SMART SAMPLING FOR RISK REDUCTION IN SEMICONDUCTOR MANUFACTURING

Gloria Luz Rodriguez Verjan

Ecole Nationale Supérieure des Mines
SAINT-ETIENNE

NNT : 2014 EMSE 0747

THÈSE

présentée par

Gloria Luz RODRIGUEZ VERJAN

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

Spécialité : Génie Industriel

SMART SAMPLING FOR RISK REDUCTION IN SEMICONDUCTOR MANUFACTURING

soutenue à Gardanne, le 11 Juillet 2014

Membres du jury

Président :	Bernard GRABOT	Professeur, Ecole Nationale d'Ingénieurs de Tarbes, Tarbes
Rapporteurs :	Philippe CASTAGLIOLA	Professeur, Université de Nantes, Nantes
	Lars MÖNCH	Professeur, FernUniversität de Hagen, Hagen
	Nathalie SAUER	Professeur, Université de Lorraine, Metz
Examinateur(s) :	Michel TOLLENAERE	Professeur, Institut Polytechnique de Grenoble, Grenoble
	Claude YUGMA	Chargé de Recherche, ENSM-SE, Gardanne
Directeur(s) de thèse :	Stéphane DAUZERE-PERES	Professeur, ENSM-SE, Gardanne
Encadrant industriel :	Jacques PINATON	Ingénieur, STMicroelectronics, Rousset
Invité(s) éventuel(s):	Philippe CAMPION	Ingénieur, STMicroelectronics, Rousset
	Philippe VIALLETELLE	Ingénieur, STMicroelectronics, Crolles

Spécialités doctorales :	Responsables :
SCIENCES ET GENIE DES MATERIAUX	K. Wolski Directeur de recherche
MECANIQUE ET INGENIERIE	S. Drapier, professeur
GENIE DES PROCEDES	F. Gruy, Maître de recherche
SCIENCES DE LA TERRE	B. Guy, Directeur de recherche
SCIENCES ET GENIE DE L'ENVIRONNEMENT	D. Graillot, Directeur de recherche
MATHEMATIQUES APPLIQUEES	O. Roustant, Maître-assistant
INFORMATIQUE	O. Boissier, Professeur
IMAGE, VISION, SIGNAL	JC. Pinoli, Professeur
GENIE INDUSTRIEL	A. Dolgui, Professeur
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		CMP
AVRIL	Stéphane	PR2	Mécanique et ingénierie	CIS
BATTON-HUBERT	Mireille	PR2	Sciences et génie de l'environnement	FAYOL
BERGER DOUCE	Sandrine	PR2		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
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
BURLAT	Patrick	PR2	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	PR1	Sciences et génie des matériaux	SMS
DESRAYAUD	Christophe	PR2	Mécanique et ingénierie	SMS
DOLGUI	Alexandre	PR0	Génie Industriel	FAYOL
DRAPIER	Sylvain	PR1	Mécanique et ingénierie	SMS
FEILLET	Dominique	PR2	Génie Industriel	CMP
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	CR	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			CMP
MALLIARAS	Georges	PR1	Microélectronique	CMP
MOLIMARD	Jérôme	PR2	Mécanique et ingénierie	CIS
MONTHEILLET	Frank	DR		SMS
MOUTTE	Jacques	CR		SPIN
NIKOLOVSKI	Jean-Pierre			CMP
PIJOLAT	Christophe	PR0	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR1	Génie des Procédés	SPIN
PINOLI	Jean Charles	PR0	Image Vision Signal	CIS
POURCHEZ	Jérémy	CR	Génie des Procédés	CIS
ROBISSON	Bruno			CMP
ROUSSY	Agnès	MA(MDC)		CMP
ROUSTANT	Olivier	MA(MDC)		FAYOL
ROUX	Christian	PR		CIS
STOLARZ	Jacques	CR	Sciences et génie des matériaux	SMS
TRIA	Assia		Microélectronique	CMP
VALDIVIESO	François	MA(MDC)	Sciences et génie des matériaux	SMS
VIRICELLE	Jean Paul	MR(DR2)	Génie des Procédés	SPIN
WOLSKI	Krzystof	DR	Sciences et génie des matériaux	SMS
XIE	Xiaolan	PR1	Génie industriel	CIS
YUGMA	Gallian	CR		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 : 03/12/2013

PR 0	Professeur classe exceptionnelle
PR 1	Professeur 1ère classe
PR 2	Professeur 2ème classe
PU	Professeur des Universités
MA (MDC)	Maître assistant
DR	Directeur de recherche

Ing.	Ingénieur
MCF	Maître de conférences
MR (DR2)	Maître de recherche
CR	Chargé de recherche
EC	Enseignant-chercheur
IGM	Ingénieur général des mines

SMS	Sciences des Matériaux et des Structures
SPIN	Sciences des Processus Industriels et Naturels
FAYOL	Institut Henri Fayol
CMP	Centre de Microélectronique de Provence
CIS	Centre Ingénierie et Santé

Gracias a la vida que me ha dado tanto,
me ha dado el sonido y el abedecedario,
con él las palabras que pienso y declaro,
madre amigo hermano y luz alumbrando...

(Canción de Violeta Parra)

Remerciements

Les travaux présentés dans cette thèse ont été réalisés dans le cadre d'une convention CIFRE (N. 2010/1440) accordée par l'Association Nationale de la Recherche Technique, en collaboration avec STMicroelectronics et l'École Nationale Supérieure des Mines de St-Etienne.

Je tiens à exprimer ma gratitude à mon directeur de thèse, le professeur Stéphane Dauzère-Pérès, pour sa disponibilité et son soutien constante. Sa confiance, son expérience, sa vision et ses indications toujours assertifs ont permis de mener à bien ce projet. Mes remerciements vont également à Jacques Pinaton, mon encadrant industriel, qui a toujours été présent pour répondre à mes questions. Merci à tous ceux qui ont participé dans le groupe d'implémentation industriel du projet W@R de STMicroelectronics. Ce travail au sein du groupe m'a permis de vivre et d'apprendre sur la gestion des changements au sein d'une organisation. Merci à Eric Tartière et Benoit Pennachio pour leur travail et leurs commentaires toujours constructifs; leur aide a été très importante pour l'implémentation des solutions dans le processus de fabrication. Merci à Sylvain Housseman pour son implication dans le projet et sa collaboration dans le développement informatique des modèles.

Mes remerciements vont ensuite aux membres du jury pour avoir accepté d'évaluer cette thèse. Merci au professeur Bernard Grabot pour avoir accepté d'être le président du jury et aux professeurs Philippe Castagliola, Lars Monch et Natalie Sauer pour leur rapport et leurs suggestions qui ont permis d'améliorer le manuscrit. Merci à Claude Yugma avec qui j'ai pu partager différentes idées au cours de ces années de travail.

Il est difficile de remercier toutes les personnes de qui j'ai beaucoup appris et à qui j'aimerai ici exprimer mon amitié. Je remercie sincèrement toutes les personnes du département SFL et tous les amis et collègues de STMicroelectronics. Je remercie également Elizabeth Zehendner pour avoir partagé sa bonne humeur avec moi tout au long de notre cohabitation au bureau. Merci à Rezvan Sadeghi avec qui j'ai pu partager le bureau les derniers mois de mon séjour chez SFL. Merci aussi à Mehdi Rowshannahad qui m'a montré par son exemple que la discipline et la persévérance sont le clé de la réussite d'un projet. Merci Anastasia Frank et les sœurs Céline et Hyacinthe Aubert pour leur amitié, leurs encouragements, les éclats de rire et pour tous les bons moments passés ensemble. Merci à ma chère Hyacinthe pour ses explications de la langue française qui m'ont permis de mieux la comprendre. Merci à Mme Bellecombe pour son accueil et sa générosité. Je tiens à remercier vivement mon frère Carlos Rodriguez pour ses conseils et les moments de partage. Enfin, un immense merci à mes parents, mes frères, ma sœur et toute ma famille pour m'avoir donnée la liberté et le soutien nécessaires à la poursuite de mes études dans les meilleures conditions. Ils ont toujours été là pour me guider et me rappeler que: *"Lo mucho o poco que haga he de hacerlo bien y con amor"*.

Gloria Luz Rodriguez Verjan
Gardanne, 2014.

Contents

List of Tables

List of Figures

Chapter 0

Résumé en français

0.1 Introduction

L'industrie de semi-conducteurs est considérée comme une des industries les plus importantes dans l'économie moderne. Actuellement, la plupart des produits utilisés dans notre vie quotidienne contiennent des circuits intégrés (*ICs Integrated Circuits* en anglais) (e.g. Téléphones, TVs, voitures, objets communicants, ordinateurs, cartes de crédit). Par ailleurs, les fabricants de circuits intégrés proposent des produits de plus en plus performants à des prix de plus en plus compétitifs. Pour cela, les fabricants de semi-conducteurs cherchent constamment des stratégies de contrôle plus efficaces pour pouvoir garantir la qualité du produit final à des coûts raisonnables.

Dans les processus de fabrication de semi-conducteurs, différents types de contrôles existent pour maîtriser les procédés. Dans cette thèse, on s'intéresse à la maîtrise et la réduction du risque sur les équipements de production. On se concentre sur les contrôles de défectivité. L'indicateur de risque utilisé concerne le nombre de produits traités par un équipement depuis la date du dernier produit contrôlé. L'introduction des différentes étapes de contrôle est indispensable pour réduire et maîtriser le risque sur les équipements de production. Par contre, la capacité d'inspection est limitée et le temps de cycle des lots inspectés peut être impacté en générant des conséquences sur le coût de fabrication. Pour éviter cela, différentes stratégies de sélection des lots existent et peuvent être classifiées selon leur capacité à intégrer la dynamique d'une unité de fabrication (*fab*). Dans les stratégies de sélection dynamique, les lots sont contrôlés en temps réel et en optimisant un critère de risque. Ces stratégies sont récentes et sont beaucoup plus efficaces que les stratégies précédentes, mais aussi plus complexes à mettre en œuvre. Dans ce cadre, le système de sélection des lots pour la défectivité à STMicroelectronics Rousset est passé d'une stratégie statique à une stratégique dynamique. Nous avons proposé et validé industriellement différents algorithmes pour identifier les lots à relâcher (à ne pas contrôler) dans les files d'attente des lots en défectivité. Nous avons aussi développé et implémenté un modèle d'optimisation de la capacité pour l'atelier de défectivité, qui permet d'évaluer l'impact de paramètres critiques (e.g. plan de production, positions des opérations de contrôles dans la gamme de fabrication, valeurs des limites de risques) dans la gestion du risque global de l'unité de fabrication.

Ce résumé est organisé comme suit: le contexte général de ce travail ainsi qu'une description des contrôles de defectivité sont présentés dans la section 0.2. L'évolution de la stratégie de sélection des lots pour la défectivité est développée dans la section 0.3. Les algorithmes proposés pour l'identification des lots à relacher sont décrits dans la section 0.3.1. L'outil pour la planification de la capacité dans l'atelier de défectivité est exposé

dans la section 0.3.2. Enfin, nous concluons ce résumé avec la section 0.4 où des résultats industriels, conclusions et perspectives sont présentés.

0.2 Contexte et problèmatique

Les circuits intégrés sont présents dans presque tous les produits de notre vie quotidienne (e.g. Informatique, automobile, communication, électroménagers, objets communicants) et sont à l'origine de nouvelles technologies qui changent notre mode de vie. C'est en cela que l'industrie des semi-conducteurs est un secteur très important de l'économie mondiale.

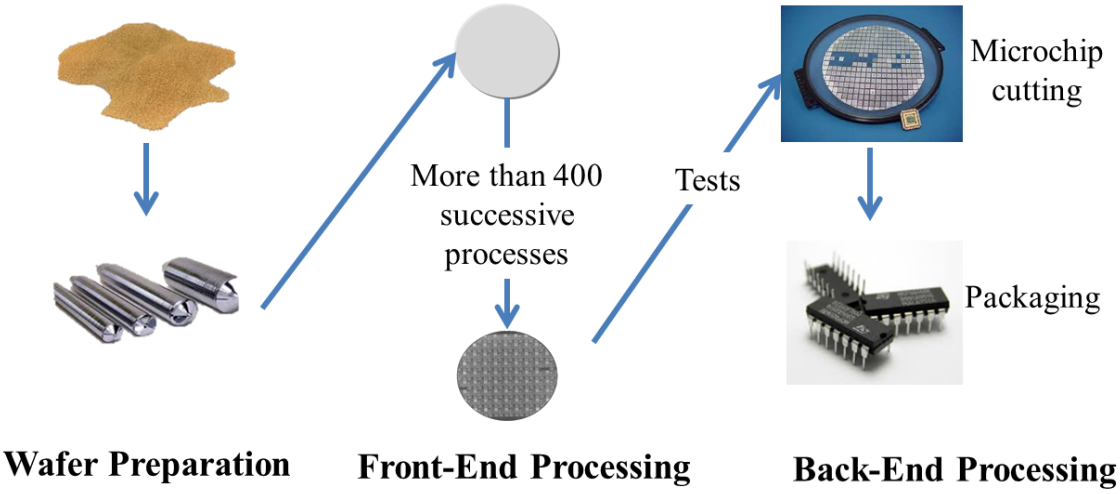


Figure 1: Fabrication des circuits intégrés

La fabrication des circuits intégrés peut être divisée en trois grandes étapes (Figure 1).

- Préparation de plaquettes de silicium (*Wafers* en anglais): Dans cette étape le silicium est extrait du sable. L'ingot de silicium monocristallin est découpé en wafers.
- Traitement Front-End: Les composants électroniques (i.e. Transisteurs, capaciteurs, resisteurs) sont fabriqués et inter-connectés. Pour cela, les wafers sont traités en différentes opérations : dépôt, lithographie, gravure, dopage, isolation et interconnexion qui se répètent plusieurs fois (voir figure 2).
- Traitement Back-End: Les plaquettes sont testées et découpées. Les circuits individuels sont assemblés et mis en boitier afin d'obtenir les circuits intégrés.

Le traitement Front-End est le plus important et complexe des trois étapes de fabrication. Cette complexité est due à : l'échelle nanométrique des opérations de fabrication, les conditions d'extrême propreté nécessaires pour éviter la contamination des produits et les flux ré-entrant des produits dans toutes les opérations de production (les produits les plus complexes peuvent avoir plus de 40 couches). C'est pour cela que différents contrôles sont

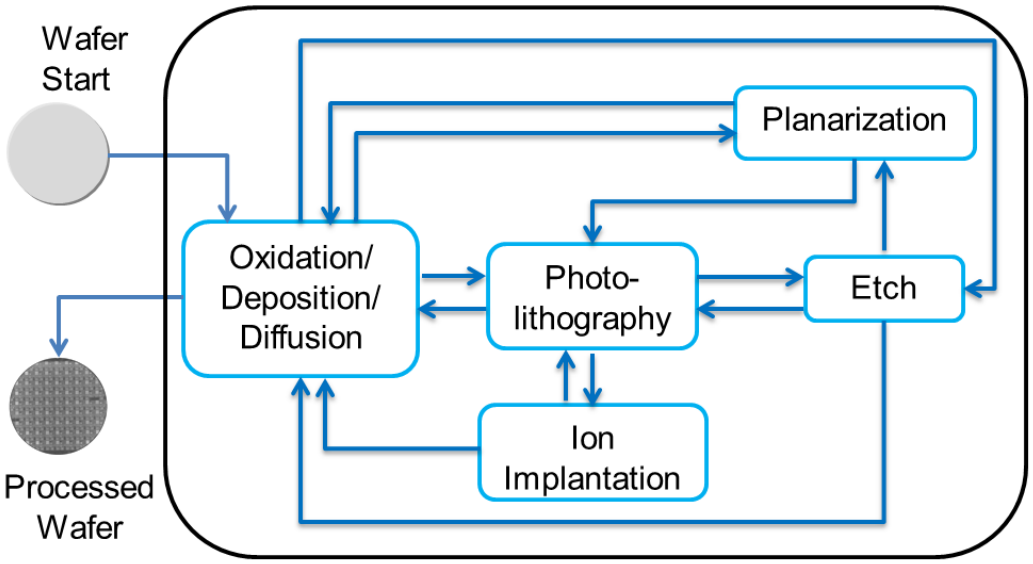


Figure 2: Représentation de traitement Front-End [1]

effectués pour sécuriser la production et pouvoir garantir la qualité des produits finaux. Différents types des contrôles sont utilisés pendant la fabrication, comme par exemple: SPC (*Statistical Process Control* en anglais), FDC (*Fault Detection and Classification* en anglais), R2R (*Run to Run* en anglais), VM (*Virtual Metrology* en anglais) et la defectivité [2]. Cependant, les contrôles peuvent affecter le temps de cycle des produits inspectés, ils sont coûteux et la capacité d'inspection est limitée. Pour ces raisons, 100% des lots de production ne peuvent pas être inspectés et l'échantillonnage (que nous appelons aussi "sélection") est nécessaire.

Les stratégies d'échantillonnage peuvent être classifiées selon leur capacité pour intégrer la dynamique de la fab [3]. Dans les stratégies statiques (*Static Sampling* en anglais) les lots sont sélectionnés au début de leur gamme de fabrication et selon un taux d'échantillonnage fixe par produit [4], [5]. Ces stratégies sont communément utilisées et relativement simples à implémenter, mais la dynamique de la fab (e.g. Panne des équipements, changement de priorités, saturation de l'atelier d'inspection, contrôles spéciales faites sur les équipements) ne peut pas être intégrée. Dans les stratégies adaptatives (*Adaptive Sampling* en anglais) le taux d'échantillonnage est adapté selon la situation. Si le processus est moins à risque, le taux d'échantillonnage peut être réduit. Si le processus dérive, le taux d'échantillonnage augmente pour inspecter plus de lots, et ainsi détecter et corriger le problème [6], [7]. Ces stratégies peuvent intégrer d'une meilleure façon la dynamique de la fab mais la capacité des équipements d'inspection est difficile à gérer. Dans les stratégies dynamiques (*Dynamic Sampling* en anglais) les lots sont sélectionnés en fonction de : l'information qui peut être obtenue en les mesurant, l'état de production et la capacité d'inspection disponible [8], [9]. Ces stratégies sont relativement récentes et plus difficiles à mettre en œuvre. Dans le cadre de cette thèse, la stratégie de sélection des lots pour les contrôles de défectivité est passée d'une stratégie statique à une stratégie dynamique. Les contrôles de défectivité sont particulièrement complexes et une description de ces caractéristiques sont rappelées par la suite.

Contrôles de défectivité

Les contrôles de défectivité sont utilisés pour : surveiller le processus de fabrication afin de détecter les défauts qui peuvent entrainer une perte de rendement, anticiper les dérives des équipements de production, réduire les excursions[1] et enfin améliorer le rendement. Par la suite, on se référera aux contrôles effectués en défectivité comme "les opérations d'inspection" ou "inspections D0" (zero défauts).

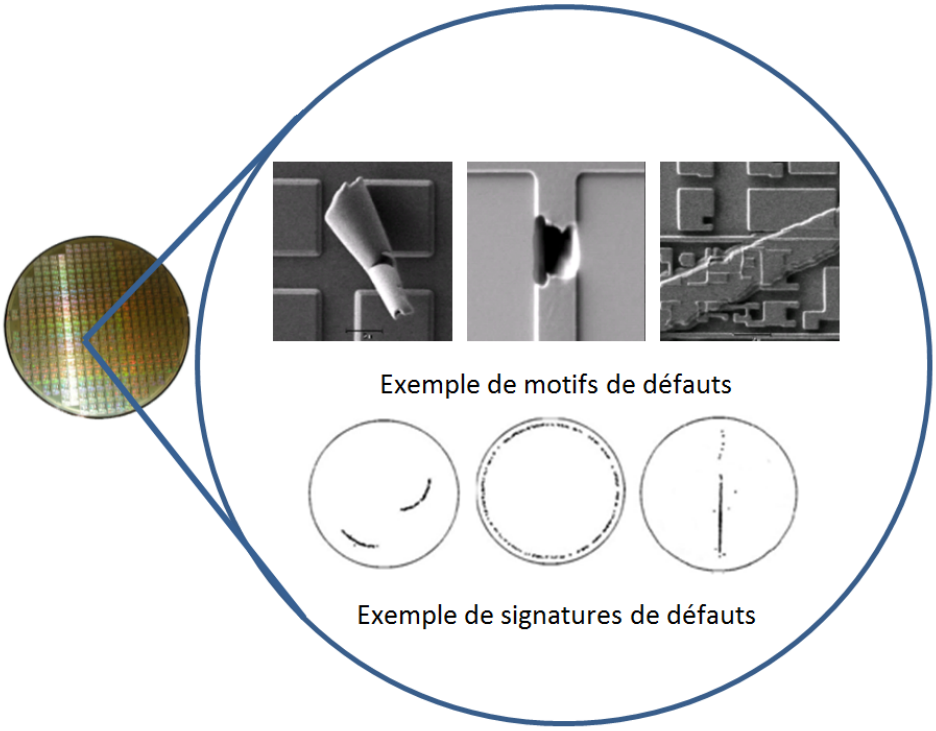


Figure 3: Exemples de motifs et de signatures des défauts

Les défauts détectés avec les inspections de défectivité peuvent avoir des motifs particuliers ou des signatures caractéristiques (voir figure 3). Les technologies les plus utilisées pour détecter les défauts sont le *brightfield*, *dakfield* et *electrobeam*. Dans les technologies *darkfield* et *brightfield* (voir figure 4) un faisceau de lumière est projeté sur le wafer. Le contrôle est effectué par comparaison puce à puce et les défauts sont détectés selon l'angle et l'intensité de la lumière reflétée ou diffractée. Pour les défauts électriques, l'*électrobeam* est utilisé. L'*électrobeam* projette un faisceau d'électrons sur le wafer. En analysant la dispersion des électrons, les défauts électriques peuvent être détectés. Le principe de détection est aussi la comparaison puce à puce.

La position des opérations d'inspection dépend de plusieurs critères. Les principaux sont : la pertinence de l'inspection, la criticité de l'étape de production et l'exposition tolérée pour chaque équipement de fabrication en termes du nombre de wafers à risque (*Wafers at Risk, W@R* en anglais).

[1]Les excursions sont une perte de rendement qui arrivent de façon aléatoire dans le temps et à cause de dérives dans les équipements de production

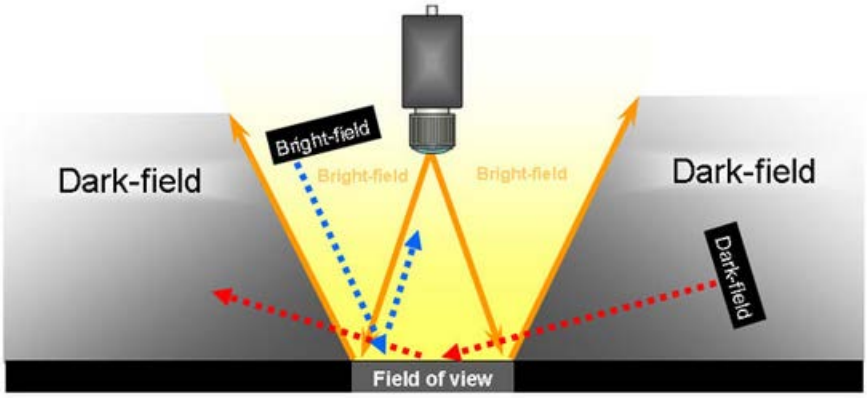


Figure 4: Technologies de détection *Brightfield* et *Darkfield*

- Pertinence: La pertinence d'une opération d'inspection dépend de la capacité à détecter les défauts et le temps nécessaire pour réaliser l'inspection. La capacité à détecter les défauts dépend de la nature de la couche. Par exemple, l'inspection sur une couche opaque ne sera pas pertinente car la technique de détection est basée sur la lumière reflétée ou diffractée. Une inspection sur une couche transparente ne sera pas non plus pertinente car toutes les couches sont observées et beaucoup de faux défauts seront détectés. De plus, il n'est pas possible d'effectuer des contrôles de défectivité à certaines étapes de production. Par exemple, il est très risqué d'introduire une opération d'inspection entre les opérations de production présentant des contraintes de temps, car lorsque la contrainte de temps n'est pas respectée le lot sera rejeté (*scrap* en anglais).
- Criticité: La criticité des opérations est basée sur le ratio des défauts "toueurs" (*Killer ratio* - *KR* en anglais). En général, les opérations d'inspection sont placées entre les opérations de production critiques pour le produit (e.g. Déposition de métal, photolithographie, processus de gravure.)
- Exposition en termes de nombre de wafers à risque: Ce critère concerne le nombre minimal de wafers considérés à risque si un problème se produit. Il est calculé comme le throughput (TH) de l'équipement de fabrication multiplié par le temps de cycle (CT) entre l'opération de production et l'opération d'inspection ($TH \times TC$).

La liste des opérations d'inspection avec leur positions et leur couvertures sont définies dans le "plan de contrôle de défectivité" (*Control plan* en anglais) (voir figure 5.1). Le plan de contrôle est établi par produit car les opérations d'inspection sont effectuées pendant la gamme de fabrication des produits. La couverture indique quelles sont les opérations de production que sont couvertes par une opération d'inspection. Une opération d'inspection de défectivité peut couvrir plusieurs opérations de production et ainsi, contrôler plusieurs équipements de production. La figure 5.1 est une représentation d'une partie de la gamme de fabrication pour un produit de la technologie "A" et un produit de la technologie "B". Dans la portion de la gamme de fabrication du produit de la technologie A, deux

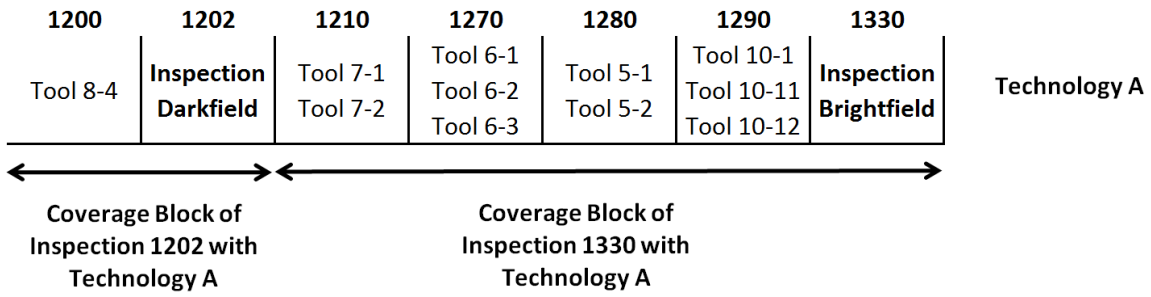
opérations d'inspection peuvent être effectuées (1202 et 1330), si un lot de la technologie "A" est inspecté dans l'opération d'inspection 1202, l'équipement de production tool 8-4 peut être contrôlé. Si un lot est inspecté à l'opération d'inspection 1330, les équipements de production qui ont traité le lot entre les opérations de production 1210 et 1290 sont contrôlés. Par conséquent, avec l'inspection d'un lot, plusieurs équipements de fabrication peuvent être contrôlés. En outre, si un lot d'un produit de la technologie "B" est inspecté à l'opération d'inspection 1330, les équipements de production qui ont traité le lot entre les opérations 1185 et 1280 sont contrôlés. C'est à dire que le bloc de couverture de l'opération d'inspection 1330 est différent entre les produits de la technologie "A" et "B".

De plus, lorsqu'une nouvelle opération d'inspection est mise en place, une recette[2] d'inspection doit être créée. La complexité pour créer une recette d'inspection augmente quand le nombre des paramètres à considérer augmente. Le nombre de paramètres augmente parce que les technologies pour détecter les défauts sont de plus en plus avancées et aussi parce que la conception des produits est plus complexe. Par conséquent, un nombre limité d'opérations d'inspection sont créées [10]. C'est-à-dire que seuls certains produits pourront être inspectés dans les étapes d'inspection où la recette existe. Les produits dits "mesurables" sont donc les produits qui peuvent être inspectés car la recette existe.

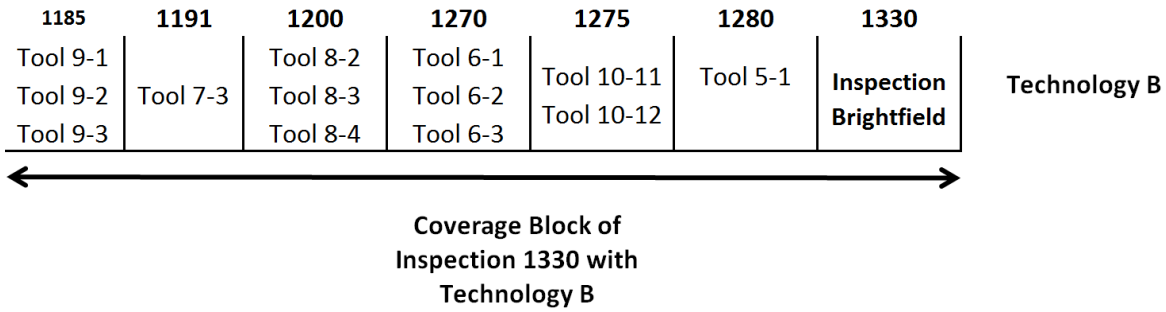
Problématique

Lorsqu'un équipement de production dérive, tous les wafers qui sont traités sur cet équipement peuvent être affectés, et, par conséquent, générer des excursions. L'impact des excursions dépend de la sévérité du problème et du nombre de wafers affectés [11], il est donc nécessaire de détecter le plus rapidement possible les problèmes. Pour cela, des inspections régulières sont définies par produit et les lots à inspecter sont sélectionnés selon une stratégie d'échantillonnage. Quand cette thèse a commencé, seulement une stratégie statique pour la sélection des lots en défectivité était utilisée. Avec cette stratégie, un pourcentage de lots était sélectionné selon un taux d'échantillonnage défini comme "$1/N$", "N" étant le nombre des lots à risque. Figure 1.8 est une représentation de la problématique des stratégies statiques. Dans cet exemple, un sur deux lots sont sélectionnés pour être inspectés en défectivité. Dans la figure 1.8, les lots L2, L4 et L6 sont sélectionnés pour être inspectés et les lots L1, L3 et L5 ne seront pas inspectés. Si les lots sont fabriqués selon la disponibilité des équipements de fabrication et que le lot L1 est traité dans la machine 1, le lot L2 sera traité dans la machine 2, le lot L3 sera assigné à la machine 1 et le lot L2 sera traité par la machine 2 et ainsi de suite. Il est donc possible que tous les lots qui ont été sélectionnés pour l'inspection soient passés seulement dans la machine 2. Il résulte de cette situation que la machine 2 est sur-contrôlée et que la machine 1 n'est pas suffisamment contrôlée. Cet exemple nous permet de constater quelques désavantages des stratégies statiques pour contrôler tous les équipements de production. Par ailleurs, la complexité augmente quand plusieurs produits sont pris en considération car le plan de contrôle et les taux d'échantillonnage sont différents entre les produits. Pour ces raisons, la stratégie d'échantillonnage a évolué vers une stratégie dynamique. Dans le cadre de cette

[2] La recette d'une opération d'inspection contient les paramètres nécessaires pour pouvoir effectuer l'inspection, comme par exemple : la configuration des signaux et de paramètres optiques.



(a) Gamme de fabrication pour un produit de la technologie "A"



(b) Gamme de fabrication pour un produit de la technologie "B"

Figure 5: Example de gammes de fabrication pour les produits des technologies "A" et "B"

thèse, des algorithmes d'identification des lots qui peuvent être relâchés de l'inspection ont été proposés et implémentés industriellement. De plus, un modèle pour planifier la capacité dans l'atelier de défectivité a été proposé et une étude concernant la conception des plans de contrôle pour la défectivité a été effectuée.

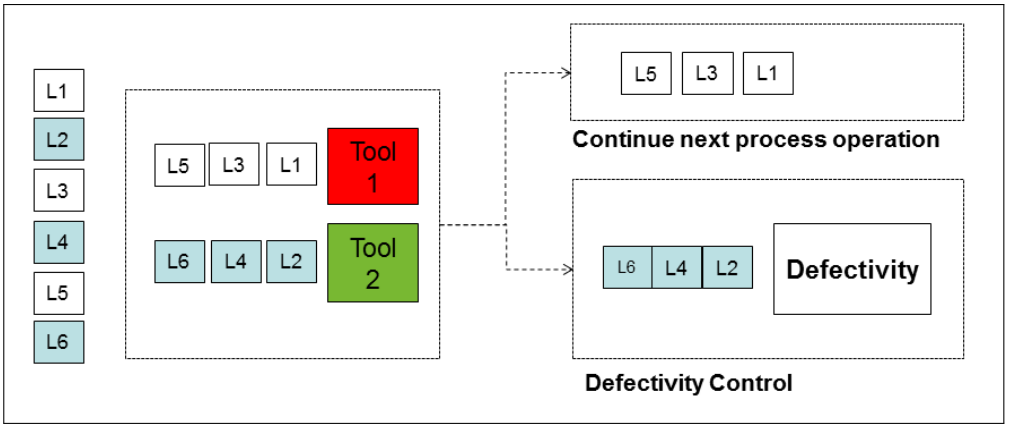


Figure 6: Désavantages des stratégies statiques [12]

0.3 Évolution du système de sélection des lots et solutions proposées

Dans la section précédente, on a décrit brièvement la problématique concernant l'utilisation des stratégies statiques pour la gestion du nombre de wafers à risque sur les équipements de fabrication. Cette section correspond au chapitre 3 de la thèse dans lequel le changement du système d'échantillonnage est abordé.

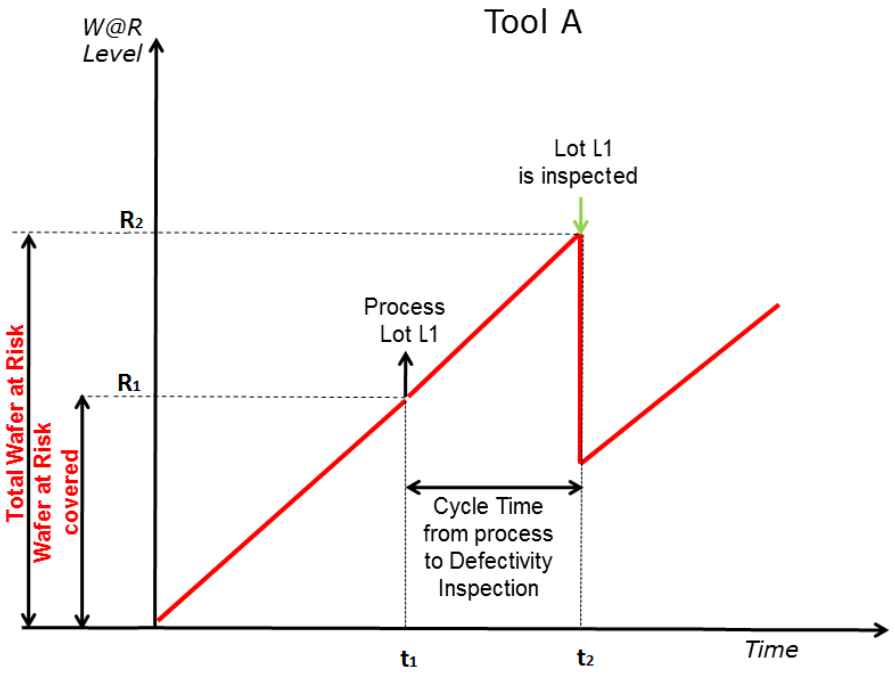


Figure 7: Compteur W@R

La figure 7 présente de quelle manière le nombre de wafers à risque sur les équipements de production est comptabilisé. L'axe y représente le nombre de wafers à risque et l'axe x représente le temps. Chaque équipement de production de l'unité de fabrication (fab) a un compteur du nombre de wafers à risque (W@R, *Wafers at Risk* en anglais). Le W@R augmente lorsque l'équipement de production traite un lot et diminue lorsque les résultats de l'inspection d'un lot sont obtenus. Dans la figure 7 quand le lot L1 est traité à l'instant t_1 le nombre de wafers couverts par le lot L_1 est équivalent à R_1. Ensuite, le lot L_1 passe par les autres opérations de fabrication jusqu'à la prochaine opération d'inspection. Pendant ce temps, l'équipement de production continue à traiter des lots et ainsi le W@R continue à augmenter. Lorsque les résultats de l'inspection du lot L_1 sont obtenus à l'instant t_2, le W@R peut être réduit. La nouvelle valeur de W@R est calculée comme la valeur actuelle de W@R (R_2) moins la valeur de W@R que le lot L_1 réduit (R_1).

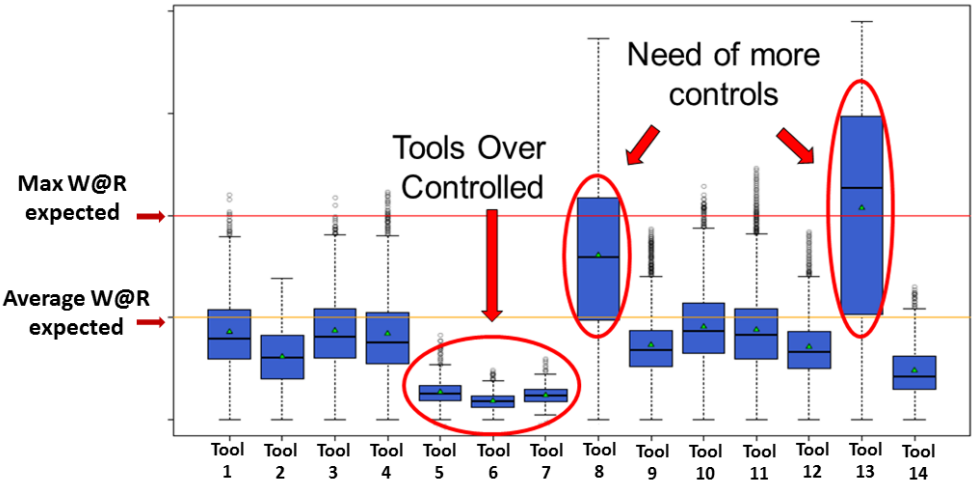


Figure 8: Rapport de W@R

La figure 8 présente les valeurs de W@R sur les équipements de production d'un même atelier de fabrication du site de Rousset de STMicroelectronics. La stratégie de sélection des lots est seulement statique. On observe que les équipements 5, 6 et 7 sont sur-contrôlés et les équipements 8 et 13 ne sont pas suffisamment contrôlés. Par conséquent, un contrôle optimal en termes de W@R sur les équipements de production ne peut pas être garanti quand la stratégie de sélection des lots pour l'inspection est statique [12].

La stratégie de sélection des lots est devenue dynamique pendant la thèse. Les lots sont sélectionnés selon l'information qui peut être obtenue en les mesurant (i.e. réduction du W@R). La figure 9 montre globalement les différentes applications développées. L'objectif du *"Dispatching for sampling"* est de s'assurer que chaque équipement de production traite au moins un lot mesurable (i.e. Un lot dont la recette d'inspection existe en défectivité). L'objectif du *"Sampling and skipping before D0"* est de sélectionner les lots avant d'entrer à l'atelier de défectivité. L'objectif du *"Skipping at D0"* est d'identifier les lots qui peuvent sortir de la file d'attente de défectivité s'ils ont des informations redondantes. Dans la section suivante une description des algorithmes développés pour l'application de *"skipping at D0"* est présentée.

0.3.1 Algorithmes de sélection des lots

Dans cette section, la méthodologie utilisée pour identifier les lots qui peuvent sortir de la file d'attente de défectivité est présentée. Cette section correspond au résumé du chapitre 4 de la thèse. Les lots sélectionnés pour la réduction du W@R portent une information de réduction de risque sur les équipements de production où ils ont été traités. Cependant, l'état de production change et le W@R des équipements évolue dans le temps. Par conséquent, les lots qui étaient sélectionnés pour l'inspection ne sont plus intéressants au bout d'un certain temps. Un lot peut ne plus être intéressant pour différentes raisons : lorsqu'il y a des nouveaux lots qui arrivent avec plus d'information, lorsqu'il y a eu des contrôles spéciaux sur les équipements de production (e.g. QTs, *Quality Tasks* en anglais), ou encore, lorsque le lot a attendu longtemps devant les machines d'inspection. Ainsi, il est

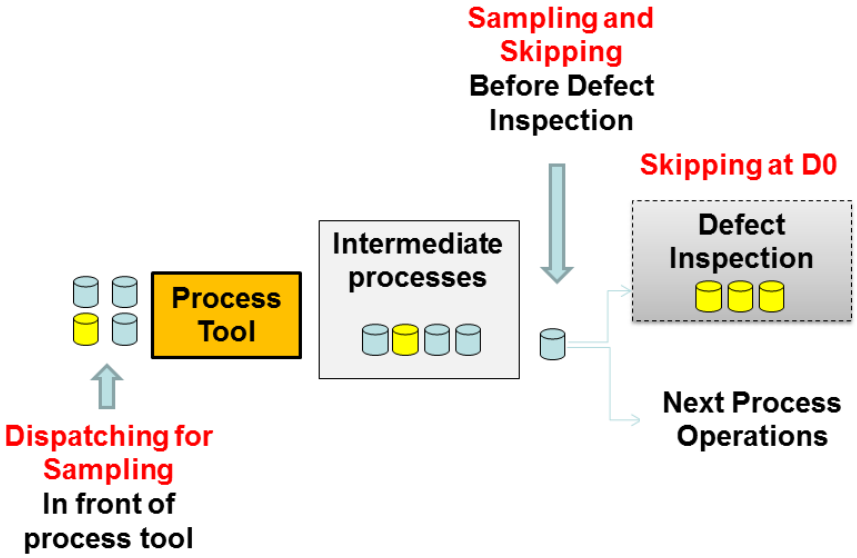


Figure 9: Description générale des applications utilisées pour la sélection dynamique des lots à STMicroelectronics, Rousset

important d'identifier les lots qui sont en défectivité et qui peuvent sortir de la file d'attente sans dégrader le risque de l'unité de fabrication (fab). Pour cela, différents algorithmes ont été développés. Afin d'identifier l'importance d'inspecter un lot ou un groupe de lots, la dégradation du risque global est évaluée chaque fois qu'un groupe de lots n'est pas inspecté. L'évaluation des risques est faite avec l'indicateur GSI (*Global Sampling Indicator* en anglais) proposé par Dauzère-Pérès et al. [9]. La décision est prise selon la variation de cet indicateur quand un groupe de lots (S) n'est pas mesuré. Cette variation est évaluée avec un indicateur d'augmentation de risque RI (*Risk increase* en anglais). Ce dernier est inspiré du LSI (*Lot Scheduling Indicator* en anglais) proposé par Nduhura Munga[12]. Par la suite, une description des indicateurs est présentée et la notation de [9] rappelée:

- T: Nombre des équipements de production considérés.
- IL_t: Limite maximale de W@R de l'équipement de production t.
- RV_t: Valeur actuelle du risque sur l'équipement de production t. Dans notre cas c'est le W@R sur l'équipement t.
- $G_{t,l}$: Réduction du risque sur l'équipement t si le lot l est inspecté.
- $NRV_{t,l}$: Nouvelle valeur de risque de l'équipement t si le lot l est inspecté, i.e. $NRV_{t,l} = RV_t - G_{t,l}$. Dans notre cas, cette valeur correspond à la nouvelle valeur de W@R (NW@R).
- $NRV_t(S)$: Nouvelle valeur du risque de l'équipement t si les lots dans l'ensemble S sont inspectés. Le $NRV_t(S)$ est calculé comme suit:

$$NRV_t(S) = \min_{l \in S} NRV_{t,l}$$

- α: Paramètre utilisé pour donner plus ou moins de poids aux lots qui permettent de s'éloigner le plus de la valeur maximale de risque (IL).

Le GSI est calculé comme suit:

$$GSI(S) = \sum_{t=1}^{T} \left(\frac{NRV_t(S)}{IL_t} \right)^{\alpha} \tag{1}$$

Avec l'indicateur RI l'impact d'un lot ou d'un ensemble de lots sur le risque global est évalué. Supposons qu'il y ait 3 lots dans la file d'attente (L_1, L_2, L_3). Si deux lots sont retirés simultanément, les possibles combinaisons listées ci-dessous sont évaluées:

1. $RI(\{L_1, L_2\}) = GSI(S \setminus S\{L_1, L_2\}) - GSI(S) = GSI(L_3) - GSI(\{L_1, L_2, L_3\})$.
2. $RI(\{L_2, L_3\}) = GSI(S \setminus S\{L_2, L_3\}) - GSI(S) = GSI(L_1) - GSI(\{L_1, L_2, L_3\})$.
3. $RI(\{L_1, L_3\}) = GSI(S \setminus S\{L_1, L_3\}) - GSI(S) = GSI(L_2) - GSI(\{L_1, L_2, L_3\})$.

Le seuil T_{Metro} est utilisé pour déterminer si un lot L ou un ensemble des lots S peuvent être retirés de la file d'attente $S_{Initial}$.

$$GSI(S_{Initial} \setminus S) - GSI(S_{Initial}) \leqslant T_{Metro} \tag{2}$$

Différents algorithmes ont été développés et implémentés industriellement. Une brève description est présentée par la suite (pour plus de détails voir le chapitre 4):

- **Algorithme 1 (Identification des lots skippables)**: Pour chaque lot dans la liste initiale $S_{Initial}$, le RI est calculé et tous les lots dont $RI(L) \leqslant T_{Metro}$ avec une itération sont retirés. Cet algorithme est utilisé pour le pré-traitement des données mais pas pour prendre la décision finale. Par exemple, s'il y a 2 lots dans la liste initiale (L_1, L_2), le RI de chaque lot sera calculé comme suit:

  1. $RI(L_1) = GSI(S_{initial} \setminus \{L_1\}) - GSI(S_{initial}) = GSI(\{L_2\}) - GSI(\{L_1, L_2\})$
  2. $RI(L_2) = GSI(S_{initial} \setminus \{L_2\}) - GSI(S_{initial}) = GSI(\{L_1\}) - GSI(\{L_1, L_2\})$

  Lorsque le RI du lot L_1 est calculé, le lot L_2 est considéré dans la file d'attente et inversement pour le calcul de RI du lot L_2. Si les deux lots ont des informations redondantes (i.e. ils réduisent le risque des mêmes équipements) leur RI peut être inférieur au T_{Metro} et, par conséquent, être retirés simultanément de la file d'attente. Pour cette raison, l'algorithme 1 est utilisé pour le pré-traitement mais pas pour prendre la décision finale.

- **Algorithme 2 (Évaluation locale)**: Pour chaque lot dans la liste initiale $S_{Initial}$, le RI est calculé, et lorsque $RI(L) \leqslant T_{Metro}$ le lot L est retiré immédiatement. Par conséquent, le RI des lots restant dans la file d'attente prendra en compte tous les lots retirés antérieurement.

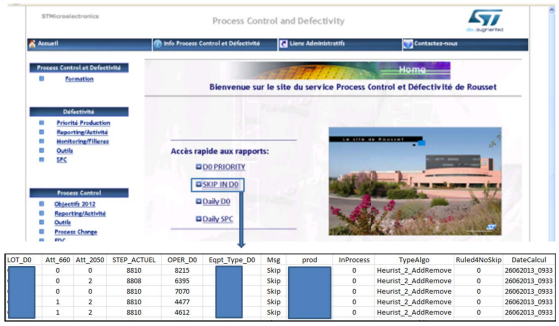
- **Algorithme 3 (Greedy)**: Pour chaque lot de la liste initiale $S_{Initial}$ le RI est calculé. Si le $min_{L \in S}(RI)$ est inférieur à T_{Metro} le lot est retiré de la liste. Ensuite, le RI des lots restants est recalculé. Cette procédure continue jusqu'à ce que le lot avec le plus petit RI ne peut plus être retiré.

- **Algorithme 4 (Add/Remouve)**: Cet algorithme est basé sur l'algorithme 3 (Greedy), mais les solutions sont améliorées grâce à une procédure de recherche locale d'ajout et suppression.
- **Algorithme 5 (Branch & Bound)**: Le RI des lots dans la liste $S_{Initial}$ sont calculés et les lots sont ordonnés selon RI croissant. Une méthode de recherche arborescente est utilisée pour trouver le meilleur ensemble de lots à retirer de la liste initiale. Les conditions d'évaluation considèrent le nombre de lots et le RI résultant.
- **Méthode d'urgence**: Les 5 premiers algorithmes considèrent la valeur de T_{Metro} comme une donnée d'entrée. Dans cette version d'urgence, on considère que le nombre des lots à retirer est connu en avance et le résultat obtenu est l'ensemble des lots avec le plus petit RI. Cette version peut être utilisée quand il y a des événements inattendus comme un changement dans la capacité d'inspection (e.g. Arrêts ou pannes des équipements d'inspection.).

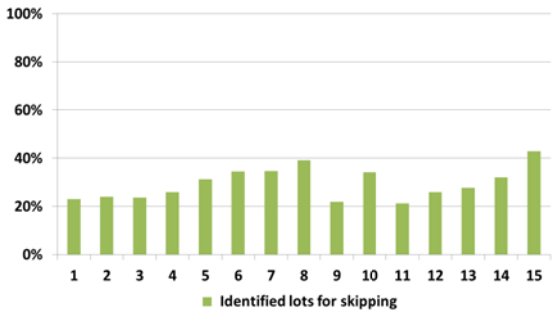
Les algorithmes ont été développés avec le logiciel R [13] et ils ont été testés avec des instances industrielles. Les résultats obtenus avec les différents algorithmes sont présentés et analysés en détail dans le chapitre 4 de la thèse. Les résultats ont montré qu'avec l'algorithme 2 le T_{Metro} est respecté, mais les solutions dépendent de l'ordre d'évaluation des lots. Seulement dans 42% de cas (60 scénarios) la meilleure solution est trouvée. L'algorithme 3 trouve la meilleure solution dans 82% des cas. Les résultats ne dépendent pas de l'ordre d'évaluation des lots en raison du fait qu'à chaque itération le lot avec le plus petit RI est retiré. Ensuite, le RI des lots restants est recalculé. Avec l'algorithme 4, les décisions sont améliorées grâce à la procédure d'ajout et suppression. La meilleure solution est trouvée dans 92% des cas analysés avec cet algorithme. L'algorithme 5 trouve les solutions optimales, avec le même nombre de lots à retirer le RI résultant est plus petit, ou encore, des solutions avec plus de lots identifiés pour être retirés et en respectant la contrainte de T_{Metro} sont trouvés. Par contre, le temps de calcul peut augmenter considérablement quand le nombre des lots à évaluer augmente. Par conséquent, l'algorithme 4 reste un bon compromis entre qualité des solutions et temps de calcul.

Implémentation Industrielle

Lorsque la stratégie de sélection des lots était seulement statique (i.e. Start Sampling) le nombre des lots arrivant en défectivité ne pouvait pas être contrôlé et il était très difficile d'identifier les lots qui pouvaient être retirés de la file d'attente. Avec l'implémentation des compteurs de risque W@R sur les équipements de production et l'évaluation des lots concernant la variation de risque global, l'identification des lots avec information redondant a été possible. La figure 4.5b présente les résultats obtenus lorsque l'application a été implémentée et utilisée industriellement. Entre 20% et 40% des lots en attente en défectivité ont pu être retirés (remplacés) pour inspecter des lots avec plus d'information. Ces gains importants ont motivé l'implémentation d'une application qui intègre les décisions de sampling (lot à échantillonner) et skipping (lots à relâcher) pour optimiser la sélection des lots avant qu'ils n'arrivent en défectivité [14].



(a) Rapport de skip manuelle



(b) Résultats skip automatique

Figure 10: Résultats de l'implémentation industrielle

0.3.2 Planification de la capacité pour la défectivité

Lorsque la stratégie de sélection des lots pour la defectivité a changé, un nouveau modèle de planification de la capacité a été nécessaire. Cette section correspond aux chapitres 5 et 6 de la thèse. Le modèle proposé prend en compte les facteurs qui affectent directement la gestion du W@R des équipements de production. Ces facteurs sont décrits ci-dessous:

- **Gamme de fabrication** : La gamme de fabrication contient la séquence des opérations de production nécessaires pour obtenir le produit final.
- **Qualification des équipements** : Chaque opération de production et d'inspection est associée à une recette. Lorsque la recette est installée et réglée pour l'équipement de production ou d'inspection, l'équipement est considéré comme qualifié. Quand plusieurs équipements de production sont qualifiés pour effectuer la même opération de production, la charge peut être mieux repartie entre les équipements. Par contre, effectuer les qualifications peut être long, coûteux et différentes restrictions techniques peuvent exister. Pour cela, seulement un nombre limité de qualifications sont effectuées. La qualification des équipements de production définit quels sont les produits qui peuvent être traités sur chaque équipement de production, il est donc important de les prendre en compte, ainsi que la qualification des équipements d'inspection.
- **Plan de contrôle de défectivité**: Avec le plan de contrôle, la position et la couverture des opérations d'inspection sont définies.
- **Mix et volume de production**: La gamme de fabrication et le plan de contrôle de défectivité sont définis par produit. Il est donc nécessaire de les considérer pour planifier la capacité de l'atelier de défectivité.
- **Limites de W@R**: Les limites de W@R définissent le nombre maximal de wafers à risque (*IL, Inhibit limit* en anglais) sur chaque équipement de production ainsi que la limite à partir de laquelle l'inspection d'un lot devrait s'effectuer (*WL, Warning limit* en anglais) pour éviter d'atteindre la limite maximale de W@R.

Pour déterminer la capacité nécessaire en défectivité, la réduction de W@R obtenue avec différents contrôles sur les équipements de production a été modélisée. La figure 11 présente l'évolution de W@R dans un équipement de production, ainsi que la réduction

qui peut être obtenu avec un lot de production et une tâche de qualité [3](*QT, Quality task* en anglais).

Lorsque le contrôle de l'équipement de production s'effectue avec l'inspection d'un lot, il y a une période de temps que doit être considéré avant d'obtenir les résultats de l'inspection. Cette période de temps correspond au temps de cycle (CT) entre l'opération de production et l'opération d'inspection. Pendant ce temps, l'équipement de production continue à produire selon une cadence déterminée par le throughput (TH). Par conséquent, lorsque les résultats de l'inspection d'un lot sont obtenus, le W@R de l'équipement est réduit jusqu'au nombre minimal de wafers à risque ($CT \times TH$). Ce nombre minimal de wafers à risque correspond à l'exposition de l'équipement de production par rapport à une opération d'inspection. Lorsque le niveau de W@R sur l'équipement de production atteint la limite IL, l'équipement est arrêté et une QT doit être effectuée. Par conséquent, le W@R de l'équipement peut être réduit à zéro quand le contrôle est effectué avec une QT. Bien que la réduction de W@R soit importante avec les QTs, ce type de contrôle est très coûteux et affecte la disponibilité de l'équipement de production. Pour cela, une grande attention est accordée à la réduction de l'utilisation des QTs. Trois modèles ont été proposés avec différents niveaux de détail. Le modèle présenté dans ce résumé est le modèle 3 du chapitre 5 de la thèse. Dans ce modèle, plus de détails industriels sont considérés.

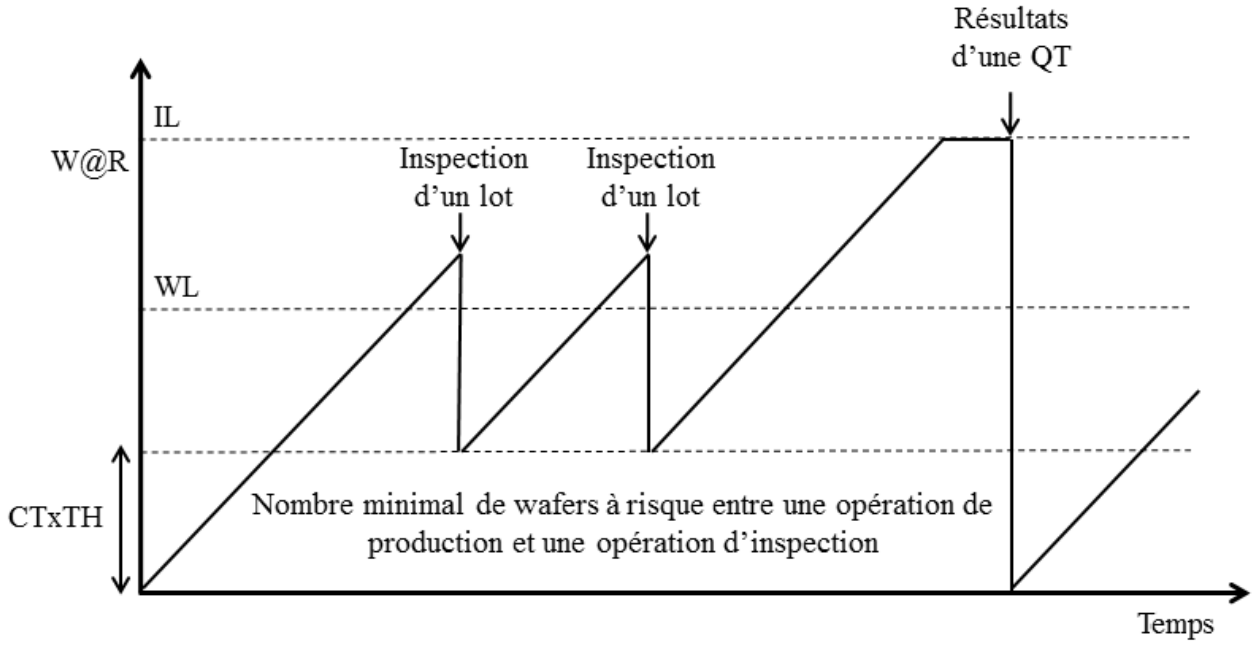


Figure 11: Réduction du W@R avec de lots de production et de QT (Tâche de qualité)

[3]Les tâches de qualité (QT) sont des contrôles spéciaux effectués avec wafers de test (*NPW, Non product wafers* en anglais) et sont utilisés pour plusieurs objectifs, comme par exemple: la vérification de l'état de l'équipement, la qualification des équipements de production, la vérification de l'état du procédé de fabrication, etc.

Paramètres et notation

I : Ensemble des produits indexés avec i,

P : Ensemble des opérations de production indexés avec p,

C : Ensemble des opérations d'inspection indexés avec c,

K : Ensemble des types d'équipements d'inspection indexés avec k,

T : Ensemble des équipements de production indexés avec t.

V^i : Volume total de production du produit i,

IL_t : Limite maximale de W@R sur l'équipement de production t,

h_p^i : Gammes de fabrication,

$$= \begin{cases} 1 \text{ si le produit } i \text{ est traité dans l'opération de production } p, \\ 0 \text{ sinon.} \end{cases}$$

$hh_{p,t}^i$: Qualification des équipements de production,

$$= \begin{cases} 1 \text{ si l'équipement de production } t \text{ est qualifié pour traiter le produit } i \text{ dans} \\ \text{l'opération de production } p, \\ 0 \text{ sinon.} \end{cases}$$

$b_{c,k}^i$: Qualification des équipements d'inspection k dans l'opération d'inspection c,

$$= \begin{cases} 1 \text{ si l'opération d'inspection } c \text{ du produit } i \text{ est qualifié sur le type d'équipement} \\ \text{d'inspection } k, \\ 0 \text{ sinon.} \end{cases}$$

R_k : Capacité réservée pour les inspections $W@R$ effectuées dans le type d'équipement d'inspection k.

$WD_{c,p}^i$: Nombre de wafers à risque entre l'opération de production p et l'opération d'inspection c du produit i. Cette valeur est calculée comme $TC \times TH$ (voir figure 11),

$e_{c,p}^i$: Bloc de couverture de l'opération d'inspection c du produit i,

$$= \begin{cases} 1 \text{ si l'opération d'inspection } c \text{ du produit } i \text{ couvre les équipements de production} \\ \text{qualifiés dans l'opération de production } p, \\ 0 \text{ sinon.} \end{cases}$$

ob_c^i : Opérations d'inspection obligatoires pour le produit i,

$$= \begin{cases} 1 \text{ si l'opération d'inspection } c \text{ est obligatoire pour le produit } i \\ 0 \text{ sinon} \end{cases}$$

SS^i : Taux d'échantillonnage de la stratégie de sélection statique pour le produit i,

QTP : Pénalité associée à l'utilisation des QTs additionnels,

$DefP_k$: Facteur pour exprimer les restrictions d'utilisation d'un type d'équipement d'inspection k pour les contrôles W@R,

α : Pourcentage des inspections obligatoires qui réduisent aussi le W@R sur les équipements de production,

γ^i : Pourcentage du volume du produit i considéré comme mesurable,

$TimeQT$: Temps d'inspection d'une QT,

$TimeD0_c^i$: Temps d'inspection du produit i dans l'opération d'inspection c,

$CapaMax_k$: Capacité maximale exprimée en termes de temps pour le type d'équipement d'inspection k,

$ScQt_t$: Nombre des QTs planifiés en avance sur l'équipement de production t,

ρ_t : Réduction de W@R considéré pour les QTs planifiés en avance sur l'équipement de production t,

Variables de décision

$X_{p,t}^i$: Volume de production du produit i traité dans l'équipement de production t dans l'opération de production p,

$Y_{c,t}^i$: Nombre des inspections du produit i effectués dans l'opération d'inspection c qui couvre l'équipement de production t,

Z_c^i : Nombre total des inspections effectués dans l'opération d'inspection c du produit i,

A_k : Capacité additionnelle nécessaire pour le type d'équipement d'inspection k.

QT_t : Nombre de QTs additionnelles effectués pour contrôler l'équipement de production t.

Le modèle considère la capacité allouée pour les deux stratégies de sélection des lots: statique et dynamique. Pour la stratégie dynamique, une capacité réservée en avance est allouée. Lorsque la capacité réservé en avance n'est pas suffisante pour garantir les limites de W@R, une capacité additionnelle est assignée. De plus, les QTs sont utilisés si les limites de W@R ne peuvent pas être garanties avec des lots de production. L'objectif est donc de minimiser la capacité additionnelle et le nombre de QTs allouées.

$$\min \sum_k DefP_k \cdot A_k + QTP \cdot TimeQT \cdot \sum_t QT_t$$

s.t.

$$\sum_{t} X^i_{p,t} \cdot hh^i_{p,t} = h^i_p \cdot V^i \qquad \forall i,p \tag{3}$$

$$\sum_{i,c} Y^i_{c,t} \cdot (\sum_{p} e^i_{c,p} \cdot hh^i_{p,t} \cdot IL_t - \sum_{p} e^i_{c,p} \cdot hh^i_{p,t} \cdot WD^i_{c,p}) +$$
$$QT_t \cdot IL_t + \rho_t \cdot ScQt_t \cdot IL_t \cdot \geqslant \sum_{i,p} X^i_{p,t} \qquad \forall t \tag{4}$$

$$Y^i_{c,t} \leqslant \sum_{p} e^i_{c,p} \cdot hh^i_{p,t} \cdot X^i_{p,t} \cdot \boldsymbol{\gamma_i} \qquad \forall i,c,t \tag{5}$$

$$Z^i_c + \alpha \cdot V^i \cdot SS^i \cdot ob^i_c \geqslant \sum_{t} Y^i_{c,t} \cdot e^i_{c,p} \cdot hh^i_{p,t} \qquad \forall i,c,p \tag{6}$$

$$A_k + R_k \geqslant \sum_{i,c} Z^i_c \cdot b^i_{c,k} \cdot TimeD0^i_c \qquad \forall k \tag{7}$$

$$A_k + R_k + \sum_{c,i} V^i \cdot SS^i \cdot ob^i_c \cdot b^i_{c,k} \cdot TimeD0^i_c \leqslant CapaMax_k \qquad \forall k \tag{8}$$

$$X^i_{p,t} \geqslant 0 \qquad \forall i,p,t \tag{9}$$

$$Y^i_{c,t} \geqslant 0 \qquad \forall i,c,t \tag{10}$$

$$Z^i_c \geqslant 0 \qquad \forall i,c \tag{11}$$

$$A_k \geqslant 0 \qquad \forall k \tag{12}$$

Les contraintes 5.20 définissent comment le volume de production du produit i est reparti entre les équipements de production t qualifiés pour effectuer l'opération de production p. Les contraintes 5.21 expriment comment le nombre d'inspections $Y^i_{c,t}$ effectuées pour contrôler l'équipement de production t peuvent réduire le risque en considèrant l'exposition de l'équipement de production entre l'opération de production p et l'opération d'inspection c ($WD^i_{c,p}$). Ces contraintes considèrent aussi le nombre de QTs utilisés pour contrôler l'équipement de production t lorsque les inspections avec les lots de production ne permettent pas de garantir les limites W@R. Les contraintes 5.22 assurent que le nombre d'inspections effectuées dans l'équipement de production t ne soit pas supérieur au nombre de lots mesurables. Les contraintes 5.23 définissent le bloc de couverture de l'opération d'inspection c. Le nombre total d'inspections allouées pour la stratégie dynamique de sélection des lots n'est pas connu à l'avance et il est définit à travers l'optimisation. Les contraintes 5.23 sont la linéarisation de l'expression:

$$Z^c_i = \max_{i,c,p}(\sum_{t} Y^i_{c,t} \cdot e^i_{c,p} \cdot hh^i_{p,t} - \alpha \cdot V^i \cdot SS^i \cdot ob^i_c)$$

Les contraintes 5.24 définissent le nombre total d'inspections allouées à chaque type d'équipement d'inspection k. Ces contraintes considèrent la capacité nécessaire pour l'inspection des lots sélectionnés avec les stratégies dynamique et statique. Les contraintes 5.25 assurent que la capacité allouée soit inférieure à la capacité maximale par type d'équipement d'inspection.

Les détails des expérimentations et les résultats sont présentés dans le chapitre 6 de la thèse. Ce modèle permet d'anticiper l'impact des différents scénarios sur la maîtrise des risques des équipements de production. C'est un outil d'aide à la décision qui répond à différentes questions aux niveaux tactique et stratégique. Au niveau tactique, il permet de déterminer si les objectifs en termes de réduction des limites W@R peuvent être atteint avec la capacité actuelle de défectivité. De plus, il permet d'évaluer l'impact que les changements de mix et volume de production peuvent avoir sur la maîtrise du risque des équipements de production. Les résultats montrent que lorsque les QTs sont utilisées, même s'il y a de la capacité disponible en défectivité, une revue du plan de contrôle est souhaitable, car la position des opérations de contrôle ne permet pas de garantir entièrement les limites de risque. De plus, les changements de mix et volume de production peuvent entrainer des situations dont les limites de W@R ne sont plus garanties, car les plans de contrôle de défectivité sont définis par produit. Au niveau stratégique, les questions concernant l'augmentation de la capacité peuvent être aussi analysées.

Implémentation industrielle

La figure 12 présente un schéma du modèle de capacité. L'information concernant les gammes de fabrication, les plans de contrôle de défectivité, la qualification des équipements de production et inspection, les QTs planifiés par équipement de production et les limites de W@R sont obtenus à partir de la base des données. Un pré-traitement des données est fait avec le logiciel R et la résolution est faite avec les solveurs GLPK ou CPLEX. L'ingénieur prépare les différents scénarios avec l'information du mix et volume de production à considérer. Le modèle donne comme résultats: la capacité nécessaire par type d'équipement d'inspection, les équipements de production dont les limites peuvent être garantis ou non et le nombre potentiel des QTs générées pour respecter les limites W@R.

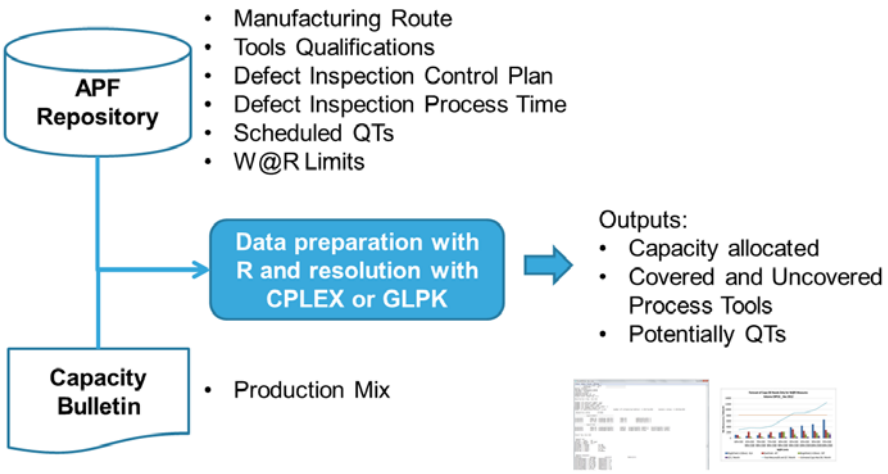
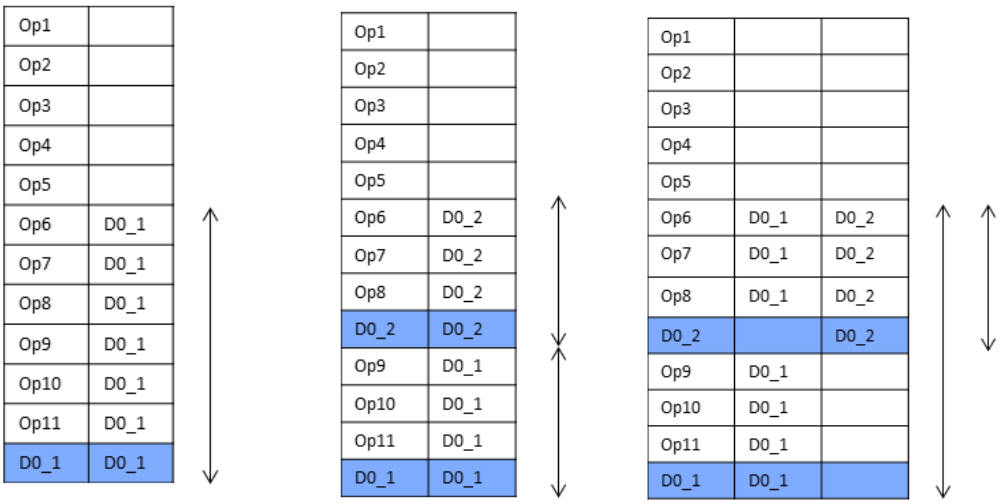


Figure 12: Schéma général du modèle de capacité

0.3.3 Étude de la conception des plans de contrôle pour la défectivité

Le W@R des équipements de production peut être réduit avec l'introduction de nouvelles opérations d'inspection. Cette section correspond au chapitre 7 de la thèse. Dans cette section, on s'intéresse à la manière dont la position et la couverture des opérations d'inspection du plan de contrôle de défectivité impacte la maîtrise des risques. Cette étude a été effectué à l'aide du simulateur S5 développé par le département de Sciences de la fabrication et logistique (SFL) de l'École Nationale Supérieure des Mines de St-Étienne (ENSM-SE). L'objectif a été d'analyser comment l'introduction de nouvelles opérations d'inspection peut avoir un impact positif ou négatif sur la maîtrise du risque global de l'unité de fabrication (fab). Deux stratégies de couverture des opérations de production ont été considérées au moment d'introduire les nouvelles opérations d'inspection: *"Overlapping"* et *"No overlapping"* (voir la figure 13).



(a) Couverture du plan de contrôle original (b) *"No Overlapping"* (c) *"Overlapping"*

Figure 13: Exemple de la couverture du plan de contrôle original et modification de la couverture avec l'introduction d'une nouvelle opération d'inspection

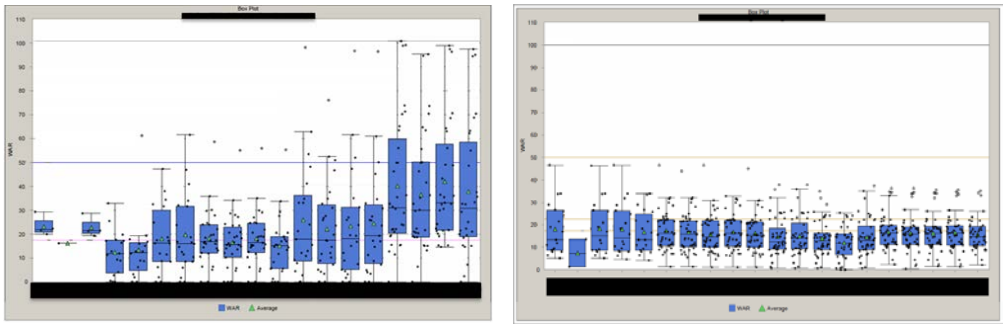
La figure 13 présente un exemple du plan de contrôle et de couverture des opérations d'inspection. Dans le plan de contrôle original (figure 13a) l'opération d'inspection D0_1 couvre les opérations de production du Op6 au Op11. Une nouvelle opération d'inspection est introduite. Si la stratégie *"No Overlapping"* est utilisée pour introduire une nouvelle opération d'inspection (figure 13b) la couverture de l'opération initiale (D0_1) sera réduite, l'opération d'inspection D0_1 couvrira les opérations de production Op9 à Op11 et

la nouvelle opération d'inspection D0_2 couvrira les opérations de production Op6 à Op8. C'est-à-dire qu'avec *"No overlapping"* le nombre des opérations de production qui sont couvertes est divisé par le nombre d'inspections introduites. Avec la stratégie *"Overlapping"*, la couverture du plan de contrôle original est maintenue et les nouvelles opérations sont incluses avec leur couverture. Dans la figure 13c la couverture de l'opération D0_1 est maintenue (i.e. les opérations de production Op6 à Op11 sont couvertes) et la nouvelle opération D0_2 est introduite avec la couverture des opérations de production Op6 à Op8. Les expérimentations ont été faites avec les données industrielles correspondant à un mois d'activité. Les résultats sont présentés et détaillés dans le chapitre 7 de la thèse. Les résultat ont montré que lorsque la capacité d'inspection est limitée, la stratégie *"Overlapping"* permet d'avoir un meilleur impact sur la réduction de risque qu'avec la stratégie *"No overlapping"*. De plus, quand les nouvelles opérations d'inspection couvrent des nouvelles opérations de production, l'impact sur la réduction du risque est plus significatif. Le chapitre se termine avec la proposition d'un modèle mathématique pour optimiser le nombre et le positionnement des opérations d'inspection.

0.4 Conclusions et perspectives

Avec l'implémentation du système de sélection dynamique des lots pour la défectivité, une meilleure maîtrise du risque a été possible. Les résultats industriels montrent l'efficacité du système. L'implémentation industrielle a été possible grâce au travail collectif des tous les participants du projet W@R[4] entre SMicroelectronics et l'ENSM-SE. La plupart des cas dans lesquels les équipements de production n'étaient pas suffisamment contrôlés ou qui étaient sur-contrôlés ont été réduits. La figure 14 présente le niveau de risque dans un groupe d'équipements d'un même atelier avant et après l'implémentation du système dynamique de sélection des lots. Le nombre de wafers potentiellement impactés lorsqu'un problème arrive est limité par l'IL (*Inhibit limit* en anglais) des équipements de production. La figure 15 présente la valeur moyenne des IL de la fab, cette valeur a pu être réduite de 66% depuis le début du projet. Enfin, d'importantes économies ont été obtenues non seulement en termes de réduction de risque global de la fab, mais aussi en termes d'une meilleure utilisation de la capacité en défectivité grâce à une meilleure sélection des lots à inspecter.

[4]Sponsors: P. Campion and M. Le Gall. Leaders: E. Tartière and J. Pinaton.
ST Rousset Team: A. Thieullen, G. Rodriguez-Verjan, P. Palouar, S. Detivaud, B. Pennachio, J.C. Mattlin, F. Chairat, C. Klingelschmidt, V. Lemaire, J. De-selle, D. Courilleau, B. Mari, C. Giuliani, D. Viard.
ENSM-SE SFL Team: S. Dauzère-Pérès, C. Yugma, J.L. Rouveyrol, S. Housseman.



(a) Wafers à Risque (W@R) Avant (b) Wafers à Risque (W@R) Après

Figure 14: W@R sur les équipements de production avant et aprés l'évolution du système de sélection des lots

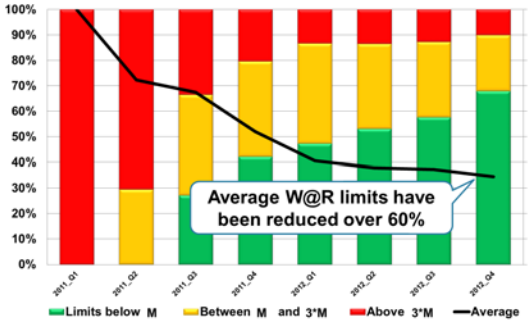


Figure 15: Déploiement et réduction des limites maximales des risque (IL) des équipements de production dans l'unité de fabrication (fab).

En ce qui concerne les perspectives, plusieurs axes de travail ont été identifiés. La réduction de W@R dans les équipements de production dépend du temps passé entre l'opération de production et le moment d'obtention des résultats de l'inspection. Par conséquent, optimiser l'ordonnancement des lots en défectivité permettra une réduction du temps pour obtenir les résultats des inspections. Le problème peut être modélisé comme un problème d'ordonnancement sur machines parallèles avec temps de production différents en prenant comme objectif la réduction du nombre de tâches en retard. Une deuxième perspective concerne l'extension de W@R à d'autres types de contrôles, notamment les contrôles de métrologie classique. Cette perspective fait l'objet des travaux de thèse en cours de M. Alejandro Sendon entre STMicroelectronics et le département SFL de l'école de Mines de St-Etienne. Finalement, optimiser la position et le nombre des opérations d'inspection est aussi un axe important de travail. Ce problème d'optimisation est proche d'un problème de couverture maximale. Les demandes des clients seraient les besoins en termes d'inspections pour les équipements de production, et le nombre de wafers à risque entre l'opération de production et l'opération d'inspection représenterait la distance à parcourir pour satisfaire les besoins de clients.

General Introduction

The technological advances in semiconductor industry are among of the most robust in modern economy. Nowadays, almost every product in our everyday's life has integrated circuits (ICs) (e.g. Phones, cars, computers, TVs, communicating objects). The strong competition and market conditions force manufacturers to provide high quality products at a competitive price. However, semiconductor fabrication is extremely complex due to the continuous shrinking of the critical dimension, the extreme process conditions and the large variety of product mixes. In this way, semiconductor manufacturers seek to develop efficient and effective control strategies to propose high quality products at competitive prices. Actually, the significant progress in semiconductor manufacturing would not be possible without the technological improvement of in-line inspections.

In this thesis, we focus on how to manage and reduce the risk (i.e. number of wafers at risk) during fabrication. With the introduction of inspection operations the cycle times of product are directly impacted with consequences on production costs. Therefore, sampling strategies are used to reduce the number of inspection operations while satisfying quality objectives. Several sampling techniques exist and can be classified according to their capability to deal with factory dynamics. Static sampling strategies are commonly used, because they are relatively easy to implement but the dynamics of the fab cannot be correctly managed, which often leads to cases of inspection operations being performed without real added value and unexpected levels of risk on process tools. Dynamic sampling strategies are relatively recent and aim at integrating the dynamics of the fab. Decisions on whether to select lots or not are taken in real time and considering the current production state. In semiconductor manufacturing, several types of controls exist. In this thesis, we focus on defectivity inspections, which aims at monitoring the processes for defect reduction and yield improvement. During this thesis, the management policy of STMicroelectronics was to implement a strategy to reduce the impact of excursions. Therefore,the sampling system used for defect inspection changed in the site of Rousset of STMicroelectronics from a static sampling strategy to a dynamic sampling strategy. Results showed significant improvements in terms of risk reduction of the fab without the need to increase defect inspection capacity.

In this thesis, a novel approach to select the lots to inspect is proposed, it is based on a skipping mechanism to efficiently manage the defect inspection queues and enable the release of lots with redundant information. Then, a model for defect inspection capacity planning is proposed, which considers both static sampling and dynamic sampling strategies. Finally, a study on how the design of defect inspection control plans can impact risk is performed. The manuscript is organized as follows:

- Chapter 1 presents the industrial context. It provides a general description of the semiconductor manufacturing process and the principal types of controls that are performed during fabrication. Then the problems are described and the thesis objectives are discussed.

- Chapter 2 proposes a literature review for the inspection allocation problem, and discusses the different research axes that are considered in this thesis.
- Chapter 3 describes how the sampling system changed from static sampling to dynamic sampling in the Rousset factory of STMicroelectronics.
- Chapter 4 details the algorithms developed for optimizing the selection of lots through a skipping mechanism. Numerical experiments on industrial data are presented and discussed.
- In chapters 5 and 6, a model is proposed for the capacity planning problem when static sampling and dynamic sampling strategies are considered. As in chapter 4, numerical experiments on industrial data are presented and discussed.
- Chapter 7 presents a study on how the defect inspection control plan design can impact the tool risk management.

The manuscript concludes with a general discussion of the conclusions, industrial results after implementing the dynamic sampling approaches and perspectives for future work.

Chapter 1

Industrial Context

This chapter provides a general introduction of process control in semiconductor manufacturing. In particular, we describe the role of defect inspection controls within the manufacturing process and the main criteria to design the defect inspection control plan. Finally, we introduce our problem and the objectives of the thesis.

1.1 Introduction

The semiconductor term refers to the availability of a material to conduct electricity. Nowadays, more than 85% of Integrated Circuits (ICs) are made from silicon semiconductor material. *IC devices* refer to the transistors, diodes, resistors and capacitors that are manufactured on the silicon surface and that are connected into a circuit to define how the chip will function [15].

The industry of semiconductors was born with the invention of the first transistor in 1947 at Bell Laboratories. Some years later, in 1957 the first commercial transistor was fabricated at Fairchild Semiconductor Corps. Then, in 1959 the integration of multiple electronic components on one silicon substrate was co-invented by Robert Noyce at Fairchild Semiconductor and Jack Kilby at Texas Instruments. These two inventions represent the basis for the IC that we know nowadays [16]. The extraordinary development and innovation of semiconductor industry is one of the most robust in modern economy (more than U.S $ 2 trillion industry [17]). Actually, electronics have been an enabler for productivity and growth in all areas of economic activity. Almost each product in our everyday's life has some microelectronic components: Phones, printers, cars, TVs, PCs, electrodomestics, etc. Moreover, consumers receive products with higher performances and ever lower prices. Since 1970, the number of components per chip has doubled every two years following the historical trend known as *Moore's Law*. With this continuous down scaling of the critical dimension in the integrated circuit, the controls performed during the manufacturing of ICs have a key role to achieve the expected yield[1]. Indeed, the significant progress during the last 30 years in the semiconductor industry would not be possible without the science of yield enhancement and defect reduction [18]. In this chapter, we introduce the semiconductor manufacturing processes and the controls that are performed during fabrication.

[1] *Yield*: Is one of the most important metrics to evaluate the performance of a wafer fab. It expresses the ratio of good produced parts over the total produced parts.

This chapter is organized as follows: section 1.2 provides a general introduction of the semiconductor manufacturing process. Section 1.3 describes the role of process control and defines the framework of this work. In particular, this section introduces defect inspection and how defect inspection control plans are defined. Finally, the problem and scientific objectives are presented in Sections 1.4 and 1.5.

1.2 IC Fabrication

The fabrication of an Integrated Circuit (IC) can be divided into three general stages:

- Wafer Preparation: A silicon ingot with the appropriate diameter is sliced into thin wafers. A circular shape is used in order to minimize losses due to wafer handling during fabrication.
- Front-end Processing: In this stage the wafer fabrication is performed. It is divided into Front-End Of Line (FEOL) and Back-End Of Line (BEOL). In the FEOL the electronic components are fabricated (i.e. transistors, capacitors, resistors), and in the BEOL the electronic components are interconnected.
- Back-end Processing: It refers to the testing, sorting, assembling and packaging of each die of the wafer. In the testing and sorting steps, each die is probed and electrically tested, bad dies are marked to be sorted later. In the assembling and packaging steps, the wafers are cut to separate each die, the marked dies are scraped or rejected and the others are packed. Then, the metal connections are bounded to each die, which is encapsulated in a protective package. At the end, the final test is performed to ensure that the product meets the electrical and environmental specifications.

This thesis focuses on the Front-end Processing. Therefore, we introduce the different processes performed during the wafer fabrication. The wafer fabrication in semiconductor manufacturing is characterized by its complexity and highly expensive processes. An IC is made layer by layer and the number of layers depends on the product technology. Complex technologies can have more than 40 layers, which implies more than 400 successive processes. Figure 1.1 is a schematic representation of the wafer fabrication processes. A general explanation of the different processing steps to create a layer is presented.

- *Oxidation process*: A defect-free uniform layer of Silicon Dioxide (SiO_2) is grown by heating the wafer at very high temperatures and with O_2. Thin oxides, such as gate oxide, are grown with dry oxygen. Gate oxide is an important layer under which a conduct channel is formed between the source and the drain. With a field oxide, isolation from other devices can be provided [19].
- *Deposition process*: Thin films of different materials are deposited on the wafer through several processes, such as: Chemical Vapor Deposition (CVD), Physical Vapor Deposition (PVD), Plasma Enhanced Chemical Vapor Deposition (PECVD),

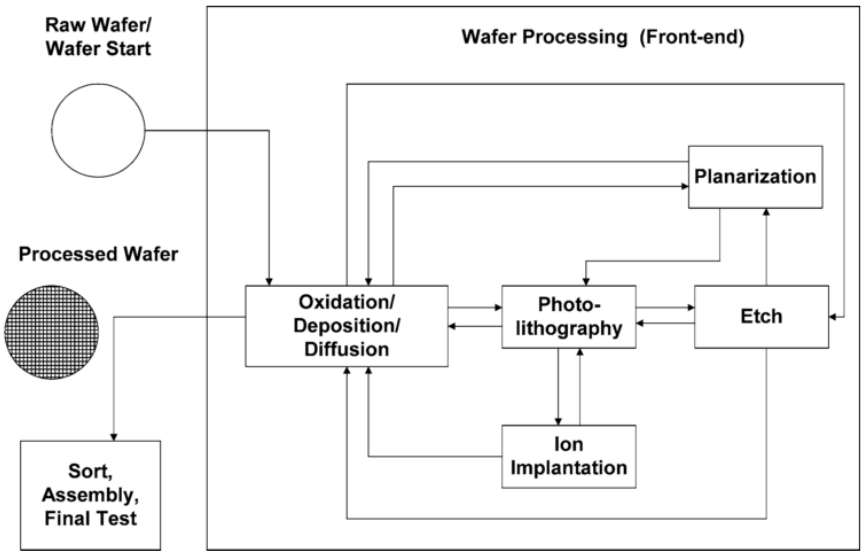


Figure 1.1: Schematic representation of the Wafer Processing [1]

Epitaxy or Metalization. CVD is used for depositing dielectric and metal films. PVD is used for applying metals such as aluminum. The CVD and PVD are performed on cluster tools[2].

- *Photolithography process*: It is used to add the patterns on the wafer. First, the wafer is coated with a film of photosensitive polymer. Then, the wafer pattern is transferred from a photo mask, also called *reticle*, onto the photosensitive polymer by projecting light through the reticle and exposing the wafer using ultraviolet light. Finally, the polymerized sections of photoresist material are removed from the wafer to develop the pattern.

- *Etch Processes*: With the etch processes, the areas defined by the patterns can be removed. In wet etch, liquids are used such as, acids, bases and solvents to chemically remove wafer surface material. Whereas in dry etch, the wafer surface is exposed to a plasma created in the gaseous state. Dry etch is the most common etch process and a high-density, low-pressure etch reactor, such as the ICP (Inductively-coupled Plasma) is widely used [15].

- *Planarization processes*: In order to achieve a flat layer, the wafer surface is polished. This is critical for the follow-on process steps (e.g. better linewidth control during photolithography) and can serve to increase device yields by removing undesirable foreign material on the wafer surface. The most common planarization technique is Chemical Mechanical Polishing (CMP) [15].

- *Ion Implantation*: With ion implantation, the doping process of the wafer is performed to obtain the right electrical properties (i.e. specific regions with positive

[2] Cluster tool: A tool that combines several process chambers within a closed environment together with a handling robot.

or negative charge). The most used gases carrying the desire dopant are arsenic, phosphorus, boron, boron difluoride, indium, antimony, germanium, silicon, nitrogen, hydrogen, and helium [20]. Ions are accelerated for implantation over an energy range from $\approx$ 100eV to nearly 10MeV and a dose range from 10^{10} to over 10^{18} ions/cm^2 [21].

Wafer fabrication is one of the most costly, complex and time consuming stages of the IC manufacturing. The complexity factors can be summarized in three categories:

- Process conditions, which refer to re-entrant flow, mix of different process types, unrelated parallel machines, sequence-dependent setup times, etc. [22, 23].
- Product evolution, which refers to the continuous change of products design in order to face the increasing demand for more powerful and faster devices [17, 24].
- Market conditions, which refers to the highly uncertain demand, the long lead times and the high cost of capacity increase [23, 25].

For all these reasons, it is crucial to guarantee the quality of the final products. The *yield* is one of the most significant metrics to evaluate the ability of a wafer fab to produce high quality products [26]. It is expressed as the ratio of good produced parts over the total produced parts. The types of controls performed in semiconductor manufacturing can be classified into six main levels [27]: Facilities and technical installations, equipment sensors, fab and in-line measurements, parametric test, functional test and physical characterization. In this thesis, we focus on the in-line measurements and the process control performed during fabrication. The next section introduces the key areas for process control.

1.3 Process Control in Semiconductor Manufacturing

In order to achieve the challenging goals of the semiconductor industry, wafers fabs have improved the control over process parameters and reduced the source of defects during fabrication. The key areas for process control are *Metrology* and *Defect Inspection* [10]. The objective of *Metrology* is to control the physical and electrical properties of the wafers during fabrication. Many of the improvements on *Metrology* came from the development of sophisticated equipment that can provide real-time data. The objective of *Defect Inspection* is to detect and reduce the defects produced by particles or process drifts that affect the production *yield*. The main improvements in defect reduction came from the development of tools with higher sensibility and techniques for defect classification and source identification [18]. Metrology and inspection tools can be divided into two categories: stand alone measurement tools and integrated measurement tools (See Table 1.1).

The information acquired from the process and the product need to be properly processed in order to create knowledge, increase equipment efficiency and accelerate yield improvement. Advance Process Control (APC) includes a wide range of techniques to obtain and treat the information in order to improve the electrical performance of devices by reducing variations and correcting drifts during the manufacturing process. The information can be classified as follows [28]: i) Real-time equipment data which are typical

Table 1.1: Measurement tools classification [15]

Category	**Measurement System**	**Description**
Stand-Alone Tools	Off-line	Available only outside fab (i.e. Laboratory)
	At-line	Available in the fab (measure monitor wafers that are destructive, contaminated of unpatterned
	In-line	Used during production (can measure patterned wafers)
Integrated Tools	On-line	Available at the process workstation to measure patterned wafers
	In-situ	Measure wafer, process or equipment during processing (real-time measurement)

from in-situ sensors, ii) Data on geometric properties, which are obtained with integrated metrology tools and with stand-alone metrology tools, iii) Wafer Acceptance Test data (WAT), which are provided by the electrical tests at the end of the line and is available after long time delays because many steps are required to complete the structure before measurement, iv) Yield data, which give the percentage of good dies over the total produced dies, usually available after almost one month of production.

The different types of methods used to obtain and analyze the information can be summarized as follows:

- Statistical Process Control (SPC): It is based on statistical tools to ensure the stability of the process. SPC is used to decide whether the process is operating under statistical control or not. The SPC was born with the "control chart" proposed by Walter Shewart in 1930. Thanks to the availability of automated in-situ data collection and real time data processing, SPC has significant evolve over the years (e.g. multivariate control, model-based SPC, time series models) [29].

- Fault Detection and Classification (FDC): It is a statistical method focused on equipment data. The data is collected in real-time and, when a problem is detected, the tool is stopped and actions are taken. [29]

- Run-to-Run (R2R): It is a loop control technique based on equipment and lot information. Depending on results obtained in a run, the controller adjusts the recipe variables to reduce the output variability. In general R2R controllers are models coupled with a mechanism to observe the variables of interest. There are two types of control loops: Feed-forward and Feed-back. The Feed-forward loop adjusts the recipe parameters using the results of previous measurements. The Feed-back loop adjusts recipe parameters based on results of post-measurements in order to counteract process drifts [30, 31].

- Virtual Metrology (VM): It is a technique that aims at predicting metrology measurements and forecasting electrical and physical parameters on wafers. It is based

on predictive models generated with process sensor data and previous metrology measurements. An accurate VM model has several benefits, such as time and cost reduction due to less direct measurements on wafers [32]. However, development of accurate models and robust over time is a challenge due to the large number of input variables and the lack of real measured lots in the learning phase of the model [33], [34].

- Defect Inspections or Defectivity controls: They are used to monitor the process for defect excursions and to drive continuous improvement of the yield. Excursions are temporal yield losses that randomly happen (in time) as a consequence of an out-of-control condition on a process tool. In this thesis we focus on how to manage and reduce the number of wafers at risk regarding defect inspections. Therefore, only the control operations performed in defect inspection area denoted as "inspection operations" are considered. This type of controls are presented in detail in the remaining of this section.

1.3.1 Description of Defect Inspections

The main role of defect inspections (or "*Defectivity controls*") during manufacturing is to monitor the process for defect excursions and to drive continuous improvement of the yield. Inspections are performed either on patterned wafers or non-patterned wafers. The inspections performed on patterned wafers are used to decrease *defectivity* during the ramp-up and full production phases, to monitor the processes that introduce contamination, scratches or pattern defects and to predict the yield. The inspections performed on non-patterned wafers are used to monitor contamination and surface quality of the wafer and to monitor the cleanliness of tools (e.g. process and metrology) [10]. In general, the control operations performed in the "*Defect Inspection Area*" are addressing all the production tools of the fab. Figure 1.2 is an operational description of the defect inspection area.

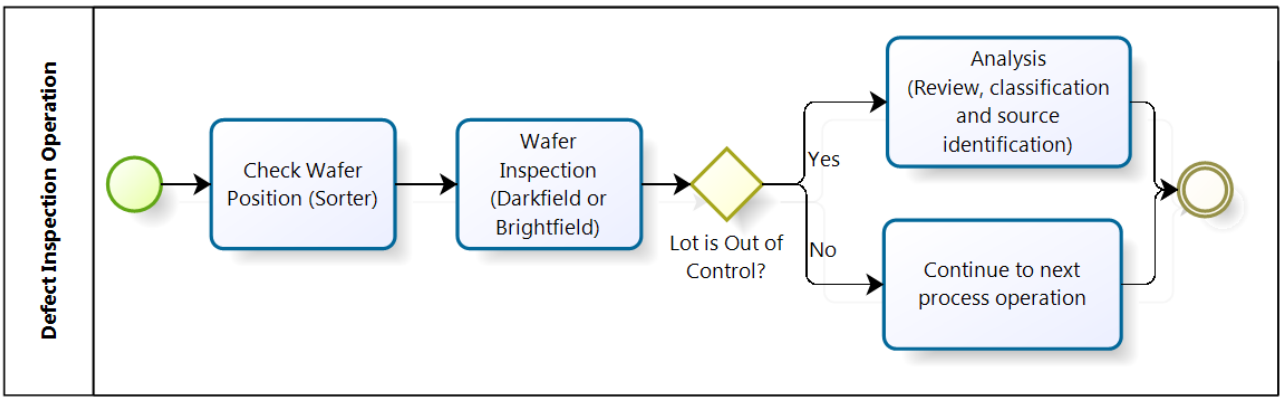


Figure 1.2: Operational Description of the Defectivity Area

1. **Check Wafer Position.** In general, the same wafers from the same lot are inspected at different stages of their manufacturing route, the objective is to compare the added defects from previous inspections. Therefore, the wafer ID (Identification Number) within the slot is verified with a Sorter tool.

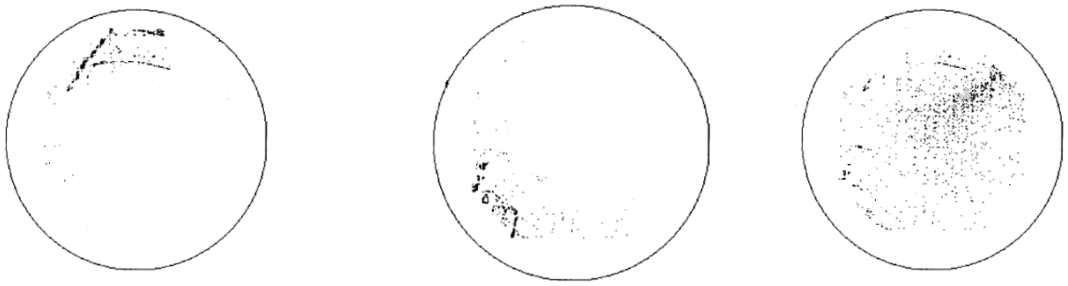
2. **Wafer Inspection.** The number of defects over the surface are counted (defect density) and the localization of defects over the wafer is generated (mapping of defects). This is done with optical detection tools (i.e. Brightfield, Darkfield) or tools for electrical test (i.e. Electro beam). The optical detection principle is based on the die to die comparison. If the intensity of the reflected light is larger than a predefined limit, a defect is detected. Results are compared with predefined control limits and if the lot is considered under control, it can continues to the next process operation. Otherwise, the lot cannot exit defect inspection and further analysis is performed.

3. **Analysis.** In this phase, the defects are reviewed, classified and the root cause identified. Defects can have a signature or a pattern that is characteristic of the process step and/or the process tool (Figures 1.3 and 1.4). At this stage, there are some defects whose root cause can be identified based on a typical signature. However when several failures occur simultaneously, detection becomes more complex.

   (a) **Review.** In the review, an image of the defects is obtained. This is done with specialized microscopes that use different technologies (e.g. Optic microscope, Scanning Electron Microscope (SEM)).

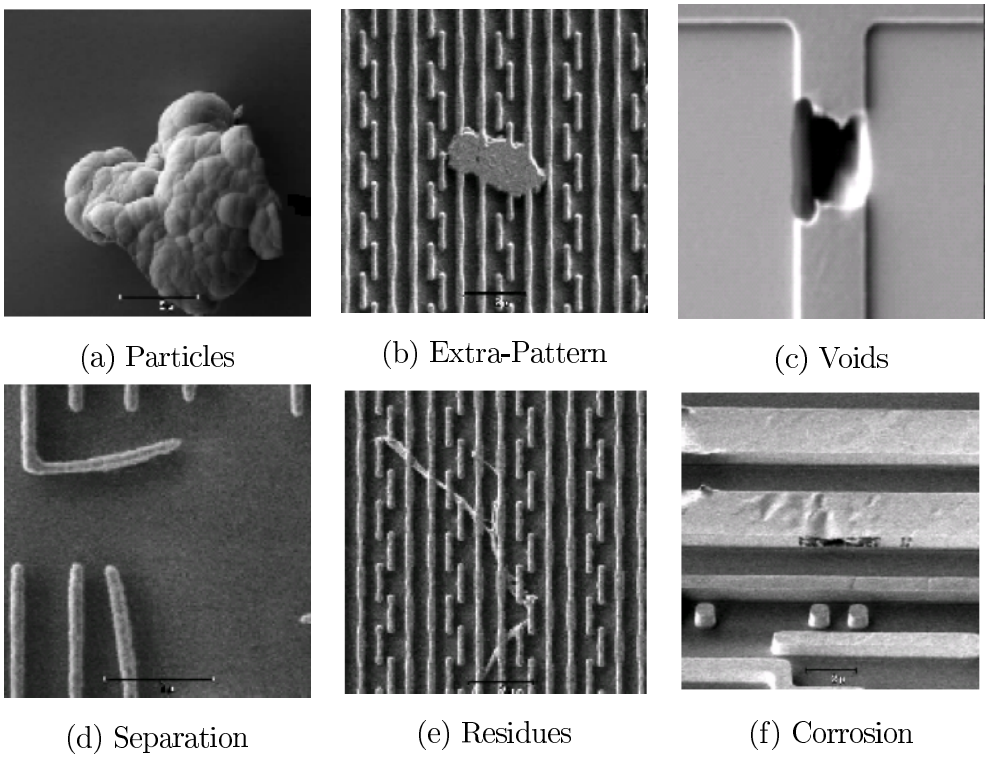
   (b) **Defect classification.** In this phase, the defects are classified and their relative importance is characterized based on their frequency and size. In the past, this phase was entirely manual, i.e. a trained operator sorted the defects into categories using a reference book that contained the image of typical defects. Nowadays, Automatic Defect Classification (ADC) systems are available which significantly reduce the subjectivity and errors from operators [10]. Future trends include real-time defect classification, higher detection sensitivity, improvements on data management for yield learning and *in-situ* particle monitoring [18].

   (c) **Defect source identification.** The objective of this phase is to identify the process tool that generates the defect. This is performed using a variety of techniques such as Spatial Signature Analysis (SSA), Automated Image Retrieval (AIR) [35]. The root cause of some defects can be identified based on typical signatures, figure 1.3 is an example of three types of signatures. The signature of figure 1.3(a) is a double-slot signature. This was produced by a robot handler that attempts to place a wafer in a slot that is already occupied, this information is enough to locate the source of the problem without further analysis. This is the same for figure 1.3(b), which is due to a robot handler that scratched the wafer. However, there are some signatures that need further analysis. For example, figure 1.3(c) presents the signature of a CVD contamination. In this case, the defect engineer isolates a particular CVD but there is not enough information on the composition and source of the contamination defects [36].



(a) Double Slot (b) Mechanical Scratch (c) CVD Contamination

Figure 1.3: Examples of defect signatures [36]



(a) Particles (b) Extra-Pattern (c) Voids

(d) Separation (e) Residues (f) Corrosion

Figure 1.4: Examples of defect patterns

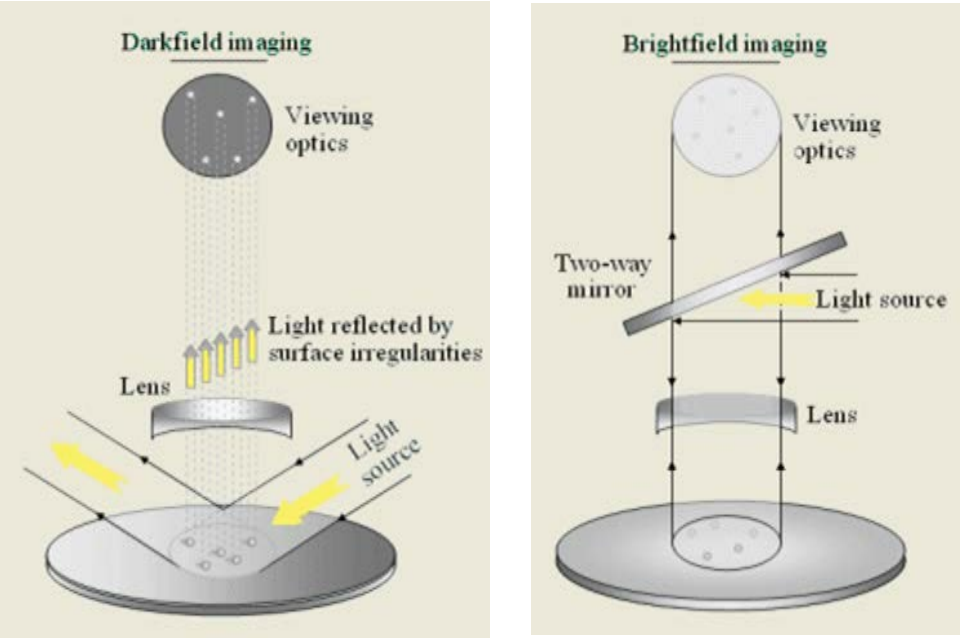
1.3.2 Defect Inspection Tools

There are several types of tools used to perform defect inspection controls, the most popular are those based on Dark-field, Bright-field and Electron-beam technologies.

Darkfield (Figure 1.5a). A laser beam is used. The laser is projected onto the wafer surface and the intensity of the diffracted light (or *scattered*) is detected. The most common type of angle-resolved scatterometer is called a "2θ" scatterometer due to the two angles (incident and measurement) associated with the method [19]. The darkfield tools have higher throughput compared with other tools. Hence, as many recipes for defect inspection as possible are qualified[3] on this type of tools.

Brightfield (Figure 1.5b): A lamp is used which resembles a microscope. A broadband light source is used to illuminate the wafer. Then, the system collects both the scattered light and the reflected light through the same aperture to obtain an image[15].

Electro Beam: Electrical defects cannot be detected with darkfield and brightfield tools. Therefore, electrical tests are performed to identify possible flaws. This tool projects an electron beam on the wafer surface and detects electrons that scatter back. With this tool, it is possible to detect whether a material conducts or insulates as it is supposed to. Compared to the two other types of tools, the Electro Beam is the most sensitive but has the longest inspection time [15].



(a) Darkfield technology (b) Brightfield technology

Figure 1.5: Defect inspection techniques [15]

[3]Tool qualification is a kind of setup that assures the right conditions for the process [37]. Qualification of inspection tools, refers to the set of instructions or control parameters that are necessary to inspect a wafer.

1.3.3 Defect Inspection Control Plan

The defect inspection control plan is the list of inspection operations that are performed throughout the manufacturing route [4] of products. It states the position and coverage of inspection operations. The position refers after which process operations a defect inspection operation is performed. The coverage refers to the set of process operations that can be controlled with an inspection operation. The defect inspection control plan is defined by product. Depending on the production volume, maturity and process criticality, the defect inspection control plan can be *complete*, *standard* or *relaxed*. A *complete* defect inspection control plan means that lots must be inspected in all the defined inspection operations of their route. A *standard* and *relaxed* defect inspection control plan means that lots must be inspected only in some inspections operations of the route. In the *relaxed* control plan, there are fewer mandatory inspections operations compared with the *standard* control plan. Figure 1.6 is a representation of the control plan for three products from the same technology (i.e. A-11, A-22, A-33). Product A-11 has a *complete* control plan while product A-22 has a *standard* control plan and product A-33 has a *relaxed* control plan.

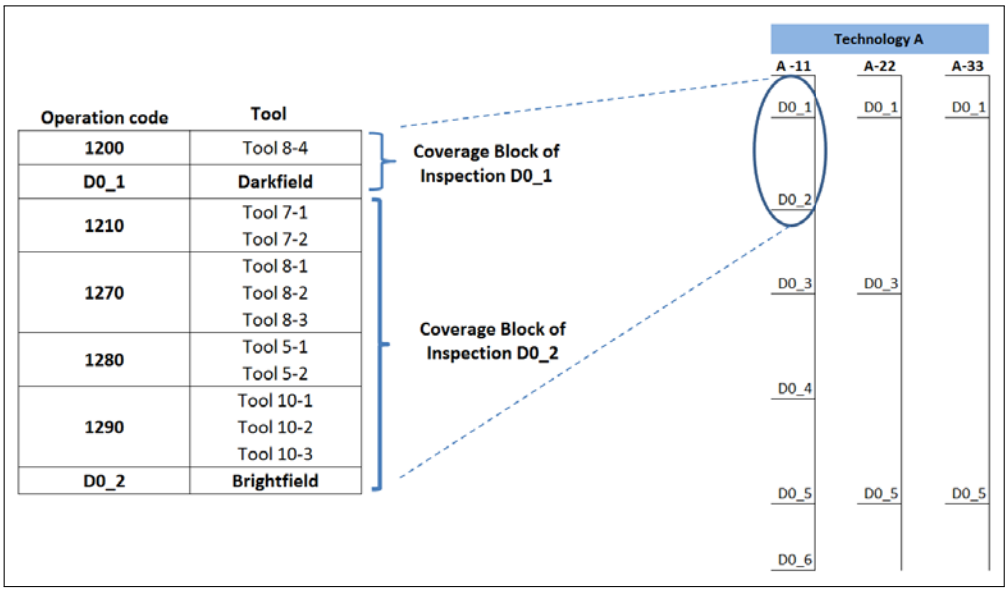


Figure 1.6: Representation of a defect inspection control plan and coverage block

In figure 1.6 an example of the coverage block of inspection operations D0_1 and D0_2 is illustrated. The process operation 1200 can be controlled with the inspection operation D0_1 and process operations 1210 to 1290 can be controlled with the inspection operation D0_2. Therefore, with the inspection of one lot, several process tools can be controlled. This is an important characteristic of defect inspection operations and one of the main concepts to keep in mind for the remaining of this work.

The coverage of an inspection operation is defined according to detection capability across successive layers. The main criteria to create defect inspections are the criticality,

[4] The manufacturing route is the sequence of process operations required to obtain the final product.

relevance and exposure which depend on the product (e.g. Phases of integration, maturity, design).

- **Criticality.** The criticality of a process is based on the killer ratio (KR). The KR defines whether a defect is killer or not depending on the size and the pattern density where it is located[5] [38]. In general, defect inspections are located between critical processes for the product (e.g. Metal deposition, photolitography, etching steps, film deposition).

- **Relevance.** It refers to the capability to detect defects and the time required to perform the inspection. The capability of an inspection operation depends on the nature of the layer. For example, after an opaque layer, it is not possible to inspect the wafer because the light cannot be reflected. After a transparent layer, the defects of all layers can be observed, hence locating the defect is too complex. After a metallic layer, the material reflects too much light, thereby many false defects are detected. In addition, there are some process conditions that restrict the creation of inspection operations, such as time constraints between consecutive process steps. These constraint loops are established to prevent the native oxidation and contamination effects on the wafer surface. If the time constraint is violated, the lot will be scrapped [39].

- **Exposure.** It refers to the minimum number of wafers that are exposed if a problem occurs. It is calculated as the throughput (TH) of the process tool multiplied by the cycle time (CT) from the process operation until the inspection operation ($CT \times TH$). With the introduction of dynamic controls, this criterion is now very important to locate defect inspection operations.

In order to inspect a new product or to control a given process, it is necessary to create an inspection recipe. The inspection recipe contains the measurement parameters such as optical and signal processing configurations. Due to the number of parameters to consider, the complexity for creating inspection recipes increases as the inspection technology advances [10]. This is why only the necessary number of inspection operations are created. Therefore, when a lot of a given product can be inspected, it is said "measurable", otherwise, the lot is "not measurable".

1.4 Problem Description

When a process or a tool drifts out its control limits, all the wafers that are processed can be affected and generate excursions[6]. Considering that the impact of an excursion depends on the severity and the number of wafers that are affected before the problem is identified and corrected [11], the need to find defective products as fast as possible is critical. This is

[5] A defect is considered Killer when its presence can cause the dice structure to fail.

[6] Excursions: Temporal yield losses that randomly happen (in time) as a consequence of an out-of-control condition on a process tool.

achieved with regular controls throughout production. In consequence, sampling strategies are defined to select whether a lot is controlled or not.

Sampling strategies can be classified according to their capability to react depending on the production state (e.g. Static, Adaptive and Dynamic Sampling) [3]. Until recently, a Static Sampling strategy named "*Start Sampling*" was used in the site of Rousset of STMicroelectronics (and is still being used in other factories) to select lots for defect inspection. This strategy consists in selecting at the beginning of the manufacturing process the lots that will be inspected. The sampling rate is different among products, and it is defined according to the production volume, criticality, maturity of the product and customer requirements.

In Start Sampling, the sampling rate is defined as "$1/N$", "N" being the number of wafers that are considered at risk. Figure 1.7 is a representation of the expected risk on process tool 1. Let us suppose that the throughput of this process tool is 4 lots per hour and the sampling rate specifies that 1 lot out of 4 should be inspected (in Figure 1.7 the yellow lots are selected for inspection). If the time for the selected lots to reach the inspection operation is approximately 2 hours, when a lot is inspected, the process tool has at least already processed 8 new lots. Hence, if a problem is detected, the number of wafers that are potentially impacted will be larger than 4. In order to reduce the number of wafers at risk, it is necessary to accurately calculate this indicator. Chapter 3 presents the strategy used to tackle this problem. The elapsed time before the selected lot arrives at the defect inspection operation and is inspected, is considered as the cycle time (CT) from the process operation until the inspection operation. The CT multiplied by the throughput of the process tool (CT×TH) is the exposure level of the process tool for a specific inspection operation (See section 1.3.3).

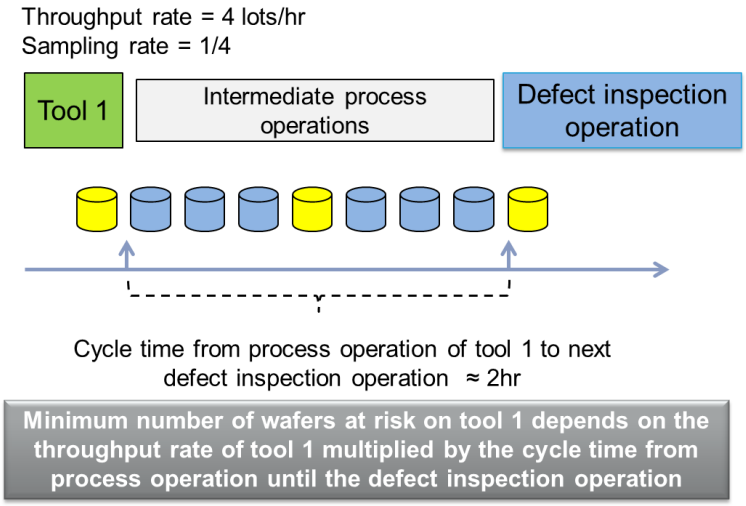


Figure 1.7: Evaluation of the minimal number of wafers at risk on a process tool

The exposure level is very important to reduce the number of wafers at risk on process tools. In order to reduce this exposure, CT should be reduced. Strategies to decrease CT combines the reduction of the waiting time in the inspection area, which can be done

with new scheduling mechanisms that consider the number of wafers at risk on process tools and the introduction of new inspection operations within the manufacturing route. However, in the case of Start Sampling, the selected lots are systematically inspected in all the inspection operations of their route. Hence, the creation of new inspection operations can drastically impact the inspection capacity workload and the total cycle time of the selected lots. For example, in a simplified case of one production route of 300 successive process operations and 5 hours per process operation, the total cycle time would be 62,5 days. If there are 40 initial inspection operations, the total cycle time of the selected lots will be 13% larger than the lots without inspection. If additional inspection operations are required to divide by two the number of wafers at risk on all process operations, 80 inspection operations should be performed, which leads to an increase of 27% of the cycle time for the selected lots. But the complexity of the system increases with the number of parameters to consider (e.g. different sampling rates, several products, different defect inspection control plans, tool qualifications). Moreover, increasing inspection capacity can cost between $ 3 and $ 8 millions of dollars depending on the technology required [40]. A reduction of the manufacturing cycle time can very significantly contribute to yield improvement by reducing the elapsed time between the occurrence of the excursion and its detection since corrective actions can be taken earlier [41].

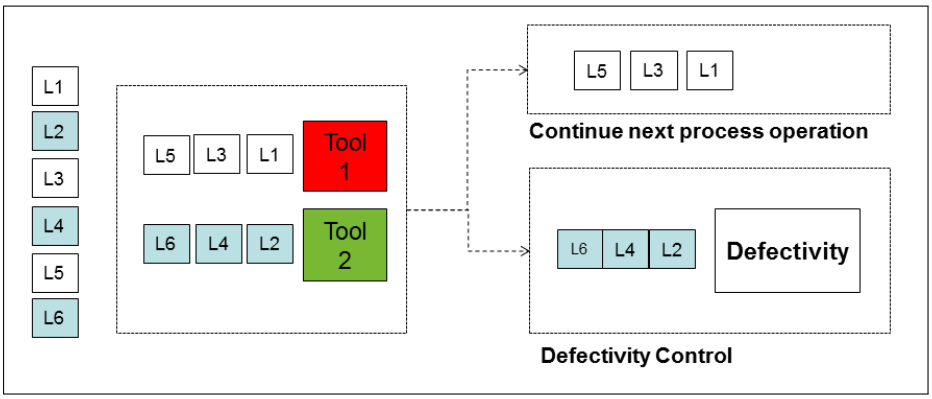


Figure 1.8: Drawback of Static Sampling [12]

An additional aspect to consider is that process tools can have different qualifications and lots are processed according to the availability of tools. Hence, not all the selected lots for inspection are processed on the tools that need to be controlled. This situation can result in cases of over-control and lack of controls on process tools [42]. Figure 1.8 is an example of the problem, where tool 1 is not controlled and tool 2 is over controlled because all the selected lots for inspection are processed on tool 2. In order to tackle this situation, new methods to dynamically select the lots for control were introduced (i.e. Dynamic sampling). The objective is to limit the number of wafers that are potentially impacted if a problem occurs on process tools.

1.5 Thesis Objectives

Within the framework of this thesis, we focus on how to manage and reduce the number of wafers at risk on process tools regarding defect inspection controls. As previously discussed, defect inspection controls are particularly complex because they address all the production tools of the fab. During the realization of this thesis, the sampling system in the site of Rousset of STMicroelectronics changed from a sampling strategy that was only focusing on yield enhancement (achieved with start sampling) to a sampling strategy that also focuses on faster detection and reduction of excursion impacts (achieved with dynamic sampling). Hence, the following objectives were defined:

- To propose and implement new methods to reduce and manage the number of wafers at risk on process tools.
- To propose and implement a new model for defect inspection capacity planning that considers both sampling strategies (i.e. Static and Dynamic sampling).
- To study how the design of defect inspection control plans influence the risk on tools.

1.6 Conclusions

In this chapter, we introduced the different processes performed during wafer fabrication and the importance of process control during manufacturing. Control operations are necessary to ensure the quality of products, however they are costly and directly impact cycle times of products. Hence, only a limited number of lots can be controlled. It is thus critical to cleverly select these lots. In this thesis, we focus on how to manage and reduce the number of wafers at risk on process tools regarding defect inspections. At the beginning of the project, only a Static sampling strategy named "Start Sampling" was used in the site of Rousset of STMicroelectronics to select the lots for inspection. This strategy consists in selecting the lots at the beginning of their manufacturing process that will later be controlled in all the inspections operations defined in their defect inspection control plan. The main concern is yield enhancement by controlling critical processes for the product which are evaluated with the killer ratio. However, with "Start Sampling", the dynamics of the fab (e.g. tool qualification, dispatching rules, lot priorities) cannot be correctly handled, generating cases of over-control and lack of control on process tools. In order to better control process tools by reducing the number of wafers at risk, different research axes are investigated: New methods to dynamically select lots for inspection, a new model for inspection capacity allocation and the analysis of the impact of the control plan design on the risk on tools. In the next chapter, a literature review on these different axes is provided, and chapter 3 presents a description of how the sampling system changed.

Chapter 2

Literature Review on Inspection Allocation

This chapter presents the state of the art for the different research axes that are considered in this thesis.

2.1 Introduction

Inspection allocation problems on multi-stage manufacturing systems have been widely studied in the literature. Different surveys can be found in [43], [44], [45], [46] and [3]. The main objective of inspection allocation is to determine: i) The number and location of inspection stations, ii) The sampling frequency and sample size and iii) The rigor of inspections (acceptance limits) [46]. These three questions are interrelated. Thereby, when the problem aims at finding the optimal location of inspections, it is assumed a fixed sampling rate. When the problem aims at finding the optimal sampling mechanism, it is assumed that the inspection locations are already established. Works that address the two questions simultaneously use an iterative mechanism [47], [48]. According to the life cycle of the product, different inspection allocation objectives are considered. During the ramp-up phase, most of the problems are unknown, and as many data as possible are necessary to characterize defects and systematic issues. Therefore, a high sampling frequency is required to define the critical layers to inspect. When the product is in the full production phase, the critical layers are already defined and the focus is on an economic sampling strategy to monitor excursions [49].

In the following, a review of inspection allocation problems is presented. The scope of the literature is excursion monitoring and control. The reviewed papers are classified according to the following criteria: i) System Characteristics, ii) Solution approach and iii) Sampling Strategy.

2.2 System Characteristics

According to the configuration of the system, three general categories can be identified: i) Serial systems, ii) Assembly systems and iii) Non-serial systems. In Serial Systems, the products pass through a sequence of processing stations where each processing activity has a single immediate predecessor [50]. In an Assembly System, products from different manufacturing lines are assembled [51]. Configurations that are neither serial nor assembly

are Non-Serial systems. A semiconductor manufacturing system can be considered as a typical example of a Non-serial manufacturing system. Therefore, we focus on this system configuration. Concerning the risk associated to performing an inspection, there are two type of errors: Type I error and type II error. Type I is the risk associated to the rejection of good items. Type II error refers to the acceptance of non-conforming items. In general, the type II error is more serious [46]. Some authors considered both errors [52], [40], while others considered only one type of error [49], and others considered that inspection is error free [53], [5], [48].

Another important characteristic is the defect mechanism, which can be classified into two categories. In the first category, the probability of an item to become defective is independent of the processing of previous items [54], [55]. In the second category, the process is modeled as a two-state Markov process, because during the time that the process is performing out-of-control, it can generate defects on wafers, and the state changes when the problem is detected and corrected [40]. In this work it is considered that inspections are error free and the excursion monitoring and control problem is a two-state Markov Process. Several works consider that the inspection can be performed immediately after any process station [5]. But the type of controls considered in this thesis can address several process tools and cannot be performed after each process operation. The location of control operations (i.e. inspection operations) are determined by the criticality and capability to detect defects and the exposure of process tools, which is established in the defect inspection control plan design.

2.3 Sampling Strategies

In order to verify the quality of products and processes, in-line inspection operations are introduced in the manufacturing process. With this, an early detection of problems can be guaranteed. However, unnecessary inspection operations can have a negative impact on the yield due to long waiting times of lots in front of metrology or inspection tools, which result in delays to take corrective actions [56]. In addition, some excursions can only be detected at the probing test. Hence, long manufacturing cycle times may have a detrimental effect on yield [41], [57]. For all these reasons, sampling strategies are used to find a trade-off between product yield, costs and cycle time. Several sampling strategies exist and can be classified according to their capability to deal with factory dynamics and variability. An excellent review is presented in [3] where authors propose three categories to classify sampling strategies: Static, Adaptive and Dynamic.

Static sampling is based on rules that do not change throughout production and different static sampling strategies exist. Inspecting every constant number of lots arriving at a workstation or selecting a fixed number of lots at the beginning of their manufacturing route are the most common strategies used in practice. With these strategies the inspection capacity is allocated in advance via the sampling rate of lots. Since the selected lots are systematically inspected, a limited number of inspection operations are introduced in the manufacturing routes to avoid saturating the inspection capacity. The main advantage of this strategy is that the added defect density can be identified between sequential steps [18]. This technique has been widely used in semiconductor manufacturing thanks

to the relative simplicity for implementation. However, with this strategy, the dynamics and variability of the fab cannot be correctly handled [42].

Adaptive Sampling is based on Static Sampling but the sampling rate is adjusted according to the production state [58],[49],[59]. When a problem occurs, the sampling rate increases in order to inspect more lots and correct the process variations. When process is under control or the risk is not significant, the sampling rate decreases to better use the inspection capacity. However, managing inspection capacity is more complex due to variations on the workload of metrology and inspection tools during production.

Dynamic Sampling is a strategy that selects the lots in real time. No rule is defined in advance and decisions are taken according to the manufacturing state (e.g. information obtained by inspecting lots, workload of inspection tools, level of risk in the fab.) [9], [60], [61]. With this strategy, the workload of metrology and inspection tools is better controlled contrary to the adaptive and static sampling strategies. Table 2.1 summarizes the classification of surveyed papers.

2.4 General discussion and thesis approach

The scope of this thesis is the inspection allocation in order to manage the number of Wafers at Risk (W@R) on process tools. The concept of material at risk (MAR) was introduced by Bean [11]. He proposed a model based on probabilities of excursion to allocate inspection operations. First an identification of the risk operations and their probability of excursion is performed. Then, the re-allocation of inspection operations is proposed and the minimum level of material at risk is evaluated in order to compare the gain of re-allocating inspection operations. Their model assumes that there is only one inspection location per defect source. Elliot *et al.* [52] study the optimal Critical Dimension (CD) sample planning that aims at identifying problems more quickly. The inspection errors I and II are considered. Through data analysis, the systematic and random components of variations between lots, wafers, fields and sites and their baseline distributions is determined and used as inputs for the sample planning model. The work of Chien *et al.* [63] also addressed the CD measurements. The error types I and II are considered, and after inspection two decisions are possible, to accept or to reject the wafer. A cost-based heuristic for statistically determining the sampling rate is proposed.

Nurani *et al.* [66] propose a cost-based sampling methodology to allocate inspection capacity. The model aims at specifying the process operations to inspect, the number of lots, the number of wafers within a lot and a percentage area of the wafer. They discuss the possibility of implementing dynamic sampling for deteriorating processes. They point out that the problem is to find when to schedule inspections and cleans to minimize cost or to maximize the number of good dies. An extension of this work can be found in [40], where an analytic model and a genetic algorithm are proposed for inspection allocation in full production and ramp-up phases.

Williams *et al.* [47] propose a cost based sample planning optimization. The inspection cost and the associated excursions costs are considered. Once the inspection sampling plan is close to the theoretical sample plan, the inspection operations are re-allocated to support alternate inspection activities.

Table 2.1: Classification of models for inspection allocation

Article	**Model Configuration**	**Inspection Error**	**Sampling Strategy**	**Decisions**	**Solution Approach**
Raz et Kaspi [62]	Serial Multi-Stage	I and II	Static	Inspection Location	NLP and B&B
Chien et al[63]	Serial Multi-Stage	I and II	Static	Sampling rate	Heuristic
Shiau [54]	Serial Multi-Stage	I and II	Static	Inspection Location	EM and Heuristic
Kakade et al [64]	Serial Multi-Stage	Error free	Dynamic	Inspection rate	SA and B&B
Van Volsem et al[50]	Serial Multi-Stage	I and II	Static	Location, Sampling rate and inspection limits	EA and Simulation
Vaghefi et al [65]	Non-serial Multi-Stage	Error free	Static	Inspection Location	Analytical and Simulation
Naharani et Khan [53]	Non-serial Multi-Stage	Error free	Static	Inspection location	AT and Simulation
Nurani et al [66]	Non-serial Multi-Stage	I and II	Static	Sampling	Statistical
Nurani et al [4]	Non-serial Multi-Stage	I and II	Static	Sampling	Heuristic
Bean [11]	Non-serial Multi-Stage		Static	Inspection location	ILP
Elliot et al [52]	Non-serial Multi-Stage	I and II	Static	Sampling rate	Statistical
Rau et al [5]	Non-serial Multi-Stage	I and II	Static	Inspection Location	Heuristic
Gudmundsson [40]	Non-serial Multi-Stage	I and II	Static	Sampling rate and Inspection Location	GA
Emmons et Rabinowitz [67]	Non-serial Multi-Stage	Error free	Static	Sampling rate and Inspection Location	NLP
Hall et al [48]	Non-serial Multi-Stage	Error free	Static	Inspection location	Statistical and Linear cost model
Rau and Cho [68]	Non-serial Multi-Stage	I and II	Static	Inspection Location	GA
William et al [47]	Non-serial Multi-Stage	Error free	Adaptive	Sampling rate and Inspection Location	Statistical and Simulation
Mouli et al [7]	Non-serial Multi-Stage	Error free	Adaptive	Sampling	Statistical
Bousseta and Cross [49]	Non-serial Multi-Stage	I	Adaptive	Sampling rate	Statistical
Kuo [6]	Non-serial Multi-Stage	I	Adaptive	Sampling rate	Mathematical
Verduzco et al [69]	Non-serial Multi-Stage	I and II	Dynamic	Sampling based on Cost model	Heuristic
Purdy et al [60]	Non-serial Multi-Stage	Error free	Dynamic	Sampling	Rule based
Holfeld et al [70]	Non-serial Multi-Stage	Error free	Dynamic	Sampling	Statistical
Su A.J et al [61]	Non-serial Multi-Stage	Error free	Dynamic	Sampling	Mathematical
Pfeffer et al [71]	Non-serial Multi-Stage	Error free	Dynamic	Sampling	Simulation
Bettayeb et al [72] [73]	Non-serial Multi-Stage	Error free	Dynamic	Inspection location	Heuristic
Nduhura et al [42]	Non-serial Multi-Stage	Error free	Dynamic	Risk Calculation	Mathematical
Nduhura et al [74]	Non-serial Multi-Stage	Error free	Dynamic	Key Parameters for sampling	MIP
Dauzère-Pérès et al [9]	Non-serial Multi-Stage	Error free	Dynamic	Sampling, Skipping Scheduling	Mathematical
Purdy et al [75]	Non-serial Multi-Stage	Error free	Dynamic	Sampling	MILP
Hyung [76]	Non-serial Multi-Stage	Error free	Dynamic	Sampling	Mathematical
Lin et al [77]	Non-serial Multi-Stage	Error free	Dynamic	Sampling	Ruled Based
Rodriguez-Verjan et al [78]	Non-serial Multi-Stage	Error free	Dynamic	Skipping	Heuristic and B&B

B&B = Branch and Bound, EA = Evolutionary Algorithm, EM = Enumeration Method, NLP = Non Linear Programming, AT = Analytic Technique, ILP = Integer Linear Programming, GA = Genetic Algorithm, MIP = Mixed Integer Programming.

Bousetta and Cross [49] propose a method to monitor key parameters to adjust the sampling plan for effective inspection capacity utilization. The key parameters are the variance ratio, excursion frequency and normalized mean shift. Mouli *et al.* [7] propose a score mechanism based on APC and SPC to weight each lot and wafer within a lot to decide whether to sample or not. Sun and Johnson [61] also use a score mechanism which is based on weighted objectives to optimize the sampling decisions. Verduzco *et al.* [69] study a real-time Automated Visual Inspection in an electronic assembly system. They propose a model for a real-time inspection allocation based on the information gained by inspecting one additional component. They propose a model as an information maximization problem, to define a real-time inspection allocation based on the information gained if one additional component is inspected.

Purdy *et al.* [60] propose a method to release lots in metrology queue, each lot is evaluated individually and the objective is to guarantee the measurement of lots with more recent information. If two or more lots providing the same type of information are in the metrology queue and the one more recently processed is measured, the older lot (lots) can be removed. The developed application is part of a sampling system which combines a number of separate sampling rules into a single sampling decision. Holfed*et al.* [70] propose a detail of lot level and wafer level sampling application. At the lot level: the application combines several restrictions with past sampling decisions to find a well balanced lot sampling conclusion. The result is that a lot is only measured if ultimately required by a certain state, e.g. exceeding a risk threshold. At the wafer level: The wafer sampling application collects all wafer context information and incorporates historic wafer properties and APC application wafer properties. All of these contexts are associated to wafer sampling rules. These rules carry weights (violation penalty) and are then merged with historic rule violations to find a well balanced wafer sampling decision. Nduhura-Munga *et al.* [42] propose a global indicator that allows very quick and fast computations of material at risk. This indicator (named IPC, Permanent Index per Context) is a generic solution that can be used to calculate different types of risk. An industrial prototype was developed where results showed a reduction of more than 70% of material at risk compared with previous sampling strategies [12]. This index can be used to compute the Global Sampling Indicator used in the sampling, skipping and scheduling algorithm proposed by Dauzère-Pérès *et al.* [9]. Pfeffer *et al.* [71] propose a predictive sampling algorithm, they also conclude that, by using a dynamic sampling strategy, the number of required defect density measurements could be reduced by 67% compared to the previous sampling strategy. Then, Nduhura-Munga *et al.* [74] propose a mixed integer linear program to calculate key parameters required in Dynamic Sampling. Their model aims at allocating inspection capacity to manage the risk on process tools. Close to this work is the work of Bettayeb *et al.* [79]. They present an approach to control risk in two stages. The first stage consist of a minimum allocation of inspection operations that ensures a level of risk during a specific time horizon. In the second stage, additional inspection operations are allocated according to the remaining available capacity of metrology tools. Results show that the pre-allocation of inspection operations based on risk, allows the risk exposure to be balanced among the operations in the process flow. In Bettayeb *et al.* [72], an extension of their approach for allocation of inspection operations within the manufacturing route

of products is presented. Their algorithm provides a predicted quality control plan which aims at managing the risk exposure and quality control effectiveness. Then, in [73], an evaluation of their model is performed with industrial instances that consider the process tools of two workshops (i.e. 35 process tools). The results provide the maximum exposure that can be expected using the sampling algorithm of [9]. The proposed approach in this thesis aims at determining whether a set of predefined limits of risk exposure on process tools can be achieved with the available inspection capacity and if necessary, the required additional inspection capacity to meet these limits. Rodriguez-Verjan *et al.* [78] present an industrial application of skipping algorithms to effectively manage inspection queues. Concerning the problem of locating inspection operations, Hall *et al.* [48] propose a methodology to optimize based on statistical process control (SPC) model for defect excursion monitoring. They use a cost function that considers the power of the inspection, the interval between inspections and the yield impact (costs). This function is optimized for all inspection allocations in a given process flow. The proposed methodology can be used to allocate the inspections based upon the risk of yield excursions and also to estimate a return on investment of inspections. The sampling strategy is Static Sampling and the risk of adding or removing inspection operations in the flow is evaluated with a cost-of-risk function.

Narahari and Khan [53] propose an analytical model of a non-serial multi-stage manufacturing system in presence of inspections. Their model aims at predicting the mean steady-state cycle time and throughput based on a Mean Value Analysis (MVA) and under various scheduling policies. They include probabilistic routing after inspections for the cases of accept, reject or rework at some previous stages. Results show that a small number of strategically located inspection stations can perform better than a larger number of poorly located inspections. An extension of their work is presented in Rau and Cho [5]. They propose a mathematical model to optimize the location of inspections and solve it with heuristic methods. Then, Rau and Cho [68] propose a genetic algorithm to solve the problem of inspection allocations. They compare a complete enumeration method, a heuristic and a genetic algorithm. Solutions obtained with the genetic algorithm give better total profit than solutions obtained with other methods. Shin *et al.* [80] study the impact on throughput due to variation in the inspection time. They show that allocating more inspection operations helps to maintain a low bottleneck time. This is actually one of the advantages when selecting the lots dynamically. With more inspection operations, the system is more flexible to manage the risk because lots are selected when information is required while satisfying the inspection capacity constraints.

In summary, Dynamic Sampling is considered to be one of the most suitable strategies for modern high-mix semiconductor fabs to increase yield while limiting the impacts on cycle times [3]. Hence, our research work focus on new strategies for dynamic sampling that can be implemented as in [12]. Moreover, using these strategies, the inspection capacity to be allocated is not known in advance because lots are selected in real time in any of the inspection operations of their manufacturing route. Hence, new models for inspection capacity planning are proposed in chapters 5 and 6, and finally a study concerning the design of defect inspection control plan is performed in chapter 7.

Chapter 3

System Analysis and Evolution Strategy

This chapter introduces the system description and evolution from static sampling strategies (called Start Sampling) to dynamic sampling strategies (called Smart Sampling). Until recently in the site of Rousset of STMicroelectronics the selection of lots to be inspected was only done at the beginning of their manufacturing process and according to fixed sampling rates per product. However, static sampling strategies cannot handle the dynamics of the fab (i.e. tool qualifications, process flows, lot priorities). In order to reduce the number of wafers at risk on process tools, dynamic sampling has been implemented.

The evolution of the system has been possible thanks to the work of all the participants of the W@R implementation group in ST Rousset and the people of SFL department at EMSE-CMP. [1]

3.1 Introduction

Inspections are necessary to guarantee the early detection of defective products. However, more inspections do not always result in more quality. When the sampling rate increases, the yield also increases but only until to a certain point. If the sampling rate increases without taking inspection capacity into account, the yield can be negatively impacted due to longer queues of lots waiting for inspection and thus longer delays for corrective actions [56]. This is why it was decided to implement an effective sampling strategy that reduces the level of material at risk in the fab and considers tool capacity constraints. In this chapter, we present how the sampling strategy has evolved from a static sampling to dynamic sampling in the site of Rousset of STMicroelectronics.

This chapter is organized as follows, section 3.2 presents an analysis of the previous system and the motivation to change from start sampling to dynamic sampling. Section 3.3 describes the different phases required for the system evolution. Finally, section 3.4 is devoted to the concluding remarks.

[1] Sponsors: P. Campion and M. Le Gall. Leaders: E. Tartière and J. Pinaton.
ST Rousset Team: A. Thieullen, G. Rodriguez-Verjan, P. Palouar, S. Detivaud, B. Pennachio, J.C. Mattlin, F. Chairat, C. Klingelschmidt, V. Lemaire, J. De-selle, D. Courilleau, B. Mari, C. Giuliani, D. Viard.
EMSE-CMP SFL Team: S. Dauzère-Pérès, C. Yugma, J.L. Rouveyrol, S. Housseman.

3.2 System Analysis

At the beginning of the project, only the start sampling strategy was used at the site of Rousset of STMicroelectronics to select lots for defect inspection. The main characteristics of this strategy are listed below:

- Lots are selected at the beginning of their manufacturing process.
- A sampling rate is used to determine the ratio of lots that are selected. The sampling rate is defined as "$1/N$", "N" being the number of lots that are considered at risk.
- The selected lots systematically visit all the inspection operations defined in their defect inspection control plan.

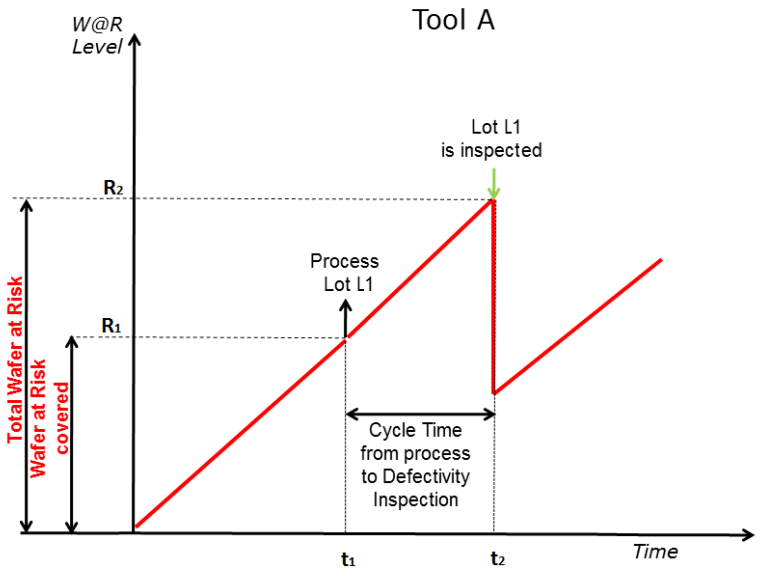


Figure 3.1: W@R Counter of a process tool

In order to manage the number of wafers at risk on process tools an indicator, called "Wafers at Risk" (W@R) was introduced. The W@R is the number of wafers processed on a tool since the process date of the latest inspected lot. The W@R represents the number of wafers that are potentially impacted if a problem occurs. It can be calculated according to different contexts, such as: process tool, recipe, type of control, technology. Figure 3.1 illustrates the evolution of the W@R on a process tool. The $y-$axis represents the W@R level and the $x-$axis the time. Each time a lot is processed on the tool, the W@R is incremented and is decremented when the results of an inspection are obtained. Let us assume that Lot L1 is inspected at the next inspection operation. When Lot L1 is processed on the process tool at time t_1, the risk that it can cover is the current W@R value (R_1). This value is called "*W@R covered*". Then, Lot L1 goes through all the intermediate process operations of its route before arriving at the next defect inspection operation. Meanwhile, the W@R level increases because the tool processes other lots. When Lot L1 is inspected at time t_2, the new W@R level is the previous W@R level (R_2) minus the

W@R covered by Lot L1 (R_1). By observing the wafers at risk in the fab, several cases of overcontrol and lack of control on process tools were found. Figure 3.2 presents the results of the W@R on tools in the same area during 4 weeks. It can be observed that tools 5, 6 and 7 are over controlled because their average W@R levels are well below the expected average W@R level. This means that most of the selected lots to be inspected were processed on these tools. On the other hand, tools 8 and 13 are not sufficiently controlled.

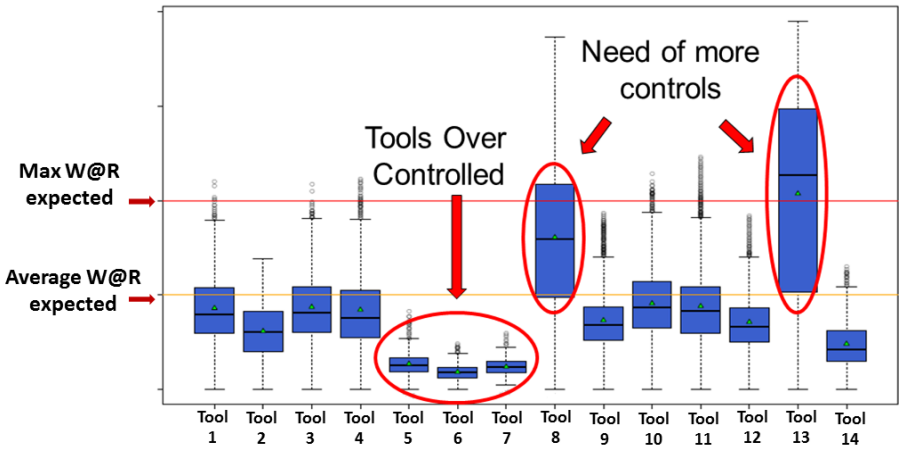


Figure 3.2: W@R Levels (on industrial data)

In order to reduce those cases of lack of control and over control on process tools, an analysis of the sampling strategies was performed. The objective was to check whether the static sampling (called start sampling) strategy could reduce the W@R on tools above a given limit (i.e. $100 * L$) with the current inspection capacity. Data are normalized for confidentiality reasons. Experiments were performed using industrial data on the Smart Sampling Scheduling and Skipping Simulator (S5) developed by the EMSE-SFL[2]. Results are presented in figure 3.3. The x-axis is the available capacity in the defect inspection area (data are normalized with the value "A"), the y-axis is the average of the maximum W@R value for all the tools in the fab during one month of production. The red trend (i.e. the first from the top) is the W@R obtained when static sampling (i.e. start sampling) is used and the blue trend is the W@R obtained when dynamic sampling (i.e. smart sampling) is used. It can be observed that with the same available capacity, the difference changes from 60% to 24% (i.e. from capacity A to 2,5*A). When the capacity is reduced, the impact on the risk highly increases with static sampling. In other words, if the defect inspection capacity changes due to unexpected events (e.g. breakdowns), the risk increases because the selected lots systematically visit the inspection operations of their routes. They will be directed to the inspection area even if they cannot be inspected. Moreover, lots can be redundant in terms of W@R information because they may have been processed on the same tools.

With dynamic sampling, the system also depends on the defect inspection capacity but is less sensitive to unexpected events. If the capacity changes, the risk will be adapted

[2]EMSE-SFL: École de Mines de St-Etienne, Département Science de la Fabrication et Logistique

because the selection of lots is done in real time and according to the available capacity. Moreover, the redundant lots in terms of W@R will not be selected or will be skipped from the defectivity inspection queue. Using both strategies the maximum value of W@R was not lower to $100 * L$. In order to reduce the exposure on process tools, additional inspection operations were required coupled with a dynamic selection of lots.

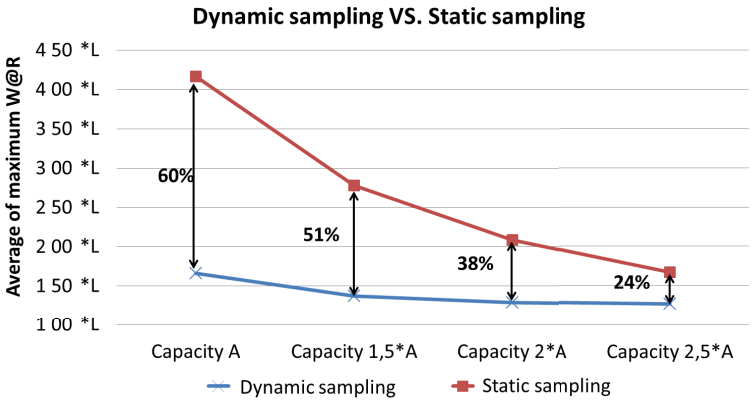


Figure 3.3: Average of maximum W@R obtained with static sampling vs. dynamic sampling

The implementation strategy was divided in three main phases that are listed below.

1. Phase I. In order to follow the wafers at risk on process tools, counters of W@R were implemented on each process tool.

2. Phase II. Once the W@R counters were operational, a mechanism to balance the measurable lots (i.e. lot which recipe exists in defect inspection) over the process tools was implemented. This mechanism is called Dispatching for Sampling [81] and the objective was to guarantee that all process tools process at least one measurable lot .

3. Phase III. An algorithm to optimize the selection of the best set of lots to measure was developed.

The mechanism to follow the W@R counters (Phase I) is the same as the one described in figure 3.1 (section 3.2). The next section describes Phase II and introduces the Phase III. Chapter 4 details the algorithms used in the dynamic selection of lots based on a skipping mechanism (Phase III).

3.3 System Evolution

Phase II - Dispatching for Sampling. Based on the W@R counters and two limits (i.e. Warning Limit - WL and Inhibit Limit - IL), the operator is informed of when it is necessary to process a measurable lot on a given tool. The Inhibit Limit (IL) is the

maximum value of acceptable W@R on the process tool. The Warning Limit (WL) is the limit after which actions have to be taken in order to control the situation. These limits have been calculated taking into account the current defect inspection capacity, the cycle time to reach an inspection operation and the throughput rates of process tools. More details concerning the optimization of control limits can be found in [74]. The system to dispatch the lots that are later inspected is described in figure 3.4. Each time the W@R level is updated, several conditions are tested and the information is gathered and presented with the color code below.

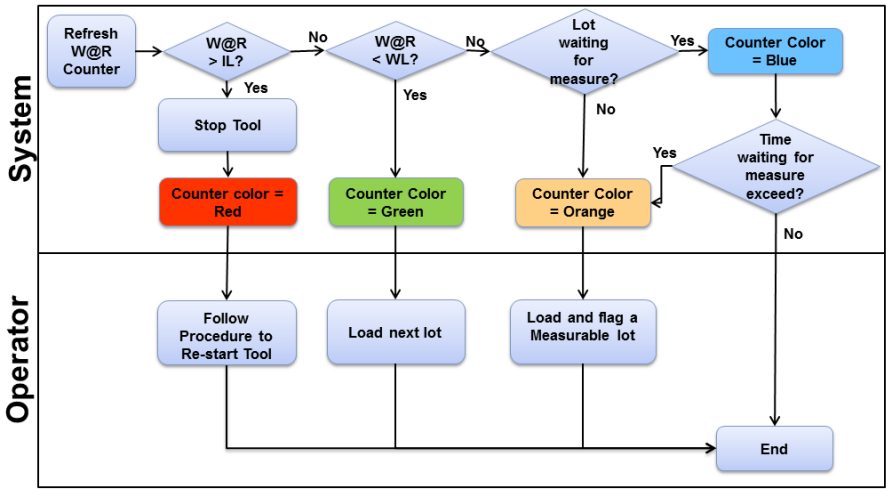
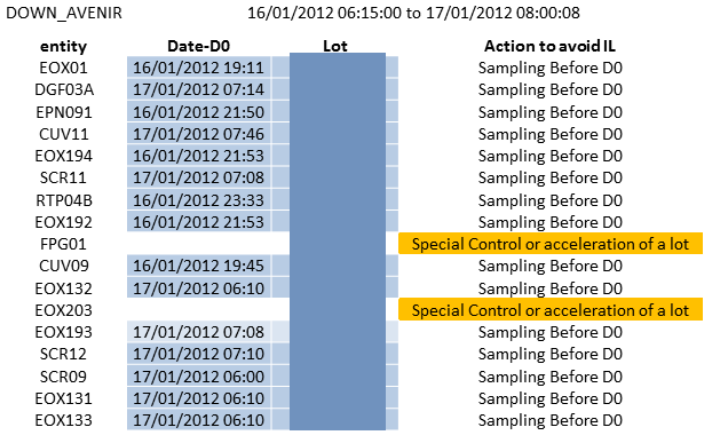


Figure 3.4: Flow diagram of the dispatching for sampling system

- *Green*: The W@R level is considered as normal. No specific actions are necessary.
- *Orange*: The W@R level is larger than the Warning Limit (WL). Hence, the operator should load on the process tool a lot that is measurable in the defectivity area. If the lot has been selected in previous process operations it is more suitable to be processed on the tool. Note that the lot can also reduce the risk on several process tools from previous operations, which is defined with the coverage block of inspection operations.
- *Blue*: The W@R level is larger than the WL, but a lot has already been selected and the results of the inspection are expected.
- *Red*: The W@R level achieved the Inhibit Limit (IL). Therefore, the situation is critical and the tool is automatically stopped. Once a tool is stopped a Quality Task (QT) is performed. A QT is a special control performed on non-product wafers (NPW), they are used for qualification, testing, stabilization of process performance, tool qualification or other purposes [82].

Phase III - Dynamic sampling. Due to the dynamics of the fab, when lots are selected in front of process tools, some events cannot be controlled (e.g. selected lots may be stopped, the time to get to the inspection is longer than expected, breakdown of

inspection tools). Moreover, dispatching and sampling efforts can be useless if the arrival rate of selected lots results in either too high or too small workload of inspection tools. Figure 3.5 shows a list of process tools with a W@R level close to the IL. In order to avoid that the W@R level reaches the IL, two actions are possible, either to accelerate the already selected lots that reduce the W@R on the process tools or to select a new lot whose next operation is an inspection operation. In most of the cases, it is more simple and efficient to select a new lot which next operation in the manufacturing route can be an inspection operation. The reason is that the resulting W@R reduction is significant (more than 50% of the current W@R) and the time to get the inspection results only depend on the waiting time in the inspection area. In order to make the sampling system more reactive and effective, dynamic sampling algorithms were implemented. Chapter 4 details the algorithms that were developed to optimize the selection of lots to inspect.



DOWN_AVENIR 16/01/2012 06:15:00 to 17/01/2012 08:00:08

entity	Date-D0	Lot	Action to avoid IL
EOX01	16/01/2012 19:11		Sampling Before D0
DGF03A	17/01/2012 07:14		Sampling Before D0
EPN091	16/01/2012 21:50		Sampling Before D0
CUV11	17/01/2012 07:46		Sampling Before D0
EOX194	16/01/2012 21:53		Sampling Before D0
SCR11	17/01/2012 07:08		Sampling Before D0
RTP04B	16/01/2012 23:33		Sampling Before D0
EOX192	16/01/2012 21:53		Sampling Before D0
FPG01			Special Control or acceleration of a lot
CUV09	16/01/2012 19:45		Sampling Before D0
EOX132	17/01/2012 06:10		Sampling Before D0
EOX203			Special Control or acceleration of a lot
EOX193	17/01/2012 07:08		Sampling Before D0
SCR12	17/01/2012 07:10		Sampling Before D0
SCR09	17/01/2012 06:00		Sampling Before D0
EOX131	17/01/2012 06:10		Sampling Before D0
EOX133	17/01/2012 06:10		Sampling Before D0

Figure 3.5: Report of process tools with W@R levels close to their IL

Finally, when changes are introduced in organizations, one of the most important aspects to tackle is the fear of change. In order to overcome this, the different participants should have a clear understanding of the benefits of the new system [83]. Moreover, a strong leadership, commitment and participation of top level management is often a key for a successful implementations of new systems [84]. Table 3.1 shows the defined key metrics of the project that address the goals of the different participants.

Table 3.1: Key metrics of the project

Item	Service Department of Participants
Scrap Reduction	Process and Yield Control
W@R Inhibit Limit Reduction	Process and Yield Control
Quality Task (QT) Reduction	Production
Deployment on Process Tools	Engineering, Device and Defectivity

3.4 Conclusions

In this chapter we described how, in the site of Rousset of STMicroelectronics the sampling system for defect inspection changed from static sampling to dynamic sampling. Figure 3.6 summarizes the implemented applications for dynamic sampling. The dispatching for sampling guides operators in selecting the lots that will later be inspected in defect inspection. This ensures that a process tool that needs to be controlled will process a measurable lot (i.e. a lot for whose recipe exists in defect inspection). Moreover, the W@R reduction of a lot can be enhanced throughout its route until the next defect inspection operation. However, dispatching and sampling efforts can be lost if the arrival rate results in an unbalanced workload of inspection tools. In addition, while a selected lot is waiting for inspection, the production state changes and it may no longer be interesting to inspect this lot. This is why, a skipping mechanism for lots in defect inspection area was implemented. In summary, dispatching for sampling and smart sampling and skipping are complementary strategies that enable the system to be more effective.

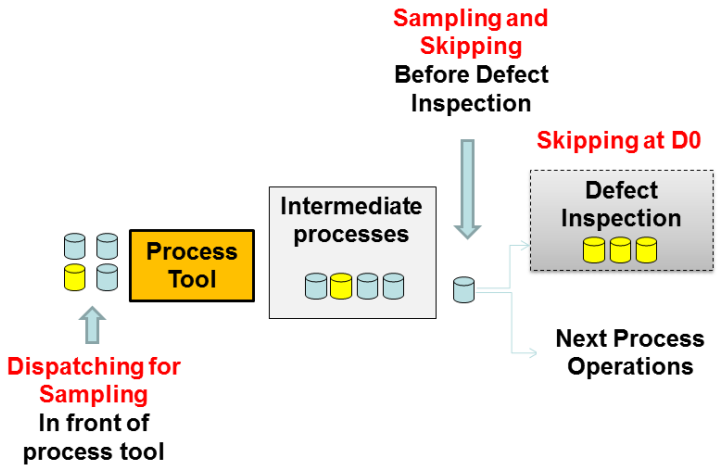


Figure 3.6: Implemented applications for dynamic sampling

With a new methodology to sample lots and control the risk on process tools, new challenges arose at different decision levels (e.g. Operational, tactical and strategic). At the operational level, the problem of how to optimize the selection of lots to be inspected is addressed in chapter 4. At the tactical and strategic levels, the problem of planning the capacity of the defect inspection area is addressed with a capacity model proposed in chapters 5 and 6. Finally, how the defect inspection control plan impacts the W@R on process tools is illustrated in chapter 7.

Chapter 4

Dynamic Selection of Lots for Defect Inspection

In this chapter a new methodology to manage defect inspection queues is presented[1]*. This work is an extension of [9] and [12] and is based on their proposed risk indicator (i.e. Global Sampling Indicator, GSI). The objective is to identify the lots that can be removed from the inspection queue with limited impacts on the global risk of the fab. Significant industrial results have been obtained by implementating this skipping algorithms*[2]*.*

4.1 Introduction

In chapter 3 we presented how the sampling system changed from static sampling to dynamic sampling in the site of Rousset of STMicroelectronics. In this chapter we describe the algorithms used for selecting the lots that can remain in the inspection queue. The proposed approach is a skipping mechanism of lots that are already in the inspection area but can be released from inspection because the production state changed. It is based on the Global Sampling Indicator (GSI) proposed in [9] and an indicator of Risk Increase (RI) inspired from the Lot Scheduling Indicator (LSI) proposed in [12]. These indicators are used to identify the impact of inspecting or not a lot or a set of lots on the global risk of the fab.

The chapter is organized as follows: Section 4.3 presents the developed algorithms to skip lots that are already in the waiting queue. The decision to skip a lot or a set of lots is taken by checking whether inspecting a lot do not reduce the global risk, or if lots have redundant information compared with other lots in the waiting queue. Numerical experiments on industrial instances are presented and discussed in section 4.4. The industrial implementation of the skipping mechanism is discussed in section 4.5. Finally, concluding remarks and perspectives are presented in section 4.6 .

[1] Part of this chapter was published in the 9th International Conference on Modeling and Analysis of Semiconductor Manufacturing (included in the 2013 Winter Simulation Conference) [78]

[2] Special thanks to Sylvain Housseman for his participation in the industrial implementation of these algorithms.

4.2 Problem description and solution approach

As described in section 1.4 . Two control limits are used to manage the W@R on process tools: The Warning Limit (WL) and the Inhibit Limit (IL). Figure 4.1 recalls how the W@R evolves on a process tool and illustrates why controlling a lot does not necessarily reduce the W@R of a process tool. Let us consider that lot L1 is processed before lot L2 on the same process tool (i.e. Tool 1). Suppose that, after a while, both lots are waiting in the inspection queue. Lot L2 is inspected first and the W@R of the process tool is thus reduced by the number of wafers that were processed on the tool before processing lot L2. When lot L1 is inspected, the W@R is not decreased because the information of W@R reduction was already obtained with lot L2. Hence, lot L1 can be removed from the inspection queue without impacting the W@R on process tool 1. Due to several situations in the fab (e.g. long waiting times in inspection queues, arrival of new lots with more recent information, special controls performed on process tools), some lots that were previously selected for inspection can be removed with limited impacts on the global risk.

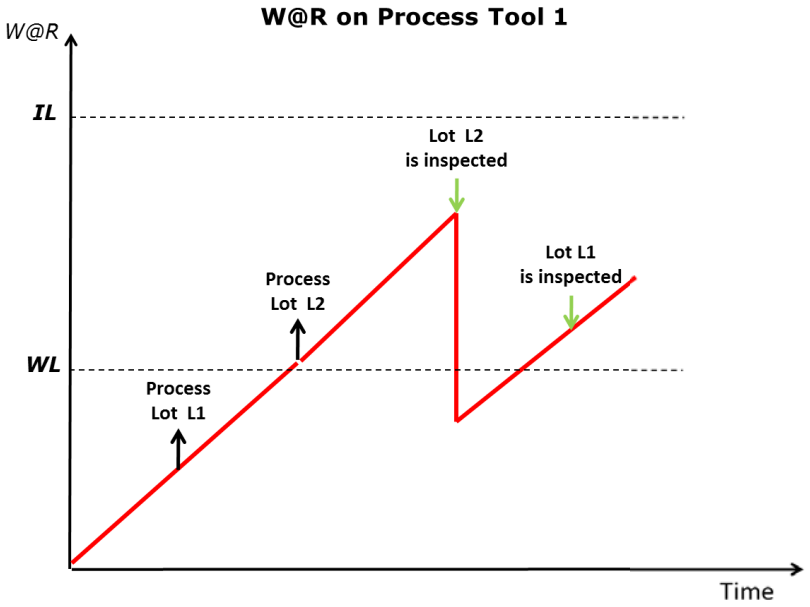


Figure 4.1: Wafers at risk evolution on process tool 1

Our problem is to identify which lots can be released from inspection due to redundant information in terms of W@R. We use the Global Sampling Indicator (GSI) to evaluate the risk in the fab when a set S of lots are waiting to be inspected or being inspected. Let us recall the following notations from [9]:

- T: Number of considered process tools.
- IL_t: Inhibit Limit for tool t.
- RV_t: Current risk value on tool t. In our case it is the W@R on process tool t
- $G_{t,l}$: Gain on risk of tool t if lot l is inspected.
- $NRV_{t,l}$: New risk value of tool t if lot l is inspected, i.e. $NRV_{t,l} = RV_t - G_{t,l}$. In our case, it is the new value of Wafers at Risk (NW@R).

- $NRV_t(S)$: New risk value of tool t if lots in set S are inspected. It is calculated as follows:

$$NRV_t(S) = \min_{l \in S} NRV_{t,l}$$

- α: Parameter used to give more or less emphasis on getting as far as possible from Inhibit Limits.

The GSI is calculated as follows:

$$GSI(S) = \sum_{t=1}^{T} \left(\frac{NRV_t(S)}{IL_t} \right)^{\alpha} \tag{4.1}$$

Through simulation, Nduhura-Munga *et al.* [12] studied the impact of the parameter α on the performance of the GSI-based sampling algorithms. They observed that satisfactory results are achieved by setting $\alpha = 6$. We use a Risk Increase (RI) indicator to evaluate the associated risk of not inspecting a lot or a set of lots. The smaller the value of RI, the less important is the lot or the set of lots. When the RI is calculated for a single lot, it is equal to the LSI (Lot Scheduling Indicator) proposed in [12]. Let us suppose that there are 3 lots L_1, L_2 and L_3 in the waiting queue of the defect inspection area. To define the impact of skipping a single lot, three combinations are evaluated. These combinations are obtained by removing each lot from the initial set of lots (S):

1. $GSI(S) = GSI(\{L_1, L_2, L_3\})$
2. $RI(L_1) = GSI(S \setminus \{L_1\}) - GSI(S) = GSI(\{L_2, L_3\}) - GSI(\{L_1, L_2, L_3\})$
3. $RI(L_2) = GSI(S \setminus \{L_2\}) - GSI(S) = GSI(\{L_1, L_3\}) - GSI(\{L_1, L_2, L_3\})$
4. $RI(L_3) = GSI(S \setminus \{L_3\}) - GSI(S) = GSI(\{L_1, L_2\}) - GSI(\{L_1, L_2, L_3\})$

When a set of lots is removed simultaneously, the RI is associated to that set of lots and not only to each lot removed independently. For instance, if two lots are going to be removed, the combinations that need to be evaluated are:

1. $RI(\{L_1, L_2\}) = GSI(S \setminus S\{L_1, L_2\}) - GSI(S) = GSI(L_3) - GSI(\{L_1, L_2, L_3\})$.
2. $RI(\{L_2, L_3\}) = GSI(S \setminus S\{L_2, L_3\}) - GSI(S) = GSI(L_1) - GSI(\{L_1, L_2, L_3\})$.
3. $RI(\{L_1, L_3\}) = GSI(S \setminus S\{L_1, L_3\}) - GSI(S) = GSI(L_2) - GSI(\{L_1, L_2, L_3\})$.

A threshold named T_{Metro} is used to decide whether a lot or a set of lots can be skipped. It can be interpreted as the maximum risk value that can be tolerated for degrading the initial GSI. It can also be seen as the minimum gain in terms of GSI that a lot or a set of lots should bring to remain in the waiting queue.

In order to understand the setting of T_{Metro}, let us consider the GSI of a single tool which Inhibit Limit is 750 wafers. Figure 4.2 shows the GSI values for different W@R values and using $\alpha = 6$. Let us suppose that the current W@R value is near the Inhibit Limit (e.g. 650 wafers). If the T_{Metro} is equal to 0.5 it represents a maximum risk value of 90 wafers that is tolerated to degrade the GSI. If T_{Metro} is equal to 0.3 the GSI can be

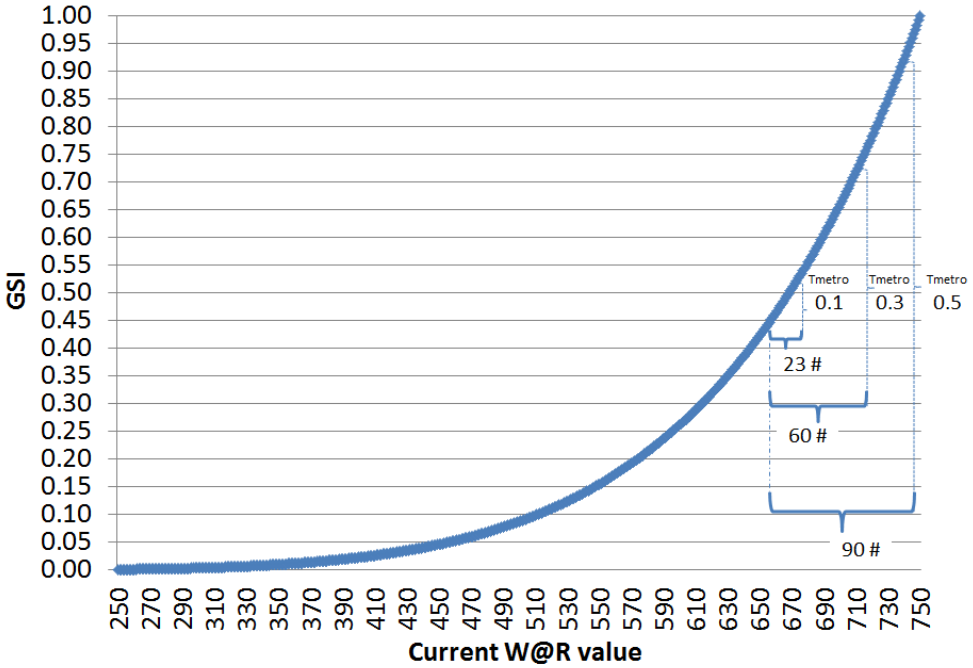


Figure 4.2: Impact of differentT_{Metro} values

degraded of 60 wafers, if the T_{Metro} is equal to 0.1 the GSI can be degraded by 23 wafers, and if T_{Metro} is equal to 0.01 the GSI can be degraded by 2.5 wafers. Hence, the smaller the value of T_{Metro}, the more restricted is the system for skipping lots. Aside from the RI indicator to decide if a lot or a set of lots can be skipped, the defect inspection team needs to define additional rules that can be temporal or permanent in order to focus on measuring a particular group of lots with a given attribute. This information is used to create a subset of lots that can skip the inspection. The next section describes the five algorithms that we developed for the skipping mechanism.

4.3 Skipping Mechanism

This section summarizes the implemented algorithms. The objective is to maximize the number of skipped lots while satisfying the threshold (T_{Metro}). When equivalent solutions are found in terms of number of lots to skip, the solution that minimizes the impact on the global risk is selected (i.e. $min(RI)$). The algorithms are listed according to their complexity and their capability to obtain solutions of better quality. Let us consider the following notations:

- $S_{Initial}$: Set of lots already in the inspection queue,
- $S_{Skippable}$: Set of lots that are in the inspection queue and can be skipped $S_{Skippable} \subset S_{initial}$. When there are not additional rules for not skipping lots, $S_{Skippable} = S_{Initial}$,
- $S_{SkipList}$: Set of lots identified for skipping,

- S^* : Set of lots remaining in the inspection queue, $S^* = \{S_{Initial} \setminus S_{SkipList}\}$,
- T_{Metro}: Maximum risk value tolerated for degrading the initial GSI. It can also be seen as the minimum gain in terms of GSI that a lot or a set of lots should bring to remain in the waiting queue.

4.3.1 Algorithm 1 - Identification of Skippable Lots

For each lot l in set $S_{Initial}$ the RI is calculated. Those lots whose RI value is lower than or equal to the threshold T_{Metro} can potentially be skipped.

Algorithm 1 Identification of Skippable Lots

1: $S^* = S_{initial}$
2: $S_{SkipList} = \emptyset$
3: **for** each lot $l \in S_{initial}$ **do**
4: $RI(l) = GSI(S_{initial} \setminus \{l\})$ - $GSI(S_{initial})$
5: **end for**
6: **for** each lot $l \in S_{initial}$ **do**
7: **if** $RI(l) < T_{Metro}$ **then**
8: $S^* = \{S^* \setminus \{l\}\}$
9: $S_{SkipList} = \{S_{SkipList} \cup \{l\}\}$
10: **end if**
11: **end for**
12: return $S_{SkipList}$

Note that $GSI(S_{initial})$ is calculated considering that all lots in the waiting queue are inspected. Then, the impact of removing a lot is compared to the same initial situation. Consider an example where, only the lots L_1 and L_2 are in the waiting queue and both reduce the W@R on the same process tools. The evaluation of RI is as follows:

1. $RI(L_1) = GSI(S_{initial} \setminus \{L_1\}) - GSI(S_{initial}) = GSI(\{L_2\}) - GSI(\{L_1, L_2\})$
2. $RI(L_2) = GSI(S_{initial} \setminus \{L_2\}) - GSI(S_{initial}) = GSI(\{L_1\}) - GSI(\{L_1, L_2\})$

When RI(L_1) is evaluated, the lot L_2 is in the waiting queue and, by removing L_1, the GSI is not impacted because the information is obtained with L_2. When RI(L_2) is calculated, the lot L_1 is in the waiting queue, hence by skipping L_2 the GSI is not impacted. If the final decision of whether to skip or not is taken with this algorithm, both lots (L_1 and L_2) would be released. Therefore, the final decision of skipping lots cannot be taken with this algorithm. However, the list of lots than can potentially skip the inspection operation is reduced. If the RI of a given lot is strictly larger than T_{Metro}, it means that not enough W@R reduction can be obtained with the inspection of the remaining lots.

4.3.2 Algorithm 2 - Local Evaluation

The RI for each lot in the set $S_{Initial}$ is calculated. The decision to skip a lot or not is done immediately after the calculation of the RI. Therefore, once a lot is identified for skipping, it is not considered for the evaluation of the remaining lots.

Algorithm 2 Local Evaluation

1: $S^* = S_{initial}$
2: $S_{SkipList} = \emptyset$
3: **for** each lot $l \in S_{initial}$ **do**
4: $RI(l) = GSI(S^* \setminus \{l\})$ - $GSI(S_{initial})$
5: **if** $RI(l) < T_{Metro}$ **then**
6: $S^* = \{S^* \setminus \{l\}\}$
7: $S_{SkipList} = \{S_{SkipList} \cup \{l\}\}$
8: **end if**
9: **end for**
10: return $S_{SkipList}$

Note that the RI of all lots is compared to the same initial situation ($GSI(S_{initial})$). The main difference between algorithm 1 (Identification of Skippable Lots) and algorithm 2 (Local Evaluation) is that the list of the remaining lots in the waiting queue is updated when a lot is identified for skipping. Let us consider that lots L_1, L_2 and L_3 are in the waiting queue ($S_{initial} = \{L_1, L_2, L_3\}$). Lots L_1 and L_2 reduce the W@R on the same process tools. Then, the evaluation of RI is as follows:

1. $RI(L_1) = GSI(S_{initial} \setminus \{L_1\}) - GSI(S_{initial}) = GSI(\{L_2, L_3\}) - GSI(\{L_1, L_2, L_3\})$. Since lot L_2 is in the waiting queue lot L_1 is immediately skipped. When the RI of the remaining lots is determined, the L_1 is not considered in the waiting queue.
2. $RI(\{L_2, L_1\}) = GSI(S_{initial} \setminus \{L_2, L_1\}) - GSI(S_{initial}) = GSI(\{L_3\}) - GSI(\{L_1, L_2, L_3\})$
3. $RI(\{L_3, L_1\}) = GSI(S_{initial} \setminus \{L_3, L_1\}) - GSI(S_{initial}) = GSI(\{L_2\}) - GSI(\{L_1, L_2, L_3\})$

Since the waiting queue is updated when a lot is identified for skipping, the final decision can be taken with this algorithm. The drawback is that the quality of the resulting solution depends on the list order. In this example, lot L_1 is evaluated first and therefore is skipped. If L_2 is evaluated first, it would be skipped.

4.3.3 Algorithm 3 - Greedy

The RI for each lot l in set $S_{Initial}$ is calculated. The lot l with the smallest RI is identified and, if $RI(l) < T_{Metro}$ this lot is immediately skipped. Then, the RI of the remaining lots are recalculated. This procedure is performed until the lot with the smallest RI cannot be skipped.

Algorithm 3 Greedy

1: $S^* = S_{initial}$
2: $S_{SkipList} = \emptyset$
3: $l' = \arg\min RI\{S^*\}$
4: **while** $RI(l') < T_{Metro}$ **do**
5: $S^* = \{S^* \setminus \{l'\}\}$
6: $S_{SkipList} = \{S_{SkipList} \cup \{l'\}\}$
7: **for** each lot $l \in S^*$ **do**
8: $RI(l) = GSI(S^* \setminus \{l\})$ - $GSI(S_{initial})$
9: **end for**
10: $l' = \arg\min RI\{S^*\}$
11: **end while**
12: return $S_{SkipList}$

The main difference between algorithm 2 (Local Evaluation) and algorithm 3 (Greedy) is that the lots with smallest RI are skipped first. Therefore, the order of the list does not influence the final solution. Let us consider three lots L_1, L_2 and L_3 and assume that lots L_1 and L_2 reduce the W@R on the same process tools, but lot L_1 has more information in terms of W@R reduction compared to L_2. In the first iteration, all the combinations are evaluated:

1. $RI(L_1) = GSI(S_{initial} \setminus \{L_1\}) - GSI(S_{initial}) = GSI(\{L_2, L_3\}) - GSI(\{L_1, L_2, L_3\})$
2. $RI(L_2) = GSI(S_{initial} \setminus \{L_2\}) - GSI(S_{initial}) = GSI(\{L_1, L_3\}) - GSI(\{L_1, L_2, L_3\})$
3. $RI(L_3) = GSI(S_{initial} \setminus \{L_3\}) - GSI(S_{initial}) = GSI(\{L_1, L_2\}) - GSI(\{L_1, L_2, L_3\})$

If the smallest RI is obtained with lot L_2, it is skipped and the waiting queue is updated. Therefore, the combinations that are evaluated in the next iteration are the following:

1. $RI(\{L_1, L_2\}) = GSI(S_{initial} \setminus \{L_1, L_2\}) - GSI(S_{initial}) = GSI(\{L_3\}) - GSI(\{L_1, L_2, L_3\})$
2. $RI(\{L_3, L_2\}) = GSI(S_{initial} \setminus \{L_3, L_2\}) - GSI(S_{initial}) = GSI(\{L_1\}) - GSI(\{L_1, L_2, L_3\})$

With this algorithm, the identification of lots to skip is improved compared to algorithm 2. However, once a lot is skipped, the decision is not reviewed. The objective of the next algorithm is to improve the solutions with a local search.

4.3.4 Algorithm 4 - Add-Remove

This algorithm is based on the Greedy algorithm. Once the solution obtained with Greedy a local search is performed. Each lot of the solution ($S_{SkipList}$) is removed from the list and all the candidates (S^*) are evaluated to replace it. If the lot with the smallest RI is different from the lot that was removed, the procedure restarts with the modified list. The procedure stops when the list ($S_{SkipList}$) is not modified and when a new lot cannot be added.

Algorithm 4 Add-Remove

1: $S^* = S_{initial}$
2: $S_{SkipList} = Algorithm3(S_{initial})$
3: **repeat**
4: **for** $l \in S_{SkipList}$ **do**
5: remove l from $S_{SkipList}$ and replace it with k=arg min $RI\{S^* \setminus S_{Skiplist}\}$
6: **if** $k \neq l$ **then**
7: Restart evaluation with an updated list of $S_{SkipList}$ and S^*
8: **end if**
9: **end for**
10: **until** A new lot is added to the $S_{SkipList}$
11: return $S_{SkipList}$

Let us consider four lots in the waiting queue, i.e. $S_{Initial} = \{L_1, L_2, L_3, L_4\}$ and assume that the greedy algorithm identify lots L_2 and L_3 to skip and the last added lot was L_3. The local search procedure tries to replace the previous lots identified for skipping (i.e. L_2). Therefore, the evaluated combinations are:

1. $RI(\{L_1, \boldsymbol{L_3}\}) = GSI(S_{initial} \backslash \{L_1, \boldsymbol{L_3}\}) - GSI(S_{initial}) = GSI(\{L_2, L_4\}) - GSI(\{L_1, L_2, L_3, L_4\})$
2. $RI(\{L_2, \boldsymbol{L_3}\}) = GSI(S_{initial} \backslash \{L_2, \boldsymbol{L_3}\}) - GSI(S_{initial}) = GSI(\{L_1, L_4\}) - GSI(\{L_1, L_2, L_3, L_4\})$
3. $RI(\{L_4, \boldsymbol{L_3}\}) = GSI(S_{initial} \backslash \{L_4, \boldsymbol{L_3}\}) - GSI(S_{initial}) = GSI(\{L_1, L_2\}) - GSI(\{L_1, L_2, L_3, L_4\})$

Let us assume that $RI(\{L_4, L_3\}) < RI(\{L_2, L_3\})$. Then, lot L_2 is replaced by L_4 and the evaluation restarts with an updated list. Hence, the procedure reevaluates the previous decisions (i.e. L_3) and the following combinations are considered:

1. $RI(\{\boldsymbol{L_4}, L_1\}) = GSI(S_{initial} \backslash \{\boldsymbol{L_4}, L_1\}) - GSI(S_{initial}) = GSI(\{L_2, L_3\}) - GSI(\{L_1, L_2, L_3, L_4\})$
2. $RI(\{\boldsymbol{L_4}, L_2\}) = GSI(S_{initial} \backslash \{\boldsymbol{L_4}, L_2\}) - GSI(S_{initial}) = GSI(\{L_1, L_3\}) - GSI(\{L_1, L_2, L_3, L_4\})$
3. $RI(\{\boldsymbol{L_4}, L_3\}) = GSI(S_{initial} \backslash \{\boldsymbol{L_4}, L_3\}) - GSI(S_{initial}) = GSI(\{L_1, L_2\}) - GSI(\{L_1, L_2, L_3, L_4\})$

Let us assume that $RI(\{L_4, L_3\})$ has the smallest value. Then, the local search procedure stops, because the last added lot is the same that the one to replace (i.e. L_3). Since the set $S_{SkipList}$ changed, the algorithm will try to add a new lot (i.e. a solution with three lots). If a new lot is added, the local search starts again but with the set of three

lots. If a new lot cannot be added, the algorithm stops. With this algorithm, the solutions obtained from the Greedy can be improved. However, each lot is evaluated individually and different solutions can be obtained if sets of lots are considered. The next algorithm aims at finding the best set of lots to skip.

4.3.5 Algorithm 5 - Branch and Bound

The RI is calculated for each lot in $S_{Initial}$. Lots are sorted by increasing RI and a Branch and Bound method is applied. Bounds consider both the number of lots that can be skipped and the RI of the solution. Let us consider the set of lots $S_{SkipList}$ as the local solution and the set $S_{Solution}$ as the best solution found. A recursive function is implemented to explore the nodes. The index of the evaluated lot in the list $S_{Skippable}$ is expressed with the parameter IndexLot.

Algorithm 5 Branch & Bound

1: Order the set of lots $S_{initial}$ by increasing RI
2: $S_{Solution} = \emptyset$
3: $RI(S_{Solution}) = T_{Metro} + 1$
4: Explore$(S_{Solution}, 0, S_{Initial})$
5: return $S_{Solution}$

Function Explore $(S_{SkipList}, IndexLot, S_{Initial})$

1: Evaluation of $S_{SkipList}$
2: **if** $S_{SkipList}$ is better than $S_{Solution}$ **then**
3: $S_{Solution} = S_{SkipList}$
4: $nb_{Solution} = |S_{SkipList}|$
5: **for** NextIndex=IndexLot+1 until $|S_{SkipList}|$ **do**
6: **if** $S_{Solution}$ can be improved in terms of number of lots or RI **then**
7: Explore$(\{S_{SkipList} \cup \{l_{NextIndex}\}\}, NextIndex, S_{initial})$
8: **else**
9: Stop branching on this node

The handling of a search tree node stops if the solution cannot be improved in terms of number of lots or resulting RI. The number of the remaining lots to evaluate in a node is calculated as the total number of lots in $S_{Initial}$ minus the value of $IndexLot$ (i.e. $|S_{Initial}| - IndexLot$). Hence, if the number of lots in the local solution ($S_{SkipList}$) plus the remaining number of candidate lots is smaller than the number of lots in the best solution ($S_{Solution}$) the search stops; because we know that the current solution cannot be improved to contain as many lots as the best known solution. The second condition to stop the branching of a node corresponds to the evaluation of the RI. The gap in terms of number

of lots between the local solution ($S_{SkipList}$) and the best solution ($S_{Solution}$) is calculated. If the RI of the local solution ($S_{SkipList}$) plus the RI of lots considered in the gap is larger than T_{Metro}, the search on that branch is stopped (i.e. $RI(S_{SkipList}) + \sum_{i=indexLot+1}^{indexLot+gap+1} RI(l_i) > T_{Metro}$).

4.3.6 Emergency Mode

In practice, there can be unexpected events (e.g. breakdown of tools) that change the available capacity in defect inspection area. Hence, the total number of lots in the waiting queue cannot be inspected. These versions of the skipping algorithms can be used when the engineer would like to define the number of lots to skip.

In the previous algorithms, T_{Metro} is considered as a parameter which is used as a to satisfy for the Risk Increase indicator. In the emergency mode, the number of lots to skip is defined in advance and the algorithm identifies the set of lots to skip with minimum impact on the overall risk.

Algorithm 6 Branch & Bound Emergency Mode

1: Order the set of lots $S_{initial}$ by increasing RI
2: $S_{Solution} = \emptyset$
3: $RI(S_{Solution}) = 999999$
4: Explore$(S_{Solution}, 0, S_{Initial})$
5: return $S_{Solution}$

Function Explore $(S_{SkipList}, IndexLot, S_{Initial})$

1: Evaluation of $S_{Solution}$
2: **if** $S_{SkipList}$ is better than $S_{Solution}$ **then**
3: $S_{Solution} = S_{SkipList}$
4: $nb_{Solution} = |S_{SkipList}|$
5: **for** each lot l in the subset $S_{Initial}$ from indexLot until last lot of $S_{Initial}$ **do**
6: **if** $S_{Solution}$ can be improved in terms of number of lots or RI **then**
7: explore$(\{S_{SkipList} \cup l\}, index(l), S_{initial})$
8: **else**
9: Do not explore the branch

In this version of the branch and bound skipping algorithm, the conditions to stop the handling of a search tree node are different. The number of the remaining lots to evaluate in a node is calculated as the total number of lots in $S_{Initial}$ minus the value of $IndexLot$ (i.e. $|S_{Initial}| - IndexLot$). Hence, if the number of lots in the local solution ($S_{SkipList}$) plus the remaining lots to explore in a node is smaller than the number of lots to skip, the search tree node stops. If the RI of the local solution ($S_{SkipList}$) considering the remaining

lots of the branch is larger than the RI of the best solution ($S_{Solution}$), the search tree node is also stopped.

The section below presents a numerical example of the skipping mechanism. More numerical experiments are presented in section 4.4.

4.3.7 Numerical Example

In the following, we illustrate the proposed algorithms using a numerical example. Let us suppose that there are 5 lots in the defect inspection queue. Table 4.1 shows the W@R reductions that can be obtained with the inspection of each lot. The column W@R represents the current W@R level of process tools. The column NW@R shows the new risk value if the lot is inspected and the column IL gives the value of the Inhibit Limit on the associated process tool. As previously mentioned, a lot can reduce the W@R of several process tools. It depends on the product and the coverage block of the inspection operation. In this example, if lot L1 is inspected, the value of W@R on process tools 07, 08 and 12 is reduced.

Table 4.1: Example of lots waiting to be inspected

Lot	Process Tool	W@R	NW@R	IL
L1	Tool 07	960	481	1100
L1	Tool 08	948	486	1100
L1	Tool 12	625	425	2500
L2	Tool 05	179	104	500
L2	Tool 06	622	349	1200
L3	Tool 03	82	56	500
L3	Tool 04	79	52	500
L3	Tool 06	622	274	1200
L3	Tool 07	960	456	1100
L4	Tool 08	948	462	1100
L4	Tool 11	737	274	2500
L5	Tool 01	226	104	500
L5	Tool 02	31	1	500
L5	Tool 06	622	299	1200
L5	Tool 09	306	293	1100
L5	Tool 10	302	290	1100
L5	Tool 12	625	425	2500

Table 4.2 gives the GSI for the initial set of lots and the RI for each lot. Let us consider that $T_{Metro} = 0.007$. If all lots are inspected, $GSI(S_{initial}) = 0.01159$. If lot L1 is removed from the queue, $GSI(S_{initial} \setminus \{L1\}) = 0.01159$ and $RI(L1) = 0.000$, thus lot L1 can be skipped. This is because, although the $W@R$ on process tools (i.e. Tools 7, 8 and 12) can be reduced by inspecting L1, the $W@R$ on the same tools is also reduced by inspecting other lots that are in the queue (i.e. L3, L4 and L5). If lot L5 is removed from the queue, $GSI(S_{initial} \setminus \{L5\})$ would be 0.002024 and $RI(L5) = 0.00865$. Since $RI(L5)$ is larger than T_{Metro}, L5 cannot be skipped because it is the only lot that reduces the risk on tools 1, 2, 9 and 10. It is important to note that the RI has been calculated only with one

Table 4.2: Example of GSI and RI calculations (Iteration 1)

	$(NW@R/IL)^{\alpha}$					
Tools	$S_{initial}$	$\{S_{initial} \setminus \{L1\}\}$	$\{S_{initial} \setminus \{L2\}\}$	$\{S_{initial} \setminus \{L3\}\}$	$\{S_{initial} \setminus \{L4\}\}$	$\{S_{initial} \setminus \{L5\}\}$
Tool 1	0.00008	0.00008	0.00008	0.00008	0.00008	0.00853
Tool 2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Tool 3	0.00000	0.00000	0.00000	0.00002	0.00000	0.00000
Tool 4	0.00000	0.00000	0.00000	0.00002	0.00000	0.00000
Tool 5	0.00008	0.00008	0.00211	0.00008	0.00008	0.00008
Tool 6	0.00014	0.00014	0.00014	0.00024	0.00014	0.00014
Tool 7	0.00507	0.00507	0.00507	0.00699	0.00507	0.00507
Tool 8	0.00549	0.00549	0.00549	0.00549	0.00744	0.00549
Tool 9	0.00036	0.00036	0.00036	0.00036	0.00036	0.00046
Tool 10	0.00034	0.00034	0.00034	0.00034	0.00034	0.00043
Tool 11	0.00000	0.00000	0.00000	0.00000	0.00066	0.00000
Tool 12	0.00002	0.00002	0.00002	0.00002	0.00002	0.00002
GSI	0.01159	0.01159	0.01361	0.01363	0.01419	0.02024
RI	0.00000	0.00000	0.00202	0.00204	0.00260	0.00865

Table 4.3: Example of RI calculations

(a) Iteration 2

Lot	*RI*	Decision
L2	0.002024	Skip
L3	0.436901	Not Skip
L4	0.404892	Not Skip
L5	0.008865	Not Skip

(b) Iteration 3

Lot	*RI*	Decision
L2	–	–
L3	0.438925	Not Skip
L4	0.406916	Not Skip
L5	0.010890	Not Skip

iteration. If Algorithm 1 is used to skip lots that have their RI smaller than T_{Metro}, then the final decision would be to skip lots L1, L2, L3 and L4. However, when the RI of lot L2 is calculated, lots L3 and L5 are in the queue. When the RI of L3 is calculated, lots L1, L2 and L5 are in the queue. Thus, skipping simultaneously all the lots with $RI < T_{Metro}$ can lead to uncontrolled (and thus undesirable) situations. However, with Algorithm 1 we can establish that lot L5 will not be skipped. Therefore, the number of lots to evaluated is reduced, which reduce the number of iterations for the other algorithms. In the following, the mechanism of Algorithm 3 is explained. In each iteration, the lot with the smallest RI is identified. In this example, lot L1 is skipped (i.e. $RI(L1) = 0.000$) and the RI of the remaining lots is recalculated (See Table 4.3). In the second iteration, lot L2 has the smallest RI (i.e. $RI(L2) = 0.002024$). Since $RI(L2)$ is smaller than T_{Metro}, lot L2 can be skipped. In the third iteration, the smallest RI is obtained with lot L5 which cannot be skipped because $RI(L5)$ is larger than T_{Metro}. Therefore, the final decision is to skip lots L1 and L2.

Table 4.4: Example of RI calculations for sets of lots

Set of Lots	RI
L1, L2, L4	0.406916
L1, L2, L5	0.010890
L1, L3, L4	0.841793
L1, L3, L5	0.446132
L1, L4, L5	0.413757
L2, L3, L4	0.006673
L2, L3, L5	0.031869
L2, L4, L5	0.013273
L3, L4, L5	0.013660

Different lots may be selected when computing RI for sets of lots rather than evaluating lots individually. Table 4.4 gives the resulting RI when sets of lots are considered. It can be observed that the set of lots {L2, L3, L4} has the smallest RI (0.006673) hence, it can be skipped. Let us note that, compared to the previous solution, lot L1 is not skipped. The reason is that, when $RI(L1)$ is calculated, lots L3 and L4 are in the queue and if L1 is skipped $RI(L3)$ and $RI(L4)$ increase. When the RI is calculated for sets of lots, it is preferable to keep lot L1 in the queue and to skip lots L3 and L4. Then, evaluating sets of lots for skipping performs better than evaluating each lot individually. The complexity of the problem was not deeply studied due to time considerations. Evidence (calculation time and structure of the problem) let us think that it is a particular instance of a quadratic assignment problem, indicating NP complexity. The decision about which lot should be skipped depends on an underlying location problem. However, a deep study could be interesting. Next section presents the results on industrial instances.

4.4 Numerical Results and Discussion

The algorithms were developed with the R software [13]. The computational experiments in this section compare the efficiency of the different algorithms on a set of 3 industrial instances. The results of 12 industrial instances are presented in appendix B . Table 4.5 details the results for different values of T_{Metro} (0.001, 0.005, 0.01, 0.05 and 0.1).

With Algorithm 1 the lots for which $RI \leq T_{Metro}$ in the first iteration are identified. As previously mentioned, skipping all those lots can lead to uncontrolled situations. As it can be observed, the resulting RI on all the instances is larger than T_{Metro}. However, the list of lots that can be skipped is reduced, which reduces the number of possible combinations to be evaluated for the other algorithms, and therefore, the calculation times are also reduced. Algorithm 2 only screens once the list of lots and, each time a lot is identified for skipping, the lot is immediately removed. With this algorithm T_{Metro} is satisfied. However, it works as a blind search because solutions highly depend on the list order. In finds the best solution only in 42% of the analyzed cases (i.e. 60 scenarios). With algorithm 3 solutions are improved compared with algorithm 2. For example, the solution for instance 1 and $T_{Metro} = 0.05$ has a larger number of lots and a smaller RI (i.e. 12 lots and $RI = 0.04545$

Table 4.5: Number of skipped lots and final RI for different values of T_{Metro}

Instance	TypeAlgo	**Number of skipped lots** (Related RI)				
		T_{Metro}				
		0.001	0.005	0.01	0.05	0.1
1	Algorithm 1	**6** (0.001279)	**12** (0.284293)	**13** (0.291487)	**19** (0.435451)	**22** (0.722157)
	Algorithm 2	**5** (0.000653)	**7** (0.004930)	**7** (0.008011)	**10** (0.049931)	**14** (0.095832)
	Algorithm 3	**5** (0.000626)	**7** (0.004930)	**8** (0.008628)	**12** (0.045453)	**15** (0.098124)
	Algorithm 4	**5** (0.000626)	**7** (0.004930)	**8** (0.008628)	**12** (0.035398)	**15** (0.093280)
	Algorithm 5	**5** (0.000626)	**7 (0.004219)**	**8 (0.007871)**	**12** (0.035398)	**15** (0.093280)
2	Algorithm 1	**4** (0.070563)	**10** (0.131550)	**11** (0.167974)	**13** (0.194752)	**15** (0.312755)
	Algorithm 2	**3** (0.000882)	**5** (0.004762)	**7** (0.007275)	**10** (0.047746)	**11** (0.084170)
	Algorithm 3	**3** (0.000115)	**6** (0.003970)	**7** (0.006507)	**10** (0.046978)	**11** (0.083402)
	Algorithm 4	**3** (0.000115)	**6** (0.003970)	**7** (0.006507)	**10** (0.043171)	**11** (0.079595)
	Algorithm 5	**3** (0.000115)	**6** (0.003970)	**7** (0.006507)	**10** (0.043171)	**11** (0.079595)
7	Algorithm 1	**7** (0.001262)	**13** (0.396196)	**14** (0.401685)	**18** (0.522087)	**19** (0.595998)
	Algorithm 2	**6** (0.000512)	**7** (0.004225)	**8** (0.009715)	**12** (0.045763)	**14** (0.099489)
	Algorithm 3	**6** (0.000385)	**7** (0.001262)	**9** (0.009661)	**12** (0.038781)	**14** (0.081529)
	Algorithm 4	**6** (0.000385)	**7** (0.001262)	**9** (0.009661)	**12** (0.037227)	**14** (0.081529)
	Algorithm 5	**6** (0.000385)	**8 (0.004317)**	**9 (0.008450)**	**12** (0.037227)	**14** (0.081529)

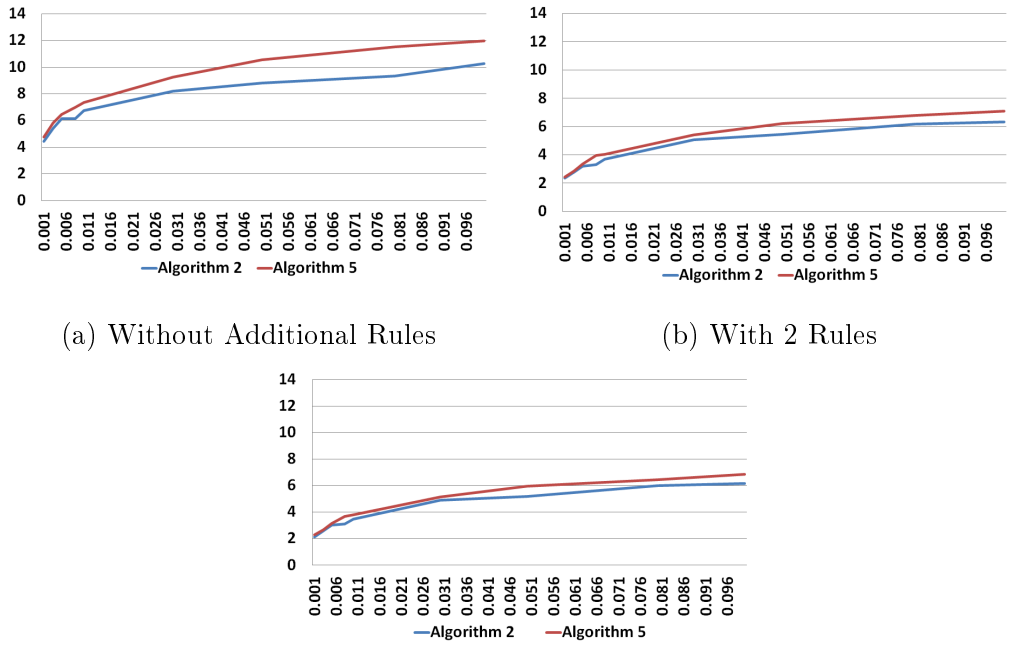
Table 4.6: Average Calculation Time (sec)

	T_{Metro}				
Type Algorithm	0.001	0.005	0.01	0.05	0.1
Algorithm 1	0.347	0.287	0.336	0.311	0.284
Algorithm 2	0.338	0.281	0.335	0.306	0.277
Algorithm 3	1.194	1.471	2.072	2.392	2.732
Algorithm 4	2.717	5.184	7.905	12.358	17.493
Algorithm 5	0.316	8.544	17.510	439.952	2250.984

vs. 10 lots and $RI = 0.04993$). Algorithm 3 finds the best solution in 82% of the analyzed cases. However, lots are evaluated individually and once a lot is selected for skipping, previous decisions will not be reconsidered. These decisions are improved with algorithm 4 trough a local search. Actually, it can find the best solution in 92% of the analyzed cases, but it is still myopic since lots are evaluated individually. Algorithm 5 is an exact method and thus finds the optimal solutions. For example, considering instance 1 and $T_{Metro} = 0.005$, the same number of skipped lots is obtained but the resulting RI is smaller (i.e. 7 lots for skipping with $RI = 0.004219$). It also finds solutions with more lots while satisfying the parameter T_{Metro} (i.e. Instance 7 with $T_{Metro} = 0.005$, 8 lots to skip with $RI = 0.004317$).

Table 4.6 presents the average calculation times for the different algorithms. When T_{Metro} increases, the calculation time increases. This is explained by the fact that more combinations are possible because more lots can be skipped (i.e. The size of $S_{Skippable}$ increases with the value of T_{Metro}). When the set $S_{Skippable}$ is reduced, algorithm 5 gives the best solutions in a reasonable time. Algorithm 4 is a good trade-off between quality of solutions and calculation time when the set $S_{Skippable}$ increases.

Another case in which the set $S_{Skippable}$ is reduced is when additional rules are defined. During manufacturing, the defectivity group may focus on controlling a specific group of lots according to a given attribute (i.e. technology, product, process operation). Hence, a predefined set of additional rules (temporal or permanent) are defined. The problem associated to increasing the number of rules is that the number of lots that can be selected by the skipping algorithms is reduced. Figure 4.3 shows the average number of lots identified for skipping with two algorithms (i.e. Algorithms 2 and 5) and when additional rules are considered. In Figure 4.3a there are not additional rules, the only criterion to select lots is the impact on the global risk of the fab. It can be observed that the larger T_{Metro}, the larger the difference between algorithms 2 and 5. The reason is that more lots can be considered in the set $S_{Skippable}$ hence more combinations are evaluated. When additional rules are considered, the number of lots that can be skipped decreases. In figures 4.3b and 4.3c, two and four additional rules are considered respectively. It can be observed that the difference between the algorithms is reduced. Hence, the quality of solutions is dominated by the number of rules that are considered. The same conclusions can be drawn for sampling or skipping, when the number of rules increases it is difficult to find solutions that satisfy all the rules [8].



(a) Without Additional Rules

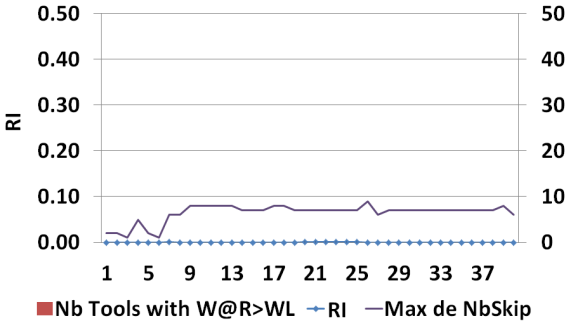
(b) With 2 Rules

(c) With 4 Rules

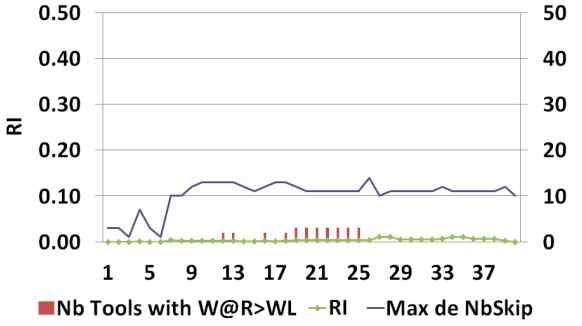
Figure 4.3: Average number of skipped lots with algorithm 2 and 5

In order to evaluate the emergency mode algorithm (see section 4.3.6), a set of industrial instances were selected, associated to cases where the defect inspection area was saturated. In the emergency mode, the engineer defines the number of lots to skip and the algorithm

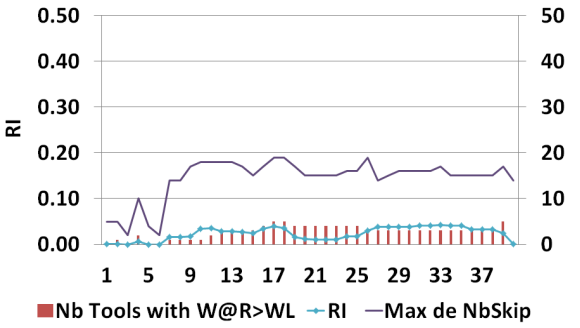
identifies the set of lots with the minimum resulting RI. In order to define the number of lots to skip, the number of lots that exceed capacity was calculated and four cases were tested. In the first one, only 30% of the lots exceeding capacity were skipped. In the second case, only 50% of lots were skipped, in the third and fourth cases, the 70% and 100% of lots exceeding capacity were skipped.



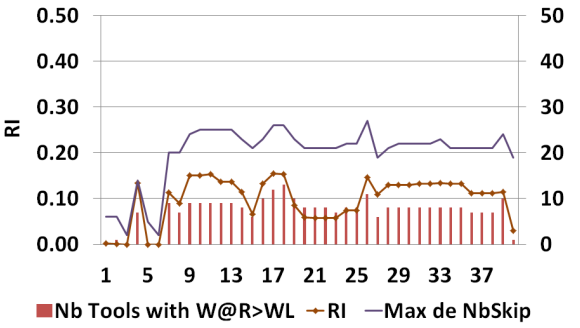
(a) Skip 30% of lots exceeding capacity



(b) Skip 50% of lots exceeding capacity



(c) Skip 70% of lots exceeding capacity



(d) Skip 100% of lots exceeding capacity

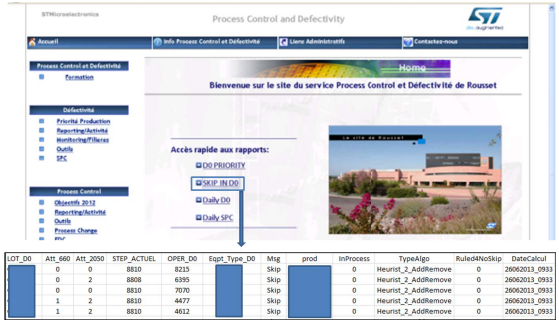
Figure 4.4: Resulting RI for Emergency Mode Algorithm

Figure 4.5 shows the results for the different instances. The first line from the bottom shows the resulting RI and its y-axis on the left side. The second line from the bottom is the number of skipped lots and the bar graph is the resulting number of process tools whose $W@R$ would remain over the WL (i.e. Warning Limit). The axis for both graphs is the y-axis on the right side. Figure 4.4a shows the results when 30% of the lots exceeding capacity are skipped. It can be observed that for the studied instances, between 1 and 9 lots can be skipped without impacting the risk of the fab (i.e. $RI = 0.0$). Moreover, there are not process tools whose W@R remain over their WL. When 50% of lots exceeding capacity are skipped, the risk is not impacted, except on some instances in which $RI = 0,015$, the maximum number of process tools which $W@R > WL$ is 3. The larger the number of skipped lots, the higher is the impact on the risk and the potentially impacted tools. When 70% of lots exceeding capacity are skipped, the RI varies from 0.02 to 0.04, and the maximum number of process tools whose $W@R > WL$ is 5. Finally, when all the lots

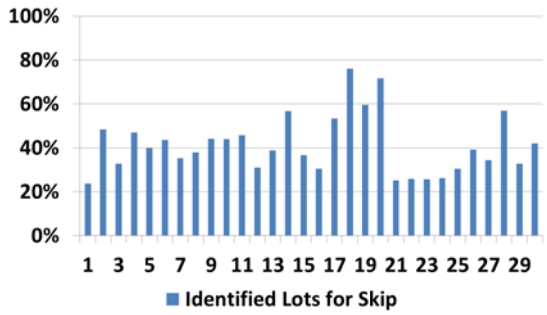
that exceed capacity are skipped, the impacts on the risk level (i.e. RI) varies from 0 to 0.15, with a maximum number of 13 tools that remain over their WL. This information is key to the engineer when deciding whether to skip or not a given number of lots, since it shows the impacts on the global risk of the fab and the possible impacted tools. It is important to note that the results presented for the Emergency Mode algorithm depend on the current production state and the information that can be obtained by inspecting each lot. It also shows how the decisions of skipping or not directly influence the productivity on the other production areas by avoiding process tools to be stopped due to W@R control. Hence, if the sampling strategy is not adapted to the current available capacity at the defect inspection area, it can generate situations where the W@R on process tools will reach the IL leading to process tools to stop.

4.5 Industrial Implementation

Industrialization of new solutions is always a challenge, since it is necessary to change habits and work methods. With the previous strategy to select lots for inspection (i.e. Static Sampling) the number of arriving lots at defect inspection area could not be handled. Moreover, in cases in which skipping lots was necessary, it was difficult to identify which lots to skip, and also if a given set of lots was skipped the risk associated to the fab activity. Hence, a set of rules were defined and lots that satisfied those rules were not skipped. The drawback of this approach is that when the number of rules increases, the remaining lots that can be skipped is reduced. When an arbitrary number of rules are used, it is difficult to find solutions that satisfy all of them [8].



(a) Skip report



(b) Percentage of lots identified for skipping

Figure 4.5: Manual Skipping tool

With the introduction of W@R counters and of the dynamic selection of lots for inspection, the identification of lots to skip has changed. A lot that looked interesting for inspection due to its W@R information, may not longer be interesting after a while due to changes in the production state (e.g. arrival of new lots with more recent information, long waiting times in inspection, tools that are stopped or special controls performed on

process tools). Hence, it is essential to identify the lots with redundant information that can skip inspection.

The first prototype for skipping lots was included in the S5 Simulator (Smart Sampling Skipping Scheduling Simulator) [85] developed by EMSE-SFL [3] during the IMPROVE [4] project and was used to compare different sampling policies using historical data. In this simulator, the skipping decisions were based on events. Each time a lot was evaluated to be sampled, the skipping of a lot was considered if there was not available space in the waiting queue. For the industrialization phase, the decision of whether to skip a lot or not cannot be done based on events due to automation constraints. Therefore, the skipping application is triggered once every period of time, e.g. once every hour. Results of the first version of the skipping application were published in the web site of Process Control of the site of Rousset of STMicroelectronics and also sent via e-mail. Figure 4.5a illustrates how the report was presented. With this information, the defectivity engineer considered if the list of proposed lots had to be skipped or not. Figure 4.5b presents the percentage of identified lots for skipping over the total inspected lots in defectivity area during 30 days. In the period during which the skipping process was performed manually, all the identified lots were not always skipped. Hence, the number of lots identified for skipping is over-estimated due to lots with redundant information. For example, assume that lots L_x and L_y are in the waiting queue and that they have redundant information (i.e. they reduce the W@R on the same process tools). If L_x is identified for skipping and is actually inspected. The next time that the application runs, it will propose the lot L_y for skipping. Hence, the results show two lots identified for skipping but, in reality, it would be only one. Although the observations overestimate the potential gains, they still show a large potential improvement when managing of defect inspection queues. Hence, automatic skipping of lots were adopted in the defect inspection area.

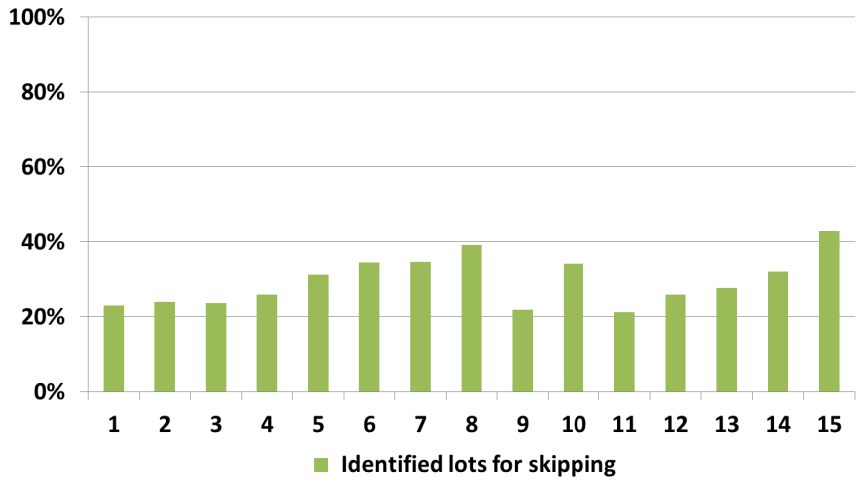


Figure 4.6: Results of automatic Skipping

[3] EMSE-SFL: École de Mines de St-Etienne, Département Sciences de la Fabrication et Logistique

[4] IMPROVE: **I**mplementing **M**anufacturing science solutions to increase equi**P**ment p**RO**ducti**V**ity and fab p**E**rformance

Figure 4.6 shows the results related to using the automatic selection of lots for skipping during for several weeks. Between 20% and 40% of lots were removed from the inspection queues, allowing the inspection of lots with more relevant information in terms of W@R. The gains in terms of number of lots to inspect motivated the development of an integrated application that considers the sampling and skipping decisions simultaneously [14].

4.6 Conclusion and perspectives

In this chapter we presented a new approach for managing defect inspection queues. The objective is to identify the lots that can skip inspection with limited impacts on the overall risk of the fab. Various skipping algorithms are proposed and evaluated on industrial instances. The Add-Remove and Branch&Bound algorithms (algorithms 4 and 5) give the best solutions compared with the other algorithms. However, the calculation time for algorithm 5 quickly increases with the number of lots to consider. Even-though algorithm 4 does not give in all the cases the optimal solution, 92% of the analyzed cases were solved to optimality. Hence, Algorithm 4 provides a good trade-off between quality of solutions and calculation time when the number of lots to consider increases.

This skipping mechanism has been implemented and is currently being used in the site of Rousset of STMicroelectronics. Gains in terms of a better selection of lots to inspect have been obtained and the work methods changed. Thanks to the new system, between 20% and 40% of lots were removed from inspection with limited impacts on the global risk of the fab. Moreover, because of the implementation of the W@R on process tools and the mechanism to evaluate the pertinence of measuring sets of lots (i.e. by using GSI and RI), the potential impacts of skipping lots or not can now be directly linked to the fab productivity. This can be expressed in terms of saved Quality Tasks (QT) incurred when process tools are stopped.

These encouraging results have motivated the improvement of the application and various research directions have been identified. Sampling and Skipping decisions are interrelated. The sampling of lots depends on the available capacity and the information obtained with lots already in the inspection queue. The skipping decision is influenced by the arrival of new lots sampled for inspection. Therefore, considering both decisions simultaneously is the next step to improve the selection of lots. Moreover, considering inspection times, instead of number of lots, should help to improve the workload on inspection tools.

Chapter 5

Allocation of Defect Inspection Capacity

With the evolution of the sampling strategy, from static sampling to dynamic sampling, new models are required to estimate the associated inspection capacity. The models proposed in this chapter aim at determining whether a set of predefined W@R limits (i.e. Warning limits and Inhibit limits) can be satisfied and, if not, the additional inspection capacity that is required to keep the W@R on process tools below the limits[1].

5.1 Introduction

With the introduction of dynamic sampling a new model to estimate the required defect inspection capacity is necessary. In this chapter we are interested in the dual problem of [42]. They propose a mixed integer linear program to determine the optimal value of W@R limits on process tools according to a fixed inspection capacity. Their model is aggregated and does not consider the details of the manufacturing routes. With the models proposed in this chapter we aim at determining whether a set of predefined W@R limits can be satisfied and, if not, the additional inspection capacity that is required to keep the W@R on process tools below the limits. Our model takes into account all the process operations of the manufacturing routes, the qualifications of process and inspection tools, and the design of defect inspection control plans.

The chapter is structured as follows: the problem is detailed in section 5.2. Section 5.3 presents the proposed mathematical model. Sections 5.4 and 5.5 introduce two improved versions of the model that include more industrial details. This chapter ends with a description of the industrial implementation in section 5.6 and the concluding remarks in section 5.7.

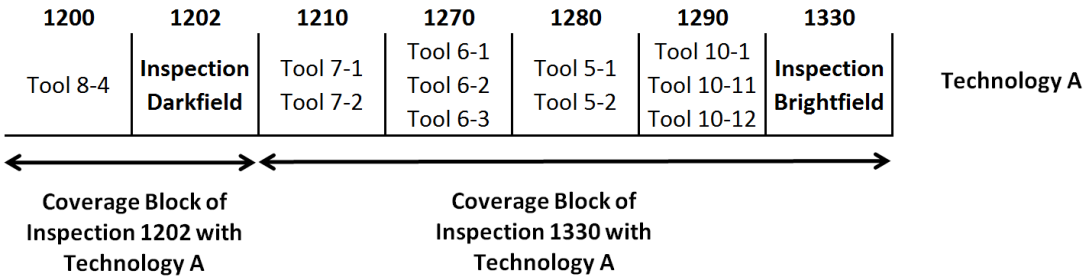
5.2 Problem Description

When fabs use static sampling to select lots for inspection, the defect inspection capacity is allocated in advance via the sampling rates of lots. The sampled lots are systematically

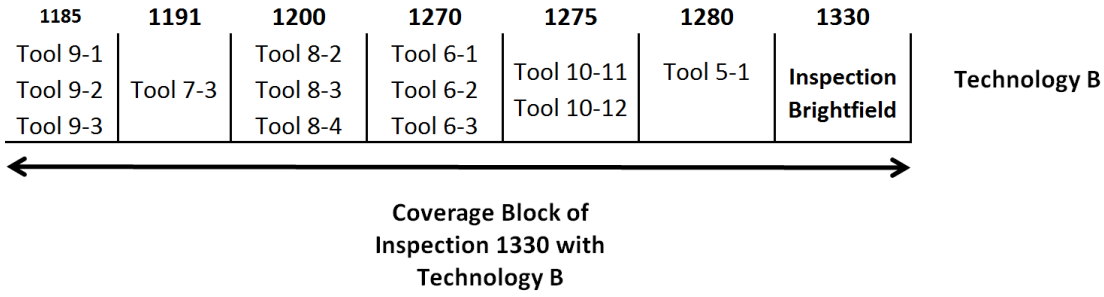
[1]Part of this chapter was published in the 8th International Conference on Modeling and Analysis of Semiconductor Manufacturing (included in the 2012 Winter Simulation Conference) [86] and was submitted to the international journal of Computers & Operations Research

inspected in all the inspection operations defined in their defect inspection control plan. With the inspection of the same wafers at different inspection operations, it is possible to identify whether defects were recently added or not since the last inspection operation. However, by only using static sampling, the levels of W@R on process tools cannot be correctly managed. With dynamic sampling strategies, the W@R on process tools can be better controlled but the defect inspection capacity to be allocated is not known in advance. The reason is that lots are selected in real time in any of the inspection operations of their routes. The problem addressed in this chapter is to estimate the defect inspection capacity necessary to satisfy a given set of W@R limits. Because the W@R refers to the number of wafers processed on the tool since the last inspected lot, the lots that are selected for inspection do not require “pre" and “post" inspection results.

In the model proposed in this chapter, the following factors are considered because they directly impact the W@R on process tools: Production routes, defect inspection control plans, qualifications of process and inspection tools, W@R limits, mix and volume of products.



(a) Route Technology A



(b) Route Technology B

Figure 5.1: Exemple of the manufacturing routes for profucts of Technologies A and B

- **Manufacturing Route:** It refers to the sequence of process operations that are necessary to obtain the final product. The manufacturing route depends on the specifications of the technology to produce (i.e. type of devices). Products from the same technology may have similar manufacturing route, but not equal because it

depends on the specification of the product design. Lots of a given product follow the same sequence of process operations. Figure 5.1 shows a portion of the manufacturing routes for products of technologies "A" and "B". All lots from a product of technology "A" (figure 5.1a) have to be processed on operation 1200 and then go through operations 1210 to 1290. Lots from a product of technology "B" are processed through process operations 1185 to 1280 (figure 5.1b). The process tools that can process lots of a given product are defined through recipe qualifications.

- **Tool Qualifications:** The qualification of process tools refers to a kind of setup that assures the right conditions for the process (e.g. right temperature, metal composition, gas pressure). Each process operation is associated with a recipe. The qualification of recipes on several process tools allows the workload to be better allocated. However, not all the process tools of a toolset can be qualified on the same process operations. This is due to technical restrictions (e.g. machine hardware or software restrictions) and also because performing a qualification is costly and time consuming. Therefore, only a limited number of qualifications are performed [37], [87]. The qualification on inspection tools refers to the set of instructions or control parameters that are necessary to inspect a wafer. Figure 5.1 is a representation of the route of two products from Technologies “A” and “B”. The products of both technologies have some common process operations, but the qualifications of process tools are different. For example, the products of Technology "B" can only be performed on tool 5-1 in process operation 1280. Whereas, for the products of Technology "A", the same process operation 1280 can be performed on two process tools (i.e. Tools 5-1 and 5-2). Concerning the qualification of inspection tools, in our example, the inspection operation 1202 can be performed with a Darkfield tool and the inspection operation 1330 can be performed with a Brightfield tool. Qualifications have to be considered in the model because they define which products can be processed or controlled on which tool throughout the manufacturing route.

- **Defect Inspection Control Plan**: It defines the position and coverage of defect inspection operations within the manufacturing route (see section 1.3.3). The coverage block refers to the process operations that can be controlled with a given inspection operation. In figure 5.1a the process tools that are qualified to perform operation 1200 can be controlled with lots inspected in inspection operation 1202 of Technology "A". The process tools that are qualified for operations 1210 to 1290 can be controlled with lots inspected in inspection operation 1330 of Technology "A". This implies that, by inspecting one lot, several process tools can be controlled. In addition, process tools can be qualified to perform several operations and thus can be controlled through several inspection operations.

- **Product mix and production volumes**: The manufacturing route and the defect inspection control plan are defined by product. Hence, the product mixes and the volumes must be considered because they directly affect the resulting W@R on process tools.

- **W@R Limits on process tools (i.e. WL and IL):** The W@R limits are key

parameters to manage the W@R on process tools. WL and IL play different roles and co-exist in the dynamic sampling strategy. They define the maximum number of wafers at risk that is acceptable on each process tool. Hence, with a continuous reduction of W@R limits, the frequency of lots to inspect increases.

5.3 Defect Inspection Capacity Planning - Model 1

This section introduces a Linear Program (LP) to calculate the requirements in terms of number of inspections to ensure W@R limits on process tools. The model allocates the production volumes on process tools while taking into account their qualifications. Then, the number of inspections that need to be performed on a process tool is calculated as the total processed volume divided by the warning limit. Using this information, the model determines where to allocate inspections and how many to allocate considering the predefined control plans. Process tools are modeled individually because qualifications and W@R limits are defined for each process tool (see figure 5.1). Inspection tools are modeled by type because all inspection tools of the same type are identical. This assumption is valid since the type refers to the detection capability and inspection technology (e.g. darkfield, brightfield).

In order to estimate the capacity required for W@R, an initial percentage of capacity is reserved. If it is not enough, additional capacity will be required. In this version of the model it is supposed that, once a lot is inspected, the W@R on the process tools that were covered is set to zero.

Sets

I : Set of products indexed by i,

P : Set of process operations indexed by p,

C : Set of inspection operations indexed by c,

K : Set of inspection tools type indexed by k,

T : Set of process tools indexed by t.

Parameters

V^i : Total production volume of product i,

WL_t : Warning Limit of process tool t,

IL_t : Inhibit Limit of process tool t,

h_p^i : Product route,

$$= \begin{cases} 1 \text{ if product } i \text{ is processed on process operation } p, \\ 0 \text{ otherwise.} \end{cases}$$

$hh^i_{p,t}$: Qualification of process tools,

$$= \begin{cases} 1 \text{ if process tool } t \text{ is qualified to process product } i \text{ on process operation } p, \\ 0 \text{ otherwise.} \end{cases}$$

$b^i_{c,k}$: Qualification of inspection tool k on inspection operation c,

$$= \begin{cases} 1 \text{ if the inspection operation } c \text{ of product } i \text{ is qualified on inspection tool type } k, \\ 0 \text{ otherwise.} \end{cases}$$

R_k : Reserved Capacity for $W@R$ measures on inspection tool type k.

Decision Variables

$X^i_{p,t}$: Production volume of product i processed on tool t in process operation p,

$Y^i_{c,p,t}$: Number of inspections of product i in inspection operation c, that covers process tool t of process operation p,

Z^i_c : Total number of inspections performed in inspection operation c of product i,

A_k : Additional capacity required for inspection tool type k.

The objective is to minimize the additional defect inspection capacity required to satisfy the Warning Limits on process tools.

$$\min \sum_k A_k$$

s.t.

$$\sum_t X^i_{p,t} \cdot hh^i_{p,t} = h^i_p \cdot V^i \qquad \forall i, p \tag{5.1}$$

$$\sum_{i,c,p} Y^i_{c,p,t} \geqslant \frac{\sum_{i,p} X^i_{p,t}}{WL_t} \qquad \forall t \tag{5.2}$$

$$Y^i_{c,p,t} \leqslant X^i_{p,t} \cdot hh^i_{p,t} \qquad \forall i, c, p, t \tag{5.3}$$

$$Z^i_c \geqslant \sum_t Y^i_{c,p,t} \cdot hh^i_{p,t} \qquad \forall i, c, p \tag{5.4}$$

$$A_k + R_k \geqslant \sum_{i,c} Z^i_c \cdot b^i_{c,k} \qquad \forall k \tag{5.5}$$

$$X^i_{p,t} \geqslant 0 \qquad \forall i, p, t \tag{5.6}$$

$$Y^i_{c,p,t} \geqslant 0 \qquad \forall i, c, p, t \tag{5.7}$$

$$Z^i_c \geqslant 0 \qquad \forall i, c \tag{5.8}$$

$$A_k \geqslant 0 \qquad \forall k \tag{5.9}$$

Constraints (5.1) define how the volume of product i is processed among tools t that are qualified for process operation p. The equality sign in this constraint is important,

since we want to check how a given product mix (quantities of each product) covers the risk on process tools. If an inequality sign was used, the model could artificially produce more than required of some products to cover the risk on process tools.

Constraints (5.2) express the requirements in terms of number of inspections for process tool t. This number is calculated as the total volume processed on tool t divided by its WL_t. Constraints (5.3) ensure that the number of lots inspected for process tool t cannot be larger than the total number of lots processed on tool t. Constraints (5.4) consider the coverage block of an inspection operation. They define how the total number of inspections performed at inspection operation c are assigned to control the process tools which are covered. These constraints are not summed over p because, by inspecting one lot, the process operations p where the lot was processed can be covered (if they belong to the coverage block of inspection operation c). The number of inspections performed in inspection operation c of product i is not known in advance and is defined through the optimization. Constraints (5.4) are actually the linearization of:

$$Z_i^c = \max_{i,c,p}(\sum_t Y_{c,p,t}^i \cdot hh_{p,t}^i)$$

Constraints (5.5) calculate the total number of inspections allocated to each inspection tool type k and define the additional capacity required on inspection tool k when reserved capacity R_k is not enough. Decision variables should be integer. However, the model is resolved for industrial instances with the manufacturing routes and defect inspection control plans of 10 technologies. This corresponds to 1800 process operations, 170 inspection operations and more than 700 process tools. Moreover, since this model is used to support decisions at the tactical level, using continuous variables is enough to relevant satisfactory results in a reasonable time.

5.4 Defect Inspection Capacity Planning - Model 2

In this model, the exposure of process tools is considered. The exposure refers to the number of wafers processed on a tool before the results of inspection are obtained. When a process tool is controlled with the inspection of a production lot, the time before obtaining the results of the inspection must be considered. This time corresponds to the cycle time (CT) between the process operation p and the inspection operation c. It depends on the manufacturing route of the product and it is calculated considering the intermediate process operations that must be performed before the lot reaches the defect inspection operation. This is illustrated in figure 5.1a. If tool 7-1 at process operation 1210 is controlled with a lot

of technology "A", the results of the inspection will be obtained after the lot is processed on operations 1270, 1280, 1290 and inspected on operation 1330. During that time, tool 7-1 continues to process. For practical reasons, the considered throughput (TH) of the tool is the average production rate when the tool is loaded. Therefore, when the results of an inspection are obtained, the W@R cannot be smaller than the average number of wafers processed on the tool between the process operation and the inspection operation ($CT \times TH$). We denote the exposure as $WD^i_{c,p}$ which is the average number of wafers between process operation p and inspection operation c of product i. Figure 5.2 shows an example of the W@R reduction resulting from the inspection of production lots.

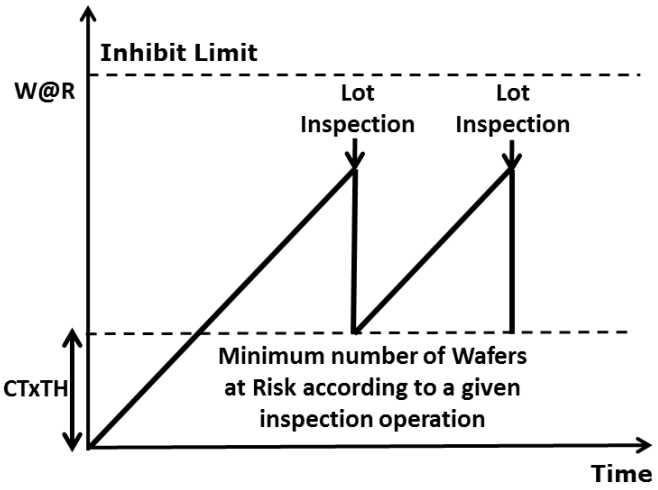


Figure 5.2: W@R reduction obtained with the inspection of production lots

In order to reduce the size of the problem, the index p on the decision variable $Y^i_{c,p,t}$ is no longer included. Instead, it is considered with a new parameter $e^i_{c,p}$ that defines the coverage block of inspection operation c with product i. Therefore, the exposure on process tools are included in constraint (5.2) as follows:

$$\sum_{i,c,p} Y^i_{c,t} \cdot e^i_{c,p} \cdot hh^i_{p,t} \geqslant \frac{\sum_{i,c,p} WD^i_{c,p} \cdot Y^i_{c,t} \cdot e^i_{c,p} \cdot hh^i_{p,t} + \sum_{i,p} X^i_{p,t}}{WL_t} \qquad \forall t$$

New Parameters

$WD^i_{c,p}$: Average number of wafers between process operation p and inspection operation c of product i. It is equal to $TC \times TH$,

$e^i_{c,p}$: Coverage block of inspection operation c of product i,

$$= \begin{cases} 1 \text{ if the inspection operation } c \text{ of product } i \text{ covers the tools of process operation } p, \\ 0 \text{ otherwise.} \end{cases}$$

Modified Decision Variables

$Y^i_{c,t}$: Number of inspections of product i in inspection operation c that covers process tool t,

The new model is given below:

$$\min \sum_k A_k$$

s.t.

$$\sum_t X^i_{p,t} \cdot hh^i_{p,t} = h^i_p \cdot V^i \qquad \forall i,p \tag{5.10}$$

$$\sum_{i,c} Y^i_{c,t} \cdot (\sum_p e^i_{c,p} \cdot hh^i_{p,t} \cdot WL_t - \sum_p e^i_{c,p} \cdot hh^i_{p,t} \cdot WD^i_{c,p}) \geqslant \sum_{i,p} X^i_{p,t} \qquad \forall t \tag{5.11}$$

$$Y^i_{c,t} \leqslant \sum_p X^i_{p,t} \cdot hh^i_{p,t} \cdot e^i_{c,p} \qquad \forall i,c,t \tag{5.12}$$

$$Z^i_c \geqslant \sum_t Y^i_{c,t} \cdot hh^i_{p,t} \cdot e^i_{c,p} \qquad \forall i,c,p \tag{5.13}$$

$$A_k + R_k \geqslant \sum_{c,i} Z^i_c \cdot b^i_{c,k} \qquad \forall k \tag{5.14}$$

$$X^i_{p,t} \geqslant 0 \qquad \forall i,p,t \tag{5.15}$$

$$Y^i_{c,t} \geqslant 0 \qquad \forall i,c,t \tag{5.16}$$

$$Z^i_c \geqslant 0 \qquad \forall i,c \tag{5.17}$$

$$A_k \geqslant 0 \qquad \forall k \tag{5.18}$$

The problem of the model at its present form is that the selected term of constraints (5.11) can be negative when the average number of wafers between process operations and inspection operation is larger than the Warning Limit.

$$\sum_{i,c} Y^i_{c,t} \cdot \boxed{(\sum_p e^i_{c,p} \cdot hh^i_{p,t} \cdot WL_t - \sum_p e^i_{c,p} \cdot hh^i_{p,t} \cdot WD^i_{c,p})} \geqslant \sum_{i,p} X^i_{p,t} \qquad \forall t$$

In order to tackle this situation, two considerations have been taken into account:

- The Warning Limit gives an alarm, but the objective is to ensure that the W@R on the process tool will not reach the maximum value of acceptable W@R for the process tool. Hence, the Inhibit Limit (IL_t) should be used when the average number of wafers between process operations and inspection operations is considered.
- When the W@R reaches the IL, the process tool is stopped (also the W@R counter) and a Quality Task (QT) is performed.

These remarks are considered in next section.

5.5 Defect Inspection Capacity Planning - Model 3

In this version of the model, Quality Tasks (QTs) are included. QTs are special controls performed on Non-Product Wafers (NPWs) (e.g. Test, Monitoring or Dummy wafers). Most tools in a fab use NPWs for process qualification, tool qualification, process conditioning or other purposes [82]. Hence, when QTs are performed it is assumed that the process tool is stopped until the results of the associated QT are obtained (figure 5.3). Therefore, for each QT that is performed, the W@R drops to zero and at most a number of wafers equal to IL can be produced before the W@R reaches the IL again. Although from the point of view of W@R reduction, QTs are important (i.e. they set the W@R to zero), performing QTs are costly and reduce process tool availability. Actually, QTs are important contributors to the production costs and major attention is paid to reduce their use [88], [89].

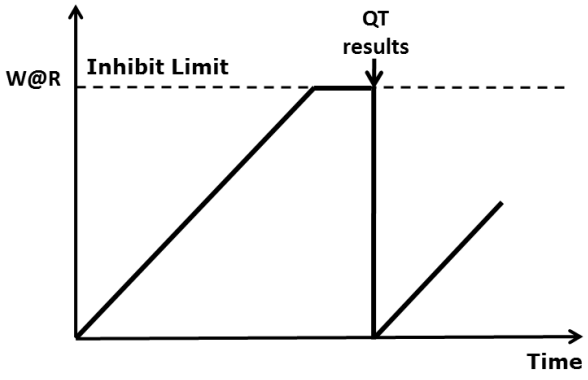


Figure 5.3: W@R reduction obtained with a QT

The QTs are modeled as extra controls that need to be performed when W@R limits cannot be satisfied. QTs are introduced in constraints (5.19). A penalty QTP is included in the objective function to reduce this type of controls. In addition, we have considered

the planned quality task (i.e. $ScQT_t$), they are not fully coordinated with the W@R controls, hence, when a planned QT is performed, the W@R level on the process tool is not necessarily at the maximum acceptable level (IL). Therefore, the W@R reduction that can be obtained is modeled as a percentage of the IL (i.e. ρ_t).

$$QT_t + \rho_t \cdot ScQt_t + \sum_{i,c,p} Y^i_{c,t} \cdot e^i_{c,p} \cdot hh^i_{p,t} \geqslant \frac{\sum_{i,c,p} WD^i_{c,p} \cdot Y^i_{c,t} \cdot e^i_{c,p} \cdot hh^i_{p,t} + \sum_{i,p} X^i_{p,t}}{IL_t} \qquad \forall t \tag{5.19}$$

In the previous models, we have considered that all the production volume can be inspected. However, a technology is composed of different types of products with the same route but not with the same defect inspection control plan (see section 1.3.3). Hence, there is only a percentage of the production volume of each technology that is measurable. This is included with the parameter γ_i.

Concerning the total defect inspection capacity, it is shared between two sampling strategies: Dynamic sampling and static sampling. Lots selected with static sampling are systematically inspected in all inspection operations defined as mandatory in their defect inspection control plan. This is included with the parameter SS^i, that gives the predefined sampling rate of product i and the parameter ob^i_c represents the set of mandatory inspections operations for product i. Moreover, the inspection time depends on the technology and the qualified inspection tool. For example, a Brightfield tool is slower than a Darkfield tool but can detect more types of defects. In consequence, the capacity is considered in terms of time. Finally, there are some inspection tools that are preferably used for some types of inspections other than W@R (e.g. Static sampling), therefore a penalty per inspection tool type ($DefP_k$) was introduced.

New Parameters

QTP : Penalty associated to the total number of additional QTs performed,

ob^i_c : Mandatory inspection operations for product i,

$$= \begin{cases} 1 \text{ if inspection } c \text{ is mandatory for product } i \\ 0 \text{ Otherwise} \end{cases}$$

SS^i : Static sampling rate for product i,

$DefP_k$: Factor to express a restriction to use the defect inspection tool k for W@R measures,

α : Percentage of mandatory inspections that also reduce the W@R on process tools,

γ^i : Percentage of volume of product i that is measurable,

$TimeQT$: Inspection time of a QT,

$TimeD0_c^i$: Inspection time of product i on inspection operation c,

$CapaMax_k$: Total capacity of inspection tool k given in time,

$ScQt_t$: Number of QTs that are scheduled on tool t,

ρ_t : W@R reduction considered for the planned QTs on process tools,

New Decision variables

QT_t : Number of additional QTs performed on process tool t.

The model is given below:

$$\min \sum_k DefP_k \cdot A_k + QTP \cdot TimeQT \cdot \sum_t QT_t$$

s.t.

$$\sum_t X_{p,t}^i \cdot hh_{p,t}^i = h_p^i \cdot V^i \qquad \forall i,p \tag{5.20}$$

$$\sum_{i,c} Y_{c,t}^i \cdot (\sum_p e_{c,p}^i \cdot hh_{p,t}^i \cdot IL_t - \sum_p e_{c,p}^i \cdot hh_{p,t}^i \cdot WD_{c,p}^i) +$$
$$QT_t \cdot IL_t + \rho_t \cdot ScQt_t \cdot IL_t \cdot \geqslant \sum_{i,p} X_{p,t}^i \qquad \forall t \tag{5.21}$$

$$Y_{c,t}^i \leqslant \sum_p e_{c,p}^i \cdot hh_{p,t}^i \cdot X_{p,t}^i \cdot \boldsymbol{\gamma_i} \qquad \forall i,c,t \tag{5.22}$$

$$Z_c^i + \alpha \cdot V^i \cdot SS^i \cdot ob_c^i \geqslant \sum_t Y_{c,t}^i \cdot e_{c,p}^i \cdot hh_{p,t}^i \qquad \forall i,c,p \tag{5.23}$$

$$A_k + R_k \geqslant \sum_{i,c} Z_c^i \cdot b_{c,k}^i \cdot TimeD0_c^i \qquad \forall k \tag{5.24}$$

$$A_k + R_k + \sum_{c,i} V^i \cdot SS^i \cdot ob_c^i \cdot b_{c,k}^i \cdot TimeD0_c^i \leqslant CapaMax_k \qquad \forall k \tag{5.25}$$

$$X_{p,t}^i \geqslant 0 \qquad \forall i,p,t \tag{5.26}$$

$$Y_{c,t}^i \geqslant 0 \qquad \forall i,c,t \tag{5.27}$$

$$Z_c^i \geqslant 0 \qquad \forall i,c \tag{5.28}$$

$$A_k \geqslant 0 \qquad \forall k \tag{5.29}$$

Constraints (5.20) are the same as constraints (5.1). Constraints (5.21) express how the number of inspections $Y^i_{c,t}$ for process tool t reduces the level of risk taking into account the number of wafers between the process operation p and the control operation c (i.e. $WD^i_{c,p}$). These constraints also consider the number of QTs that should be performed on tool t when inspections with production lots are not enough to satisfy the Inhibit Limit (IL_t). Constraints (5.22) ensure that the number of inspections on process tool t cannot be larger than the total number of measurable lots processed on t. Constraints (5.23) define the coverage block of an inspection operation by stating how the inspections performed on inspection operation c are assigned to the process tools that are covered. These constraints are not summed over p because, by inspecting one lot, all the process operations p where the lot was processed can be covered (if they belong to the coverage block of inspection operation c). The number of controls for dynamic sampling performed in inspection operation c of product i is not known in advance and is defined through the optimization. Constraints (5.23) are the linearization of:

$$Z^c_i = \max_{i,c,p}(\sum_t Y^i_{c,t} \cdot e^i_{c,p} \cdot hh^i_{p,t} - \alpha \cdot V^i \cdot SS^i \cdot ob^i_c)$$

Constraints (5.24) assign the total number of inspections on each inspection tool k and define the additional capacity required on inspection tool k when the reserved capacity R_k for W@R measures is not enough. This constraint also includes the capacity used for inspections of lots selected with the static sampling strategy. Since this model includes more industrial aspects, we added constraints (5.25), which ensure that the total inspection capacity allocated is smaller than the maximal capacity per inspection tool k. Hence, if the total required capacity per inspection tool k is larger than the maximal ($CapaMax_k$) the model allocates additional QTs. When the model is used to analyze the inspection capacity increases (i.e. decisions at the strategic level), this constraint can be ignored or $CapaMax_k$ increased.

5.6 Industrial Implementation

The data preparation is performed with R. The information of manufacturing routes, defect inspection control plans, tools qualifications, planned QTs, processing times and W@R limits are extracted from the data base. The defectivity engineer prepare the different scenarios with the information of mix and volume of products to consider. Figure 5.4 shows the general scheme of the application.

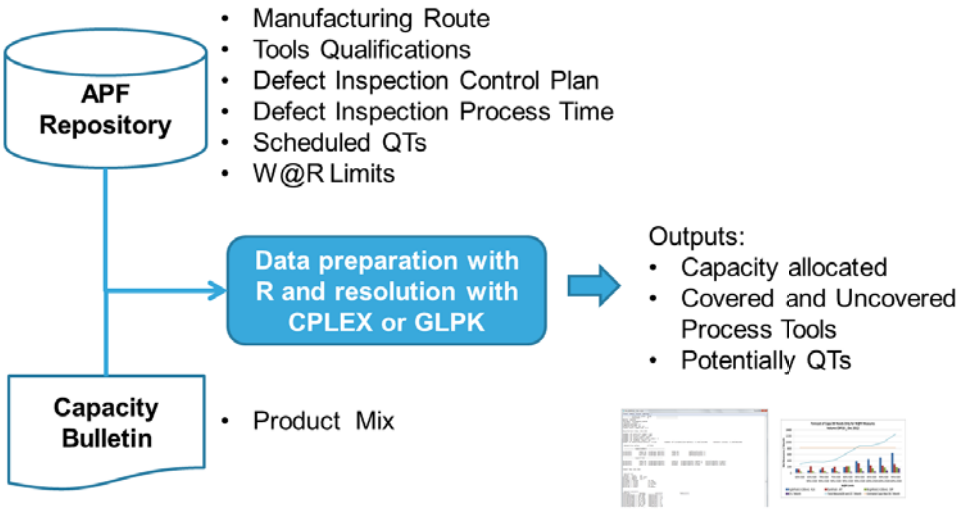


Figure 5.4: General Scheme of the Defectivity Capacity Model

With the proposed model it has been possible to justify the definition of W@R objectives in terms of Inhibit Limits (IL) reduction. Moreover, to understand the impacts of reducing W@R limits without taking into account the design of control plans, in particular in terms of additional QTs that can be generated due to process tools which W@R limits could not be satisfied. The model is currently used by defectivity engineers to anticipate the necessary actions to be taken when the product mix changes (e.g. creation of new inspection operations, setting of W@R limits). It has also supported the analysis of the potential impacts on the W@R management when any of the following factors changes:

1. What happens if the reserved capacity for W@R measures changes?
2. What happens if the ratio of lots selected with static sampling changes?
3. What happens if the volume of measurable products changes?
4. What happens if the product mix changes?
5. What happens if the W@R limits change?
6. What happens if the defect inspection control plans change?

5.7 Conclusions

In this chapter, three models for defect inspection capacity planning are presented. With the evolution of the sampling system in the site of Rousset of STMicroelectronics, a new approach to calculate the required capacity to satisfy the W@R limits was necessary.

When static sampling is used to select the lots for inspection, the capacity is allocated in advance via the sampling rates of lots. With the inspection of the same lots at different

inspection operations of their routes, the identification of yield-limiting process operations is possible thanks to the added-defect analysis [18]. When dynamic sampling is used, the W@R on process tools can be better controlled, because lots are selected in real time. Therefore, the two sampling strategies (i.e. Static sampling and dynamic sampling) are still being used in the factory, and are considered in the model. Industrial details that are critical for W@R control on process tools are included, such as: The configuration of control plans, qualifications of tools (i.e. process and inspection tools), the QTs performed on process tools and the W@R limits. Numerical experiments on industrial data are presented in chapter 6.

Chapter 6

Experiments on the Capacity Model

Numerical experiments on industrial data are presented in this chapter. The model can be used at different decision levels. At the tactical level, it shows if W@R limits can be satisfied when the product mix changes and/or if planned W@R reductions can be met with the available inspection capacity. At the strategic level, the model helps to justify capacity investments if the objectives in terms of W@R reduction cannot be achieved with the available capacity.

6.1 Introduction

In the proposed model, the capacity allocated with dynamic sampling is divided into two categories: (i) Reserved capacity (R_k) and (ii) Additional capacity (A_k). R_k is a percentage of the available capacity dedicated for dynamic sampling. If the reserved capacity is not enough to satisfy the inspection requirements additional capacity will be allocated. Therefore, the model proposed in chapter 5 aims at minimizing the total additional capacity required per inspection tool types.

Section 6.2 presents the results obtained with the capacity model version 1 (proposed in section 5.3). Since it is not possible to solve the second model without including additional parameters, we then only present in section 6.3 the results obtained with the model version 3 (proposed in section 5.5). Finally the concluding remarks and future work are presented in section 6.4.

6.2 Experiments on capacity model version 1

Numerical experiments on the model proposed in section 5.3 are presented. In this version of the model, the average number of wafers between the process operation p and the inspection operation c of product i is not considered (i.e $(WD^i_{c,p})$), hence, the W@R on process tools can be set to zero with the inspection of a production lot. The inspection time is similar for all lots and the considered capacity is only for the dynamic sampling strategy. Other type of sampling strategies are not explicitly included. We considered more

than 350 process tools, 1 800 process operations, 170 inspection operations and 12 product families. The model was developed using CPLEX 12.3 and experiments were run on a PC Intel Core i5 (2.40 GHz). In order to calculate the total capacity required, the reserved capacity (R_k) was set to 0 for all inspection tool types. In the following the inspection tool type 1 will be referred as T1, the same for inspection tool types 2 and 3 (T2 and T3).

6.2.1 Reduction of W@R Limits

The W@R limits are constraints to the quality department and the objective is to reduce these limits. The proposed model verifies if a given set of W@R limits can be satisfied with the available capacity. Let us recall that the W@R limits considered in this version of the model are the Warning Limits (WL). If the W@R level is larger than the WL the situation is considered as critical and actions have to be taken. Figure 6.1 shows the impact when W@R limits are reduced. The $y-axis$ is the percentage of defect inspection capacity that is required and the $x-axis$ is the number of products considered in each experiment. The limits in Group A correspond to the current W@R limits of process tools when these experiments were performed. The objective is to conduct a campaign of limit reductions. The limits in group B correspond to the first set of reductions, while the limits in group C correspond to the last set of reductions.

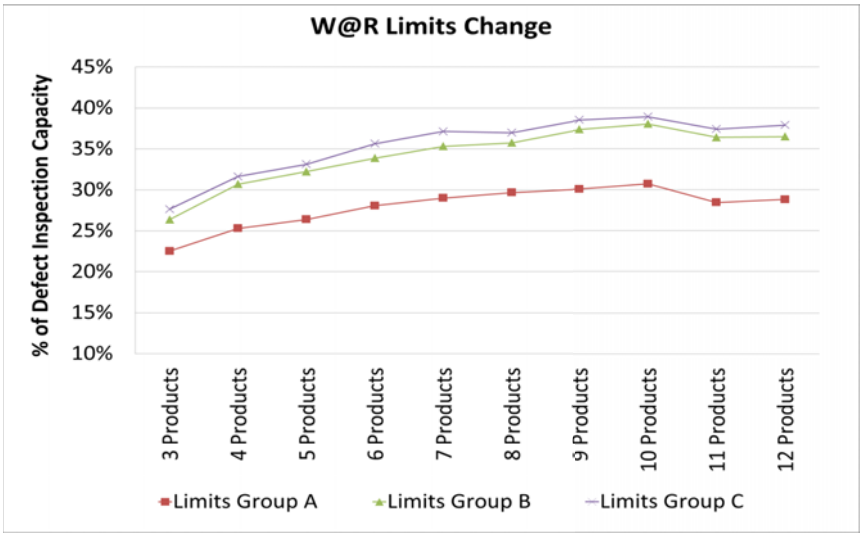


Figure 6.1: W@R limit reductions Vs. Defect inspection capacity

The gaps between limits A and B are explained by the fact that the warning limits for some process tools were reduced by over 60%. The gaps between limits B and C correspond to an additional reduction of 30% of the warning limits. It can be observed that when changing from 10 products to 11 products, there is a reduction on the capacity

requirements. Even if the total volume considered increases, the required capacity for W@R inspections is reduced. The reason is that the defect inspection control plan of the new product in group 11 has more inspection operations for which the coverage block is larger. Consequently, with the inspection of one lot more process tools can be covered.

6.2.2 Impact of Mix and Volume of Products

Table 6.1 presents the total required capacity with different product mixes and the value of the W@R limits when the experiments were conducted. To satisfy the W@R limits on process tools with product Mix 1, it is necessary to reserve 30.7% of the total defect inspection capacity, which represents 16.0% of T1 capacity, 10.0% of T2 capacity and 4.7% of T3 capacity. These results help to anticipate the qualification of inspection tools. The required capacity among inspection tool types is not balanced. Inspection capacity on T1 is the most required while inspection capacity on T3 is the less required. The main reason is that T1 is qualified on more inspection operations than the other tool types. However, for Mix 5, T2 is slightly more required than T1. The reason is that some inspection operations qualified on T2 can have similar W@R reduction than some inspection operations qualified on T1. Hence, these inspection operations are chosen when solving the model. The total required capacity with Mix 4 and Mix 7 is very close (46.9% and 46.3%) but the balance among inspection tool types changes. In particular, the capacity required on T1 decreases from 22.6% with Mix 4 to 17.0% with Mix 7 and the capacity required on T2 increase from 16.7% to 19.0%. The reason is that the manufacturing route of each technology is different. Therefore, the position, number and qualification of inspection operations are also different.

Table 6.1: Current W@R limits and different product mixes

Inspection tool types	Mix 1	Mix 2	Mix 3	Mix 4	Mix 5	Mix 6	Mix 7
T1	16.0%	15.8%	12.4%	22.6%	14.0%	19.9%	17.0%
T2	10.0%	10.2%	11.4%	16.7%	14.8%	12.6%	19.0%
T3	4.7%	4.5%	8.8%	7.2%	7.4%	5.8%	10.4%
Total	30.7%	30.5%	32.6%	46.5%	36.2%	38.3%	46.4%

Table 6.2 presents the results obtained when the W@R limits change. The objective is to check whether W@R limit targets can be achieved. The total required capacity in Mix 4 and Mix 7 is similar (59.7% and 59.3%) and the requirement among inspection tool types is also similar (32.7% to 31.7% for T1, 19.6% to 19.8% for T2 and 7.4% to 7.8% for T3) contrary to Table 6.1. With these new limits, more lots are selected and the inspection

operations where lots are sampled are also different. This explains why the balance on the capacity requirements among inspection tool types can change when W@R limits change.

Moreover, the results illustrate the impact of the defect inspection control plan design over the defect inspection allocation. For example, with the previous W@R limits and product mix 5 (Table 6.1) the capacity allocated on T2 is larger than T1, because with the defect inspection control plan configuration, it was possible to choose between inspection operations qualified on T1 and T2. However, when limits are more restricted (Table 6.2), both inspection tool types are chosen at the same time. This is the reason why T1 is the most used in all mixes.

Table 6.2: Target W@R limits and different product mixes

Inspection tool types	Mix 1	Mix 2	Mix 3	Mix 4	Mix 5	Mix 6	Mix 7
T1	19.6%	23.4%	23.4%	32.7%	25.5%	26.5%	31.7%
T2	11.5%	9.2%	10.7%	19.6%	15.5%	16.3%	19.8%
T3	7.4%	6.0%	6.6%	7.4%	5.7%	6.0%	7.8%
Total	38.5%	38.6%	40.7%	59.7%	46.7%	48.8%	59.3%

In this version of the model when a production lot is inspected the W@R drops to zero, i.e. the average number of wafers between process operation p and inspection operations c of product i, $WD^i_{c,p}$ is not included. Considering this parameter, the decisions regarding the selection of inspection operations and therefore the allocated inspection capacity change.

6.3 Experiments on capacity Model Version 3

This version of the model includes the average number of wafers between process operations and inspection operations ($WD^i_{c,p}$), the QTs that need to be performed if the W@R level reaches the IL, the planned QTs and the capacity allocated to the different sampling strategies: (i) Static sampling and (ii) Dynamic sampling. We tested scenarios in which the utilization of QTs are allowed and scenarios in which QTs are highly restricted. Also, the preference for an inspection tool type can be set via a specific penalty (i.e. $DefP_k$). For these experiments, a new inspection tool has been considered because new inspection operations were created and an additional tool qualified.

If the problem is feasible and limits can be satisfied, solving the model shows the utilization rate per inspection tool type (i.e. Total allocated time/ Total capacity in terms of time) and estimates the additional required QTs. If limits cannot be satisfied, the potentially uncovered process tools are listed with the estimated number of additional QTs. The sets of W@R limits presented in Table 6.3 are evaluated. For confidentiality

reasons all limits are normalized with the value M. Let us recall that in this version of the model we consider the Inhibit limits (IL). Set B corresponds to 65% of process tools with their ILs smaller than M, 90% of process tools with their ILs smaller than 3× M and 100% of process tools with their ILs smaller than $7 \times M$. The parameters used for the experiments are presented in tale 6.4. The production volume considered in experiments corresponds to one month of activity. Therefore, the number of additional QTs that can be generated due to W@R controls are generated during a month.

Table 6.3: Sets of W@R limits for process tools

Set B	**Set C**	**Set D**	**Set E**
$65\% \leqslant M$	$70\% \leqslant M$	$75\% \leqslant M$	$80\% \leqslant M$
$90\% \leqslant 3 \times M$	$90\% \leqslant 3 \times M$	$90\% \leqslant 3 \times M$	$90\% \leqslant 3 \times M$
$100\% \leqslant 7 \times M$	$100\% \leqslant 7 \times M$	$100\% \leqslant 7 \times M$	$100\% \leqslant 7\times$M

Table 6.4: Parameters used for scenarios with Model 3

Parameters	**Values**	**Interpretation**
Product Mixes	Mix1, Mix2, Mix3, Mix4, Mix5, Mix6	Different mixes and volumes of products
Limits	B, C, D, E	Set of W@R Limits on process tools
$DefP_k$ Penalty	[1, 5, 10]	Restriction of using a given inspection tool type for W@R control
QT Penalty (QTP)	[1, 10]	Restriction of using QTs that are not planned
Alpha	[0, 0.2, 0.5, 0.7, 1]	Percentage of lots selected with Static Sampling that can also reduce the W@R
Reserved Capacity (R_k)	[0, 0.2, 0.5, 0.7]	Capacity reserved in advance on inspection tool types for W@R control

The model was solved using IBM ILOG 12.5.1 and experiments were run on a PC Intel Core i5 (2.40 Ghz). Instances include more than 700 process tools (cluster tools are considered at the chamber level), 1800 process operations, 170 inspection operations and 10 product families.

6.3.1 Impact of penalty values

As presented in section 5.5, two penalty factors are used in the capacity model: $DefP_k$ and QTP. The penalty $DefP_k$ is used to restrict the utilization of a given inspection tool type over another inspection tool to inspect lots dynamically selected. The interest of this

penalty is to analyze scenarios where a given inspection tool type is reserved to be used for an activity other than W@R measures or where a different balance in terms of utilization rate among inspection tool types is explored. Concerning the penalty QTP penalty, it is used to restrict the utilization of additional QTs due to W@R control.

a) Variations of $DefP_k$ value

To analyze the impact of restricting the utilization of some inspection tool types over others, different combinations of this penalty were evaluated (i.e. 1, 5 and 10). Concerning the setting of the other parameters, we chose the values corresponding to a standard scenario of the fab when experiment were conducted. These values are summarized in table B.2. Limits C is the set of W@R limits when experiments were conducted and $Alpha = 0.2$ is the percentage of lots selected with static sampling that can reduce the W@R to its minimum level (i.e. $TC \times TH$).

Table 6.5: Parameters for evaluating $DefP_k$ variations

Parameters	
Product Mix	Mix 1
Limits	C
QTP	10
Alpha	0.2
Reserved Capacity	0

All experiments are presented in Appendix B. In this section we discuss the results of six scenarios presented in table 6.6. The performance measure is the utilization rate of each inspection tool type and the number of additional QTs. In the following, the inspection tool type 1 will be referred as T1, the same for inspection tool types 2, 3 and 4 (T2, T3 and T4). The columns named as "$DefP_k$ values", refer to the penalty selected on each inspection tool type. The columns named as "Utilization Rates for Static Sampling", refer to the time used for the inspection of lots selected only with the static sampling strategy over the total capacity in terms of time. The "Total Utilization Rates for Static and Dynamic Sampling", refer to the time used to inspect the total number of lots selected with both strategies. The column "QTs" refers to the number of additional QTs that are estimated. In this scenario, performing QTs due to W@R is highly restricted. For that, we set $QTP = 10$ and we consider that performing a QT would be six times longer than the longest inspection time of a production lot.

When there is no restriction for using a particular tool type (i.e. Scenario 1), the utilization rates of T1 and T4 are larger than the other two inspection tool types. The

reason is that the inspection time on T1 is larger than the others (e.g. average inspection time of 30 minutes on T1 versus 15 minutes on T2) and because there are more inspection operations qualified on T1 than on T3. However, the inspection time depends on the recipe, the product to inspect and the required inspection technology (e.g. Darkfield or Brightfield). Inspection tool type T4 corresponds to only one tool. Hence, the total capacity is smaller than the other tool types and it is easier to achieve 100% utilization rate. Moreover, an important aspect to consider is that the utilization rates dedicated to static sampling are 70.4% on T1 and 62.8% on T4. This implies that, even if there is a high penalty for using these tools, the utilization rates will not be smaller than the capacity already assigned with static sampling. When the use of T1 is restricted (i.e. scenario 2 with penalties 10,1,1,1) the workload is mainly transferred to T2 and T3. Since the utilization rates of T1 and T4 are 100% without restriction, only penalizing T1 will not impact T4 and vice-versa (scenarios 2 and 3).

When there is a restriction on T1 and T4 (i.e. scenarios 4, 5 and 6) the workload is transferred on T2 and T3. These results illustrate the flexibility of the control plan to cover the process tools. If there is a change on the available capacity for T1 or T4, there are other inspection operations qualified on T2 and T3 that can cover the process tools that were previously controlled with T1 or T4. Nevertheless, this is only true for the lots dynamically selected, because the dynamic sampling strategy takes the available capacity on the inspection tools into account before sampling a lot. It is important to note that the number of additional QTs does not change among the scenarios. This means that, with the current design of defect inspection control plans (i.e. position and coverage of inspection operations), there are some process tools that cannot be controlled only with the inspection of production lots.

Table 6.6: Impact of $DefP_k$ Penalties

Scenario	$DefP_k$ Values				Utilization Rates for Static Sampling				Total Utilization Rates for Static and Dynamic Sampling				QTs
	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	
1	1	1	1	1	70.4%	16.1%	15.3%	62.8%	100.0%	56.2%	23.3%	100.0%	45
2	1	1	1	10	70.4%	16.1%	15.3%	62.8%	100.0%	64.0%	40.3%	62.8%	45
3	10	1	1	1	70.4%	16.1%	15.3%	62.8%	86.6%	65.3%	48.4%	100.0%	45
4	5	1	1	10	70.4%	16.1%	15.3%	62.8%	87.3%	65.5%	74.9%	66.8%	45
5	10	1	1	5	70.4%	16.1%	15.3%	62.8%	86.6%	65.1%	73.5%	70.9%	45
6	10	1	1	10	70.4%	16.1%	15.3%	62.8%	86.6%	64.9%	80.1%	66.8%	45

b) Variation of QTP value

In the capacity model, we consider two types of QTs: The planned QTs and the additional QTs. The planned QTs are defined in advance and are thus inputs in our model. The additional QTs are allocated in the model when the W@R limits on process tools cannot be satisfied. Hence, the penalty QTP is used for QTs due to W@R control. The values of the other parameters in this scenario can be found in Table 6.7

Table 6.7: Parameters for evaluating QTP variations

Parameters	
Product Mix	Mix 1
Limits	C
$DefP_k$	[1, 5, 10]
Alpha	0.2
Reserved Capacity	0

Table 6.8 presents the results for $QTP = 1$. We consider that the time to perform a QT is as long as the time to perform the longest inspection operation of a production lot. When there is no restriction for using a particular tool type (i.e. Scenario 7), the utilization rates of T1 and T4 are 100%. Let us note that the utilization rates due to lots selected with start sampling are already 70.4% on T1 and 62.8 % on T4. Hence, it is easier to achieve 100% of utilization compared with T2 and T3. If there is a restriction on using an inspection tool type the estimated number of QTs can increase in the solution of the model (i.e. scenarios from 8 to 10). Restricting the use of T1 will generate more QTs than restricting the use of T4 (i.e. scenario 8 with 303 QTs and scenario 9 with 361 QTs). The reason is that T1 is qualified on more inspection operations than T4. The number of additional QTs generated vary according to the inspection tool type that is restricted. The number of QTs is reduce from 361 in scenario 9 to 354 in scenario 10. The reason is that with the restriction of T1 and T2, the utilization rate of T2 and T3 increases.

Table 6.8: Impact of QTs without penalty

Scenario	$DefP_k$ **Values**				**Utilization rates Static sampling**				**Total utilization rates Static and Dynamic sampling**				**QTs**
	T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**	
7	1	1	1	1	70.4%	16.1%	15.3%	62.8%	100.0%	46.6%	16.7%	100.0%	248
8	1	1	1	10	70.4%	16.1%	15.3%	62.8%	100.0%	51.4%	19.0%	62.8%	303
9	10	1	1	1	70.4%	16.1%	15.3%	62.8%	70.4%	71.7%	38.4%	100.0%	361
10	10	1	1	10	70.4%	16.1%	15.3%	62.8%	70.4%	76.5%	58.5%	62.8%	354

6.3.2 Impact of lots selected with static sampling

In this set of experiments, we aim at studying the impact of the W@R reduction that can be obtained with lots selected with the static sampling strategy. For this, we use the parameter *alpha*. This parameter expresses the percentage of lots selected with static sampling that can reduce the W@R to its minimum level (i.e. resulting W@R is $CT \times TH$). Table 6.9 gives the other parameters.

Table 6.9: Parameters for evaluating Alpha variation

Parameters	
Product Mix	Mix 1
Limits	C
$DefP_k$	[10,1,1,10]
QTP	10
Reserved Capacity	0

The results are presented in table 6.10. For these experiments, the utilization of additional QTs are highly restricted. When the percentage of lots from static sampling that reduce the W@R increases the resulting utilization rate is reduced. However, having all lots from static sampling that reduce the W@R levels to its minimum level, implies that these lots will need to be dispatched on the tools that need to be controlled. Moreover these lots should have a high priority to reduce the waiting times on intermediate process operations and to ensure that the maximum W@R reduction can be obtained. However, the scheduling of lots in each process area is already complex and having an additional criterion to accelerate all the lots selected for static sampling would increase the complexity. With the dispatching for sampling application (see section 3.3), the operators can identify when it is necessary to process a measurable lot on a process tool that needs to be controlled. That is why, scenarios with an alpha between 0.2 and 0.5 were representative of the fab activity.

Table 6.10: Impact of alpha variation

Scenario	**Alpha**	$DefP_k$ **Values**				**Utilization rates Static sampling**				**Total utilization rates Static and Dynamic sampling**				**QTs**
		T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**	
12	0	10	1	1	10	70.4%	16.1%	15.3%	62.8%	91.1%	69.0%	88.0%	66.9%	45
14	0.2	10	1	1	10	70.4%	16.1%	15.3%	62.8%	86.6%	64.9%	80.1%	66.8%	45
16	0.5	10	1	1	10	70.4%	16.1%	15.3%	62.8%	80.0%	59.9%	65.8%	66.8%	45
18	0.7	10	1	1	10	70.4%	16.1%	15.3%	62.8%	75.6%	55.8%	58.5%	66.8%	45
20	1	10	1	1	10	70.4%	16.1%	15.3%	62.8%	71.0%	51.0%	43.8%	66.8%	45

6.3.3 Impact of changes in the mix of products

To analyze the impact of the product mixes, we studied two different cases:

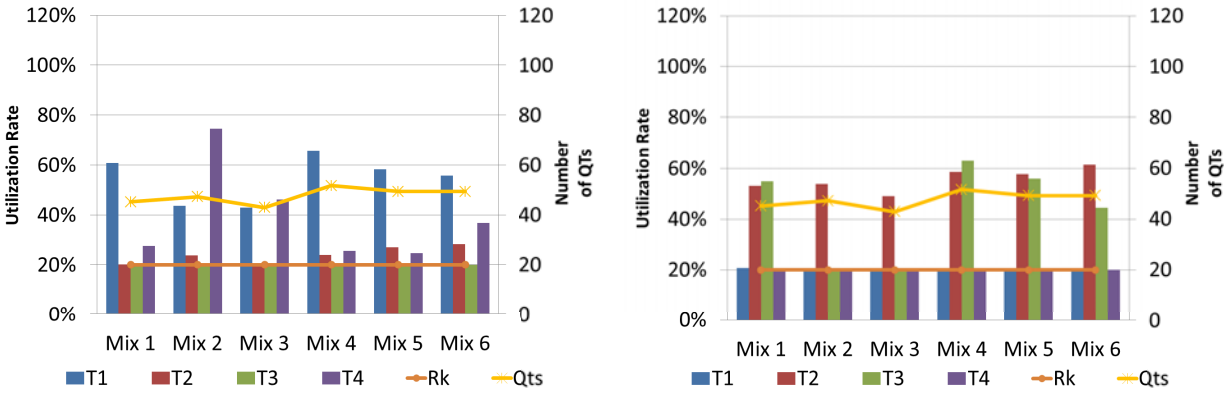
1. Lots selected with static sampling are not considered and the percentage of reserved capacity for W@R measures change.
2. Lots selected with static sampling are considered and the total required capacity is analyzed.

Table 6.11 gives the values of the other parameters. We have tested six different product mixes, the W@R limits correspond to the limits in the fab when experiments were conducted. Finally, two sets for the parameter $DefP_k$ are considered.

Table 6.11: Parameters for evaluating product mix variation

Factors	
Product Mixes	Mix1, Mix2, Mix3, Mix4, Mix5, Mix6
Limits	C
$DefP_k$	[10,1,1,10] [1,1,1,1]
Alpha	0.2
QTP	10
R_k	0, 20%, 30%, 40%

The following results correspond to the case where a percentage of the capacity is reserved in advance for W@R measures. Only the lots dynamically selected are considered.

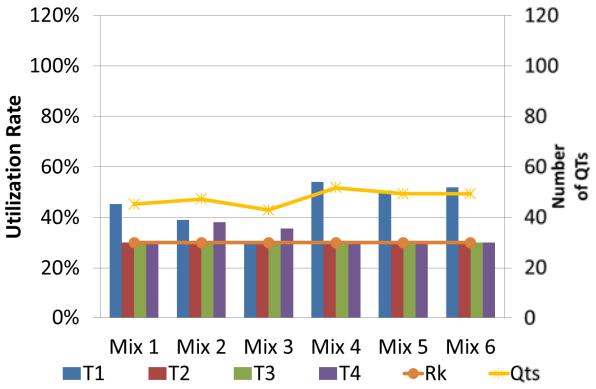


(a) $DefP_k$=[1,1,1,1] (b) $DefP_k$=[10,1,1,10]

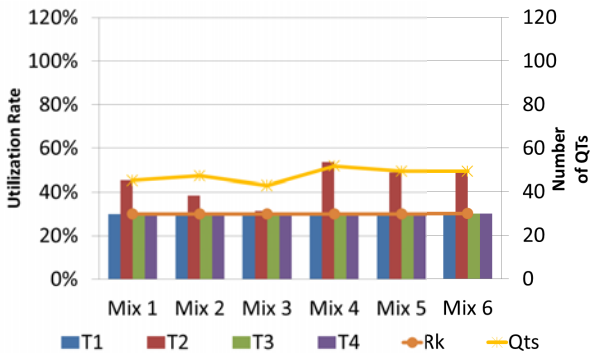
Figure 6.2: Utilization rates of inspection tool types when R_k=20%

Figures 6.2 shows the results when 20% of capacity is reserved in advance in all the inspection tool types. The *y-axis* of the right side of each graph is the utilization rate of

inspection tool types, while the *y-axis* of the left side represents the number of additional QTs. The *x-axis* corresponds to the different product mixes that were evaluated. The utilization rates of inspection tool types are illustrated with the bar graph and the additional QTs are illustrated with the line graph. Figure 6.2a presents the resulting utilization rates when there is no restriction of using a particular tool type (i.e. $DefP_k$=[1,1,1,1]). It can be observed that, in all groups of product mixes it is necessary to use additional capacity. In particular T1 and T4 need more additional capacity compared with the other tool types. As expressed before, this is a consequence of the inspection time and the number of inspection operations that are qualified on these tool types. However, when there is a penalty on T1 and T4, see figure 6.2b, the utilization rates of T2 and T3 increase depending on the product mix. For example, in product mixes 2 and 3, by restricting the use of T1 and T3, only the utilization rate of T2 increases. The reason is that, with the defect inspection control plan and mixes 2 and 3, the process tools can be controlled with the inspection operations qualified on T2.



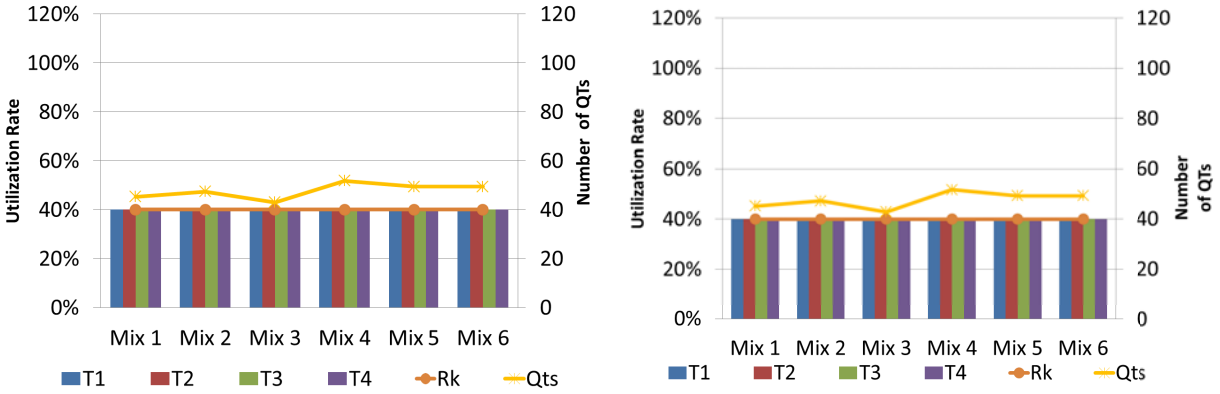
(a) R_k = 30% and $DefP_k$=[1,1,1,1]



(b) R_k = 30% and $DefP_k$=[10,1,1,10]

Figure 6.3: Utilization rates of inspection tool types when R_k=30%

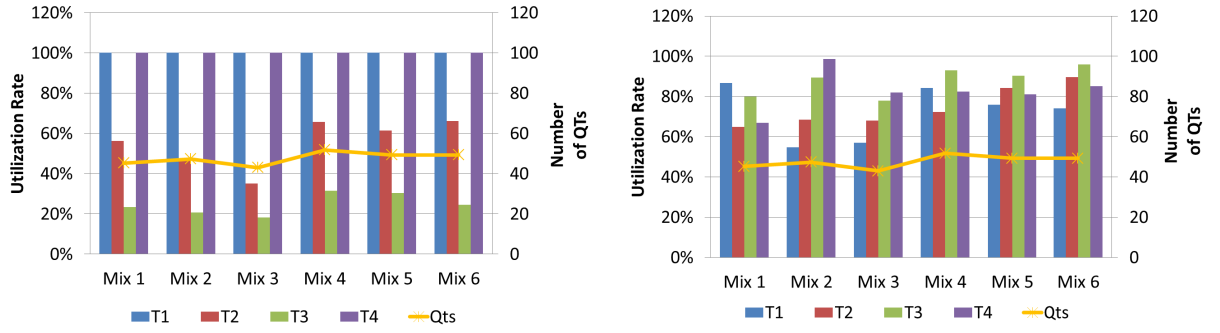
Figure 6.3 presents the results when the reserved capacity is 30% on all inspection tool types. In figure 6.3a there is no restriction of using a particular tool type. It is important to note that T1 is the only inspection tool type which requires additional capacity for all product mixes, contrary to T4 for which additional capacity is only required for mixes 2 and 3. But if there is a restriction on T1 and T4 (figure 6.3b), the only impacted inspection tool type is T2. These results are consistent with the reality because inspection operations qualified on T1 and T2 have a better coverage than the inspection operations qualified on the other tool types. In addition, if a new inspection operation needs to be created, T2 would be preferably qualified.



(a) R_k = 40% and $DefP_k$=[1,1,1,1] (b) R_k = 40% and $DefP_k$=[10,1,1,10]

Figure 6.4: Utilization rates of inspection tool types when R_k=40%

When 40% of capacity is reserved for W@R on all inspection tool types (figure 6.4), the W@R limits of process tools can be satisfied, excepting for those tools where QTs are generated. When additional QTs are still generated shows that W@R limits of some process tools cannot be satisfied only with production lots as a consequence of the defect inspection control plan configuration.



(a) $DefP_k$=[1,1,1,1] (b) $DefP_k$=[10,1,1,10]

Figure 6.5: Total utilization rates for static and dynamic sampling

Figure 6.5 shows the utilization rates required when both sampling strategies are considered (Static and Dynamic). In these scenarios, we did not reserve capacity for dynamic sampling. On the contrary, we consider that 20% of lots selected with static sampling reduce the W@R to its minimum level. In figure 6.5a, there is no restriction related to using a specific inspection tool type. As in previous cases, T1 and T4 have the highest utilization rates on all the considered product mixes. When there is a restriction on these inspection tool types (figure 6.5b), the utilization rates of T2 and T3 increase and the additional QTs do not change. In summary, different solutions for capacity allocation can be obtained. In

all the product mixes that were evaluated, most of the W@R limits on process tools can be satisfied. However, there are some process tools with limits that cannot be satisfied and thus additional QTs are necessary. Using this information, defect inspection engineers can anticipate the required modifications in the defect inspection control plan to stay within the W@R limits on the process tools. Moreover, if new inspection operations cannot be created, QTs associated with violated W@R control limits are estimated.

6.3.4 Impact of W@R limit reduction

In this section, we analyze the impact of reducing the W@R limits. Experiments have been divided in two different cases:

1. The group of W@R limits B, C, D and E are evaluated with different product mixes, only the lots dynamically selected are considered and a capacity of 20% is reserved in advance (i.e R_k) on all inspection tool types. The objective of the first case is to analyze the additional requirements only for dynamic sampling.

2. Lots selected with both sampling strategies (static sampling and dynamic sampling) are considered. Concerning the set of limits for the second case, we consider all the inspection operations that can cover each process tool, and then reduce the IL of each process tool to its minimum exposure (min $TC \times TH$). The new value of the Inhibit Limit is set 10% larger than the minimum exposure.

In the following, the limits reductions in the first case are presented. The values for the other parameters are summarized in table 6.12.

Table 6.12: Parameters for evaluating W@R limit reductions

Parameters	
Product Mixes	Mix1, Mix2, Mix3, Mix4, Mix5, Mix6
Limits	B, C, D, E
$DefP_k$	[1,1,1,1]
QTP	10
R_k	20%

Table 6.13 shows the utilization rates required only with lots dynamically selected and the set B of W@R limits. There is no restriction for using a particular inspection tool (i.e. $DefP_k$=[1,1,1,1]). In the first set of W@R limit reductions, from limits B (table 6.13) to limits C (table 6.14), the impact in terms of utilization rates is similar and the additional number of QTs does not change. Hence, it is possible to reduce the limits with

the current defect inspection control plan configuration. However, changing the limits from C (table 6.14) to D (table 6.15) leads to an increase of the utilization rates, in particular on inspection tool types T2 and T4. The number of QTs drastically increases (e.g on mix 1 it increases from 45 to 509 QTs), which generates that the utilization rates for T2 do not change. The main reason is that the inspection operations are no longer as close as required to the process tools. Hence the new W@R limits cannot be satisfied with the inspection of production lots. Finally, if there is an additional W@R reduction, from D (table 6.15) to E (table 6.16), the utilization rate will be 100% and the number of additional QTs is multiplied by 20 compared to the QTs generated with limits B or C. In summary, it can be observed that limits D and E can generate a very large number of additional QTs if the control plan is not modified by adding new inspection operations or better positioning the inspection operations (see chapter 7).

Produc Mix	Utilization Rates Dynamic Sampling				QTs
	T1	T2	T3	T4	
Mix 1	61.9%	21.5%	20.0%	20.8%	45
Mix 2	45.0%	22.8%	20.0%	73.1%	47
Mix 3	42.0%	20.2%	20.0%	46.9%	43
Mix 4	65.5%	25.3%	20.0%	24.8%	52
Mix 5	59.6%	28.0%	20.0%	22.7%	49
Mix 6	56.6%	27.5%	20.0%	38.6%	49

Table 6.13: Impact of W@R Limits B

Product Mix	Utilization Rates Dynamic Sampling				QTs
	T1	T2	T3	T4	
Mix 1	60.7%	20.1%	20.0%	27.7%	45
Mix 2	43.6%	23.8%	20.0%	74.6%	47
Mix 3	43.0%	20.3%	20.0%	46.1%	43
Mix 4	65.7%	24.0%	20.0%	25.7%	52
Mix 5	58.3%	27.1%	20.0%	24.7%	49
Mix 6	55.7%	28.3%	20.0%	36.8%	49

Table 6.14: Impact of W@R Limits C

Product Mix	Utilization Rates Dynamic Sampling				QTs
	T1	T2	T3	T4	
Mix 1	67.8%	53.7%	20.0%	33.7%	509
Mix 2	45.0%	58.6%	20.0%	89.9%	557
Mix 3	47.4%	50.7%	20.0%	54.9%	489
Mix 4	72.6%	62.9%	20.0%	45.2%	582
Mix 5	64.9%	71.1%	20.0%	49.2%	564
Mix 6	69.0%	69.5%	20.0%	50.5%	881

Table 6.15: Impact of W@R Limits D

Product Mix	Utilization Rates Dynamic Sampling				QTs
	T1	T2	T3	T4	
Mix 1	100%	100%	100%	100%	1 054
Mix 2	100%	100%	100%	100%	1 184
Mix 3	100%	100%	100%	100%	896
Mix 4	100%	100%	100%	100%	1 436
Mix 5	100%	100%	100%	100%	1 407
Mix 6	100%	100%	100%	100%	1 771

Table 6.16: Impact of W@R Limits E

In the set of experiments conducted in the second case, the value of ILs are calculated according to the minimum exposure on process tools. For each process tool, the new value IL is 10% larger than the minimum exposure (i.e. $TC \times TH$). With this rule, 77% of the process tools have their ILs smaller than or equal to M, 94% of process tools have their ILs smaller than or equal to $3 \times M$ and 100% of process tools have their ILs smaller than or equal to $\leqslant 6 \times M$. Table 6.17 shows the results with the product mix 1. It can be observed that, with this strategy, all inspection tools are required at 100%. When the two strategies to select lots are considered (Static and Dynamic), it is better to dispatch the

lots selected with static sampling. The larger alpha, the smaller the number of additional quality tasks (i.e. from 5 972 QTs to 4 274 QTs). However, it is more effective in terms of W@R management to reduce the number of lots selected with static sampling and to dedicate the available capacity to inspect lots that are dynamically selected (i.e. 2 886 QTs).

Product Mix	Alpha	Total Utilization Rates Static and Dynamic Sampling				QTs
		T1	T2	T3	T4	
Mix 1	0	100%	100%	100%	100%	5 972
	0.2	100%	100%	100%	100%	5 622
	0.5	100%	100%	100%	100%	5 102
	0.7	100%	100%	100%	100%	4 764
	1	100%	100%	100%	100%	4 274
Mix 1 without static sampling	-	100%	100%	100%	100%	2 886

Table 6.17: Total utilization rates when W@R limits change to the minimum exposure value

In these scenarios, it is important to note that the number of generated QTs highly increases compared with the previous experiments. The main reason is that, by choosing the minimum exposure as the IL, the inspection operations whose exposures are the highest will not be selected when solving the model.

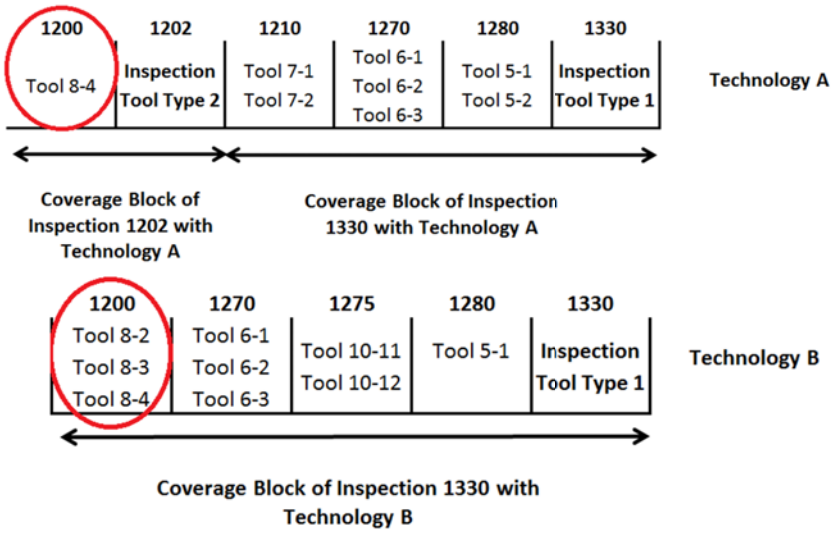


Figure 6.6: Exemple of Coverage Blocks

Figure 6.6 is an example of the coverage block obtained with a product of technology A and a product of technology B. For instance, with the initial set of W@R limits, the process tool 8-4 can be controlled with inspection operation 1202 of technology A and

inspection operation 1330 of technology B. When the new IL is the minimum exposure, the only inspection operation that can control this tool is 1202 of technology A. The reason is that, the time to get the inspection results from inspection 1330 is too long to avoid the W@R of process tool 8-4 to reach the IL. Hence this process tool will be controlled with QTs.

6.4 Conclusions

In this chapter, numerical experiments conducted with industrial data on the models for defect inspection capacity planning were presented and discussed. The model is used as a decision support tool to help answering the main question: Is it possible or not to satisfy a given set of W@R limits under different scenarios? the approach takes into account the design of control plans, the qualifications of process and inspection tools, the mix and volumes of products and the W@R limits. The approach can be used at different decision levels:

- At the tactical level, it helps anticipating the impact of product mix changes on the required defect inspection capacity, but also defining the objectives in terms of feasible reductions of W@R limits.
- At the strategic level, it can support decisions on the capacity increase of the defect inspection workshop.

Since reductions of W@R levels highly depend on the defect inspection control plans, the system depends on the product mix. Therefore, situations in which W@R limits are ensured may change if the product mix changes. Quality tasks are special controls performed on process tools when the W@R levels reaches Inhibit Limit. When quality tasks are necessary, even if inspection capacity is available, it means that the W@R limits of all the process tools cannot be satisfied only with the inspection of production lots. In the next chapter, a first study on how the defect inspection control plans design can impact the W@R level on process tools is presented.

Chapter 7

Impact of Control Plan Design on Tool Risk Management

In this chapter[1]*, we aim at analyzing how the position and coverage of inspection operations may influence the manufacturing robustness from the point of view of the risk on tools (i.e. W@R). This study was performed with the simulation tool S5*[2] *developed by the EMSE-SFL department*[3]*. Results show that not only the number and position of inspection operations impact the risk on tools, but also how each inspection operation covers process operations. The chapter concludes with the proposition of a mathematical model for the inspection operation location and allocation problem.*

7.1 Introduction

In the previous chapters we analyzed the inspection capacity allocation problem based on predefined defect inspection control plans. Let us recall that a defect inspection control plan, corresponds to the position and coverage of inspection operations within the manufacturing route (see Section 1.3.3). Control plans and sampling strategies are highly related. However, few methods link risk analyses and actual control plan design in a detailed manner [79]. In this chapter we aim at analyzing the impact of control plan design on the overall Wafers at Risk of the fab.

The chapter is structured as follows. Section 7.2 introduces the problem. Section 7.3 is devoted to the experimental study and analysis of results. Section 7.4 presents the formulation of a mathematical model for the location and allocation problem. Finally, conclusions and perspectives are discussed in Section 7.5.

[1] Part of this chapter was published in the 7th International Conference on Modeling and Analysis of Semiconductor Manufacturing (included in the 2011 Winter Simulation Conference) [90]

[2] S5: **S**mart **S**ampling, **S**kipping and **S**cheduling **S**imulator

[3] EMSE-SFL: École des Mines de St-Etienne, Département Science de la Fabrication et Logistique.

7.2 Problem Description and Solution Approach

This chapter focuses on how the defect inspection control plan can impact the W@R levels in the fab. As shown in previous chapters, the W@R on process tools is influenced by the throughput of process tools and the time to get the inspection results. Figure 7.1 illustrates the minimum W@R (i.e. exposure) on a process tool with two different defect inspection control plans. In order to reduce the exposure of process tool 1 two factors can be managed (i.e. reduced): (1) The throughput of the process tool, but this would not be acceptable from a productivity point of view or (2) the delay to get the inspection results. Hence, a new inspection operation is introduced which enables the delay to get the inspection results to be reduced. In this study, we analyze how the configuration of new control plans can have a positive or negative effect on the W@R levels.

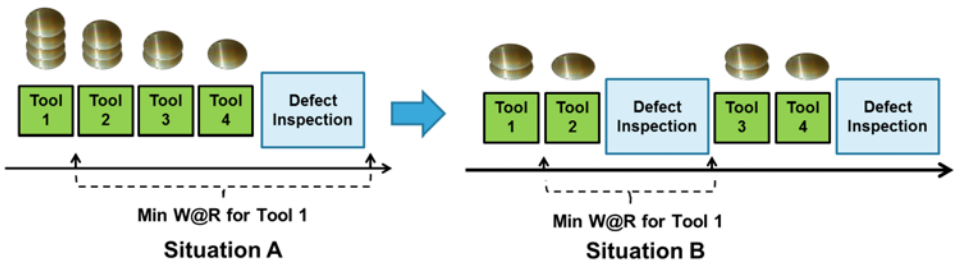


Figure 7.1: Illustration of minimal W@R

In order to analyze the impact of having new inspection operations, the original control plans were modified. Let us recall that defect inspection control plans are specified for each product (i.e. they are defined according the manufacturing route of the corresponding product). Hence, several defect inspection control plans are considered. Originally, in our industrial data, there were only X number of inspection operations in the defect inspection control plans. We modified them to include additional inspection operations (i.e. X+20 and X+50). Two strategies for including new inspection operations are studied, the “Overlapping” and “No Overlapping” of inspections. Hence, for these experiments, the technical restrictions in terms of capability of inspections were not considered when modifying defect inspection control plans. Figure 7.2 is a representation of both strategies to include more inspection operations. With the "Overlapping" strategy, the coverage of the original defect inspection control plan is maintained and additional inspection operations are included. Therefore, the impact of the original defect inspection control plan is conserved. With the "No Overlapping" strategy, the coverage of each inspection operation is reduced, hence the number of process operations that are covered is divided by the number of inspection operations. Results on industrial instances are presented in the next section.

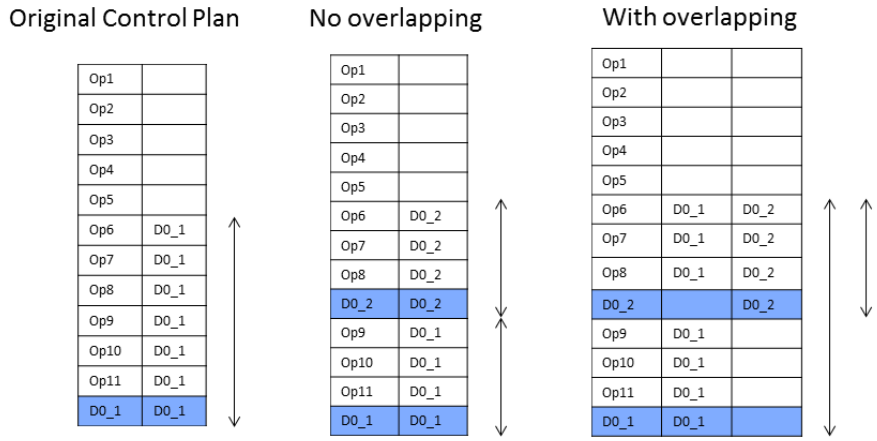


Figure 7.2: Defect inspection control plans with and without overlapping

7.3 Experimental Results

Experiments were conducted on industrial data, and the instances we use correspond to the activity of the fab during one month of production, i.e. they include more than 700 production tools. The experiments were conducted at the beginning of the thesis, hence the only available sampling strategy was "Static Sampling" (i.e. Start Sampling). For confidentiality reasons, all results are normalized. The selected indicator is the "max W@R average", which refers to the average of the maximum Wafers at Risk level for each process tool in the fab. All performance measures are compared with historical data, which correspond to the results obtained with the static sampling strategy in the fab. Therefore, the column "static Sampling" is set to 100% and results with the dynamic sampling strategy are compared with the use of a static sampling strategy. Different capacity values for the defect inspection area are considered. Capacity A corresponds to the available capacity when the experiments were conducted. Capacities A2, A3, A4 and A5 correspond to a reduced number of inspection tools. Infinite Capacity, refers to the case where all measurable lots are inspected, it represents a lower bound for W@R levels that can be achieved with each configuration.

Table 7.1 presents the results of max W@R averages with different configurations of defect inspection control plans. Results for X+20 and X+50 correspond to defect inspection control plans with 20 and 50 new inspection operations. These new inspection operations cover the same process operations than the initial defect inspection control plan. Results for capacity A and the original defect inspection control plan (i.e. X inspection operations) show the large gains that can be obtained by only changing the sampling strategy. The maximum W@R average is reduced from 100% with static sampling to 72.6% with dynamic

sampling. Results obtained when the capacity is reduced (i.e. A4 and A5) show that the maximum W@R average can increase when considering additional inspection operations and no overlapping (155.2% and 104.2% with X+50 inspection operations respectively). The reason is that without overlapping, the coverage block of an inspection operation is reduced. As illustrated in figure 7.2, with the original defect inspection control plan, the inspection operation $D0\_1$ validates six process operations and, with an additional inspection operation $D0\_2$ (without overlapping) it only validates three process operations. This difference between overlapping and no overlapping is detailed in table 7.2. In the case of capacity A5, we observe that an additional reduction on W@R is obtained when overlapping is considered, 3.2% and 15.0% with X+20 and X+50 respectively. Therefore, when the number of inspection operations increases and the capacity is reduced, the influence of the overlapping of inspection operations becomes more important.

Table 7.1: Impact of defect inspection control plans with and without overlapping

Maximum W@R average	Static S. Capa A	Dynamic S. Capa A5	Dynamic S. Capa A4	Dynamic S. Capa A3	Dynamic S. Capa A2	Dynamic S. Capa A	Infinite Capacity
X inspection operations	100.0%	140.3%	95.4%	78.8%	74.3%	72.6%	57.7%
X+20 no overlap.	100.0%	142.5%	95.1%	78.9%	73.6%	72.6%	57.0%
X+50 no overlap.	100.0%	155.2%	104.2%	86.5%	81.7%	80.6%	66.0%
X+20 with overlap.	100.0%	139.3%	92.5%	78.0%	73.4%	71.6%	56.3%
X+50 with overlap.	100.0%	140.1%	93.3%	77.8%	73.3%	71.8%	56.3%

Table 7.2: Difference between defect inspection control plans with and without overlapping

Delta between defect inspection control plan with and without overlapping	Dynamic S. Capa A5	Dynamic S. Capa A4	Dynamic S. Capa A3	Dynamic S. Capa A2	Dynamic S. Capa A	Infinite Capacity
X+20 inspection operations	3.2%	2.6%	0.9%	0.3%	1.0%	0.7%
X+50 inspection operations	15.0%	10.9%	8.7%	8.3%	8.8%	9.7%

In table 7.3, the positions of inspection operations are studied. It presents the results of defect inspection control plans with X+20 and X+50 inspection operations. The new inspection operations are included with different positions within the manufacturing route and all of them with overlapping. In configuration 1, the throughput of process tools are considered but no in an exhaustive way. In particular, some inspection operations are placed near process tools with high throughput, which leads to a reduction on the maximum W@R levels for these process tools and better results compared with the other configurations. In configuration 2, the throughput is not considered, hence the W@R levels are degraded compared to the other configurations, which reflects that inspecting a lot would not be efficient enough in terms of W@R reduction. These results show that

not always more inspection operations result in less risk because it highly depends on the position of these inspection operations within the manufacturing route. The scenario with "infinite capacity" shows that the maximum W@R average cannot not be smaller than 42.0% with the current configurations of defect inspection control plans.

Table 7.3: Impact of new inspection operations with overlapping

Maximum W@R average	Dynamic S. Capacity A5	Dynamic S. Capacity A4	Dynamic S. Capacity A3	Dynamic S. Capacity A2	Dynamic S. Capacity A	Infinite Capacity
X inspection operations	100.0%	140.3%	95.4%	78.8%	74.3%	72.6%
X+20 Configuration (1)	139.3%	92.5%	78.0%	73.4%	71.6%	56.3%
X+20 Configuration (2)	167.5%	109.4%	86.2%	75.3%	70.1%	46.4%
X+20 Configuration (3)	140.1%	93.4%	78.0%	73.0%	71.6%	56.3%
X+50 Configuration (1)	140.1%	93.3%	77.8%	73.3%	71.8%	56.3%
X+50 Configuration (2)	197.5%	128.9%	95.5%	81.1%	72.4%	42.0%
X+50 Configuration (3)	140.1%	93.2%	77.5%	73.3%	71.9%	56.3%

In the results presented until now, when additional inspection operations are included, the coverage of new inspection operations is similar to the original defect inspection control plan. As illustrated in figure 7.3, the W@R on process tools that were not covered with the original defect inspection control plan was not impacted with the new inspection operations.

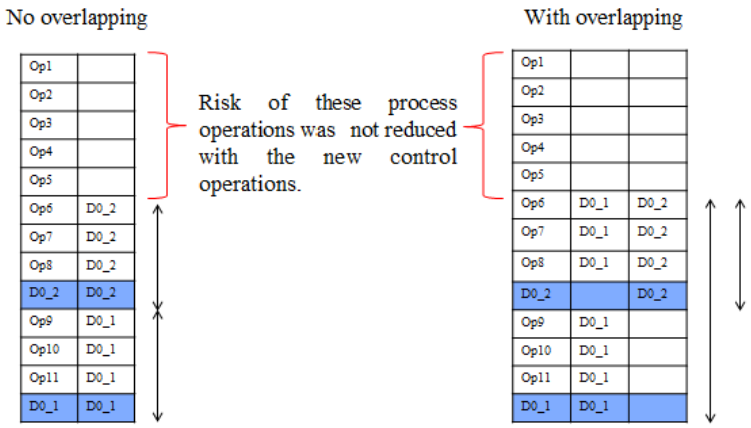


Figure 7.3: Coverage of inspection operations

Results in table 7.4, show the maximum W@R average obtained with a defect inspection control plan with X+20 inspection operations. The coverage of new inspection operations include the process operations that were not covered before. The value "B" represents the number of process operations covered by each inspection operation. Hence, a defect inspection control plan with B+6 covers more process operations than a defect inspection control plan with B.

Table 7.4: Impact of new inspection operations that cover new process operations

Maximum W@R average	Dynamic S. Capacity A5	Dynamic S. Capacity A4	Dynamic S. Capacity A3	Dynamic S. Capacity A2	Dynamic S. Capacity A	Infinite Capacity
X inspection operations	140.3%	95.4%	78.8%	74.3%	72.6%	57.7%
X+20 without overlapping (B)	156.7%	102.7%	74.7%	62.6%	54.3%	26.2%
X+20 with overlapping (B+2)	146.6%	93.1%	73.9%	59.8%	53.0%	26.2%
X+20 with overlapping (B+4)	143.7%	90.9%	69.9%	61.2%	52.5%	26.0%
X+20 with overlapping (B+6)	137.4%	89.1%	70.0%	58.7%	51.2%	25.9%

Results with infinite capacity show that, when new process operations are covered, the impact of additional inspection operations is significant. The maximum W@R average decreases from 57.7% with the original defect inspection control plan (X inspection operations) to 26.2% with a defect inspection control plan of X+20 inspection operations. Therefore, when capacity is increased, the factor that enables the reduction of W@R is the number of inspection operations. When capacity is reduced (i.e. case of capacity A4), overlapping is an important factor that helps to reduce the overall W@R. The maximum W@R average obtained with the original defect inspection control plan is 95.4%, compared with a maximum W@R average of 102.7% without overlapping and 89.1% with overlapping.

7.4 Mathematical Model for the Location and Allocation Problem

In this section, a model is proposed to optimize the location of inspection operations within the manufacturing routes of products. The model is based on the inspection capacity allocation model proposed in chapter 5.

The following sets are used to write the model:

I : Set of products indexed by i,

P : Set of process operations indexed by p,

C : Set of inspection operations indexed by c,

T : Set of process tools indexed by t,

K : Set of inspection tools indexed by k,

The parameters below are necessary:

IL_t : Inhibit Limit of process tool t,

$WD^i_{c,p}$: Average number of wafers of product i between process operation p and inspection operation c,

$WDQT_t$: Average number of wafers that cannot be produced while a Quality Task is being performed on process tool t. It is calculated as the throughput of t multiplied by the time required to perform a Quality Task on t,

$hh^i_{p,t}$: Qualification of process tools,

$$= \begin{cases} 1 \text{ if process tool } t \text{ is qualified to process product } i \text{ in process operation } p, \\ 0 \text{ otherwise.} \end{cases}$$

$e^i_{c,p}$: Coverage block of inspection operation c in product i,

$$= \begin{cases} 1 \text{ if the inspection operation } c \text{ of product } i \text{ covers the tools of process operation } p, \\ 0 \text{ otherwise.} \end{cases}$$

$b^i_{c,k}$: Qualification of inspection tool k in inspection operation c of product i,

$$= \begin{cases} 1 \text{ if the inspection operation } c \text{ of product } i \text{ is qualified on inspection tool } k \\ 0 \text{ otherwise.} \end{cases}$$

$CapaMax_k$: Total capacity of inspection tool k, calculated in terms of total number of inspections that can be performed on k,

$X^i_{p,t}$: Production volume of product i processed on tool t in process operation p,

γ : Penalty for the lost in the process tool availability due to a Quality Task performed on process tool t.

The following variables are used in the model:

U^i_c : Decision of creating an inspection operation,

$$= \begin{cases} 1 \text{ if inspection operation } c \text{ of product } i \text{ is created} \\ 0 \text{ otherwise.} \end{cases}$$

Z^1_c : Total number of inspections performed on inspection operation c of product i,

$Y^i_{c,t}$: Number of inspections performed for product i in inspection operation c that covers process tool t,

QT_t : If the Inhibit Limit of process tool t cannot be satisfied with the locations of inspection operations a Quality Task is performed,

The main objective of our model is to determine the right number and the locations of inspection operations to minimize the exposure in terms of $W@R$ on process tools. A QT is performed on tool t if there is no inspection operations that can be located. Since a process tool is stopped while a QT is performed, the exposure is considered as the number of wafers that cannot be produced ($WDQT_t$).

$$\min \sum_{i,c,p,t} WD^i_{c,p} \cdot Y^i_{c,t} + \gamma \cdot \sum_t WDQT_t \cdot QT_t$$

s.t.

$$\sum_{i,c,p} U^i_c \cdot e^i_{c,p} \cdot hh^i_{p,t} + QT_t \geqslant 1 \qquad \forall \{t \in T | WD^i_{c,p} \leqslant IL_t\} \tag{7.1}$$

$$Y^i_{c,t} \leqslant \sum_p e^i_{c,p} \cdot hh^i_{p,t} \cdot X^i_{p,t} \qquad \forall i,c,t \tag{7.2}$$

$$QT_t \cdot IL_t + \sum_{i,c,p} Y^i_{c,t} \cdot (IL_t \cdot e^i_{c,p} \cdot hh^i_{p,t} - e^i_{c,p} \cdot hh^i_{p,t} \cdot WD^i_{c,p}) \geqslant \sum_{i,p} X^i_{p,t} \qquad \forall t \tag{7.3}$$

$$U^i_c \cdot Z^i_c \geqslant \sum_t Y^i_{c,t} \cdot hh^i_{p,t} \cdot e^i_{c,p} \qquad \forall i,c,p \tag{7.4}$$

$$CapaMax_k \geqslant \sum_{c,i} U^i_c \cdot Z^i_c \cdot b^i_{c,k} \qquad \forall k \tag{7.5}$$

$$Y^i_{c,t} \geqslant 0 \qquad \forall i,c,p \tag{7.6}$$

$$Z^i_c \geqslant 0 \qquad \forall i,c \tag{7.7}$$

$$U^i_c \in \{0,1\} \qquad \forall i,c \tag{7.8}$$

Constraints (7.1) ensure that the position of the selected inspection operation U^i_c is used for the control of tool t if the distance $WD^i_{c,p}$ is smaller than or equal to the Inhibit Limit IL_t. If there are no possible locations, a QT would be performed on process tool t. Constraints (7.2) state that the total number of inspected lots $Y^i_{c,p}$ is smaller than or equal to the quantity of product i processed on tool t. Constraints (7.3) define the number of inspections performed on production lots ($Y^i_{c,t}$) that are needed to satisfy the W@R limits (IL_t) of process tool t. Constraints (7.4) define the number of inspections performed for inspection operation c of product i if the inspection operation c is created

($U_c^i = 1$). Constraints (7.5) ensures that the total number of inspections performed for inspection operation c is lower than or equal to the capacity of inspection tool k.

This model is not linear because of the term $U_c^i \cdot Z_c^i$ in constraints (7.4) and (7.5). To linearize our model, we define a the new variable below:

$$\lambda_{i,c} = U_c^i \cdot Z_c^i \tag{7.9}$$

A set of constraints are included in the model to replace the non-linear term, where M is a big number:

$$\lambda_{i,c} \leqslant M \cdot U_c^i \qquad \forall i, c \tag{7.10}$$

$$CapaMax_k \geqslant \sum_{c,i} \lambda_{i,c} \cdot b_{c,k}^i \qquad \forall k \tag{7.11}$$

In constraints (7.10), if the inspection operation c of product i is not created ($U_c^i = 0$) $\lambda_{i,c}$ is equal to zero. When the inspection operation is created ($U_c^i = 1$), the variable $\lambda_{i,c}$ can take any value smaller than M but not larger than the total capacity of inspection tool k. The later condition is verified with constraints (7.11). Therefore, the final model is as follows:

$$\min \sum_{i,c,p,t} WD_{c,p}^i \cdot Y_{c,t}^i + \gamma \cdot \sum_t WDQT_t \cdot QT_t$$

s.t.

$$\sum_{i,c,p} U_{c,i} \cdot e_{c,p}^i \cdot hh_{p,t}^i + QT_t \geqslant 1 \qquad \forall \{t \in T | WD_{c,p}^i \leqslant IL_t\} \tag{7.12}$$

$$Y_{c,t}^i \leqslant \sum_p e_{c,p}^i \cdot hh_{p,t}^i \cdot X_{p,t}^i \qquad \forall i, c, t \tag{7.13}$$

$$QT_t \cdot IL_t + \sum_{i,c,p} Y_{c,t}^i \cdot (IL_t \cdot e_{c,p}^i \cdot hh_{p,t}^i - e_{c,p}^i \cdot hh_{p,t}^i \cdot WD_{c,p}^i) \geqslant \sum_{i,p} X_{p,t}^i \qquad \forall t \tag{7.14}$$

$$\lambda_{i,c} \geqslant \sum_t Y_{c,t}^i \cdot hh_{p,t}^i \cdot e_{c,p}^i \qquad \forall i, c, p \tag{7.15}$$

$$CapaMax_k \geqslant \sum_{c,i} \lambda_{i,c} \cdot b_{c,k}^i \qquad \forall k \tag{7.16}$$

$$\lambda_{i,c} \leqslant M \cdot U_c^i \qquad \forall i, c \tag{7.17}$$

$$\lambda_{i,c} \geqslant 0 \qquad \forall i, c, t \tag{7.18}$$

$$U_{c,i} \in \{0, 1\} \qquad \forall c, i \tag{7.19}$$

7.5 Conclusion and perspectives

In this chapter we analyze the impact of defect inspection control plans on the W@R of process tools. Experiments were performed on industrial instances with the simulation tool S5 developed by EMSE-SFL. The results showed that more inspection operations in the defect inspection control plan do not always reduce the overall W@R. The W@R reduction highly depends on the position of inspection operations and how they cover process operations. Moreover, when inspection capacity is reduced, the overlapping of inspection operations can enhance the W@R reduction. When inspection capacity is increased, the number of inspection operations is a key factor to consider. The chapter concludes with a mathematical model to optimize the location of new inspection operations within the manufacturing route. Our perspectives include the implementation and the validation of the proposed model.

General Conclusion and Perspectives

General Conclusion

This thesis was conducted within the framework of a joint collaboration between industrial and academics. We have faced the problem of how to efficiently manage and reduce the risk on process tools. The notion of risk considered in this thesis refers to the number of Wafers at Risk (W@R) on process tools regarding defect inspection operations, which corresponds to the number of wafers that are potentially impacted if a problem occurs. Hence, by limiting the W@R on process tools, the impact of excursions can be better controlled. Sampling strategies are used to find a trade-off on the number of inspections. More inspections lead to high product yields, and thus reduced costs for scraps and re-works, while fewer inspections lead to lower cycle times, and thus reduced production cost. When a static sampling strategy is used, several factors related to the dynamics of the fab cannot be handled resulting in cases of over-control or lack of control on process tools. Therefore, the sampling system for defect inspection has changed from a static sampling to a dynamic sampling strategy in the site of Rousset of STMicroelectronics. In a dynamic sampling strategy, lots are selected in real time and according to the information that can be obtained by inspecting sampled lots. Results showed that dynamic sampling strategies are more suitable for modern fabs to stay competitive by increasing yield through an efficient selection of lots to sample.

An application that dynamically selects the lots for inspection was developed. It is based on a skipping mechanism that aims at identifying the lots that can be released from inspection due to redundant information in terms of W@R. This mechanism helps to avoid inspections without added value for reducing the W@R of process tools. The W@R management highly depends on the product mix. Therefore, situations were W@R limits are satisfied may change if the product mix changes. In order to anticipate the production changes that directly affect W@R management, a model that optimizes the inspection capacity allocation was proposed. It takes into account the key factors that influence the W@R on process tools (e.g. manufacturing routes, tools qualifications, W@R limits, product mix, defect inspection control plans). It helps to identify the capacity required in the defect inspection workshop to satisfy the W@R limits on process tools. When these limits cannot be satisfied, the model gives an estimation of the additional inspection capacity that is required and the potentially uncovered tools. Results showed that not always more inspections means less risk, since the W@R reduction highly depends

on the position of inspection operations within the manufacturing routes of products and how inspection operations cover process operations. Important savings were obtained with the industrial implementation of the system, not only in terms of overall W@R reduction but also in terms of number of measurements, thanks to a better selection of lots to inspect.

Industrial Results

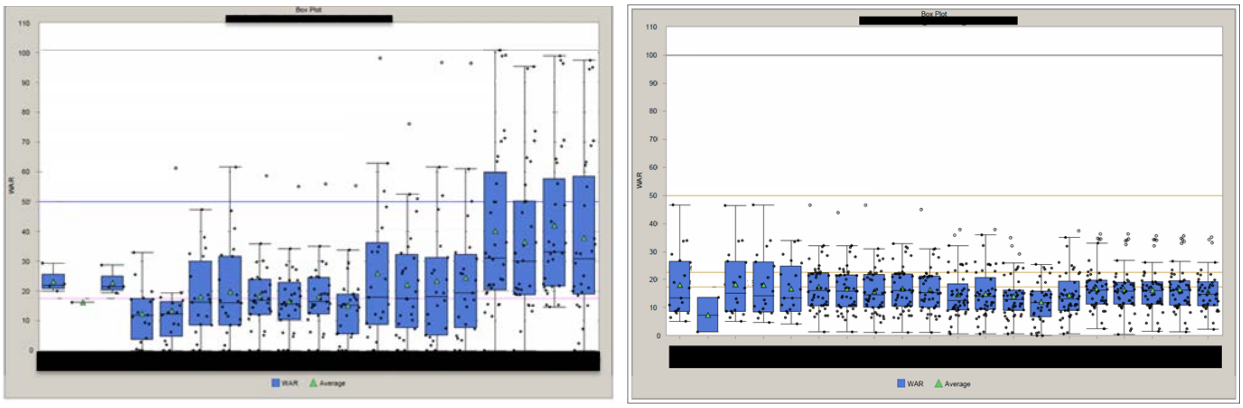
In the following, the industrial results obtained after the implementation of the W@R and dynamic sampling are discussed. The evolution of the system has been possible thanks to the collaborative work of all the participant of the "W@R Implementation" group[4]. Special thanks to the people who are in charge of the W@R defectivity project in the site of Rousset of STMicroelectronics: **Eric Tartière and Jacques Pinaton**, whose commitment and daily work guarantee the good performance of the system.

Table 7.5 presents the key metrics one year after the implementation of the various components of the project. A scrap reduction of 11% has been obtained thanks to the early detection of excursions. The scheduled Quality Tasks (QTs) were reduced by 15% thanks to the systematic control of production lots that is assured with the W@R limits. Figures 7.4 present the resulting W@R on process tools from the same area. Each boxplot represents the value of the W@R on each process tool. The cases of lack of control and over control on process tools were reduced. The number of wafers potentially impacted if a problem occurs are limited with the Inhibit Limit (IL) of each process tool. Figure 7.5 shows the average value of the IL in the fab, which has been reduced by 66% since the beginning of the project.

Table 7.5: Key Metrics of the project

Item	Results
Scrap Reduction	11%
Average W@R Inhibit Limit Reduction	66%
Quality Task (QT) Reduction	15%
Fab deployment (Covered process tools)	100%

[4]Sponsors: P. Campion and M. Le Gall. Leaders: E. Tartière and J. Pinaton.
ST Rousset Team: A. Thieullen, G. Rodriguez-Verjan, P. Palouar, S. Detivaud, B. Pennachio, J.C. Mattlin, F. Chairat, C. Klingelschmidt, V. Lemaire, J. De-selle, D. Courilleau, B. Mari, C. Giuliani, D. Viard.
EMSE-CMP SFL Team: S. Dauzère-Pérès, C. Yugma, J.L. Rouveyrol, S. Housseman.



(a) Wafer at Risk Before (b) Wafer at Risk After

Figure 7.4: W@R on process tools before and after the sampling system evolution

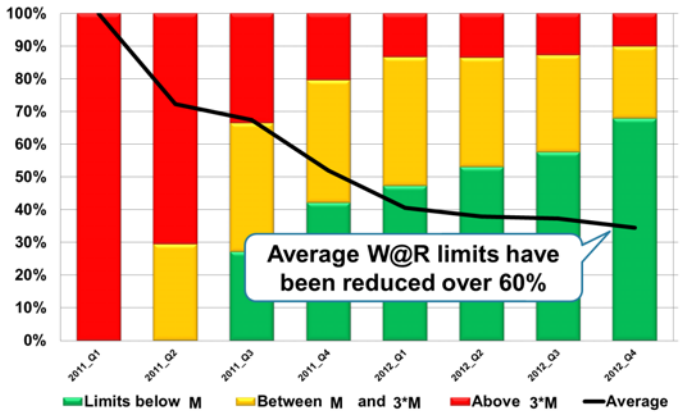


Figure 7.5: Deploiment and reduction of fab W@R Inhibit Limits

Perspectives

Different perspectives for future work are identified. Concerning the reduction of the delay to obtain the inspection results, optimizing the schedule of lots in the defect inspection area can be considered. Lots that arrive in the defect inspection area have different priorities. There are some lots for which inspection is mandatory (i.e. lots with special requirements for analysis, lots selected with static sampling) and lots that are dynamically selected (i.e. skipping is allowed). The problem can be modeled as a scheduling problem on parallel machines and the objective would be to reduce the number of tardy jobs. Mandatory lots would have a strict deadline strict and the non-mandatory lots would have tardiness. The tardiness of a lot can be defined by the remaining time before the W@R of the process tool reaches its IL. Note that the scheduling problem on parallel machines with different processing times is NP-hard.

A second perspective concerns the extension of W@R to other type of controls such as metrology. Several differences between defectivity inspection and metrology need to be considered. The objective of metrology is to control the physical and electrical properties of the wafers during fabrication. Hence, metrology steps are defined according to the property to control (e.g. Thickness, critical dimension, uniformity, electrical tests). A first challenge concerning the extension to metrology is to define the context in order to develop the W@R counters. Due to the sensibility of the product several W@R counters would be required on the same process tool. An additional difference is that the physical locations of metrology tools need to be considered. Contrary to defect inspection tools that are located in a single workshop, the metrology tools are located in different areas of the fab.

An additional perspective concerns the optimisation of the number and location of inspection operations. The optimisation problem can be seen as a maximal covering location problem, where the clients are the process tools that need to be controlled. The demands of clients are the requirements in terms of lots to inspect. Selecting a facility corresponds to selecting an inspection operation to allocate and the distance between a site and client is the average number of wafers between the process operations and the inspection operations.

Appendix A

Glossary

In the following a glossary of the terminology used in this thesis is presented. References [15], [91], [92] should be consulted for more extensive glossaries.

- APC: Advanced Process Control. A set of techniques used to control processes and machines.
- BEOL: Back End Of Line. It refers to the processes performed from the contact through completion of the wafer and prior to electrical tests.
- Brightfield: In optical detection or photolithography tools, it refers to the illumination and detection technique. The incident light and the reflected light are parallel to each other and the illuminated object appears on a bright background.
- Cluster Tool: A tool that combines several process chambers within a closed environment together with a handling robot.
- Coverage Block: In this thesis, the term coverage block refers to the set of process operations that can be controlled with an inspection operation. The coverage block is defined in the defect inspection control plan.
- Darkfield: In optical detection or photolithography tools, it refers to the illumination and detection technique. The incident light and the reflected light has an angle and the illuminated object appears on a dark field background.
- Defect inspection: Type of control that aims at detecting the defects produced on wafers during the production.
- Defect inspection control plan: Is the list of inspection operations that are performed throughout the manufacturing route of products. It states the position within the manufacturing route and the coverage block of each inspection operation.
- Dies: Individual chips cut from a wafer before they are packaged.
- Excursions: Temporal yield losses that randomly happen (in time) as a consequence of an out-of-control condition on a process tool.

- Fab: It refers to a semiconductor fabrication plant. Usually refers to the front-end process of making the devices in semiconductor wafers.

- FEOL: Front end of line. It refers to all the processes from the wafer start through contact etch.

- FOUP: Front Opened Unified Pod, it is a pod used to transport wafers, usually contains 25 wafers.

- GSI: Global Sampling Indicator. Score used to evaluate the global risk of the fab and helps to define the best set of lots to inspect.

- IL: Inhibit Limit. Is the maximum value of wafers at risk, if the W@R level achieves this limit, the process tool should be stopped and a special control performed.

- IMPROVE: Implementing Manufacturing science solutions to increase PRoductiVity and fab pErformance.

- Killer Defect: Defects that cause the chip to fail.

- Lot: It is a group of wafers that is manipulated with a FOUP. Usually it contains 25 wafers.

- Manufacturing Route: Sequence of processes that are necessary to obtain the final product.

- Qualification: It refers to the setting of a recipe on the tools, in order to be able to perform the processing or the inspection of the wafers.

- Quality Tasks: It refers to a special control performed on Non-patterned wafers. These type of controls are used to monitor contamination and surface quality of the wafer and monitor the cleanliness of tools.

- R2R: Run to Run, is a loop control technique based on equipment and lot information.

- Recipe: Set of instructions and parameters required in order to perform semiconductor processing or control on a given tool.

- RI: Risk Increase indicator. It is the the associated risk (i.e. GSI variation) of not inspecting a lot or a set of lots.

- S5: Smart Sampling, Skipping and Scheduling Simulator. It is a simulator developed by the department SFL (Science de la Fabrication et Logistique) of the Ecole de Mines de St Etienne, Microelectronic Center of Provence. It is used to compare different sampling strategies for defect inspection and based on historical data.
- SPC: Statistical Process Control, it is based on statistical tools to ensure the stability of the process.
- VM: Virtual Metrology, it is a control technique that aims to predict metrology measurements and forecast electrical and physical parameters on the wafer.
- Wafer: Slice of silicon used to manufacture semiconductor devices.
- W@R: Wafers at Risk, it is the number of wafers produced on a tool since the process date of the latest inspected lot.
- WL: Warning Limit, it is a limit after which actions have to be taken in order to control a lot that can reduce the W@R level on a process tool.
- Yield: It is the number of good produced units (e.g. wafers, dies, etc) over the total produced units. It is one of the most important metrics to evaluate the efficiency of a fab.

Appendix B

Experimental Results

Skipping Experimentation

Table B.1: Number of skipped lots and final LSI for different values of T_{Metro}

Instance	TypeAlgo	**Number of skipped lots** (Related LSI)				
		T_{Metro}				
		0.001	0.005	0.01	0.05	0.1
1	Algorithm 1	**6** (0.001279)	**12** (0.284293)	**13** (0.291487)	**19** (0.435451)	**22** (0.722157)
	Algorithm 2	**5** (0.000653)	**7** (0.004930)	**7** (0.008011)	**10** (0.049931)	**14** (0.095832)
	Algorithm 3	**5** (0.000626)	**7** (0.004930)	**8** (0.008628)	**12** (0.045453)	**15** (0.098124)
	Algorithm 4	**5** (0.000626)	**7** (0.004930)	**8** (0.008628)	**12** (0.035398)	**15** (0.093280)
	Algorithm 5	**5** (0.000626)	**7** (0.004219)	**8** (0.007871)	**12** (0.035398)	**15** (0.093280)
2	Algorithm 1	**4** (0.070563)	**10** (0.131550)	**11** (0.167974)	**13** (0.194752)	**15** (0.312755)
	Algorithm 2	**3** (0.000882)	**5** (0.004762)	**7** (0.007275)	**10** (0.047746)	**11** (0.084170)
	Algorithm 3	**3** (0.000115)	**6** (0.003970)	**7** (0.006507)	**10** (0.046978)	**11** (0.083402)
	Algorithm 4	**3** (0.000115)	**6** (0.003970)	**7** (0.006507)	**10** (0.043171)	**11** (0.079595)
	Algorithm 5	**3** (0.000115)	**6** (0.003970)	**7** (0.006507)	**10** (0.043171)	**11** (0.079595)
3	Algorithm 1	**4** (0.002080)	**9** (0.099745)	**9** (0.099745)	**14** (0.209593)	**14** (0.209593)
	Algorithm 2	**1** (0.000989)	**3** (0.004692)	**6** (0.009018)	**9** (0.047786)	**12** (0.093494)
	Algorithm 3	**2** (0.000532)	**5** (0.003150)	**7** (0.007983)	**10** (0.045235)	**12** (0.091388)
	Algorithm 4	**2** (0.000532)	**5** (0.003150)	**7** (0.007983)	**10** (0.045235)	**12** (0.091388)
	Algorithm 5	**2** (0.000532)	**5** (0.003150)	**7** (0.007983)	**10** (0.045235)	**12** (0.091388)
4	Algorithm 1	**0** (0.000000)	**2** (0.005700)	**3** (0.013162)	**6** (0.084101)	**6** (0.084101)
	Algorithm 2	**0** (0.000000)	**1** (0.004455)	**2** (0.005700)	**4** (0.041514)	**6** (0.084101)
	Algorithm 3	**0** (0.000000)	**1** (0.001171)	**2** (0.005700)	**4** (0.032991)	**6** (0.084101)
	Algorithm 4	**0** (0.000000)	**1** (0.001171)	**2** (0.005700)	**4** (0.032991)	**6** (0.084101)
	Algorithm 5	**0** (0.000000)	**1** (0.001171)	**2** (0.005700)	**4** (0.032991)	**6** (0.084101)
5	Algorithm 1	**4** (0.001457)	**5** (0.005294)	**5** (0.005294)	**11** (0.166730)	**12** (0.227748)
	Algorithm 2	**3** (0.000493)	**4** (0.001457)	**5** (0.005294)	**7** (0.048881)	**7** (0.091766)
	Algorithm 3	**3** (0.000493)	**4** (0.001457)	**5** (0.005294)	**7** (0.038333)	**9** (0.081920)
	Algorithm 4	**3** (0.000493)	**4** (0.001457)	**5** (0.005294)	**7** (0.038333)	**9** (0.081920)
	Algorithm 5	**3** (0.000493)	**4** (0.001457)	**5** (0.005294)	**7** (0.038333)	**9** (0.081920)
6	Algorithm 1	**0** (0.000000)	**4** (0.008165)	**5** (0.017027)	**9** (0.125484)	**10** (0.211769)
	Algorithm 2	**0** (0.000000)	**2** (0.003708)	**4** (0.008165)	**6** (0.027115)	**8** (0.096540)
	Algorithm 3	**0** (0.000000)	**2** (0.003396)	**4** (0.008165)	**6** (0.027115)	**8** (0.083012)
	Algorithm 4	**0** (0.000000)	**2** (0.003396)	**4** (0.008165)	**6** (0.027115)	**8** (0.083012)
	Algorithm 5	**0** (0.000000)	**2** (0.003396)	**4** (0.008165)	**6** (0.027115)	**8** (0.083012)

Instance	TypeAlgo	**Number of skipped lots** (Related LSI)				
		T_{Metro}				
		0.001	0.005	0.01	0.05	0.1
7	Algorithm 1	**7** (0.001262)	**13** (0.396196)	**14** (0.401685)	**18** (0.522087)	**19** (0.595998)
	Algorithm 2	**6** (0.000512)	**7** (0.004225)	**8** (0.009715)	**12** (0.045763)	**14** (0.099489)
	Algorithm 3	**6** (0.000385)	**7** (0.001262)	**9** (0.009661)	**12** (0.038781)	**14** (0.081529)
	Algorithm 4	**6** (0.000385)	**7** (0.001262)	**9** (0.009661)	**12** (0.037227)	**14** (0.081529)
	Algorithm 5	**6** (0.000385)	**8** (0.004317)	**9** (0.008450)	**12** (0.037227)	**14** (0.081529)
8	Algorithm 1	**1** (0.000000)	**2** (0.003037)	**2** (0.003037)	**5** (0.091761)	**5** (0.091761)
	Algorithm 2	**1** (0.000000)	**2** (0.003037)	**2** (0.003037)	**4** (0.042092)	**5** (0.091761)
	Algorithm 3	**1** (0.000000)	**2** (0.003037)	**2** (0.003037)	**4** (0.042092)	**5** (0.091761)
	Algorithm 4	**1** (0.000000)	**2** (0.003037)	**2** (0.003037)	**4** (0.042092)	**5** (0.091761)
	Algorithm 5	**1** (0.000000)	**2** (0.003037)	**2** (0.003037)	**4** (0.042092)	**5** (0.091761)
9	Algorithm 1	**8** (0.001878)	**10** (0.007715)	**10** (0.007715)	**10** (0.007715)	**12** (0.087647)
	Algorithm 2	**4** (0.000447)	**6** (0.004665)	**10** (0.007715)	**10** (0.007715)	**12** (0.087647)
	Algorithm 3	**4** (0.000000)	**8** (0.004741)	**10** (0.007715)	**10** (0.007715)	**12** (0.087647)
	Algorithm 4	**4** (0.000000)	**8** (0.004741)	**10** (0.007715)	**10** (0.007715)	**12** (0.087647)
	Algorithm 5	**4** (0.000000)	**8** (0.004741)	**10** (0.007715)	**10** (0.007715)	**12** (0.087647)
10	Algorithm 1	**9** (0.111135)	**13** (0.120418)	**13** (0.12418)	**19** (0.488045)	**21** (0.682193)
	Algorithm 2	**6** (0.000985)	**9** (0.004572)	**11** (0.009628)	**13** (0.041486)	**13** (0.098187)
	Algorithm 3	**7** (0.000724)	**9** (0.003832)	**10** (0.006745)	**13** (0.029081)	**15** (0.077084)
	Algorithm 4	**7** (0.000724)	**9** (0.003832)	**11** (0.008887)	**13** (0.029081)	**15** (0.077084)
	Algorithm 5	**7** (0.000724)	**9** (0.003759)	**11** (0.008887)	**13** (0.029081)	**15** (0.077084)
11	Algorithm 1	**6** (0.001644)	**9** (0.29586)	**13** (0.128561)	**13** (0.128561)	**13** (0.128561)
	Algorithm 2	**4** (0.000670)	**7** (0.003612)	**8** (0.009209)	**11** (0.043792)	**12** (0.074949)
	Algorithm 3	**5** (0.000951)	**7** (0.003612)	**8** (0.006475)	**11** (0.043792)	**12** (0.074949)
	Algorithm 4	**5** (0.000951)	**7** (0.003612)	**8** (0.006475)	**11** (0.043792)	**12** (0.074949)
	Algorithm 5	**5** (0.000951)	**7** (0.003612)	**8** (0.006475)	**11** (0.043792)	**12** (0.074949)
12	Algorithm 1	**0** (0.000000)	**2** (0.003039)	**4** (0.024960)	**6** (0.074434)	**7** (0.124853)
	Algorithm 2	**0** (0.000000)	**2** (0.003039)	**2** (0.009072)	**4** (0.044788)	**6** (0.074434)
	Algorithm 3	**0** (0.000000)	**2** (0.003039)	**2** (0.009072)	**5** (0.047339)	**6** (0.074434)
	Algorithm 4	**0** (0.000000)	**2** (0.003039)	**2** (0.009072)	**5** (0.047339)	**6** (0.074434)
	Algorithm 5	**0** (0.000000)	**2** (0.003039)	**2** (0.009072)	**5** (0.047339)	**6** (0.074434)

Results capacity planning

Table B.2: Parameters of the scenario parameters for $DefP_k$ variation

Factors	
Product Mix	Mix 1
Limits	C
QTP	10
Alpha	0.2
Reserved Capacity	0

Table B.3: Variation of $DefP_k$ Penalties

$DefP_k$ Values				Total Utilization Rates Static and Dynamic Sampling				QTs
T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**	
1	1	1	1	100.0%	56.2%	23.3%	100.0%	45
1	1	1	5	100.0%	61.1%	31.2%	70.9%	45
1	1	1	10	100.0%	64.0%	40.3%	62.8%	45
1	1	5	1	100.0%	62.3%	15.3%	100.0%	45
1	1	5	5	100.0%	64.1%	15.3%	93.0%	45
1	1	5	10	100.0%	63.8%	18.6%	84.0%	45
1	1	10	1	100.0%	62.3%	15.3%	100.0%	45
1	1	10	5	100.0%	64.1%	15.3%	93.0%	45
1	1	10	10	100.0%	67.6%	15.3%	89.9%	45
1	5	1	1	100.0%	42.0%	95.3%	100.0%	45
1	5	1	5	100.0%	42.0%	95.3%	100.0%	45
1	5	1	10	100.0%	46.6%	100.0%	66.8%	45
1	5	5	1	100.0%	56.2%	23.3%	100.0%	45
1	5	5	5	100.0%	56.2%	23.3%	100.0%	45
1	5	5	10	100.0%	58.9%	23.3%	88.2%	45
1	5	10	1	100.0%	60.2%	17.4%	100.0%	45
1	5	10	5	100.0%	60.2%	17.4%	100.0%	45
1	5	10	10	100.0%	60.2%	17.4%	100.0%	45
1	10	1	1	100.0%	41.6%	100.0%	100.0%	45
1	10	1	5	100.0%	41.6%	100.0%	100.0%	45
1	10	1	10	100.0%	41.6%	100.0%	100.0%	45
1	10	5	1	100.0%	51.3%	39.4%	100.0%	45
1	10	5	5	100.0%	51.2%	39.5%	100.0%	45
1	10	5	10	100.0%	51.3%	39.3%	100.0%	45
1	10	10	1	100.0%	56.2%	23.3%	100.0%	45
1	10	10	5	100.0%	56.2%	23.3%	100.0%	45
1	10	10	10	100.0%	56.2%	23.3%	100.0%	45
5	1	1	1	86.6%	65.3%	48.4%	100.0%	45
5	1	1	5	88.1%	65.1%	64.9%	70.9%	45
5	1	1	10	87.3%	65.5%	74.9%	66.8%	45
5	1	5	1	98.8%	64.1%	15.3%	100.0%	45
5	1	5	5	100.0%	64.1%	15.3%	93.0%	45
5	1	5	10	100.0%	63.8%	18.6%	84.0%	45
5	1	10	1	98.8%	64.1%	15.3%	100.0%	45
5	1	10	5	100.0%	64.1%	15.3%	93.0%	45

$DefP_k$ Values				Total Utilization Rates Static and Dynamic Sampling				QTs
T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**	
5	1	10	10	100.0%	67.6%	15.3%	89.9%	45
5	5	1	1	100.0%	42.0%	95.3%	100.0%	45
5	5	1	5	100.0%	42.0%	95.3%	100.0%	45
5	5	1	10	100.0%	46.6%	100.0%	66.8%	45
5	5	5	1	100.0%	56.2%	23.3%	100.0%	45
5	5	5	5	100.0%	56.2%	23.3%	100.0%	45
5	5	5	10	100.0%	58.9%	23.3%	88.2%	45
5	5	10	1	100.0%	60.2%	17.4%	100.0%	45
5	5	10	5	100.0%	60.2%	17.4%	100.0%	45
5	5	10	10	100.0%	60.2%	17.4%	100.0%	45
5	10	1	1	100.0%	41.6%	100.0%	100.0%	45
5	10	1	5	100.0%	41.6%	100.0%	100.0%	45
5	10	1	10	100.0%	41.6%	100.0%	100.0%	45
5	10	5	1	100.0%	51.2%	39.6%	100.0%	45
5	10	5	5	100.0%	51.3%	39.3%	100.0%	45
5	10	5	10	100.0%	51.3%	39.4%	100.0%	45
5	10	10	1	100.0%	56.2%	23.3%	100.0%	45
5	10	10	5	100.0%	56.2%	23.3%	100.0%	45
5	10	10	10	100.0%	56.2%	23.3%	100.0%	45
10	1	1	1	86.6%	65.3%	48.4%	100.0%	45
10	1	1	10	86.6%	64.9%	80.1%	66.8%	45
10	1	1	5	86.6%	65.1%	73.5%	70.9%	45
10	1	5	1	94.3%	64.3%	23.3%	100.0%	45
10	1	5	5	94.3%	64.3%	23.3%	100.0%	45
10	1	5	10	96.3%	64.3%	23.3%	89.0%	45
10	1	10	1	98.8%	64.1%	15.3%	100.0%	45
10	1	10	5	98.8%	64.1%	15.3%	100.0%	45
10	1	10	10	100.0%	67.6%	15.3%	89.9%	45
10	5	1	1	91.6%	49.8%	93.1%	100.0%	45
10	5	1	5	91.6%	49.8%	93.1%	100.0%	45
10	5	1	10	91.6%	54.3%	100.0%	66.8%	45
10	5	5	1	100.0%	56.2%	23.3%	100.0%	45
10	5	5	5	100.0%	56.2%	23.3%	100.0%	45
10	5	5	10	100.0%	58.9%	23.3%	88.2%	45
10	5	10	1	100.0%	60.2%	17.4%	100.0%	45
10	5	10	5	100.0%	60.2%	17.4%	100.0%	45
10	5	10	10	100.0%	60.2%	17.4%	100.0%	45
10	10	1	1	100.0%	41.6%	100.0%	100.0%	45
10	10	1	5	100.0%	41.6%	100.0%	100.0%	45
10	10	1	10	100.0%	41.6%	100.0%	100.0%	45
10	10	5	1	100.0%	51.3%	39.3%	100.0%	45
10	10	5	5	100.0%	51.3%	39.3%	100.0%	45
10	10	5	10	100.0%	51.3%	39.3%	100.0%	45
10	10	10	1	100.0%	56.2%	23.3%	100.0%	45
10	10	10	5	100.0%	56.2%	23.3%	100.0%	45
10	10	10	10	100.0%	56.2%	23.3%	100.0%	45

Bibliography

[1] L. Monch, J. Fowler, S. Dauzère-Pérès, S. Mason, and O. Rose, "A survey problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations," *Journal of Scheduling*, vol. 14, no. 6, pp. 583–599, 2011. (Cité en pages 3 et 27.)

[2] H. Bettayeb, "Conception et évaluation des plans de surveillance basés sur le risque, limitation des incertitudes qualité avec des ressources limitées de maîtrise," Ph.D. dissertation, Université de Grenoble, laboratoire G-Scop, 2012. (Cité en page 3.)

[3] J. Nduhura Munga, G. Rodriguez-Verjan, S. Dauzère-Pérès, C. Yugma, P. Vialletelle, and J. Pinaton, "Literature Review on Sampling Techniques in Semiconductor Manufacturing," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 26, 2013, pp. 188–195. (Cité en pages 3, 36, 39, 40 et 44.)

[4] R. K. Nurani, R. Akella, and A. J. Strojwas, "In-Line Defect Sampling Methodology in Yield Management: an Integrated Framework," *IEEE Transactions on Semiconductor Manufacturing*, vol. 9, no. 4, pp. 506–517, 1996. (Cité en pages 3 et 42.)

[5] H. Rau and K. Cho, "Layer modelling for the inspection allocation problem in reentrant production systems," *Internation Journal of Production Research*, vol. 43, no. 17, pp. 3633–3655, 2005. (Cité en pages 3, 40, 42 et 44.)

[6] W. W. Kuo, R. Akella, and D. Fletcher, "Adaptive Sampling for Effective Multi-Layer Defect Monitoring," in *IEEE International Symposium on Semiconductor Manufacturing*, 1997, pp. 289–293. (Cité en pages 3 et 42.)

[7] C. Mouli and M. J. Scott, "Adaptive Metrology Sampling Techniques Enabling Higher Precision in Variability Detection and Control," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 2007, pp. 12–17. (Cité en pages 3, 42 et 43.)

[8] R. P. Good and M. A. Purdy, "An MILP Approach to Wafer Sampling and Selection," *IEEE Transactions on Semiconductor Manufacturing*, vol. 20, no. 4, pp. 400–407, 2007. (Cité en pages 3, 67 et 69.)

[9] S. Dauzère-Pérès, J.-L. Rouveyrol, C. Yugma, and P. Vialletelle, "A Smart Sampling Algorithm to Minimize Risk Dynamically," in *Proceedings of the 2010 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 2010, pp. 307–310. (Cité en pages 3, 10, 41, 42, 43, 44, 53 et 54.)

[10] M. Keefer, R. Pinto, C. Dennison, and J. Turlo, *The Role of Metrology and Inspection in Semiconductor Manufacturing.* Noyes Publications, 2002, 2th Edition. (Cité en pages 6, 28, 30, 31 et 35.)

[11] J. W. Bean, "Variation reduction in a wafer fabrication line through inspection optimization," Master's thesis, Massachussets Institute of Technology, USA, 1997. (Cité en pages 6, 35, 41 et 42.)

[12] J. Nduhura Munga, "Implementing and optimizing dynamic control plans in semiconductor manufacturing," Ph.D. dissertation, Ecole Nationale Supérieure des Mines de Saint-Etienne, Gardanne, France, 2012. (Cité en pages 7, 9, 10, 37, 43, 44, 53 et 55.)

[13] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org (Cité en pages 12 et 65.)

[14] S. Housseman, S. Dauzère-Pérès, G. Rodriguez-Verjan, and J. Pinaton, "Smart dynamic sampling for wafer at risk reduction in semiconductor manufacturing," in *IEEE/CASE International Conference on Automation Science and Engineering (Submitted)*, 2014. (Cité en pages 12 et 71.)

[15] M. Quirk and J. Serda, *Semiconductor Manufacturing technology.* Prentice Hall, 2001, 1st Edition. (Cité en pages vii, 25, 27, 29, 33 et 117.)

[16] M. Riordan, L. Hoddeson, and C. Herring, "The invention of the transistor," *Rev. Mod. Phys.*, vol. 71, pp. S336–S345, 1999. (Cité en page 25.)

[17] C. Mack, "Fifty years of moore's law," *IEEE Transactions on Semiconductor Manufacturing*, vol. 24, no. 2, pp. 202–207, 2011. (Cité en pages 25 et 28.)

[18] R. Guldi, "In-Line Defect Reduction from a Historical Perspective and Its Implications for Future Integrated Circuit Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 17, no. 4, pp. 629–640, 2004. (Cité en pages 25, 28, 31, 40 et 86.)

[19] G. S. May and C. J. Spanos, "Fundamentals of Semiconductor Manufacturing and Process Control." John Wiley & Sons, 2006. (Cité en pages 26 et 33.)

[20] L. Rubin and J. Poate, "Ion implantation in silicon technology," *Industrial Physicist*, vol. 9, no. 3, pp. 12–15, 2003. (Cité en page 28.)

[21] L. A. Larson, J. M. Williams, and M. I. Current, "Ion implantation for semiconductor doping and materials modification," *Reviews of Accelerator Science and Technology*, vol. 4, no. 1, pp. 11–40, 2011. (Cité en page 28.)

[22] L. Monch, J. Fowler, and S. Mason, *Production Planning and control for semiconductor wafer fabrication facilities*. Springer New York, 2013, 1th Edition. (Cité en page 28.)

[23] J. Gupta, R. Ruiz, J. Fowler, and S. Mason, "Operational planning and control of semiconductor wafer production," *Production Planning & Control*, vol. 17, no. 7, pp. 639–647, 2006. (Cité en page 28.)

[24] W. Arden, M. Brillouët, P. Cogez, M. Graef, B. Huizing, R. Mahnkopf, J. Pelka, J. Pfeiffer, A. Rouzaud, M. Tartagni, C. Hoof, and J. Wagner, *Towards a "More than Moore" Roadmap*, 2011. (Cité en page 28.)

[25] N. Geng and Z. Jiang, "A review on strategic capacity planning for the semiconductor manufacturing industry," *International Journal of Production Research*, vol. 47, no. 13, pp. 3639–3655, 2009. (Cité en page 28.)

[26] R. Leachman, *Competitive Semiconductor Manufacturing: Summary Report on Findings from Benchmarking Eight-inch, sub-350nm Wafer Fabrication Lines*. Competitive Semiconductor Manufacturing Program, 2002. (Cité en page 28.)

[27] S. Bassetto, "Contribution Ã la quantification et Ã l'amélioration des moyens de production," Ph.D. dissertation, Ecole Nationale Supérieure d'Arts et Métiers, Metz, France, 2005. (Cité en page 28.)

[28] A.-J. Su, C.-C. Yu, and B. A. Ogunnaike, "On the Interaction between Measurement Strategy and Control Performance in Semiconductor Manufacturing," *Journal of Process Control*, vol. 18, no. 3-4, pp. 266–276, 2008. (Cité en page 28.)

[29] C. J. Spanos, "Statistical process control in semiconductor manufacturing," *Procedings of the IEEE*, vol. 80, no. 6, pp. 819–830, 1992. (Cité en page 29.)

[30] K. Stoddard, P. Crouch, M. Kozicki, and K. Tsakalis, "Application of feed-fordward and adaptive feed back control to semiconductor device manufacturing," in *Proceedings of the american control conference*, vol. 1, 1994, pp. 892–896. (Cité en page 29.)

[31] T. Edgar, S. Butler, W. Campbell, C. Pfeiffer, C. Bode, S. Hwang, K. Balakrishnan, and J. Hahn, "Automatic control in microelectronic manufacturing: Practices, challenges and possibilities," *Automatica*, vol. 30, pp. 1567–1603, 2000. (Cité en page 29.)

[32] P. Kang, H. Lee, S. Cho, D. Kim, J. Park, C. Park, and S. Doh, "A virtual metrology system for semiconductor manufacturing," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12 554–12 561, 2009. (Cité en page 30.)

[33] J. Besnard, D. Gleispach, H. Gris, A. Ferreira, A. Roussy, C. Kernaflen, and G. Hayderer, "Virtual metrology modeling for cvd film thickness," *International Journal of Control Science and Engineering*, vol. 2, no. 3, pp. 26–33, 2012. (Cité en page 30.)

[34] A. Ferreira, A. Roussy, and L. Conde, "Virtual Metrology Models for Predicting Physical Measurement in Semiconductor Manufacturing," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 2009, pp. 149–154. (Cité en page 30.)

[35] K. W. Tobin, "Integrated applications of inspection data in the semiconductor manufacturing environment," in *Proc. SPIE 4275, Metrology-based Control for manufacturing environment*, vol. 31, 2001. (Cité en page 31.)

[36] S. Gleason, K. W. Tobin, and T. Karnowski, "An integrated spatial signature analysis and automatic defect classification system," *191st Meeting Electrochemical Society, Inc*, 1997. (Cité en pages 31 et 32.)

[37] C. Johnzén, S. Dauzère-Pérès, and P. Vialletelle, "Flexibility Measures for Qualification Management in Wafer Fabs," *Production Planning & Control*, vol. 22, no. 1, pp. 81–90, 2011. (Cité en pages 33 et 75.)

[38] D. Pepper, O. Moreau, and G. Hennion, "Inline automated defect classification: a novel approach to defect management," in *EEE/SEMI Conference and Workshop on Advanced Semiconductor Manufacturing*, 2005, pp. 43–48. (Cité en page 35.)

[39] A. Klemmt and L. Mönch, "Scheduling jobs with time constraints between consecutive process steps in semiconductor manufacturing," in *Proceedings of 2012 Winter Simulation Conference*, 2012, pp. 2173–2182. (Cité en page 35.)

[40] D. Gudmundsson, "Inspection and metrology capacity allocation in the full production and ramp phases of semiconductor manufacturing," Ph.D. dissertation, University of California, Berkeley, USA, 2005. (Cité en pages 37, 40, 41 et 42.)

[41] R. Leachman and S. D. Ding, "Excursion Yield Loss and Cycle Time Reduction in Semiconductor Manufacturing," *Transactions on Automation Science and Engineering*, vol. 8, no. 1, pp. 112–117, 2011. (Cité en pages 37 et 40.)

[42] J. Nduhura Munga, S. Dauzère-Pérès, P. Vialletelle, and C. Yugma, "Dynamic Management of Controls in Semiconductor Manufacturing," in *Proceedings of the 22nd Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 2011, pp. 18–23. (Cité en pages 37, 41, 42, 43 et 73.)

[43] T. Raz, "A survey of models for allocating inspection effort in multistage production systems," *Journal of Quality Technology*, vol. 18, no. 4, pp. 239–247, 1986. (Cité en page 39.)

[44] S. Mandroli, A. Shrivastava, and Y. Ding, "A survey of inspection strategy and sensor distribution studies in discrete-part manufacturing processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 38, no. 4, pp. 309–328, 2006. (Cité en page 39.)

[45] J. Shi and S. Zhou, "Quality control and improvement for multistage systems : A survey," *IIE Transactions*, vol. 41, pp. 744–753, 2009. (Cité en page 39.)

[46] A. Shetwan, V. Vitanov, and B. Tjahjono, "Allocation of qualigy controls stations in multistage manufacturing systems," *Computer and Industrial Engineering*, vol. 60, no. 4, pp. 473–484, 2011. (Cité en pages 39 et 40.)

[47] R. Williams, D. Gudmundsson, K. Monahan, and J. G. Shanthikumar, "Optimized Sample Planning for Wafer Defect Inspection," in *IEEE International Symposium on Semiconductor Manufacturing*, 1999, pp. 43–46. (Cité en pages 39, 41 et 42.)

[48] G. Hall, R. Young, M. Dunne, and M. Muro, "A quality-cost model of in-line inspections for excursion detection and reduction," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 2008, pp. 273–277. (Cité en pages 39, 40, 42 et 44.)

[49] A. Bousetta and A. J. Cross, "Adaptive Sampling Methodology for In-Line Defect Inspection," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 2005, pp. 25–31. (Cité en pages 39, 40, 41, 42 et 43.)

[50] S. Van Volsem, W. Dullaert, and H. Van Landeghem, "An evolutionary algorithm and discrete event simulation for optimizing inspection strategies for multi-stage processes," *European Journal of Operational Research*, vol. 179, no. 3, pp. 621–633, 2007. (Cité en pages 39 et 42.)

[51] M. Penn and T. Raviv, "Optimizing the quality control station configuration," *Naval Research Logistics (NRL)*, vol. 54, no. 3, pp. 301–314, 2007. (Cité en page 39.)

[52] R. C. Elliott, R. K. Nurani, D. Gudmundsson, M. Preil, R. Nasongkhla, and J. G. Shanthikumar, "Critical Dimension Sample Planning for Sub-0.25 Micron Processes," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 1999, pp. 139–142. (Cité en pages 40, 41 et 42.)

[53] Y. Narahari and L. Khan, "Modeling Re-entrant Manufacturing Systems with Inspections," *Journal of Manufacturing Systems*, vol. 15, no. 6, pp. 367–378, 1996. (Cité en pages 40, 42 et 44.)

[54] Y. Shiau, "Inspection resource assignment in a multistage manufacturing system with an inspection error model," *International Journal of Production Research*, vol. 40, no. 8, pp. 1787–1806, 2002. (Cité en pages 40 et 42.)

[55] J. Lee and S. Unnikrishnan, "Planning quality inspection operations in multistage manufacturing systems with inspection errors," *International Journal of Production Research*, vol. 36, pp. 141–155, 1998. (Cité en page 40.)

[56] I. Tirkel, N. Reshef, and G. Rabinowitz, "In-Line Inspection Impact on Cycle Time and Yield," *IEEE Transactions on Semiconductor Manufacturing*, vol. 22, no. 4, pp. 491–498, 2009. (Cité en pages 40 et 45.)

[57] A. Wein, "Random yield and scrap in a multistage batch manufacturing environment," *Operation Research*, vol. 40, no. 3, pp. 551–563, 1992. (Cité en page 40.)

[58] L. Song-bor, R. L. Ta-Yung, L. Janson, and C. Yu-Ching, "A Capacity-Dependence Dynamic Sampling Strategy," in *Proceedings of the 2003 IEEE International Symposium on Semiconductor Manufacturing*, 2003, pp. 312–314. (Cité en page 41.)

[59] C. Mouli, "Adaptive Sampling Technology - the next step to factory efficiency," *EuroAsia Semiconductor Magazine*, pp. 1–3, 2005. (Cité en page 41.)

[60] M. Purdy, C. Nicksic, and K. Lensing, "Method for Efficiently Managing Metrology Queues," in *Proceedings of the 2005 IEEE International Symposium on Semiconductor Manufacturing*, 2005, pp. 71–74. (Cité en pages 41, 42 et 43.)

[61] S. Sun and K. Johnson, "Method and System for Determining Optimal Wafer Sampling in Real-Time Inline Monitoring and Experimental Design," in *IEEE International Symposium on Semiconductor Manufacturing*, 2008, pp. 44–47. (Cité en pages 41, 42 et 43.)

[62] T. Raz and M. Kaspi, "Location and sequencing of imperfect inspection operations in serial multi-stage production systems," *International Journal of Production Research*, vol. 29, no. 8, pp. 1645–1659, 1991. (Cité en page 42.)

[63] C.-F. Chien, S.-C. Hsu, S. Peng, and C.-H. Wu, "A Cost-Based Heuristic for Statistically Determining Sampling Frequency in a Wafer Fab," in *Semiconductor Manufacturing Technology Workshop*, 2000, pp. 217 –229. (Cité en pages 41 et 42.)

[64] V. Kakade, J. F. Valenzuela, and J. Smith, "An optimization model for selective inspection in serial manufacturing systems," *International journal of production research*, vol. 42, no. 18, pp. 3891–3909, 2004. (Cité en page 42.)

[65] A. Vaghefi and V. Sarhangian, "Contribution of simulation to the optimization of inspection plans for multi-stage manufacturing systems," *Computers & Industrial Engineering*, vol. 57, pp. 1226–1234, 2009. (Cité en page 42.)

[66] R. K. Nurani, R. Akella, A. J. Strojwas, R. Wallace, M. G. McIntyre, J. Shields, and I. Emami, "Development of an Optimal Sampling Strategy for Wafer Inspection," in *Proceedings of the 1994 International Symposium on Semiconductor Manufacturing*, 1994, pp. 143–146. (Cité en pages 41 et 42.)

[67] H. Emmons and G. Rabinowitz, "Inspection allocation for multistage deteriorating production systems," *IIE Transactions*, vol. 34, no. 12, pp. 1031–1041, 2002. (Cité en page 42.)

[68] H. Rau and K. Cho, "Genetic algorithm modeling for the inspection allocation in reentrant production systems," *Expert Systems with Applications*, vol. 36, no. 8, pp. 11 287–11 295, 2009. (Cité en pages 42 et 44.)

[69] A. Verduzco, J. R. Villalobos, and B. Vega, "Information-based inspection allocation," *Journal of Manufacturing Systems*, vol. 2, no. 1, pp. 13–22, 2001. (Cité en pages 42 et 43.)

[70] A. Holfeld, R. Barlović, and P. Good, "A Fab-Wide APC Sampling Application," *IEEE Transactions on Semiconductor Manufacturing*, vol. 20, no. 4, pp. 393–399, 2007. (Cité en pages 42 et 43.)

[71] M. Pfeffer, R. Oechsner, S. Eckert, A. Hartmann, H. Gold, G. Biebl, and J. Kaspar, "Predictive sampling approach to dynamically optimize defect density control operations," in *Advanced Semiconductor Manufacturing Conference (ASMC)*, 2012, pp. 367–370. (Cité en pages 42 et 43.)

[72] B. Bettayeb, S. Bassetto, P. Vialletelle, and M. Tollenaere, "Quality and exposure control in semiconductor manufacturing. part I : Modelling," *International Journal of Production Research*, vol. 50, no. 23, pp. 6835–6851, 2012. (Cité en pages 42 et 43.)

[73] ——, "Quality and exposure control in semiconductor manufacturing. part II : Evaluation," *International Journal of Production Research*, vol. 50, no. 23, pp. 6852–6869, 2012. (Cité en pages 42 et 44.)

[74] J. Nduhura Munga, S. Dauzère-Pérès, C. Yugma, and P. Vialletelle, "A Mathematical Programming Approach for Determining Control plans in Semiconductor Manufacturing," in *Proceedings of the International Conference on Industrial Engineering and Systems Management*, 2011, p. 9 pages. (Cité en pages 42, 43 et 49.)

[75] M. Purdy, "Dynamic, Weight-Based Sampling Algorithm," in *IEEE International Symposium on Semiconductor Manufacturing*, 2007, pp. 1–4. (Cité en page 42.)

[76] B. Hyung Joo Lee, "Advanced process control and optimal sampling in semiconductor manufacturing," Ph.D. dissertation, The University of Texas at Austin, USA, 2008. (Cité en page 42.)

[77] C.-T. Lin, C.-C. Huang, C.-Y. Yang, Y.-W. Wu, C.-S. Lu, P.-Y. Tsai, C.-M. Huang, and Y.-L. Wang, "Defect Intelligent Sampling System," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 2010, pp. 162–164. (Cité en page 42.)

[78] G. L. Rodriguez-Verjan, S. Dauzère-Pérès, S. Housseman, and J. Pinaton, "Skipping algorithms for defect inspection using a dynamic control strategy in semiconductor manufacturing," in *Proceedings of MASM 2013 (9th International Conference on Modeling and Analysis of Semiconductor Manufacturing), included in the 2013 Winter Simulation Conference*, 2013, pp. 3684 – 3695. (Cité en pages 42, 44 et 53.)

[79] B. Bettayeb, P. Vialletelle, S. Bassetto, and M. Tollenaere, "Optimized design of control plans based on risk exposure and resources capabilities," *International Symposium on Semiconductor Manufacturing*, 2010. (Cité en pages 43 et 103.)

[80] W. Shin, S. Hart, and H. Lee, "Strategic allocation of inspection stations for a flow assembly line: a hybrid procedure," *IIE Transactions*, vol. 27, no. 6, pp. 707–715, 1995. (Cité en page 44.)

[81] G. L. Rodriguez-Verjan, E. Tartère, J. Pinaton, S. Dauzère-Pérès, and A. Thieullen, "Dispatching of lots to dynamically reduce the wafers at risk in semiconductor man-

ufacturing," in *International Conference on Automation Science and Engineering (IEEE CASE)*, 2012, pp. 920–923. (Cité en page 48.)

[82] B. Foster, D. Meyersdorf, J. Padillo, and R. Brenner, "Simulation of test wafer consumption in a semiconductor facility," *IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pp. 298–302, 1998. (Cité en pages 49 et 81.)

[83] D. Chrusciel and D. Field, "Success factors in dealing with significant change in an organization," *Business Process Management Journal*, vol. 12, no. 4, pp. 503 – 516, 2006. (Cité en page 50.)

[84] E. J. Umble, R. R. Haft, and M. M. Umble, "Enterprise resource planning : Implementation procedures and critical success factors," *European Journal of Operational Research*, vol. 146, pp. 241–257, 2003. (Cité en page 50.)

[85] C. Yugma, S. Dauzère-pérès, J. Rouveyrol, P. Vialletelle, J. Pinaton, and C. Relliaud, "A smart sampling scheduling and skipping simulator ans its evaluation on real data sets," in *Proceedings of MASM 2011 (7th International Conference on Modeling and Analysis of Semiconductor Manufacturing), included in the 2011 Winter Simulation Conference*, 2011, pp. 1908–1917. (Cité en page 70.)

[86] G. L. Rodriguez-Verjan, S. Dauzère-Pérès, and J. Pinaton, "A mathematical model for estimating defect inspection capacity with a dynamic control strategy," in *Proceedings of MASM 2012 (8th International Conference on Modeling and Analysis of Semiconductor Manufacturing), included in the 2012 Winter Simulation Conference*, 2012, pp. 1–9. (Cité en page 73.)

[87] M. Rowshannahad and S. Dauzère-Pérès, "Qualification management with batch size constraint," in *Proceedings of MASM 2013 (9th International Conference on Modeling and Analysis of Semiconductor Manufacturing), included in the 2013 Winter Simulation Conference*, 2013, pp. 3707–3718. (Cité en page 75.)

[88] E. C. Ozelkan and M. Çakanyildirm, "Test Wafer Management for Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 19, no. 2, pp. 241–251, 2006. (Cité en page 81.)

[89] H. Wohlwend and A. Ikemura, "Non-product wafer (npw) tracking guidelines," in *International SEMATECH Manufacturing Initiative*, 2008, pp. 1–16. (Cité en page 81.)

[90] G. L. Rodriguez-Verjan, S. Dauzère-Pérès, and J. Pinaton, "Impact of Control Plan Design on Tool Risk Management: A Simulation Study in Semiconductor Manufacturing," in *Proceedings of MASM 2011 (7th International Conference on Modeling and Analysis of Semiconductor Manufacturing), included in the 2011 Winter Simulation Conference*, 2011, pp. 1918–1925. (Cité en page 103.)

[91] SEMATECH, "Sematech dictionary of semiconductor terms." [Online]. Available: http://www.sematech.org/publications/dictionary/b.htm (Cité en page 117.)

[92] J. Ruzyllo, "Semiconductor glossary." [Online]. Available: http://www.semi1source.com/glossary/default.asp?whichpage=default (Cité en page 117.)

École Nationale Supérieure des Mines

de Saint-Étienne

NNT : 2014 EMSE 0747

Gloria Luz RODRIGUEZ VERJAN

SMART SAMPLING FOR RISK REDUCTION IN SEMICONDUCTOR MANUFACTURING

Speciality : Industrial Engineering

Keywords : Excursion monitoring, Wafers at Risk, Defect Inspection, Sampling.

Abstract :

In semiconductor manufacturing, several types of controls are required to ensure the quality of final products. In this thesis, we focus on defect inspections, which aim at monitoring the process for defect reduction and yield improvement. We are interested in managing and reducing the risk on process tools (i.e. number of wafers at risk) during fabrication. To reduce this risk, inspection operations are performed on products. However, because inspection operations directly impact the cycle times of products, sampling strategies are used to reduce the number of inspected lots while satisfying quality objectives. Several sampling techniques exist and can be classified according to their capability to deal with factory dynamics. Dynamic sampling strategies have recently been proposed, in which lots to inspect are selected in real time while considering the current production risk. These strategies are much more efficient than previous strategies but more complex to design and implement. In this thesis, a novel approach to select the lots to inspect is proposed. Multiple algorithms have been proposed and validated to efficiently manage the defect inspection queues by skipping (i.e. releasing) lots that do no longer bring enough information. In order to support strategic and tactical decisions, an optimization model for defect inspection capacity planning is also proposed. This model calculates the required defect inspection capacity to ensure the risk limits on process tools when the production conditions change. Industrial results show significant improvements in terms of risk reduction without increasing defect inspection capacity.

École Nationale Supérieure des Mines

de Saint-Étienne

NNT : 2014 EMSE 0747

Gloria Luz RODRIGUEZ VERJAN

ÉCHANTILLONNAGE DYNAMIQUE DE LOTS POUR LA RÉDUCTION DES RISQUES EN FABRICATION DE SEMI-CONDUCTEURS.

Spécialité : Génie Industriel

Mots clés : Sélection dynamique, échantillonnage dynamique, contrôle, défectivité, gestion du risque, fabrication de semi-conducteurs.

Résumé :

Dans les processus de fabrication de semi-conducteurs, différents types des contrôles existent pour maîtriser les procédés et garantir la qualité du produit final. Ces travaux de thèse s'intéressent aux contrôles de défectivité qui visent à maîtriser le risque sur les équipements de production. L'indicateur utilisé est le nombre de produits traités par un équipement depuis la date du dernier produit contrôlé. On s'intéresse à la maîtrise et la réduction du risque sur les équipements de production. Pour cela, différentes stratégies de sélection des lots existent et peuvent être classifiées selon leur capacité à intégrer la dynamique d'une unité de fabrication. Dans les stratégies de sélection dynamique, les lots sont contrôlés en temps réel et en optimisant un critère. Ces stratégies sont récentes et sont beaucoup plus efficaces que les stratégies précédentes, mais aussi plus complexe à mettre en œuvre. Dans ce cadre, nous avons proposé et validé industriellement différents algorithmes pour identifier les lots à relâcher (à ne pas contrôler) dans les files d'attente des lots en défectivité. Nous avons aussi développé et implémenté un modèle d'optimisation de la capacité pour l'atelier de défectivité, qui permet d'évaluer l'impact de paramètres critiques (e.g. plan de production, positions des opérations de contrôles dans la gamme de fabrication, valeurs des limites de risques) dans la gestion du risque global de l'unité de fabrication.