item supply chain. The classical capacitated multi-item lot-sizing problem with non-stationary costs, demands and setup times is considered in [96]. The problem is decomposed into a set of uncapacitated single product lot-sizing problems using a Lagrangian relaxation. The single-item problems are solved using a dynamic programming algorithm. A smoothing heuristic is used to make the dual solution feasible.

If there is a parent-component relationship in the item structure, the problem is classified as a multi-level lot-sizing problem. In single-level problems, only independent demands (from external customers) are considered while, in multi-level problems, the production of each final item generates a dependent demand for its components. The problem studied in this thesis is a bi-level lot-sizing problem. The problem of minimizing the setup and inventory costs in a capacitated multi-level lot-sizing problem is discussed in [6]. An interesting literature review on multi-level capacitated lot-sizing problems together with a solution approach for the dynamic multi-level capacitated lot-sizing problem are discussed in [48].

In the literature, the difference between by-product (by-production) and co-product (co-production) is not clear. Both co-products and by-products are generated during the production process or at each production run. Co-products have their own independent demand and cannot be used interchangeably. Whereas by-products are undesired "sub-products" generated during production. They must normally be reworked before being able to reuse them.

Terms such as "by-product", "co-product", "remanufacturing" and "recycling", used to designate different concepts in the literature, are close to our research. In our research, the raw material once used for production is considered as "by-product". This by-product cannot fulfill any demand and must be reworked before coming back to the manufacturing cycle. The process of restoring the generated by-products makes them reusable again as raw materials. Therefore, this process

can be considered as a "remanufacturing process". In the literature, remanufacturing concerns the returned products from the customers whereas in our case the customer is not involved.

A multi-item uncapacitated lot-sizing problem in which co-products are produced at each production run is treated in [1]. In this paper, it is considered that the co-products have their own demand and cannot fulfill the demand of the main product. Several MIP formulations are presented for the problem. Using a variant of the zero-inventory property, a dynamic program is used to solve the problem in the single-item case.

Other studies consider the co-production of a range of products with different performances in a single production run. The co-products are then sorted according to their key performance to satisfy demands of each co-product [95].

Remanufacturing in reverse logistics is considered in [76]. By remanufacturing in reverse logistics, it is meant that there is not only a one-way flow of the products to the customers but materials and products may be returned to the manufacturer for product recovery. This is what is called a "closed-loop supply chain" (see Section 9.3.2). In the proposed model, known quantities of used products are returned from customers in each period. Once reworked, the returned products are used to satisfy the customer demand as new products. Therefore, in each period, it is determined whether to produce new products or to remanufacture returned products to satisfy the demand. The supply chain is modeled using Mixed-Integer Programming (MIP). The studied problem is shown to be NP-hard. Several alternative reformulations of the original formulation are presented. The solution qualities of the LP relaxations are compared.

In a more recent study, a closed-loop supply chain with setup costs, product returns and remanufacturing is considered in [106]. The study is inspired from the paper manufacturing industry in which both virgin and deinked pulps are used to make papers. A MIP model and a Lagrangian relaxation based solution approach are further proposed. A manufacturing - remanufacturing closed-loop supply chain in a dynamic continuous time stochastic context is studied in [61].

Spengler et al. [93] propose production planning models for recycling generated by-products during production, and dismantling and recycling products at the end

of their lifetime. Several other studies use the term remanufacturing to denote the restoring or recycling of products [27].

None of the cited articles treat all aspects of our research. The problem modeling is original and to the best of our knowledge, it is not addressed in the literature. In the following chapters, the problem is unfolded in more detail and the notations are introduced.

9.6 Conclusion

In this chapter, an overview of supply chain management and production systems was presented to set the stage for the forthcoming chapters.

We are interested in production planning of two related production lines of Soitec in which the by-products generated in one production line is reworked in the second to come back to the first production line. The rework is considered as a kind of remanufacturing. That is why the production system is similar to closed-loop manufacturing systems. The production environment follow an Engineer-to-Order or Make-to-Order policy. The whole production system is intermittent with a batch production flow. The workcenters (toolsets) of the main production line consist of multi-purpose unrelated configurable parallel machines, with toolsets in a job shop configuration. The remanufacturing process line has in turn workcenters with mostly identical machines with a flow shop configuration. The production planning is done on a finite discrete time horizon with big time buckets. The mathematical models are large scale and multi-period with constrained solution space, linear functions, deterministic parameters, continuous and binary variables.

In Chapter 10, we discuss the original industrial problem with all industrial constraints. The industrial problem gives birth to an original academic problem covered in Chapter 11.

Chapter 10

Multi-Item Bi-Level Capacitated Lot-Sizing with Multiple Remanufacturing of Reusable By-Products

In this chapter, we investigate a multi-item production planning problem in which reusable by-products are generated during final product fabrication. After further processing, the generated by-products can be reused as raw materials. However, by-products can be "recycled" only a given number of times. The production and recycling processes are performed in internal and external sites with limited capacity. Each product may be produced using specific raw material (newly purchased or recycled) references. The proposed model represents a part of the supply chain of "SOI" (Silicon-On-Insulator) fabrication units. Using numerical examples based on industrial data, the model is validated and some of its characteristics are discussed. Finally, some perspectives of this work are proposed.[1]

[1]Part of this chapter has been published in the proceedings of the conference MOSIM 2014 [79].

10.1 Introduction

Silicon wafers are extensively used in semiconductor manufacturing to produce microelectronic components such as chips and integrated circuits. However, some devices require higher performance which cannot be delivered by traditional silicon-only wafers. Components built on Silicon-On-Insulator (SOI) wafers offer much more performance while consuming less energy compared to components on silicon-only wafers. SOI wafers may be produced using different technologies: *SIMOX™* (Separation by IMplantation of OXygen), *wafer bonding* or *Seed methods*. The production system studied in this chapter concerns a type of wafer bonding technology called the *Smart-Cut™ Technology* described in Section 2.3.

The chapter is organized as follows. The main aspects of the Smart-Cut™ Technology and the studied supply chain are concisely described in Section 10.2. The mathematical model corresponding to the supply chain under study is introduced in Section 10.3. The complexity analysis of the mathematical model is to be found in 10.4. Some industrial extensions and additional constraints are presented in 10.5. Using numerical experiments on test instances generated based on industrial data, the model is validated and some managerial insights are discussed in Section 10.6. Finally, conclusions are drawn and some perspectives of the study are presented in Section 10.7.

10.2 Problem Definition

In this study, we consider the supply chain of a Silicon-on-Insulator (SOI) Wafer production unit using the Smart-Cut™ Technology. In SOI Wafers, a thin layer of silicon is laid on a silicon Wafer which serves only as a physical support (or handle). These two silicon layers are separated by an insulator: *The oxide*. Figure 10.1 illustrates the main steps of the Smart-Cut™ Technology. Once Wafer **A** is oxidized and implanted, it is ready to be bonded with Wafer **B**. After the Wafers are bonded, they are split to form the SOI Wafer. Wafer **A** is the "donor" Wafer in the sense that a thin silicon layer of this substrate is deposited on Wafer **B**. In the industrial jargon, Wafer **A** is called "Top" while Wafer **B** is called "Base". As

only a thin layer of the Top Wafer is deposited on the Base Wafer, it is possible to reuse the Top Wafer several times to produce other SOI Wafers. This is one of the main advantages of the Smart-Cut™ Technology which makes the process cost competitive.

A "used Top Wafer", called "Negative Wafer", must be reworked before returning to the SOI fabrication process. This remanufacturing process is called the *refresh process* or shortly *refresh*. In industrial terminology, a new Top Wafer used for the SOI fabrication is called a "Fresh Wafer". A Fresh Wafer is purchased from silicon Wafer suppliers. After the first utilization, the generated Negative of the Fresh Wafer is called "Negative 0" or shortly "Neg 0". Refreshing "Negative 0" gives a newly usable Top Wafer called "Refresh Wafer 1". A Wafer may be refreshed only a maximum number of times (called the "maximum refresh level"). It is economically interesting to refresh a Top Wafer as many times as possible. However there is an end to the refresh process because of quality and yield constraints. Mostly, mature products have the highest refresh level because of a better understanding of the characteristics of the product and a higher expertise of the refresh process. Using this logic of numbering, a Fresh Wafer can also be called "Refresh Wafer 0". The Top Wafer can be refreshed only a limited number of times. Performing too many refresh processes leaves a deteriorating impact on the final SOI [77]. Multiple thermal treatment processes cause defectivity in the Wafers and makes the silicon transfer more and more difficult. The defectivity in the refresh process must be the least in order to keep high quality for future SOI Wafers. Hence, continuous defect inspection monitoring of the refresh process line is important.

Special yield and quality constraints or specific customer specifications may cause the SOI-Refresh planning to be more complicated. Some products may only use a Fresh Wafer as a Top Wafer and not its Refresh Wafers. Or some customers may require to use only up to a certain refresh level for their products. A rare situation which may also occur is that some products can only be produced from the Refresh Wafers of a Fresh Wafer. In this case, in order to obtain the required Refresh Wafers, Fresh Wafers are first used to produce another product. Then the generated Negative Wafers are refreshed and used further.

The refresh process, of limited capacity, can be done internally or externally.

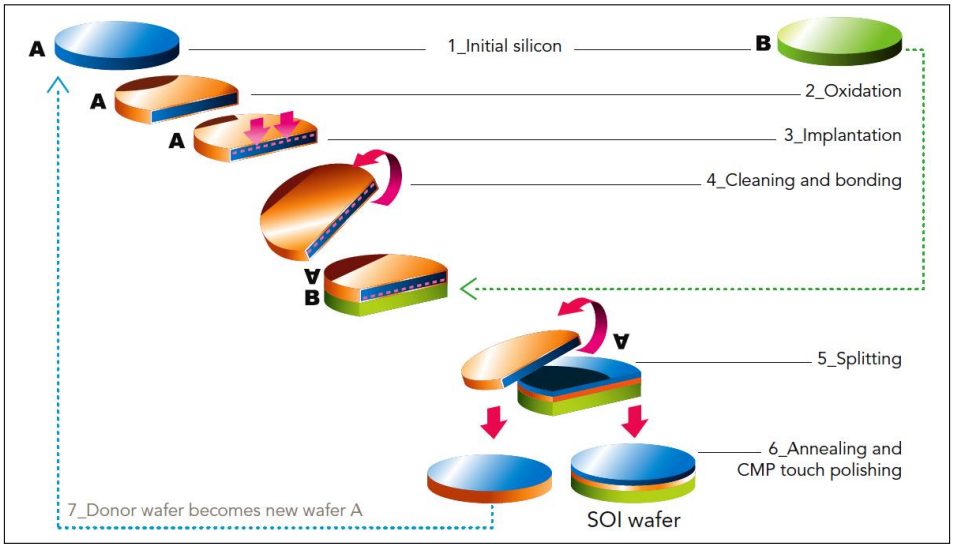


Figure 10.1: Unibond SOI Wafer Fabrication Steps Using the Smart-Cut™ Technology - [92] accessed May 2014

Internal refresh may also be performed in a different site than the one where the SOI Wafers are produced. Therefore, the shipping, planning, extra packaging, possible deterioration and increased and less certain cycle time must be considered when the refresh process is not done at the same site where SOI is produced. A yield factor is associated with the refresh process because of the manufacturing line scrap. The demand is assumed to be known over a discrete time horizon. It is possible to stock products to satisfy future demand but no backlogging is allowed.

In Figures 10.2 and 10.3, the landscape of the supply chain under study is depicted. Figure 10.2 illustrates different levels of the refresh process until the Top Wafer can no longer be refreshed. At this time, it is used as a test Wafer for qualification or test purposes or simply as a filler Wafer. Figure 10.3 depicts a simple SOI-Refresh supply chain. Top (Fresh) and Base Wafers are purchased from a bulk supplier. The SOI production site includes also a refresh line. An external refresh supplier and a customer are also present.

This study deals with production planning decisions on a discrete time horizon. The model is with big time bucket periods as multiple items can be produced in the same time period [37] and the products manufactured in a period can be used to

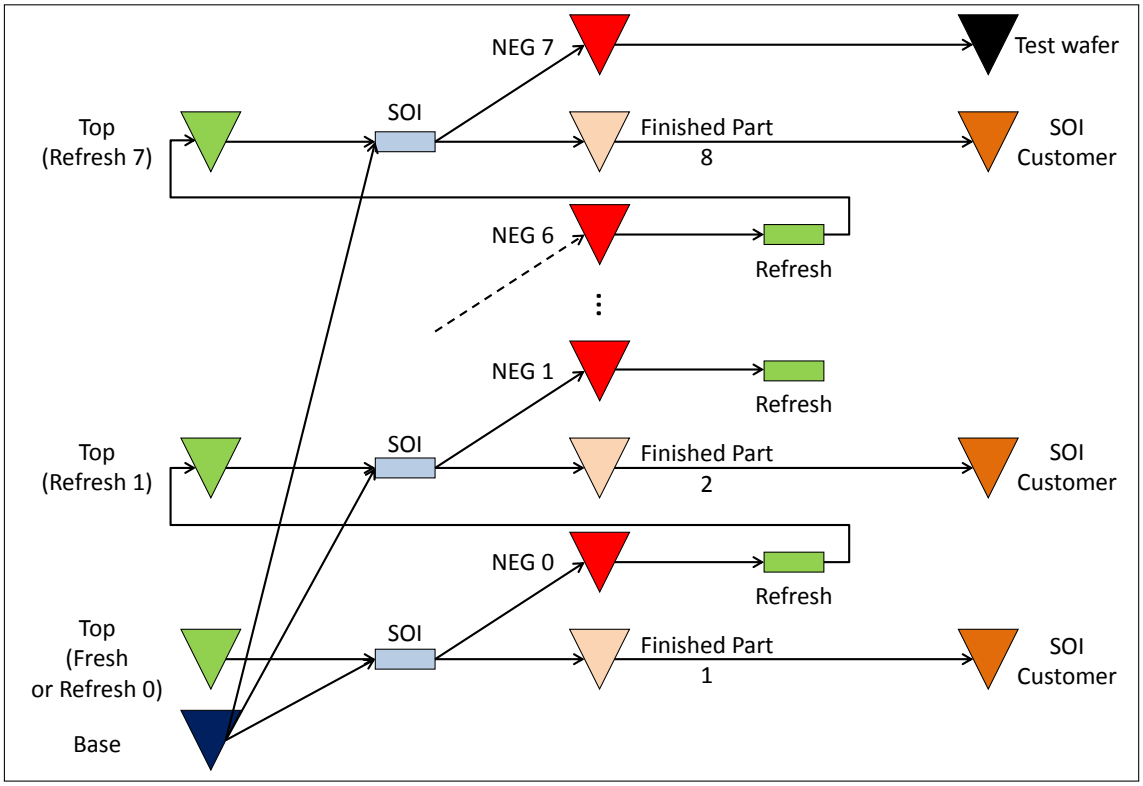


Figure 10.2: SOI Fabrication and an Example of Refresh Process up to 7 Levels ($l^{max} = 7$)

satisfy the demand of the same period.

10.3 Mathematical Model

The objective of our model is to decide when and how much to produce final products (SOI), when and how much to purchase raw materials (Base Wafer and Fresh Wafer), when and how much to refresh used Top Wafers in order to satisfy demand. The demand satisfaction is done over a discrete time horizon while minimizing the total cost (production, purchase, refresh and inventory costs). The plan must satisfy inventory, bill of materials as well as capacity constraints. The parameters, the decisions variables and the mathematical model are presented below.

The production planning model corresponding to the studied supply chain is

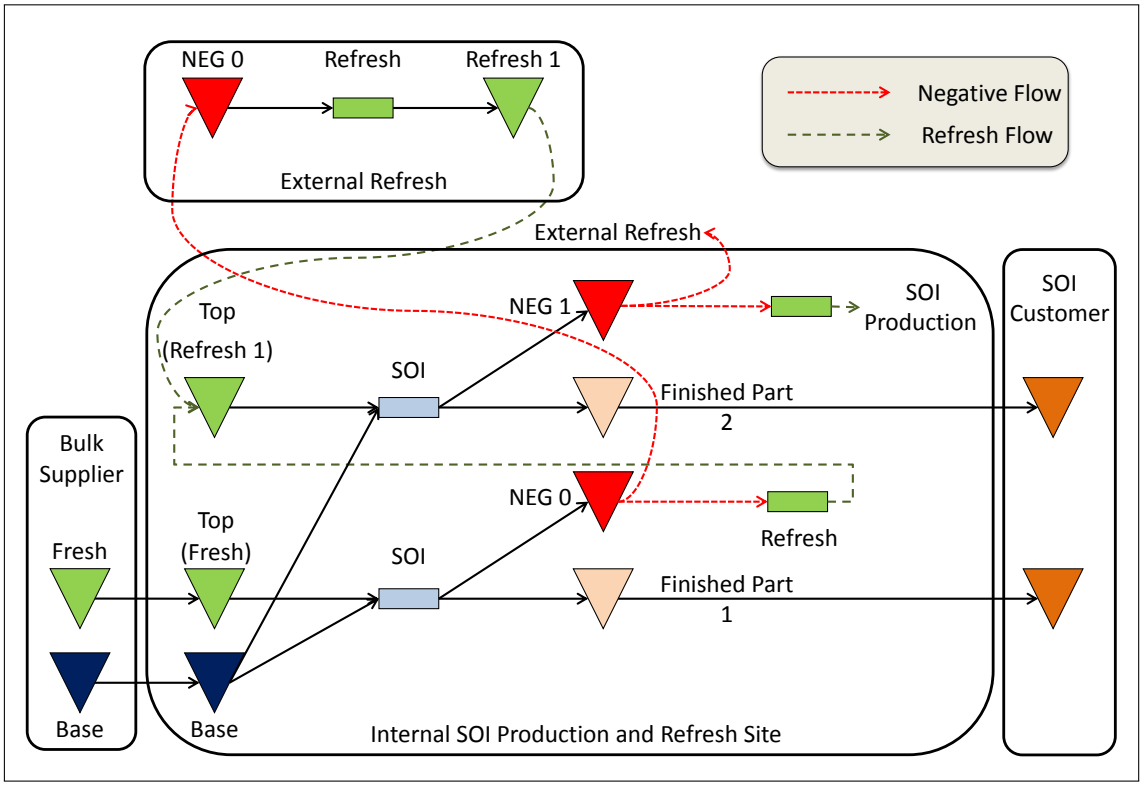


Figure 10.3: A Simple SOI Production-Refresh Supply Chain

described below. The goal is to determine the optimal raw material procurement, refreshing and production policies as well as the associated inventory levels.

Parameters

I	Total number of final products,
F	Total number of Top (Fresh and Refresh) Wafers,
B	Total number of Base Wafers,
T	Total number of periods,
M	Total number of production or refresh sites,
$d_{i,t}$	Demand of product i in period t,
n_f^{max}	Maximum refresh level of Top Wafer f,
n_f	Number of times that Top Wafer f has been refreshed,
l	Lead time of the refresh process,
$\alpha_{f,m}$	Yield of the refresh process for Top Wafer f at site m,

$a_{i,b}$ $\begin{cases} 1 \text{ if Base Wafer } b \text{ can be used in product } i, \\ 0 \text{ otherwise.} \end{cases}$

$a_{i,f}$ $\begin{cases} 1 \text{ if Top } f \text{ (Fresh or Refresh) Wafer can be used in product } i, \\ 0 \text{ otherwise.} \end{cases}$

$b_{f',f}$ $\begin{cases} 1 \text{ if Top } f' \text{ (Refresh) Wafer can be obtained via the refresh} \\ \text{process from } f \text{ (Negative of either a Fresh or Refresh Wafer),} \\ 0 \text{ otherwise.} \end{cases}$

$rc_{m,t}$ Refresh resource capacity at site m in period t,

$\beta_{f,m}$ Process time of refreshing one unit of Top Wafer f at site m,

pc_t Production resource capacity in period t,

η_i Process time of producing one unit of product i,

$S_{b,0}$ Initial inventory level of Base Wafer b,

$S^+_{f,0}$ Initial inventory level of Top Wafer f,

$S^-_{f,0}$ Initial inventory level of Negative of Top Wafer f (used Fresh or Refresh Wafer),

$S_{i,0}$ Initial inventory level of product i,

$\hat{B}_{b,t}$ Quantities of Base Wafer b already planned to be received (in transit) at the beginning of period t,

$\hat{F}_{f,t}$ Fresh quantities of f already planned to be received (in transit) at the beginning of period t,

$\hat{R}_{f,t}$ Refresh quantities of f already planned to be received (in transit) at the beginning of period t,

$h_{b,t}$ Unitary inventory cost of Base Wafer b at the end of period t,

$h^+_{f,t}$ Unitary inventory cost of Top Wafer f at the end of period t,

$h^-_{f,t}$ Unitary inventory cost of Negative of Top Wafer f at the end of period t,

$h_{i,t}$ Unitary inventory cost of product i at the end of period t,

$cp_{f,t}$ Purchase cost of Top Wafer f $(\forall f | n_f = 0)$ (Fresh Wafer) in period t,

$cp_{b,t}$ Purchase cost of Base Wafer b in period t,

$cr_{f,m,t}$ Unitary refresh cost $(\forall f | n_f \in \{1, \cdots, n_f^{max}\})$ of Top Wafer f at site m in period t (The refresh cost of all refresh levels are the same),

$cg_{i,t}$ Unitary production cost of product i in period t,
sp_t^F Fresh Wafer purchase order cost in period t,
sp_t^B Base Wafer purchase order cost in period t,
$sr_{m,t}$ Refresh setup cost at site m in period t,
sg_t Production setup cost in period t.

Variables

$G_{i,t}$ Produced quantity of product (SOI) i in period t,
$B_{b,t}$ Ordered quantity of Base Wafer b in period t,
$F_{f,t}$ Ordered quantity of Fresh Wafer f ($f|n_f = 0$) in period t,
$R_{f,f',m,t}$ Refreshed quantity f' (Refresh Wafer) obtained from Negative f at site m in period t,
$X_{i,b,t}^B$ Used quantity of Base Wafer b in period t to produce product i,
$X_{i,f,t}^F$ Used quantity of Top Wafer f in period t to produce product i,
$S_{i,t}$ Inventory level of product (SOI) i at the end of period t,
$S_{b,t}$ Inventory level of Base Wafer b at the end of period t,
$S_{f,t}^+$ Inventory level of Top Wafer f at the end of period t,
$S_{f,t}^-$ Inventory level of Negative of Top Wafer f (used Fresh or Refresh Wafer) at the end of period t,

Y_t
$$\begin{cases} 1 \text{ if production occurs in period } t, \\ 0 \text{ otherwise.} \end{cases}$$

V_t^B
$$\begin{cases} 1 \text{ if Base Wafer procurement occurs in period } t, \\ 0 \text{ otherwise.} \end{cases}$$

V_t^F
$$\begin{cases} 1 \text{ if Fresh Wafer procurement occurs in period } t, \\ 0 \text{ otherwise.} \end{cases}$$

$W_{m,t}$
$$\begin{cases} 1 \text{ if refresh process is performed in site } m \text{ in period } t, \\ 0 \text{ otherwise.} \end{cases}$$

Mathematical model

$$\begin{aligned}
\min \sum_{\forall f|n_f=0} \sum_{\forall t} cp_{f,t}F_{f,t} + \sum_{\forall b}\sum_{\forall t} cp_{b,t}B_{b,t} \\
+ \sum_{\forall m} \sum_{\forall f|n_f>0} \sum_{\forall f'}\sum_{\forall t} cr_{f,m,t}R_{f,f',m,t} + \sum_{\forall i}\sum_{\forall t} cg_{i,t}G_{i,t} \\
+ \sum_{\forall t} sp_t^B V_t^B + \sum_{\forall t} sp_t^F V_t^F + \sum_{\forall m}\sum_{\forall t} sr_{m,t}W_{m,t} + \sum_{\forall t} sg_t Y_t \\
+ \sum_{\forall b}\sum_{\forall t} h_{b,t}S_{b,t} + \sum_{\forall f}\sum_{\forall t} h_{f,t}^+ S_{f,t}^+ + \sum_{\forall f|n_f\neq n_f^{max}} \sum_{\forall t} h_{f,t}^- S_{f,t}^- + \sum_{\forall i}\sum_{\forall t} h_{i,t}S_{i,t} \qquad (10.1)
\end{aligned}$$

subject to

$$S_{i,t-1} + G_{i,t} = d_{i,t} + S_{i,t} \qquad \forall i,t \qquad (10.2)$$

$$\sum_{\forall b} a_{i,b}X_{i,b,t}^B = G_{i,t} \qquad \forall i,t \qquad (10.3)$$

$$\sum_{\forall f} a_{i,f}X_{i,f,t}^F = G_{i,t} \qquad \forall i,t \qquad (10.4)$$

$$\hat{B}_{b,t} + S_{b,t-1} + B_{b,t} - \sum_{\forall i} X_{i,b,t}^B = S_{b,t} \qquad \forall b,t \qquad (10.5)$$

$$\hat{F}_{f,t} + S_{f,t-1}^+ + F_{f,t} - \sum_{\forall i} X_{i,f,t}^F = S_{f,t}^+ \qquad \forall f|n_f = 0,t \qquad (10.6)$$

$$\hat{R}_{f,t} + S_{f,t-1}^+ + \sum_{\forall m}\sum_{\forall f'|b_{f,f'}=1} \alpha_{f,m}R_{f,f',m,t-l} - \sum_{\forall i} X_{i,f,t}^F = S_{f,t}^+ \qquad \forall f|n_f \neq 0,t \qquad (10.7)$$

$$S_{f,t-1}^- + \sum_{\forall i} X_{i,f,t}^F - \sum_{\forall m}\sum_{\forall f'} b_{f,f'}R_{f,f',m,t} = S_{f,t}^- \qquad \forall f|n_f \neq n_f^{max},t \qquad (10.8)$$

$$\sum_{\forall i} \eta_i G_{i,t} \leq pc_t \qquad \forall t \qquad (10.9)$$

$$\sum_{\forall f|n_f \neq 0} \sum_{\forall f'} \beta_{f',m} R_{f,f',m,t} \leq rc_{m,t} \qquad \forall m,t \qquad (10.10)$$

$$\sum_{\forall i} G_{i,t} \leq M \cdot Y_t \qquad \forall t \qquad (10.11)$$

$$\sum_{\forall b} B_{b,t} \leq M \cdot V_t^B \qquad \forall t \qquad (10.12)$$

$$\sum_{\forall f|n_f=0} F_{f,t} \leq M \cdot V_t^F \qquad \forall t \qquad (10.13)$$

$$\sum_{\forall f'} \sum_{\forall f|n_f>0} R_{f',f,m,t} \leq M \cdot W_{m,t} \qquad \forall m,t \qquad (10.14)$$

$$R_{f,f',m,t} = 0 \qquad \forall m,t,f,f'|b_{f,f'} = 0 \qquad (10.15)$$

$$R_{f,f',m,T} = 0 \qquad \forall m,f,f' \qquad (10.16)$$

$$Y_t,\ V_t^B,\ V_t^F \in \{0,1\} \qquad \forall t \qquad (10.17)$$

$$W_{m,t} \in \{0,1\} \qquad \forall m,t \qquad (10.18)$$

$$G_{i,t} \geq 0 \qquad \forall i,t \qquad (10.19)$$

$$S_{b,t},\ B_{b,t} \geq 0 \qquad \forall b,t \qquad (10.20)$$

$$S_{f,t}^+,\ S_{f,t}^-,\ S_{i,t},\ F_{f,t} \geq 0 \qquad \forall f,t \qquad (10.21)$$

$$R_{f,f',m,t} \geq 0 \qquad \forall f,f',m,t \qquad (10.22)$$

$$X_{i,b,t}^B \geq 0 \qquad \forall i,b,t \qquad (10.23)$$

$$X_{i,f,t}^F \geq 0 \qquad \forall i,f,t \qquad (10.24)$$

The objective function (10.1) minimizes the total cost which is the sum of the purchase cost of new Top Wafers (Fresh Wafers) as well as Base Wafers, refresh cost at all sites, final products (SOI Waferss) production cost, raw material (Bulk, i.e. Fresh and Base Wafers) procurement cost, refresh setup cost, production setup cost, and inventory costs of Base Wafers, Top Wafers (either Fresh or Refresh Wafers), generated Negative Wafers and final products (SOI Wafers).

Constraint (10.2) models the flow conservation of finished goods. Constraints (10.3) and (10.4) respectively determine the amount of Base Wafers and Top Wafers (either Fresh or Refresh Wafers) which are used to satisfy the production plan of

product i in period t ($G_{i,t}$). Constraints (10.5) and (10.6) respectively model the flow conservation for Base Wafers and Fresh Wafers. Constraint (10.7) refers to Refresh Wafer inventory balance. It indicates that the inventory of the Refresh Wafers f in period t ($S^{+}_{f,t}$) is equal to the Refresh Wafers to be received in this period (in transit Refresh Wafers) ($\hat{R}_{f,t}$) plus the Refresh Wafer inventory in the previous period ($S^{+}_{f,t-1}$) plus the Refresh Wafers of that period in all sites (refreshed Negative Wafers obtained in period t) ($\sum_{\forall m} \sum_{\forall f' | b_{f,f'}=1} \alpha_{f,m} R_{f,f',m,t-l}$) minus the used Top Wafers of f in that period ($\sum_{\forall i} X^{F}_{i,f,t}$). Constraint (10.8) models the inventory balance of Negative Wafers in period t. It specifies that the inventory of the Negative of the Top Wafer f in period $t-1$ ($S^{-}_{f,t-1}$) is equal to the Negative inventory of the Top Wafer f at the previous period ($S^{-}_{f,t-2}$) plus the Negatives generated at period $t-1$ ($\sum_{\forall i} X^{F}_{i,f,t-1}$) minus the Negative Wafers sent to be refreshed in all sites ($\sum_{\forall m} \sum_{\forall f'} b_{f,f'} R_{f,f',m,t}$). Note that the Negative Wafers of f taken in Constraint (10.8) are returned back refreshed in (10.7). However, because of the refresh line scraps, not all of the Negative Wafers are transformed into Refresh Wafers. This is why the yield factor $\alpha_{f,m}$ is used in Constraint (10.7).

Constraints (10.9) and (10.10) respectively restrict the production and refresh line capacities in each period t. Constraints (10.11) through (10.14) respectively model the production setup cost, raw material (Base Wafers and Fresh Wafers) procurement cost and refresh setup cost, where M is a large positive number (classical big-M in mixed integer linear programming formulations). Constraint 10.15 is added to prevent refreshing non-refreshable Negative Wafers. Constraint 10.16 avoids refreshing in the last period of the planning horizon. Binary and non-negativity sign restrictions are ensured using Constraints (10.17) through (10.24).

Note that, if Wafers f' can be obtained from several Wafers f ($\sum_{\forall f'} b_{f,f'} > 1$), the condition $b_{f,f'} = 1$ is added to avoid removing several times f to obtain Refresh Wafer f'. However if $\sum_{\forall f'} b_{f,f'} = 1 \quad \forall f$, there exists only one f for each f' and, therefore, only one unit of f' is removed by removing one unit of f. In this case, the refresh terms in Constraints (10.7), (10.8), (10.10) and the objective function could simply be written as $R_{f,m,t}$ instead of $R_{f',f,m,t}$.

10.4 Complexity Analysis

An instance of our problem is NP-hard without considering its bi-level nature and all refresh constraints, only because resource capacities are time-dependent. In fact, [29] and [9] demonstrate that the classical capacitated single-item problem is NP-hard in the weak sense.

Moreover, in Section 11.3.1, we show that the problem in the mono-level and single-item case (i.e. only one raw material with its refresh levels) without capacity constraints is still NP-hard.

10.5 Industrial Extensions

Here, we address some industrial specificities and restrictions in the studied SOI supply chain. Some of them are already treated in the model. Otherwise, the necessary constraints are presented.

- **Market share rules.** The procurement department negotiates based on long-term agreement the bulk consumption demand from a supplier. These agreements define the "market share" rules which aim at increasing material procurement reliability, decreasing lead-time variability and guaranteeing a better acquisition price. The market share rules define that the bulk purchased from a supplier must at least be x percent of the total bulk consumption. While planning, it is desired that the market share is considered and, in case of violation, that the gap is minimized.

  The $\bullet$ character represents F and B.

Parameters

Q	Total number of suppliers,
$a_{q,b}$	$\begin{cases} 1 \text{ if Base Wafer } b \text{ can be purchased from supplier } q, \\ 0 \text{ otherwise.} \end{cases}$
$a_{q,f}$	$\begin{cases} 1 \text{ if Fresh Wafer } f \text{ can be purchased from supplier } q, \\ 0 \text{ otherwise.} \end{cases}$
$UB^{\bullet}_{q,t}$	Market share upper bound for supplier q in period t $(\sum_{q=1}^{Q} UB^{\bullet}_{q,t} \geq 1 \quad \forall t)$,
$LB^{\bullet}_{q,t}$	Market share lower bound for supplier q in period t $(\sum_{q=1}^{Q} LB^{\bullet}_{q,t} \leq 1 \quad \forall t)$.

Variables

$B_{q,b,t}$	Ordered quantity of Base Wafer b to supplier q in period t,
$F_{q,f,t}$	Ordered quantity of Fresh Wafer f $(f\|n_f = 0)$ to supplier q in period t.

If market share is considered for Top and Base Wafers separately.

$$\sum_{q=1}^{Q} a_{q,b} B_{q,b,t} = B_{b,t} \qquad \forall b,t \qquad (10.25a)$$

$$\sum_{b=1}^{B} LB^{B}_{q,t} B_{b,t} \leq \sum_{b=1}^{B} a_{q,b} B_{q,b,t} \leq \sum_{b=1}^{B} UB^{B}_{q,t} B_{b,t} \qquad \forall t,q \qquad (10.25b)$$

$$\sum_{q=1}^{Q} a_{q,f} F_{q,f,t} = F_{f,t} \qquad \forall f,t \qquad (10.26a)$$

$$\sum_{f=1}^{F} LB^{F}_{q,t} F_{f,t} \leq \sum_{f=1}^{F} a_{q,f} F_{q,f,t} \leq \sum_{f=1}^{F} UB^{F}_{q,t} F_{f,t} \qquad \forall t,q \qquad (10.26b)$$

If market share is considered at the same time for Top and Base Wafers.

$$\sum_{q=1}^{Q} a_{q,b} F_{q,b,t} = B_{b,t} \quad \forall b, t \quad (10.27a)$$

$$\sum_{q=1}^{Q} a_{q,f} F_{q,f,t} = F_{f,t} \quad \forall f, t \quad (10.27b)$$

$$LB_{q,t}(\sum_{f=1}^{F} F_{f,t} + \sum_{b=1}^{B} B_{b,t}) \leq \sum_{f=1}^{F} a_{q,f} F_{q,f,t} + \sum_{b=1}^{B} a_{q,b} B_{q,b,t} \quad \forall t, q \quad (10.27c)$$

$$\sum_{f=1}^{F} a_{q,f} F_{q,f,t} + \sum_{b=1}^{B} a_{q,b} B_{q,b,t} \leq UB_{q,t}(\sum_{f=1}^{F} F_{f,t} + \sum_{b=1}^{B} B_{b,t}) \quad \forall t, q \quad (10.27d)$$

- **Purchase-refresh constraint** Some bulk suppliers also propose the refresh process (external refresh). The suppliers do not agree that their competitors manipulate their wafers. Therefore, Fresh Wafers bought from one specific supplier may only be refreshed internally or by the same supplier; and not by one of its competitors.

  We add the index q to the refresh variable to consider the origin of the Top Wafer. We also need a matrix which defines the relationship between suppliers and production or refresh sites.

  Parameters

  $a_{q,m}$ $\begin{cases} 1 \text{ if supplier } q \text{ and production or refresh site m} \\ \text{are compatible,} \\ 0 \text{ otherwise.} \end{cases}$

  Variables

  $R_{f,f',q,m,t}$ Refreshed quantity f' (Refresh Wafer) obtained from Negative f purchased originally from supplier q at site m in period t.

  In constraints (10.7) and (10.8) as well as the objective function (10.1), the terms $R_{f,f',m,t}$ must be replaced with $a_{q,m} R_{f,f',q,m,t}$ while summing over q.

- **Raw materials, remanufactured by-products and final product connectivity** It may happen that the Refresh Wafers (at all levels) of one Fresh Wafer are used to produce an SOI product but the Fresh itself is not used. This is due to yield issues. A Fresh Wafer has a different geometry from all its Refresh Wafers. The refresh process acts as a normalizing process. This means that once a Negative Wafer 1 is refreshed, it has a different geometry than its previous Fresh Wafer and all the subsequent Refresh Wafers have the same geometry. Therefore, for some Top Wafers, the Fresh Wafer is used to produce one SOI reference and all its Refresh Wafers are used to produce other SOI reference(s). In this way, the yield is increased. In other words, once the Negative of a Fresh is refreshed, it becomes "connectable" to all references for which it was prohibited. In another case, there exist products which use only Fresh Wafers and their refreshes may not at all be used for any other product. All these constraints can be modeled using the parameters $a_{i,b}$, $a_{i,f}$ and $b_{f',f}$.

  For some products, a Base Wafer b is associated with a Top Wafer f for the production of the SOI Wafer i.

  $$a_{b,f} \begin{cases} 1 \text{ if Base Wafer } b \text{ can be used with Top Wafer } f, \\ 0 \text{ otherwise.} \end{cases}$$

  $$a_{b,f,i} \begin{cases} 1 \text{ if Base Wafer } b \text{ can be used with Top Wafer } f \text{ to produce} \\ \text{SOI product } i, \\ 0 \text{ otherwise.} \end{cases}$$

- **Maximum refresh level** Due to SOI customer requirements or product characteristics, not all refresh levels of a Fresh Wafer are allowed to be used. Therefore the maximum refresh level is determined from two standpoints. First, the Fresh Wafer f may be refreshed n_f^{max} times. Second, an SOI Wafer i may be produced from the Refresh Wafers of the Fresh Wafer f until $n_{f\to i}^{max}$ level ($n_{f\to i}^{max} \leq n_f^{max}$). If $n_{f\to i}^{max} < n_f^{max}$, the Refresh Wafers of f not used for SOI product i may be used to produce another SOI product i'. This constraint can be modeled using

the parameter $a_{i,f}$.

- **Yield monitoring - Minimum Fresh Wafer usage** In the SOI production line, there must be a minimum level of Fresh Wafers. In other words, the SOI production line must not be run with *only* Refresh Wafers. This is related to the SOI production line yield monitoring. Once a Fresh Wafer is refreshed (even once), its geometry and characteristics change. Having this in mind, when only Refresh Wafers are used in the SOI production line, if the line performance drops, it is difficult to judge whether the problem comes from the SOI production line or from Refresh Wafers. Hence, for securing the yield monitoring of the SOI production line, a minimum level of Fresh Wafers are to be used with Refresh Wafers. This constraint can be satisfied using a minimum production level as in Constraint (10.28).

$$\sum_{\forall i} X^F_{i,f,t} \geq F_{min} \qquad \forall f | n_f = 0, t \qquad (10.28)$$

- **Monitor wafers.** The Negative Wafer of the last level may also be refreshed and used as monitor wafer. The monitor wafers may be used internally or sold to external companies. However, as monitor wafers have less value than Refresh Wafers, the refresh process line capacity must first be allocated to get Refresh Wafers and then -if capacity remains unused- to get monitor wafers.

10.6 Numerical Experiments

10.6.1 Data Sets

Based on industrial data, small, medium and large test instances are constructed to run experiments on the model and study its behavior. The parameters for generating data sets are listed in Table 10.1. In order to avoid infeasibility, coefficients

called *capacity tightness factors* (CTF), q^p and q^r are used for adjusting production and refresh capacities respectively. Both of the capacity tightness factors are fixed to 1.0 (tight), 1.2 (normal), 1.6 (large), and 2.0 (very large). Initial inventories and scheduled wafers to be received (in transit) are set to zero. The refresh process can be performed in different sites. Therefore, the refresh cost and refresh setup cost ($cr_{f,m,t}$ and $sr_{m,t}$) are defined based on the refresh site m. The refresh site can be internal, external, close or remote.

The instances are generated by fixing one of the parameters and by considering all the combinations of the other parameters. In total, 2304 instances are generated. The reduced Mixed Integer linear Program (MIP) of the smallest instance has 2383 constraints and 9822 variables from which 280 are binaries, whereas the reduced MIP of the largest instance has 16452 constraints and 149356 variables from which 526 are binaries.

10.6.2 Experimental Results

The test instances are solved using *IBM ILOG CPLEX™ 12.5.1*. All computational experiments have been run on an AMD Phenom II X2 B57 3.20 GHz with 3.24GB of RAM. The relative MIP gap tolerance is set to 0.5%. A summary of the results can be found in Tables 10.2 and 10.3, where the average results of all instances are provided for a given value of each of the parameters CTF, T, I, F, B and M.

Table 10.2 shows the average resolution time. Large CPU Times are observed and rather independently of the variations of most parameters. A significant increase of the CPU time is observed when the length of the planning horizon increases. The average CPU Time is multiplied by more than 17 when the number of periods increases from 6 to 48. The CPU Time also significantly increases when the number of Top Wafer references increases. Table 10.3 shows the percentage of each cost component in the total optimal cost. The percentage of Base Wafer purchase and procurement cost is significantly larger than the percentage of the Fresh Wafer purchase and procurement cost. The reason is that less Fresh Wafer procurement is needed as Refresh Wafers are produced using the refresh process. However, since

Table 10.1: Parameters for generating data sets

Parameter	**Value**
I	10, 20, 50, 100
F	6, 12, 18
B	4, 5, 6, 7
T	6, 12, 24, 48
M	2, 3, 4
$d_{i,t}$	Uniformly drawn from$[1000, 3000]$
n_f^{max}	5
l	1
$a_{i,b}$	$a_{i,b} \in [0,1] \vert P(a_{i,b} = 1) = 0.90$
$a_{i,f}$	$a_{i,f} \in [0,1] \vert P(a_{i,f} = 1) = 0.90$
$b_{f',f}$	$b_{f',f} \in [0,1] \vert P(b_{f',f} = 1) = 0.70$
$\alpha_{f,m}$	0.98
$h_{b,t}$	1
$h_{f,t}^{+}$	2
$h_{f,t}^{-}$	2
$h_{i,t}$	4
$rc_{m,t}$	$q^r(\sum_{\forall i} d_{i,t}/(\alpha_{f,m} n_f^{max}))$
$\beta_{f,m}$	1
pc_t	$q^p(\sum_{\forall i} d_{i,t})$
η_i	1
$cg_{i,t}$	150
sg_t	150000
$cr_{f,m,t}$	20, 30, 40, 50
$sr_{m,t}$	40000, 80000, 120000, 160000
$cp_{b,t}$	50
$cp_{f,t}$	150
sp_t^B	30000
sp_t^F	30000

the refresh process is relatively cheaper than Fresh Wafer procurement, relatively small refresh process and setup costs are incurred. The largest production and setup costs concern the SOI fabrication. Inventory costs are negligible for each cost component even if inventory costs are slightly larger for SOI production. Except for the length of the planning horizon (T), the variation of all other parameters do

Table 10.2: Average resolution times

Parameters	Result (Avg)
CTF	**CPU Time (sec.)**
1	4
1.2	4
1.6	4
2	4
T	
6	**1**
12	**2**
24	**4**
48	**11**
I	
10	4
20	3
50	4
100	**7**
F	
6	**2**
12	**3**
18	**6**
B	
4	4
5	4
6	4
7	4
M	
2	**4**
3	**4**
4	**5**

not significantly alter the percentages of the cost components. By increasing the planning horizon, the Fresh Wafer purchase (and also procurement and inventory) cost drastically reduces while the refresh process and setup costs increase. Note that the maximum refresh level n_f^{max} is 5. This means that a newly bought Fresh Wafer can only be refreshed five times. As the lead time is equal to one period

Table 10.3: Cost components for different data sets

Parameters		Results											
		Top (Fresh) Wafer Procurement			Base Wafer Procurement			SOI Production			Refresh Process		
CTF		Purchase	Procurement	Inv.	Purchase	Procurement	Inv.	Production	Setup	Inv.	Refresh	Setup	Neg. Inv.
	1	9.36%	0.03%	0.01%	20.43%	0.20%	0.00%	61.30%	1.02%	0.34%	7.07%	0.23%	0.00%
	1.2	9.33%	0.03%	0.01%	20.43%	0.21%	0.00%	61.29%	1.04%	0.36%	7.07%	0.23%	0.00%
	1.6	9.06%	0.03%	0.01%	20.50%	0.20%	0.00%	61.49%	1.01%	0.34%	7.13%	0.23%	0.00%
	2	9.17%	0.03%	0.01%	20.46%	0.21%	0.00%	61.39%	1.03%	0.36%	7.11%	0.24%	0.00%
T													
	6	**18.63%**	0.06%	0.02%	**18.65%**	0.18%	0.00%	**55.95%**	0.92%	0.34%	**5.08%**	0.16%	0.00%
	12	**10.17%**	0.03%	0.01%	**20.28%**	0.20%	0.00%	**60.83%**	1.01%	0.35%	**6.89%**	0.22%	0.00%
	24	**5.32%**	0.02%	0.01%	**21.20%**	0.21%	0.00%	**63.60%**	1.07%	0.37%	**7.93%**	0.26%	0.00%
	48	**2.72%**	0.01%	0.00%	**21.71%**	0.22%	0.00%	**65.14%**	1.08%	0.34%	**8.49%**	0.28%	0.00%
I													
	10	9.05%	0.05%	0.02%	20.19%	0.31%	0.01%	60.57%	1.54%	0.91%	7.01%	0.35%	0.00%
	20	9.33%	0.04%	0.01%	20.37%	0.26%	0.00%	61.11%	1.29%	0.26%	7.05%	0.29%	0.00%
	50	9.23%	0.02%	0.00%	20.67%	0.12%	0.00%	62.01%	0.62%	0.00%	7.18%	0.14%	0.00%
	100	9.38%	0.01%	0.00%	20.74%	0.06%	0.00%	62.23%	0.31%	0.00%	7.19%	0.07%	0.00%
F													
	6	9.32%	0.04%	0.03%	20.22%	0.27%	0.00%	60.66%	1.35%	0.78%	6.98%	0.33%	0.01%
	12	9.18%	0.03%	0.01%	20.45%	0.21%	0.00%	61.36%	1.03%	0.39%	7.10%	0.23%	0.00%
	18	9.26%	0.03%	0.01%	20.48%	0.20%	0.00%	61.45%	0.98%	0.27%	7.10%	0.22%	0.00%
B													
	4	9.20%	0.03%	0.01%	20.47%	0.20%	0.00%	61.40%	1.00%	0.35%	7.10%	0.23%	0.00%
	5	9.14%	0.03%	0.01%	20.47%	0.21%	0.00%	61.42%	1.03%	0.34%	7.11%	0.23%	0.00%
	6	9.30%	0.03%	0.01%	20.44%	0.21%	0.00%	61.32%	1.03%	0.36%	7.08%	0.23%	0.00%
	7	9.27%	0.03%	0.01%	20.45%	0.21%	0.00%	61.34%	1.03%	0.34%	7.08%	0.23%	0.00%
M													
	2	9.38%	0.03%	0.01%	20.42%	0.20%	0.00%	61.27%	1.02%	0.37%	7.06%	0.23%	0.00%
	3	9.11%	0.03%	0.01%	20.48%	0.20%	0.00%	61.45%	1.02%	0.34%	7.12%	0.23%	0.00%
	4	9.20%	0.03%	0.01%	20.46%	0.21%	0.00%	61.39%	1.02%	0.34%	7.10%	0.23%	0.00%

(without any SOI and refresh capacity restrictions), it takes 7 periods to fully use a purchased Fresh Wafer (one period for Negative generation and one for the refresh process). Therefore, as the planning horizon increases, the refresh process becomes more important. Therefore, when the planning horizon increases, the Fresh Wafer procurement cost decreases and the refresh process cost increases. However, as the Fresh Wafer purchase cost is relatively larger than the refresh process cost, the decrease of the Fresh Wafer procurement cost components is larger than the increase of the refresh process cost components. This illustrates the economical importance of the refresh process and that an efficient production planning contributes to a substantial cost decrease. Note that, as the percentage of the Fresh Wafer procurement cost decreases, the Base Wafer procurement and SOI production cost percentages increase in the total planning cost.

10.6.3 Closed-Loop Production Planning

Now that we have a clear idea of the problem, it is possible to make a more detailed comparison of the presented research with the literature. We skimmed through

the ideas of Closed-Loop Supply Chain (9.3.2) and product recovery (9.3.3.2). In production systems with remanufacturing or some kind of reverse material flow, the production planning system must be consequently adapted.

[42] and [40] enumerate the complicating characteristics of remanufacturing. We merely mention them by telling to which extend the problem we tackle is different.

The complicating characteristics of remanufacturing based on [40] are:

- **The uncertain timing and quantity of returns.** This is not true in our case, since by-products are generated after the splitting step. So the timing can be estimated, and the quantity depends on the yield of the production steps until splitting.
- **The need to balance returns of items from consumers with demand for remanufactured items.** As the demand can be satisfied using the Fresh Wafers as well as Refresh Wafers, this problem does not really exist.
- **The need to disassemble the returns.** There is no disassembly in our case.
- **The uncertainty in materials recovered from returned products.** According to the refresh line yield, we can exactly estimate the successfully refreshed Negative Wafers (by products).
- **The requirement for a reverse logistics network.** In case of refresh of by-products generated in the same site, there is no need for a logistics network.
- **The complications of material matching restrictions.** The Top Wafer-Finished product matching restrictions are modeled using parameters ($a_{i,b}$, $a_{i,f}$ and $b_{f',f}$) in Section 10.3.
- **The problems of stochastic routings for materials for remanufacturing operations and highly variable processing times.** In our case, all routings are deterministic.

For a more recent survey, analysis and classification of the literature treating production planning and remanufacturing, see [57].

10.7 Conclusions and Perspectives

In this chapter, the supply chain of a SOI fabrication unit using the Smart-Cut™ Technology was modeled. Using this technology, one of the two purchased raw materials can be used several times after reprocessing. The reprocessing which is considered as a kind of "remanufacturing" can be done internally or externally. A mixed-integer linear program (MILP) is proposed to model the production planning problem. The model is validated and the optimal solution behavior is studied using generated data sets based on industrial data.

The refresh capacity is constrained by both the refresh line capacity and the SOI production line capacity. In fact, the SOI production determines the rate of the generation of Negative Wafers. Therefore, the refresh process throughput depends on the available refresh capacity and the SOI production (generation of Negative Wafers) rate.

Due to the purchase and refresh cost structure, optimal solutions in our numerical experiments usually include purchase and refresh campaigns, which cause an irregular and fluctuating cost profile along the planning horizon. This may not be desirable both financially and for workforce management, for which a more stable cost expense over the whole planning horizon is preferable. A possible approach is to use a non-linear cost objective function in order to make the sum of all costs closer to their average. However, it makes the model non-linear and its resolution much more difficult.

Chapter 11

Single-Item Bi-Level Uncapacitated Lot-Sizing with Multiple Remanufacturing of Reusable By-Products

In this chapter, we study bi-level lot-sizing problems in which reusable by-products are generated during production. The generated by-products can be reused as raw materials after further processing. However, by-products can be "recycled" only a given number of times. The industrial problem concerns the production system of "SOI" (Silicon-On-Insulator) fabrication units. Based on the industrial model discussed in Chapter 10, an uncapacitated single-item version of the problem is derived and analyzed. We also propose a dynamic programming algorithm for the resolution of a restricted version of the problem.[1]

[1]Part of this chapter has been presented in IWLS 2014 [80].

11.1 Introduction

Sustainable development has come into attention due to environmental and economical issues. Closed-loop supply chain and product recovery are emerging fields in the domain of sustainable development. Reverse flow of materials and remanufacturing add complexity to production systems. Planning strategies must be adapted to globally optimize production flows.

We introduce a problem inspired from an industrial case in semiconductor manufacturing. Silicon-On-Insulator (SOI) wafers are used instead of silicon-only wafers in semiconductor manufacturing where high performance and efficiency is needed. SOI wafers can be produced using different technologies. Among them, *Smart-Cut™ Technology* is a very economic and efficient way of fabricating SOI Wafers. Using this technology, a thin crystalline layer of a so-called donor wafer (Top Wafer) is laid on another silicon wafer (called the handle, support or Base Wafer) using bonding and splitting processes [15]. As only a thin layer of the Top Wafer is transferred to the final product (SOI Wafer), the Top Wafer can be reused to produce other SOI Wafers. Before reusing the *used Top Wafer* (the by-product) in the SOI production line, it must be reprocessed (or remanufactured). The production and remanufacturing processes form a so-called "closed-loop manufacturing system".

In this chapter, we discuss the production planning of closed-loop manufacturing systems, an economic lot-sizing problem with multiple remanufacturing of reusable by-products. The discussed problem can be enhanced to model more complicated real world problems.

In the next section, the problem is formally set. The mathematical model is presented and analyzed in Section 11.3. In Sections 11.4 and 11.5 simplified versions of the problem are presented and thoroughly discussed. The chapter is closed with conclusions and multiple perspectives.

11.2 Problem Definition

The academic problem we tackle is an uncapacitated single-item bi-level lot-sizing problem in which by-products are generated during production. The generated by-

products can be remanufactured to be reused again to satisfy the final product demand. Final products fabricated from both newly purchased and remanufactured raw materials can be used to fulfill the same demand. However, a newly purchased raw material may be remanufactured and reused only a limited number of times.

As stated, we define the problem based on SOI production using the Smart-Cut™ Technology. The interested reader can refer to Section 10.2 for a more detailed discussion of the manufacturing process. Here, we only define the necessary elements for our discussion in this chapter.

In SOI production, a thin layer of a donor wafer (Top Wafer) is transferred to a support wafer (Base Wafer). As only a thin layer of the Top Wafer is deposed on the final product (SOI), the Top Wafer may be reused several times. The purchased raw material -which is not yet used in production- is called "Fresh Wafer". The generated by-product during SOI production is called "Negative Wafer". The Negative Wafers are not immediately reusable as raw material. The process of making by-products reusable is called the "refresh process" or shortly "refresh". The refreshed Negative Wafer which can be used again as raw material in production is called "Refresh Wafer". A purchased raw material (Fresh Wafer) has a limited refresh (or remanufacturing) life. It means that the generated by-product (Negative Wafer) can be reused (refreshed) only a limited number of times. The production and refresh flows are is depicted in Figure 11.1. The Base Wafer flow belongs to

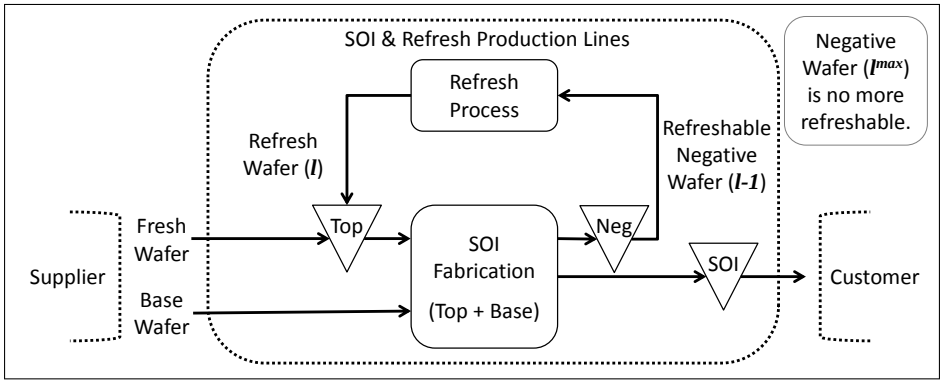
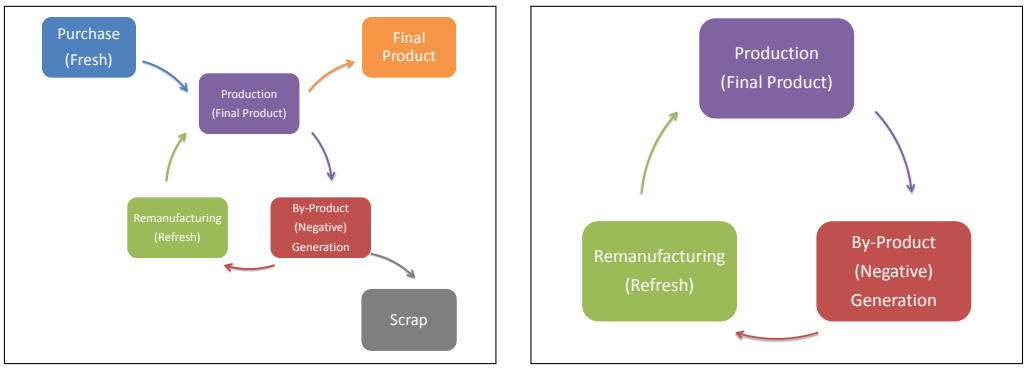


Figure 11.1: Simplified Production and Refresh Flow Schema of a SOI Fabrication Unit Using the Smart-Cut™ Technology

classic lot-sizing problems. Therefore, we consider only the Top Wafer flow. The

whole cycle is shown in Figure 11.2a. The core of the problem which makes the production planning complicated is the internal cycle (Figure 11.2b). We consider



(a) Whole Cyle (b) Internal Cycle

Figure 11.2: SOI Fabrication and Refresh Process Cycles

only one product and one raw material which is reusable (refreshable) l^{max} times. A newly purchased Top Wafer is called a *Fresh Wafer*. Once used, it is called *Negative Wafer 1x* or *Neg 1*. The remanufacturing process to make the Negative Wafer usable in the SOI manufacturing is called the *refresh process* or shortly *refresh*. The obtained wafer from *Neg 1* is called *Refresh Wafer 2x*. The parameter l (refresh level) indicates how many times the raw material has been used and how many more times before reaching its maximum refresh level l^{max}. Concerning Top Wafers, $l = 0$ indicates that the raw material is a Fresh Wafer (has never been used), and $l > 0$ that the raw material is a Refresh Wafer. In Negative Wafers, $l = 0$ designate the by-product generated after SOI production using a Fresh Wafer. Negative Wafers of level l^{max} are no longer refreshable.

It is assumed that the replenishment, production and refresh alternatives are restricted to the beginning of each period[2]. The lead time of the refresh process is known with certainty and is considered to be one period. We suppose that the Negative Wafer generated during a production period can be refreshed at the same period but will only be available at the beginning of the next period. The generation (return) of the by-products is deterministic. By fabricating one final product, one by-product is generated. The demand rate may vary from period to period, but it is

[2]The case where replenishment decisions can be made at any time is discussed in [88].

known. The unit variable purchase, production and refresh costs do not depend on the purchase, production or refresh quantity; i.e. no discounts or economy of scale exist. No capacity restriction is considered for production or for refresh process. Fresh Wafer purchase cost is implicitly considered to be larger than the refresh process cost. Otherwise, the refresh process loses its economic interest and the problem reduces to a classical economic lot-sizing problem. All Fresh Wafer ordering, final product demand, final product fabrication occur at the beginning of the period. While inventory costs are charged based on the end of the period inventory. No shortages are permitted. The entire replenishment, production and refresh happen at once. With these assumptions, we define a basic model which can be used for more complicated studies.

11.3 Uncapacitated Lot-Sizing with Multiple Remanufacturing (ULS-MR)

In this section we first present an *Uncapacitated Single-Item Bi-Level Lot-Sizing Problem with Multiple Remanufacturing of Reusable By-Products* model (hereafter, *ULS-MR*). We show that the ULS-MR problem is NP-hard. Let us simplify the model by omitting two binary setup variables (corresponding to final product fabrication and by-product remanufacturing) while leaving only one (corresponding to purchase) to obtain a problem with a *single setup*. The properties of the optimal solution are discussed after. In order to make the model still more tractable and find the optimal solution through a dynamic programming algorithm, we make some assumptions, called "Full Push Policy Assumptions" (ULS-$MR_1{}^{FP}$). ULS-$MR_1{}^{FP}$ offers additional properties. Based on these properties, a model is presented in Section 11.5. The resolution approach is given in Section 11.5.3.

Parameters

T	Total number of periods in the planning horizon,
l^{max}	Maximum refresh level of the Top Wafer,
d_t	Demand of final product (SOI) in period t,
cf_t	Fresh Wafer purchase cost in period t,
cr_t	Refresh process cost in period t,
cp_t	Production cost of the final product (SOI) in period t,
csf_t	Setup cost of Fresh Wafer purchase in period t,
csr_t	Setup cost of the refresh process line in period t,
csp_t	Setup cost of the final product (SOI) production line in period t,
$ch_t^{l^-}$	Inventory cost of Negative Wafer at level l at the end of period t,
$ch_t^{l^+}$	Inventory cost of Top (Fresh or Refresh) Wafer at level l at the end of period t,
ch_t	Inventory cost of the final product (SOI) at the end of period t.

Variables

p_t	Quantity of Fresh Wafers purchased in period t,
x_t^l	Production quantity of the final product (SOI) using Top Wafer at level l in period t,
z_t^l	Quantity of Negative Wafers at level l to be refreshed in period t,
s_t	Final product (SOI) inventory at the end of period t,
$s_t^{l^+}$	Top (Fresh or Refresh) Wafer inventory at level l at the end of period t,
$s_t^{l^-}$	Negative Wafer inventory at level l at the end of period t,
v_t	$\begin{cases} 1 \text{ if raw materials procurement occurs in period } t, \\ 0 \text{ otherwise.} \end{cases}$
w_t	$\begin{cases} 1 \text{ if the refresh process occurs in period } t, \\ 0 \text{ otherwise.} \end{cases}$
y_t	$\begin{cases} 1 \text{ if final product (SOI) production occurs in period } t, \\ 0 \text{ otherwise.} \end{cases}$

Mathematical Model of ULS-MR

$$\begin{aligned}
\min \sum_{\forall t} cf_t p_t + \sum_{\forall t} \sum_{l=0}^{l^{max}-1} cr_t z_t^l + \sum_{\forall t} \sum_{\forall l} cp_t x_t^l \\
+ \sum_{\forall t} csf_t v_t + \sum_{\forall t} csr_t w_t + \sum_{\forall t} csp_t y_t \\
+ \sum_{\forall t} \sum_{l=0}^{l^{max}-1} ch_t^{l^-} s_t^{l^-} + \sum_{\forall t} \sum_{\forall l} ch_t^{l^+} s_t^{l^+} + \sum_{\forall t} \sum_{l=1}^{l^{max}} ch_t^{l^+} z_{t-1}^{l-1} \\
+ \sum_{\forall t} ch_t s_t \qquad (11.1)
\end{aligned}$$

subject to

$$s_{t-1} + \sum_{l=0}^{l^{max}} x_t^l = d_t + s_t \qquad \forall t \qquad (11.2)$$

$$s_{t-1}^{l^+} + p_t = x_t^l + s_t^{l^+} \qquad \forall t, l = 0 \qquad (11.3)$$

$$s_{t-1}^{l^+} + z_{t-1}^{l-1} = x_t^l + s_t^{l^+} \qquad \forall t, l > 0 \qquad (11.4)$$

$$s_{t-1}^{l^-} + x_t^l = z_t^l + s_t^{l^-} \qquad \forall t, l | l \neq l^{max} \qquad (11.5)$$

$$p_t \leq M_1 \cdot v_t \qquad \forall t \qquad (11.6)$$

$$\sum_{l=0}^{l^{max}} x_t^l \leq M_2 \cdot y_t \qquad \forall t \qquad (11.7)$$

$$\sum_{l=0}^{l^{max}-1} z_t^l \leq M_3 \cdot w_t \qquad \forall t \qquad (11.8)$$

$$\sum_{\forall t} z_t^{l^{max}} = 0 \qquad (11.9)$$

$$w_T = 0 \qquad (11.10)$$

$$x_t^l, z_t^l, s_t^{l^+}, s_t^{l^-} \geq 0 \qquad \forall t, l \qquad (11.11)$$

$$p_t, s_t \geq 0 \qquad \forall t \qquad (11.12)$$

$$y_t, v_t, w_t \in \{0, 1\} \qquad \forall t \qquad (11.13)$$

The objective function (11.1) minimizes the raw material (Fresh Wafer) purchase cost, the refresh cost at all levels, the production cost using the raw materials (Top Wafers) at all levels (Fresh and Refresh Wafers), the raw material procurement cost, the refresh setup cost, the final product (SOI) production setup cost, the Negative

Wafer inventory cost at each level, the Fresh and Refresh Wafer inventory cost and the final product inventory cost. Constraint (11.2) models the final product (SOI) inventory balance. Constraint (11.3) models the purchased raw material (Fresh Wafer) flow conservation. Each time production takes place in period t, ($x_t^l > 0$ in Constraint (11.2)), a Negative Wafer is generated. The generated Negative Wafer enters the Negative Wafer flow conservation Constraint (11.5). If the Negative Wafer has not yet reached its maximum refresh level, it can possibly be refreshed. Constraint (11.4) models the Refresh Wafer flow conservation. The correctness of the flow conservation constraints can be verified by summing Constraints (11.2), (11.3) and (11.4) to obtain $\sum_{t=1}^{T}\sum_{l=0}^{l^{max}-1} z_t^l + \sum_{t=1}^{T} p_t = \sum_{t=1}^{T} d_t \quad \forall t$.

The z_{t-1}^{l-1} and x_t^l variables of the Refresh Wafer flow conservation constraint (11.4) appear for the first time with $t \geq 2$. As an example, the Negative Wafers x_1^0 are refreshed in the same period z_1^0 (Constraint (11.5)). The refreshed Wafers become available in the next period z_2^0 to be used in SOI production x_2^1 (Constraint (11.4)).

Note that the flow conservation constraints (11.4) and (11.5) do not allow to record the Refresh Wafer inventories which are consumed right after leaving the refresh process. Therefore, the term $(\sum_{\forall t}\sum_{l=1}^{l^{max}} ch_t^{l^+} z_{t-1}^{l-1})$ is added to the objective function to take into account the immediately used Refresh Wafers.

Constraints (11.6), (11.7) and (11.8) are respectively raw material procurement, final wafer production and refresh process setup constraints. M_1, M_2 and M_3 are large positive numbers, which can be substituted as follows to give a tighter formulation: $M_1 = M_2 = \sum_{t'=t}^{T} d_{t'}$ and $M_3 = \sum_{t'=t+1}^{T} d_{t'}$.

Constraints (11.9) and (11.10) are added to prevent the boundary (edge) effects. Constraint (11.9) ensures that the Negative Wafer of the last level is not refreshed. It can also be replaced with $z_t^{l^{max}} = 0 \quad \forall t$. Constraint (11.10) ensures that no refresh occurs in the last period. It can also be replaced with $\sum_{l=0}^{l^{max}-1} z_{t=T}^l = 0$.

As we only produce to satisfy the demand, stationary production costs have no impact on the production plan. Hence, the term $\sum_{\forall t}\sum_{\forall l} cp_t x_t^l$ in the objective function (11.1) can be substituted by the constant $\sum_{\forall t} cp_t d_t$.

In this chapter, a resolution method is proposed. Hence, an example of the decision variables is presented. Given $l^{max} = 2$, diagrams 11.14 and 11.15 show how the variables (in particular x_t^l, z_t^l, $s_t^{l^+}$ and $s_t^{l^-}$) evolve with each final product

fabrication and Negative Wafer refresh (remanufacturing) process. The notations below the arrows refer to the corresponding constraints.

$$\xrightarrow[(11.6)]{p_t,v_t} \text{Fresh (R0)} \xrightarrow[(11.3),(11.7)]{x_t^0,y_t} \text{SOI 0} \xrightarrow[(11.2)]{x_t^0} \text{Neg0} \xrightarrow[(11.5),(11.8)]{z_t^0,w_t} \text{R1} \xrightarrow[(11.4),(11.7)]{x_{t+1}^1,y_{t+1}} \text{SOI 1}$$
$$\xrightarrow[(11.2)]{x_{t+1}^1} \text{Neg1} \xrightarrow[(11.5),(11.8)]{z_{t+1}^1,w_{t+1}} \text{R2} \xrightarrow[(11.4),(11.7)]{x_{t+2}^2,y_{t+2}} \text{SOI 2} \xrightarrow[(11.2)]{x_{t+2}^2} \text{Neg2} \quad \downarrow \tag{11.14}$$

$$\text{Fresh (R0)} \xrightarrow{s_t^{0^+}} \text{SOI 0} \xrightarrow{s_t} \text{Neg0} \xrightarrow{s_t^{0^-}} \text{R1} \xrightarrow{s_t^{1^+}} \text{SOI 1} \xrightarrow{s_t} \text{Neg1} \xrightarrow{s_t^{1^-}} \text{R2} \xrightarrow{s_t^{2^+}} \text{SOI 2}$$
$$\xrightarrow{s_t} \text{Neg2} \quad \downarrow \tag{11.15}$$

The material flow representation of the model is illustrated in Figure 11.3. Note that it is not a classical network flow in which the flow conservation constraint is satisfied in the nodes representing the final product (SOI Wafers) fabrication.

11.3.1 Complexity Analysis of ULS-MR

In the following, it is demonstrated that the ULS-MR problem is NP-hard even for only one level of refresh ($l^{max} = 1$) and without considering the production setup of the final product (in a mono-level configuration). To prove the NP-hardness of ULS-MR, we perform a polynomial reduction from the Partition Problem.

Partition Problem: Given the set S containing N positive integers $a_1, a_2, \cdots, a_N$. Can the set S be partitioned into two separated subsets S_1 and S_2 ($S = S_1 \cup S_2$ and $S_1 \cap S_2 = \emptyset$) such that the sum of the elements of each subset are equal to A (i.e. $\sum_{i \in S_1} a_i = \sum_{i \in S_2} a_i = A$)?

Proposition 11.3.1. *The ULS-MR problem is NP-hard.*

Proof. Let us consider an instance I^{ULS-MR} of ULS-MR with $T = 2N + 3$ periods (see Figure 11.4). Let $l^{max} = 1$ and $\forall t$: $cf_t = 1$, $cr_t = 0$, $cp_t = 0$, $csf_t = csr_t = 1$, $csp_t = 0$, $ch_t^{l^-} = 0$, $ch_{2k-1}^{l^+} = 3 \quad k = 1, \cdots, N+1$, , $ch_{2k}^{l^+} = 0 \quad k = 1, \cdots, N+1$, $ch_t = 3$. The final product demand associated with each period t is d_t. $d_1 = A$, $d_{2k} = 0 \quad k = 1, \cdots, N+1$, $d_{2k+1} = a_k \quad k = 1, \cdots, N$ and $d_T = A$. We show that

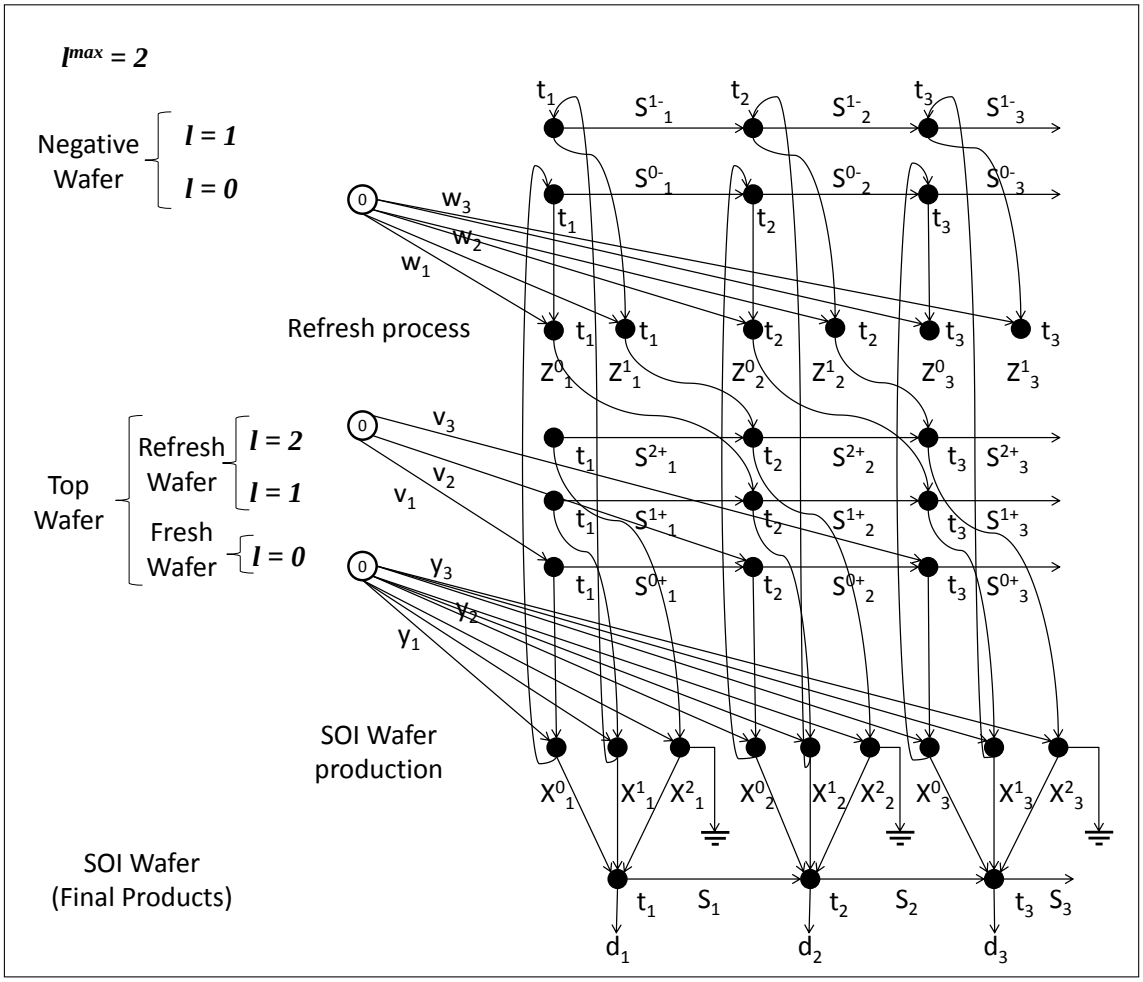


Figure 11.3: Material Flow Representation

the answer to the instance I of the Partition Problem is positive if and only if the reduced instance I^{ULS-MR} of ULS-MR has a cost of at most $2A + N + 2$.

First, assume that a solution to the instance I^{ULS-MR} has a cost of at most $2A + N + 2$. In this case, the inventory costs are sufficiently large to avoid keeping stocks of SOI and Top (Fresh and Refresh) Wafers. Let us begin with the first period ($t = 1$). As the demand must be satisfied, for the first period, A final products must be fabricated ($x_1^0 = A$). Having no Top Wafers in initial inventory, A Fresh Wafers are ordered ($p_1 = A$). The final product fabrication generates A Negative Wafers ($s_1^{0^-} = A$) which does not cost anything ($ch_1^{l^-} = 0$). The cost associated with the

first period is equal the Fresh Wafer purchase and procurement setup costs $(A+1)$. The demand of the first period is now satisfied.

As $d_{2k} = 0 \quad k = 1, \cdots, N+1$ and $d_{2k+1} = a_k \quad \forall k$, the sum of the demands of the periods $t = 2, \cdots, 2N+2$ is equal to $2A$. Now, the question is whether to satisfy the demand by (re-)procuring Fresh Wafers or refreshing the A Negative Wafers (generated in period 1) or both. The answer is that, as the refresh process cost (cr_t) is zero and the Fresh Wafer procurement setup and refresh process setup costs are equal $(csf_t = csr_t = 1)$ (and as we are minimizing the costs), the refresh process is more interesting than Fresh Wafer re-procurement. But, by refreshing the A Negative Wafers generated in the first period, only A Top Wafers are produced. So the requirements for other A Top Wafers until period $2N+2$ must still be satisfied. Note that the final period $(2N+3)$ has also a demand of A units. We show that the remaining demand of $2A$ until the end of the planning horizon must be satisfied with purchase and refresh of A Fresh Wafers. Let us first calculate the cost after the first period. Note that large inventory costs avoid to keep stocks (making campaigns). So the A generated Negative Wafers resulting from the first period are refreshed in even periods to satisfy the demand of uneven periods. The demand of each period in the interval $2, \cdots, 2N+2$ is satisfied by either refreshing A Negative Wafers or purchasing A Fresh Wafers. Finally, the demand of the last period is satisfied by refreshing the A Negative Wafers generated during the periods before $2N+2$. The refresh occurs in period $2N+2$ and satisfies the Top Wafer need of the final period.

As neither SOI Wafers nor Top Wafers are kept in stock and that the demand of every period is positive, either purchase or refresh occur in each period. As we minimize the cost and tend to satisfy only the demand (and not more), even one period with both purchase and refresh makes the total cost going beyond $2A+N+2$. Hence, purchase and refresh periods form the sets S_1 and S_2.

Now, it is explained why any other production plan will be either infeasible or more costly. Ordering less Fresh Wafers would lead to unsatisfied demand. However, one may suggest to order more Fresh Wafer and refresh less. We show that the cost of such a plan exceeds $2A + N + 2$. Ordering more Fresh Wafers in the first period is penalized with high Top Wafer holding cost of 3 units per each extra item. This penalty will never be equal to the setup cost of the refresh process, which is only

1 unit. Now imagine the case where the demand of the final period $(2N + 3)$ is not purely satisfied from refreshing. In this case, suppose that we purchase ϵ Fresh Wafers in the last period ($p_{2N+3} = \epsilon$) to complete the Top Wafer need. So, we do not need to purchase exactly A Fresh Wafers in periods $2, \cdots, 2N + 2$; and we order $A - \epsilon'$ Fresh Wafers. We must satisfy the remaining demand of A of periods $2, \cdots, 2N + 1$ using the ordered $A - \epsilon'$ Fresh Wafers. This is only possible if we refresh ϵ' of the generated Negative Wafers. This leaves us with $A - 2\epsilon'$ Negative Wafers which we refresh in period $2N + 2$. As we must provide A Top Wafers for the satisfaction of the final period demand, we must order the missing $2\epsilon'$. So ϵ is equal to $2\epsilon'$.

The cost associated with this production plan is calculated as follows: $A + 1$ for purchasing A Fresh Wafers in the first period; $N + (A - \epsilon')$ for periods $2, \cdots, 2N + 1$; 1 cost unit for the setup cost of refreshing $A - 2\epsilon'$ in period $2N + 2$ and $2\epsilon' + 1$ for purchase unit cost and ordering setup of $2\epsilon'$ Fresh Wafers in final period $2N + 3$. The total cost is $2A + N + 2 + \epsilon' + 1$, which is larger than $2A + N + 2$. Hence, it is proved that if I^{ULS-MR} is positive, then I is positive.

Conversely, it is shown that if I is positive, then I^{ULS-MR} is positive. In other words, if an instance I is positive, then it exists S_1 and S_2 such that $\sum_{i \in S_1} a_i = \sum_{i \in S_2} a_i = A$. Having I, we can build a valid solution at the cost of $2A + N + 2$. The solution is to procure A Fresh Wafers to satisfy the demand of the first period, to refresh A generated Negative Wafers and to purchase A Fresh Wafers to satisfy the Top Wafer need of periods $2, \cdots, 2N + 1$, and to refresh A generated Negative Wafers in period $2N+2$ to satisfy the final period demand. The refresh and purchase make sets S_1 and S_2, separately. □

Considering the fact that the optimal solution to ULS-MR and to the PARTITION PROBLEM is independent of the ordering of $a_i \quad \forall i$, the ULS-MR problem is still NP-hard even with decreasing or increasing demand trends over the planning horizon.

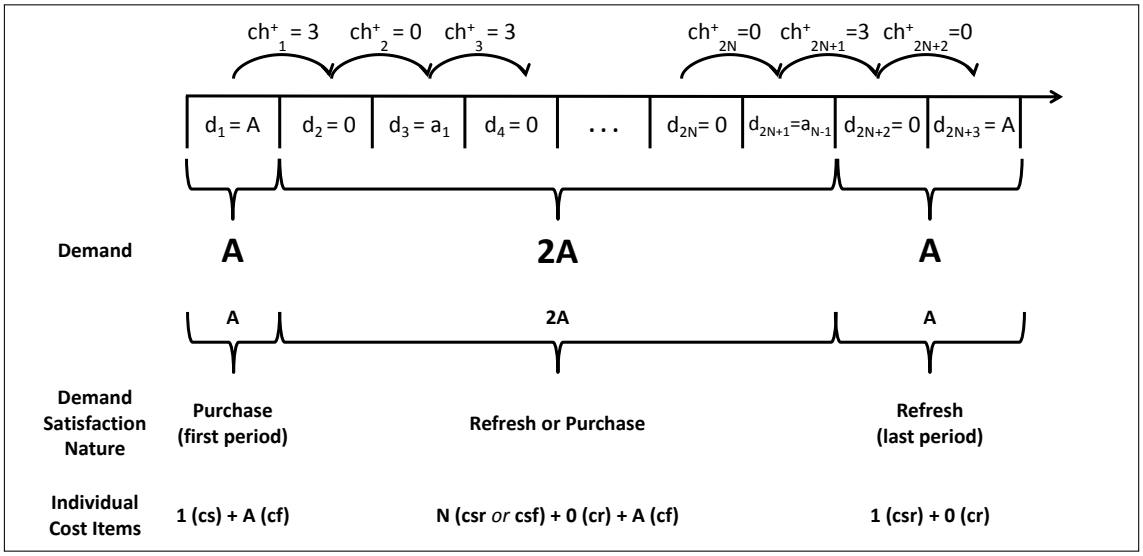


Figure 11.4: An Instance of I^{ULS-MR}

11.4 ULS-MR with only Procurement Setup (ULS-MR$_1$)

In a first step, we simplify the ULS-MR problem by relaxing two setup constraints, i.e. the SOI production and refresh process setup constraints (Constraints (11.7) and (11.8), respectively). The only setup constraint considered is the Fresh Wafer procurement setup cost (Constraint (11.6)). We refer to this new model as ULS-MR$_1$.

Inventory cost determination The inventory cost is composed of physical holding cost and capital cost. The physical holding cost is negligible and considered to be the same for final product, Fresh, Refresh and Negative Wafers. However, the capital cost decreases for Negative Wafers as the refresh level approaches l^{max}. The Negative Wafer inventory value is illustrated in Figure 11.5. Among raw materials, Fresh Wafer has the highest value. As it is used and refreshed further, its value decreases by losing its potential to be refreshed and reused again in production. In reality, a Negative Wafer which is no longer refreshable is used further as test wafer with a scrap value. However, for simplicity the Negative Wafer scrap value is considered to be zero.

Each amortization value can be determined as $\frac{\text{Fresh Wafer value} - \text{Negative Wafer scrap value}}{l^{max}+1}$. The same reasoning holds for Top (Fresh and Refresh) Wafers. Therefore, in the

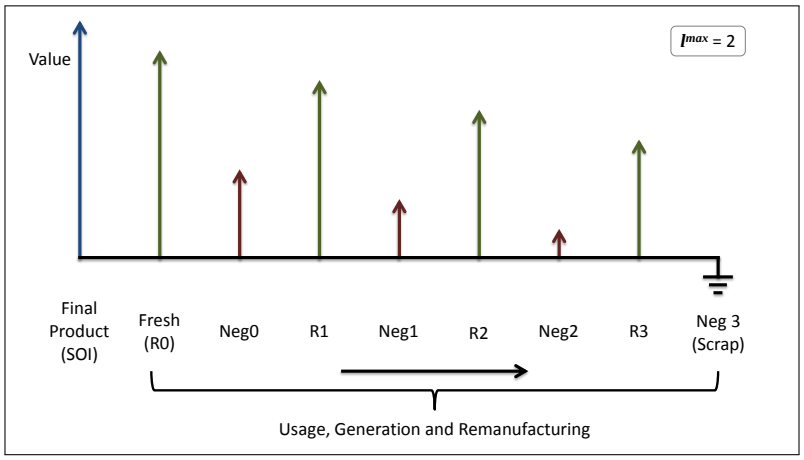


Figure 11.5: Fresh and Negative Wafer Degressive Value

model, the Negative and Top Wafer inventory costs are assumed to be degressive.

Assumptions Before tackling the structure of the optimal solutions, we will make the following assumptions. They are well-grounded and correspond to industrial facts.

- In the reduced model, we have three *stock keeping units (SKUs)*: *SOI Wafer (final product)*, *Top Wafer (raw material)* and *Negative Wafer (by-product)*. A Base Wafer and a Top Wafer are used to make an SOI Wafer. In view of the used raw material and numerous process steps to make an SOI Wafer, it is the most expensive SKU in comparison to Top and Negative Wafers. Between Top and Negative Wafers, the Top Wafer is more valuable as it can directly be used in the final product fabrication. The least value goes to the Negative Wafer, as it needs remanufacturing (refreshing) to return to the SOI production process. By virtue of the values of the SKUs, the inventory cost of the SOI Wafers is larger than the Top Wafer inventory cost. And the Top Wafer inventory cost is larger than the Negative Wafers inventory cost (see Figure 11.5).

- Each Top Wafer has a potential value stored in it. At each refresh process, this value decreases. Thus, a Fresh Wafer (R0) is worthier than a Refresh Wafer 1 (R1) as the Fresh Wafer keeps in itself R1 (and also all other levels of Refresh Wafer, if any). As a result, the Top Wafer inventory cost decreases with the refresh level (see Figure 11.5).
- The same reasoning holds for the Negative Wafers. The Negative Wafers of the higher levels are more valuable than those of the lower levels. Therefore, the fixed capital (and hence inventory) cost associated with higher level Negative Wafers is larger (see Figure 11.5).
- As the value of Negative wafers reduces after each refresh process (by-product value depreciation), in terms of the inventory cost, higher Top Wafer levels are used for SOI production when a choice can be made. By choosing Top Wafers of higher levels, we also guarantee the Negative Wafer generation. For instance, first we would satisfy the Top Wafer requirements (to produce SOI Wafers) from R1 rather than R2, rather than R3 and so on.
- For simplicity reasons, we consider that the Negative Wafer generation and refresh process can occur in the same period.
- An implicit assumption which is considered is that the refresh process cost is lower than the Fresh Wafer purchase cost. Otherwise the refresh process loses its economical value and the model reduces to a classical lot-sizing problem.
- The costs are assumed to be stationary over time.

A summary of the inventory cost comparison can be found in Figure 11.6.

11.4.1 Structure of the Optimal Solutions for ULS-MR$_1$

Here, we discuss the optimal solution properties of ULS-MR$_1$. To simplify the presentation of this section, we suppose that no initial and final inventories are

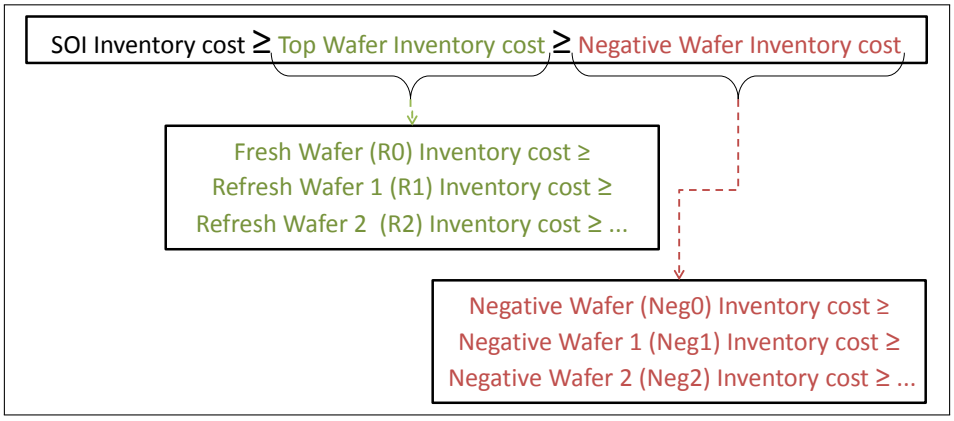


Figure 11.6: Inventory Cost Assumptions

determined[3].

Definition 11.4.1. The *Fresh Wafer Life Cycle* is defined as the interval between the Fresh Wafer order until its full utilization (i.e. the last refresh level l^{max}), obviously if $T \geq l^{max} + 2$. One period is needed for the Negative Wafer generation and refreshing. The refreshed wafer is available at the beginning of the next period.

Proposition 11.4.1. *In an optimal solution of ULS-MR$_1$, with $l^{max} \geq 1$ and $T > 1$,* $\frac{\sum_{t=1}^{T} d_t}{\min\{T, l^{max}+1\}} \leq \sum_{t=1}^{T} p_t \leq \sum_{t=1}^{T} d_t$.

Proof. First consider the right hand side of the inequality ($\sum_{t=1}^{T} p_t \leq \sum_{t=1}^{T} d_t$) which states that the total raw material purchase cannot exceed the total demand over the whole planning horizon. By contradiction, assume there is an optimal policy in which the total purchase exceeds the total demand of all periods, i.e. $\sum_{t=1}^{T} p_t > \sum_{t=1}^{T} d_t$. Let us demonstrate the existence of another policy with lower total cost (or at least the same cost). As no (final product) production and refresh process setup costs are considered and since the refresh cost is lower than the Fresh Wafer purchase cost, if we instead order a portion of the total demand and use the by-products (Negative Wafers) to satisfy the Top Wafer need, we save the difference between the

[3]If the initial and ending inventory values and stock lower bounds are fixed, the demand requirements can be reformulated to set the initial and ending inventory values as well as lower bounds to zero [74].

refresh process and the Fresh Wafer purchase costs multiplied by the amount saved in purchase and sent to refresh.

Now consider the left hand side of the inequality ($\frac{\sum_{t=1}^{T} d_t}{\min\{T, l^{max}+1\}} \leq \sum_{t=1}^{T} p_t$). It is technically impossible to refresh more than $\min\{T, l^{max}+1\}$ times. This means that, in the best case, we may use the Fresh Wafer and all its Refresh Wafers (if the planning horizon length allows) to satisfy the demand and no more. □

Note 1. Here, the aim is not to discuss the optimal solution properties of the original model. However, Proposition 11.4.1 also holds for the original problem. In the same way, the bounds for the refresh process are defined as $\frac{\sum_{t=1}^{T} d_t}{\min\{T, l^{max}+1\}} \leq \sum_{t=1}^{T} \sum_{l=0}^{l^{max}-1} z_t^l \leq \sum_{t=1}^{T} \frac{l^{max}}{l^{max}+1} d_t$.

The following proposition states that if Fresh Wafers are ordered in period t, then final product fabrication (using the purchase) takes place in the same period t.

Proposition 11.4.2. *In an optimal solution of ULS-MR$_1$, if $p_t > 0$, then $x_t^0 > 0$.*

Proof. The proof is done by contradiction. Let us consider an optimal solution in which the production takes place in period t' after a Fresh Wafer purchase campaign in period t. In this case, the Fresh Wafer quantity (relating to the purchase campaign in period t) can be moved to period t' (where the production takes place), saving the Fresh Wafer inventory cost between t and t' since the Fresh Wafer purchase and holding costs are stationary. This leads to a total cost at least equal to the total cost of the given optimal solution contradicting the initial assumption. □

Proposition 11.4.3. *In an optimal solution of ULS-MR$_1$, if $x_t^{l+1} > 0$, then $z_{t-1}^l = x_t^{l+1} \quad \forall l | l \neq l^{max}$.*

The proposition states that in case of SOI production from Refresh Wafers in a given period t, the required Refresh Wafer will be satisfied through the refresh process of Negative Wafers of the previous level in the previous period $t-1$ available in t.

Proof. As no refresh process setup cost is considered, with stationary unitary refresh cost and as the Negative Wafer inventory cost is lower than any Refresh Wafer

inventory cost, it is profitable not to refresh a Negative Wafer just to keep its Refresh Wafer in stock. Hence, if we refresh Negative Wafers, it is only to satisfy the SOI production need. □

Note 2. In general, if $x_t^l > 0 \quad \forall l$, then $(z_{t-1}^{l-1} > 0)$ and/or $(p_t > 0$ or $s_{t-1}^{0^+} > 0)$.

The conclusion can be detailed as follows:

- if $x_t^l > 0, \quad l = 0$, then either $p_t > 0$ or $s_{t-1}^{0^+} > 0$;
- if $x_t^l > 0, \quad l > 0$, then $z_{t-1}^{l-1} > 0$ and $z_{t-1}^{l-1} = x_t^l$.

Proposition 11.4.4. *There exists an optimal solution to ULS-MR$_1$ in which* $s_t^{l^+} = 0 \quad \forall t, l > 0$.

Proof. It follows from Proposition 11.4.3 and Constraint (11.4). □

The Refresh Wafer inventory must be defined as $s_t^{l^+} + z_t^{l-1}$ where $s_t^{l^+}$ is only to maintain the inventory balance. $s_t^{l^+}$ alone does not allow to record the immediately consumed Refresh Wafers. Considering Proposition 11.4.3, for the single setup model, only z_t^{l-1} allows Refresh Wafer inventory cost charging.

Proposition 11.4.5. *There exists an optimal solution to ULS-MR$_1$ in which* $s_{t-1}^{0^+} p_t = 0 \quad \forall t$.

Proof. This is the newly purchased raw material Zero Inventory Ordering property. The proof is done by contradiction. Given an optimal solution for which this property does not hold. It is possible to construct a feasible solution with a lower or equal total cost by moving $s_\bullet^{0^+}$ from the last purchase period (before t) to the current purchase period t. This contradicts the assumptions. Note that this property holds even with stationary purchase costs. □

Proposition 11.4.5 can also be proved intuitively. Remember that Top Wafers of higher levels are first used to make SOI Wafers, i.e. firstly Fresh Wafer, if not available R1, if not available R2 and so on. If Fresh Wafer stock is on hand in period t, either no production has occurred in the precedent periods (before t) or production has occurred in the precedent periods (before t). First case where no production

has occurred in the precedent periods: It means that no refreshable Negative Wafers are on-hand, so that they can be refreshed and used in the SOI production. In this case, the Fresh Wafer purchase p_t can be grouped with the precedent Fresh Wafer purchase. Second, if production has occurred in the precedent periods: In this case, refreshable Negative Wafers must be on-hand, as Fresh Wafers are still available on stock. Therefore, in order to satisfy the demand in terms of SOI, on-hand Negative Wafers must be refreshed before any order is placed, since the refresh cost is lower than Fresh Wafer purchase cost. Some conclusions can be drawn from Proposition 11.4.5.

Note 3. Constraint (11.3) of the Model is expanded using Proposition 11.4.3.

$$s_{t-1}^{l^+} + p_t = x_t^l + s_t^{l^+} \qquad \forall t, l = 0 \tag{11.16a}$$

$$s_{t-1}^{0^+} - s_t^{0^+} = x_t^0 - p_t \qquad \forall t \tag{11.16b}$$

Several cases may happen:

$$\text{If } p_t > 0 \text{ and } p_t = x_t^0 \xrightarrow[11.4.5]{Prop.} s_{t-1}^{0^+} = s_t^{0^+} = 0 \tag{11.17}$$

$$\text{If } p_t > 0 \text{ and } p_t > x_t^0 \rightarrow s_{t-1}^{0^+} < s_t^{0^+} \xrightarrow[11.4.5]{Prop.} s_t^{0^+} > 0 \tag{11.18}$$

$$\text{If } p_t = 0 \text{ and } x_t^0 > 0 \rightarrow s_{t-1}^{0^+} > 0 \rightarrow s_{t-1}^{0^+} > s_t^{0^+} \tag{11.19}$$

$$\text{If } p_t = 0 \text{ and } x_t^0 = 0 \rightarrow s_{t-1}^{0^+} - s_t^{0^+} = 0 \rightarrow$$

$$\begin{cases} s_{t-1}^{0^+} = s_t^{0^+} = 0, \text{ then } d_t \geq 0 \rightarrow (11.21) & (11.20a) \\ s_{t-1}^{0^+}, s_t^{0^+} > 0, \text{ then } d_t = 0. \text{ However, } d_{t-1} \text{ might be larger than 0.} & (11.20b) \end{cases}$$

$$\text{if } d_t > 0, \text{ then } \sum_{l=1}^{l^{max}} x_t^l > 0 \xrightarrow[11.4.4]{Prop.} \sum_{l=0}^{l^{max}-1} z_{t-1}^l > 0 \tag{11.21}$$

(11.20b) comes from the fact that Refresh Wafers will not be used for the production of SOI Wafers until Fresh Wafers are on hand. Therefore, if no SOI production using Fresh Wafers are made ($x_t^0 = 0$) while Fresh Wafer inventory is positive $s_t^{0^+} > 0$, there must have been no demand for SOI wafers.

Proposition 11.4.6. *There exists an optimal solution to ULS-MR$_1$ in which* $\sum_{l=0}^{l^{max}-1} s_{t-1}^{l^-} p_t = 0 \quad \forall t.$

Proof. As no refresh process setup cost is considered, it is possible to refresh in period $t-1$ and order Fresh Wafers in t to complete the raw material need in order to satisfy d_t. But it is not possible to keep refreshable Negative Wafers as soon as the refresh process cost is less expensive than Fresh Wafer procurement. Therefore, it is economically interesting to refresh all refreshable Negative Wafers on hand to satisfy the SOI production need instead of purchasing any Fresh Wafers and keeping refreshable Negative Wafers in stock. □

Note than Fresh Wafer procurement (say in period t, i.e. p_t) may occur within a Fresh Wafer Life Cycle, and when $z_{t-1}^l > 0$. Nevertheless, if refreshable Negative Wafer inventory exists, it would be preferable to refresh these Negative Wafers instead of acquiring Fresh Wafers which cost considerably more than refreshing the Negative Wafers. To conclude, it is possible to order Fresh Wafers and to refresh Negative Wafers at the same time. But it is not possible to order Fresh Wafers while keeping refreshable Negative Wafers in inventory.

Note 4. As no SOI production setup and refresh process setup costs are considered and also because the refresh process lead time is assumed to be equal to 1 period, the Fresh Wafer purchase is to complete the Top Wafer need of the final product demand (see (11.22)).

$$\sum_{l=0}^{l^{max}-1} z_{t-1}^l + p_t - d_t \geq 0 \quad \forall t \tag{11.22}$$

11.4.2 Mathematical Model of ULS-MR$_1$

The optimal solution properties allow us to simplify ULS-MR by removing the refresh variables z_t^l as follows. Based on Propositions 11.4.3 and 11.4.4, we get:

$$z_t^l = x_{t+1}^{l+1} = s_{t-1}^{l^-} - s_t^{l^-} + x_t^l \qquad \forall t, l | l \neq l^{max} \tag{11.23}$$

The mathematical model for ULS-MR$_1$ is written:

$$\begin{aligned}
&\min \sum_{\forall t}(cf + cp)p_t + \sum_{\forall t}\sum_{l=1}^{l^{max}} cpx_t^l \\
&+\sum_{\forall t}\sum_{l=1}^{l^{max}} cr(s_{t-1}^{l^-} + x_t^l - s_t^{l^-}) + \sum_{\forall t} csfv_t \\
&+\sum_{\forall t}\sum_{l=0}^{l^{max}-1} ch^{l^-} s_t^{l^-} + \sum_{\forall t} ch^{0^+} s_t^{0^+} \\
&+\sum_{\forall t}\sum_{l=1}^{l^{max}} ch^{l^+}(s_{t-1}^{l^-} + x_t^l - s_t^{l^-}) + \sum_{\forall t} chs_t \qquad (11.24)
\end{aligned}$$

subject to

$$s_{t-1} + \sum_{l=0}^{l^{max}} x_t^l = d_t + s_t \qquad \forall t \qquad (11.25)$$

$$s_{t-1}^{0^+} + p_t = x_t^0 + s_t^{0^+} \qquad \forall t \qquad (11.26)$$

$$s_{t-1}^{l^-} + x_t^l = x_{t+1}^{l+1} + s_t^{l^-} \qquad \forall t, l | l \neq l^{max} \qquad (11.27)$$

$$p_t \leq M_1 \cdot v_t \qquad \forall t \qquad (11.28)$$

$$\sum_{\forall t} x_t^{l^{max}+1} = 0 \qquad (11.29)$$

$$x_t^l, s_t^{l^-} \geq 0 \qquad \forall t, l \qquad (11.30)$$

$$p_t, s_t, s_t^{0^+} \geq 0 \qquad \forall t \qquad (11.31)$$

$$v_t \in \{0, 1\} \qquad \forall t \qquad (11.32)$$

Note that in the objective function (11.24), we have immediately counted the SOI production cost of all purchased Fresh Wafers ($(cf + cp)p_t$). The reason is that no Fresh Wafer is bought only for stocking purpose. The first two terms could also be written as $\sum_{\forall t} cfp_t + \sum_{\forall t}\sum_{l=0}^{l^{max}} cpx_t^l$.

The material flow representation of ULS-MR$_1$ is illustrated in Figure 11.7. By only considering the raw material procurement setup and the cost structure assumptions, the problem is simplified. However, still important decisions must be made in each period regarding the SKUs (Fresh Wafers, Negative Wafers and Refresh Wafers). In fact, the Refresh Wafers do not really create complications since according to Proposition 11.4.4, the Refresh Wafer inventory is null over the whole

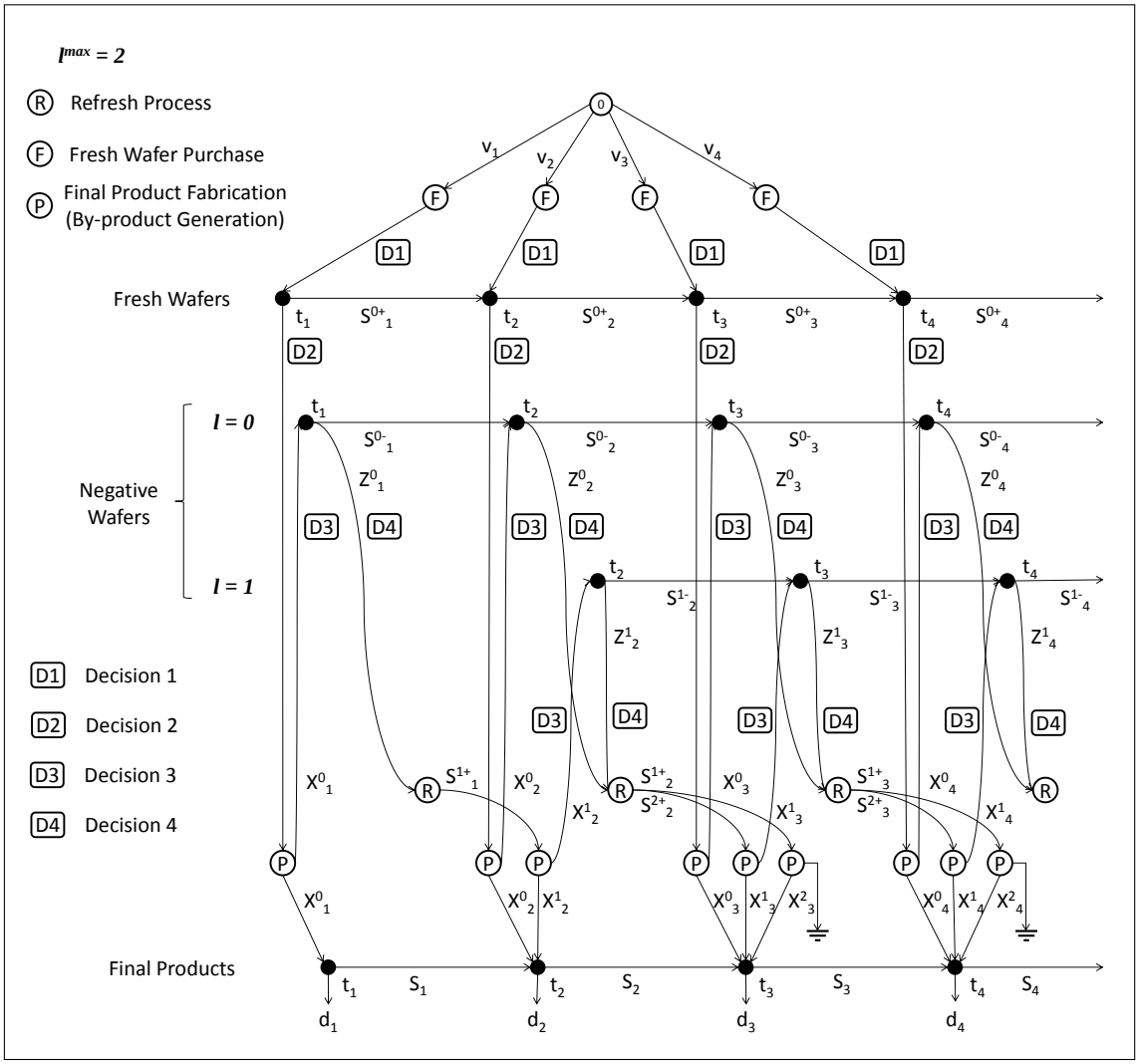


Figure 11.7: Material Flow Representation of ULS-MR$_1$ for $l^{max} = 2$ and $T = 4$

planning horizon. In other words, the refresh process follows a just-in-time policy. However, the two other SKUs make the planning more difficult. Each additional decision is shown in Figure 11.7.

Regarding Fresh Wafers, at each procurement or production period, we must take two more decisions. Decision 1: How many more Fresh Wafers are to be bought to keep in stock for future use? Decision 2: How many of the stocked Fresh Wafer inventory are to be used for production in a production period? Constraint (11.26)

is at the origin of these decisions.

Concerning Negative Wafers, at each production or refresh period, we must take two decisions. Decision 3: How many final products must be fabricated to assure that Negative Wafer generation is enough for future raw material requirements? Decision 4: How many of the generated Negative Wafers in the inventory are to be sent to refresh? Constraint (11.27) is at the origin of these decisions. Note that Decision 3 is one of the most complicating characteristics of this problem, i.e. final product fabrication is not only to satisfy the final product demand but also to create the by-products needed to satisfy the final product demand later.

In order to simplify these decisions and the model, we introduce "push" policies in the next section.

11.5 ULS-MR$_1$ under "Full Push Policy" (ULS-MR$_1^{FP}$)

To make the problem more tractable, we consider three additional assumptions which eliminate decisions regarding three SKUs: Fresh, Negative and Refresh Wafers. We "push" the manufacturing system to produce and remanufacture in order to avoid keeping any inventory of these three SKUs.

"Full Push Policy" Assumptions

1. All Fresh Wafers (purchased raw material) are systematically transformed to SOI Wafers (final products).
2. The generated Negative Wafers are directly refreshed after generation.
3. The Refresh Wafers are also systematically transformed to final products.

It follows that neither Top Wafer inventory nor Negative Wafer inventory is kept (see Figure 11.8). Hence, imposing these assumptions eliminates decisions on whether to keep or to use inventories of these SKUs. By ordering p_t units of Fresh Wafer, p_t units of final products systematically appear the following periods until period $t + l^{max} + 1$.

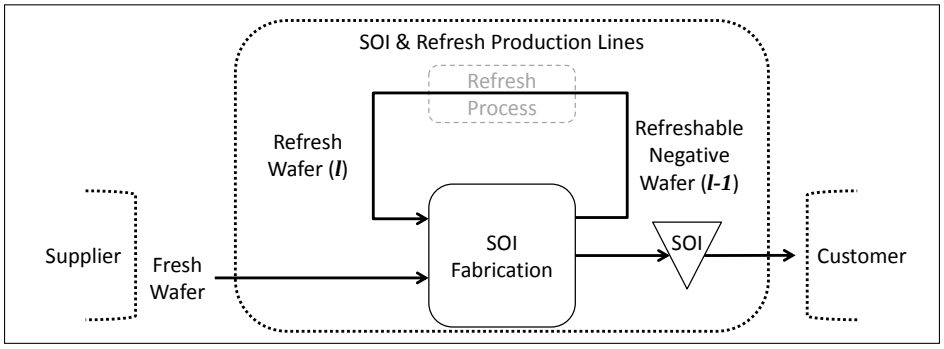


Figure 11.8: Simplified SOI-Refresh Production Schema after "Full Push Policy" Assumptions

11.5.1 Structure of the Optimal Solutions for ULS-MR$_1^{FP}$

We now study the structure of an optimal solution of the Single Setup model after imposing the "Full Push Policy" Assumptions.

Proposition 11.4.2 changes to Proposition 11.5.1. It states that if Fresh Wafers (raw materials) are ordered in period t, then SOI production (using *all* of the purchased Fresh Wafers) takes place in the same period t.

Proposition 11.5.1. *In an optimal solution of ULS-MR$_1^{FP}$, if $p_t > 0$, then $x_t^0 = p_t$.*

Proposition 11.4.3 changes to Proposition 11.5.2.

Proposition 11.5.2. *In an optimal solution of ULS-MR$_1^{FP}$, $z_{t-1}^l = x_t^{l+1} \quad \forall l | l \neq l^{max}$.*

Proposition 11.4.4 changes to Proposition 11.5.3, meaning that no Top Wafer inventory (neither Fresh Wafer nor Refresh Wafer) is held.

Proposition 11.5.3. *There exists an optimal solution to ULS-MR$_1^{FP}$ in which $s_t^{l+} = 0 \quad \forall t, l \geq 0$.*

According to Proposition 11.5.3, Proposition 11.4.5 always holds.

As Negative Wafers are systematically refreshed, no Negative Wafer inventory is kept.

Proposition 11.5.4. *There exists an optimal solution to ULS-MR$_1^{FP}$ in which* $s_t^{l^-} = 0 \quad \forall t, l | l \neq l^{max}$.

Based on Proposition 11.5.4, Proposition 11.4.6 always holds as no Negative Wafer inventory exists. Without loss of generality, the initial inventory vector is considered to be null. If stocks $s_{t-1}^{l^+} \quad \forall l$ and s_{t-1} are on-hand, we can pre-treat the final products demands after t. Under the Full Push assumptions, we know with certainty when the final products are fabricated. This makes the data pre-treatment of the final demand possible. After data pre-treatment, the stock vector is null and we have new demands ($d'_\bullet$).

Proposition 11.5.5. *If the demand is non-decreasing over the planning horizon, there exists an optimal solution to ULS-MR$_1^{FP}$ such that* $p_t s_{t-1} = 0 \quad \forall t$.

Proof. The proof is done by contradiction. Assume two periods t and t', where $t < t'$ and t is a procurement period. With non-decreasing demand, let us consider an optimal solution for which this property does not hold. This means that we have satisfied demand and the final product inventory is still on-hand at the end of period t', i.e. $(\frac{\sum_{k=t}^{t'} d'_k + s_{t'}}{l^{max}+1})$. It is possible to construct a feasible solution with a lower than or equal total cost by ordering less Fresh Wafers of amount $\frac{s_{t'}}{l^{max}+1}$ in the last purchase period (t) while still satisfying demand. This contradicts the assumptions. □

This property is a kind of Zero Inventory Ordering Property. It allows regeneration intervals to be defined. In general, we end up with a minimum final product inventory level.

Note that this property does not necessarily hold for decreasing demand trend.

Proposition 11.5.6. *If the demand is non-decreasing over the planning horizon, there exists an optimal solution to ULS-MR$_1^{FP}$ such that* $s_T = 0$.

Proof. The proof is done by contradiction. With non-decreasing demand, let us consider an optimal solution for which this property does not hold. This means that the demand is satisfied and a final product inventory is still on-hand at the end of the planning horizon, i.e. $s_T > 0$. It is possible to construct a feasible solution with a lower or equal total cost by ordering less Fresh Wafers of amount $\frac{s_T}{l^{max}+1}$ in

the previous purchase period while still satisfying demand. This contradicts the assumptions. □

In other words, Proposition 11.5.6 states raw materials are exhausted by making full Fresh Wafer life cycles, leading to zero final product inventory at the end of the planning horizon. Note that Propositions 11.5.5 and 11.5.6 also hold for ULS-MR$_1$.

11.5.2 Mathematical Model for ULS-MR$_1^{FP}$

When considering the Full Push Policy Assumptions, flow conservation constraints can be simplified. Flow conservation Constraints (11.2) to (11.5) are repeated below (Constraints (11.33) to (11.37)).

$$s_{t-1} + x_t^0 = d_t + s_t \qquad \forall t \tag{11.33}$$

$$s_{t-1} + \sum_{l=1}^{l^{max}} x_t^l = d_t + s_t \qquad \forall t \tag{11.34}$$

$$s_{t-1}^{l^+} + p_t = x_t^l + s_t^{l^+} \qquad \forall t, l = 0 \tag{11.35}$$

$$s_{t-1}^{l^+} + z_{t-1}^{l-1} = x_t^l + s_t^{l^+} \qquad \forall t, l > 0 \tag{11.36}$$

$$s_{t-1}^{l^-} + x_t^l = z_t^l + s_t^{l^-} \qquad \forall t, l | l \neq l^{max} \tag{11.37}$$

Top Wafer and Negative Wafer inventories are eliminated. Since no refresh process is possible in period 0, by changing Constraint (11.36) to Constraint (11.41), t must be at least 2.

$$s_{t-1} + x_t^0 = d_t + s_t \qquad \forall t \tag{11.38}$$

$$s_{t-1} + \sum_{l=1}^{l^{max}} x_t^l = d_t + s_t \qquad \forall t \tag{11.39}$$

$$p_t = x_t^0 \qquad \forall t \tag{11.40}$$

$$z_{t-1}^{l-1} = x_t^l \qquad \forall t > 1, l > 0 \tag{11.41}$$

$$x_t^l = z_t^l \qquad \forall t, l | l \neq l^{max} \tag{11.42}$$

Constraint (11.44) is obtained by replacing x_t^0 with p_t (using Constraint (11.40)) in Constraint (11.38). Let us consider a simple numerical example to see to which extent we may still reduce the constraints. The purchased Fresh Wafers, in say period 1, are used to make final products (based on Constraint (11.40): $p_1 = x_1^0$). The final products that are fabricated are used to satisfy the demand in period 1 (d_1) in Constraint (11.38). The fabrication of final products generates Negative Wafers (Neg0) which are sent to refresh directly (based on Constraint (11.42): $x_1^0 = z_1^0$). The refreshed Negative Wafers (Refresh Wafers at level 1, or simply R1) are used at the beginning of the next period to fabricate final products (based on Constraint (11.41): $z_1^0 = x_2^1$). Final Wafers produced from Refresh Wafer at level 1 (x_2^1) is used in Constraint (11.39) to satisfy the demand of the second period (d_2). This cycle continues. Constraint (11.43) shows the Top Wafer need for final product fabrication.

$$\sum_{l=1}^{l^{max}} z_t^{l-1} + p_t = \sum_{l=0}^{l^{max}} x_t^l \qquad \forall t \tag{11.43}$$

The flow conservation Constraints (11.38) to (11.42) reduce to Constraints (11.44) and (11.45). As the by-product generation and refresh process happen step by step in successive periods after a procurement, the upper limit of the sum is defined as $(\min\{t, l^{max}+1\}) - 1$. This phenomenon can be observed in the beginning of the planning horizon in Figure 11.9. Constraint (11.42) is needed to charge the Refresh Wafer inventory cost at the end of each period. In order to get rid of x_t^l, let us re-write the constraint as (11.46).

$$s_{t-1} + p_t = d_t + s_t \qquad \forall t \tag{11.44}$$

$$s_{t-1} + \sum_{l=1}^{(\min\{t,l^{max}+1\})-1} p_{t-l} = d_t + s_t \qquad \forall t \tag{11.45}$$

$$p_{t-l} = z_t^l \qquad \forall t, l | l \neq l^{max} \tag{11.46}$$

As SOI production and refresh process can be expressed through Fresh Wafer purchase, the objective function can also be simplified.

$$\sum_{\forall t} cfp_t + (l^{max} - 1) \sum_{\forall t} crp_t + l^{max} \sum_{\forall t} cpp_t + \sum_{\forall t} csfv_t + \sum_{\forall t} \sum_{l=1}^{l^{max}} ch^{l^+} z_t^{l-1} + \sum_{\forall t} chs_t \tag{11.47}$$

Constraints (11.44) and (11.45) yield the single flow conservation Constraint (11.49). Without loss of generality, we consider all initial inventories equal to zero. The model is presented below:

Mathematical Model for ULS-MR$_1^{FP}$

$$\begin{aligned}
\min \quad & (cf - cr + l^{max}(cr + cp)) \sum_{\forall t} p_t + \sum_{\forall t} csfv_t + \\
& \sum_{\forall t} \sum_{l=1}^{(\min\{t, l^{max}+1\})-1} ch^{l^+} p_{t-l} + \sum_{\forall t} chs_t && && (11.48) \\
& \text{subject to} \\
& s_{t-1} + \sum_{l=0}^{(\min\{t, l^{max}+1\})-1} p_{t-l} = d_t + s_t && \forall t && (11.49) \\
& p_t \leq M_1 \cdot v_t && \forall t && (11.50) \\
& p_t, s_t \geq 0 && \forall t && (11.51) \\
& v_t \in \{0, 1\} && \forall t && (11.52)
\end{aligned}$$

The material flow representation of the model is illustrated in Figure 11.9. It is not a classical network flow in which flow conservation constraint is satisfied in the nodes representing the final product fabrication.

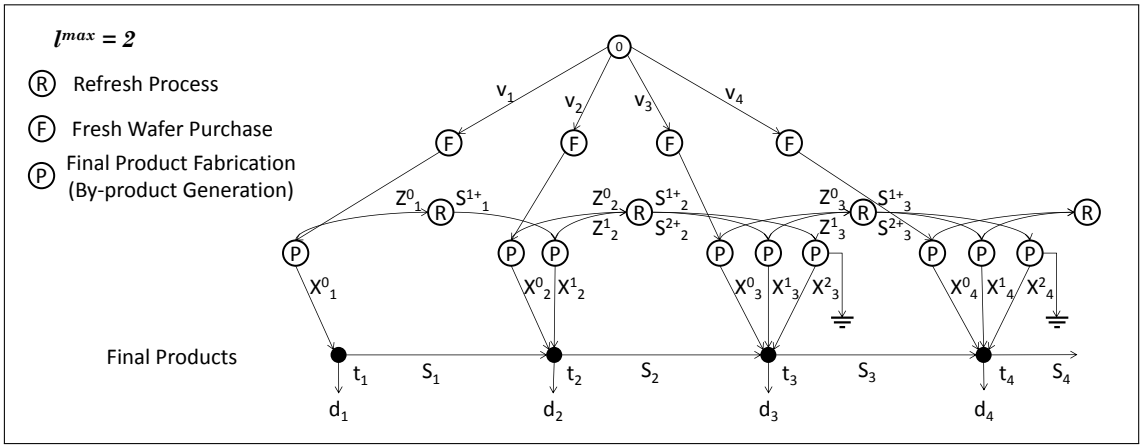


Figure 11.9: Material Flow Representation of ULS-MR$_1$$^{\text{FP}}$ for $l^{max} = 2$ and $T = 4$

11.5.3 Exact Resolution Approach for ULS-MR$_1^{FP}$: A Dynamic Programming Algorithm

A dynamic programming (DP) algorithm is presented to solve the model. Let $C(t)$ be the total minimum cost of solving the problem over the first t periods. It corresponds to the best Fresh Wafer procurement strategy that satisfies the final product (SOI Wafer) demand requirements in periods $1, \cdots, t$.

The Shortest Path Representation

The DP states can be presented by a directed graph with $\sum_{t=1}^{T} 2^{(t-1)}$ nodes. Each node represents a status. Each arc (i, j) where $i \leq j$ represents the regeneration interval $[i-1, j]$. In such a regeneration interval, with the raw material (Fresh Wafer) procurement setup in period i, the final product (SOI Wafer) demand of the interval $[i-1, j]$ is covered. A cost $C(i-1, j)$ is associated with each arc (i, j). It corresponds to the purchase unitary and setup costs and the resulting inventory costs of remanufactured by-products and final products.

The number of nodes increases exponentially. Even three indices are not enough for unique labeling of the nodes. To show this, let use define the indices of each node as (i, j, k) with $i < j$ and $j \leq k$, where i is the previous production period, j is the current period and j is the final period covered by this replenishment campaign.

Figure 11.10 shows an instance of four periods. Even in this simple example, we see that at the period 4, problems with duplicates in labeling arise. Therefore, for

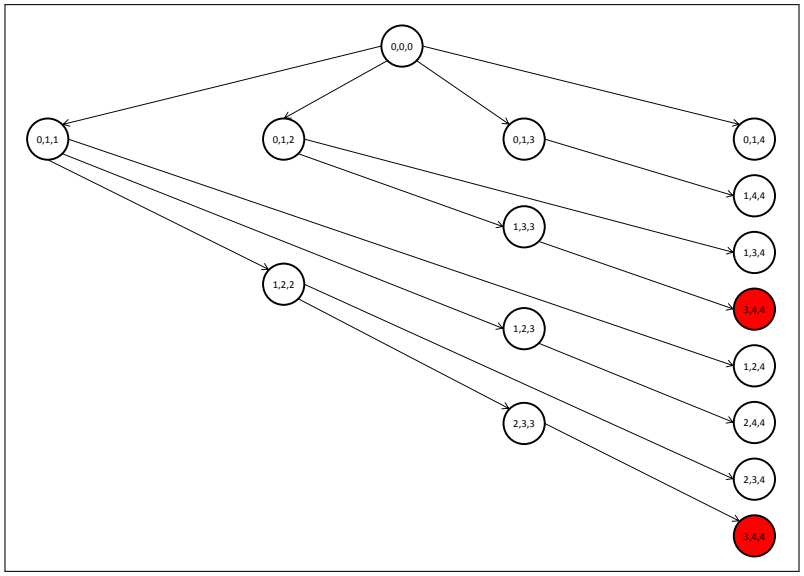


Figure 11.10: Tree Representing the States of the DP for an Instance of Four Periods

unique labeling of each node, we use T indices. It is not practical, but we have not yet found a dominance rule which allows us to reduce the number of nodes ahead. The cost associated with each period has T positions $C(1, 2, \cdots, T)$. Each position may take either no value or a value of 1 or 0. If a purchase setup occurs in period t, its position takes value 1. If the demand of a period is covered but no setup occurs, its position is set to 0. Otherwise the position is vacant. For instance, $C(1, 0, 1, -, \cdots)$ is the cost associated with a node in the third period. It means that, with a procurement setup in the first period, we cover the demand of the first and second periods. A setup in the third period also occurs to cover the demand of the third period. An example of this type of labelling for a tree representing the states of the DP for four periods is depicted in 11.11. In order to calculate the recurrence equations of the DP, we need to define the state at each node. A state is defined by a vector $\pi = (p_{i,j}, s_{i,j}, s_{i,j}^{l+})$. We define each element of the status vectors as follows.

The procurement quantity $p_{i,j}$ (with $i \leq j$) placed in period i to cover the demand

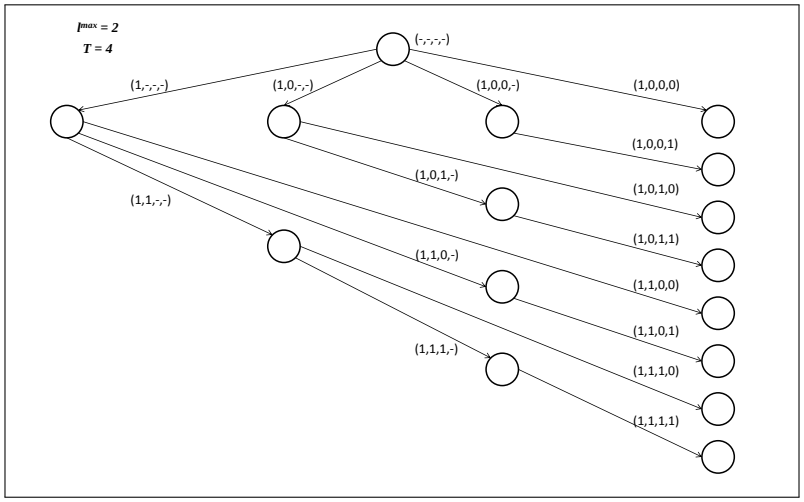


Figure 11.11: Tree Representing the States of the DP for an Instance of Four Periods with Correct Labelling

of periods i to j is calculated using (11.53) below.

$$p_{i,j} = \max\{0; \max_{1\leq k\leq j-i+1}\{\frac{\sum_{k'=0}^{k-1} d_{i+k'} - s_{i-1} - \sum_{k'=0}^{k-1}\sum_{l=1}^{l^{max}+1-k'} s_{i-1}^{l+}}{\min\{k, l^{max}+1\}}\}\} \tag{11.53}$$

The final product inventory at the end of period j resulting from ordering $p_{i,j}$ is calculated using (11.54) below.

$$s_{i,j} = (\min\{j-i+1, l^{max}+1\})p_{i,j} + s_{i-1} + \sum_{k'=1}^{j-i}\sum_{l=1}^{l^{max}-k'+1} s_{i-1}^{l+} - \sum_{u=i}^{j} d_u \tag{11.54}$$

The Refresh Wafer inventory at level l (with $l \neq 0$) at the end of period j resulting from ordering $p_{i,j}$ is calculated using (11.55) below.

$$s_{i,j}^{(\min\{j-i+1,l^{max}\})^+} = p_{i,j} \tag{11.55}$$

Recurrence equation

The costs are calculated based on vector π at each state. We begin with (11.56a) and define each position one by one (for instance, the first position is filled as

(11.56b)) until all positions are determined.

$$C(-,\cdots,-) = 0 \tag{11.56a}$$

$$C(1,\cdots,-) = \text{By ordering } p_{1,1} \text{ using (11.53) and (11.54) and substituting in the objective function (11.48).} \tag{11.56b}$$

The optimal solution is obtained when all positions of $C(\bullet)$ are set to 1 or 0. The minimum value of $C(\bullet)$ provides the optimal solution value and purchasing periods.

Dominance Rules

The number of nodes increases exponentially in the tree representing the DP. Some filtering rules may help to discard some nodes and branches.

- With non decreasing demand over the planning horizon, Proposition 11.5.5 and 11.5.6 help to define the regeneration intervals more easily.
- If for all periods (t'') between two periods of t and t', we have (and generate) enough raw materials to satisfy the demand, then we do not need new raw material procurement. Formally, we set periods $(t+1,\cdots,t')$ as non-purchase periods if (11.57) holds:

$$\sum_{k=1}^{t-t''+1} \sum_{l=1}^{l^{max}-k} s_{t-1}^{l^+} + s_{t-1} + (\min\{t''-t+1, l^{max}\})p_t \geq \sum_{u=t}^{t''} d_u \quad t \leq t'' \leq t' \tag{11.57}$$

- The raw materials (Top Wafer) stock of the arcs departing from the same node exhausts after $l^{max}+1$ periods. However, the final product stock of the nodes may not always be null. If the stock levels are equal, the node with the minimum cost is kept and the other nodes are discarded. If the final product stock levels are different, the node with the least final product stock value is kept and the others are discarded.
- In each period, among the nodes ending with the same inventory vector, the least expensive is kept and the others are discarded.

Using Proposition 11.5.6 a backward DP can be proposed. An analysis should then be conducted on several variants of DP to determine their efficiency.

As argued before, by relaxing "push" assumptions new decisions arise. In order to solve ULS-MR$_1$, the proposed DP can be adapted. The maximum number of the states of the DP will be the same. However at each node, the mentioned decisions regarding Fresh Wafer and Negative Wafers must be taken.

11.6 Conclusions and Perspectives

In this chapter, we introduced in detail a novel closed-loop production planning problem derived from a real world industrial case. The model can be extended to cover other real applications, for instance in recycling or circular economy. The lot-sizing model contains three setup constraints associated with *purchase*, *remanufacturing* and *production*. The complexity analysis shows that the uncapacitated single-item model is NP-hard. Note that one characteristic which makes the problem difficult is that the SOI Wafer (final product) fabrication is not only driven by the demand satisfaction but also by Negative Wafer generation. The generated Negative Wafers are then used to satisfy SOI Wafers.

In a first attempt to solve the problem, we simplified it by eliminating two setup constraints of *production* and *remanufacturing* and keeping only *purchase* setup constraints. Properties of this problem were explored and a model with less variables and constraints was proposed. In order to develop a dynamic programming to solve the problem, we considered new assumptions, called "Full Push Policy". It implies that once a replenishment (or refresh) occurs, we push the production system to fabricate final products. In the same way, we immediately remanufacture the generated by-products. We proposed a dynamic programming resolution algorithm. Note that the "Full Push Policy" assumptions help to simplify the problem by eliminating several decisions.

Multiple perspectives can be imagined. First, it is necessary to study the complexity of the reduced models. Then, improving the dynamic programming algorithms may lead to pseudo-polynomial algorithms. Also, the problem may have other properties which could be identified. Regarding the dynamic programming

algorithm, based on extensive experimentation, it could be analyzed whether the number of states explodes in practice. The idea is to check the number of explored states in comparison to the worst case.

As the original model is NP-hard, heuristics can be proposed for its resolution. It is worth to observe that some of the properties of an optimal solution discussed for the model with single purchase setup are also valid for the original problem. An adaptation of the Silver-Meal heuristic, or least cost heuristics or other classic heuristics with worst case performance guarantee could be derived. The traditional formulation proposed in this chapter provide weak lower bounds. In order to obtain better lower bounds, alternative formulations of the problem can be proposed. The tightness of formulations can be evaluated by comparing the LP relaxations and MIP computational times.

Chapter 12

General Conclusions and Perspectives

With this final chapter, we come to the end of our journey, even if an endless way is still to explore. Conclusions and perspectives relevant to each subject are presented at the end of the chapters. Sections 3.6, 4.5, 5.5, 6.6, 7.5 together with Section 8.1 of the conclusive Chapter 8 are dedicated to Part I. Therefore, in this brief chapter, we do not aim to repeat all conclusions and perspectives, but to present a more global and comprehensive picture of the various contributions of the thesis while pointing out how the two parts can be connected.

12.1 Conclusions

Early in Chapter 2, the framework of the thesis was outlined. It was conducted in the company Soitec, a world leader for high performance semiconductor material. The flagship product of the company, called SOI (Silicon-On-Insulator) wafer, is used to make efficient microelectronic components.

The microelectronics business environment is unstable and highly competitive. Moreover, the fabrication costs are high and the production is dynamic. Therefore, flexibility and agility are identified as the ants to play any game in semiconductor manufacturing systems.

In semiconductor manufacturing, at each operation step, a recipe is defined for each product. A recipe must be qualified on a machine in order to be able to allocate the production volume of the recipe to the machine. This restriction has a direct impact on capacity utilization of toolsets. Chapter 3 discusses the impact

of qualification management (QM) on capacity optimization and flexibility increase which is the subject of the first part of this thesis. By the concept "flexibility", we try to evaluate the workload balance on a toolset; the better the workload balance, the higher the flexibility. An adequate qualification configuration gives flexibility for recipe workload allocations to machines. Hence, we aim at increasing flexibility in workload allocation. Increasing flexibility leads to capacity optimization.

Industrial constraints affect the QM. First, the influence of the limited equipment capacity on capacity allocation and QM was studied in 4. Capacitated flexibility measures were presented to evaluate the flexibility of a workload balance on a toolset. Additional measures, called "deviation ratio" are necessary to help the decision making process for performing new qualifications.

The presented work is interesting from the point of view of workload balancing and capacity allocation. Chapter 6 also considers workload balancing with capacity restrictions. Our criteria is to minimize the deviation of the total workload allocated to a machine from the maximum available time of each machine *while balancing the workload on the whole toolset*. In Section 6.4.2, weighted variability measures were discussed. It is possible modify the criteria to minimize the number of overloaded (or underloaded) machines by considering a big weight for corresponding cases.

Batching is a frequent production characteristic. The influence of batching on qualification management was studied in Chapter 5. The sequencing of allocations were not considered while workload balancing. But when making full batches, the production volume of each recipe (of product) is as if it is "packed" in a package, which is the batch size. Therefore, the batches allocated to each machines can be sequenced freely on the same machine. Hence, workload allocation under batch size constraint can be used for scheduling. In scheduling literature, qualification restrictions are equivalent to machine eligibility restrictions. Qualifications management is similar to configurability in flexible manufacturing systems (FMS).

Finally, the industrialization of the concept at Soitec was discussed in Chapter 7. This decision making process is new in the company. Therefore, data exchange interfaces were created to facilitate the continuous usage of the concept.

In conclusion, by a better QM, we try to decreases the WIP on the shopfloor (leading to more floor space), decrease the inventory level, increase machine capacity

utilization, reduce the cycle time, improve the machine investment amortization, decrease of production time (for instance, by allocating products to machines with shorter process times). The objective is to increase flexibility at a minimum cost to accommodate demand variability, product mix variations and process and design changes.

In the second part of the thesis, the production planning of the SOI fabrication and refresh process lines were discussed. The production planning problem tackled is novel as Soitec disposes of its exclusive the production procedure for SOI fabrication. The two production lines are related and make a so-called *closed-loop manufacturing system*. While fabricating final products, by-products are generated in the main production line. Once remanufactured, the by-products return to the main production line to make final products. The manufacturing system was modeled with all specific constraints of Soitec.

Based on the industrial problem, a single-item uncapacitated lot-sizing model was defined to study the production planning of a special type of closed-loop manufacturing systems. To the best of our knowledge, this problem has never been treated in previous studies. The problem is NP-hard. Therefore, several reduced variants of the original problem were defined and studied. For one of the reduced models, based on the optimal solution structure, an exact resolution method using dynamic programming was proposed.

12.2 Perspectives

Various perspectives were discussed in the end of each chapter. In Section 4.5, the bi-objective optimization of capacitated flexibility and deviation ratio measures was proposed. Bi-objective optimization aims at finding a trade-off compromise between flexibility (workload balance) increase and decrease of the workload allocation deviation from the equipment capacity. In Section 5.5, possible improvements the proposed workload balancing algorithms with batching were discussed. Moreover, Appendix A is dedicated to batching when considering minimum and maximum batch sizes. Section 6.6 considers (local) workload variability in a workcenter and the qualification configuration flexibility in workload allocation. The proposed ap-

proach can be used to measure the incurred variability due to other variability factors such as batching. Section 7.5 proposed other industrial constraints while QM, some examples are recipe priority, production yield, auxiliary resource management, consumables, material handling system, etc.

Losing a qualification may decrease the toolset flexibility. Using the discussed approach, it is possible to evaluate the flexibility loss related to a disqualification. We named this *recipe-to-machine criticality*. In the same way it is possible to calculate *machine criticality*. This valuable measure can help for preventive maintenance (PM) planning. By calculating the machine criticality in advance, it is possible to schedule PM in a period where machine criticality is at the least. Another interesting perspective is to study QM over time instead of considering one production period. The qualification setup cost was not taken into account. In dynamic QM, the setup costs and the disqualification issue should also be considered. The perspectives relating to Part II of the thesis were examined in Sections 10.7 and 11.6.

Now perspectives relating both parts of the thesis are presented. In the first part, the optimal toolset capacity utilization is calculated to determine new qualification(s). Therefore, new qualifications are levers for increasing capacity. On the other hand, we know that production planning is constrained by production capacity. Therefore, the two concepts are complementary and can be combined.

Qualification Management and Production Planning

So far, it was shown that the recipe-to-machine qualification configuration directly affects the production capacity. Here, we open the discussion of how qualification management and production planning may be combined by adding new constraints to a capacitated lot-sizing model. It is not our goal to re-write a complete capacitated lot-sizing model. Merely necessary constraints are mentioned. In Part I, we have always used throughput TP and not process time PT. Both are equivalent, i.e. $TP = 1/PT$. If a recipe is not qualified (or qualifiable) on a machine, its throughput tends to infinity or, in other words, its process time is zero.

Dynamic versions of the QM model and QM models for multiple qualifications were proposed in [53]. Without rewriting the whole model, we open the discussion

of how on combine lot-sizing and qualification management. Consider the following parameters and variables.

Parameters

$PQ_{r,m}$	$\begin{cases} 1 \text{ if recipe } r \text{ is qualifiable on machine } m, \\ 0 \text{ if recipe } r \text{ is not qualifiable on machine } m. \end{cases}$
$pt_{r,m}$	Process time of recipe r on machine m: $\begin{cases} > 0 \text{ if recipe } r \text{ is qualifiable or already qualified on machine } m, \\ 0 \quad \text{Otherwise.} \end{cases}$
c_m^t	Capacity of machine m in period t,
d_r^t	Demand of recipe r in period t,
$cp_{r,m}$	Production cost of recipe r in period t,
h_r	Holding cost of recipe r,
$sp_{r,m}^t$	Setup cost if recipe r is produced on machine m in period t,
$sq_{r,m}^t$	Setup cost if recipe r is qualified on machine m.

Variables

$X_{r,m}^t$	Production volume of recipe r assigned to machine m in period t,
$Y_{r,m}^t$	Setup binary variable of recipe r assigned to machine m in period t,
s_r^t	Inventory of recipe r in the end of period t,
$OQ_{r,m}$	$\begin{cases} 1 \text{ if recipe } r \text{ is to be qualified on machine } m, \\ 0 \text{ otherwise.} \end{cases}$

Two cases may be considered. First, we do not intend to perform new qualifications. In this case the qualification must be considered in the setup constraint and workload allocation. However, if additional new qualifications are allowed, the

mathematical programming would be as follows:

$$\min \sum_{t=1}^{T}\sum_{r=1}^{R}\sum_{m=1}^{M} cp_{r,m}X_{r,m}^{t} + \sum_{t=1}^{T}\sum_{r=1}^{R} h_r s_r^t + \sum_{r=1}^{R}\sum_{m=1}^{M} sq_{r,m}OQ_{r,m} + \sum_{t=1}^{T}\sum_{r=1}^{R}\sum_{m=1}^{M} sp_{r,m}^{t}Y_{r,m}^{t} \tag{12.1a}$$

subject to

$$s_r^{t-1} + \sum_{m=1}^{M} X_{r,m}^{t} = d_r^t + s_r^t \qquad \forall r,t \tag{12.1b}$$

$$\sum_{r=1|Q_{r,m}=1,PQ_{r,m}=1}^{R} (pt_{r,m}^{t}X_{r,m}^{t} + OQ_{r,m}cq_{r,m}) \leq c_m^t \qquad \forall m,t \tag{12.1c}$$

$$X_{r,m}^{t} \leq M(Q_{r,m} + OQ_{r,m})Y_{r,m}^{t} \qquad \forall r,m,t \tag{12.1d}$$

$$\sum_{r=1}^{R}\sum_{m=1|PQ_{r,m}=1}^{M} OQ_{r,m} \leq k \qquad \forall t \tag{12.1e}$$

$$\sum_{r=1}^{R}\sum_{m=1|PQ_{r,m}=0}^{M} OQ_{r,m} = 0 \qquad \forall t \tag{12.1f}$$

$$X_{r,m}^{t}, s_r^t \geq 0 \qquad \forall r,m,t \tag{12.1g}$$

$$Y_{r,m}^{t}, OQ_{r,m} \in \{0,1\} \qquad \forall r,m,t \tag{12.1h}$$

The objective function (12.1a) tries to minimize the production, holding, new qualification and production setup costs. Constraint (12.1b) guarantees the flow conservation balance. Production capacity is respected using Constraint (12.1c). Constraint (12.1d) represents the production setup.

Note that performing new qualifications is costly. Therefore, the maximum number of qualifications can be directly determined in each period. The qualification cost in the objective function also limits the number of qualifications. The term $(Q_{r,m} + OQ_{r,m}^{t})Y_{r,m}^{t}$ in (12.1d) contains the product of two binary variables which can be replaced by an additional binary variable, say $Z_{r,m}^{t}$. The term must be substituted by $Q_{r,m}Y_{r,m}^{t} + Z_{r,m}^{t}$ and the set of Constraints (12.2) are to be added to the model. Constraint (12.1e) defines the allowed number of qualifications (k). Note that k must be smaller than or equal to the number of qualifiable recipe-to-

machine couples. Constraint (12.1f) ensures that if a recipe-to-machine couple is not qualifiable ($PQ_{r,m} = 0$), it is not considered to be qualified.

$$Z^t_{r,m} \leq OQ^t_{r,m} \qquad \forall r, m, t \qquad (12.2a)$$
$$Z^t_{r,m} \leq Y^t_{r,m} \qquad \forall r, m, t \qquad (12.2b)$$
$$Z^t_{r,m} \geq OQ^t_{r,m} + Y^t_{r,m} - 1 \qquad \forall r, m, t \qquad (12.2c)$$
$$Z^t_{r,m} \in \{0, 1\} \qquad \forall r, m, t \qquad (12.2d)$$

Based on the qualification difficulty, monetary cost, time consumption, importance, etc., the qualification decision may belong to two different decisions levels, tactical or more operational. Lot-sizing decisions are made at a tactical level. So if the qualification decision is more operational, it must be carefully though of how to relate these two types of decisions. Studies considering the integration of decisions at different levels attract the attention of researchers and practitioners, for instance integration of lot-sizing and distribution, or lot-sizing and scheduling (see [22] and [36]). Therefore, it may be possible to inspire from similar researches to relate both decision levels.

Flexible Production Planning - Limit Changes

Demands, customer expectations, production factors, lead times and other parameters are subject to constant changes. Therefore, the production plan must be revised constantly. However, abrupt changes in the production plan are not desired. It makes the scheduling, material sourcing and in general the functioning difficult. Moreover, the users of the production plans lose their confidence in the results. Hence, it is preferable to have robust production plans throughout the planning horizon. A similar approach as what was presented in Part I and in particular in Section 6.4.2 could be used i.e. to penalize the differences between a new production plan and the current one. By adjusting linear and non-linear weights, the discrepancies of the current production plan from the previous can be reduced. In this case, the objective function needs to be adapted consequently.

Flexible Production Planning - Cost Profile

Due to purchase and refresh cost structure, it is possible that the model recommends campaigns of Fresh Wafer purchase and refresh process. These campaigns cause irregular cost profiles along the whole planning horizon. In order to better manage the financial elements of the planning, it is preferable to create an almost constant cost expense over the whole horizon instead of periodic costs. Here also, it is possible to extend the approach proposed in Chapter 6 to overcome the undesired abrupt cost fluctuations.

Flexible Production Planning - End-of-horizon Effects

One of the undesired effects which one may encounter while planning production is the *End-of-horizon Effects*. In a lot-sizing problem, the model tends to end up with zero inventory levels. As we all know that the world does not come to its end with the end of planning horizon, it causes difficulties for planner. Several approaches are proposed to overcome this undesired end-of-horizon effect. One of them is to consider sufficiently long (to be also defined!) planning horizons, i.e. more than needed. Another approach consists in considering some kind of safety stock or minimum ending inventory position.

In the case where backlogging is possible, the model may tend to backlog instead of satisfying demand as we approach the end of horizon. One possible way is to penalize more the backlogging cost as we approach the end of horizon. Of course, this penalization can be controlled by some sort of linear and/or nonlinear weights (see for instance Section 6.3.3).

Flexibility in Supply Chain Management

The concept of flexibility increasingly attracts the attention of practitioners and researchers, in particular in supply chain management. Some reasons motivating why a supply chain must be flexible due to new conditions are discussed in [90]. Globalization, increasing customer expectations as well as lead times, labor cost increase in developing countries, energy cost fluctuations, increase in logistic costs,

increased risk in supply and many other factors cause that decisions in supply chain management may need to be revised in relatively short times. That is why supply chains must be more flexible and agile enough to quickly respond to changes. The new competition environment enumerated by [90] is dynamic, complex and uncertain. Note that these elements exist in a fab in semiconductor manufacturing. The answer of is to increase flexibility to overcome the described supply chain environment. The flexibility is defined by [90] as the ability to respond quickly and in a cost effective way to changes which may come in many forms such as demand volume and mix, prices, costs, etc. The objective is defined as to increase the service level (decrease the amount of unsatisfied demand) by improving the capacity utilization while reducing cost, of course as quickly as possible. Note that we treated in Part I exactly the capacity utilization increase, while reducing (or even eliminating) overloaded machines (which decrease the service level by leaving some demand unsatisfied). All this by reducing costs incurred by under-utilization of machines, operator performance decrease, increase of WIP and buffer stock at the shop floor, unsatisfied demand, induced variability in the production line, etc. with the least possible time by performing the best (or a good) qualification. [90] mentions three ways to achieve flexibility: flexibility through *product design*, *process design* and *system design*. Two mentioned points under system design are to increase service level by coping with high forecast error and better resource utilization. Note that these points are treated in Part I. Interestingly, similar results are obtained in case studies in SCM as in Section 6.5, i.e. too much manufacturing flexibility does not necessarily decrease the cost (in this section the results are illustrated by other KPIs such as variability, overload decrease, etc.). As in SCM, the objective is to match supply with demand, in Part I, it was tried to match production capacity with demand, which come actually to the same concepts.

Appendices

Appendix A

Batching: Minimum and Maximum Batch Sizes

A.1 Minimum Batch Size

When we talk about batch size, normally the maximum batch size is considered. By maximum batch size, we mean the maximum number of wafers (or products in general) which can be loaded into a machine during a production run. In semiconductor manufacturing, for furnaces, the maximum batch size is determined according to boat slots, yield and quality issues. For implantors, the size and layout of the wheel determine the maximum batch size.

Apart from the maximum batch size, sometimes the minimum batch size is also defined for each production run. Minimum batch size is defined due to several reasons:

The production run becomes economically feasible Several costs are incurred at each production run:

- Setup time,
- Qualification cost, (if qualification required),
- Filler wafers (for filling vacant places),
- Consumables (if used).

The production quality is guaranteed Minimum batch size also guarantees the production yield and quality.

Avoiding frequent setups Minimum batch size sets a minimum profitability production threshold while avoiding frequent setups.

Keeping machine qualified Qualification consumes time and energy. Some machines are disqualified automatically if the recipe is not executed after a defined time period. Therefore, it is sometimes logical not to wait to make full (maximum) batches and to make a minimum batch size run, in order to keep the machine qualified.

Avoiding WIP bubbles and reduction of variability Full (maximum) batch size runs make the production more profitable. However waiting for the lots to arrive and constitution of a full batch may slow the production flow, clutter the fab and cause WIP bubbles. WIP bubbles propagate the variability in the fab. The larger the batch size, the more intense the impact. So, it is sometimes interesting to make runs with minimum batch sizes.

Not enough production volume A simple reason to use minimum batch size is that the whole production volume does not constitute a full batch run but may fit a minimum batch run.

Increasing machine utilization while decreasing operator performance The minimum batch size guarantees a minimum work level while avoiding the machine being idle and loosing operators performance.

Among the possible approaches, we present two short indications.

A.2 Step Length Adjusting in the Active Set Method

The same approaches described in Chapter 5 can be adapted to consider minimum and maximum batch sizes. One of the heuristics which can be easily adapted is the Batch-Feasibility Heuristic. The other heuristic discussed in Chapter 5 is the Adapted Active-Set Method. We described in Section 5.3.2.1 that in case where full-batches can be made, the constraint becomes active as the variable equals its upper bound [35]. The same reasoning holds for the minimum batch size. Minimum batch size acts like a lower bound. So, when the step length equals the lower bound, the corresponding constraint becomes active. Now the question is how to define the

step length. The answer is that, if with the WIP on-hand, we make (if possible) a maximum batch. In case some remainder is still on-hand, if a minimum batch can be made, we do it. Else we distribute the remainder to the least loaded machine.

A.3 Mathematical Model using Semi-Continuous Variables

The model proposed in Section 5.2.1 can be adapted to consider minimum batch sizes. However, we present here another modeling approach using Semi-Continuous Variables. First, the single machine case is considered and then the case with multiple machines is developed.

A.3.1 Case of a Single Machine

The production volume of each production run must lie between Minimum Batch Size ($minBS$) and Maximum Batch Size ($maxBS$).

Parameters

R	set of all recipes
K	production run ($k \in 1, ..., \subseteq \frac{\sum_r wip_r}{minBS_r}$),
M	set of all machines,
$q_{r,m}$	$\begin{cases} 1 \text{ if recipe } r \text{ is qualified on machine } m, \\ 0 \text{ otherwise.} \end{cases}$
$pq_{r,m}$	$\begin{cases} 1 \text{ if recipe } r \text{ is qualifiable on machine } m, \\ 0 \text{ if recipe } r \text{ is not qualifiable on machine } m. \end{cases}$
wip_r	total production volume of recipe r,
$minBS_{r,m}$	minimum batch size of recipe r on machine m,
$maxBS_{r,m}$	maximum batch size of recipe r on machine m,
v	number of desired qualifications.

Variables

$Y_{r,k,m}$ $\begin{cases} 1 \text{ if recipe } r \text{ is in the production run } k \text{ on machine } m, \\ 0 \text{ otherwise.} \end{cases}$

$X_{r,k,m}$ production volume of recipe r in production run k on machine m,

$OQ_{r,m}$ $\begin{cases} 1 \text{ if recipe } r \text{ is to be qualified on machine } m \\ 0 \text{ otherwise.} \end{cases}$

Constraints

$$\sum_k Y_{r,k} X_{r,k} = wip_r \qquad \forall r \tag{A.1}$$

$$\sum_r Y_{r,k} = 1 \qquad \forall k \tag{A.2}$$

$$X_{r,k} \geq minBS_r Y_{r,k} \qquad \forall k \tag{A.3}$$

$$X_{r,k} \leq maxBS_r Y_{r,k} \qquad \forall k \tag{A.4}$$

Constraint A.1 guarantees that the sum of the production volume of each recipe of type r in each production run k is equal to the total amount of production volume of recipe r. Constraint A.2 states that each production run consists of only one one recipe and no empty runs are authorized. In order to allow empty runs and at the same time oblige each production run to only consist of one recipe, Constraint A.2 must be changed to:

$$\sum_r Y_{r,k} \leq 1 \qquad \forall k \tag{A.5}$$

Constraint A.3 and A.4 state respectively that the number of wafers of each recipe at each production run must lie between $minBS_r$ and $maxBS_r$. $Y_{r,k}$ is binary, and $X_{r,k}$ is continuous. The product of a binary and a continuous variable creates a *semi-continuous variable*. Constraint A.1 is nonlinear.

A *semi-continuous variable* can take the value zero or any continuous value between its lower and upper bounds. The lower bound is any positive value while the upper bound can be infinite. Semi-continuous variables are often used in real-world modeling, such as production planning, transportation, electrical power generation, portfolio selection, etc. Our case is similar to production planning in which either no products are fabricated or if fabricated, the production must be higher than a min-

imum amount. In transportation is the same, no shipment is done until a minimum shipment amount is reached.

For our problem, the semi-continuous variable $Z_{r,k}$ is defined as follows:

$$Z_{r,k} \in \{0\} \cup [minBS_r, maxBS_r] \Leftrightarrow minBS_r Y_{r,k} \leq Z_{r,k} \leq MaxBS_r Y_{r,k}. \tag{A.6}$$

The semi-continuous variable is transformed under the form of a system of mixed inequalities:

$$\begin{cases} Y_{r,k}X_{r,k} = X_{r,k} \text{ if } Y_{r,k} = 1 \\ Y_{r,k}X_{r,k} = 0 \text{ if } Y_{r,k} = 0 \\ a = Min_r(X_{r,k}) \\ b = Max_r(X_{r,k}) \end{cases} \Leftrightarrow \begin{cases} Y_{r,k}X_{r,k} \leq bX_{r,k} \\ Y_{r,k}X_{r,k} \geq aX_{r,k} \\ a = Min_r(X_{r,k}) \\ b = Max_r(X_{r,k}) \end{cases} \tag{A.7}$$

$$Y_{r,k}X_{r,k} \leq bY_{r,k} \quad \forall r \tag{A.8}$$

$$Y_{r,k}X_{r,k} \geq aY_{r,k} \quad \forall r \tag{A.9}$$

$$Y_{r,k}X_{r,k} \leq X_{r,k} - a(1 - Y_{r,k}) \quad \forall r \tag{A.10}$$

$$Y_{r,k}X_{r,k} \geq X_{r,k} - b(1 - Y_{r,k}) \quad \forall r \tag{A.11}$$

$$\sum_k Y_{r,k}X_{r,k} = wip_r \quad \forall r \tag{A.12}$$

$$\sum_r Y_{r,k} = 1 \quad \forall k \tag{A.13}$$

$$X_{r,k} \geq minBS_r Y_{r,k} \quad \forall k \tag{A.14}$$

$$X_{r,k} \leq maxBS_r Y_{r,k} \quad \forall k \tag{A.15}$$

a and b, defined respectively as the minimum and maximum values of the variable $X_{r,k}$ over r, corresponds to $minBS_r$ and $maxBS_r$.

According to A.7, Constraints A.14 and A.15 can be written as:

$$0 \leq Y_{r,k}X_{r,k} \leq maxBS_r. \tag{A.16}$$

$Y_{r,k}X_{r,k}$ is substituted by the semi-continuous variable $Z_{r,k}$.

$$Z_{r,k} \leq maxBS_r Y_{r,k} \qquad \forall r,k \tag{A.17}$$
$$Z_{r,k} \geq minBS_r Y_{r,k} \qquad \forall r,k \tag{A.18}$$
$$Z_{r,k} \leq X_{r,k} - minBS_r(1 - Y_{r,k}) \qquad \forall r,k \tag{A.19}$$
$$Z_{r,k} \geq X_{r,k} - maxBS_r(1 - Y_{r,k}) \qquad \forall r,k \tag{A.20}$$
$$\sum_k Z_{r,k} = wip_r \qquad \forall r \tag{A.21}$$
$$\sum_r Y_{r,k} = 1 \qquad \forall k \tag{A.22}$$
$$0 \leq Z_{r,k} \leq maxBS_r \qquad \forall r,k \tag{A.23}$$

A.3.2 Case of Multiple Machines

$$\max \quad F \tag{A.24}$$
$$\text{Subject to}$$
$$Z_{r,k,m} \leq maxBS_{r,m} Y_{r,k,m} \qquad \forall r,k,m \tag{A.25}$$
$$Z_{r,k,m} \geq minBS_{r,m} Y_{r,k,m} \qquad \forall r,k,m \tag{A.26}$$
$$Z_{r,k,m} \leq X_{r,k,m} - minBS_{r,m}(1 - Y_{r,k,m}) \qquad \forall r,k,m \tag{A.27}$$
$$Z_{r,k,m} \geq X_{r,k,m} - maxBS_{r,m}(1 - Y_{r,k,m}) \qquad \forall r,k,m \tag{A.28}$$
$$(\sum_k \sum_m Z_{r,k,m}) q_{r,m} = wip_r \qquad \forall r \tag{A.29}$$
$$\sum_r Y_{r,k,m} = 1 \qquad \forall k,m \tag{A.30}$$
$$0 \leq Z_{r,k,m} \leq maxBS_{r,m} \qquad \forall r,k,m \tag{A.31}$$
$$\sum_k Z_{r,k,m} \leq M q_{r,m} \qquad \forall r,m \tag{A.32}$$

In the Flexibility, Variability and Deviation Ratio Measures, $WIP_{r,m}$ is to be replaced by $\sum_k Z_{r,k,m}$.

Remark In the above model, the minimum production volume of each recipe (wip_r) must be equal to its minimum batch size ($minBS_r$). It is possible to define

supplementary variables to cover the case where $wip_r \leq minBS_r$.

A.3.3 Case of Multiple Qualifications

$$\max \quad F \tag{A.33}$$

Subject to

$$Z_{r,k,m} \leq maxBS_{r,m} Y_{r,k,m} \qquad \forall r,k,m \tag{A.34}$$

$$Z_{r,k,m} \geq minBS_{r,m} Y_{r,k,m} \qquad \forall r,k,m \tag{A.35}$$

$$Z_{r,k,m} \leq X_{r,k,m} - minBS_{r,m}(1 - Y_{r,k,m}) \qquad \forall r,k,m \tag{A.36}$$

$$Z_{r,k,m} \geq X_{r,k,m} - maxBS_{r,m}(1 - Y_{r,k,m}) \qquad \forall r,k,m \tag{A.37}$$

$$\sum_k \sum_m Z_{r,k,m} = (q_{r,m} + OQ_{r,m}) wip_r \qquad \forall r \tag{A.38}$$

$$\sum_r Y_{r,k,m} = 1 \qquad \forall k,m \tag{A.39}$$

$$0 \leq Z_{r,k,m} \leq maxBS_{r,m} \qquad \forall r,k,m \tag{A.40}$$

$$\sum_k Z_{r,k,m} \leq M(q_{r,m} + OQ_{r,m}) \qquad \forall r,m \tag{A.41}$$

$$\sum_{r=1}^{R} \sum_{m=1;pq_{r,m}=1}^{M} OQ_{r,m} = v \qquad \forall r,m \tag{A.42}$$

$$\sum_{r=1}^{R} \sum_{m=1;pq_{r,m}=0}^{M} OQ_{r,m} = 0 \qquad \forall r,m \tag{A.43}$$

Using the Constraint A.42, number of desired qualifications are determined. Constraint A.43 avoids to qualify recipes-to-machine couples which are not qualifiable.

A.3.4 Best Qualification Selection: Alternative Criterion

Flexibility, Variability and Deviation Ratio Measures try to balance the best the workload on the toolset. These measures can be used in the precedent model as the objective function. However, while scheduling with minimum and maximum batch size, it would be preferable to choose a qualification which maximizes the total

number of maximum and minimum batch size runs on the toolset. As perspective, it is interesting to make the model choose between minimum and maximum batch size by favoring bigger batch sizes, for instance with a nonlinear piece-wise function A.1.

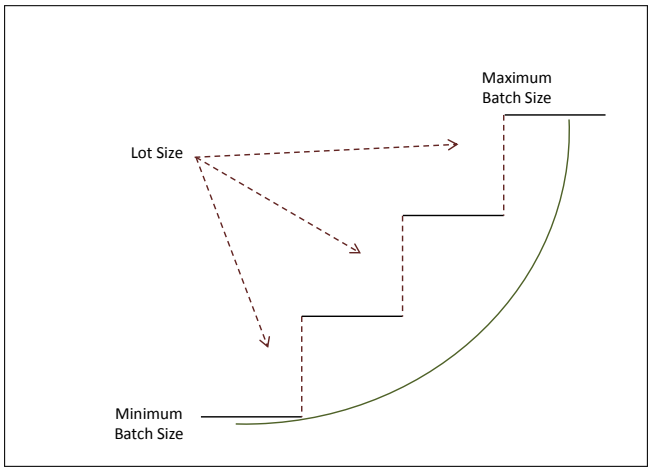


Figure A.1: Nonlinear Piecewise Function to Model the Batch Sizes between the Minimum and the Maximum Batch Sizes

Appendix B

SOI-Refresh Lines Modeling

Consider the refresh and SOI production processes as a whole system. This system has different states, such as Fresh, Refresh 1, Negative 1x, Scrap, etc. The models described until now does not consider all of the refresh process line charachteristics in detail. The system is simplified and reduced to its essential elements (Figure B.1). Now, let us consider the system in more detail. The characteristics

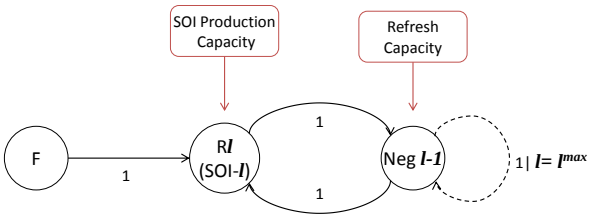


Figure B.1: Graphical Representation of the Fundamental Elements of the Refresh-SOI Processes

of the system evolve at discrete points in time, for instance a Fresh is used to make SOI wafers (and there are some scraps), the generated Negative 1x is refreshed to give some Refresh 1 wafers and some scraps, and so on. Let X_t represent the characteristics of the system at time t. X_t is a random variable as its value is not known certainly before time t. A description of the relationship between the random variables $X_t \quad \forall t$ is a **discrete-time stochastic process**.

Here, we consider a special type of discrete-time stochastic processes called a *Markov Chain*. In a Markov chain, the probability distribution of the state at time $t+1$ depends only and only of the probability distribution of the state at time t and not other states through which the chain has passed. Lets take a look at our problem. Once a Negative 1x is generated, it only depends upon its precedent state, i.e. SOI production and not two states before, i.e. Fresh state; Or production of

Fresh 1x depends only on its precedent state, i.e. Negative 1x and not two states before, i.e. SOI production.

A special type of Markov chain are **absorbing chains**. An absorbing chain is a Markovian chain in which some of the states are absorbing and the rest are transient states. In such a chain, we begin in a transient state and eventually we are sure to leave the transient state and finishes finally in an absorbing state.

A Fresh Wafer enters for the first time the SOI production line. After the splitting step, a Negative Wafer is generated. The Negative wafer is then sent to the refresh line in order to be used again in the SOI production. The generated Negative Wafer is refreshed successfully with a defined probability which is equal to the refresh line yield. And the Negative Wafer will be scraped with a probability equal to the complement of the success probability. If the Negative Wafer is successfully refreshed, if reenters the SOI production line and has the chance to be refreshed again until its maximum allowed (qualified) level. Note that rework exists only for refresh process but not SOI production. The refresh process has several states: Fresh, first level Refresh, second level Refresh until last level Refresh, Monitor and Scrap. Refresh last level, Monitor and Scrap are absorbing states while all other states are transient states. For example, first level Refresh is a transient state while there is a path from first level Refresh to Scrap state but there is no path returning from Scrap to first level Refresh.

The transition matrix is written in the following order, first the listed transient states and finally the absorbing states.

$$P = \left(\begin{array}{c|c} Q & R \\ \hline 0 & I \end{array}\right)$$

The transition matrix P consists of fours sub-matrices. Q represents the transition between the transient states. R represents the transition from the transient to absorbing states. 0 is a matrix consisting of zero elements meaning that the transition from an absorbing state to a transient state is impossible. I is an identity matrix meaning that it is never possible to leave an absorbing state. Some facts about absorbing chains can be answered using this kind of matrix decomposition [60]. The

matrix $(I_n - Q)^{-1}$ denoted by F is called the **Markov chain's fundamental matrix**. I_n has the same size as Q. The ijth element of Matrix F gives the expected number of visits to state j before absorption occurs, given that the current state is state i.

F can be expanded as:

$$F = I + Q + Q^2 + ... = (I - Q)^{-1} \tag{B.1}$$

F is a square matrix where rows and columns correspond to the *non-absorbing states*.

Equation B.1 implies:

$$F = Q^0 + Q^1 + Q^2 + ... \tag{B.2}$$

Therefore $Q^t(i, j)$ gives the expected proportion of period t spent in the jth state or in other words, the probability that the process occupies the jth non-absorbing state in period t, given that the process began in the ith non-absorbing state.

The product FR gives the probability that a particular initial non-absorbing state will end up in a particular absorbing state, in case several absorbing states exist.

The expected number of steps (t_i) before the chain is absorbed, given that the starting state is state i, is called *Time to Absorption*. It is calculated as follows:

$$t = Fc \tag{B.3}$$

t is a column vector whose ith element is t_i and c is a column vector with all elements equal to 1. t is simply the sum of all elements of each row of the fundamental matrix F.

$P =$

	Fresh	Refresh 1x	Refresh 2x	Refresh 3x (last level)	Scrap
Fresh	0	1	0	0	0
Refresh 1	0	0	0.97	0	0.03
Refresh 2	0	0	0	0.97	0.03
Refresh 3	0	0	0	1	0
Scrap	0	0	0	0	1

$$F = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \end{array} \begin{array}{c} \begin{array}{ccc} \text{F} & \text{R1} & \text{R2} \end{array} \\ \begin{pmatrix} 1 & 1 & 0.97 \\ 0 & 1 & 0.97 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$FR = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \end{array} \begin{array}{c} \begin{array}{cc} \text{R3} & \text{Scrap} \end{array} \\ \begin{pmatrix} 0.9409 & 0.0591 \\ 0.9409 & 0.0591 \\ 0.97 & 0.03 \end{pmatrix} \end{array}$$

$$Fc = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \end{array} \begin{array}{c} \\ \begin{pmatrix} 2.97 \\ 1.97 \\ 1 \end{pmatrix} \end{array}$$

$$P = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \\ \text{R3} \\ \text{Monitor} \\ \text{Scrap} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{F} & \text{R1} & \text{R2} & \text{R3X} & \text{Monitor} & \text{Scrap} \end{array} \\ \left(\begin{array}{cccc|cc} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.97 & 0 & 0.02 & 0.01 \\ 0 & 0 & 0 & 0.97 & 0.02 & 0.01 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}$$

The corresponding fundamental matrix (F) is as follows:

$$F = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \end{array} \begin{array}{c} \begin{array}{ccc} \text{F} & \text{R1} & \text{R2} \end{array} \\ \begin{pmatrix} 1 & 1 & 0.97 \\ 0 & 1 & 0.97 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$FR = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \end{array} \begin{array}{c} \begin{array}{ccc} \text{R3} & \text{Monitor} & \text{Scrap} \end{array} \\ \begin{pmatrix} 0.9409 & 0.0394 & 0.0197 \\ 0.9409 & 0.0394 & 0.0197 \\ 0.97 & 0.02 & 0.01 \end{pmatrix} \end{array}$$

$$Fc = \begin{matrix} \text{F} \\ \text{R1} \\ \text{R2} \end{matrix} \begin{pmatrix} 2.97 \\ 1.97 \\ 1 \end{pmatrix}$$

Thus, about 94.09% of the Fresh wafers which enter the SOI-and-Refresh Lines are fully refreshed, i.e. until $R3$, about 3.94% can be used as monitor wafers and 1.97% are to be scraped. In reality, the refresh process is not always successful and some

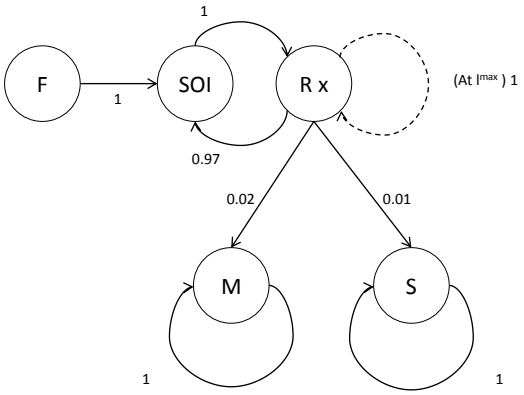


Figure B.2: Graphical Representation of Transition Matrix

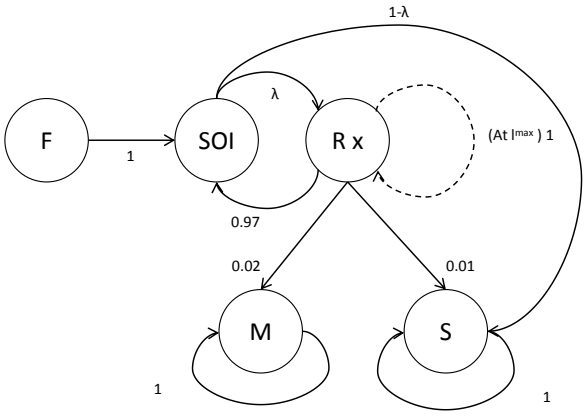


Figure B.3: Graphical Representation of Transition Matrix

wafers are to be refreshed again. Therefore, there is a chance to stay in the refresh state and repeat the process before going to the next step.

$$P = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \\ \text{R3} \\ \text{M} \\ \text{S} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{F} & \text{R1} & \text{R2} & \text{R3} & \text{Monitor} & \text{Scrap} \end{array} \\ \left(\begin{array}{cccccc} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0.05 & 0.95*0.98 & 0 & 0.95*0.015 & 0.95*0.005 \\ 0 & 0 & 0.05 & 0.95*0.98 & 0.95*0.015 & 0.95*0.005 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}$$

$$F = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \end{array} \begin{array}{c} \begin{array}{ccc} \text{F} & \text{R1} & \text{R2} \end{array} \\ \begin{pmatrix} 1 & 1.052 & 1.031 \\ 0 & 1.052 & 1.031 \\ 0 & 0 & 1.052 \end{pmatrix} \end{array}$$

$$FR = \begin{array}{c} \\ \text{R3} \\ \text{Monitor} \\ \text{Scrap} \end{array} \begin{array}{c} \begin{array}{ccc} \text{R3} & \text{Monitor} & \text{Scrap} \end{array} \\ \begin{pmatrix} 0.9604 & 0.0297 & 0.0099 \\ 0.9604 & 0.0297 & 0.0099 \\ 0.98 & 0.015 & 0.005 \end{pmatrix} \end{array}$$

$$Fc = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \end{array} \begin{array}{c} \\ \begin{pmatrix} 3.0842 \\ 2.0842 \\ 1.0526 \end{pmatrix} \end{array}$$

$$P = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \\ \text{R3} \\ \text{M} \\ \text{S} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{F} & \text{R1} & \text{R2} & \text{R3} & \text{Monitor} & \text{Scrap} \end{array} \\ \left(\begin{array}{cccccc} 0 & \lambda & 0 & 0 & 0 & 1-\lambda \\ 0 & \lambda.05 & \lambda(.95*.98) & 0 & \lambda(.95*.015) & \lambda(.95*.005)+(1-\lambda) \\ 0 & 0 & \lambda.05 & \lambda(.95*.98) & \lambda(.95*.015) & \lambda(.95*.005)+(1-\lambda) \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}$$

Let λ be 90%, then F and FR are calculated as follows:

$$F = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \end{array} \begin{array}{c} \begin{array}{ccc} \text{F} & \text{R1} & \text{R2} \end{array} \\ \begin{pmatrix} 1 & 0.9424 & 0.8268 \\ 0 & 1.0471 & 0.9187 \\ 0 & 0 & 1.0471 \end{pmatrix} \end{array}$$

$$FR = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \end{array} \begin{array}{c} \begin{array}{ccc} \text{R3} & \text{Monitor} & \text{Scrap} \end{array} \\ \begin{pmatrix} 0.6928 & 0.0226 & 0.2844 \\ 0.7697 & 0.0252 & 0.2049 \\ 0.8773 & 0.0134 & 0.1091 \end{pmatrix} \end{array}$$

$$Fc = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R2} \end{array} \begin{array}{c} \\ \begin{pmatrix} 2.7692 \\ 1.9658 \\ 1.0471 \end{pmatrix} \end{array}$$

However, when the wafers repeat the refresh process, the associated yield factor is

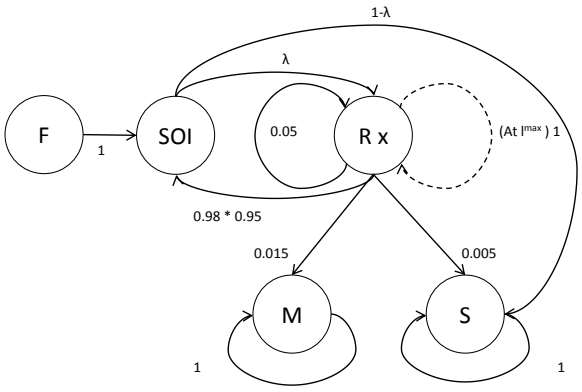


Figure B.4: Graphical Representation of Transition Matrix

not the same. By considering the previous space of states, X_t is no more a Markov chain. This is because if for instance state R1 appears, we don't know if it has already been once processed (1st pass) or twice. Therefore supplementary states

must be defined.

$$P = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R'1x} \\ \text{R2} \\ \text{Monitor} \\ \text{Scrap} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{F} & \text{R1} & \text{R'1x} & \text{R2} & \text{Monitor} & \text{Scrap} \end{array} \\ \left(\begin{array}{cccccc} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.05 & 0.95*0.98 & 0.95*0.015 & 0.95*0.005 \\ 0 & 0 & 0 & 0.8 & 0.15 & 0.05 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}$$

$$F = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R'1x} \end{array} \begin{array}{c} \begin{array}{ccc} \text{F} & \text{R1} & \text{R'1x} \end{array} \\ \left(\begin{array}{ccc} 1 & 1 & 0.05 \\ 0 & 1 & 0.05 \\ 0 & 0 & 1 \end{array}\right) \end{array}$$

$$FR = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R'1x} \end{array} \begin{array}{c} \begin{array}{ccc} \text{R2} & \text{Monitor} & \text{Scrap} \end{array} \\ \left(\begin{array}{ccc} 0.971 & 0.0217 & 0.0072 \\ 0.971 & 0.0217 & 0.0072 \\ 0.8 & 0.15 & 0.05 \end{array}\right) \end{array}$$

$$Fc = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R'1x} \end{array} \begin{array}{c} \\ \left(\begin{array}{c} 2.05 \\ 1.05 \\ 1 \end{array}\right) \end{array}$$

$P =$

$$\begin{array}{c} \\ \text{F} \\ \text{R1} \\ R'1x \\ \text{R2} \\ R'2x \\ \text{R3} \\ \text{M} \\ \text{S} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{F} & \text{R1} & R'1x & \text{R2} & R'2x & \text{R3} & \text{Monitor} & \text{Scrap} \end{array} \\ \left(\begin{array}{cccccccc} 0 & \lambda & 0 & 0 & 0 & 0 & 0 & 1-\lambda \\ 0 & 0 & \lambda 0.05 & \lambda 0.95*0.98 & 0 & 0 & \lambda 0.95*0.015 & \lambda 0.95*0.005+(1-\lambda) \\ 0 & 0 & 0 & 0.8 & 0 & 0 & 0.15 & 0.05 \\ 0 & 0 & 0 & 0 & \lambda 0.05 & \lambda 0.95*0.98 & \lambda 0.95*0.015 & \lambda 0.95*0.005+(1-\lambda) \\ 0 & 0 & 0 & 0 & 0 & 0.8 & 0.15 & 0.05 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}$$

$$F = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R'1x} \\ \text{R2} \\ \text{R'2x} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{F} & \text{R1} & \text{R'1x} & \text{R2} & \text{R'2x} \end{array} \\ \begin{pmatrix} 1 & 0.9 & 0.0405 & 0.8158 & 0.0367 \\ 0 & 1 & 0.045 & 0.9065 & 0.0407 \\ 0 & 0 & 1 & 0.8298 & 0.0373 \\ 0 & 0 & 0 & 1.0373 & 0.0466 \\ 0 & 0 & 0 & 0.8298 & 1.0373 \end{pmatrix} \end{array}$$

$$FR = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R'1x} \\ \text{R2} \\ \text{R'2x} \end{array} \begin{array}{c} \begin{array}{ccc} \text{R3} & \text{Monitor} & \text{Scrap} \end{array} \\ \begin{pmatrix} 0.6836 & 0.0335 & 0.2827 \\ 0.7595 & 0.0373 & 0.2030 \\ 0.6953 & 0.1662 & 0.1384 \\ 0.8691 & 0.0203 & 0.1105 \\ 0.6953 & 0.1662 & 0.1384 \end{pmatrix} \end{array}$$

$$Fc = \begin{array}{c} \\ \text{F} \\ \text{R1} \\ \text{R'1x} \\ \text{R2} \\ \text{R'2x} \end{array} \begin{array}{c} \\ \begin{pmatrix} 2.793 \\ 1.9923 \\ 1.084 \\ 1.05 \\ 1.8672 \end{pmatrix} \end{array}$$

With $\lambda = 0.9$

$P =$

	F	N1	$N^{'}1x$	R1	N2	$N^{'}2x$	R3	N3	Monitor	Scrap
F	0	λ	0	0	0	0	0	0	0	$1-\lambda$
N1	0	0	0.05	$0.95*0.98$	0	0	0	0	$0.95*0.015$	$0.95*0.005$
$N^{'}1x$	0	0	0	0.8	0	0	0	0	0.15	0.05
$R1$	0	0	0	0	λ	0	0	0	0	$1-\lambda$
N2	0	0	0	0	0	0.05	$0.95*0.98$	0	$0.95*0.015$	$0.95*0.005$
$N^{'}2x$	0	0	0	0	0	0	0.8	0	0.15	0.05
$R2$	0	0	0	0	0	0	0	λ	0	$1-\lambda$
N3	0	0	0	0	0	0	0	1	0	0
M	0	0	0	0	0	0	0	0	1	0
S	0	0	0	0	0	0	0	0	0	1

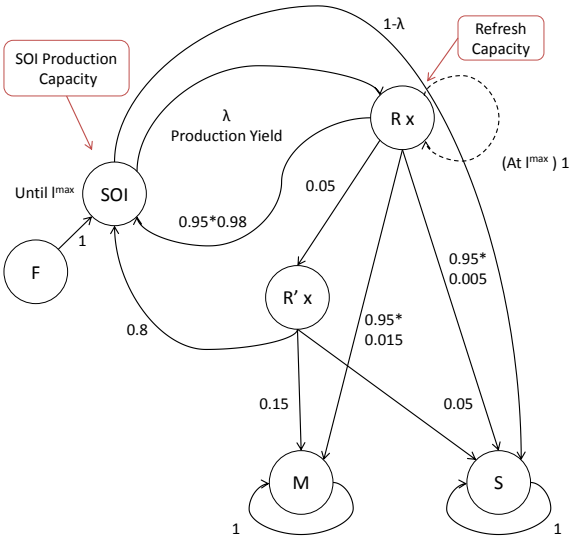


Figure B.5: Graphical Representation of Transition Matrix

$$F = \begin{array}{c} \\ \text{F} \\ \text{N1} \\ \text{N'1} \\ \text{R1} \\ \text{N2} \\ \text{N'2x} \\ \text{R2} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{F} & \text{N1} & \text{N'1} & \text{R1} & \text{N2} & \text{N'2} & \text{R2} \end{array} \\ \begin{pmatrix} 1 & 0.9 & 0.045 & 0.8739 & 0.78651 & 0.0393255 & 0.76370121 \\ 0 & 1 & 0.05 & 0.971 & 0.8739 & 0.043695 & 0.8485569 \\ 0 & 0 & 1 & 0.8 & 0.72 & 0.036 & 0.69912 \\ 0 & 0 & 0 & 1 & 0.9 & 0.045 & 0.8739 \\ 0 & 0 & 0 & 0 & 1 & 0.05 & 0.971 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0.8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

$$FR = \begin{array}{c} \\ \text{F} \\ \text{N1} \\ \text{N'1x} \\ \text{R1} \\ \text{N2} \\ \text{N'2x} \\ \text{R2} \end{array} \begin{array}{c} \begin{array}{ccc} \text{N3} & \text{Monitor} & \text{Scrap} \end{array} \\ \begin{pmatrix} 0.687331089 & 0.036681593 & 0.275987319 \\ 0.76370121 & 0.040757325 & 0.195541465 \\ 0.629208 & 0.16566 & 0.205132 \\ 0.78651 & 0.019575 & 0.193915 \\ 0.8739 & 0.02175 & 0.10435 \\ 0.72 & 0.15 & 0.13 \\ 0.9 & 0 & 0.1 \end{pmatrix} \end{array}$$

Thus, about 76.37% of the Fresh wafers which enter the SOI-and-Refresh Lines are fully refreshed, i.e. until $N3$, about 3.85% can be used as monitor wafers and 1.97%

are to be scraped.

$$Fc = \begin{matrix} \text{F} \\ \text{N1} \\ \text{N'1x} \\ \text{R1} \\ \text{N2} \\ \text{N'2x} \\ \text{R2} \end{matrix} \begin{pmatrix} 4.40843671 \\ 3.7871519 \\ 3.25512 \\ 2.8189 \\ 2.021 \\ 1.8 \\ 1 \end{pmatrix}$$

$P =$

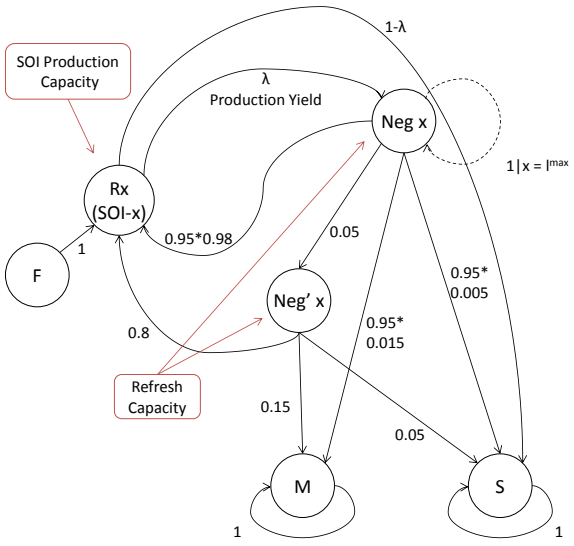


Figure B.6: Graphical Representation of Transition Matrix

	F	R1	R1 1st pass	R1 2nd pass	R2	R2 1st pass	R2 2nd pass	R3	Monitor	Scrap
F	0	λ	0	0	0	0	0	0	0	$1-\lambda$
R1	0	0	0.95	0.05	0	0	0	0	0	0
R1 1st	0	0	0	0	$0.98*\lambda$	0	0	0	$0.015*\lambda$	$(0.005*\lambda)+(1-\lambda)$
R1 2nd	0	0	0	0	$0.8*\lambda$	0	0	0	$0.15*\lambda$	$(0.05*\lambda)+(1-\lambda)$
R2	0	0	0	0	0	0.95	0.05	0	0	0
R2 1st	0	0	0	0	0	0	0.98	0	0.015	0.005
R2 2nd	0	0	0	0	0	0	0.8	0	0.15	0.05
R3	0	0	0	0	0	0	0	1	0	0
Monitor	0	0	0	0	0	0	0	0	1	0
Scrap	0	0	0	0	0	0	0	0	0	1

The refresh first pass can be at the same time a transient and an absorbent state,

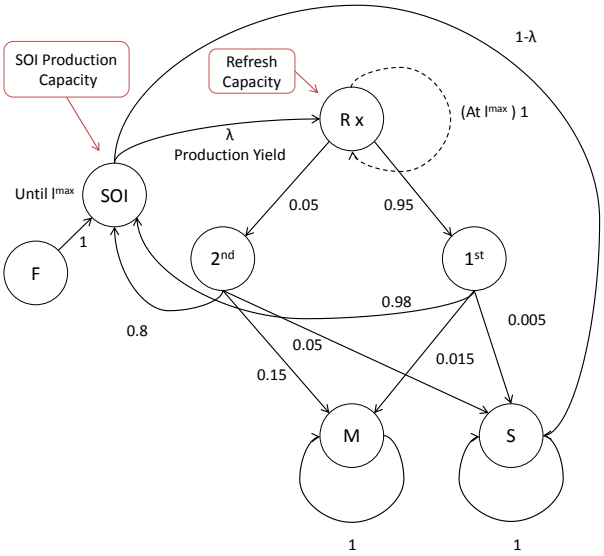


Figure B.7: Graphical Representation of Transition Matrix

with respectively very low and high probabilities.

The value of monitor wafers, also called NPW (Non-productive Wafers) is not considered in our models. One of the advantages of considering the refresh process line in its details is a better management of NPWs.

Appendix C

Acronyms

APS Advanced Planning System

ATO Assemble-to-Order Manufacturing Environment

BOM Bill of Materials

CAD Computer Aided Design

CAM Computer Aided Manufacturing

CLSC Closed-Loop Supply Chain

CLSP Capacitated Lot-Sizing Problem

CMP Chemical Mechanical Polishing/Planarization

CRP Capacity Requirements Planning

CSLP Continuous Setup Lot-Sizing Problem

DLSP Discrete Lot-Sizing and Scheduling Problem

DSP Double Side Polishing

ELSP Economic Lot Scheduling Problem

ELSRj Economic Lot-Sizing with Remanufacturing and Joint setups

ELSRs Economic Lot-Sizing with Remanufacturing and Separate setups

EOQ Economic Order Quantity

ERP Enterprise Resource Planning

ETO Engineer-to-Order Manufacturing Environment

FMS Flexible Manufacturing System

FOUP Front Opening Unified Pod

GLSP General Lot-Sizing and Scheduling Problem

IC Integrated Circuit

JIT Just-in-Time

KPI Key Performance Indicator

LT Lead Time

MES Manufacturing Execution Systems

MF Manufacturing Flexibility

MLCLSP Multi-Level Capacitated Lot-Sizing Problem

MLDLSP Multi-Level Discrete Lot-Sizing and Scheduling Problem

MLGLSP Multi-Level General Lot-Sizing and Scheduling Problem

MLPLSP Multi-Level Proportional Lot-Sizing and Scheduling Problem

MPC Manufacturing Planning and Control

MRP Material Requirements Planning

MRPII Manufacturing Resource Planning

MTO Make-to-Order Manufacturing Environment

MTS Make-to-Stock Manufacturing Environment

NPW Non-productive Wafers

OHT Overhead Hoist Transfer

PLSP Proportional Lot-Sizing and Scheduling Problem

PM Preventive Maintenance

QM Qualification Management

RATO Reassemble-to-Order

RMTO Remanufacture-to-Order

RMTS Remanufacture-to-Stock

ROI Return on Investment

SCC Supply Chain Council

SCM Supply Chain Management

SCOR Supply Chain Operations Reference-Model

SKU Stock Keeping Unit

SQL Structured Query Language

TP Throughput

ULS-MR Uncapacitated Single-Item Bi-Level Lot-Sizing with Multiple Remanufacturing of Reusable By-Products

ULS-MR$_1$ ULS-MR with only Procurement Setup

ULS-MR$_1^{FP}$ ULS-MR$_1$ under "Full Push Policy" Assumptions

WIP Work-In-Progress or Work-In-Process

WW Wagner-Whitin

Bibliography

[1] S. Agrali. A Dynamic Uncapacitated Lot-Sizing Problem with Co-Production. *Optimization Letters*, 6(6):1051–1061, Feb. 2011. (cited on page 166)

[2] R. Anthony. *Planning and Control: a Framework for Analysis*. Cambridge MA: Harvard University Press, 1965. (cited on page 150)

[3] A. Aubry, M.-L. Espinouse, and M. Jacomino. Robust Load-Balanced Configuration with Fixed Costs for the Parallel Multi-Purpose Machines Problem. *International Conference on Service Systems and Service Management*, 2:990–995, Oct. 2006. (cited on page 52)

[4] A. Aubry, M. Jacomino, A. Rossi, and M. L. Espinouse. Maximizing the Configuration Robustness for Parallel Multi-Purpose Machines Under Setup Cost Constraints. *Journal of Scheduling*, 15:457–471, 2012. (cited on page 52)

[5] A. Aubry, A. Rossi, M.-L. Espinouse, and M. Jacomino. Minimizing Setup Costs for Parallel Multi-Purpose Machines Under Load-Balancing Constraint. *European Journal of Operational Research*, 187(3):1115–1125, June 2008. (cited on page 52)

[6] P. J. Billington, J. O. McClain, and L. J. Thomas. Mathematical Programming Approaches to Capacity-Constrained MRP Systems: Review, Formulation and Problem Reduction. *Management Science*, 29(10):1126–1141, 1983. (cited on page 165)

[7] A. Bilyk, L. Mönch, and C. Almeder. Scheduling Jobs with Ready Times and Precedence Constraints on Parallel Batch Machines Using Metaheuristics. *Computers & Industrial Engineering*, 78:175–185, Dec. 2014. (cited on page 54)

[8] G. R. Bitran and D. Tirupati. Hierarchical Production Planning. In *Handbooks in operations research and management science*, pages 523–568. 4 edition, 1993. (cited on page 164)

[9] G. R. Bitran and H. H. Yanasse. Computational Complexity of the Capacitated Lot Size Problem. *Management Science*, 28(10):1174–1186, 1982. (cited on page 180)

[10] K. K. Boyer and G. Leong. Manufacturing Flexibility at the Plant Level. *Omega-International Journal of Management Science*, 24(5):495–510, Oct. 1996. (cited on pages 51 and 54)

[11] S. P. Bradley, A. C. Hax, and T. L. Magnanti. *Applied Mathematical Programming*. Addison-Wesley Pub. Co., 1977. (cited on page 161)

[12] N. Brahimi. *Production Planning : New Lot-Sizing Models and Algorithms*. PhD thesis, Université de Nantes. (cited on page 164)

[13] N. Brahimi, S. Dauzère-Pérès, N. M. Najid, and A. Nordli. Single Item Lot Sizing Problems. *European Journal of Operational Research*, 168(1):1–16, Jan. 2006. (cited on page 164)

[14] J. Browne, D. Dubois, and K. Rathmill. Types of Flexibilities and Classification of Flexible Manufacturing Systems. *The university of Michigan*, pages 1–8, 1984. (cited on page 50)

[15] M. Bruel. Silicon on Insulator Material Technology. *Electronics Letters*, 31(14):1201–1202, 1995. (cited on pages 37 and 192)

[16] L. Buschkühl, F. Sahling, S. Helber, and H. Tempelmeier. Dynamic Capacitated Lot-Sizing Problems: a Classification and Review of Solution Approaches. *OR Spectrum*, 32(2):231–261, Oct. 2008. (cited on page 164)

[17] J. Buzacott and M. Mandelbaum. Flexibility and Productivity in Manufacturing Systems. In *Proceedings of the Annual IIE Conference, Los Angeles, CA.*, pages 404–413, 1985. (cited on page 50)

[18] T.-L. Chen, J. T. Lin, and C.-H. Wu. Coordinated Capacity Planning in Two-Stage Thin-Film-Transistor Liquid-Crystal-Display (Tft-Lcd) Production Networks. *Omega-International Journal of Management Science*, 42(1):141–156, Jan. 2014. (cited on page 49)

[19] P. Ciprut, M.-O. Hongler, and Y. Salama. On the Variance of the Production Output of Transfer Lines. *IEEE Transactions on Robotics and Automation*, 15(1):33–43, 1999. (cited on page 54)

[20] J. J. Coyle, C. J. Langley Jr., R. A. Novack, and B. J. Gibson. *Supply Chain Management: A Logistics Perspective*. South-Western College Pub, 9th edition, 2013. (cited on pages 148 and 149)

[21] P. Damodaran, P. Kumar Manjeshwar, and K. Srihari. Minimizing Makespan on a Batch-Processing Machine with Non-Identical Job Sizes Using Genetic Algorithms. *International Journal of Production Economics*, 103(2):882–891, Oct. 2006. (cited on page 53)

[22] S. Dauzère-Pérès and J.-B. Lasserre. *An Integrated Approach in Production Planning and Scheduling*. Springer-Verlag Berlin Heidelberg, 1st edition, 1994. (cited on page 231)

[23] S. Dauzère-Pérès and L. Mönch. Scheduling Jobs on a Single Batch Processing Machine with Incompatible Job Families and Weighted Number of Tardy Jobs Objective. *Computers & Operations Research*, 40:1224–1233, 2013. (cited on page 53)

[24] S. Dowlatshahi. A Strategic Framework for the Design and Implementation of Remanufacturing Operations in Reverse Logistics. *International Journal of Production Research*, 43(16):3455–3480, 2005. (cited on page 151)

[25] A. EL HAJJ DIAB. *Novel Pseudo-MOSFET Methods for the Characterization of Advanced SOI Substrates*. PhD thesis, Grenoble INP, 2012. (cited on page 36)

[26] D. Erlenkotter. Ford Whitman Harris and the Economic Order Quantity Model. *Operations Research*, 38(6):937–946, 1990. (cited on page 162)

[27] G. Ferrer and D. C. Whybark. Material Planning for a Remanufacturing Facility. *Production and Operations Management*, 10(2):112–124, 2001. (cited on page 167)

[28] R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons Ltd., 2nd edition, 2000. (cited on page 78)

[29] M. Florian, J. K. Lenstra, and A. H. G. Rinnooy Kan. Deterministic Production Planning: Algorithms and Complexity. *Management Science*, 26(7):669–679, 1980. (cited on page 180)

[30] E. Frazelle. *Supply Chain strategy: The Logistics of Supply Chain Management*. McGraw-Hill, 2002. (cited on page 149)

[31] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979. (cited on page 89)

[32] Gaurav Akrani. Types of Production System - Intermittent and Continuous http://kalyan-city.blogspot.com/2012/02/types-of-production-system-intermittent.html, 2012. (cited on pages 17 and 156)

[33] D. Gerwin. Do's and Don'ts of Computerized Manufacturing. *Harvard Business Review*, 60(2):107–116, 1982. (cited on page 50)

[34] S. Ghosh and C. Gaimon. Routing Flexibility and Production Scheduling in a Flexible Manufacturing System. *European Journal of Operational Research*, 60(3):344–364, 1992. (cited on page 51)

[35] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Emerald Group Publishing Limited, 1982. (cited on pages 77, 95 and 238)

[36] E. D. Gomez Urrutia. *Optimisation Intégrée des Décisions en Planification et Ordonnancement dans une Chaine Logistique*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne, 2014. (cited on pages 164 and 231)

[37] S. C. Graves. Manufacturing Planning and Control. 1999. (cited on page 172)

[38] S. C. Graves and B. T. Tomlin. Process Flexibility in Supply Chains. *Management Science*, 49(7):907–919, 2003. (cited on pages 16, 54 and 112)

[39] I. Griva, S. G. Nash, and A. Sofer. *Linear and Nonlinear Optimization*. 2nd edition, 2008. (cited on pages 77, 78 and 115)

[40] V. D. R. Guide Jr. Production Planning and Control for Remanufacturing: Industry Practice and Research Needs. *Journal of Operations Management*, 18:467–483, 2000. (cited on page 189)

[41] V. D. R. Guide Jr., V. Jayaraman, and J. D. Linton. Building Contingency Planning for Closed-Loop Supply Chains with Product Recovery. *Journal of Operations Management*, 21:259–279, 2003. (cited on page 155)

[42] V. D. R. Guide Jr., V. Jayaraman, and R. Srivastava. Production Planning and Control for Remanufacturing : A State-Of-The-Art Survey. *Robotics and Computer-Integrated Manufacturing*, 15(3):221–230, 1999. (cited on pages 151 and 189)

[43] V. D. R. Guide Jr. and L. N. V. Wassenhove. The Reverse Supply Chain. *Harvard Business Review*, pages 10–11, 2014. (cited on page 151)

[44] A. Gupta and A. Sivakumar. Optimization of Due-Date Objectives in Scheduling Semiconductor Batch Manufacturing. *International Journal of Machine Tools and Manufacture*, 46(12-13):1671–1679, Oct. 2006. (cited on page 53)

[45] Y. P. Gupta and S. Goyal. Flexibility of Manufacturing Systems Concepts and Measurements. *European Journal of Operational Research*, 43(2):119–135, 1989. (cited on pages 50 and 51)

[46] F. W. Harris. Operations and Cost. *Factory management series*, pages 48–52, 1915. (cited on page 162)

[47] X.-F. He, S. Wu, and Q.-L. Li. Production Variability of Production Lines. *International Journal of Production Economics*, 107(1):78–87, May 2007. (cited on page 54)

[48] S. Helber and F. Sahling. A Fix-and-Optimize Approach for the Multi-Level Capacitated Lot Sizing Problem. *International Journal of Production Economics*, 123(2):247–256, Feb. 2010. (cited on page 165)

[49] P. Higgins, P. Le Roy, and L. Tierney. *Manufacturing Planning and Control: Beyond MRP II*. Chapman and Hall, 1996. (cited on pages 17, 152 and 153)

[50] W. Hoyer. Variants of the Reduced Newton Method for Nonlinear Equality Constrained Optimization Problems. *Optimization*, 17(6):757–774, 1986. (cited on pages 79 and 96)

[51] J. P. Ignizio. Cycle Time Reduction via Machine-to-Operation Qualification. *International Journal of Production Research*, 47(24):6899–6906, Dec. 2009. (cited on page 52)

[52] Z.-h. Jia and J. Y.-T. Leung. A Meta-Heuristic To Minimize Makespan for Parallel Batch Machines with Arbitrary Job Sizes. *European Journal of Operational Research*, 240(3):649–665, Feb. 2015. (cited on page 54)

[53] C. Johnzén. *Modeling and Optimizing Flexible Capacity Allocation in Semiconductor Manufacturing*. PhD thesis, Department of Manufacturing Sciences and Logistics - Center of Microelectronics in Provence, École des Mines de St-Étienne, 2009. (cited on pages 50, 51, 52, 75, 77, 78, 89, 95, 113, 115, 127 and 228)

[54] C. Johnzén, S. Dauzère-Pérès, and P. Vialletelle. Flexibility Measures for Qualification Management in Wafer Fabs. *Production Planning & Control*, 22(1):81–90, Jan. 2011. (cited on pages 7, 52, 54, 58, 59, 60, 61, 62, 74, 89, 90, 112, 127 and 138)

[55] C. Johnzén, S. Dauzère-Pérès, P. Vialletelle, and C. Yugma. Importance of Qualification Management for Wafer Fabs. In *Advanced Semiconductor Manufacturing Conference (ASMC)*, pages 166–169, 2007. (cited on pages 7, 48 and 49)

[56] W. C. Jordan and S. C. Graves. Principles on the Benefits of Manufacturing Process Flexibility. *Management Science*, 41(4):577–594, 1995. (cited on page 54)

[57] M. L. Junior and M. G. Filho. Production Planning and Control for remanufacturing literature Review and Analysis. *Production Planning & Control*, 23(6):419–435, 2012. (cited on page 190)

[58] K. E. Kabak, C. Heavey, V. Corbett, and P. J. Byrne. Impact of Recipe Restrictions on Photolithography Toolsets in an ASIC Fabrication Environment. *IEEE Transactions on Semiconductor Manufacturing*, 26(1):53–68, Feb. 2013. (cited on page 53)

[59] B. Karimi, S. Fatemi Ghomi, and J. Wilson. The Capacitated Lot Sizing Problem: A Review of Models and Algorithms. *Omega-International Journal of Management Science*, 31(5):365–378, Oct. 2003. (cited on page 164)

[60] J. G. Kemeny and L. J. Snell. *Finite Markov Chains*, volume 40. 1960. (cited on page 246)

[61] J.-P. Kenné, P. Dejax, and A. Gharbi. Production Planning of a Hybrid Manufacturing-Remanufacturing System under Uncertainty within a Closed-Loop Supply Chain. *International Journal of Production Economics*, 135(1):81–93, Jan. 2012. (cited on page 166)

[62] D. S. Kim and J. Alden. Estimating the Distribution and Variance of Time to Produce a Fixed Lot Size Given Deterministic Processing Times and Random Downtimes. *International Journal of Production Research*, 35(12):3405–3414, Dec. 1997. (cited on pages 55 and 110)

[63] S. Kim and R. Uzsoy. Heuristics for Capacity Planning Problems with Congestion. *Computers & Operations Research*, 36(6):1924–1934, June 2009. (cited on page 49)

[64] B. G. Kingsman. Modeling Input - Output Workload Control for Dynamic Capacity Planning in Production Planning Systems. *International Journal of Production Economics*, 68(1):73–93, Oct. 2000. (cited on page 49)

[65] G. F. Knolmayer, P. Mertens, A. Zeier, and J. T. Dickersbach. *Systems, Supply Chain Management Based on SAP: Architecture and Planning Processes*. Springer-Verlag Berlin Heidelberg, 2009. (cited on page 149)

[66] O. Kononchuk and B.-Y. Nguyen. *Silicon-On-Insulator (SOI) Technology: Manufacture and Applications*. Woodhead Publishing Series in Electronic and Optical Materials, 1st edition, 2014. (cited on pages 3, 36, 37 and 39)

[67] X. Li, Y. Huang, Q. Tan, and H. Chen. Scheduling Unrelated Parallel Batch Processing Machines with Non-Identical Job Sizes. *Computers & Operations Research*, 40(12):2983–2990, Dec. 2013. (cited on page 54)

[68] D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming*. Springer, 3rd edition, 2008. (cited on pages 76 and 95)

[69] L. Mönch, J. W. Fowler, S. Dauzère-Pérès, S. J. Mason, and O. Rose. A Survey of Problems, Solution Techniques, and Future Challenges in Scheduling Semiconductor Manufacturing Operations. *Journal of Scheduling*, 14(6):583–599, Jan. 2011. (cited on page 86)

[70] L. Mönch, J. W. Fowler, and S. J. Mason. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems*. Springer-Verlag New York Inc., 2013. (cited on pages 15, 35 and 40)

[71] A. Muriel, A. Somasundaram, and Y. Zhang. Impact of Partial Manufacturing Flexibility on Production Variability. *Manufacturing & Service Operations Management*, 8(2):192–205, Apr. 2006. (cited on pages 16, 54, 110 and 112)

[72] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006. (cited on pages 76, 77 and 78)

[73] A. Obeid. *Ordonnancement et Contrôle Avancé des Procédés en Fabrication de Semiconducteurs*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne, 2012. (cited on page 141)

[74] Y. Pochet and L. A. Wolsey. *Production Planning by Mixed Integer Programming*. Springer Series in Operations Research and Financial Engineering, 2006. (cited on pages 162 and 206)

[75] D. Quadt and H. Kuhn. Capacitated Lot-Sizing with Extensions: A Review. *4OR*, 6(1):61–83, 2008. (cited on page 164)

[76] M. J. Retel Helmrich, R. Jans, W. van den Heuvel, and A. P. Wagelmans. Economic Lot-Sizing with Remanufacturing: Complexity and Efficient Formulations. *IIE Transactions*, 46(1):67–86, Jan. 2014. (cited on page 166)

[77] F. Robert. Impact des Recyclages Successifs du Substrat Donneur sur la Qualité Finale du SOI. Technical report, Université de Rennes 1, 2006. (cited on page 171)

[78] G. L. Rodriguez-Verjan, S. Dauzère-Pérès, and J. Pinaton. Optimized Allocation of Defect Inspection Capacity with a Dynamic Sampling Strategy. *Computers & Operations Research*, 53:319–327, 2015. (cited on page 49)

[79] M. Rowshannahad, N. Absi, S. Dauzère-Pérès, and B. Cassini. A Multi-Item Bi-Level Production Planning Problem with Reusable By-Products. In *10th International Conference on MOdeling, Optimization and SIMulation - MOSIM14*, Nancy, France, 2014. (cited on page 169)

[80] M. Rowshannahad, N. Absi, S. Dauzère-Pérès, and B. Cassini. Bi-Level Lot-Sizing Problems with Reusable By-Products. In *IWLS 2014 - International Workshop on Lot Sizing*, pages 2–5, Porto, Portugal, 2014. (cited on page 191)

[81] M. Rowshannahad and S. Dauzère-Pérès. Qualification Management with Batch Size Constraint. In *Proceedings of the 2013 Winter Simulation Conference. IEEE*, number 2007, pages 3707–3718, 2013. (cited on pages 85 and 90)

[82] M. Rowshannahad, S. Dauzère-Pérès, and B. Cassini. Qualification Management and its Impact on Capacity Optimization. *Advanced Semiconductor Manufacturing Conference (ASMC)*, (Available Online):174–179, May 2013. (cited on pages 54, 83 and 123)

[83] M. Rowshannahad, S. Dauzère-Pérès, and B. Cassini. Qualification Management to Reduce Workload Variability in Semiconductor Manufacturing. In *Proceedings of the 2014 Winter Simulation Conference. IEEE*, 2014. (cited on pages 87 and 109)

[84] M. Rowshannahad, S. Dauzère-Pérès, and B. Cassini. Capacitated Qualification Management in Semiconductor Manufacturing. *Omega*, 54:50–59, 2015. (cited on page 57)

[85] A. J. Ruiz-Torres, J. C. Ho, and J. H. Ablanedo-Rosas. Makespan and Workstation Utilization Minimization in a Flowshop with Operations Flexibility.

Omega-International Journal of Management Science, 39(3):273–282, June 2011. (cited on page 51)

[86] SEMI-E10-0304E. Specification for Definition and Measurement of Equipment Reliability, Availability, and Maintainability (RAM), 2004. (cited on pages 15, 16, 8 and 58)

[87] D. Shin, J. Park, N. Kim, and R. Wysk. A Stochastic Model for the Optimal Batch Size in Multi-Step Operations with Process and Product Variability. *International Journal of Production Research*, 47(14):3919–3936, July 2009. (cited on page 86)

[88] E. A. Silver and H. C. Meal. A Simple Modification of the EOQ for the Case of a Varying Demand Rate. *Production and Inventory Management*, 10(4):52–65, 1969. (cited on page 194)

[89] E. A. Silver, D. F. Pyke, and R. Peterson. *Inventory Management and Production Planning and Scheduling*. John Wiley and Sons, 3 edition, 1998. (cited on pages 17, 49, 150, 155 and 156)

[90] D. Simchi-Levi. *Operations Rules: Delivering Customer Value through Flexible Operations*. The MIT Press, 2013. (cited on pages 232 and 233)

[91] D. Simchi-Levi, P. Kaminsky, and E. Simchi-Levi. *Designing and Managing the Supply Chain*. 2007. (cited on page 149)

[92] Soitec. Soitec Smart Cut Technology - www.soitec.com, Date accessed: May 2014, 2014. (cited on pages 17 and 172)

[93] T. Spengler, H. Puechert, T. Penkuhn, and O. Rentz. Environmental Integrated Production and Recycling Management. *European Journal of Operational Research*, 97(2):308–326, Mar. 1997. (cited on page 166)

[94] H. Stadtler and C. Kilger. *Supply Chain Management and Advanced Planning*. Springer, 3rd edition, 2005. (cited on pages 148, 149 and 160)

[95] B. Tomlin and Y. Wang. Pricing and Operational Recourse in Coproduction Systems. *Management Science*, 54(3):522–537, Mar. 2008. (cited on page 166)

[96] W. W. Trigeiro, L. J. Thomas, and J. O. McClain. Capacitated Lot Sizing with Setup Times. *Management Science*, 35(3):353–366, 1989. (cited on page 165)

[97] T. E. Vollmann, W. Berry, D. C. Whybark, and F. R. Jacobs. *Manufacturing Planning and Control for Supply Chain Management*. McGraw-Hill/Irwin, 5th edition, 2005. (cited on pages 40, 49, 149, 152 and 160)

[98] S. Voss and D. L. Woodruff. *Introduction to Computational Optimization Models for Production Planning in a Supply Chain*. Springer, 2nd edition, 2006. (cited on pages 17 and 161)

[99] H. M. Wagner and T. M. Whitin. Dynamic Version of the Economic Lot Size Model. *Management Science*, 5(1):89–96, 1958. (cited on page 163)

[100] U. Wemmerlöv. Assemble-to-Order Manufacturing: Implications for Materials Management. *Journal of Operations Management*, 4(4):347–368, 1984. (cited on page 152)

[101] S. C. Wheelwright and R. H. Hayes. Link Manufacturing Process and Product Life Cycles. *Harvard Business Review*, 1979. (cited on page 155)

[102] C. Wolosewicz. *Approche Intégrée en Planification et Ordonnancement de la Production*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne, 2008. (cited on pages 17 and 163)

[103] E. Wong. *Active-Set Methods for Quadratic Programming*. PhD thesis, University of California, San Diego, 2011. (cited on page 77)

[104] C. Yugma, S. Dauzère-Pérès, C. Artigues, A. Derreumaux, and O. Sibille. A Batching and Scheduling Algorithm for the Diffusion Area in Semiconductor Manufacturing. *International Journal of Production Research*, 50(8):2118–2132, 2012. (cited on page 53)

[105] Z. Zhang, M. T. Zhang, S. Niu, and L. Zheng. Capacity Planning with Reconfigurable Kits in Semiconductor Test Manufacturing. *International Journal of Production Research*, 44(13):2625–2644, 2006. (cited on page 49)

[106] Z.-H. Zhang, H. Jiang, and X. Pan. A Lagrangian Relaxation Based Approach for the Capacitated Lot Sizing Problem in Closed-Loop Supply Chain. *International Journal of Production Economics*, 140(1):249–255, 2012. (cited on page 166)

École Nationale Supérieure des Mines de Saint-Étienne

NNT : 2015 EMSE 0784

Mehdi ROWSHANNAHAD

QUALIFICATION MANAGEMENT AND CLOSED-LOOP PRODUCTION PLANNING IN SEMICONDUCTOR MANUFACTURING

Speciality: Industrial Engineering

Keywords: Semiconductor manufacturing, qualification management, flexibility, variability, workload balancing, batching, capacity planning, bi-level capacitated lot-sizing, by-production, remanufacturing, closed-loop production planning, optimization.

Abstract: The thesis is composed of two parts. In the first part, we take a binding restriction, called "qualification", present in semiconductor manufacturing as a lever for increasing flexibility and optimizing capacity utilization. A qualification determines the processing authorization of a product on a machine. It acts like an eligibility constraint that allows production volume allocation of a product to a machine. In order to define the best qualification, the production volume should be allocated to parallel non-identical machines which are partially reconfigurable. Capacitated flexibility measures are introduced to define the best qualification which increases machine capacity utilization at most. Batching is another industrial constraint encountered in semiconductor industry. It influences workload balancing and hence qualification management. Several workload balancing algorithms are proposed to find the optimal workload balance of a workcenter. Variability measures are also proposed to evaluate the workload variability of a workcenter. The concept is industrialized and is continuously used at Soitec.

The second part of the thesis deals with closed-loop production planning. Soitec uses Smart-Cut™ Technology to fabricate SOI wafers. Using this technology, one of the two raw materials used to fabricate SOI wafers can be reused several times to make other SOI wafers. However, before coming back to the SOI fabrication line, the used raw material (designated as by-product) must be reworked (remanufactured) in another production line. An original closed-loop production planning model adapted to the supply chain specificities of Soitec is proposed, and is validated using industrial data. Based on this industrial model, a single-item uncapacitated closed-loop lot-sizing model is defined, analyzed, and a dynamic programming algorithm is proposed for a simplified version of the problem.

École Nationale Supérieure des Mines de Saint-Étienne

NNT : 2015 EMSE 0784

Mehdi ROWSHANNAHAD

GESTION DES QUALIFICATIONS ET PLANIFICATION DE PRODUCTION EN BOUCLE FERMÉE DANS LA FABRICATION DES SEMICONDUCTEURS

Spécialité: Génie Industriel

Mots Clés : Industrie des semi-conducteurs, gestion des qualifications, flexibilité, variabilité, équilibrage de charge, batching, planification de la capacité, dimensionnement de lots bi-niveau à capacité finie, produits dérivés, refabrication, planification de production en boucle fermée, optimisation.

Résumé: La thèse est composée de deux parties. La première partie traite de la gestion des qualifications dans l'industrie des semi-conducteurs. La contrainte de qualification définit l'éligibilité d'une machine à processer un produit. La gestion des qualifications nécessite de résoudre un problème d'allocation et d'équilibrage des charges sur des machines parallèles non-identiques et partiellement reconfigurables. Nous avons défini et introduit des indicateurs pour la gestion des qualifications en tenant compte de la capacité des équipements ainsi que la contrainte de regroupements de lots (batching). Plusieurs algorithmes d'équilibrage de charge sont proposés et validés pour le calcul de la charge optimale sur un parc d'équipements. Ce concept est industrialisé au sein de l'entreprise Soitec et fait partie du processus de prise de décision.

La deuxième partie de la thèse porte sur la planification de production en boucle fermée. Le processus de fabrication des plaques SOI à Soitec s'appuie sur la Technologie Smart-Cut™. En utilisant cette technologie, une des deux matières premières peut être réutilisée à plusieurs reprises pour la fabrication des produits finis. Le couplage de deux lignes de production crée un système manufacturier en boucle fermée. Nous avons proposé un modèle de dimensionnement de lots original pour la planification de production de ce système manufacturier, que nous avons validé avec des données industrielles. En se basant sur le problème industriel, un problème mono-produit et sans contrainte de capacité est défini, analysé et résolu pour une version simplifiée du problème.