MISE EN ŒUVRE ET OPTIMISATION DES PLANS DE CONTRÔLE DYNAMIQUE DANS LA FABRICATION DES SEMI-CONDUCTEURS

Justin Nduhura Munga

To cite this version:

Justin Nduhura Munga. MISE EN ŒUVRE ET OPTIMISATION DES PLANS DE CONTRÔLE DYNAMIQUE DANS LA FABRICATION DES SEMI-CONDUCTEURS. Autre. Ecole Nationale Supérieure des Mines de Saint-Etienne, 2012. Français. NNT : 2012EMSE0663 . tel-00848578

HAL Id: tel-00848578

https://theses.hal.science/tel-00848578

Submitted on 26 Jul 2013

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

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

NNT : **2012 EMSE 0663**

THÈSE

présentée par

Justin NDUHURA MUNGA

pour obtenir le grade de

Docteur de l'Ecole Nationale Supérieure des Mines de Saint-Étienne

Spécialité : **Génie Industriel**

IMPLEMENTING AND OPTIMIZING DYNAMIC CONTROL PLANS IN SEMICONDUCTOR MANUFACTURING

soutenue à Gardanne, le 09 Octobre 2012

Membres du jury

Président	Mireille JACOMINO	Professeur, Institut National Polytechnique, Grenoble
Rapporteurs	Oliver ROSE	Professeur, Universität der Bunderswehr Műnchen, Allemagne
	Farouk YALAOUI	Professeur, Université de Technologie de Troyes, Troyes
Examinateurs		
Directeur de thèse	Stéphane DAUZÈRE-PÉRÈS	Professeur, EMSE, Saint-Étienne
Directeur industriel	Philippe VIALLETELLE	Ingénieur, STMicroelectronics, Crolles
Encadrants scientifiques	Claude YUGMA	Chargé de Recherche, EMSE, Saint-Étienne
	Samuel BASSETTO	Professeur-adjoint, Ecole Polytechnique de Montréal, Canada
Invités	Jacques PINATON	Ingénieur, STMicroelectronics, Rousset
	Léon VERMARIEN	Ingénieur, STMicroelectronics, Rousset

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	Ph. Collot, Professeur

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

AVRIL	Stéphane	MA	Mécanique & Ingénierie	CIS
BATTON-HUBERT	Mireille	MA	Sciences & Génie de l'Environnement	Fayol
BENABEN	Patrick	PR 1	Sciences & Génie des Matériaux	CMP
BERNACHE-ASSOLLANT	Didier	PR 0	Génie des Procédés	CIS
BIGOT	Jean-Pierre	MR	Génie des Procédés	SPIN
BILAL	Essaïd	DR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR 1	Informatique	Fayol
BORBELY	Andras	MR	Sciences et Génie des Matériaux	SMS
BOUCHER	Xavier	MA	Génie Industriel	Fayol
BRODHAG	Christian	DR	Sciences & Génie de l'Environnement	Fayol
BURLAT	Patrick	PR 2	Génie industriel	Fayol
COLLOT	Philippe	PR 1	Microélectronique	CMP
COURNIL	Michel	PR 0	Génie des Procédés	SPIN
DARRIEULAT	Michel	IGM	Sciences & Génie des Matériaux	SMS
DAUZERE-PERES	Stéphane	PR 1	Génie industriel	CMP
DEBAYLE	Johan	CR	Image, Vision, Signal	CIS
DELAFOSSE	David	PR1	Sciences & Génie des Matériaux	SMS
DESRAYAUD	Christophe	MA	Mécanique & Ingénierie	SMS
DOLGUI	Alexandre	PR 1	Génie Industriel	Fayol
DRAPIER	Sylvain	PR 2	Sciences & Génie des Matériaux	SMS
FEILLET	Dominique	PR 2	Génie Industriel	CMP
FOREST	Bernard	PR 1	Sciences & Génie des Matériaux	CIS
FORMISYN	Pascal	PR 1	Sciences & Génie de l'Environnement	Fayol
FRACZKIEWICZ	Anna	DR	Sciences & Génie des Matériaux	SMS
GARCIA	Daniel	MR	Sciences de la terre	SPIN
GIRARDOT	Jean-Jacques	MR	Informatique	Fayol
GOEURIOT	Dominique	MR	Sciences & Génie des Matériaux	SMS
GRAILLOT	Didier	DR	Sciences & Génie de l'Environnement	Fayol
GROSSEAU	Philippe	MR	Génie des Procédés	SPIN
GRUY	Frédéric	MR	Génie des Procédés	SPIN
GUY	Bernard	MR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HAN	Woo-Suck	CR		SMS
HERRI	Jean-Michel	PR 2	Génie des Procédés	SPIN
INAL	Karim	PR 2	Microélectronique	CMP
KLÖCKER	Helmut	DR	Sciences & Génie des Matériaux	SMS
LAFOREST	Valérie	CR	Sciences & Génie de l'Environnement	Fayol
LERICHE	Rodolphe	CR CNRS	Mécanique et Ingénierie	SMS
LI	Jean-Michel	EC (CCI MP)	Microélectronique	CMP
MALLIARAS	George Grégory	PR 1	Microélectronique	CMP
MOLIMARD	Jérôme	PR2	Mécanique et Ingénierie	SMS
MONTHEILLET	Frank	DR 1 CNRS	Sciences & Génie des Matériaux	SMS
PERIER-CAMBY	Laurent	PR 2	Génie des Procédés	SPIN
PIJOLAT	Christophe	PR 1	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR 1	Génie des Procédés	SPIN
PINOLI	Jean-Charles	PR 0	Image, Vision, Signal	CIS
ROUSTANT	Olivier	MA		Fayol
STOLARZ	Jacques	CR	Sciences & Génie des Matériaux	SMS
SZAFNICKI	Konrad	MR	Sciences & Génie de l'Environnement	Fayol
TRIA	Assia		Microélectronique	CMP
VALDIVIESO	François	MA	Sciences & Génie des Matériaux	SMS
VIRICELLE	Jean-Paul	MR	Génie des procédés	SPIN
WOLSKI	Krzysztof	DR	Sciences & Génie des Matériaux	SMS
XIE	Xiaolan	PR 1	Génie industriel	CIS

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

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

Glossaire :

PR 0	Professeur classe exceptionnelle	Ing.	Ingénieur
PR 1	Professeur 1ère classe	MCF	Maître de conférences
PR 2	Professeur 2ème classe	MR(DR2)	Maître de recherche
PU	Professeur des Universités	CR	Chargé de recherche
MA(MDC)	Maître assistant	EC	Enseignant-chercheur
DR	Directeur de recherche	IGM	Ingénieur général des mines

Centres :

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

Dernière mise à jour le : 02 avril 2012

À mon père NDUHURA SEKIYOBA
et ma mère NTAHONDI KABAMI.

Remerciements

Arrivé à la fin de ce travail de thèse, je me dois de remercier tous ceux qui y ont contribué de manière directe ou indirecte.

Tout d'abord, je tiens à remercier l'**A**ssociation **N**ationale de la **R**echerche **T**echnique (**ANRT**) qui, en collaboration avec l'entreprise **STMicroelectronics** a financé cette thèse. Les travaux présentés dans ce manuscrit de thèse s'inscrivent dans le cadre d'une convention **CIFRE** (**C**onventions **I**ndustrielles de **F**ormation par la **R**echerche) et font aussi partie du projet Européen **IMPROVE** (**I**mplementing **M**anufacturing science solutions to increase equi**P**ment p**RO**ducti**V**ity and fab p**E**rformance).

Je remercie ensuite tous les membres du jury qui ont accepté d'évaluer cette thèse. Merci à Madame Mireille Jacomino pour avoir accepté de présider ce jury de thèse. Merci à M. Oliver ROSE et M. Farouk YALAOUI pour le temps qu'ils ont consacré à la relecture et l'évaluation du présent manuscrit. Leurs rapports, commentaires et remarques ont permis d'enrichir davantage le présent manuscrit de thèse et pousser plus loin les réflexions concernant les perspectives. Merci à M. Jacques Pinaton et M. Léon Vermarien pour leur suivi régulier des travaux de la thèse lors des journées d'échanges et leurs remarques sur le manuscrit de thèse.

Je ne saurai poursuivre sans remercier mes responsables directs avec qui j'ai énormément appris et sans qui les différentes pistes explorées dans ces travaux n'auraient jamais été menées à bout. À mon directeur de thèse M. Stéphane Dauzère-Pérès pour son encadrement, ses conseils, son exigence sans cesse de la qualité, et ses relectures continues des différents documents. Sa compréhension très rapide des différents problèmes et ses propositions de solutions très pertinentes ont été un vrai moteur pour cette thèse. À mon directeur industriel M. Philippe Vialletelle pour son expertise, sa très grande disponibilité malgrès son emploi du temps bien chargé, et ses conseils souvent au-délà du cadre professionnel. Son expérience, ses idées, et sa capacité à présenter des problèmes complexes de manière simple et précise ont permis l'exploration d'un très grand nombre de pistes au cours de cette thèse. À mon encadrant M. Claude Yugma pour ses encouragements, sa disponibilité, ses corrections des différents articles, et nos multiples discussions sur divers sujets. À mon co-encadrant M. Samuel Bassetto pour ses conseils ainsi que ses critiques et remarques sur le manuscrit de thèse.

Durant les trois années de thèse, mon temps a été partagé entre l'entreprise STMicroelectronics et le laboratoire SFL (Sciences de la Fabrication et Logistique)

de l'Ecole des Mines de Saint-Etienne. Je tiens à remercier tous ceux qui m'ont accueilli, conseillé, et offert leur amitié. Je remercie M. Emmanuel Gomez, responsable du groupe **MFA** (**M**ove to **F**ull **A**utomation) chez STMicroelectronics pour m'avoir accueilli au sein de son équipe. Un grand merci à Soidri Bastoini pour son encadrement, sa très grande disponiblité, la formation et son expérience dont j'ai pu bénéficier. Merci à Aymen Mili et Jean-Etienne Kiba pour leur accueil, l'encadrement, les conseils, les différents échanges, et leur grande disponibilité. Merci à Guillaume Lepelletier et Guillaume Palichleb pour leur contribution significative dans tous mes développements et prototypes. Je pense notamment la DMToolbox, outil sans lequel la validation des solutions proposées dans cette thèse aurait été plus laborieuse. Merci à Manuel Cali ainsi qu'à tous les membres du département *Industrial Engineering* pour leur disponibilité et les solutions à tous mes problèmes. Merci du fond du cœur tous les membres du départment SFL. Pour tous les moments passés ensemble et l'amitié qui m'a été offerte. Merci beaucoup pour les cadeaux, tous les messages et les encouragements. Un spécial merci à Mme Agnès Roussy pour son encadrement, sa gentillesse, et son aide durant ma conférence aux USA.

Tout ceci n'aurait pas été possible sans ma famille, mes proches, et mes amis qui m'ont accompagné dans la vie de tous les jours. Je ne pourrai citer des noms de peur d'en oublier. À vous tous qui avaient été présents dans les meilleurs moments ainsi que dans moments difficiles, puissiez-vous trouver ici mes sincères remerciements. À la famille SEBINWA. À Jean-François, Elly, Uzam, Fabrice, Nsenga, Aurlie, Uwimana, Régis. À tous mes frères et sœurs, Safi, Aline, Mihigo, Baraka, Amina, Providence, Benjamin Nduhura. À mon père Constantin Nduhura et ma mère Ernerstine Ntahondi, à vous tous je dédie ce travail.

Justin NDUHURA MUNGA
Gardanne, le 09 Octobre 2012

Science... never solves a problem without creating ten more!

Georges Bernard

Contents

1 Résumé (Summary in French) 1

1.1 Introduction 1
1.2 Contexte Industriel et Problématique de la thèse 3
1.2.1 Environnement de production 4
1.2.2 Étapes de la fabrication 4
1.2.3 Contrôles durant la fabrication et problématique de la thèse . 6
1.3 Généralisation de la Problématique 9
1.3.1 Échantillonnage statique 10
1.3.2 Échantillonnage adaptatif 11
1.3.3 Échantillonnage dynamique 11
1.4 Solutions Générales : IPC - GSI - MILP 13
1.4.1 Analyse et évaluation d'un échantillonnage statique 13
1.4.2 Gestion dynamique des contrôles : indicateur IPC 15
1.4.3 Échantillonnage dynamique : GSI 19
1.4.4 Optimisation d'échantillonnage dynamique : MILP 29
1.5 Solutions Spécifiques : Prototypes et Industrialisation 31
1.6 Conclusion et Perspectives 33

General Introduction 35

2 Industrial Context 41

2.1 Introduction 42

2.2 Semiconductor Manufacturing 42
2.2.1 Production environment 44
2.2.2 Manufacturing stages 45
2.2.3 Integrated circuit 47
2.3 Controls in Semiconductor Manufacturing 50
2.3.1 Levels of controls 50
2.3.2 Types of controls 51
2.4 Conclusion 57

3 Problem Identification and Research Issues 59
3.1 Introduction 60
3.2 Defectivity Controls 60
3.2.1 Defectivity activity and defects types 61
3.2.2 Techniques and tools 64
3.2.3 Defectivity control plans and complexity 65
3.3 Research Issues and Solving Approaches 74
3.3.1 Research issues 74
3.3.2 Solving approaches 75
3.4 Conclusion 76

4 Literature Review on Sampling Techniques 77
4.1 Introduction 78
4.2 Sampling Techniques 79
4.3 Static or Start Sampling 81
4.4 Adaptive Sampling 85
4.5 Dynamic Sampling 90
4.6 Conclusion 94

5 Analyzing and Optimizing Control Plans **95**

5.1 Introduction . . . 96

5.2 Factory Dynamics and Variability . . . 96

5.3 Permanent Index per Context (IPC) . . . 100

5.4 Real-Time Risk Assessment: CMP-WAR . . . 103

5.5 Excursion Management . . . 113

5.6 Conclusion . . . 119

6 Implementing Smart Sampling Policies **121**

6.1 Introduction . . . 122

6.2 Smart sampling mechanism . . . 122

6.2.1 Sampling mechanism . . . 123

6.2.2 Skipping mechanism . . . 123

6.2.3 Scheduling mechanism . . . 125

6.3 Global Sampling Indicator (GSI) . . . 126

6.3.1 GSI 1 . . . 130

6.3.2 GSI 2 . . . 137

6.4 GSI sampling algorithms . . . 141

6.4.1 Threshold definitions . . . 141

6.4.2 GSI sampling algorithm 1 (GSI-SA-1) . . . 147

6.4.3 GSI sampling algorithm 2 (GSI-SA-2) . . . 148

6.5 Numerical experiments . . . 150

6.5.1 S5 simulator . . . 150

6.5.2 Evaluating GSI sampling algorithms . . . 151

6.5.3 Analyzing the impact of GSI parameters . . . 154

6.6 Conclusion . . . 172

7 Optimizing Smart Sampling Policies **173**

7.1 Introduction . . . 174

7.2 Warning Limit and Inhibit Limit 174
7.3 Impact of the Warning Limit and Inhibit Limit 177
7.4 Mixed-Integer Linear Programming model 1 182
7.5 Mixed-Integer Linear Programming model 2 185
7.6 Mixed-Integer Linear Programming model 3 187
7.7 Numerical experiments 192
7.7.1 Evaluating MILP model 1 192
7.7.2 Evaluating the MILP model 2 195
7.8 Conclusion 197

8 Industrial Developments and Implementations 199

8.1 Introduction 200
8.2 CMP-WAR Prototype 200
8.3 Excursion Management Prototype 203
8.4 Financial Metrics 204
8.5 Conclusion 207

General Conclusion and Perspectives 209

A 213

A.1 Glossary 213

B 217

B.1 Clean room - ISO standard Classification 217

C 219

C.1 S5 prototype 219
C.1.1 Input data 222
C.1.2 Output results - Statistics 223
C.2 Impact of parameter β 225

Bibliography **243**

List of Tables

1.1 Situation initiale. 22

1.2 Exemple si deux ensembles des lots S1 ou S2 sont sélectionnés pour une inspection. 22

4.1 Survey on static or start sampling 82

4.2 Mathematical techniques or approaches for static or start sampling . 83

4.3 Survey on adaptive sampling 85

4.4 Mathematical techniques or approaches for adaptive sampling 88

4.5 Survey on dynamic sampling 90

4.6 Mathematical techniques or approaches for dynamic sampling 93

5.1 IPC computation and mechanism 106

5.2 RI computations 107

6.1 Initial situation. 126

6.2 Example if sets of lots S1 or S2 are selected for inspection. 126

6.3 Example1 - Evaluating two different set of lots **S1** and **S2** with the GSI. 132

6.4 Example2 - Evaluating two different sets of lots **S3** and **S4** using the GSI. 133

6.5 Example3 - Evaluating two different sets of lots **S5** and **S6** using the GSI. 134

6.6 Example4 - Evaluating two different sets of lots **S7** and **S8** using the GSI. 135
6.7 Example5 - Evaluating two different sets of lots **S7** and **S8** using the GSI. 138
6.8 Evaluating the GSI sampling algorithms. 153
6.9 Impact of β when $\alpha = 6$. 159
6.10 Impact of $T_{Max} \in$ [0,20%]. 162
6.11 Impact of $T_{Max} \in$ [0,0.5%]. 163
6.12 Impact of $T_{Max} \in$ [0.5,1%]. 164
6.13 Impact of $T_{Max} \in$ [0,0.1%]. 165
6.14 Impact of $T_{Min} \in$ [0,0.5%]. 166
6.15 Impact of $T_{Min} \in$ [0.5,1%]. 167
6.16 Impact of $T_{Metro} \in$ [0,0.5%]. 168
6.17 Impact of $T_{Metro} \in$ [0.5,1%]. 168

7.1 Impact of the WL and IL values on the sampling plan policy. 178
7.2 Impact of the WL value on the sampling plan policy. 179
7.3 Impact of the IL value on the sampling plan policy. 180
7.4 Evaluating the WL and IL obtained with the MILP model 1 (Exposure=1 for all production tools). 193
7.5 Evaluating the WL and IL obtained with the MILP model 1 for different values of the exposure. 194
7.6 Evaluating the WL and IL obtained with MILP model 2 (delay defined per workshop and per tool). 196

B.1 Clean room - ISO Standard Classification [97]. 217

C.1 Impact of β when $\alpha = 1$. 226
C.2 Impact of β when $\alpha = 2$. 226
C.3 Impact of β when $\alpha = 4$. 227
C.4 Impact of β when $\alpha = 6$. 227

C.5 Impact of β when $\alpha = 8$. 228
C.6 Impact of β when $\alpha = 10$. 228
C.7 Impact of β when $\alpha = 12$. 229

List of Figures

1.1 Circuits électroniques dans vie de tous les jours. 3
1.2 Salle blanche. 4
1.3 Fabrication des puces électroniques. 5
1.4 Problème de l'échantillonnage statique. 14
1.5 Mécanisme IPC. 17
1.6 Mécanisme de *sampling*. 20
1.7 Mécanisme de *skipping*. 21
1.8 Vue générale du prototype CMP-WAR. 31
1.9 Vue générale du prototype de gestion des excursions. 32
1.10 General problem solving model (TRIZ approach) [90]. 37
1.11 Thesis reading plan. 38

2.1 Integrated Circuits or chips in everyday's life. 43
2.2 Clean room. 44
2.3 Wafer size evolution. 45
2.4 Front-End Processing [101]. 46
2.5 Back-End Processing [101]. 47
2.6 Transistor size - scale factors. 48
2.7 Impact of Moore law (Cost of 1MB of memory on silicon). 49
2.8 Interaction of APC elements [85]. 52
2.9 Statistical Process Control [75]. 53

2.10 Fault Detection and Classification [75]. 53
2.11 Run-to-Run. 54
2.12 Virtual Metrology [39]. 55
2.13 Examples of defects on wafers. 55

3.1 Particles and arcing on wafers. 62
3.2 Voids on wafers. 62
3.3 Scratches on wafers. 62
3.4 Extra and Missing patterns on wafers. 63
3.5 Corrosions and plate-block on wafers. 63
3.6 Residues on wafers. 63
3.7 Other types of defects. 63
3.8 Dark-Field and Bright-Field systems [31]. 64
3.9 Example of a control plan for the CMOS065 technology. 67
3.10 Example of depth of control for the CMOS065 technology. 68
3.11 Example of qualified tools per process operation. 71
3.12 Killer defects. 72
3.13 Non killer defects. 72

5.1 Drawbacks of static sampling. 97
5.2 Impact of delay between process and measurement steps. 99
5.3 IPC mechanism. 101
5.4 Overview of the CMP-WAR prototype. 104
5.5 Risk Indicator and best lot for control. 105
5.6 Depth of control. 108
5.7 Production tool history. 110
5.8 Lots waiting in front of the CMP area. 110
5.9 Lots waiting in front of the tool. 111
5.10 Example of an excursion management problem. 113

5.11 Example of excursion analysis. 115
5.12 Overview of the Excursion Management prototype. 117
5.13 Concept of dominating sets. 119

6.1 Sampling mechanism. 123
6.2 Skipping mechanism. 124
6.3 GSI combinations. 127
6.4 GSI-1 Evolution. 131
6.5 GSI-2 Evolution. 138
6.6 Inspection queue not empty. 143
6.7 Inspection queue empty. 143
6.8 Threshold metrology (T_{Metro}) - Skipping lots after inspection. 145
6.9 Impact of α on the number of measured lots. 156
6.10 Impact of α on the Medium WAR and Maximum WAR. 157
6.11 Impact of α on the number of sampled lots and skipped lots. 157

7.1 Evolution of the W@R on a production tool. 175

8.1 General schema – Databases. 201
8.2 Overview of the CMP WAR prototype. 201
8.3 Results on global Risk Indicator (RI) reduction. 202
8.4 Impact of the prototype on the overall risk. 203
8.5 Overview of the Excursion Management prototype. 204
8.6 Concept of dominating sets. 212

C.1 S5 interfaces (Modules DATA INPUT and SIMULATIONS). 221
C.2 S5 interfaces (Modules GRAPHS and PARAMETERS). 221
C.3 S5 (Module RESET). 222
C.4 Process input data. 223
C.5 Measurement input data. 223
C.6 Defectivity models. 223

Chapter 1

Résumé (Summary in French)

1.1 Introduction

Le développement très rapide de nouvelles technologies ces dernières années, les exigences de plus en plus fortes des marchés internationaux, et la recherche de rentabilité de plus en plus élevée par les entreprises multinationales ont donné place à une très forte concurrence au niveau mondial. Dans le monde de la microélectronique où les principaux produits fabriqués (circuits et puces électroniques) sont des éléments majeurs de la vie quotidienne[1], proposer les meilleurs produits à des prix compétitifs est vital pour les entreprises.

Plusieurs pistes sont explorées par les entreprises dans le but de réduire les coûts de production sans impacter la qualité finale du produit. Parmi les différentes pistes, une des principales concerne les contrôles durant la fabrication des circuits électroniques. En effet, la taille des circuits ou puces électroniques à fabriquer devient tellement petite (de l'ordre du nanomètre)[2] que plusieurs contrôles sont nécessaires pour s'assurer que les procédés de fabrication sont correctement réalisés et que le produit satisfait aux spécifications du client. Cependant, parmi tous les contrôles réalisés, certains sont principalement destinés à anticiper toute dérive po-

[1] On estime qu'une personne utilise environ 250 circuits électroniques par jour [40].

[2] La taille des composants électroniques (transistors, résistances, condensateurs, etc.) nécessaires à la fabrication des circuits électroniques est environ 5000 fois plus petite que le diamètre d'un cheveu.

tentielle et limiter les pertes potentielles en cas problème durant la production. Ils sont donc jugés *non-obligatoires* car ils n'ajoutent rien à la fonctionnalité finale du produit à livrer au client. D'où le challenge pour les entreprises d'arriver à mieux maîtriser, répartir, et limiter ce nombre de contrôles *non-obligatoires* sans augmenter le risque (c'est-à-dire la perte potentielle) sur la production.

Différentes techniques d'échantillonnage existent et ont été développées par les entreprises et dans la littérature dans le but de trouver le meilleur compromis entre le nombre de contrôle et le risque toléré au sein de la production. Entre les techniques statiques et dynamiques, les techniques d'échantillonnage dynamiques sont jugées plus robustes de part leur capacité à intégrer la dynamique de la production et la variabilité. Le problème qui se pose concerne l'industrialisation de ces techniques d'échantillonnage dynamiques. L'investissement requis (ressources informatiques, formation des opérateurs et ingénieurs, système de production, etc.) et la complexité sont tels que la plupart des entreprises préfèrent rester sur des techniques d'échantillonnage statiques alors que l'inefficacité de ces dernières à anticiper les dérives potentielles pour les entreprises multi-produits a déjà été démontrée [78] [12].

Dans le cadre ma thèse, je m'intéresse à l'évaluation de l'efficacité des différentes techniques d'échantillonnage, l'identification des points de sur- et sous-contrôles, et la mise en œuvre concrète des plans de contrôles ou d'échantillonnage dynamiques au sein de l'entreprise STMicroelectronics.

Ce premier chapitre est un résumé global en Français de ma thèse qui est rédigée en Anglais. Je commence par présenter rapidement le contexte industriel et la problé-matique de ma thèse. Ensuite, je synthétise la révue de la littérature qui généralise mon problème et le positionne parmi les différentes techniques développées au cours des 20 dernières années. Après cette synthèse de la révue de la littérature, je présente les solutions générales que je développe dans la cadre de ma thèse. Je termine ensuite ce résumé en Français en donnant un aperçu des solutions spécifiques que j'ai développées pour valider les solutions générales au sein du site de 300mm de la société STMicroélectronics basée à Crolles, en France. **L'originalité des travaux de ma thèse repose sur le fait que toutes les solutions proposées et**

présentées dans ce manuscrit ont été validées industriellement et certaines d'entr'elles ont été industrialisées sur le site de Crolles et les autres sites de STMicroelectronics (Rousset, Italie-Catania).

1.2 Contexte Industriel et Problématique de la thèse

La principale activité d'une entreprise de semiconducteurs est de fabriquer des composants électroniques, les interconnecter, et obtenir ainsi des puces ou circuits électroniques qui sont utilisés dans plusieurs domaines de la vie de tous les jours (téléphone, voiture, régulateur de température, ordinateurs, etc.). La Figure 1.1 donne un aperçu de l'utilisation des circuits électroniques dans la vie de tous les jours. Dans quasiment toute activité, chaque produit, nous utilisons des puces ou circuits électroniques.

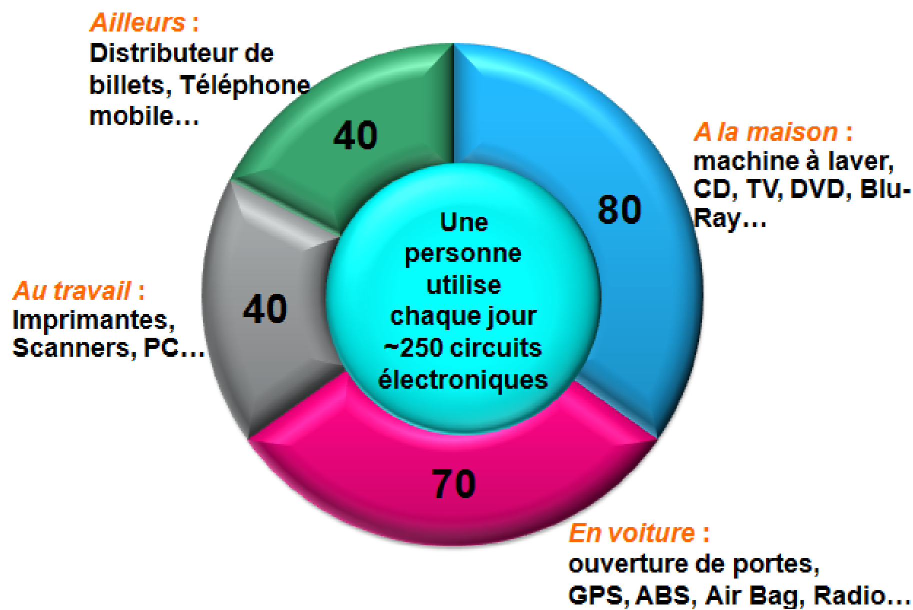


Figure 1.1: Circuits électroniques dans vie de tous les jours.

1.2.1 Environnement de production

Les dimensions des composants électroniques fabriqués sont tellement petites qu'un environnement spécial appelé "**salle blanche**" est nécessaire pour éviter toute contamination et assurer un bon fonctionnement des circuits électroniques. La Figure 1.2 donne un aperçu de l'environnement de production. Dans cet environnement, les opérateurs et ingénieurs sont couverts de la tête aux pieds et l'air ambiant est renouvellé toutes les 30 secondes. En comparaison avec une salle d'opération chirurgicale, la plus "crasseuse" des **salles blanches** est au moins 3 fois plus propre qu'une salle de chirurgie [74].

Figure 1.2: Salle blanche.

1.2.2 Étapes de la fabrication

Le processus de fabrication des puces ou circuits électroniques[3] se résume en deux principales étapes [101] : **Front-End** et **Back-End**. Le Front-End regroupe les étapes de fabrication des éléments de base de la puce tandis que le Back-End celui des interconnexions des éléments de base et la mise en boitier (Figure 1.3).

[3] Une puce électronique se compose de plusieurs composants électroniques qui sont fabriqués sur des plaques de silicium. Ces composants électroniques sont interconnectées entr'elles de diverses manières pour réaliser diverses fonctions.

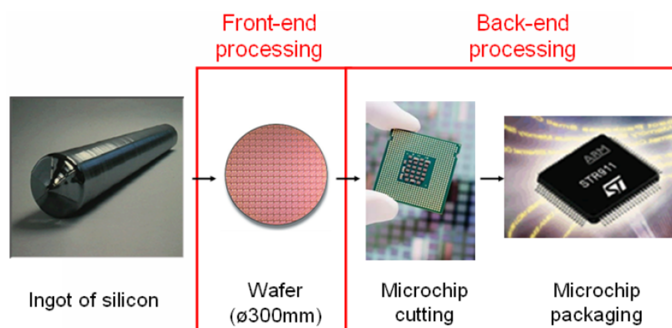


Figure 1.3: Fabrication des puces électroniques.

Cette thèse est réalisée au sein du site 300mm de STMicroelectronics où seulement les opérations de Front-End sont réalisées. Nous ne nous intéresserons donc qu'à la partie Front-End durant laquelle plusieurs opérations sont réalisées sur des plaquettes circulaires de silicium appelées *wafers*. Parmi les principales opérations du Front-End, nous pouvons citer [40] [101] :

1. **Oxidation** : on oxide la plaquette de silicium sur toute sa surface. Plusieurs fours spécifiques sont utilisés.

2. **Dépôt de résine** : on dépose de façon uniforme une couche de résine photo sensible sur la couche d'oxide. Cette couche de résine se transforme sous l'action de la lumière.

3. **Photolithographie** : comme son nom l'indique, on utilise le principe de la photo. On se sert des jeux de masque pour créer des motifs sur la plaquette de silicium. On aligne le masque sur la plaquette et le tout est exposé à une source de lumière. La résine "s'imprègne" comme une pellicule photographique normale dans les zones laissées libres par le masque.

4. **Développement** : comme pour le développement photographique, on enlève la résine qui a été exposée à la lumière (dans les zones laissées libres par le masque).

5. **Gravure** : on enlève l'oxyde laissé libre par la résine, sans attaquer le silicium de départ.

6. **Dopage** : on introduit des éléments chimiques pour modifier les caractéristiques du silicium et le rendre capable de conduire le courant électrique.

Toutes ces opérations réalisées durant le Front-End se repètent plusieurs fois. Avant, pendant, et après chaque opération de fabrication, une ou plusieurs opérations de contrôle sont nécessaires pour vérifier que le procédé a bien été réalisé et les spécifications clients respectées.

1.2.3 Contrôles durant la fabrication et problématique de la thèse

1.2.3.1 Contrôles durant la production

Plusieurs niveaux de contrôles existent durant la fabrication des composants électroniques. Pour chaque niveau de contrôle, plusieurs types de contrôle existent. Parmi les principaux niveaux de contrôle, nous pouvons citer :

1. **Installations techniques** : on contrôle les paramètres liés à l'environnement de production (température ambiante, contamination, luminosité, liquides, gaz, etc.) pour garantir des conditions les plus optimales possibles au sein de la production.

2. **Capteurs sur les équipements de contrôle** : on contrôle l'état des capteurs placés sur les différents équipements de production pour s'assurer d'une remontée correcte des données au moment de l'analyse des différents paramètres.

3. **Mesures ou contrôles en ligne** : on contrôle, tout au long de la production, les *wafers* sur lesquelles les composants électroniques sont réalisés. Plusieurs techniques sont utilisées (ellipsométrie, réflectivité, scatterométrie, etc.) donnant lieu à plusieurs types de contrôles. La plupart des techniques sont regroupées sous le nom **"APC"** (***A****dvanced* ***P****rocess* ***C****ontrol* - Contrôle avancé des procédés).

4. **Tests paramétriques** : on analyse les paramètres électriques des composants électroniques (courant de fuite, tension de claquage, tension de seuil des transistors, etc.).

5. **Tests finaux ou fonctionnels** : ces tests interviennent à la fin du Front-End. On vérifie que les circuits réalisés fonctionnent correctement avant leur mise en boitier (Back-End) et envoie aux différents clients.

6. **Caractérisations physiques** : on évalue la durée de vie des différents composants électroniques sous l'influence des perturbations extérieures (température, humidité, corrosion, etc.).

Parmi ces 6 niveaux de contrôle, ma thèse s'intéresse au troisième niveau de contrôle et principalement aux **contrôles défectivité** où un des principaux objectifs est **la détection des défauts générés par les équipements de production sur les *wafers***.

1.2.3.2 Problématique de la thèse

Ma thèse aborde de manière générale les différents contrôles qui ont lieu sur les *wafers* et les lots[4] durant la fabrication des composants électroniques. Je me suis focalisé sur les contrôles défectivité à cause de la forte complexité qui est principalement due à la **profondeur de contrôle** et le **champ d'investigation**. La **profondeur de contrôle** donne le nombre d'opérations de fabrication qui sont couvertes[5] par une opération de contrôle, et le **champ d'investigation** donne le nombre d'équipements à considérer lors de l'analyse des défauts détéctés sur les *wafers*. Tous les équipements de production sont concernés car ils ont tous des parties mécaniques qui génèrent à la fois des particules et des défauts sur les *wafers*.

L'autre niveau de complexité vient du nombre de produits (plus de 200 produits différents sont fabriqués en parallèle), des technologies (plus de 20 technologies différentes), ou encore des priorités à considérer (la criticité des produits, les exigences clients, les délais de livraison, les coûts de production, l'environnement, le management, etc.) lors de la sélection des *wafers* ou des lots à contrôler. Cela constitue un véritable challenge pour l'échantillonnage. En effet, intégrer tous les paramètres est tout simplement impossible et l'enjeu majeur réside donc dans la

[4] Un lot est un groupe d'au plus 25 *wafers*.

[5] Une opération de fabrication est dite *couverte* par une opération de contrôle lorsque cette dernière permet d'avoir l'information sur l'opération de fabrication.

capacité à utiliser de la manière la plus optimale possible la capacité de contrôle disponible.

L'objectif de ma thèse est d'arriver à comprendre le mécanisme d'échantillonnage statique[6] en place au sein de STMicroelectronics, analyser son efficacité, détecter les différents points de sur- et sous-contrôle, et arriver à proposer des solutions pouvant supporter la mise en place et le déploiement "**concret**" (industriel) des plans de contrôle et d'échantillonnage dynamique. Un des principaux challenges étant d'arriver à manipuler en temps réel un très grand volume de données, et proposer des algorithmes compréhensibles, maintenables, généralisables, et par-dessus tout **industrialisables**.

[6]Au début de ma thèse, l'échantillonnage était à 100% statique, c'est-à-dire qu'un certain nombre de lots/*wafers* était désigné au lancement de la production pour subir des contrôles réguliers tout au long du cycle de fabrication. Seuls ces lots "pré-désignés" pouvaient être contrôlés et ce, à toutes les étapes possibles de contrôle.

1.3 Généralisation de la Problématique : Révue de la Littérature

L'échantillonnage n'est pas un concept récent dans le monde du semiconducteur pour deux raisons principales [19] [23] [33] :

1. Un contrôle à 100% est impossible à cause du coût que cela engendre sur le produit final [76].

2. Un contrôle à 100% ne pourra jamais garantir une qualité 100% dans le semiconducteur à la cause de la taille des particules, des défauts, et de la fiabilité des différents procédés de contrôle [23].

Il est donc indispensable de limiter le nombre de contrôle durant tout processus de fabrication tout en s'assurant de faire les bons contrôles au bon moment. Plusieurs types ou méthodes d'échantillonnage existent en fonction des objectifs recherchés. On en distingue trois principaux [48] :

1. **Contrôle du matériel à risque et gestion des excursions**[7] [83] [9] [61] : le but est de sélectionner des lots à contrôler suivant une fréquence bien définie pour d'un côté limiter la perte potentielle en cas de problème, et de l'autre côté arriver à détecter le plus rapidement possible les différents défauts générés au cours de la production.

2. **Intégration de nouveaux procédés et amélioration du rendement** [46] : le but est d'ajuster le pourcentage des lots à contrôler pour mieux identifier les principaux détracteurs pour les différentes technologies et les éliminer au fur et à mesure. Dans les usines de petite taille, on cherche à ajuster le nombre de lots lancés au début de la production pour compenser les pertes éventuelles.

3. **Statistiques et apprentissage** [26] : le but est d'apprendre sur les différents types des défauts détectés ainsi que leur mécanismes.

[7] Une excursion se produit lorsqu'un problème est détecté sur un *wafer*, un lot, ou un équipement après une opération de contrôle.

Dans ma thèse, je m'intéresse au premier groupe d'échantillonnage qui vise à contrôler le matériel à risque durant la production et détecter le plus rapidement possible les différentes excursions. L'objectif est double : limiter le nombre de contrôle (par échantillonnage) sans augmenter le matériel à risque en cas de problème, et détecter très rapidement les différents problèmes. Il y a un compromis nécessaire à trouver car, si d'un côté on se focalise sur la réduction du nombre de contrôle (ce qui permet de réduire le coût final du produit), on prend le risque de ne pas détecter rapidement les différents défauts et donc d'avoir des pertes significatives en cas de problème. D'un autre côté, si on se focalise uniquement sur la détection rapide des défauts, on risque d'augmenter le nombre de contrôle et donc augmenter le coût du produit final.

Plusieurs politiques d'échantillonnage ont été développées dans la littérature. Nous les classifions en trois principaux groupes : statiques, adaptatives, et dynamiques.

1.3.1 Échantillonnage statique

Un échantillonnage statique consiste à définir un nombre fixe et limité des lots à contrôler tout au long de la production. Le nombre de lots à contrôler est fixé par la capacité disponible de contrôle. Les lots à contrôler sont pré-sélectionnés au début de la production et subissent systematiquement une opération de contrôle devant chaque étape de contrôle [45] [70].

Cette politique d'échantillonnage a été largement utilisée par les entreprises dans les années 1990 car en contrôlant toujours les mêmes lots, il est possible de quantifier les défauts apportés par chaque opération de fabrication et donc identifier rapidement la source des défauts [31]. De nos jours, cette politique d'échantillonnage statique est fortement critiquée à cause de son incapacité à ajuster les paramètres d'échantillonnage en fonction de la dynamique de la production [12]. Ceci est d'autant plus vrai dans les entreprises multi-produits où plusieurs produits sont fabriqués en parallèle et où la production n'est jamais linéaire, c'est-à-dire que le premier produit qui entre dans la chaîne da fabrication n'est pas toujours le premier

à en sortir [57].

Pour prendre en compte la dynamique de la production, de nouvelles politques d'échantillonnage appelées "adaptatives" ont été dévéloppées.

1.3.2 Échantillonnage adaptatif

Un échantillonnage adaptatif est basé sur un échantillonnage statique mais la différence majeure avec cette dernière est que le nombre de lots à contrôler est ajusté en fonction de l'état de la production [98]. Lorsque la production est considérée comme étant "sous-contrôle", le taux d'échantillonnage est réduit, et lorsque il y a suspicion de dérive, le nombre de lots à contrôler est augmenté pour confirmer rapidement la dérive et limiter ainsi les pertes potentielles [99] [71].

Cette technique d'échantillonnage s'avère plus efficace que la technique d'échantillonnage statique mais le problème ici est que l'on ne maîtrise plus la charge de travail (ou ressources nécessaires) des ingénieurs responsables des opérations de contrôle. Le nombre de lots à contrôler n'étant plus constant en fonction de l'état de la production, le risque est d'avoir des périodes avec beaucoup de lots à contrôler, ce qui remettrait en cause l'efficacité des contrôles [58].

Pour faire face à ce problème de gestion de ressources, de nouvelles politiques d'échantillonnage (très récentes) dites "dynamiques ou intelligentes" ont été développées.

1.3.3 Échantillonnage dynamique

Un échantillonnage dynamique consiste à la sélection en temps réel des lots à contrôler. Le nombre total de contrôles est fixé par la capacité de contrôle disponible [79]. Contrairement aux techniques d'échantillonnage précédentes (statiques et adaptatives), aucun lot n'est pré-sélectionné à l'avance. La sélection se fait lorsque le lot arrive devant une opération de contrôle. En fonction de l'information contenue dans le lot, de la capacité disponible de contrôle, ou des priorités au sein de la production, le lot est soit contrôlé, soit dirigé directement à l'opération de fabrication suivante [29] [78] [50]. L'avantage d'une telle technique est qu'en prenant

en compte l'état réel de la production et en choisissant dynamiquement les bons lots à contrôler, il est possible de détecter très rapidement toute dérive potentielle sans augmenter le risque sur la production ou la charge de travail des ingénieurs responsables des opérations de contrôle [20].

Ma thèse se focalise donc sur cette troisième technique et l'enjeu est d'arriver à mettre concrètement en place une telle technique au sein du site 300mm de STMicroelecronics. La technique est très récente et les auteurs qui ont travaillé dessus ne donnent pas toujours assez de détails sur la complexité technique de la mise en œuvre d'une telle technique au sein d'une usine où plus de 200 produits différents sont fabriqués en parallèle. De plus, l'environnement du semiconducteur est tellement particulier qu'une technique peut s'avérer efficace dans une usine A et s'avérer inutilisable pour une usine B.

Pour arriver à mettre en œuvre une telle technique d'échantillonnage et donc des plans de contrôle dynamiques, il a fallu proposer successevement un indicateur permettant de manipuler en temps réel un très grand nombre de données sans consommer trop de ressources informatiques, développer un indicateur pour arriver à choisir dynamiquement les bons lots à contrôler, et optimiser les différents paramètres pour s'assurer de la robustesse de la solution. Je résume dans la section suivante les principales solutions générales que je propose dans ma thèse pour la mise en œuvre des techniques de contrôle dynamiques. Le lecteur pourra trouver plus de détails dans les chapitres 5, 6, et 7 de ce manuscrit de thèse.

1.4 Solutions Générales : IPC - GSI - MILP

Dans ma thèse, je propose trois solutions générales permettant de mettre en œuvre des plans des contrôles dynamiques dans une unité avancée de fabrication des semiconducteurs. Ce trois solutions sont : l'**IPC** (**I**ndice **P**ermanent par **C**ontexte), le **GSI** (**G**lobal **S**ampling **I**ndicator), et un modèle **MILP** (**M**ixed **I**nteger **L**inear **P**rogramming). L'**IPC** est un indicateur qui permet de manipuler un volume important de données avec une très faible consommation de ressources informatiques. Cet indicateur permet de simplifier l'analyse de plusieurs types de risques et donc de supporter l'industrialisation des algorithmes de contrôles ou d'échantillonnage dynamiques qui manipulent un volume important de données. Le **GSI** est un indicateur qui permet de choisir dynamiquement le meilleur lot à contrôler et défnir la priorité de contrôle sur les équipements de contrôle. Le **MILP** est un modèle qui calcule les paramètres clés utilisés par le **GSI** pour une sélection dynamique et optimisée des lots à contrôler.

Avant de pouvoir proposer les solutions générales résumées dans cette section, il était nécessaire d'analyser l'efficacité du plan de contrôle "statique" en place pour en cerner les avantages et inconvénients.

1.4.1 Analyse et évaluation d'un échantillonnage statique

L'analyse du plan de contrôle statique en place chez STMicroelectronics a été faite en partant de l'hypothèse selon laquelle *"un contrôle sans valeur ajoutée est à la fois une perte de temps et une perte d'argent"*. Considérons l'exemple de la Figure 1.4 qui met en évidence un des principaux inconvénients de l'échantillonnage statique.

Il y a 6 lots (L1, L2, L3, L4, L5, et L6) qui arrivent dans l'atelier 1 pour subir diverses opérations de fabrication avant d'aller dans l'atelier 2 pour d'autres opérations de fabrication. Le plan de contrôle statique défini par les ingénieurs au début de la production est de contrôler un lot sur deux. Dans le cas de la Figure 1.4, les lots L2, L4, et L6 ont été identifiés au début de la production pour des contrôles ponctuels devant chaque étape de contrôle. Cela signifie qu'une fois passés dans l'atelier 1, ces

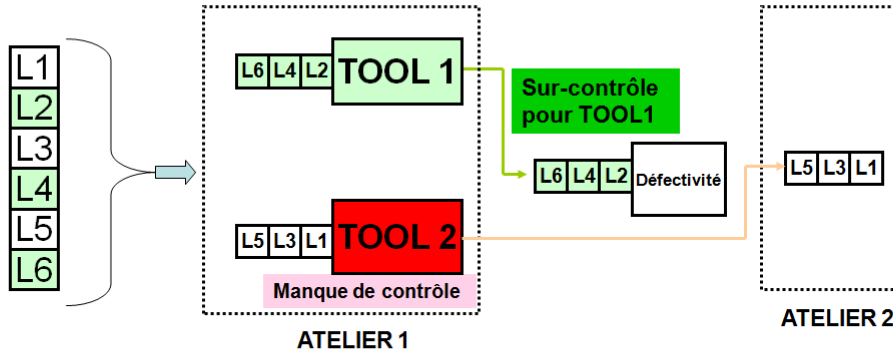


Figure 1.4: Problème de l'échantillonnage statique.

lots (L2, L4, L6) doivent subir une opération de contrôle dans l'atelier de défectivité avant de subir d'autres opérations de fabrication dans l'atelier 2.

Comme introduit dans les sections précédentes, un contrôle en défectivité a pour objectif de détecter les défauts ou particules générées par les équipements de production sur les *wafers*, c'est-à-dire qu'à chaque fois qu'un contrôle en défectivité est réalisé sur un lot ou plusieurs *wafers*, on analyse le nombre de particules ou des défauts présents sur les *wafers*. Si aucune alarme n'est déclenchée (nombre de particules en dessous d'un certain seuil ou défauts non critiques), on relâche l'incertitude sur l'ensemble des équipements de production sur lesquels les *wafers* contrôlés ont subis des opérations de fabrication avant d'arriver au contrôle défectivité. Dans le cas décrit dans la Figure 1.4, un plan de contrôle optimal consisterait à faire passer sur chaque équipement (TOOL1 et TOOL2) au moins un lot échantillonné (c'est-à-dire L2 ou L4 ou L6), ce qui permettrait, en cas de contrôle “bon” en défectivité, de relâcher l'incertitude sur l'ensemble des équipements de l'atelier 1.

Cependant, dans un environnement multi-produits (cas du site de 300mm de STMicroelectronics) où plusieurs produits différents sont fabriqués simultanément sur les mêmes équipements, le cas tel que celui décrit sur la Figure 1.4 s'est avéré être un cas fréquent. Tous les lots échantillonnés (c'est-à-dire pré-sélectionnés pour un contrôle en défectivité) passent sur un seul équipement (TOOL1) alors que sur l'autre équipement (TOOL2) il n'en passe aucun. Cela résulte donc, après un

contrôle défectivité, à un sur-contrôle pour l'équipement TOOL1 et un manque de contrôle pour l'équipement TOOL2. Le phénomène est encore plus accentué avec le nombre d'équipements disponibles en production (plus de 300 équipements chez STMicroelectronics).

Pour éviter ce phénomène, une solution simple serait de définir une limite maximale de lots échantillonnés pour chaque équipement de production. Le problème qui se pose concerne la disponibilité des équipements et leur qualifications respectives car, tous les lots ne peuvent pas passer sur tous les équipements et tous les équipements ne peuvent pas réaliser les mêmes opérations de fabrication à la même vitesse. De plus, en fonction du produit, de l'opération à laquelle se trouve chaque lot, de la technologie, de l'état de la production, les responsables de la production ne peuvent pas se permettre d'arrêter un équipement pour respecter un taux prédéfini d'échantillonnage. D'où l'intérêt même d'un échantillonnage en mode dynamique.

Convaincu alors de la nécessité d'un échantillonnage en mode dynamique, plusieurs questions ont été soulévées : comment arriver à analyser en temps réel un très grand volume de données (300 équipements, 200 produits, 20 technologies, plusieurs contraintes de production, etc.) et arriver à identifier "instantanément", en temps réel le meilleur lot à contrôler ? Quel coût cela engendrerait-il en terme d'investissement et de ressources informatiques ? Serait-ce vraiment rentable pour l'entreprise ?

Pour répondre à cette question, l'indicateur **IPC** a été dévéloppé pour accélérer les calculs, généraliser les solutions et supporter les différentes décisions en mode dynamique. La section suivante décrit brièvement cet indicateur **IPC**, son utilisation, et sa généralisation à plusieurs types de risques au sein de la production.

1.4.2 Gestion dynamique des contrôles : indicateur IPC

L'IPC est un compteur qui est incrémenté à chaque fois qu'un contexte est vérifié. Le contexte peut être un équipement, une chambre[8], une recette[9], une technologie,

[8]Une chambre est une partie à l'intérieur d'un équipement où est réalisé une opération de fabrication.

[9]Une recette est un ensemble de données nécessaires à un équipement pour le traitement physique d'un *wafer* ou d'un lot.

un type de résine, la combinaison d'une opération de fabrication et d'une technologie, etc. Ce compteur n'est jamais remis à zéro sauf lorsqu'un événement particulier se produit (maintenance préventive, qualification d'un équipement, changement de recette, etc.). Le but de l'IPC est d'avoir un indicateur général, standard, et simple qui permette d'évaluer très rapidement différents types de risques en fonction du contexte sans consommer trop de ressources informatiques ni nécessiter des développements informatiques complexes.

Lors de la première implémentation dans le cadre de la validation industrielle de l'IPC, le contexte avait été défini au niveau de l'équipement, c'est-à-dire qu'on s'intéressait à évaluer en temps réel le risque de faire passer un lot sur un équipement de production. Ce risque s'appelle *Wafer-At-Risk* et représente le nombre de *wafers* qui ont subis une opération de fabrication entre deux opérations de contrôle. À chaque lot l et équipement de production m est associé un IPC, qui est égale à 0 si l n'a subi aucune opération de fabrication sur l'équipement m. Définissons M comme le nombre d'équipements de production, et $NW(l)$ comme le nombre de *wafers* contenus dans le lot l. L'objectif est de mettre à jour les paramètres suivants en temps réel :

- $LLM(m)$: indice du dernier lot qui a été contrôlé pour valider l'équipement de production m.
- IPC_l^m : IPC du lot l pour l'équipement de production m.
- RI_m : indicateur du risque sur l'équipement de production m.
- NI_l^m : nombre de *wafers* potentiellement impactés sur l'équipement m si le lot l est contrôlé.
- NI_l : nombre de *wafers* potentiellement impactés dans l'ensemble de la production si le lot l est contrôlé.

Quand le lot l passe sur l'équipement de production m, un IPC est associé à l. Cet IPC du lot l est égal à l'IPC du lot l' passé juste avant l sur m plus le nombre de *wafers* contenus dans l ($NW(l)$) :

$$IPC_l^m = IPC_{l'}^m + NW(l) \tag{1.1}$$

L'indicateur du risque (c'est-à-dire le matériel à risque ou *Wafer-At-Risk*) sur l'équipement de production m est donc donné par :

$$RI_m = IPC_l^m - IPC_{LLM(m)}^m \tag{1.2}$$

L'utilisation de l'IPC simplifie largement les calculs des différents types de risque car tout est réduit à une simple différence entre deux valeurs entières. Cela implique une faible consommation des ressources informatiques, la possibilité d'analyser en temps un très grand nombre de types de risques sans passer par des développements complexes. Au lieu d'aller rechercher à chaque fois l'historique de la production pour analyser le risque en temps réel, on assigne à chaque lot un indice (IPC du lot) lorsque le contexte est vérifié.

La Figure 1.5 montre une séquence des lots ayant subis des opérations de fabrication sur l'équipement de production m.

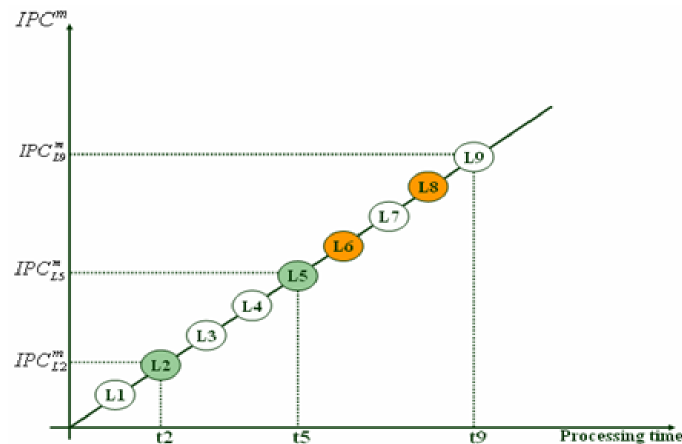


Figure 1.5: Mécanisme IPC.

Les lots $L1$, $L2$, ..., $L9$ sont passés sur m. Parmi ces lots, $L2$ et $L5$ ont été

validés par un contrôle "bon" en défectivité et dans le cas décrit dans la Figure 1.5, $L5$ correspond au dernier lot qui a été contrôlé, c'est-à-dire $L5 = LLM(m)$. Selon les équations (1.1) et (1.2), l'indicateur du risque sur m à $t9$ est donné par :

$$RI_m = IPC^m_{L9} - IPC^m_{L5}$$

où :

$$IPC^m_{L9} > IPC^m_{L5}$$

En utilisant l'IPC, il est aussi possible d'identifier rapidement le meilleur lot l à contrôler devant une étape de contrôle. Ce lot l est choisi tel que son IPC vérifie la propriété suivante :

$$IPC^m_l = Max\{0, \{IPC^m_{ll} \setminus IPC^m_{ll} > IPC^m_{LLM}, ll \in LM\}\} \quad (1.3)$$

où LM est l'ensemble des lots en attente devant une étape de contrôle.

Dans la Figure 1.5, les lots $L6$ et $L8$ sont passés sur m et sont en attente devant une étape de contrôle. Selon (1.3), le meilleur lot à contrôler pour m sera $L8$ car $IPC^m_{L8} > IPC^m_{L6}$ et $IPC^m_{L8} > IPC^m_{L5}$.

Un contrôle est défini comme une opération de mesure plus une action [7]. Cela signifie qu'il est primordial d'être capable d'évaluer en temps réel le nombre de lots potentiellement impactés si un problème est détecté après une opération de mesure sur un lot l. Ce nombre peut être déterminé pour chaque équipement de production m (NI^m_l) et pour l'ensemble de la production (NI_l) :

$$NI^m_l = Max\{0, IPC^m_l - IPC^m_{LLM}\} \quad (1.4)$$

et

$$NI_l = \sum_m NI^m_l \quad (1.5)$$

Dans la Figure 1.5, à $t9$, NI_l^m sera donné par $(IPC_{L9}^m - IPC_{L5}^m)$ correspondant à la somme des *wafers* contenus dans $L6$, $L7$, $L8$, et $L9$.

Ce mécanisme d'IPC a été implémenté dans un prototype et déployé en production pour un atelier avant d'être industrialisé sur l'ensemble de la production. Une fois industrialisé, ce mécanisme IPC a été utilisé pour supporter l'implémentation des algorithmes intelligents d'échantillonnage dynamique dont le principe est décrit dans la section suivante.

1.4.3 Échantillonnage dynamique : GSI

Un échantillonnage dynamique ou intelligent consiste à sélectionner en temps réel et au meilleur moment des lots à contrôler. "Aucune règle" n'est définie au début de la production et les lots à contrôler sont sélectionner à leur arrivée devant une étape de contrôle. Trois types de décisions sont nécessaires pour réaliser ce type d'échantillonnage : le *sampling*, le *skipping*, et le *scheduling*. Ces trois décisions sont liées aux contraintes de la production et à la capacité de contrôle disponible. L'ordre des différentes décisions n'est pas nécessairement séquentielle, c'est-à-dire d'abord le *sampling*, ensuite le *skipping*, et finalement le *scheduling*. Certaines décisions peuvent être prises simultanément. Le principal objectif est de choisir des lots à contrôler pour minimiser le risque dans la production en fonction de la capacité de contrôle disponible.

A. Mécanisme de *sampling*

Le *sampling* consiste à sélectionner un lot pour le contrôle et le placer dans la file d'attente d'inspection. La Figure 1.6 nous donne une illustration du *sampling*. À chaque fois qu'un lot **Lx** arrive devant une étape de contrôle (défectivité), une décision est prise quant à l'ajout de **Lx** dans la liste des lots déjà en attente.

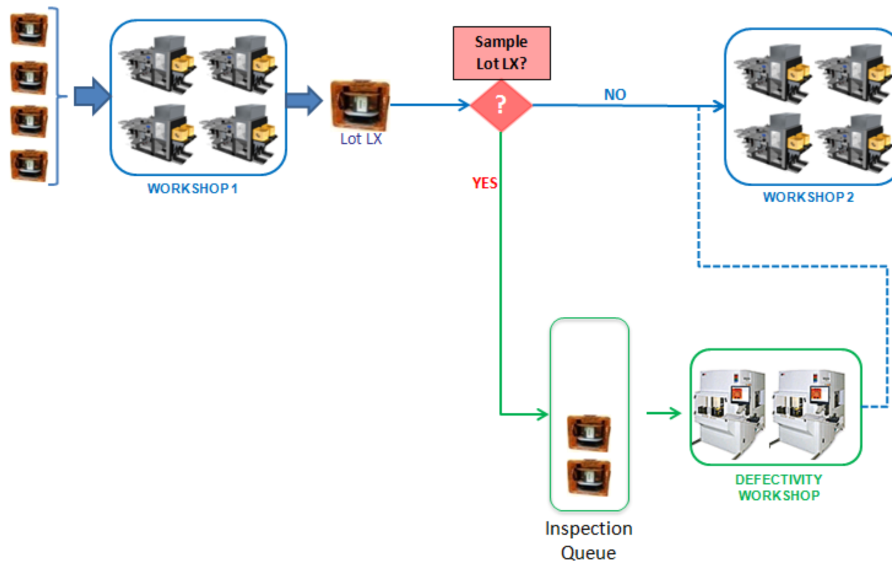


Figure 1.6: Mécanisme de *sampling*.

B. Mécanisme de *skipping*

Le *skipping* consiste à ne pas mesurer ou contrôler un lot **Lx** déjà présent dans la file d'attente d'inspection (Figure 1.7). Le lot est rétiré de la file d'attente et dirigé à la prochaine étape de fabrication. Ce type de décision peut se produire lorsque :

(1) La taille maximale de la file d'attente est atteinte et un lot "important" et prioritaire vient d'arriver devant l'étape de contrôle et doit être absolument contrôlé,

(2) La capacité de contrôle disponible est réduite à cause d'un équipement qui vient de tomber en panne,

(3) Certains lots viennent d'être mesurés et par conséquent, un ou plusieurs lots présents dans la file d'attente d'inspection perdent leur valeur ajoutée.

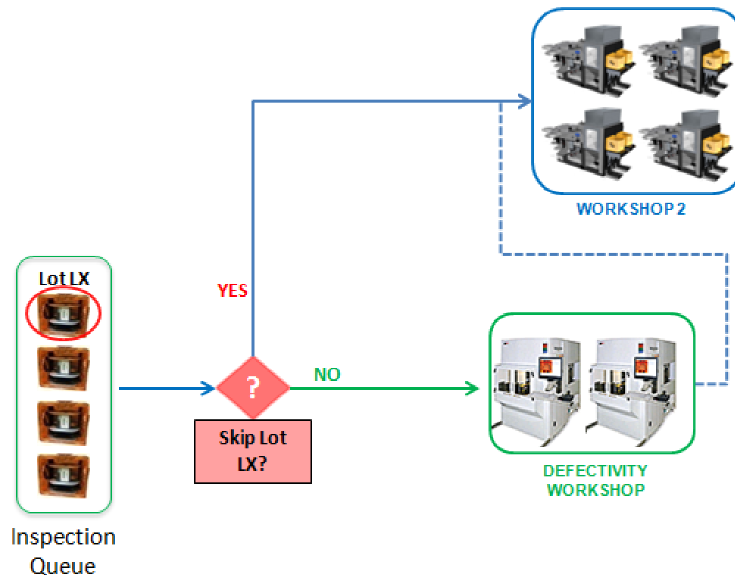


Figure 1.7: Mécanisme de *skipping*.

C. Mécanisme de *scheduling*

Le *scheduling* consiste à ordonnancer les lots présents dans la file d'attente, c'est-à-dire définir la priorité de passage sur les équipements de contrôle. En fonction du gain (en terme de réduction du risque) apporté par chaque lot et des contraintes de la production, certains lots sont plus prioritaires que d'autres.

Ces trois décisions (*sampling*, *skipping*, *scheduling*) sont prises en utilisant un indicateur (GSI) qui donne un score permettant d'évaluer le niveau de risque "futur" si un lot ou un ensemble de lots était mesuré.

1.4.3.1 Global Sampling indicator GSI

Le GSI est un indicateur qui donne un score à différents ensembles de lots. À chaque ensemble des lots S est associé un niveau de risque attendu au sein de la production si les lots dans S étaient sélectionnés pour une inspection ou un contrôle. Considérons les exemples dans Table 1.1 et Table 1.2. Table 1.1 correspond à une

situation initiale où aucun lot n'est sélectionné. Table 1.2 montre deux situations si deux ensembles différents de lots $S1$ et $S2$ sont sélectionnés pour une inspection.

Équipements de production	Niveau de risque
M1	300
M2	250
M3	450
M4	450

Table 1.1: Situation initiale.

Équipements	Niveau de risque
M1	**50**
M2	**10**
M3	450
M4	450

(a) Ensemble des lots **S1** sélectionné.

Équipements	Niveau de risque
M1	**200**
M2	**200**
M3	**200**
M4	**200**

(b) Ensemble des lots **S2** sélectionné.

Table 1.2: Exemple si deux ensembles des lots S1 ou S2 sont sélectionnés pour une inspection.

Si l'ensemble des lots $S1$ est sélectionné et inspecté, le niveau de risque résultant sera celui décrit dans Table 1.2a, c'est-à-dire une réduction du niveau de risque sur les équipements de production $M1$ et $M2$. Si l'ensemble des lots $S2$ est sélectionné, on observe une réduction du niveau de risque sur tous les équipements ($M1$, $M2$, $M3$, and $M4$). Dans le premier cas, lorsque $S1$ est sélectionné, le niveau de risque est fortement réduit pour les équipements de production $M1$ (=50) et $M2$ (=10) alors que $M3$ et $M4$ conservent un niveau de risque très élevé (=450). Dans le second cas, quand l'ensemble $S2$ est sélectionné, le niveau de risque est réduit pour tous les équipements. Cependant, dans ce deuxième cas, le niveau de risque reste très élevé pour les équipements $M1$ et $M2$ en comparaison au premier cas où l'ensemble $S1$ est sélectionné. D'où la question suivante : *est-il plus intéressant de sélectionner un*

ensemble des lots dont la mesure ou l'inspection permettrait de fortement réduire le niveau de risque sur un ou deux équipements, ou bien sélectionner un ensemble de lots qui permettrait de réduire de très peu le niveau risque sur tous les équipements ?

Pour répondre à cette question, l'indicateur GSI a été développé pour donner un poids ou score à chaque ensemble des lots S en fonction des paramètres de contrôle et de la capacité de contrôle disponible. L'ensemble des lots S peut être vide (Table 1.1) ou non (Table 1.2).

Notations :

- R : nombre de risques,
- WL_r : *Warning Limit* pour le risque r,
- IL_r : *Inhibit Limit* pour le risque r,
- RV_r : valeur actuelle pour le risque r,
- $G_{r,l}$: gain sur le risque r si le lot l est inspecté,
- $NRV_{r,l}$: nouvelle valeur du risque si le lot l est inspecté, c'est-à-dire $NRV_{r,l} = RV_r - G_{r,l}$.
- $NRV_r(S)$: nouvelle valeur du risque si les lots dans l'ensemble S sont inspectés, c'est-à-dire $NRV_r(S) = Min_{l \in S} NRV_{r,l}$.

Pour les contrôles défectivité, le risque RV_r correspond au *Wafer-At-Risk* (WAR) pour l'équipement de production r. Le WAR est le nombre de *wafers* ayant subi une opération de fabrication sur l'équipement de production r depuis le dernier contrôle réalisé en défectivité. Le gain $G_{r,l}$ est la valeur de réduction du WAR sur l'équipement r si le lot l est contrôlé. Deux paramètres de contrôle sont définis : *Warning Limit* et *Inhibit Limit*. La *Warning Limit* WL_r correspond à la valeur du WAR au-délà de laquelle la situation commence à devenir critique en terme de contrôle. L'*Inhibit Limit* IL_r est le nombre maximum de *wafers* qui peuvent subir une opération de fabrication sur un équipement de production entre deux

opérations de contrôles. Le dépassement de cette limite (IL_r) pour le WAR peut entraîner l'arrêt de l'équipement de production r.

En utilisant les paramètres ci-dessus, deux formules **GSI** calculant un score ont été proposées (voir Chapter 6 pour plus de détails). Plus faible est le score GSI associé à un ensemble S, meilleure sera la situation au sein de la production si l'ensemble S est sélectionné et inspecté. La première formule GSI ne considère que la valeur IL_r dans la détermination du score ou poids à associer à chaque ensemble des lots S :

$$GSI(S) = \sum_{r=1}^{R} \left[\left(\frac{NRV_r(S)}{IL_r} \right)^{1/\beta} + \left(\frac{NRV_r(S)}{IL_r} \right)^{\alpha} \right]$$

La deuxième formule GSI intègre, en plus de la valeur de l'IL_r, la valeur de WL_r dans le calcul du score à associer à S :

$$GSI(S) = \sum_{r=1}^{R} \left[\left(Min \left(1, \frac{\frac{NRV_r}{IL_r}}{\frac{WL_r}{IL_r}} \right) \right)^{1/\beta} + \left(Max \left(0, \frac{\frac{NRV_r}{IL_r} - \frac{WL_r}{IL_r}}{1 - \frac{WL_r}{IL_r}} \right) \right)^{\alpha} \right]$$

ou

$$GSI(S) = \sum_{r=1}^{R} \left[\left(Min \left(1, \frac{NRV_r}{WL_r} \right) \right)^{1/\beta} + \left(Max \left(0, \frac{NRV_r - WL_r}{IL_r - WL_r} \right) \right)^{\alpha} \right]$$

Les deux formules **GSI** sont utilisées dans deux algorithmes GSI (GSI-SA-1 et GSI-SA-2) pour l'échantillonnage dynamique et intelligent des lots (c'est-à-dire *sampling*, *skipping*, et *scheduling*).

1.4.3.2 Algorithmes du GSI (GSI-SA-1 et GSI-SA-2)

Les algorithmes GSI sont basés sur les formules GSI, la *Warning Limit*, l'*Inhibit Limit*, et certaines valeurs de seuil appelées *thresholds*. Les valeurs de seuil permettent de maîtriser le temps de cycle des lots en évitant de sélectionner et placer dans la file d'attente des lots ayant un gain, mais qui risquent de ne jamais être mesurés

à cause de l'arrivée sans cesse des lots ayant une valeur ajoutée plus importante. L'objective est double :

1. Échantillonner dynamiquement les lots tout en assurant une utilisation optimale de la capacité de contrôle disponible.

2. Minimiser le risque sur l'ensemble de la production tout en évitant au maximum d'atteindre ou dépasser la valeur de l'*Inhibit Limit* qui pourrait entraîner l'arrêt des équipements de production.

Trois différents seuils (*threshold*) ont été définis. Ils correspondent à des valeurs au-délà desquelles des actions spécifiques doivent être prises rapidement :

1. **Seuil minimum (T_{Min})** = gain minimum que doit apporter la mesure d'un lot pour que le lot soit sélectionné et placé dans la file d'attente lorsque cette dernière est vide.

2. **Seuil maximum (T_{Max})** = gain minimum que doit apporter la mesure d'un lot pour que le lot soit sélectionné et placé dans la file d'attente lorsque cette dernière est pleine.

3. **Metrology threshold (T_{Metro})** = gain minimum que doit apporter la mesure d'un lot pour rester dans la file d'attente lorsqu'un autre lot vient d'être mesuré[10].

Le seuil minimum (T_{Min}) est utilisé lorsque la file d'attente est vide. Le seuil maximum (T_{Max}) est utilisé lorsque la file d'attente est pleine. Lorsque la file d'attente est partiellement remplie, le seuil utilisé est proportionnelle la taille de la file d'attente, c'est-à-dire :

$$Threshold = T_{Min} + \left[\frac{NBQ}{SQ} * (T_{Max} - T_{Min})\right].$$

[10] À chaque fois qu'un lot est mesuré, les gains apportés par les lots présents dans la file d'attente sont impactés et donc recalculés.

où NBQ est le nombre des lots dans la file d'attente, et SQ la taille maximale de la file d'attente.

Les trois seuils (T_{Min}, T_{Max} et T_{Metro}) sont basés sur le GSI, c'est-à-dire que le gain de chaque lot l est toujours évalué parmi un ensemble S des lots :

$$Gain(l) = 1 - \frac{GSI(S \cup \{l\})}{GSI(S)} \in [0, 1] .$$

$Gain(l)$ est strictement positif car l'inspection d'un lot ne peut qu'améliorer la situation au sein de la production. En d'autres termes, l'inspection d'un lot ne peut pas augmenter le niveau de risque dans la production.

Les deux algorithmes GSI présentés dans cette section ont été implementés et évalués en utilisant le simulateur **S5** développé par l'École des Mines de Saint-Étienne [104] dans la cadre du projet Européen **IMPROVE**[11]. Plus de détails sont disponibles dans Chapter 6.

Notations :

- $S_{initial}$: ensemble des lots présents dans la file d'attente,
- NBQ : nombre de lots dans $S_{initial}$ ($NBQ = |S_{initial}|$), c'est-à-dire le nombre des lots dans la file d'attente,
- SQ : taille de la file d'attente,
- $NbIL(S)$: nombre des *Inhibit Limits* violés (c'est-à-dire $NRV_r > IL_r$) si l'ensemble des lots S est sélectionné pour une inspection,
- $NbWL(S)$: nombre des *Warning Limits* violés (c'est-à-dire $NRV_r > WL_r$) si l'ensemble des lots S est sélectionné pour une inspection.

A. Algorithme du GSI-1

Le premier algorithme GSI (GSI-SA-1) détermine le meilleur ensemble des lots $S^\star$ et utilise la formule GSI-1, c'est-à-dire :

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

$$GSI(S) = \sum_{r=1}^{R} \left[\left(\frac{NRV_r(S)}{IL_r} \right)^{1/\beta} + \left(\frac{NRV_r(S)}{IL_r} \right)^{\alpha} \right]$$

Si le nombre des lots déjà présents dans la file d'attente d'inspection est strictement inférieure à la taille maximale de file d'attente, c'est-à-dire $NBQ < SQ$, alors seul l'ajout d'un lot l dans $S_{initial}$ est évalué et comparé au non-ajout de l. Sinon, c'est le cas où $NBQ = SQ$, et donc toutes les combinaisons associées au retrait d'un lot $l' \in S_{initial}$ dans $S_{initial}$ et l'ajout de l dans $S_{initial}$ sont évaluées.

Dans l'algorithme ci-dessous, SS représente un ensemble d'ensembles des lots.

GSI-SA-1 – Sélection du meilleur ensemble des lots $S^\star$ en utilisant IL, WL, et la formule GSI-1

1: **Initialisation** : $S^\star = S_{initial}$
2: **Si** $NBQ = SQ$ **alors**
3: $SS = \emptyset$
4: **Pour** chaque lot $l' \in S_{initial}$
5: $SS = SS \cup \{S_{initial} \setminus \{l'\} \cup \{l\}\}$
6: **Fin Pour**
7: **Sinon si** $NBQ < SQ$ **alors**
8: $SS = \{S_{initial} \cup \{l\}\}$
9: **Fin Si**
10: **Pour** chaque ensemble des lots $S \in SS$
11: **Si** $NbIL(S) < NbIL(S^\star)$ **alors**
12: $S^\star = S$
13: **Sinon si** $NbIL(S) = NbIL(S^\star)$ **et** $NbWL(S) < NbWL(S^\star)$ **alors**
14: $S^\star = S$
15: **Sinon si** $NbIL(S) = NbIL(S^\star)$ **et** $NbWL(S) = NbWL(S^\star)$ **alors**
16: **Si** $GSI(S) < GSI(S^\star)$ **et** $[1 - GSI(S)/GSI(S_{initial})] \geqslant \left[T_{Min} + \frac{NBQ}{SQ} * (T_{Max} - T_{Min}) \right]$ **alors**
17: $S^\star = S$
18: **Fin Si**
19: **Fin Si**
20: **Fin Pour**

B. Algorithme du GSI-2

Le second algorithme GSI (GSI-SA-2) utilise la formule GSI-2 :

$$GSI(S) = \sum_{r=1}^{R} \left[\left(Min\left(1, \frac{\frac{NRV_r}{IL_r}}{\frac{WL_r}{IL_r}}\right)\right)^{1/\beta} + \left(Max\left(0, \frac{\frac{NRV_r}{IL_r} - \frac{WL_r}{IL_r}}{1 - \frac{WL_r}{IL_r}}\right)\right)^{\alpha} \right]$$

Contrairement au premier algorithme GSI, les valeurs de *Warning Limit* et *Inhibit Limit* ne sont plus des limites à éviter.

GSI-SA-2 – Sélection du meilleur ensemble des lots $S^\star$ en utilisant la formule GSI-2

1: **Initialisation**: $S^\star = S_{initial}$
2: **Si** $NBQ = SQ$ **alors**
3: $SS = \emptyset$
4: **Pour** chaque lot $l' \in S_{initial}$
5: $SS = SS \cup \{S_{initial} \setminus \{l'\} \cup \{l\}\}$
6: **Fin Pour**
7: **Sinon Si** $NBQ < SQ$ **alors**
8: $SS = \{S_{initial} \cup \{l\}\}$
9: **Fin Si**
10: **FPour** chaque ensemble des lots $S \in SS$
11: **Si** $GSI(S) < GSI(S^\star)$ **et** $[1 - GSI(S)/GSI(S_{initial})] \geqslant \left[T_{Min} + \frac{NBQ}{SQ} * (T_{Max} - T_{Min})\right]$ **alors**
12: $S^\star = S$
13: **Fin Si**
14: **Fin Pour**

La prochaine section présente un résumé du programme MILP développé pour calculer les valeurs optimales de *Warning Limit* et *Inhibit Limit* dans le but d'optimiser l'échantillonnage dynamique au travers des algorithmes GSI.

1.4.4 Optimisation d'échantillonnage dynamique : MILP

Cette section présente un partie du programme linéaire mixte que je propose dans ma thèse pour calculer les valeurs de *Warning Limit* et *Inhibit Limit* pour chaque équipement de production en fonction d'un historique de production. Deux autres versions améliorées du programme intégrant le délai entre les opérations de fabrication et les qualifications des équipements ont été proposées et le lecteur pourra trouver plus de détails dans Chapter 7.

L'objectif est de déterminer des limites **"réalistes"** qui permettent aux algorithmes GSI de prendre des décisions pertinentes et donc sélectionner les meilleurs ensembles des lots à contrôler. On cherche donc à minimiser l'*exposure* (risque global) en prenant en compte le volume total de la production, la criticité de chaque équipement, et le temps nécessaire pour valider chaque équipement.

Paramètres :

- E_t : *exposure* pour l'équipement de production t (c'est-à-dire le coût financier associé à chaque *wafer* ayant subi une opération de fabrication sur un équipement de production t).
- V_t : volume de la production sur l'équipement t.
- Pm_t : temps de mesure pour valider l'équipement de production t.
- K_{MAX} : nombre maximum de mesure pour chaque équipement de production.
- $CAPA$: capacité totale (donnée en temps) pour la mesure.
- M : nombre des équipements de production.

Variables :

- IL_t : *Inhibit Limit* de l'équipement de production t.
- d_t^k : variable binaire égale à 1 si le nombre de mesure pour valider l'équipement de production t est k, 0 sinon.

- E_{MAX} : *exposure* maximum.

Le modèle MILP est le suivant :

$$Minimiser\, E_{MAX} \tag{1.6}$$

Sujet à :

$$E_{MAX} \geq E_t * IL_t \quad \forall t \in \{1 \dots M\}. \tag{1.7}$$

$$IL_t \geq \sum_{k=1}^{K_{MAX}} \frac{V_t}{k} * d_t^k \quad \forall t \in \{1 \dots M\}. \tag{1.8}$$

$$\sum_{k=1}^{K_{MAX}} d_t^k = 1 \quad \forall t \in \{1 \dots M\}. \tag{1.9}$$

$$\sum_{t=1}^{M} \sum_{k=1}^{K_{MAX}} Pm_t * k * d_t^k \leq CAPA. \tag{1.10}$$

$$IL_t \geq 0 \quad \forall t \in \{1 \dots M\}. \tag{1.11}$$

$$d_t^k \in \{0, 1\} \quad \forall t \in \{1 \dots M\}, \quad \forall k \in \{1 \dots K_{MAX}\}. \tag{1.12}$$

$$E_{MAX} \geq 0. \tag{1.13}$$

Les contraintes 1.7 définissent l'*exposure* maximum parmi tous les équipements de production. Cet *exposure* est minimisé dans la fonction objectif. Les contraintes 1.8 expriment que l'*Inhibit Limit* de l'équipement de production t (IL_t) est supérieur ou égal au volume de la production sur t divisé par le nombre de mesure sélectionné pour valider l'équipement de production t. Les contraintes 1.9 spécifient le nombre de mesure pour l'équipement de production t, c'est-à-dire que une et une seule variable doit être égale à 1. La contrainte 1.10 assure que la capacité disponible de mesure ou de contrôle est respectée.

1.5 Solutions Spécifiques : Prototypes et Industrialisation

Dans cette section, je donne un aperçu des solutions spécifiques que j'ai développées dans la cadre de ma thèse pour valider les solutions générales que je propose. Plusieurs prototypes ont été développés mais je ne donne qu'un aperçu des deux principaux (Figure 1.8 et Figure 1.9) qui ont conduits à une industrialisation des concepts généraux de la thèse.

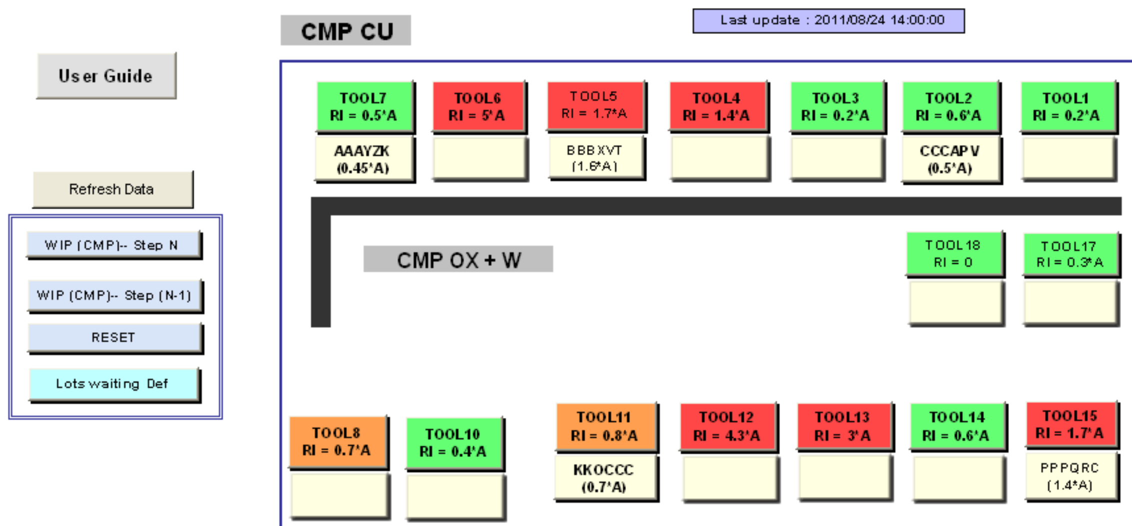


Figure 1.8: Vue générale du prototype CMP-WAR.

Le premier prototype (Figure 1.8) a été développé pour l'évaluation en temps réel du risque (*Wafer-At-Risk*) et l'amélioration du *dispatching* (c'est-à-dire la répartition des *wafers* ou lots sur les équipements de production) au sein de la production. Le second prototype (Figure 1.9) a été développé pour optimiser la gestion des excursions[12] en fournissant d'une part la liste des équipements de production les probables de la source de l'excursion, et d'autre part, la liste des lots à contrôler rapidement

[12]Une excursion intervient dans la production lorsque le contrôle sur un lot ou un équipement est jugé hors spécifications.

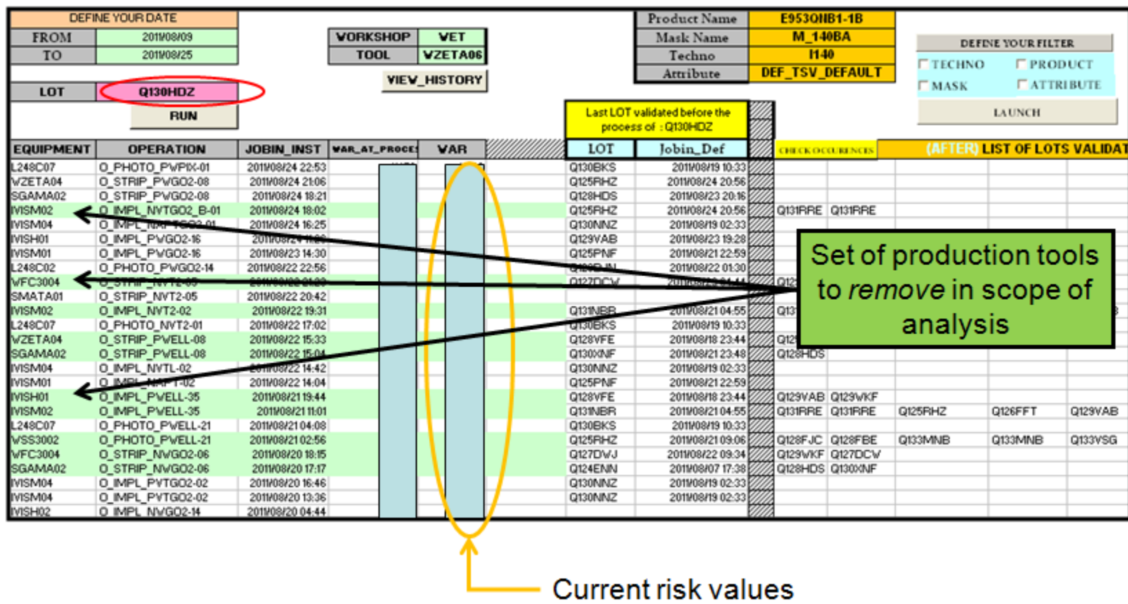


Figure 1.9: Vue générale du prototype de gestion des excursions.

pour confirmer ou infirmer l'excursion.

1.6 Conclusion et Perspectives

Dans ma thèse, je me suis intéressé au problème de la mise œuvre industrielle des plans de contrôle dynamique. Après avoir analysé et mis en évidence la complexité de la conception des plans de contrôle dans une industrie multi-produits, j'ai développé et proposé de nouvelles solutions que j'ai fait validé au travers des prototypes, simulations, et intéractions avec différents experts. Toutes les solutions proposées ont été validées industriellement et certaines ont été industrialisées au sein du site 300mm de STMicroelectronics à Crolles, en France. Les différentes solutions ont été comminquées et publiées dans des congrès, conférences et journaux internationaux. Une des communications a été recompensée avec le prix de *Best Student Paper Award* [60].

Plusieurs pistes ont été explorées ouvrant la voie à diverses perspectives. Les deux principales perspectives concernent l'**optimisation de la gestion des excursions** et l'**échantillonnage prédictif**. Concernant la gestion des excursions, le champ d'analyse (investigation ou recherche de la source du problème) pourrait être réduit en utilisant la notion d'ensemble dominant où le lot prioritaire à inspecter serait celui qui apporte le maximum d'informations sur l'ensemble des lots potentiellement impactés. En ce qui concerne l'échantillonnage prédictif, l'idée serait de ne plus sélectionner les lots en considérant uniquement les lots devant une étape de contrôle mais d'intégrer aussi des "lots futurs" c'est-à-dire des lots supposés arriver devant l'étape de contôle dans un "futur" très proche.

General Introduction

Semiconductor manufacturing is made of numerous and repetitive processing steps resulting in cycle times of more than two months. With the reduction in device sizes, re-entrant flows (repetition of similar processing steps), and the variety of products to be manufactured (more than 200 products in high-mix plants), the complexity has strongly increased in recent years. This complexity brings semiconductor manufacturers to introduce several layers of controls in order to guarantee high yield within production. However, most control operations are considered as non-added value and thus, when a control operation is introduced, cycle times increase with consequences on the final product costs. In the context of worldwide competition, companies have to provide pricing power against competitors. This implies that companies have to be able to sustain high yield with a minimum number of control operations.

Several works have been conducted on sampling techniques with the aim of minimizing the number of control operations without increasing the risk (i.e. material at risk) in production. Compared to static techniques, dynamic sampling techniques are more suitable for modern and high-mix semiconductor plants because they integrate factory dynamics and variability. However, the problem is in the industrial implementation of dynamic sampling approaches. The specificity of each semiconductor plant, the IT infrastructure, the variability of production flows, the heterogeneity of information systems, and the customer requirements are factors that strongly increase the complexity, leading to impracticability of many sampling algorithms proposed in the literature. The required investments are such that companies prefer to keep static sampling strategies whereas their inability to quickly

detect process drifts has already been pointed out.

This thesis aims at analyzing the efficiency of sampling policies, identifying breaches of controls, i.e. places throughout the process flow where control operations might be introduced or removed, assessing the added-value of each control operation, understanding why dynamic sampling techniques are seen efficient but most of the time impracticable, and providing novel solutions and approaches that can be industrialized. The thesis is realized within the framework of the *Conventions Industrielles de Formation par la REcherche* (CIFRE), in accordance with the *Association Nationale de la Recherche Technique* (ANRT) which supports companies that hire PhD students. The thesis is also written as a part of the European Union project IMPROVE (Implementing Manufacturing science solutions to increase equiPment pROductiVity and fab pErformance).

Reading plan

Generally, a scientific work is done according to the following schema [14]:

1. Problem definition,
2. State of the art review (literature review),
3. Case study,
4. Solution proposal,
5. Tests and validation,
6. Generalization and perspectives.

However, this is a thesis in an *industrial context* through a joint collaboration between industry and academics. There is an industrial problem and a research center must define the problem and propose innovative solutions. The case study comes before the literature review and proposed solutions are based on existing systems. Our work is thus structured into 7 main chapters:

- **Chapter 1**: Industrial Context.
- **Chapter 2**: Problem Identification and Research Issues.
- **Chapter 3**: Literature Review on Sampling Techniques.
- **Chapter 4**: Analyzing and Optimizing Control Plans.
- **Chapter 5**: Implementing Smart Sampling Policies.
- **Chapter 6**: Optimizing Smart Sampling Policies.
- **Chapter 7**: Industrial Developments and Implementations.

This decomposition can be linked to the **TRIZ**[13] approach [4] developed in 1946 by Genrich S. Altshuller for solving technical problems. The **TRIZ** approach is characterized by four main steps (Figure 1.10):

1. Problem identification and formulation.
2. Concept generation and comparison.
3. General solution.
4. Specific solution embodiment.

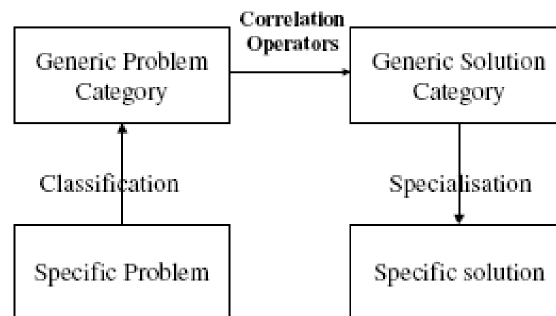


Figure 1.10: General problem solving model (TRIZ approach) [90].

Using the TRIZ approach, we can classify our work into these four main steps:

- Chapters 1 and 2 refer to **problem identification and formulation**.
- Chapter 3 refers to **concept generation and comparison**.

[13] **T**eoriya **R**esheniya **I**zobretatelskikh **Z**adatch. In English, it is defined as **T**heory of **I**nventive **P**roblem **S**olving (**TIPS**).

- Chapter 4, 5, and 6 refer to **general solutions**.
- Chapter 7 refers to **specific solution and embodiment**.

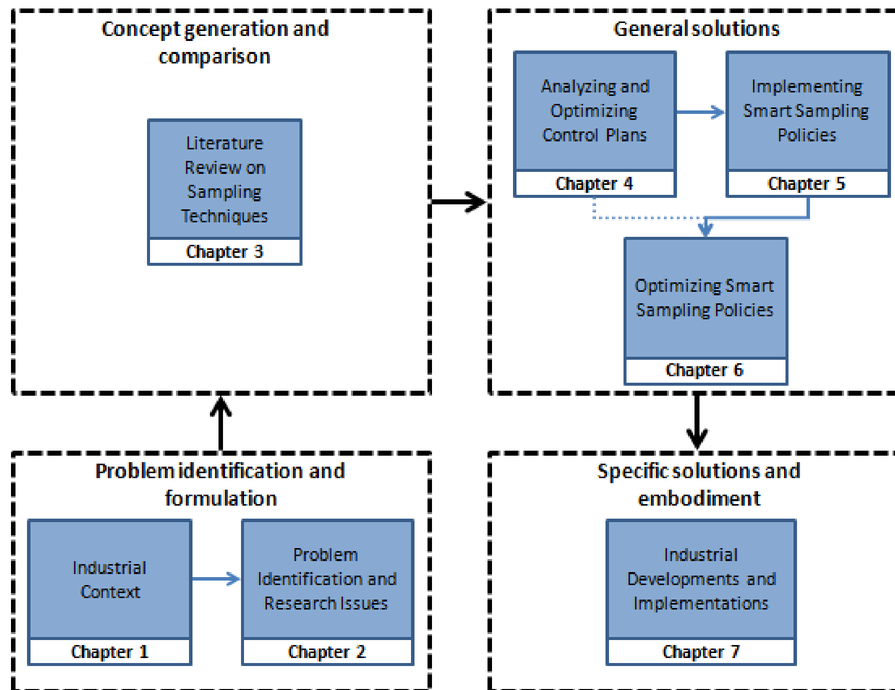


Figure 1.11: Thesis reading plan.

Chapter 1 introduces the industrial context. A description of the semiconductor industry is given, the main manufacturing steps are introduced, and controls performed throughout the production are presented.

Chapter 2 describes the problem tackled in this thesis. The specificities of STMicroelectronics Crolles are presented, and the thesis questions are introduced.

Chapter 3 surveys the literature on sampling techniques for controls in semiconductor manufacturing. Each sampling technique is reviewed through statements, critical analyses, and discussions on industrial deployments.

Chapter 4 analyses the impact of variability on static control plans, and introduces the fab-wide indicator (IPC) that has been developed to support the industrial implementation of dynamic control plans.

Chapter 5 introduces the dynamic sampling algorithms that have been developed within the framework of the European project IMPROVE.

Chapter 6 is devoted to optimizing solutions presented in chapter 4 and chapter 5.

Chapter 7 presents some prototypes that have been developed and deployed within the company during the thesis. These prototypes have been used to validate the novel approaches and algorithms that have been industrialized throughout the thesis.

The last part of the document is dedicated to a general conclusion and perspectives for further research.

Chapter 2

Industrial Context

This chapter introduces the context of the thesis: Semiconductor manufacturing and controls during the production. The focus is put on controls and especially on in-line measurements that aim at monitoring process and tool variations. The description of the different types of controls shows an important complexity linked to the size of manufactured products (Integrated Circuits). This thesis mainly addresses Defectivity controls where the objective is to detect and reduce particles generated on wafers during the production. All production tools are concerned and the variability within the production environment is such that the efficiency of a Defectivity control plan is never guaranteed. Hence our interest for this challenging problem.

2.1 *Introduction*
2.2 *Semiconductor Manufacturing*
2.3 *Controls in Semiconductor Manufacturing*
2.4 *Conclusion*

2.1 Introduction

Semiconductor industry is driven by the increasing demand of Integrated Circuits (ICs) in almost all domains (Automotive, communication, entertainment, multimedia, health care, energy saving, etc.). This strong demand leads to the pressure of delivering more and more products within reduced periods. However, manufacturing an IC requires more than 300 processing steps with a cycle time of at least two months. Moreover, with the device sizes reduction, the complexity is such that several types of controls are necessary to maintain high yield and high quality of products. Before, during, and after each processing step, several types of controls are performed to verify that the process is still under control and that products meet customer requirements. The challenge is therefore in finding the best trade-off between controls and risk[1] on the production. This chapter introduces the semiconductor environment, describes the main steps of fabrication, and the different types of controls.

Section 2.2 presents the production environment, the manufacturing stages, and the IC characteristics. In Section 2.3, we describe the levels and types of controls, and precise our focus within the framework of this thesis.

2.2 Semiconductor Manufacturing

The main activity of a semiconductor industry is to realize electronic components, interconnect them, and obtain **chips** or **ICs**. These ICs are used in quite diverse domains of everyday's life to perform different kind of functions (Temperature regulation, autopilot, television, smart-phones). Figure 2.1 shows how electronic chips drive our daily life. In almost each activity or each product, we use electronics chips.

[1] The concept of **risk** and associated actions is historically linked in industry to the area of quality and process control [7] [53]. In this thesis, the term risk is related to the material at risk, i.e. the potential loss if a problem occurs in production.

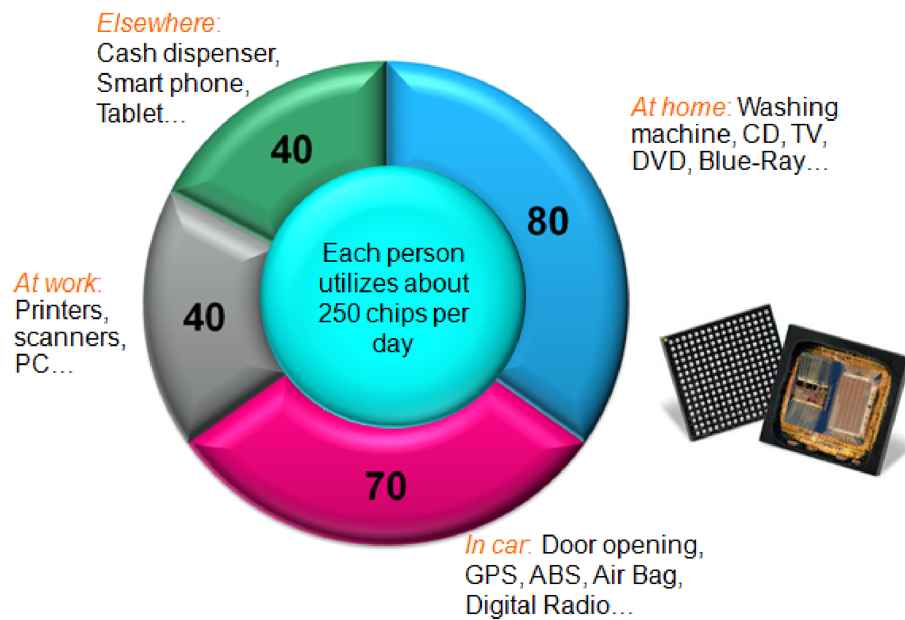


Figure 2.1: Integrated Circuits or chips in everyday's life.

The type and size of each IC varies depending on the targeted application. The trend toward mobility (reduced energy consumption) and the need for ever increasing computing power (increased speed) drives the race toward ever shrinking dimensions. The dimensions become so small that a special environment is required to avoid all kind of contamination.

2.2.1 Production environment

Integrated Circuits are manufactured in a specific environment called **clean room** (Figure 2.2). In the **clean room**, operators and engineers are covered from head to feet, and air is filtered and renewed every thirty seconds. Particles of a few hundredths of a micron in size are like meteors in this environment and might cause circuit faults and product failure. The average surgical operating room is three times dirtier than the dirtiest clean room in the world [74].

Figure 2.2: Clean room.

Several types or classes of clean rooms exist depending on the IC to be manufactured. The classification is performed based on the number of particles allowed per m^3 of air into the production environment. For example, a clean room of **class 2** corresponds to a clean room where 10^2 **particles greater than 0.1**μm of diameter are allowed. An ISO standard classification is provided in Annex B.1 and further details can be found in [97]. Several standards exist and the classification may vary from a country to another.

ICs are manufactured on silicon wafers that are sliced with a circular shape in order to minimize losses due to the wafer handling during the production. Figure 2.3 shows the evolution of the size of wafers through years. The trend is in increasing

the wafer size for producing more and more chips simultaneously (i.e. on the same wafer)[2].

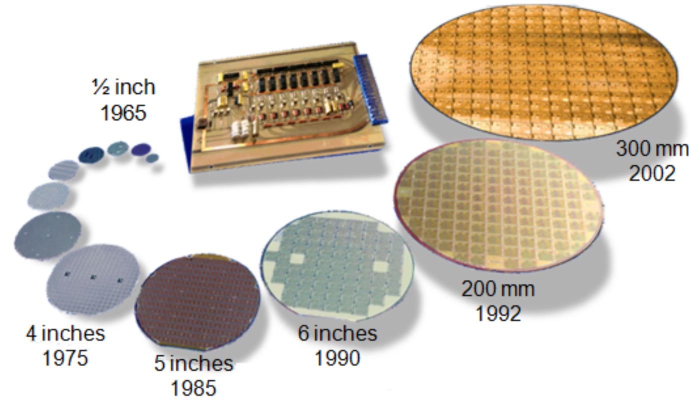


Figure 2.3: Wafer size evolution.

2.2.2 Manufacturing stages

Manufacturing stages for ICs are usually divided in two main parts [101]: **Front-End (FE)** processing and **Back-End (BE)** processing.

1. **Front-End** processing consists of several process steps that are repeated many times throughout the production[3] (Figure 2.4):

   – *Oxidation.* Silicon dioxide (SiO_2) is produced by heating the wafer to very high temperatures in the presence of oxygen.

   – *Photolithography.* Circuit patterns are formed by masking and etching processes.

   – *Implantation or doping.* After etching is completed, the exposed surfaces may be doped. Different types of dopants are added by ion implementation followed by diffusion processes.

[2]The more the number of chips produced on a wafer, the larger the reduction in the cost per die (or circuit).

[3]During the **Front-End** processing, wafers are manufactured by lots of 12, 25, or 50 wafers.

– *Chemical deposition.* Thin films of various material are deposited on the wafer through several processes (e.g. Chemical Vapor Deposition [CVD]).

– *Interconnect creation.* Sputtering or evaporation is used to create conducting circuits between individual electronic components and devices.

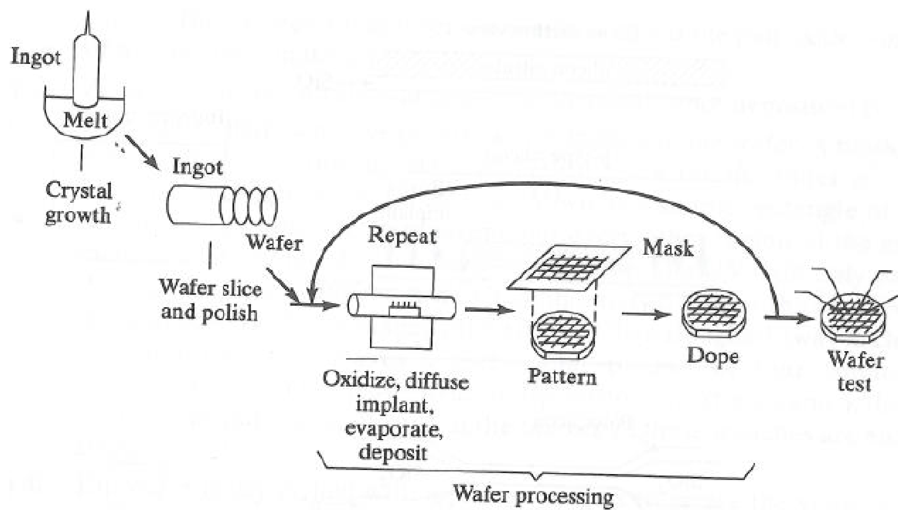


Figure 2.4: Front-End Processing [101].

2. **Back-End** processing refers to the testing, assembling, and packaging. It is performed at the end of the Front-End processing. During this second phase of fabrication, wafers are electronically tested for functionality and separated into individual dice (Figure 2.5). Each die is set into a chosen package, wire-bonded to the outer perimeter of the package, and finally tested for assembly onto a printed circuit board. Two main steps are defined: Testing-separation and Attachment-Wire Bonding-Packaging.

   – *Testing and separation.* During the Front-End processing, several tests are performed after each processing step (oxidation, etching, layering, and doping). During the Back-End processes, these test dice are put through an additional series of computer-controlled tests in which fine, needle like probes contact the aluminum bonding pads of the test dice. If

results indicate that the processing parameters were within proper limits, then each die is tested for functionality. Dice that need to be rejected are marked with an ink spot.

– *Attachment and Wire Bonding.* Good dice are seated into a desired package. Wire bonding makes the electrical contacts between the top of the die and the surrounding lead frame of the package. The package and packaging material chosen for a chip depend on the IC's size, number of external leads, power and heat dissipation, and intending operating environment.

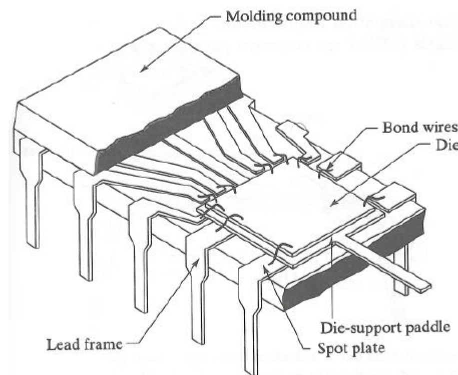


Figure 2.5: Back-End Processing [101].

In STMicroelectronics Crolles in France, only Front-End processing steps are performed. Therefore, within the framework of this thesis, we focus only on the Front End processing and especially on controls between processing steps (oxidation, etching, layering, doping, etc.).

2.2.3 Integrated circuit

An IC is generally made of four main components: Resistances, diodes, capacitors, and transistors [14]. These four main components are firstly realized on the silicon wafer before being interconnected to perform specific functions (e.g. automatic switch or regulation). Among these four components, the **transistor** is the principal component because of its ability of amplifying solids [73].

2.2.3.1 Transistor

A transistor is a combination of two diodes sharing a common region. There are two main types of transistors: Bipolar and Metal Oxide Semiconductor (MOS) transistors. Bipolar transistors are used to perform analog functions of power at very high frequency. MOS transistors are used for counting or memorizing i.e. performing logical or binary functions. Most of the integrated circuits are based on MOS transistors and in the 300mm site of STMicroelectronics Crolles, only MOS transistors are used[4].

A transistor is made of four terminal devices including a **gate**, a **source**, a **drain**, and the **bulk (silicon)**. Among these four terminals, the **gate** is the one that determines the technology node. Figure 2.6 gives a scale factor of the gate length in today's transistors.

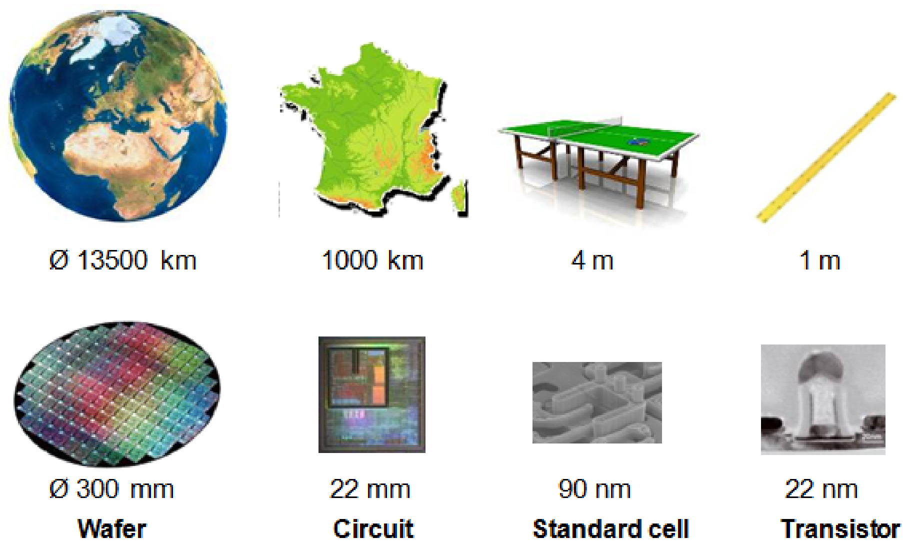


Figure 2.6: Transistor size - scale factors.

[4] In the 200mm site of STMicroelectronics, **BiCMOS** (Bipolar + MOS) transistors are used.

2.2.3.2 Technological evolution

The technological evolution in semiconductor manufacturing is linked to the transistor gate length which determines the technology node. This length decreases regularly following **Moore's Law** [6]. The Moore's law has been edited in 1965 by Gordon E. Moore. It states that the number of transistors placed in the IC will double every two years due to the size reduction. Since its edition, the Moore's law has become one of the driving principle of the semiconductor industry. All manufacturers are challenged with delivering annual breakthroughs ensuring compliance with Moore's law. In 1975, the law has been rectified by bringing to 18 months the rhythm of doubling the number of transistors within an IC. In 1997, Gordon E. Moore predicted the end of his law in 2017 because of the physical limits. Today, the trend is to do **"More than Moore"** by focusing on the system integration rather than the transistor density within the IC. For example integrating a camera into a cellphone or a cellphone into a PDA[5]. One of the main consequences of the Moore's law is the significant reduction in product prices (Figure 2.7).

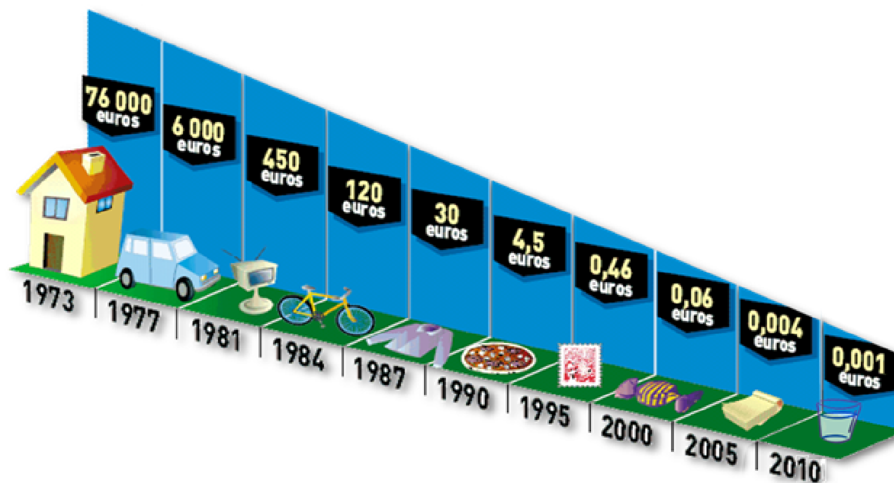


Figure 2.7: Impact of Moore law (Cost of 1MB of memory on silicon).

[5]Personal Digital Assistant.

2.3 Controls in Semiconductor Manufacturing

In semiconductor manufacturing, controls are necessary evils because of the prohibitive amount of time required to manufacture a chip [58]. Different levels of controls are defined and, for each level of control, several types of control are performed.

2.3.1 Levels of controls

Six main levels of controls can be defined [8]:

1. **Facilities or technical installations**. To guarantee the best possible environment for the fabrication of wafers, a huge number of parameters have to be monitored regarding the technical installations:
   - Clean-room *ambient* characteristics (temperature, humidity, pressure, contaminants).
   - Fluids, liquids, and gases (temperature, pressure, flow, contamination, etc.).
   - Energy (load, intensity, voltage, consumption, etc.).
   - Process outputs, wastes, gases, etc.
2. **Equipment sensors**. To ensure efficient processing operations, all variations have to be detected and analyzed. For that, several types of sensors are placed on different production tools to trigger alarms and actions in manufacturing systems.
3. **Fab or In-Line measurements**. This level of control groups measurements that are performed on silicon wafers with a large variety of techniques: Ellipsometry, reflectivity, scanning electron microscopy, visual inspection, pixel to pixel comparison, resistivity, scatterometry, etc. Measurements are classified according to the following characteristics[6]:

[6] The list is not exhaustive.

   - Impact: Destructive or non-destructive.
   - Support: Product wafers, Non-Product Wafers (NPW), monitoring wafers, dummy wafers, or test wafers.
   - Throughput or capital cost of the measurement.
   - Easiness of the required qualifications.

4. **Parametric testing**. Once transistors and other various devices are *connected* through metalization on the wafer, it is possible to control their performance versus their specifications. These measurements generally address basic parameters of electrical devices: Transistor voltage thresholds, leakage current, oxide breakdown voltage, via resistance, etc. Measurements are done on standard test structures placed in wafers *scribe lines*. All the wafers are measured on a limited number of sites per wafer.

5. **Final or Functional tests**. Once the front-end process is completed, semiconductor devices are subjected to a variety of electrical tests to determine if they function properly. The proportion of devices on the wafer found to perform properly is referred to as the yield. The fab tests the chips on the wafer with an electronic tester that presses tiny probes against the chip.

6. **Physical Characterization and Wafer level Reliability**. This last level of controls is used to evaluate component life time under various stressing conditions (humidity, temperature, corrosion, etc).

2.3.2 Types of controls

Several types of controls exist depending on the level of control. In this thesis, we focus on **In-Line measurements** and especially on the types of controls related to the process and equipment monitoring. There are five main types of controls: Fault Detection and Classification (FDC), Statistical Process Control (SPC), Run-to-Run (R2R), Virtual Metrology (VM), and Defectivity controls. The four first types of controls (FDC, SPC, R2R, and VM) are elements of **Advanced Process Control (APC)** [85] (Figure 2.8).

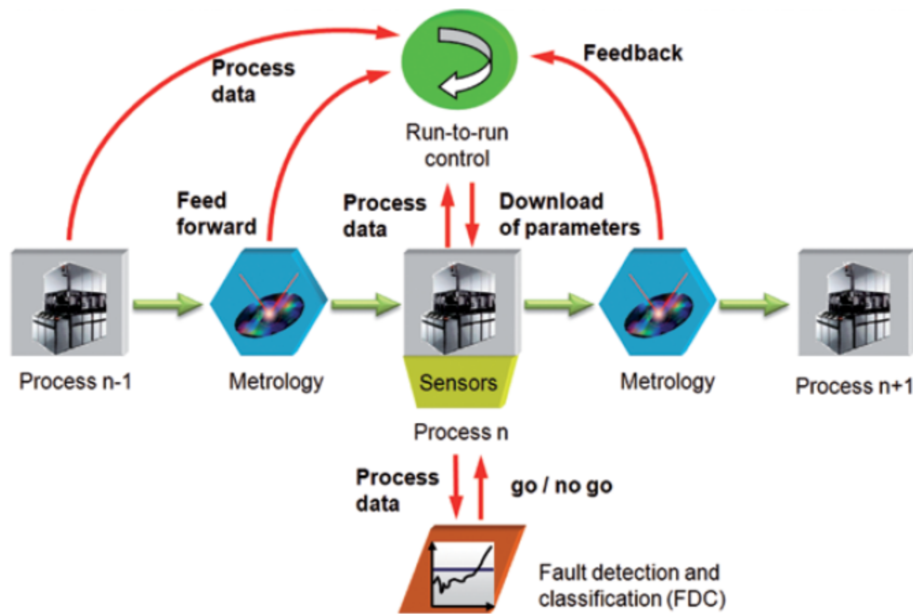


Figure 2.8: Interaction of APC elements [85].

1. **Statistical Process Control** consists in using statistical methods to analyze the process stability (Figure 2.9). Depending on the process state, different actions (stop the process tool, adjust the process parameters, etc.) are taken to achieve or maintain a state of statistical control. The objective is to continuously improve the process capability [55] [75]. Several SPC tools exist and they are based on the so-called Western Electric Rules[7] [96].

2. **Fault Detection and Classification** consists in statistically monitoring process variations by analyzing the process tool parameters (temperature, pressure, gas flow, optical emissions, etc.) [75] (Figure 2.10). During the process fabrication, all the tool parameters are collected for each processed wafer. A series of curves representing the evolution of these parameters during the time of the process are plotted for each wafer. Based on these data collected on both the tool and on wafers during their processing, different correlations are

[7]Western Electric Rules are decision rules for detecting "out-of-control" or "non-random" conditions on control charts (control charts are tools graphically displaying the process stability or instability over time).

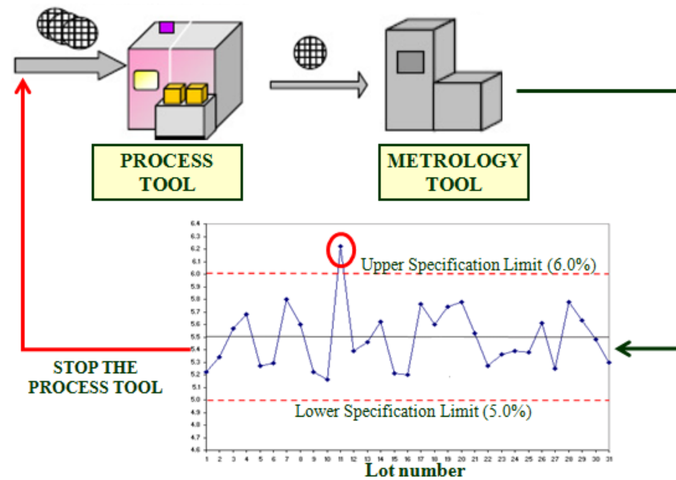


Figure 2.9: Statistical Process Control [75].

automatically computed. Whenever a problem is detected, the process tool is automatically stopped and actions immediately taken. The main difference between SPC and FDC is that FDC is a real-time based solution and can stop the process tool before the end of a processing step [2].

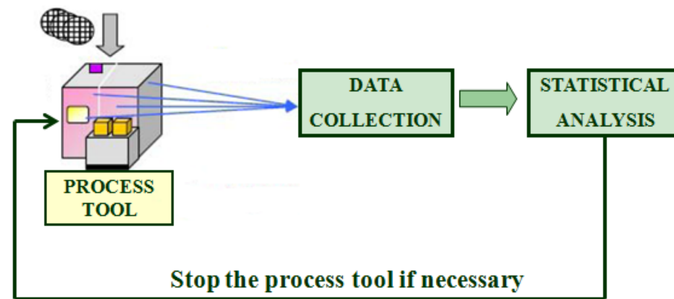


Figure 2.10: Fault Detection and Classification [75].

3. **Run-to-Run** is a closed-loop control solution to correct for process deviation from the desired target (Figure 2.11). The technique consists in modifying recipe parameters between production runs[8] to improve processing performance. In serial processing, this method can just be applied between two

[8]A **run** can be a batch, lot, or an individual wafer.

measurements [2]. It is an in-line technique like the FDC system. As illustrated in Figure 2.11, **R2R controller** is a *supervisor* that indicates whether the **automatic controller** needs adjustment. It consists in manipulating the set points (also know as recipes) of the underlying automatic controller in a supervisory manner in an attempt to reduce the output variability. Two types of control loops are used: **Feed-Forward** and **Feed-Back**. **Feed-Forward** control loops are used to reduce the impact of variability observed on run **N** by modifying the parameters of run **N+1**. **Feed-Back** control loops are used to counteract possible process drift. Using a predefined model [2], the difference between the desired target and the actual measurement value on run **N** is computed. This difference is used to adjust the parameters of run **N+1** and ensure that the desired target will be reached.

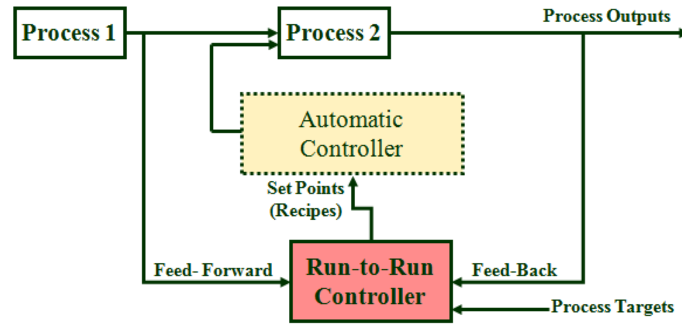


Figure 2.11: Run-to-Run.

4. **Virtual Metrology** consists in predicting measures, hence the term virtual metrology, based on previous metrology measurements and FDC data [15] (Figure 2.12). Wafer parameters are derived from upstream metrology (e.g. process state, additional sensors, temperature, pressure, gas flow, etc.) by using physical or statistical models, or hybrids models [26]. The objective is to reduce direct measurements on the wafers and to provide additional *virtual* measures to help alerting earlier when a process is drifting. The technique is based on predictive models that can forecast the electrical and physical parameters of wafers, based on data collected from the relevant process tools [39].

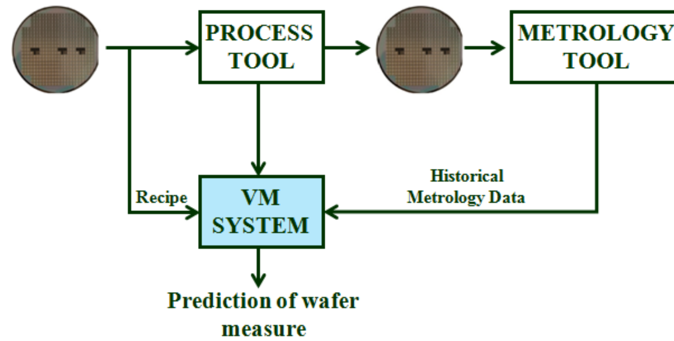
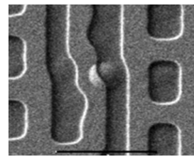
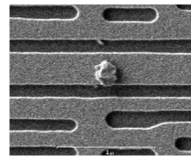


Figure 2.12: Virtual Metrology [39].

5. **Defectivity control** consists in detecting, analyzing, and reducing the number of defects generated on wafers during production (Figure 2.13). The phases of analysis and reduction are also known as the **review phase** where defects are first analyzed and then classified by type or class [46] [22].



(a) Embedded particle.



(b) Simple particle.

Figure 2.13: Examples of defects on wafers.

Detecting defects means deploying different tools and methods to capture defects generated by production tools on product or non-product wafers. Two kinds of detection are used: Optical Detection and Scanning Electron Microscope (SEM).

Analyzing defects means deploying methods to examine defects detected on wafers. Three main steps are performed during the analysis phase: Review, classification, and source identification. The review phase consists in verifying

if the defect found on a wafer is a known defect, i.e. if the defect has already been encountered in the past and classified into a specified group of "known defects". The classification phase consists in defining new groups for new types of defects or classifying them into existing groups. The source identification phase consists in locating the source of the defect, i.e. identifying the production tool that has generated the defect [61].

Reducing the number of defects means developing and deploying different tools and methodologies in order to lower the number of classified or known defects that may appear on wafers during processing steps. The main objective is to increase yield by reducing the number of bad wafers that may be discarded during the final tests.

Compared to other types of controls, defectivity controls have the specificity to potentially address all workshops and all production tools within the Fab (Chapter 3). This specificity leads to an increasing complexity when designing control plans because of the factory dynamics and variability during production (Chapter 5). This explains our interest and focus within the framework of this thesis.

2.4 Conclusion

In this chapter, we have rapidly presented the context of this thesis: The Semiconductor industry. The main activity of this industry is to design and produce ICs that are used in quite diverse domains of everyday's life. The production is divided into two main areas called *Front-End* and *Back-End*. The *Front-End* part consists in manufacturing and interconnecting different components on wafers in order to obtain ICs. In the *Back-End* part, ICs are individually tested, assembled, and packaged depending on the targeted operating environment. In STMicroelectronics in Crolles, France, the production is focused on the *Front-End* part, and more than 300 processing steps are necessary to realize a functional IC. This huge number of steps, combined the size of ICs (nanometer) and the non-reversibility of some processing steps explain the importance of controls in such an environment.

The description of the different levels and types of controls helped us to understand the source of complexity linked to both the environment and the size of products to be manufactured. In this thesis, we focus on controls related to the process and tool monitoring, and especially on one of the most complex type of control: Defectivity measurements. The particularity of defectivity controls is that all the production tools and all areas of production may be concerned. In the next chapter, we go in depth regarding defectivity controls and especially defectivity controls in STMicroelectronics in Crolles France. We aim at identifying the problem and defining methods to solve it efficiently.

Chapter 3

Problem Identification and Research Issues

Industrial research problems are not tackled the same way than theoretical research problems. The problem must be clearly identified and formulated in order to propose general solutions that can be industrialized. This chapter introduces Defectivity controls, specificities of STMicroelectonics Crolles in France, and research issues in the industrial context. We aim at concretely understanding the problem, formulating it, and defining strategies to solve it efficiently.

3.1 *Introduction*
3.2 *Defectivity Controls and STMicroelectronics Specificities*
3.3 *Research Issues and Solving Approaches*
3.4 *Conclusion*

3.1 Introduction

The strong competition in semiconductor industry is such that companies are called to search for different avenues to reduce production costs without impacting quality of final products. One of the identified tracks is controls throughout production. As more than 300 processing steps are required to produce a functional IC, the trend is to include additional control steps in order to detect as quickly as possible potential drifts. However, each additional control has impacts on the final product costs and, therefore, missing to find the right trade-off between controls and risk i.e. material at risk in production can lead to significant losses. The larger the number of products to be manufactured, the more finding a trade-off between controls and risk becomes complex. This is especially true in high-mix semiconductor plants where more than 200 products are run concurrently on hundreds of production tools.

In this thesis, we focus on defectivity controls that have the particularity to address all production tools in all process areas. Each time a defectivity control is performed, the risk related to one or several production tools is impacted i.e. reduced. The danger of having redundant controls is thus increased with the number of processing tools. Hence it is necessary to identify the right position for control operations, remove redundant controls, and optimize the use of inspection capacity. This chapter introduces defectivity controls, explains the complexity, and brings out the challenges when designing an efficient and optimized control plan. Section 3.2 introduces in depth defectivity controls and, in Section 3.3, we present research issues and solving approaches.

3.2 Defectivity Controls and STMicroelectronics Specificities

Defectivity controls consist in detecting and reducing defects generated on wafers throughout production. The aim is to increase yield and provide support for new technologies or products. As introduced in Chapter 2, a defectivity control is a type of in-line measurement performed between process operations during the fabrica-

tion process. This type of control was originally used to understand the integration issues of the main bricks[1] for a given product or, identify the main causes of yield losses. Nowadays, the control of defectivity on products is seen as the most efficient way to master yield excursions[2] in a production line. This is because measurements done on *bare* wafers i.e. *Non Production Wafers (NPW)* or without pattern, are not representative of the actual operation of a process step. Moreover, by using product wafers for defectivity controls, there is no waste of productive time (machine) or material (NPW, monitoring wafers).

One of the main characteristics of defectivity controls is that all production tools are concerned. Each and every tool is a potential contributor to defectivity (even the simplest tool has mechanical parts moving). By performing a defectivity control on wafers of a lot, information is collected on the various tools that have been used to process the considered lot so far.

3.2.1 Defectivity activity and defects types

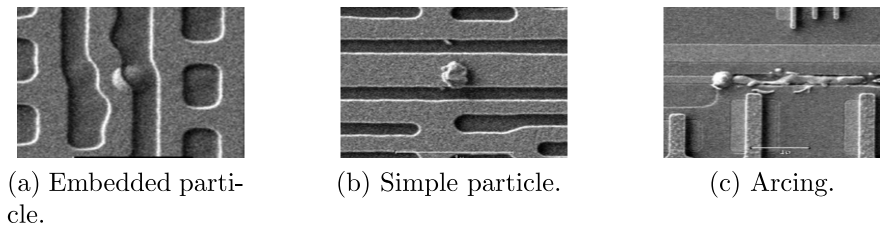
The main target addressed by defectivity controls or measurements is the reduction of the average level of defects per cm^2 and per photo-layer (also known as "**D**ϕ"), thus yield improvement. The activity of defectivity operators and engineers is twofold: **Inspection** which consists in the detection of defects (number of defects per wafer) and **Review** where subset of defects are analyzed and then classified by type or class (Section 2.3.2). Several types of defects may appear depending on the production step:

- Particles and arcing in the active zone (Figure 3.1),
- Voids (Figure 3.2),
- Scratches (Figure 3.3),
- Extra and missing patterns (Figure 3.4),
- Corrosions and plate-block (Figure 3.5),

[1]The term **brick** refers to a set of process operations that need to be performed for a given technology.

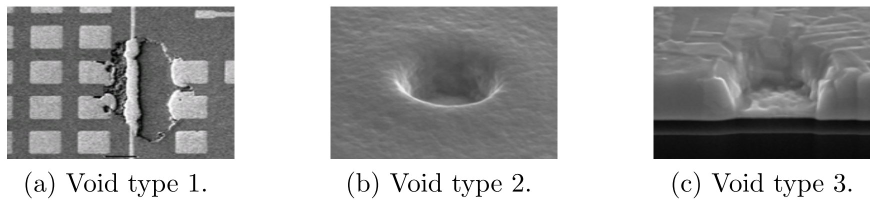
[2]An excursion is a deviation in process or product specifications. In other words, when a process or production tool is out of specifications, an *excursion* happens.

- Residues (Figure 3.6),
- Back-end type defects (Figure 3.7), etc.



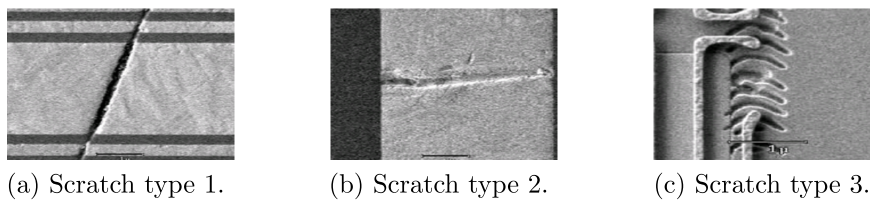
(a) Embedded particle. (b) Simple particle. (c) Arcing.

Figure 3.1: Particles and arcing on wafers.



(a) Void type 1. (b) Void type 2. (c) Void type 3.

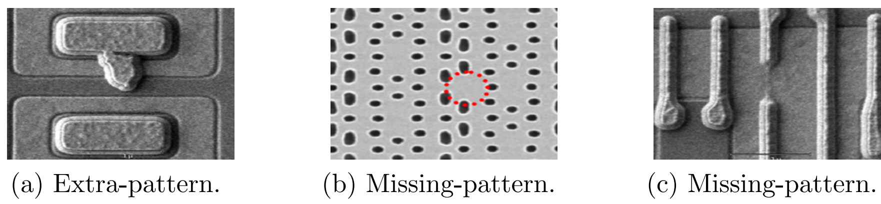
Figure 3.2: Voids on wafers.



(a) Scratch type 1. (b) Scratch type 2. (c) Scratch type 3.

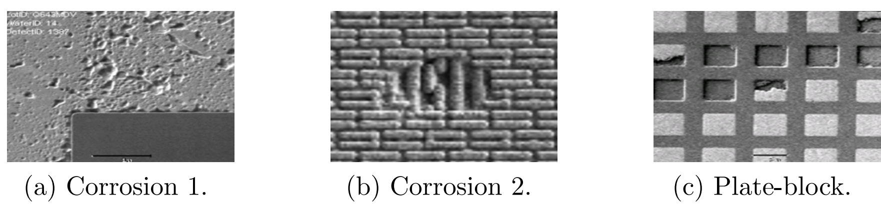
Figure 3.3: Scratches on wafers.

All defects on wafers have consequences on the final products. If they are not early detected, they can go through the production (Figure 3.7) and damage the functionality of the final product (IC). Hence it is critical to develop and set up specific methods and tools to capture the main defects as quickly as possible.



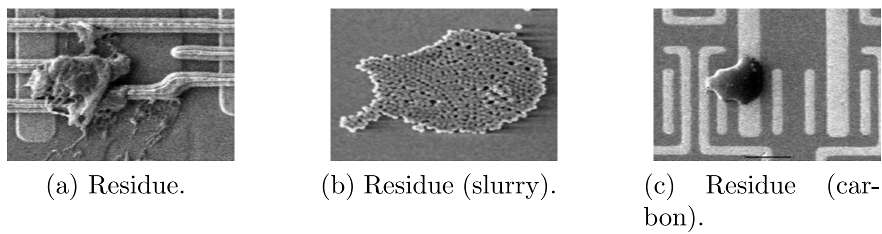
(a) Extra-pattern. (b) Missing-pattern. (c) Missing-pattern.

Figure 3.4: Extra and Missing patterns on wafers.



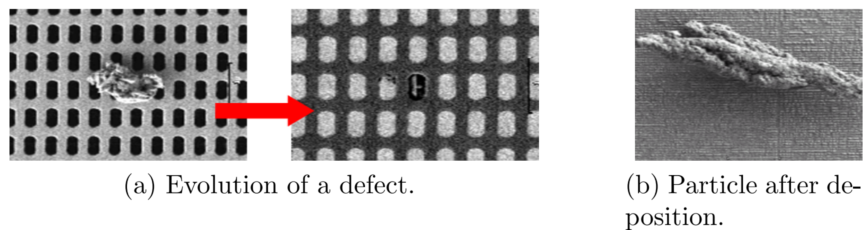
(a) Corrosion 1. (b) Corrosion 2. (c) Plate-block.

Figure 3.5: Corrosions and plate-block on wafers.



(a) Residue. (b) Residue (slurry). (c) Residue (carbon).

Figure 3.6: Residues on wafers.

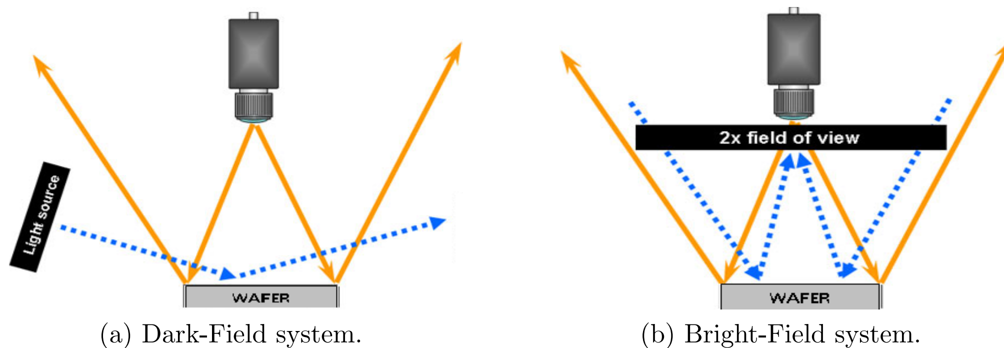


(a) Evolution of a defect. (b) Particle after deposition.

Figure 3.7: Other types of defects.

3.2.2 Techniques and tools

The diversity of defects generated by production tools on wafers is such that several types of inspection tools are necessary to capture all potential defects. The variety of tools and methodologies vary from one plant to another. Nevertheless, two main systems can be distinguished [89] [21]: Dark-Field and Bright-Field. Both systems consist in collecting scattered and reflected light on the wafer. The Bright-Field system collects both the scattered and reflected light through the same aperture to obtain an image (Figure 3.8b). The Dark-Field system only collects the scattered light. No part of the reflected light falls within the collection angle (Figure 3.8a). The difference between these two systems allows both small defects (Dark-Field system) and large defects (Bright-Field system) to be captured. The Dark-Field system has a high throughput compared to the Bright-Field system [31].



(a) Dark-Field system.

(b) Bright-Field system.

Figure 3.8: Dark-Field and Bright-Field systems [31].

Most of the tools used today for defectivity controls work on image comparison. Chips (on wafers) supposed to have the same image are compared. Based on the difference detected between images, it is possible to identify and localize defects on wafers [1]. This technique which was very expensive some years ago has been greatly enhanced with the available computing power. Nevertheless, as the technique is based on comparison of images, the set-up of measurement recipes is extremely complex and specific to each product operation (pattern, contrast, color, etc.).

Each time a defect is detected, several analyses need to be performed to identify the source of the defect, i.e. the processing tool that has generated the defect on the wafer [61] (Section 5.5).

3.2.3 Defectivity control plans and complexity

Defectivity control is instrumental to yield improvement when developing or ramping up a technology in volume (engineering phase). Once the main issues are identified and fixed, controls are relaxed and measurements are used to monitor and control production tools. These controls are considered as non-mandatory since they are not explicitly required to manufacture a functional IC. If they contribute in maintaining high yield within the production, they increase cycle times. That is why sampling is required to find a trade-off between yield and cycle time. Sampling rates are usually set by technology, at the start of production, and take into account different parameters such as the process criticality, the phase of integration, the maturity of products, or the customer requirements. For example, an old technology (e.g. CMOS120) will have a much lower sampling rate than a recent technology (e.g. CMOS028). Failing to find the right trade-off between yield and cycle time may lead to significant losses.

In a high-mix environment as in the 300mm fab of STMicroelectronics in Crolles, more than 20 technologies are run simultaneously, production tools are sometimes qualified to process more than 5 different technologies, process criticality varies from one technology (or operation) to another, processing throughput is different from one tool to another, each processing step requires a specific recipe for measurement, and customer requirements are varying. Furthermore, to each technology is associated one or several products. Each time a new measurement is introduced for a new product or operation, the corresponding recipe has to be created and engineered. Because of the variety of products running at the same time in the so-called "high-mix fab", doing this for every product is an overwhelming task. So, in such production lines, products[3] are divided into two main groups: **Measurable** and

[3] The cost of recipe set-up is so high that the only products targeted are the so called "big runners". As they are supposed to last a few months, they are seen as economically viable.

non-measurable products. By extension, lots of a product which can be measured in defectivity are said to be "measurable", and the others are "not measurable". Moreover, because of the cost of defectivity measurement tools and of the ever varying volumes of products (customer demands), it is not possible or interesting to measure all the lots of a given product. It implies that, among measurable lots, only a limited number will be chosen for a control and only at some pre-defined steps. The sampling is therefore done at lot start (beginning of the production flow)[4] and is mainly based on the experience of the engineering team. The objective is to control at least 90% of production tools in less than 24 hours.

3.2.3.1 Defectivity control plans: STMicroelectronics Crolles

In the 300mm site of STMicroelectronics in Crolles, the number of products and tools to be monitored is such that defectivity control plans are designed by technology depending on product specificities. For each technology, two matrices are designed. One matrix giving the set of control operations activated for each group of lots (Figure 3.9), and another matrix giving the **depth of control**, i.e. the set of process operations validated or covered by control operations (Figure 3.10).

Figure 3.9 shows an example of the control plan for the CMOS065 technology in the 300mm fab of STMicroelectronics. The column GENERIC_OPERATION gives the list of control operations performed in the defectivity area. Different attributes (DEF_C065_STANDARD, DEF_C065_FAST1, DEF_C065_FAST2, DEF_C065_FAST3, DEF_C065_FAST4, DEF_C065_FAST5, DEF_C065_OPTION) are defined. These attributes are associated to different measurable lots corresponding to a product of the CMOS065 technology. For each attribute, only a certain number of control operations will be performed on the lot (see the "**X**" in the graph). For example, if, to a measurable **lot** is associated an attribute "DEF_C065_OPTION", it implies that during the processing flow, only one control operation (18-O_STRIP_NSD) will be performed on this **lot**. This is the case

[4] Selecting lots at the start also guarantees that the same wafers will be inspected at operation N and N+X, thus enabling an easy identification of "added defects" and simplifying the analysis in the case a problem occurs.

for all technologies running within the fab. The processing flow varies from one technology to another, the names of attributes are different and specific to each technology, and the number of control operations activated by attributes is different. Some technologies can have more than 10 attributes. Considering an average of 20 technologies with at least 6 different attributes per technology, the engineering task consists in defining and updating 20 matrices with at least 120 attributes. **This is a challenging task** considering the high probability of missing key parameters.

	GENERIC_OPERATION	DEF_C065_STANDARD	DEF_C065_FAST1	DEF_C065_FAST2	DEF_C065_FAST3	DEF_C065_FAST4	DEF_C065_FAST5	DEF_C065_OPTION
1	O_PHOTO_ACTIVE	x						
2	O_ETCH_ACTIVE					x		
3	O_DEP_GAPFILL_STI	x	x			x	x	
4	O_OXID_SACOX	x	x			x	x	
5	O_PHOTO_PWELL							
6	O_IMPL_PWELL	x						
7	O_OXID_GO2							
8	O_DEP_POLY_GATE	x						
9	O_PHOTO_GATE							
10	O_ETCH_GATE	x	x			x	x	
11	O_PHOTO_PLDDLP							
12	O_STRIP_PLDDLP							
13	O_PHOTO_NLDDLP							
14	O_STRIP_NLDDLP							
15	O_DEP_TEO1_SPAC	x				x		
16	O_DEP_NIT_SPAC							
17	O_ETCH_NITRIDESPAC	x	x			x		
18	O_STRIP_NSD							x
19	O_PHOTO_PSD							
20	O_DEP_NIT_SIPRDIEL	x	x			x		
21	O_ETCH_SIPROT	x						
22	O_ANN_SPIKE							
23	O_RTP2_SALICIDE	x						
24	O_DEP_NIT_PMD	x	x	x		x	x	
25	O_CMP_PMD							

Figure 3.9: Example of a control plan for the CMOS065 technology.

As defects created at operation X may still be observable at operation X+D (depending on the "transparency" of successive process layers), one single defectivity

measurement may validate several process operations (tools). So, a second matrix giving the depth of control has to be designed. Figure 3.10 shows an example of this second matrix for the CMOS065 technology. Defectivity control operations (G, I,

	GENERIC_OPERATION											
A	O_OXID_PAD	x										
B	O_DEP_NIT_ACTIVE	x										
C	O_DEP_HMASK_ACTIVE	x										
D	O_DEP_AC_ACTIVE	x										
E	O_DEP_OXIDE_ACTIVE	x										
F	O_PHOTO_ACTIVE	x	x			x						
G	O_PHOTO_ACTIVE	x										
H	O_ETCH_ACTIVE		x			x						
I	O_ETCH_ACTIVE		x									
J	O_OXID_LINER			x								
K	O_DEP_GAPFILL_STI			x		x						
L	O_DEP_GAPFILL_STI			x								
M	O_ANN_STI											
N	O_CMP_STI					x						
O	O_ETCH_ON					x						
P	O_OXID_SACOX					x						
Q	O_OXID_SACOX					x						
R	O_PHOTO_NISO											
S	O_IMPL_NISO											
T	O_STRIP_NISO											
U	O_PHOTO_NWELL											
V	O_IMPL_NWELL											
W	O_STRIP_NWELL											

Figure 3.10: Example of depth of control for the CMOS065 technology.

L, Q) are those reported in the first matrix (Figure 3.9) (1, 2, 3, 4). For example, when a defectivity control operation is performed at "O_OXID_SACOX (Q)", six process operations are covered or validated (Figure 3.10): P, O, N, K, H, F. The process operation "O_ANN_STI (M)" is not validated because of the experience of the engineering team that concluded that no defect could be detected for that pro-

cess operation when measuring at "O_OXID_SACOX (Q)". Each technology has its own matrix giving the depth of control.

Consequently, during the processing flow, whenever there is a need to know if a lot has been flagged (i.e. selected for measurement at the start of production) or sampled for a defectivity measurement in the next defectivity control operation, four main steps need to be performed:

1. Identify the attribute associated to the lot.
2. Locate the current process operation of the lot.
3. Utilize the "depth-of-control" matrix (Figure 3.10) to determine whether there is a defectivity control operation or a set of defectivity control operations that can validate the process operation.
4. If there is a defectivity control operation that can validate the current processing step, then use the matrix (Figure 3.9) giving the set of defectivity control operations to be activated by each attribute to see whether the lot is flagged (**"X"**).

Let us consider **lot L1** currently in process operation O_CMP_STI (N) and having O_DEF_C065_FAST1 as attribute. To know if **lot L1** can be validated in the next defectivity operation, four steps will be performed:

1. Identify the attribute associated to the lot: O_DEF_C065_FAST1.
2. Locate the current process operation: O_CMP_STI (N).
3. Utilize the depth-of-control matrix in Figure 3.10 to determine whether the process operation O_CMP_STI (N) is validated by a control operation in defectivity. In the matrix, we can see that O_CMP_STI (N) is validated by the defectivity operation O_OXID_SACOX (Q) ("X" in the table).
4. Use the matrix in Figure 3.9 to see if the lot has a flag (or has been sampled) for a control in the next defectivity operation. We check if there is a "X"

indicating that lots having the attribute O_DEF_C065_FAST1 are sampled for a control in the defectivity operation O_OXID_SACOX (4). Since this is the case, **lot L1 will be validated by a defectivity control** in the next operation. **Lot L1** is "flagged" or "sampled" for the next defectivity operation (O_OXID_SACOX).

The complexity very quickly increases depending on the number of technologies that are run simultaneously and the number of products associated to each technology. Other parameters such as the process criticality and the capture rate[5] also play an important role in the design of the defectivity control plan. They are not explicitly modeled in the sampling or control plan but they contribute in increasing or reducing the priority of lots depending on the production state. The next section presents some of these parameters.

3.2.3.2 Factors increasing the complexity of designing control plans

Here are the main characteristics that contribute in increasing the complexity of designing defectivity control plans: Production tool qualifications, kill ratio, capture rate, Defect Work Request. These characteristics are discussed below.

- **Production tool qualifications** [36] [37] are related to the ability of production tools to perform some predefined process operations. One of the main goals when qualifying production tools is to ensure that tools are optimally used and provide flexibility for the entire production. However, by doing this, the focus is put on production tools, not on metrology or defectivity tools. The problem is that, when the sampling strategy is defined at the start of production, there is no information on the arrival of sampled lots in front of defectivity tools. As defectivity controls address the uncertainty of processing lots on production tools, some tools may have a high level of uncertainty while others will keep a low level of uncertainty. Figure 3.11 shows examples of qualifications of production tools for various operations. Operation O_OXID_PAD can be performed by three different production tools: WOASI01, WOASI02,

[5] The capture rate represents the sensibility in detecting defects for a given processing operation.

and WOASI03. These tools are also qualified for other types of operations (e.g. O_OXID_LINER). Depending on the availability of tools or types of products within the production, the priority will vary, some tools may be preferred compared to others, and it will impact the defectivity control plan.

GENERIC_OPERATION	EQPT_NAME1	EQPT_NAME2	EQPT_NAME3
O_OXID_PAD	WOASI01	WOASI02	WOASI03
O_DEP_NIT_ACTIVE	TASMI03	TINDY02	TASMI01(Ch)
O_DEP_HMASK_ACTIVE	TASMI06	TASMI05(Pi Ch Do)	
O_DEP_AC_ACTIVE	WSS3001	WSS3002	WSS3003
O_DEP_OXIDE_ACTIVE	DPROF03	DPROF06	
O_PHOTO_ACTIVE	L193C04	L193C05	L193C03(Re Pg)
O_ETCH_ACTIVE	EL23S03	EL23S04	EL23S05
O_OXID_LINER	WOASI01	WOASI02	WOASI03
O_DEP_GAPFILL_STI	DCENF03	DCENF04	
O_DEP_GAPFILL_STI	DCENF03	DCENF04	
O_ANN_STI	WFC3002	WOASI01	WOASI02
O_CMP_STI	CREFA01	CREFA05	
O_ETCH_ON	WFC3003		
O_OXID_SACOX	WFC3002	WOASI01	WOASI02
O_OXID_SACOX	WFC3002	WOASI01	WOASI02
O_PHOTO_NISO	WSS3001	WSS3002	WSS3003

Figure 3.11: Example of qualified tools per process operation.

- **Kill Ratio (KR)** defines the criticality of defects on wafers regarding the size of patterns [1]. It is generally between 0 and 1 [72] and defines whether a defect detected on a wafer is killer[6] (Figure 3.12) or not (Figure 3.13). Defectivity control plans designed by technology do not explicitly include this parameter (KR). Some process operations may be more critical than others because of the size of patterns, the product types, or the types of operations to be performed. Impacts on control are thus observed when there is a need to prioritize lots on defectivity tools or perform additional control operations. The start sampling plan is no longer respected.

- **Capture Rate (CR)** gives the percentage of defects that can be captured or detected at a given control operation [86]. All control operations (e.g. Figure 3.9) do not have the same CR because of the depth of control (e.g. Figure 3.10) and the criticality (KR) of some process operations. The priority

[6] A defect is identified as "killer" when it will definitively hinder the circuit functionality. This is linked to the size of the defect, hence its location in the circuit.

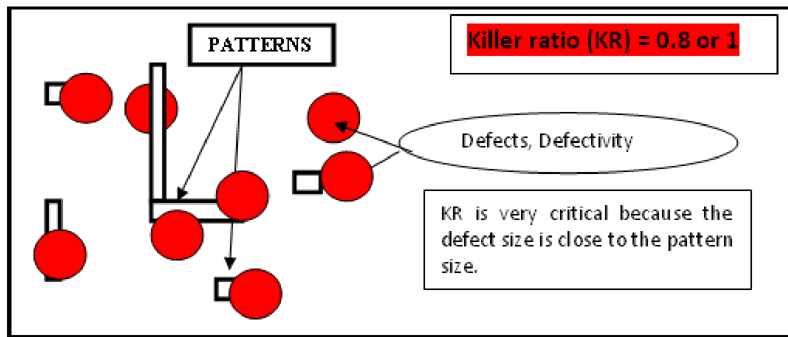


Figure 3.12: Killer defects.

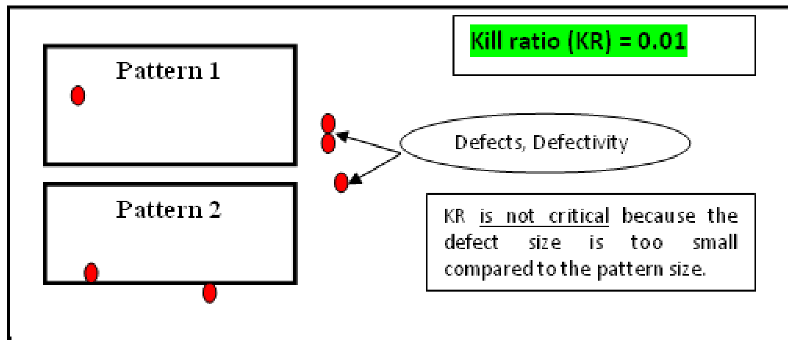


Figure 3.13: Non killer defects.

is thus different when defining sampling rate per product or technology. Hence an additional level of complexity.

- **Defect Work Request (DWR)** is a specific term used in STMicroelectronics Crolles to identify a lot that is sampled by the engineering team but information is not recorded in the Manufacturing Executing System (MES). A DWR is necessary when there is a suspicion in production and there is a need to quickly analyze a lot. This lot should normally contribute to reduce the risk level but, as no information is recorded in the MES, it is difficult to quantify the risk reduction. Moreover, this is an additional task for engineers regarding the sampling rate defined at the start of production.

All these additional factors or parameters are not explicitly included in the initial defectivity control plan but, they play an important role. They are considered by engineers when defining sampling rates, and when performing controls throughout production. It is hard to analyze and understand reasons for different sampling rates without being a member of the defectivity engineering team or working in close collaboration with defectivity experts. The large number of parameters to consider requires a high level of expertise to assess the efficiency of the control plan. This is one of the main motivations of this thesis: Try to think in a different way and find smart solutions for such problems.

3.3 Research Issues and Solving Approaches

In the previous sections, we saw that the complexity of designing a defectivity control plan increases with the number of parameters (types of products, technologies, process operations, criticality, etc.) to include or consider depending on the production state. There are situations where some controls need to be introduced throughout production, other controls need to be removed or released, and others are not even recorded in the automation system. One of the worst cases is when additional control operations do not contribute to reduce the risk at all. Increasing the number of control operations lead to instability because of the limited metrology capacity. The workload for defectivity engineers varies depending on the production state, there is no "standard" rule for prioritizing lots on defectivity tools, and the level of integration (number of parameters to consider) is such that it is impossible to assess the actual added value of controls or the efficiency of the entire control plan.

If there exist controls without actual added-value because of the complexity, can we say that there are too much controls or lack-of-control? Can we identify where control operations might be added or removed? **What can we propose to model and solve these issues?** Is it possible to have a general solution that could be understood by everybody and generalized for other types of controls than defectivity?

3.3.1 Research issues

This thesis, linked to the efficiency of control plans (especially defectivity control plans) tries to answer the three following main questions:

1. Over- or lack-of-control? And why?

2. Which kind of solution can be proposed, generalized, and industrialized?

3. How can the solution be optimized?

3.3.2 Solving approaches

This is a thesis in an industrial context. The objective is to propose solutions that can be, not only generalized, but also industrialized. We use the three following premises of the TRIZ approach to characterize our problem [4]:

1. The ideal design with no harmful functions is a goal.
2. An inventive solution involves wholly or partially eliminating a contradiction.
3. The inventive process can be structured.

Based on these three premises of the TRIZ approach, we choose as supports for the research: Interaction with experts and development of prototypes.

1. **Interacting with experts** helps us in understanding the industrial context, the origin of the problem, and the complexity we may face. We aim at avoiding traps and focusing on the final objective.
2. **Developing prototypes** helps us in testing and validating new algorithms or techniques in an industrial context. We aim at progressively validating our solutions in a production environment. As the efficiency of an algorithm or technique may vary depending on its application, we aim at avoiding developing theoretically efficient algorithms that are impracticable when aiming for industrial implementation.

3.4 Conclusion

In this chapter, we presented the problem tackled within the framework of this thesis and introduced our solving approaches. We described the defectivity control plan in the 300mm fab of STMicroelectronics in Crolles, and discussed the complexity that motivates our research. The focus is on the efficiency of controls throughout production. This thesis tries to answer whether there is over- or lack-of-control regarding the current control plan, and propose general solutions that can be industrialized and optimized.

In the next chapter, we survey the literature. We aim at identifying problems related to our thesis questions especially on sampling techniques, classifying these problems, and positioning our problem among solutions that have already been proposed.

Chapter 4

Literature Review on Sampling Techniques

This chapter provides a state-of-the-art[1] *on sampling techniques (at lots and wafers level) for non-mandatory controls in semiconductor manufacturing, and positions our problem in the literature. We observed that the specificities of each semiconductor plant is such that the efficiency of a sampling technique is directly linked to the production environment. Hence, our focus is on adaptive and dynamic sampling techniques that respond to the factory dynamics and variability.*

4.1 *Introduction*
4.2 *Sampling Techniques in Semiconductor Manufacturing*
4.3 *Static or Start Sampling*
4.4 *Adaptive Sampling*
4.5 *Dynamic Sampling*
4.6 *Conclusion*

[1]Part of this chapter has been submitted for publication in **IEEE Transactions on Semiconductor Manufacturing** [67].

4.1 Introduction

Sample measurement for a process parameter is a necessity in semiconductor manufacturing because of the prohibitive amount of time involved in 100-percent inspection while maintaining sensitivity to all types of defects and abnormality [91]. Moreover, a 100-percent inspection (or metrology) rate does not guarantee 100-percent quality since, in semiconductor manufacturing, inspection is never totally reliable and can easily introduce an error of almost the same order as the fraction of defectives [76]. If the development of sampling techniques is not recent in semiconductor manufacturing [23] [19] [33], significant improvements have been observed and, today, new challenges are being faced. Current computers offer the possibility to handle applications that were judged "infeasible" five or ten years ago. This opens the way for the development and implementation of very complex sampling techniques.

This chapter surveys the literature on sampling techniques for inspection or metrology steps (defect inspection or defectivity controls, critical dimension measurements, overlay, thickness, and step height measurements) in semiconductor manufacturing. **We aim at identifying works related to our problem, and analyzing solutions that have been proposed**. We discuss the trade-off between the cost of a measurement and the related cost in term of risk reduction, and the development of effective sampling techniques. We collected the literature from dissertations, working papers, technical reports, conference papers (Advanced Semiconductor Manufacturing and International Symposium on Semiconductor Manufacturing), and also from journals on Semiconductor Manufacturing, process control, and operational research. Each article is reviewed through statements, critical analysis, and also discussions on industrial deployments of various sampling techniques in semiconductor plants.

Through all of the papers browsed in our review, we observe that sampling techniques can be classified into three main groups: **Static or start, adaptive, and dynamic sampling**. Static or start sampling techniques are based on fixed rules

not changed throughout the production. Adaptive sampling techniques consist in adapting sampling rules defined at the start of the production. Depending on information brought by other types of controls (statistical analysis, process variations, maintenance, etc.), rules are adjusted in order to prevent potential drifts or reduce the material at risk. Dynamic sampling techniques consist in selecting in real time the best lots or wafers to inspect depending on the inspection capacity and the actual situation within the production. No rule is defined at the start of the production and the decision of selecting or not a lot is directly taken in front of the inspection step, and based on information brought by the lot.

Section 4.2 presents a general overview of sampling techniques in semiconductor manufacturing. In Section 4.3, we discuss static or start sampling techniques. Sections 4.4 and 4.5 are devoted to adaptive and dynamic sampling techniques respectively. For each group in our classification (static, adaptive, and dynamic sampling techniques), we analyze the different papers and articles using the six following indicators: Year, mathematical technique, rule-based technique, industrial deployment, simulation, and comparison with other techniques.

4.2 Sampling Techniques in Semiconductor Manufacturing

In semiconductor manufacturing, sampling techniques vary depending on the set of parameters to be monitored or production objectives. Three main groups are defined in the literature [48]:

- **Excursion monitoring and control** aim at frequently monitoring the process so that any process deviations are caught and the causes for the process excursion are fixed.

- **Process integration and yield improvement** aim at adjusting the percentage of lots flagged at the start of their production (baseline lots) in order to identify the main detractors for a given technology and eliminate them.

For low-mix semiconductor plants, the percentage of lots flagged at the start of production is adjusted in order to compensate the potential loss based on measurement results.

- **Defect detection and learning** aim at learning on different defect types and their mechanisms: Killer rates. The sampling rate has to enable defect detection at a rate that is matched to the one of root-cause analysis and problem fixing.

Among these three groups of sampling techniques, we only focus on **excursion monitoring and control**, and therefore, the classification we propose (static, adaptive and dynamic sampling techniques) concern this first group of sampling techniques that includes **defectivity controls**. The objective is twofold: Reduce the number of measurements without increasing the risk in production, and detect as quickly as possible potential excursions. Missing to reach these two objectives can lead to significant losses. Indeed, if the focus is only on the reduction of measurements, the danger can be to miss the detection of potential excursions. When a process is likely to be out of control, increasing the number of measurements can help to detect excursions as quickly as possible. Similarly, if the focus is only on excursion monitoring, the danger is to increase the number of measurements leading to increased cycle times, and therefore increased product costs.

4.3 Static or Start Sampling

Static or start sampling consists in determining a fixed number of lots to measure at different manufacturing stages. The number of lots to measure depends on the available inspection capacity, the maturity of the technology, and the process step criticality [12]. The frequency and the sensitivity of the measurement are selected in advance, at the start of production. The objective is to monitor and detect process drifts and limit the material at risk [9] between controls. For example, if the sampling plan specifies to control one lot every five lots, the objective is to limit the material at risk to not more than five. Always measuring the same lots or wafers enables the identification of the added defect density between sequential inspection steps [30]. Another advantage is the simplicity of implementation and adequate management of resources [12].

Static sampling is being widely used in most semiconductor plants. However, it does not fit high-mix semiconductor plants because of its main drawbacks of not taking into account the factory dynamics and variability. By always selecting the same lots to measure, there is, for the selected lots, a strong impact on the cycle time and an increased risk of yield losses due a higher number of steps and the significant time spent in front of each inspection step. Nevertheless, start sampling is still used during the phase of integration for some specific products. In some semiconductor industries, especially in a low-mix context, where a production tool can be qualified to process only a specific type of product, start sampling remains valid and some optimized solutions can be designed.

Among papers surveyed in Table 4.1, note that, even if all papers are applied to a case study of a semiconductor plant, very few provide industrial deployments [94] [87] [43]. In [94], the study performed in an IBM plant to determine the optimal sampling plan for the poly etch module is described. The goal is to minimize both the risk for the product and the cost of inspection. Three decision variables are considered in the study: Lot sampling interval, number of wafers per lot, and process control limits. Results indicate that an optimal sampling plan may require

	Year	Mathematical technique	Rule-based technique	Industrial deployment	Simulation	Comparison with other techniques
Lazaroff *et al.* [45]	1991		*		*	
Nurani *et al.* [70]	1994		*		*	
Nurani *et al.* [68]	1996	*				
McIntyre *et al.* [52]	1996		*			*
Tomlinson *et al.* [94]	1997		*	*		
Scanlan *et al.* [84]	1998		*			
Elliott *et al.* [25]	1999		*			
Chien *et al.* [18]	2000	*				
Lee *et al.* [48]	2001	*				
Chien *et al.* [17]	2001		*			*
Shumaker *et al.* [87]	2003		*	*		
Xumei *et al.* [103]	2003		*		*	
Wu and Pearn [102]	2006	*			*	
Kwang and Chin [43]	2008		*	*		

Table 4.1: Survey on static or start sampling

additional inspection capacity whose cost is much lower than the benefits. In [87], a sampling method developed at Motorola is discussed. The method is based on two steps: The first step consists in determining products that are good candidates for sampling, and the second step performs analysis to determine the break-even operating constraints. The method is developed and validated against historical data. Results indicate a reduction of wafer test costs by a factor of 10. Kwang and Chin [43] worked on data management. They present an industrial deployment of an automatic push-pull sampling methodology. The methodology consists in the transition from manual to automated sampling controls in order to propagate the correct sampling data to the operators and reduce sampling errors due to human interventions. Results indicate an increase of two percent in productivity.

The efficiency of an algorithm depends on its application. This is the case in semiconductor manufacturing where the environment completely changes from one

factory to another and the degree of complexity is not always the same. Different mathematical techniques have been proposed but none of them has actually been deployed. Table 4.2 presents mathematical techniques and approaches surveyed in the literature. The complexity is such that most static or start sampling

	Algorithms or Mathematical Techniques
Nurani *et al.* [68]	Heuristic approach
Chien *et al.* [18]	Bayes' theorem
Lee *et al.* [48]	Self-Organizing Feature Map (SOFM) network
Wu and Pearn [102]	Process capability index **Cpmk**

Table 4.2: Mathematical techniques or approaches for static or start sampling

techniques are rule-based and take into account some observations within the fab, personal experiences, and statistical analysis. Lazaroff *et al.* [45] present an evaluation of different defect sampling techniques using linear regression. A discussion on the strengths and weaknesses of various sampling techniques for Critical Dimension (CD) measurement is presented by Elliott *et al.* [25]. Chien *et al.* [17] and Xumei *et al.* [103] worked on optimizing sampling techniques for overlay measurements and validated their experiments through simulations using historical data from semiconductor plants. Nurani *et al.* [70] present an economic model for optimizing a sampling plan. The model aims at specifying the number of lots to inspect, the number of wafers within a lot, and the number of dies per wafer. Increasing the cost of inspection (number of lots or wafers to inspect) leads to an increased benefit by detecting excursions very quickly. However, above a certain limit, if the cost of inspection is still increasing, all revenues gained by inspections will be offset by the increased learning and subsequent defect reduction. Close to the work of Nurani *et al.* [70] are the works of McIntyre *et al.* [52] and Scanlan *et al.* [84]. McIntyre *et al.* [52] discuss key factors that influence the cost of an optimal sampling plan and Scanlan *et al.* [84] identify the use of baseline lots as a key in cost inspection reduction.

In the papers on static or start sampling, the authors try to find the best trade-off between the cost of inspections and the cost related to the material at risk. However, decisions are only taken at the start of production and do not consider unexpected

events that may occur during the production. When the process is likely to be out-of-control for example, it could be more interesting to sample more lots or wafers in order to detect potential drifts as quickly as possible. When the process is within control, metrology capacity could be saved by reducing the number of sampled lots. These main drawbacks of static sampling led to the introduction and development of adaptive sampling strategies.

4.4 Adaptive Sampling

Adaptive sampling consists in adjusting sampling decisions defined at the start of production, i.e. the number of lots or wafers to select is adjusted throughout production depending on the process state. Table 4.3 presents a survey of adaptive sampling techniques in semiconductor manufacturing.

	Year	Mathematical techniques	Rule-based	Industrial deployment	Simulation	Comparison with other techniques
Prahbu *et al.* [77]	1994	*			*	*
Nurani *et al.* [69]	1995		*		*	
Kuo *et al.* [42]	1996		*		*	*
Kuo *et al.* [41]	1997	*				*
Babikian and Engelhard [5]	1998		*			
Williams *et al.* [99]	1999		*	*		
Williams *et al.* [98]	1999		*	*		
Langford *et al.* [44]	2000		*		*	
Nurani and Shantikumar [71]	2000	*			*	*
Lee *et al.* [49]	2001	*			*	
Wootton *et al.* [100]	2001		*	*		
Allebé *et al.* [3]	2002	*			*	
Lee [47]	2002	*			*	
Song-Bor *et al.* [88]	2003		*	*		
Sullivan *et al.* [92]	2004	*		*		
Moon *et al.* [56]	2005	*			*	
Boussetta and Cross [12]	2005		*	*		
Mouli [57]	2005		*			
Shantikumar [86]	2007	*				
Mouli *et al.* [58]	2007	*		*		
Bunday *et al.* [13]	2008		*			
Veetil *et al.* [95]	2009				*	*
Chen *et al.* [16]	2009	*			*	
Sahnoun *et al.* [83]	2010	*			*	
Sahnoun *et al.* [82]	2010	*			*	*
Good *et al.* [28]	2010	*			*	
Nduhura Munga *et al.* [60]	2011	*		*		

Table 4.3: Survey on adaptive sampling

The transition from static to adaptive sampling started in the second part of the 1990's [77] and a great contribution can be noticed between 1995 and 2005. First industrial deployments can be observed in the beginning of the 2000's [99] [98] [100]. However, among twenty-seven papers browsed in this review (Table 4.3), only eight indicate an industrial deployment. Moreover, among these eight papers, no indication or comparison with other techniques or technologies is given. This shows the complexity and the particularity of the semiconductor environment. Depending on the amount of data to handle, and the strategies in semiconductor plant, a solution can be efficient when simulated but not practical because of unexpected events or factory dynamics. The specificity of each factory is such that a given solution can be efficient in a factory A and be completely impracticable for a factory B. This explains why no comparison is presented in the literature. Moreover, strong competition and confidentiality reasons explain why many works are not published. Most of the works published or patented do not detail the technical aspects, and actual performances are never published.

Among papers that indicate industrial deployments, Williams *et al.* [99] [98] present the results of a joint research project between Intel Corporation and KLA-Tencor. The project consists in evaluating and optimizing the defect inspection sampling plan for an advanced semiconductor manufacturing process. A Sample Planner is developed by KLA-Tencor to assist in the development of cost-effective defect inspection sampling strategies, and to provide an accurate assessment of whether monitor reduction and/or elimination should be pursued for cost savings. The results of the project indicate that the costs due to defect excursions could completely eradicate any projected savings from monitor reduction activities, due to the additional defect excursions that would be missed by the reduced inspection sampling plan.

Wootton *et al.* [100] present a study performed between KLA-Tencor and Motorola. The study consists in finding the best sample criteria providing the best representation of existing problems in the inspected wafers. The main drawbacks of random selection are presented and the proposed solution consists in adapting the sample size based on in-line information and priority rules (defect size). Re-

sults indicate an improvement of yield, analysis time, and sampling resolution at Motorola.

Boussetta and Cross [12] analyze the key parameters that have to be monitored for an efficient adaptive sampling plan. Their results indicate three key parameters: The variance ratio, the excursion frequency, and the normalized mean shift. They propose a general adaptive sampling plan and recommend a fab-wide strategy, a very good understanding of inspection requirements, and capacity constraints for an efficient adaptive sampling plan.

Song-Bor *et al.* [88], Sullivan *et al.* [92], Mouli *et al.* [58], and Nduhura Munga *et al.* [60] present industrial deployments of adaptive sampling plans in four different semiconductor companies: TSMC, IBM Microelectronics, Intel Corporation, and STMicroelectronics respectively. Song-Bor *et al.* [88] at TSMC present a capacity-dependence sampling strategy, based on the utilization rate of the capacity of defect inspection tools and on the WIP (Work-In-Progress) management. If the utilization of defect detection rises too high, then an automatic function that allows the execution of defect inspection is temporarily turned off and another function that allows skipping the execution of defect inspection is turned on until the utilization drops to the expected threshold pre-settled by users. If the utilization of defect detection drops too low, the function to force the execution of defect inspection is turned on to bring back the utilization level up to the threshold. Results indicate 10% enhancement in tool utilization compared to the previous static sampling plan.

Sullivan *et al.* [92] present an adaptive sampling technique for overlay measurements. The technique is based on a sampling capability ratio (CsK) analogous to the traditional CpK index[2]. The difference between the process capability (CpK) and the proposed CsK is in the selection of historical data. CpK is the process performance whereas CsK only considers data from lots that would have been available

[2]The CpK index is the process capability index. CpK takes into account both accuracy (centering) and precision (dispersion) and helps to determine the cause of failures and the need for changes in the product design, tooling, or the manufacturing process. The larger the CpK value, the greater the indication that the process is consistently under control (is within specification limits) [75].

for skipping through metrology. A sampling/skipping plan is implemented based on the results of the CsK. Results indicate significant cost savings. However, authors do not give percentage enhancement. Mouli *et al.* [58] present an Adaptive Metrology Sampling (AMS) based on a risk score evaluation. The concept consists in weighting each lot and wafer within a lot to make metrology sampling decisions and processing sequence (or priority) on metrology tools. The score varies between 0 and 1 and it is calculated based on Advanced Process Control (APC) and Statistical Process Control (SPC) analysis and observations. Results indicate a reduction of 30% of excursions without increasing tool capacity or sampling rates. Nduhura Munga *et al.* [60] present an adaptive sampling strategy based on the real time computation of the material at risk. In order to optimize the computational time, a Permanent Index per Context (IPC) is developed to reduce risk computation by simple subtractions or additions. Results indicate a risk reduction of more than 30% of material at risk compared to the previous static sampling strategy.

Concerning the technical aspects of proposed solutions, some papers are only rule-based while others are mathematical based. Table 4.4 summarizes the different mathematical techniques or approaches browsed in this review.

	Algorithms or Mathematical Techniques
Babikian and Engelhard [5]	Skip-Lot algorithm (CpK)
Nurani and Shantikumar [71]	Explicit Search algorithm
Lee *et al.* [49]	Self-Organizing Feature (SOFM) network
Lee [47]	Artificial Neural Network (ANN)
Sullivan *et al.* [92]	Skip-lot algorithm
Mouli *et al.* [58]	Risk-Score evaluation algorithm
Chen *et al.* [16]	Integer Linear Programming
Sahnoun *et al.* [83]	Skip-Lot algorithm (risk reduction)
Sahnoun *et al.* [82]	Skip-Lot algorithm (risk reduction)
Good *et al.* [28]	Sampling Compensation Algorithm (SCA)
Nduhura Munga *et al.* [60]	Permanent Index per Context (IPC)

Table 4.4: Mathematical techniques or approaches for adaptive sampling

An important point to note is that, among all papers with a mathematical technique or an algorithm, only three indicate industrial deployments [92] [58] [60]. Most works only use simulations to validate models [41] [95] but very few are industrial-

ized.

Through papers surveyed for adaptive sampling strategies, the specificities of semiconductor plants can be highlighted once again. Most of the sampling techniques browsed in this review are different. This is because of the specificity of each factory: Lot or wafer management, data storage, production tool management or qualifications, IT infrastructure, expert knowledge, company culture, etc. **Therefore, the efficiency of a sampling technique varies depending on its application** [12] [69].

Compared to static sampling strategies, adaptive sampling strategies offer two main advantages which lead to an increase in yield. The first advantage is the quick response to process variation by an increase of the number of lots to inspect when the process is likely to be out-of-control. The second advantage is a better use of metrology capacity through the reduction of the number of lots to inspect when the risk reduction is not significant or when the process is really under control. However, some drawbacks can be pointed out regarding the management of resources, the complexity of algorithms, and the industrial deployment. By modifying the number of lots to sample (increasing or reducing this number depending on the process state), the workload in metrology is no longer the same throughout production. The complexity of algorithms is such that the validation is most of the time performed through simulation and algorithms are never industrialized. To tackle the problems faced by adaptive sampling strategies, dynamic or smart sampling strategies have been introduced.

4.5 Dynamic Sampling

Dynamic sampling consists in selecting in real time the best lot or wafer to measure depending on the production state, metrology capacity, and the factory dynamics. The main difference with adaptive sampling is that no rule is defined at the start of production and the decision to sample or not a lot is taken when the lot can be selected for metrology. The metrology workload remains balanced contrary to adaptive sampling. The objective is to measure the lot that brings as much as possible information on both the risk reduction and the process variation. In high-mix semiconductor plants, where more than 200 products can be run concurrently, dynamic sampling techniques are seen as more suitable. Table 4.5 presents a survey on dynamic sampling in semiconductor manufacturing.

	Year	Mathematical techniques	Rule-based	Industrial deployment	Simulation	Comparison with other techniques
Purdy *et al.* [79]	2005	*		*		
Lensing and Stirton [50]	2007		*	*		
Holfeld *et al.* [32]	2007		*	*		
Good and Purdy [29]	2007	*		*	*	
Purdy *et al.* [78]	2007	*		*		
Kaga *et al.* [38]	2008		*	*		*
Jansen *et al.* [35]	2008		*	*		*
Hyung [34]	2008	*				*
Sun *et al.* [93]	2008	*				
Lin *et al.* [51]	2010		*			
Dauzère-pérès *et al.* [20]	2010	*			*	

Table 4.5: Survey on dynamic sampling

The first research works have been published in 2005 and a pioneer is this domain is M. A. Purdy who has authored or co-authored most of the papers found in the

literature. His works include industrial deployments [79] [32] [29] [78] and a patent can be found in [80]. Compared to adaptive sampling, dynamic sampling is mainly mathematically-based because of the levels of decision. Industrial deployments in semiconductor plants have been achieved thanks to the computing power that has strongly increased.

Among papers that indicate industrial deployments, Purdy *et al.* [79] present a Dynamic Sampling System (DSS) that combines a number of separate sampling rules into a single sampling decision. The first step consists in removing all sampling rates, i.e. making all lots measurable. For that, some defect inspection operations are defined so that all lots can enter the metrology queue. The next step consists in selecting lots to introduce in the metrology queue and lots to skip depending on the metrology capacity and on the information brought by each lot. The selection of lots to introduce in the metrology is performed based on an algorithm that analyses all rules (for example metal etchers at 30%, plasma etch at 10%, and a given product at 25%) and tries to ensure that each rule is satisfied with the minimum overall sampling rate when there are overlapping rules. The Last-In-First-Out (LIFO) principle is also used to ensure that the lots most recently added to the queue will be measured first. The aim is to get the greatest probability that the measurement of the current lot will allow for one or more other lots to be removed from the queue. Results indicate that the DSS has been rapidly adopted within the AMD company and only a small percentage of lots that entered the metrology queue were removed.

Lensing and Stirton [50], Holfeld *et al.* [32], and Purdy *et al.* [78] present and discuss the fab-wide APC sampling deployed within an AMD fab. This APC sampling system is based on the algorithm introduced by Good and Purdy [29]. The algorithm aims at selecting the best wafers to measure given a sampling rule set that can be infeasible, by assigning a penalty to each rule that is violated. This penalty is chosen such that it is larger for critical rules. The problem is written as a Mixed-Integer Linear Program (MILP) and the best wafers to measure correspond to the set that minimizes the sum of penalties. Results indicate rapid deployments within the fab and increased product yields. However, authors do not give compar-

isons with the previous system and percentage enhancement in terms of risk or cycle time reduction.

Kaga *et al.* [38] and Jansen *et al.* [35] discuss the use of design information to dynamically improve sampling for defect review. Lin *et al.* [51] discuss the benefit of developing a dynamic and intelligent sampling system in semiconductor manufacturing. Based on their experience, they point out three main benefits of a dynamic sampling system: Sampling stability, satisfactory coverage of in-line products, and comprehensive inclusion of process tools. Hyung [34] presents a model that combines the cost of sampling with the performance of control in terms of yield and cycle time. Tests are performed on different areas such as CVD (Chemical Vapor Deposition), PVD (Physical Vapor Deposition), and Photo-Lithography. Results show that the performance of dynamic sampling depends on the characteristics of the process. When the process is very stable, dynamic sampling has no effects, whereas it is effective when data set have large step disturbances.

Sun *et al.* [93] present a scoring algorithm based on weighted objectives to determine the optimal wafer sampling for maximum coverage. The algorithm is a multi-stage approach. The first stage consists in setting up various numbers of wafer samples and various numbers of equipment units. The aim is to ensure that all possible, but not redundant, combinations of wafers are captured. The second stage consists in using the scoring algorithm to evaluate and determine the preferred wafer sample based on pre-defined objectives and weighting factors. The score is calculated by multiplying individual normalized scores by associative weighted factors and summarizing them. The last stage uses the second stage results and designs a set of algorithms based on the number of experimental design group. This set of algorithms is used to select wafers in each group. No industrial assessment is mentioned.

Dauzère-pérès *et al.* [20] present a sampling, scheduling, and skipping algorithm to minimize risk dynamically. The algorithm is based on a Global Sampling Indicator (GSI) that gives a weight to each lot arriving at the measurement step, i.e. in front of metrology. This weight is computed based on the lot history and on two key

parameters, called Warning Limit (WL) and Inhibit Limit (IL). The WL indicates when the situation starts to become critical, and the IL corresponds to the maximum risk that can be tolerated for each production tool regarding the metrology capacity and production state. An Integer Linear Programming is provided in [63], and helps to compute the values of WL and IL depending on the production state. The sampling, scheduling, and skipping algorithm has been embedded in a prototype and simulated with actual data from STMicrolectronics. Results indicate a risk reduction of more than 70% compared to Fab sampling without any additional metrology capacity.

Table 4.6 summarizes the main mathematical techniques, approaches or algorithms developed for dynamic sampling.

	Algorithms or Mathematical Techniques
Good and Purdy [29]	Mix Integer Linear Programming (MILP)
Sun *et al.* [93]	Risk Scoring Algorithm
Dauzère-pérès *et al.* [20]	Global Sampling Indicator (GSI) algorithm

Table 4.6: Mathematical techniques or approaches for dynamic sampling

It is clear that dynamic sampling techniques are the most suitable techniques for modern and high-mix semiconductor fabs because of their ability to consider real-time information. **Our research is thus focused on developing and implementing dynamic sampling techniques which are one of the best ways to master the added value of each control, avoid redundant controls, and optimize the use of metrology capacity**. Between sampling at the lot level and wafer level, we focus on sampling at the lot level because the lot-to-lot variation is much higher than the wafer-to-wafer variation [94], and the processes of uploading and downloading a lot into a production tool spend more time than inspecting a wafer does [52].

The challenge is now in finding how a complex but efficient sampling algorithm can be industrialized in a high-mix semiconductor plant.

4.6 Conclusion

In this chapter, we surveyed the literature on sampling techniques in semiconductor manufacturing. We defined three main groups: Start or static, adaptive, and dynamic sampling techniques. Adaptive and dynamic sampling are more suitable for modern and high-mix semiconductor fabs. Our research is thus focused on modeling and implementing dynamic sampling control plans.

In the next chapter, we analyze the impact of variability and factory dynamics on the efficiency of a sampling plan, and introduce a fab-wide indicator that will support the implementation of dynamic and smart sampling approaches (Chapter 6 and Chapter 7).

Chapter 5

Analyzing and Optimizing Control Plans

This chapter[1] *analyzes the impact of a static control plan in a high-mix environment, highlights its main drawbacks, and introduces a fab-wide indicator called* IPC[2] *to support the industrial implementation of dynamic control plans. This IPC allows a very large amount of data to be managed, and several types of risk*[3] *indicators to be computed in real time. The simplicity and efficiency of the IPC led to its industrialization in the entire production.*

5.1 *Introduction*
5.2 *Factory Dynamics and Variability*
5.3 *Permanent Index per Context (IPC)*
5.4 *Real-Time Risk Assessment: CMP-WAR*
5.5 *Excursion Management*
5.6 *Conclusion*

[1]Part of this chapter has been submitted for publication to the **International Journal of Production Research** [64].

[2]**IPC**: ***I**ndice **P**ermanent par **C**ontexte.* In English: Permanent Index per Context.

[3]In this chapter, the risk is defined as the number of wafers processed on a production tool between two control operations. It corresponds to a potential loss in the case a problem occurs.

5.1 Introduction

What is the efficiency of an algorithm, approach, or technique if it cannot be generalized or industrialized? In the previous chapter we saw that the complexity in modern semiconductor plants is such that dynamic control plans are more suitable because of the variability throughout production. In this chapter, we first analyze the impact of variability and factory dynamics on the efficiency of control plans and then propose a global indicator (IPC) that can support the industrial implementation of dynamic control plans. The main goals of the IPC is to have a solution that can be generalized to several types of risks and simplify computations. We aim at developing solutions that can be supported by any IT infrastructure and generalized to other types of fabs, especially to the other sites of STMicroelectronics (Rousset, Italy, etc.).

Section 5.2 discusses the impact of variability in production. In section 5.3, we introduce and describe the IPC. Section 5.4 presents an industrial application where the risk i.e. material at risk is computed in real time using the IPC. In section 5.5, we present the way the IPC can be used to optimize the management of excursions.

5.2 Factory Dynamics and Variability

A high-mix semiconductor environment is characterized by several types of changes occurring in production. The qualifications as well as the availability of production tools vary depending on the production state, the number of products to be manufactured is never constant and, most of the time, changes come from intentional operational changes such as the lengthening of process flows or the addition of engineering lots. All these changes impact the static control plan that does not consider factory dynamics.

In order to concretely understand the impacts of all of these changes on a static control plan, we made some observations in the Chemical Metal Polishing (CMP) area regarding the defectivity control plan. We stated our working hypothesis on

added value of controls i.e. *"A control without a real added value is a waste of time and money"*. Observations showed us that the static sampling plan designed by the defectivity engineering team at the start of the production was completely affected (*"destroyed"*) by the factory dynamics and variability.

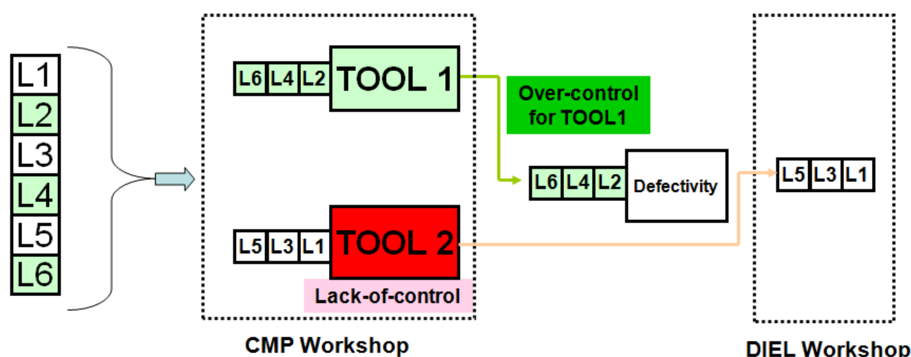


Figure 5.1: Drawbacks of static sampling.

Figure 5.1 illustrates the main drawbacks of a static sampling plan for defectivity controls. Six lots (**L1**, **L2**, **L3**, **L4**, **L5**, and **L6**) are coming into the CMP workshop to be processed. The control plan, designed by the defectivity engineering team at the start of the production is to control one lot every two lots. In this case, lots **L2**, **L4**, and **L6** are *flagged* for a defectivity control after the CMP processing operation. A control on a lot is called to reduce the risk on the production tool regarding the number of wafers processed on the tool since the latest control performed. In other words, if a defectivity control performed on a lot is validated, i.e. no critical defects are detected, the risk (i.e. the material at risk) is released on the wafers of all the lots processed in the same production tool since the latest control performed. In the case described in Figure 5.1, an optimal control plan would be to process at least one lot *flagged* on each production tool. But, because of the variability, the availability of production tools, and the complexity of process flows in high-mix semiconductor plants, situation such as the one described in Figure 5.1 is frequent. **TOOL 1** processes all *flagged* lots i.e. **L2**, **L4**, **L6** whereas **TOOL 2** does not process any. This results in over-control for production tool **TOOL 1** and

lack-of-control for production tool **TOOL 2**. As quality control is defined at product level, without taking into account tool information, the information is biased to monitor tool drifts. A simple solution could be to impose a limit set for each tool. This means to control **L3**, or **L1** and **L5** for example. The problem of this solution is that **control capacity is limited**. A better solution would be to release the control on **L4** and control **L3**. This implies to *flag* **L3** for a defectivity control operation at the next step. The question here is: *What if potential defects generated by the current processing operation cannot be captured at the next control operation for L3?* This situation is frequent since a lot, depending on its technology, will not be measurable at all (defectivity) control steps.

Another point is the variability of the delay or travel time between processing and measurement steps (Figure 5.2). Production tools are most of the time qualified to process more than one processing operations of different products [36]. The processing time varies depending on the processing operation to be performed and, for some products or technologies, additional processing steps are required in other workshops before a control in the defectivity workshop as illustrated in Figure 5.2. Consequences are such that production tools are no longer monitored by the "static (or start) sampling plan" and the situation becomes critical for all of production tools (**TOOL1** and **TOOL2**).

These situations (Figure 5.1 and Figure 5.2) led designers and defectivity engineers to put in place additional controls often redundant whereas a good repartition of different lots on production tools could allow efficient sampling rates. This results in increasing the number of controls and, most of the time, without real added value.

To overcome this problem, the risk on different production tools should be known in real time and took into account when dispatching lots on production tools. The problem lies in the large amount of data to manage. The complexity of process flows and the huge number of parameters to consider often lead many dynamic sampling algorithms to be too difficult to implement in practice. This also explains the particularity of the sampling techniques analyzed in the previous chapter.

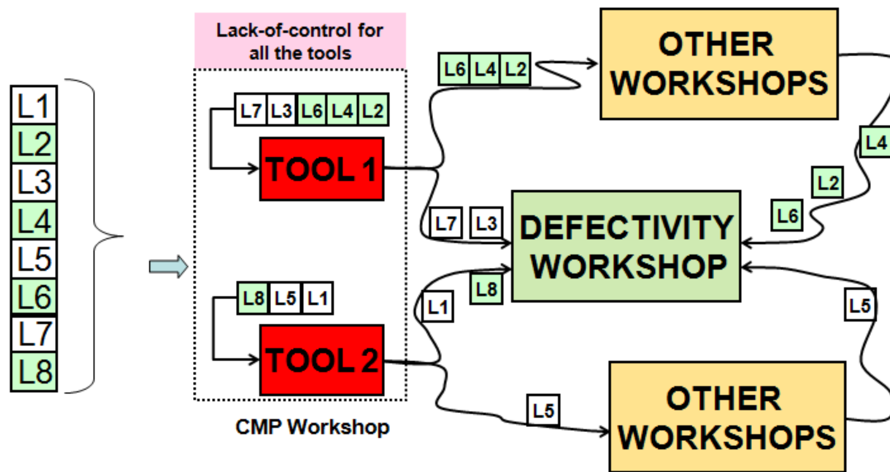


Figure 5.2: Impact of delay between process and measurement steps.

To answer this question, we developed an indicator called Permanent Index per Context (IPC) to allow very quick and fast computations of risk indicators. The risk depends on the context as described in the next section. This indicator is based on industrialization constraints and on the KISS[4] principle. We aimed at keeping things as simple as possible, minimizing the computing times of risk indicators, and thus minimizing the resource utilization.

[4]Keep It Simple Stupid. This principle states that simplicity is a key goal in design and that unnecessary complexity must be avoided.

5.3 Permanent Index per Context (IPC)

The Permanent Index per Context (IPC)[5] is a counter which is increased each time a context is verified. The context can be a tool, a chamber, a recipe, a technology, a resin, the combination of an operation and a technology, etc. This counter is never reset except when a special event occurs (Preventive Maintenance, intermediary qualification, etc.). The IPC has been introduced to allow both very quick and easy computations for any given context. In our first implementation, the context has been defined at the tool level to control the risk on production tools. Therefore, the risk is evaluated as the number of wafers processed on a production tool since the latest control performed. This is called *Wafer at risk*. To each lot l and tool m is associated an IPC, which is equal to 0 if l is not processed on m. Let M be the number of production tools, and $NW(l)$ be the number of wafers contained in lot l. The goal is to update in real time the following parameters:

- $LLM(m)$: Index of the **L**ast **L**ot that has been **M**easured for the production tool m.
- IPC_l^m: IPC of lot l for production tool m.
- RI_m: **R**isk **I**ndicator on production tool m.
- NI_l^m: **N**umber of wafers potentially **I**mpacted on tool m if lot l was measured.
- NI_l: **N**umber of wafers potentially impacted in the entire production if lot l was measured.

When lot l is processed on production tool m, an IPC is associated to l. The IPC of the lot l is equal to the IPC of the lot l' processed just before l on m plus the number of wafers in l ($NW(l)$):

$$IPC_l^m = IPC_{l'}^m + NW(l) \tag{5.1}$$

[5]Part of this section has been communicated to the **13th Scientific and Technical Meeting of ARCSIS** [59].

The risk indicator (i.e. material at risk) on production tool m is then given by:

$$RI_m = IPC_l^m - IPC_{LLM(m)}^m \tag{5.2}$$

The use of the IPC simplifies the computations of the risk indicators since these computations are reduced to calculating differences between two integer values. This implies very low resources usage, the possibility to manage a very large amount of data and quickly compute risk indicators for all of production tools. Instead of computing each time the risk indicators with complex algorithms using historical data, we assign to each lot an index (IPC of the lot) when the context is verified.

Figure 5.3 represents a sequence of different lots processed on a production tool m. Lots $L1$, $L2$, ..., $L9$ are processed on tool m.

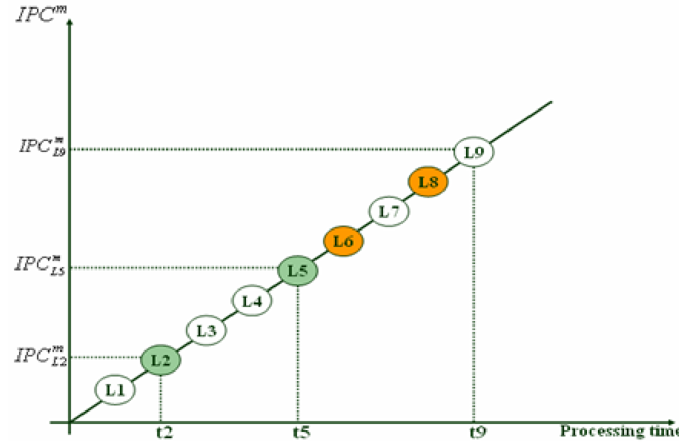


Figure 5.3: IPC mechanism.

$L2$ and $L5$ were validated by a defectivity control and, in this case, $L5$ corresponds to the last lot that has been measured $LLM(m)$. According to (5.1) and (5.2), the risk indicator on tool m at time t_9 is given by:

$$RI_m = IPC_{L9}^m - IPC_{L5}^m$$

Where:

$$IPC_{L9}^{m} > IPC_{L5}^{m}$$

It is also possible to quickly identify the best lot l to validate at the metrology step. This lot l is chosen such that its IPC verifies the following property:

$$IPC_{l}^{m} = Max\{0, \{IPC_{ll}^{m} \setminus IPC_{ll}^{m} > IPC_{LLM}^{m}, ll \in LM\}\} \tag{5.3}$$

Where LM is the set of lots waiting at the metrology step.

In Figure 5.3, lots $L6$ and $L8$ are processed on tool m and are waiting at the metrology step. According to (5.3), the best lot to select for m will be $L8$ since $IPC_{L8}^{m} > IPC_{L6}^{m}$ and $IPC_{L8}^{m} > IPC_{L5}^{m}$.

A control is defined as a measurement plus an action[7]. It is then crucial to be able to evaluate in real time the number of lots potentially impacted whenever a problem occurs on a lot l. This number can be determined for a given production tool m (NI_{l}^{m}) and for the entire production (NI_{l}):

$$NI_{l}^{m} = max\{0, IPC_{l}^{m} - IPC_{LLM}^{m}\} \tag{5.4}$$

And

$$NI_{l} = \sum_{m} NI_{l}^{m} \tag{5.5}$$

In Figure 5.3, at time $t9$, NI_{l}^{m} will be given by $(IPC_{L9}^{m} - IPC_{L5}^{m})$ corresponding to the sum of wafers in $L6$, $L7$, $L8$, and $L9$.

This IPC mechanism has been embedded in a prototype and first deployed for the CMP workshop before being industrialized in the entire fab. The next section describes the implementation of the IPC mechanism for real-time risk assessment within the CMP workshop.

5.4 Real-Time Risk Assessment: CMP-WAR

CMP-WAR is the name of an internal project that took place within the 300mm site of STMicroelectronics. The project has been initiated based on both the thesis and the European project IMPROVE. The main goal of the project was to master the risk level in production and ensure that the maximum risk (at the tool level) expected by the company would not be exceeded. For that, a *solution* had to be proposed to avoid cases of under-control as well as cases of over-control[6].

Two challenges had to be faced: Industrialization constraints and simplicity of solutions. Concerning industrialization constraints, the solution to be provided should be real time based and generalizable to other types of risks (chamber, recipe, resin, etc.). Concerning the simplicity of the solution, information should be gathered and presented in a way such that it could be easily understood by everybody.

These two challenges offered us a good opportunity to test, assess, and validate the IPC mechanism introduced in the previous section. We thus embedded the IPC mechanism in a prototype that we deployed for the CMP workshop. Figure 5.4 gives an overview of the CMP-WAR prototype that has been developed and deployed in-line during the CMP-WAR project. The prototype has been implemented in Visual Basic for Application (Excel-VBA). It shows, for each production tool, the real-time risk level value. The description of the prototype aims at illustrating the significant amount of data that had to be handled, showing the added value and efficiency of the IPC mechanism.

For the simplicity of use, three levels of alert associated to three different colors were defined:

- **Green**: The maximum risk level allowed by the company is not reached and the situation is under control for the production tool.

[6]Part of this section was presented at the **IEEE/SEMI Advanced Semiconductor Manufacturing Conference**. The presentation has been awarded with the Best Student Paper Award [60].

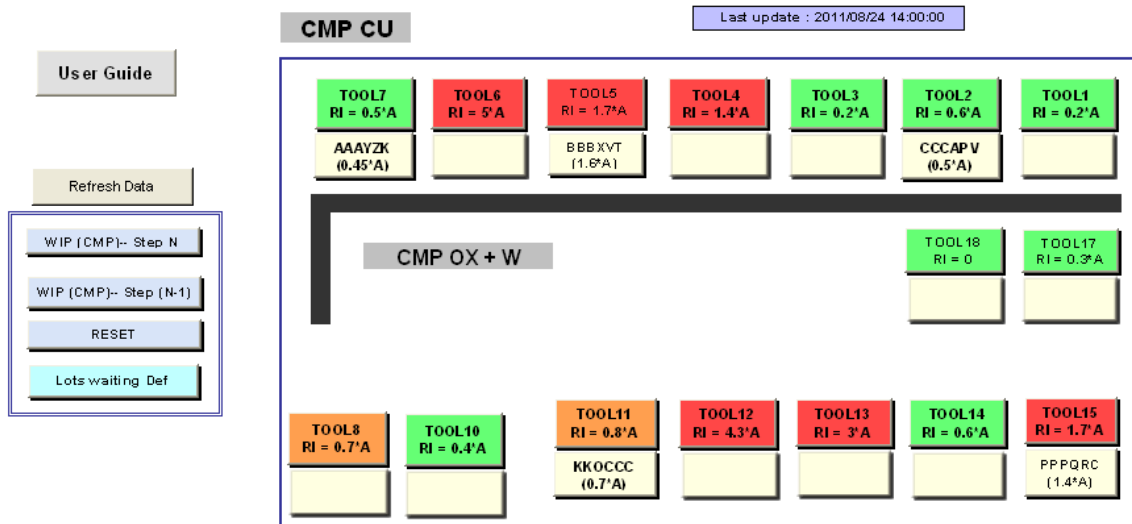


Figure 5.4: Overview of the CMP-WAR prototype.

- **Orange**: The risk level is very close to the maximum level specified by the company, and actions must be taken to reduce the risk.

- **Red**: The maximum risk level allowed by the company is reached and the tool must be stopped or actions immediately taken.

RI represents the **R**isk **I**ndicator on the production tool. It corresponds to the number of wafers processed on the production tool since the latest control performed for this tool. It is also called **Wafer at Risk (WAR)**.

To each production tool, two boxes are associated (Figure 5.5). The first box gives the value of the risk indicator and the second box gives the best lot to validate or control in the next defectivity step. This information is computed based on the IPC mechanism (Section 5.3). The maximum RI tolerated by the company is denoted **A**:

- If $RI > A$, the production tool is in **red**.

- If $(0.6 * A) < RI \leqslant A$, the production tool is in **orange**.

- If $RI \leqslant (0.6 * A)$, the production tool is in **green**.

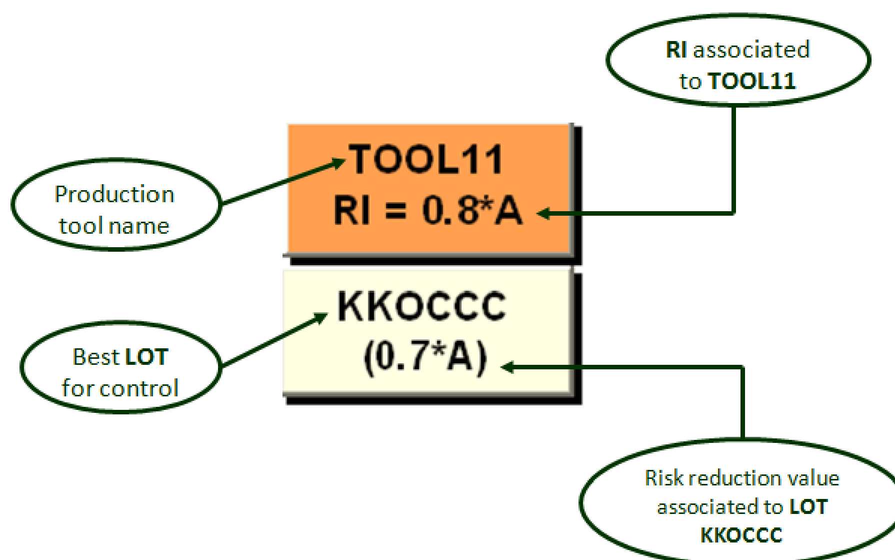


Figure 5.5: Risk Indicator and best lot for control.

RI is defined at the tool level. This means that *the context in the IPC is the production tool*. Therefore, each time a lot l is processed on a production tool m, an IPC_l^m is attached to the lot l. This IPC is equal to the IPC of the lot l' processed just before l on m plus the number of wafers of l ($NW(l)$) according to (5.1). Let us consider the example in Table 5.1. The context is the production tool **TOOL5**. Each time a lot (**L1**, **L2**, **L3**, **L4**, or **L5**) is processed on **TOOL5**, an IPC is associated to the lot for the considered context (**TOOL5**). This value is *never* reset. If, for example, **L4** is processed on another tool (e.g. **TOOL3**), another IPC (IPC_{L4}^{TOOL3}) will be attached to **L4**.

Once computed and assigned to each lot, the IPC information is then used to compute and update RI by performing simple differences between integer values. According to (5.2), RI is given by the difference between the IPC of the latest lot l processed on production tool m and the latest lot LLM validated by a defectivity control after being processed on m.

Let us consider the example in Table 5.2 that illustrates the way RI is computed:

Process date	Context	Lot ID	Number of wafers (NW)	IPC
T1	TOOL5	L1	12	$IPC_{L1}^{TOOL5} = 12$
T2	TOOL5	L2	23	$IPC_{L2}^{TOOL5} = 12 + 23 = 35$
T3	TOOL5	L3	15	$IPC_{L3}^{TOOL5} = 35 + 15 = 50$
T4	TOOL5	L4	15	$IPC_{L4}^{TOOL5} = 50 + 15 = 65$
T5	TOOL5	L5	25	$IPC_{L5}^{TOOL5} = 65 + 25 = 90$

Table 5.1: IPC computation and mechanism

- At time $T1$, $RI = IPC_{L1}^{TOOL5} = 12$.
- At time $T2$, $RI = IPC_{L2}^{TOOL5} = 35$.
- At time $T3$, $RI = IPC_{L3}^{TOOL5} = 50$.
- At time $T4$, $IPC_{L4}^{TOOL5} = 65$ and lot $L2$ is validated at the defectivity step, hence $LLM(TOOL5) = L2$. Therefore, RI is computed following the formula described in (5.2), i.e. $RI = IPC_{L4}^{TOOL5} - IPC_{L2}^{TOOL5} = 65 - 35 = 30$.
- At time $T5$, $RI = IPC_{L5}^{TOOL5} - IPC_{L2}^{TOOL5} = 90 - 35 = 55$.
- At time $T6$, $LLM(TOOL5)$ becomes $L5 \Rightarrow RI = IPC_{L6}^{TOOL5} - IPC_{L5}^{TOOL5} = 114 - 90 = 24$.
- At time $T7$, $RI = IPC_{L7}^{TOOL5} - IPC_{L5}^{TOOL5} = 136 - 90 = 46$.

RI can be computed for each context. In the CMP-WAR prototype, the context was defined at the tool level. Therefore, RI was computed for each production tool.

Process date	Context	Lot ID	IPC	LLM	RI
T1	TOOL5	L1	$IPC_{L1}^{TOOL5} = 12$		$RI = 12$
T2	TOOL5	L2	$IPC_{L2}^{TOOL5} = 35$		$RI = 35$
T3	TOOL5	L3	$IPC_{L3}^{TOOL5} = 50$		$RI = 50$
T4	TOOL5	L4	$IPC_{L4}^{TOOL5} = 65$	**L2**	$RI = 65 - 35 = 30$
T5	TOOL5	L5	$IPC_{L5}^{TOOL5} = 90$		$RI = 90 - 35 = 55$
T6	TOOL5	L6	$IPC_{L6}^{TOOL5} = 114$	**L5**	$RI = 114 - 90 = 24$
T7	TOOL5	L7	$IPC_{L7}^{TOOL5} = 136$		$RI = 136 - 90 = 46$

Table 5.2: RI computations

A defectivity control on a lot may validate more than one production tool depending on the lot history[7]. Let us consider the example of Figure 5.6, where a lot $L1$ is processed on three different production tools (from three different workshops: CMP, PHOTO, and ETCH) before being validated by a defectivity control.

The control performed on lot $L1$ gives information on the three production tools on which $L1$ was processed. If the control is **OK**, i.e. no critical defects are detected on $L1$, then RI can be reduced on the three production tools on which $L1$

[7] This is called **"depth of control"**. See Section 3.2.3 for further details.

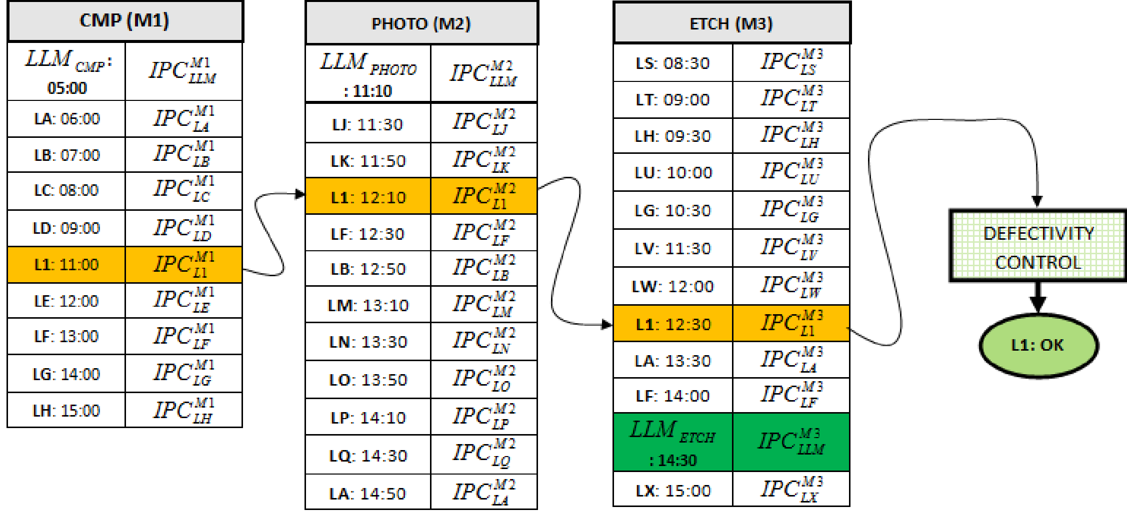


Figure 5.6: Depth of control.

was processed (**CMP(M1)**, **PHOTO(M2)**, and **ETCH(M3)**). The number of production tools that can be validated by a lot at a defectivity step depends on both the lot history and the defectivity matrices (Chapter 3). Depending on the production tool, we may have situation such that a lot l' processed after $L1$ arrives in front of a defectivity control before $L1$. This is the case for production tool **ETCH(M3)** (LLM_{ETCH}). Lot LLM_{ETCH} was processed at 14:30, after L1 (processed at 12:30), but validated before $L1$. In such a situation, $L1$ validated by the defectivity step does not bring additional information for the production tool **ETCH(M3)**. In Figure 5.6, at 15:00, the new RI will be given by:

- CMP(M1): $RI_{CMP(M1)} = IPC^{M1}_{LH} - IPC^{M1}_{L1}$.

- PHOTO(M2): $RI_{PHOTO(M2)} = IPC^{M2}_{LA} - IPC^{M2}_{L1}$.

- ETCH(M3): $RI_{ETCH(M3)} = IPC^{M3}_{LX} - IPC^{M3}_{LLM_{ETCH}}$.

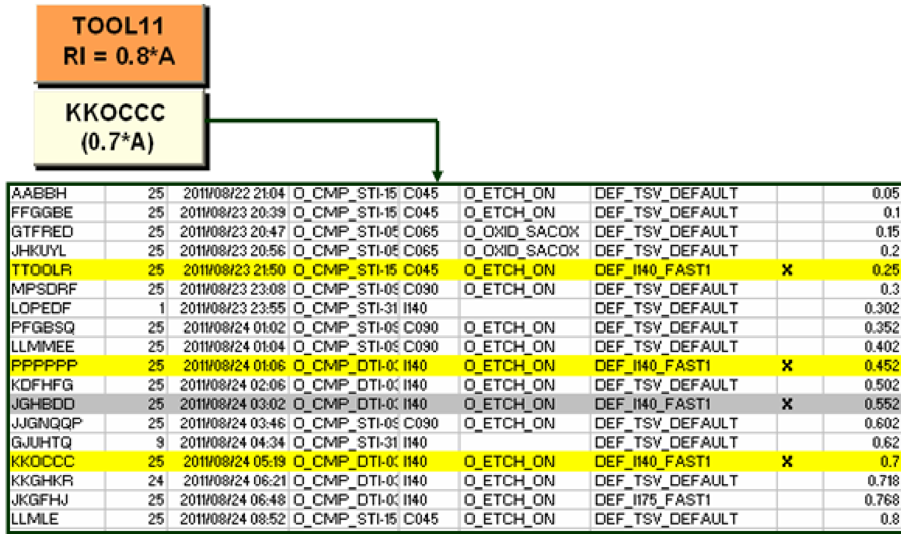
With the IPC information and using the equation (5.3), it is possible to directly and easily access the information giving the best lot to control at defectivity steps. This helps to avoid performing a control on a lot that does not bring any added value. *In the CMP-WAR prototype, the depth of control has not been considered*

when displaying the best lot to control. The lot displayed in Figure 5.5 gives the best lot to validate considering tools separately.

For the sake of simplicity, by clicking on the box giving the best lot to validate (Figure 5.7), the operator can directly access to the list of lots processed on the production tool since the last control. This production tool history explains why lot **KKOCCC** displayed is the best lot to control. In Figure 5.7, three lots are highlighted in yellow and one lot in gray. For these four lots, there is a **"X"** in the eighth column. This means that these four lots were flagged at the start of production by the defectivity engineering team for a control after the CMP workshop. Among these four lots, one has already been skipped (the one in gray), most probably because of capacity needs. The three lots in yellow (**TTOOLR**, **PPPPPP**, **KKOCCC**) were flagged for defectivity control and are not yet skipped. Among these three lots in yellow, **KKOCCC** is the most recent lot processed on production tool **TOOL11**. It implies that measuring **KKOCCC** *will* allow skipping two lots (**TTOOLR** and **PPPPPP**) processed just before and not yet skipped. This is why **KKOCCC** is displayed as the best lot to control. This information is computed in real time based on the IPC associated to each lot and stored in a database. The displayed information was adjusted to be easily understandable by operators.

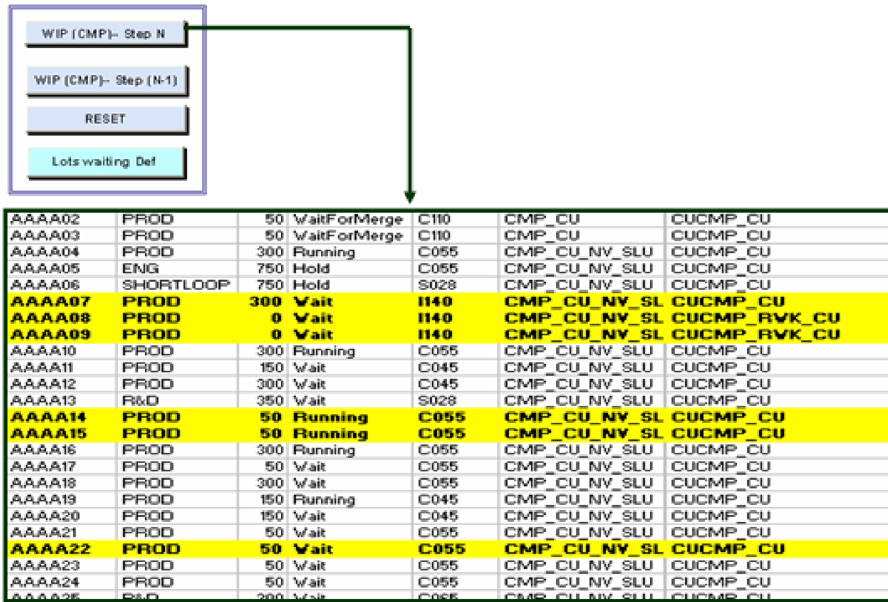
When a tool is in an orange status and there is no lot associated below the tool (see **TOOL8** in Figure 5.4), it means that, among all lots processed on the tool, there are no lots waiting at the defectivity step or there are no "flagged" lots that can be validated at the next defectivity step. As the situation starts to become critical (orange status), information is provided on lots waiting to be processed in the CMP area (WIP - Step N). If such situation arises, the operator is called to check (by a simple click) on **"WIP(CMP)–Step N"** (Figure 5.8) to identify a lot that is *flagged* for the next defectivity control step and direct this lot to the tool which is in "orange status". Processing such a lot on a tool in "orange status" will allow *RI* to be reduced for the tool once the lot is validated at the defectivity step.

In the list of lots provided in Figure 5.8, lots in yellow are *flagged* for a Defectivity control after the CMP area. However, this list concerns all production tools in CMP. Knowing that all lots cannot be processed on all production tools and for the sake



AABBH	25	2011/08/22 21:04	O_CMP_STI-15	C045	O_ETCH_ON	DEF_TSV_DEFAULT		0.05
FFGGBE	25	2011/08/23 20:39	O_CMP_STI-15	C045	O_ETCH_ON	DEF_TSV_DEFAULT		0.1
GTFRED	25	2011/08/23 20:47	O_CMP_STI-05	C065	O_OXID_SACOX	DEF_TSV_DEFAULT		0.15
JHKUYL	25	2011/08/23 20:56	O_CMP_STI-05	C065	O_OXID_SACOX	DEF_TSV_DEFAULT		0.2
TTOOLR	25	2011/08/23 21:50	O_CMP_STI-15	C045	O_ETCH_ON	DEF_I140_FAST1	X	0.25
MPSDRF	25	2011/08/23 23:08	O_CMP_STI-09	C090	O_ETCH_ON	DEF_TSV_DEFAULT		0.3
LOPEDF	1	2011/08/23 23:55	O_CMP_STI-31	I140		DEF_TSV_DEFAULT		0.302
PFGBSQ	25	2011/08/24 01:02	O_CMP_STI-09	C090	O_ETCH_ON	DEF_TSV_DEFAULT		0.352
LLMMEE	25	2011/08/24 01:04	O_CMP_STI-09	C090	O_ETCH_ON	DEF_TSV_DEFAULT		0.402
PPPPPP	25	2011/08/24 01:06	O_CMP_DTI-0	I140	O_ETCH_ON	DEF_I140_FAST1	X	0.452
KDFHFG	25	2011/08/24 02:06	O_CMP_DTI-0	I140	O_ETCH_ON	DEF_TSV_DEFAULT		0.502
JGHBDD	25	2011/08/24 03:02	O_CMP_DTI-0	I140	O_ETCH_ON	DEF_I140_FAST1	X	0.552
JJGNQQP	25	2011/08/24 03:46	O_CMP_STI-09	C090	O_ETCH_ON	DEF_TSV_DEFAULT		0.602
GJUHTQ	9	2011/08/24 04:34	O_CMP_STI-31	I140		DEF_TSV_DEFAULT		0.62
KKOCCC	25	2011/08/24 05:19	O_CMP_DTI-0	I140	O_ETCH_ON	DEF_I140_FAST1	X	0.7
KKGHKR	24	2011/08/24 06:21	O_CMP_DTI-0	I140	O_ETCH_ON	DEF_TSV_DEFAULT		0.718
JKGFHJ	25	2011/08/24 06:48	O_CMP_DTI-0	I140	O_ETCH_ON	DEF_I175_FAST1		0.768
LLMLE	25	2011/08/24 08:52	O_CMP_STI-15	C045	O_ETCH_ON	DEF_TSV_DEFAULT		0.8

Figure 5.7: Production tool history.

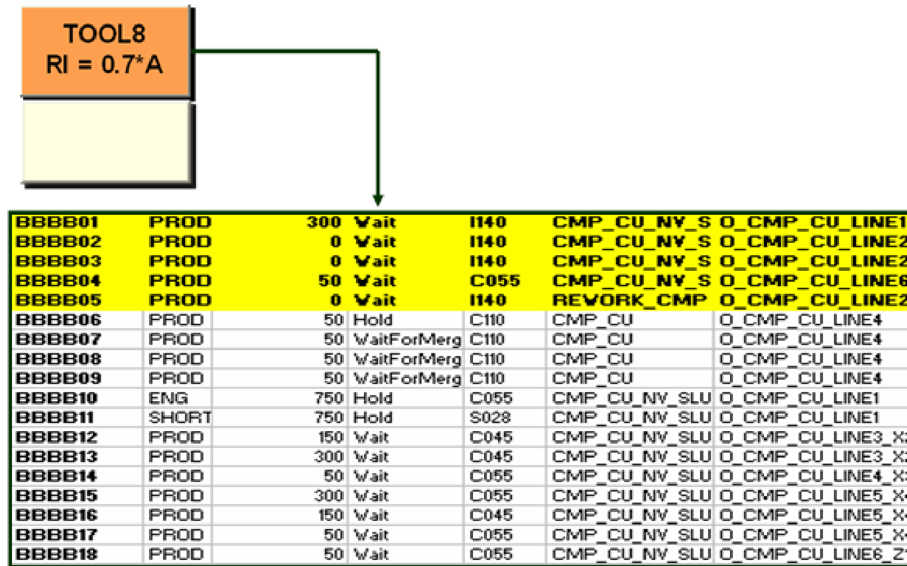


AAAA02	PROD	50	WaitForMerge	C110	CMP_CU	CUCMP_CU
AAAA03	PROD	50	WaitForMerge	C110	CMP_CU	CUCMP_CU
AAAA04	PROD	300	Running	C055	CMP_CU_NV_SLU	CUCMP_CU
AAAA05	ENG	750	Hold	C055	CMP_CU_NV_SLU	CUCMP_CU
AAAA06	SHORTLOOP	750	Hold	S028	CMP_CU_NV_SLU	CUCMP_CU
AAAA07	**PROD**	**300**	**Wait**	**I140**	**CMP_CU_NV_SL**	**CUCMP_CU**
AAAA08	**PROD**	**0**	**Wait**	**I140**	**CMP_CU_NV_SL**	**CUCMP_RVK_CU**
AAAA09	**PROD**	**0**	**Wait**	**I140**	**CMP_CU_NV_SL**	**CUCMP_RVK_CU**
AAAA10	PROD	300	Running	C055	CMP_CU_NV_SLU	CUCMP_CU
AAAA11	PROD	150	Wait	C045	CMP_CU_NV_SLU	CUCMP_CU
AAAA12	PROD	300	Wait	C045	CMP_CU_NV_SLU	CUCMP_CU
AAAA13	R&D	350	Wait	S028	CMP_CU_NV_SLU	CUCMP_CU
AAAA14	**PROD**	**50**	**Running**	**C055**	**CMP_CU_NV_SL**	**CUCMP_CU**
AAAA15	**PROD**	**50**	**Running**	**C055**	**CMP_CU_NV_SL**	**CUCMP_CU**
AAAA16	PROD	300	Running	C055	CMP_CU_NV_SLU	CUCMP_CU
AAAA17	PROD	50	Wait	C055	CMP_CU_NV_SLU	CUCMP_CU
AAAA18	PROD	300	Wait	C055	CMP_CU_NV_SLU	CUCMP_CU
AAAA19	PROD	150	Running	C045	CMP_CU_NV_SLU	CUCMP_CU
AAAA20	PROD	150	Wait	C045	CMP_CU_NV_SLU	CUCMP_CU
AAAA21	PROD	50	Wait	C055	CMP_CU_NV_SLU	CUCMP_CU
AAAA22	**PROD**	**50**	**Wait**	**C055**	**CMP_CU_NV_SL**	**CUCMP_CU**
AAAA23	PROD	50	Wait	C055	CMP_CU_NV_SLU	CUCMP_CU
AAAA24	PROD	50	Wait	C055	CMP_CU_NV_SLU	CUCMP_CU

Figure 5.8: Lots waiting in front of the CMP area.

of simplicity, we separate this list of lots by usage, tools, and *processability*. The operator can therefore access to the list of lots waiting to be processed (by a specified

group of tools) by directly clicking on the box displaying the name of the tool. The list is sorted such that lots that are *flagged* appear first in the list (in yellow), followed by other lots (Figure 5.9).



BBBB01	PROD	300	Wait	I140	CMP_CU_NV_S	O_CMP_CU_LINE1
BBBB02	PROD	0	Wait	I140	CMP_CU_NV_S	O_CMP_CU_LINE2
BBBB03	PROD	0	Wait	I140	CMP_CU_NV_S	O_CMP_CU_LINE2
BBBB04	PROD	50	Wait	C055	CMP_CU_NV_S	O_CMP_CU_LINE6
BBBB05	PROD	0	Wait	I140	REWORK_CMP	O_CMP_CU_LINE2
BBBB06	PROD	50	Hold	C110	CMP_CU	O_CMP_CU_LINE4
BBBB07	PROD	50	WaitForMerg	C110	CMP_CU	O_CMP_CU_LINE4
BBBB08	PROD	50	WaitForMerg	C110	CMP_CU	O_CMP_CU_LINE4
BBBB09	PROD	50	WaitForMerg	C110	CMP_CU	O_CMP_CU_LINE4
BBBB10	ENG	750	Hold	C055	CMP_CU_NV_SLU	O_CMP_CU_LINE1
BBBB11	SHORT	750	Hold	S028	CMP_CU_NV_SLU	O_CMP_CU_LINE1
BBBB12	PROD	150	Wait	C045	CMP_CU_NV_SLU	O_CMP_CU_LINE3_X
BBBB13	PROD	300	Wait	C045	CMP_CU_NV_SLU	O_CMP_CU_LINE3_X
BBBB14	PROD	50	Wait	C055	CMP_CU_NV_SLU	O_CMP_CU_LINE4_X
BBBB15	PROD	300	Wait	C055	CMP_CU_NV_SLU	O_CMP_CU_LINE5_X
BBBB16	PROD	150	Wait	C045	CMP_CU_NV_SLU	O_CMP_CU_LINE5_X
BBBB17	PROD	50	Wait	C055	CMP_CU_NV_SLU	O_CMP_CU_LINE5_X
BBBB18	PROD	50	Wait	C055	CMP_CU_NV_SLU	O_CMP_CU_LINE6_Z

Figure 5.9: Lots waiting in front of the tool.

In a high mix environment as in the 300mm site of STMicroelectronics in Crolles, the factory dynamics is such that we may have situations where a tool is in an orange status and there is no *flagged* lot waiting in front of the CMP area. This is why we provide the set of lots waiting to be processed or currently being processed in the area before the CMP: **"WIP(CMP)–Step (N-1)"**. If such situation arises, the operator is called to click on the box **"WIP(CMP)–Step (N-1)"** to identify a *flagged* lot for the tool in CMP that is in an orange status. This allows anticipating and accelerating some lots[8] in order to reduce RI on production tools.

If there are no *flagged* lots in WIP(N) and WIP(N-1), the information is directly sent to the defectivity engineers who take the decision to **force** a lot that is not

[8] Accelerating a lot consists in increasing the priority of lots on some processing steps.

flagged for a defectivity control to reduce RI. This **forced** lot is called DWR[9]. If the result of Defectivity control is validated, then the defectivity engineer resets RI using the box **"RESET"**[10]. An additional box **"Lots Waiting in DEF"** is also provided to show the RI reduction of each lot waiting in front of the defectivity.

Computing all of the information described above may require significant computing power if data are not *optimally organized*. This is why the **IPC mechanism** is *very efficient* since all computations are reduced to simple additions and subtractions between integer values. The mechanism was easily understood and the computing time strongly reduced compared to other algorithms previously implemented where computing RI required to manage the tool history (list of lots, number of wafers, processing time, technology, etc.) in real time.

Based on the simplicity and efficiency of the **IPC mechanism**, we decided to use it for excursion management, i.e. when a process or a production tool falls out of specifications. This is the case when a defectivity control on a lot is not validated, i.e. the result of the control is judged out of specifications. The source of the defect must be isolated as quickly as possible.

[9]Defect Work Request.

[10]Authentication is required to restrict the use of the prototype and to avoid an increase workload for defectivity engineers.

5.5 Excursion Management

An excursion happens in production when a process or tool falls outs of specification[11]. Since defectivity controls are performed on wafers, when the sum of defects on wafers exceeds a given threshold, an excursion occurs. The production tool generating the defects has to be identified and stopped as quickly as possible before too many lots are impacted. Let us consider the simplified example of Figure 5.10 where lot $L1$ is successively processed on three different tools (from three different workshops: CMP, PHOTO, and ETCH) before arriving in the defectivity workshop for a control operation.

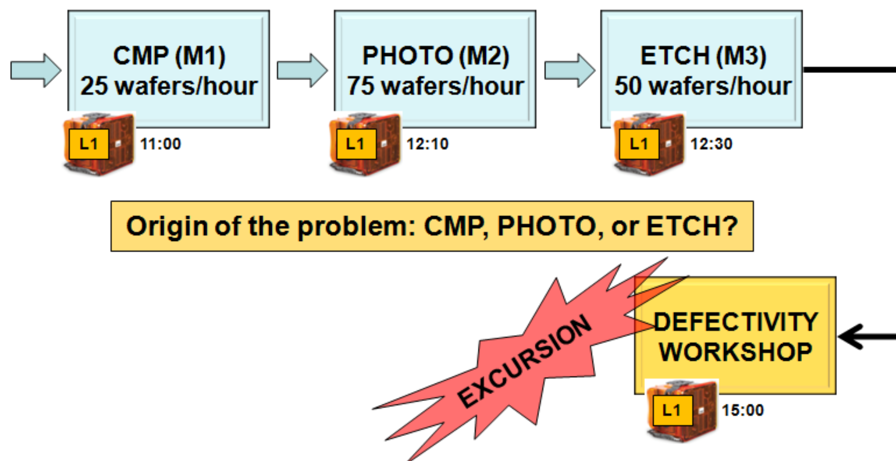


Figure 5.10: Example of an excursion management problem.

If $L1$ is judged out of specifications, the challenge is to:

1. Isolate the most probable source of excursion: CMP, PHOTO, or ETCH,
2. Select the best set of lots to measure in order to contain the excursion,
3. And quantify the number of lots potentially impacted.

[11]Part of this section was presented at the 7th International Conference on Modeling and Analysis of Semiconductor Manufacturing (included in the 2011 Winter Simulation Conference) [61].

Depending on the time elapsed between the excursion detection (on $L1$) and the source detection, the impact can be significant. In Figure 5.10, the throughput on the CMP production tool is 25 wafers/hour, 75 wafers/hour for the PHOTO production tool, and 50 wafers/hour for the ETCH tool. $L1$ is processed on the CMP tool at 11:00 and controlled in defectivity at 15:00. It means that 4 hours elapsed between the process of $L1$ on the CMP tool and its control in defectivity. This corresponds to the process of $25 * 4 = 100$ wafers on the CMP tool. In other words, related to the CMP tool, we have at least 100 wafers potentially impacted. If the defect source is isolated 10 hours later, there are 14 hours (4 + 10) between the process of $L1$ on the CMP tool and the control operation in defectivity. This implies that, instead of having 100 wafers potentially impacted, there are 25*14 = 350 wafers. If the process operation is non reversible, it results in 350 wafers impacted on the CMP tool. This is similar for all production tools on which $L1$ was processed before arriving in the defectivity workshop for a control operation.

One of the complexities in identifying the source of excursions in high mix semiconductor manufacturing lies in the large amount of data to handle as quickly as possible. Most of the time, engineers are used to navigate between different IT tools and use their experience to identify the most probable cause of an excursion. Once the source of defects is *identified*, the next step consists in determining and selecting a lot or a set of lots to measure in order to confirm or deny the source of the excursion. Depending on the current processing step of a lot, the recipe, the technology, the WIP, the lot history, the product, etc., a commonality analysis[12] is performed to identify the lot or set of lots most likely to confirm or deny the excursion source. The aim is to find a lot that has the same characteristics than the lot on which the excursion has occurred. This implies manipulating a significant amount of data leading to an overwhelming task for defectivity engineers.

To optimize the management of excursions and thus reduce the potential impact by quickly detecting the source of defects, **using the IPC is once again very**

[12] A commonality analysis consists in identifying different links that exist among lots. These links concern lot history, product, quantity, technology, mask, etc.

effective. Indeed, by simplifying the computations of several types of risk, a lot of information can quickly be gathered with little CPU effort. Let us consider another example based on Figure 5.10. We consider that an IPC is attached to each lot processed on production tools.

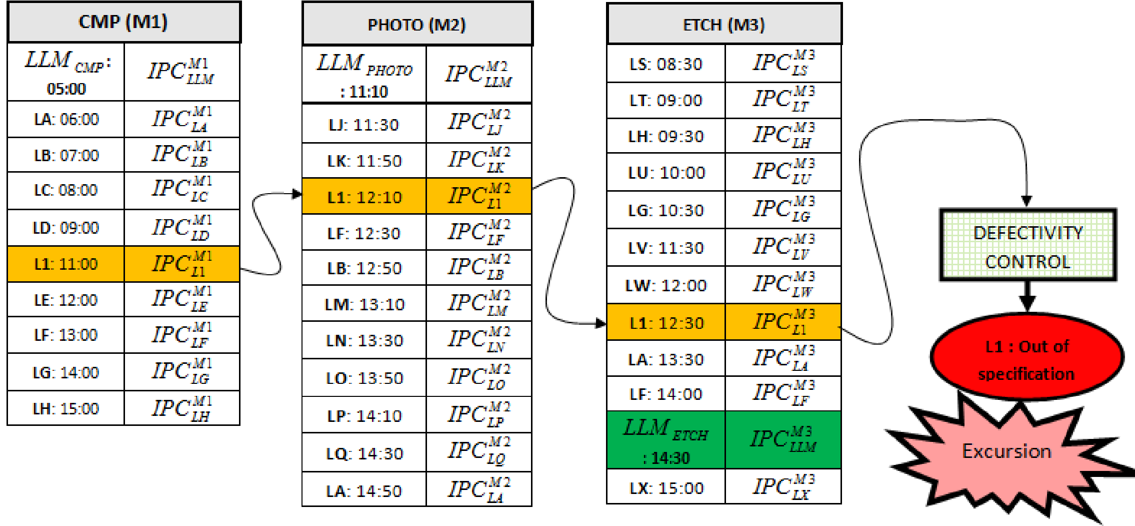


Figure 5.11: Example of excursion analysis.

Figure 5.11 illustrates how the scope of analysis can be reduced based on the IPC. $L1$ has been processed on three different production tools before being judged as *out of specifications* after a control operation in the defectivity workshop. Based on the history of tools and the information brought by the IPC, it is possible to quickly identify the set of tools that was validated by a control operation after the process of $L1$. This helps in removing this set of tools from the initial scope of analysis, and thus reducing the scope of analysis. For example, considering tool **ETCH(M3)**, a control was performed on lot LLM_{ETCH} processed after $L1$. As LLM_{ETCH} was judged within specification, **ETCH(M3)** can be removed from the initial set of analysis. To be more general, let us introduce the following notations:

- LE: Index of the **L**ot on which an **E**xcursion has occurred,
- M_{LE}: Set of **M**achines on which lot LE was processed,

According to (5.1), (5.2), and (5.3), we have:

$$MI_{LE} = \{m \in M_{LE} \setminus NI_{LE}^{m} > 0\} \tag{5.6}$$

$$LPI_{LE}^{m} = \{l \in \{1, \ldots, L\} \setminus IPC_{l}^{m} < IPC_{LE}^{m} \ and \ NI_{l}^{m} > 0\} \tag{5.7}$$

$$LPI_{LE} = \bigcup_{m=1}^{M_{LE}} LPI_{LE}^{m} \tag{5.8}$$

Where:

- MI_{LE} in (5.6) is the set of **M**achines to be considered in the analysis,
- LPI_{LE}^{m} in (5.7) is the set of **L**ots **P**otentially **I**mpacted regarding lot LE on production tool m,
- LPI_{LE} in (5.8) is the set of **L**ots **P**otentially **I**mpacted regarding lot LE in the entire production.

For the case illustrated in Figure 5.11, the set of lots potentially impacted is given by:
$LPI_{L1} = \{LPI_{L1}^{M1} \bigcup LPI_{L1}^{M2} \bigcup LPI_{L1}^{M3}\}$
$= \{LA, LB, LC, LD, LE, LF, LG, LH\} \bigcup \{LJ, LK, LF, LB, LM, LN, LO, LP, LQ, LA\} \bigcup \varnothing$
$= \{LA, LB, LC, LD, LE, LF, LG, LH, LJ, LK, LM, LN, LO, LP, LQ\}$

Tool **ETCH(M3)** is not included in the initial set of analysis since $NI_{L1}^{M3} < 0$. The number of lots to analyze is therefore reduced by reducing the set of tools to consider. **A prototype based on the IPC information was developed to help quickly identifying the set of tools most likely to be the source of the excursion.** Figure 5.12 gives an overview of the software prototype that was developed. More details on the prototype are provided in Chapter 8.

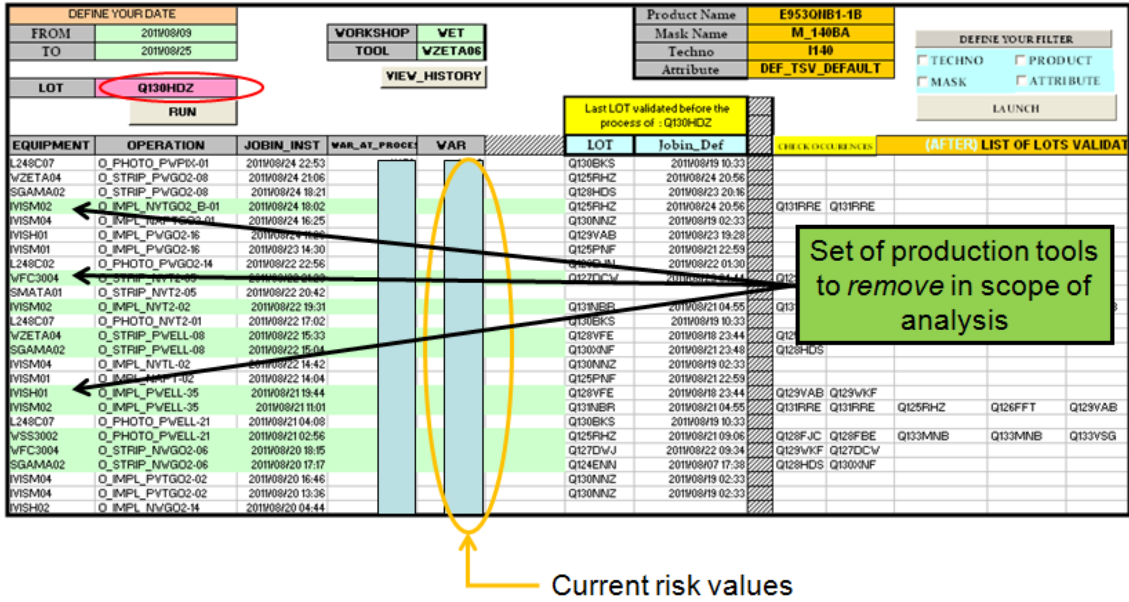


Figure 5.12: Overview of the Excursion Management prototype.

Other avenues can be explored regarding the best lot to prioritize on defectivity tools in order to contain the excursion. The lot has to be selected among the set of lots to consider in the scope of analysis as described above (see LPI_{L1})[13]:

1. The first approach could be to select lots based on the probability of a production tool to be most likely the source of the excursion. This probability is defined based on the type of defects detected on wafers in the lot. Let us consider the example introduced in Figure 5.11. If we focus on tool **CMP(M1)**, we may have the three following cases:

   a. If the probability for tool **CMP(M1)** to be the source of defects is **very high**, then the best lot to measure would be LA to confirm the excursion and then to stop all of the lots processed on the tool after LA i.e.:

   $$\text{Stop the set of lots } \{LB, LC, LD, L1, LE, LF, LG, LH\}.$$

   b. If the probability for tool **CMP(M1)** to be the source of defects is **very low**, then the best lot to measure would be LH to quickly validate all

[13]These avenues were not explored in depth in this thesis.

the lots processed before LH and thus exclude tool **CMP(M1)** from the scope of analysis.

c. If the probability for tool **CMP(M1)** to be the source of defects is **average**, then it would be interesting to choose a lot between LA and LH. This lot could be for example LE.

The selection of lots based on probabilities linked to production tools would be performed using the **IPC**.

2. The second approach would be to use the concept of dominating sets with the aim of identifying and selecting the lot that covers the maximum number of lots and production tools as illustrated in Figure 5.13. In this case, it would be lot LA.

The two perspectives described above will strongly contribute in improving the management of excursions in dynamic sampling. Indeed, using a static sampling plan, the same lots are measured throughout production. It implies that, when a problem occurs, it is possible to reduce the scope of analysis only based on the lot history. The added defects of each processing step can be quantified. However, in dynamic sampling, the selection of lots is based on the added value in term of control. A lot does no longer have to be inspected at all stages of inspection. Hence the challenge lies in identifying the best lot to inspect or measure in order to contain the excursion.

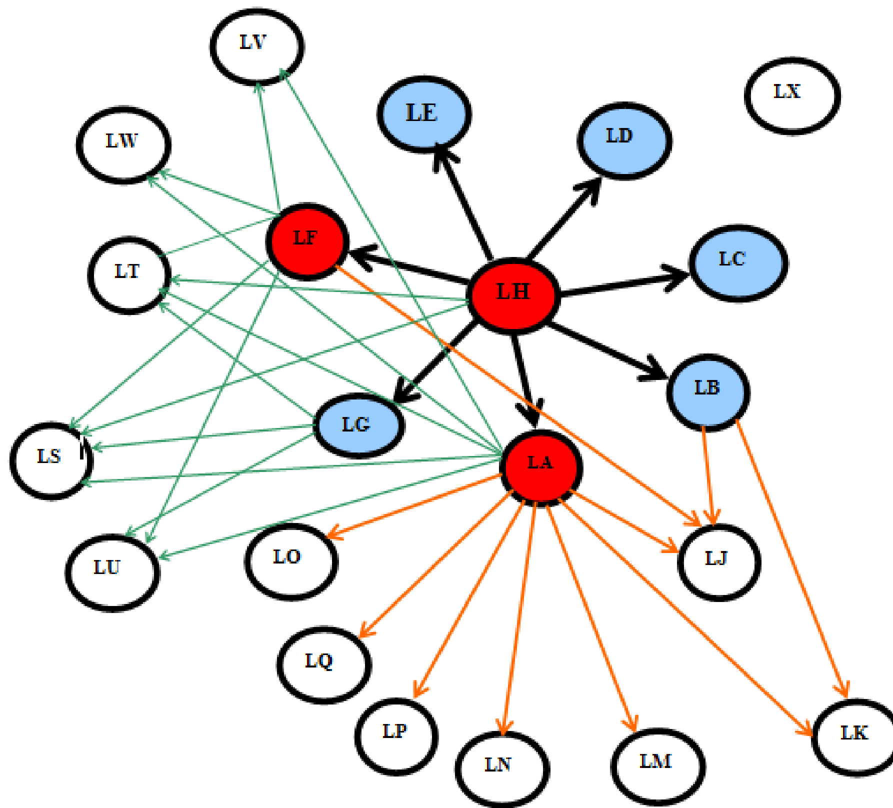


Figure 5.13: Concept of dominating sets.

5.6 Conclusion

In this chapter, we pointed out the drawbacks of static control plans, analyzed the impact of variability in a high-mix semiconductor plant, and introduced a fab-wide indicator that can support both the implementation of dynamic control plans, and the optimization of excursion management. This fab-wide indicator called **IPC** is based on industrial constraints and its efficiency is in the ability to compute in a very simple way several risk indicators with little CPU effort.

In the next chapter, we introduce dynamic sampling algorithms that have been developed within the framework of the European project IMPROVE, and industrialized using the IPC mechanism.

Chapter 6

Implementing Smart Sampling Policies

This chapter introduces the sampling algorithms that have been developed within the framework of the European project IMPROVE[1]*. The aim is to dynamically sample lots in front of metrology or inspection steps. By evaluating through simulations the different sampling algorithms, results indicate a risk reduction of more than 70% compared to Fab sampling. By defining financial metrics to assess the return on investment of such algorithms, potential gains are estimated to more than US$1,000,000.*

6.1 *Introduction*
6.2 *Smart sampling mechanism*
6.3 *Global Sampling Indicator (GSI)*
6.4 *GSI sampling algorithms*
6.5 *Numerical experiments*
6.6 *Conclusion*

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

6.1 Introduction

Premature optimization is the root of all evil [2]. In the previous chapter, we analyzed the particularities of a static sampling plan, identified its main drawbacks, and proposed an indicator (IPC) to support industrial implementation of dynamic control plans. In this chapter, we propose a smart sampling approach to optimize fab-wide sampling. It is based on a Global Sampling Indicator (GSI) that helps to dynamically identify the best set of lots to measure, skip, or prioritize on metrology tools.

The chapter is structured as follows. Section 6.2 describes the smart sampling approach. In section 6.3, we present the different GSI formulas. Section 6.4 introduces the GSI sampling algorithms that are based on GSI formulas and additional constraints linked to the production environment. Section 6.5 is devoted to numerical experiments. We analyze the efficiency of the GSI sampling algorithms versus fab sampling, and discuss the impact of GSI parameters in the case of STMicroelectronics in Crolles, France. Section 6.6 concludes the chapter.

6.2 Smart sampling mechanism

Smart sampling consists in dynamically selecting in an intelligent way the lots to inspect or measure[3]. Three types of decisions are performed: Sampling, skipping, and scheduling. These three decisions are taken based on control parameters and metrology capacity. The order of decisions is not necessarily sequential, i.e. sample, skip, and finally schedule. Some decisions can be taken simultaneously. The aim is to sample and measure lots in order to minimize some objectives based on the risks[4].

---

[2] **Donald Ervin Knuth**.

[3] In this chapter, inspecting a lot is equivalent to measuring a lot. The same for inspection steps and metrology steps.

[4] In this chapter, the risk is the number of wafers processed between two control operations. It is called *Wafer at risk* and denoted **WAR** in the sequel.

6.2.1 Sampling mechanism

Sampling a lot is based on how much is gained when adding the lot to the set of lots already waiting to be inspected. As illustrated in Figure 6.1, each time a lot **Lx** arrives in front of an inspection step (Defectivity workshop), it must be decided whether or not to include the lot in the set of lots already waiting for inspection. If the lot is selected and introduced in the queue, then the lot is sampled.

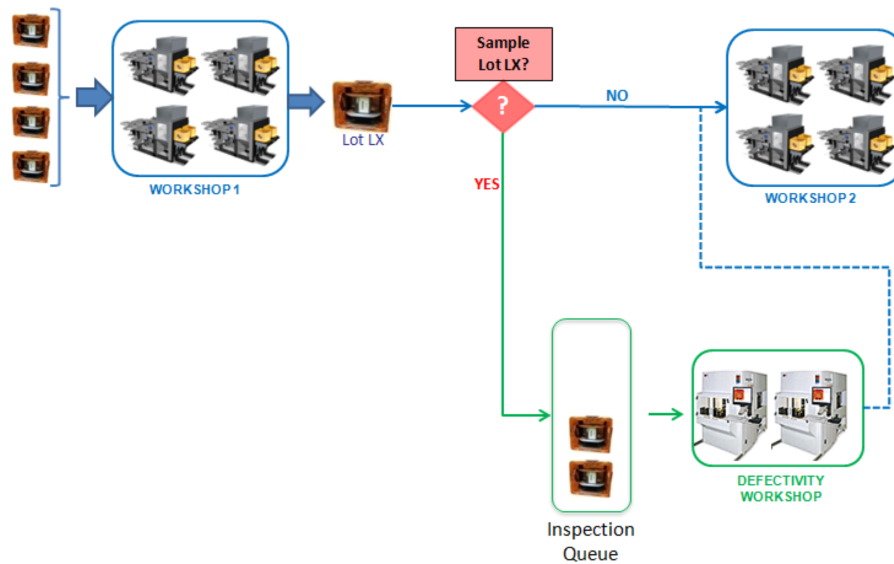


Figure 6.1: Sampling mechanism.

6.2.2 Skipping mechanism

Skipping a lot consists in avoiding inspecting a lot **Lx** that has been sampled (Figure 6.2). The lot is removed from the inspection queue and directed to the next process step (next workshop). This type of decisions may happen when: (1) The maximum size of the inspection queue is reached and there is a new lot that need to be sampled because of the significant gain, (2) when the inspection capacity is reduced because of the unavailability of an inspection tool, or (3) when the result of another inspection is within specifications.

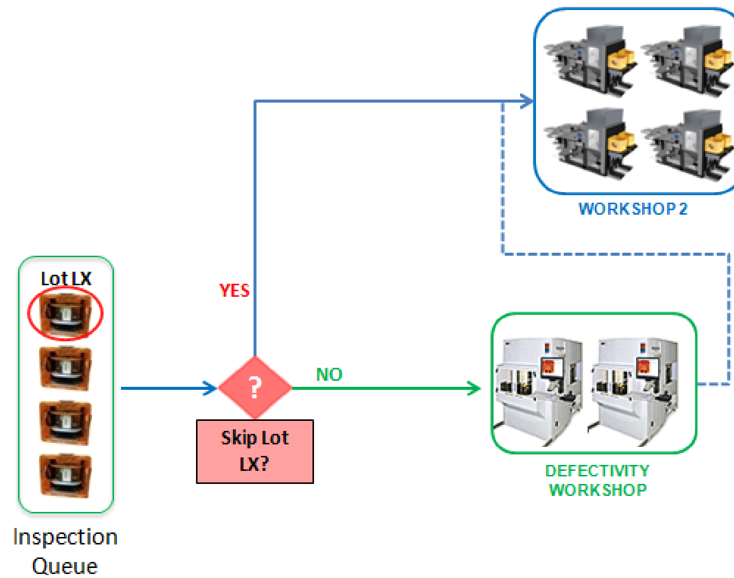


Figure 6.2: Skipping mechanism.

1. **Arrival of a new lot.** When a new lot arrives in front of an inspection step and there is no more places in the queue, it may be interesting to skip a lot already waiting in the queue and replace it by the new lot if it brings more information. This helps keeping the queue size smaller than a maximum value while inspecting the best possible lots.

2. **Unavailability of an inspection tool.** When an inspection tool is down or unavailable (preventive or corrective maintenance for example), the inspection capacity is reduced. The queue size has to be adjusted to avoid increasing the waiting time of lots for inspection. The more a lot waits for inspection, the more its cycle time increases. Therefore, if the inspection capacity is reduced, some lots should be skipped.

3. **Inspection of another lot.** Each time a lot is inspected, the situation in term of risks changes. Depending on the results of the inspection, the priority of lots may be modified or further analysis requested. Some lots may thus be prioritized and others skipped.

6.2.3 Scheduling mechanism

Scheduling lots consists in assigning sampled lots on inspection tools and sequencing them. This is performed each time an inspection tool becomes available. The objective is to prioritize lots having the largest gains.

To dynamically perform these three types of decisions, i.e. sampling, skipping, and scheduling, an indicator called **GSI** (**G**lobal **S**ampling **I**ndicator) has been developed [20] as described in the next section.

6.3 Global Sampling Indicator (GSI)

The GSI is an indicator that gives a score to different sets of lots. To each set of lots S is associated an expected level of risk on the entire production if the lots in S are selected for inspection. Let us consider examples in Table 6.1 and Table 6.2. Table 6.1 corresponds to the initial situation where no lot is selected. Table 6.2 shows two outcomes if two different sets of lots $S1$ and $S2$ are selected for inspection.

Production tools	Risk Level
M1	300
M2	250
M3	450
M4	450

Table 6.1: Initial situation.

Production tools	Risk Level
M1	**50**
M2	**10**
M3	450
M4	450

(a) Set of lots **S1** selected.

Production tools	Risk Level
M1	**200**
M2	**200**
M3	**200**
M4	**200**

(b) Set of lots **S2** selected.

Table 6.2: Example if sets of lots S1 or S2 are selected for inspection.

If the set of lots **S1** is selected and inspected, the resulting risk will be the one in Table 6.2a, i.e. the risk levels of production tools $M1$ and $M2$ are reduced. If the set **S2** is selected, the resulting risk will be the one in Table 6.2b, i.e. the risk levels of all production tools ($M1$, $M2$, $M3$, and $M4$) are reduced. In the first case, when **S1** is selected, the risk level is strongly reduced for production tools M1 (=50) and M2 (=10) whereas M3 and M4 keep a high level of risk (=450). In the second

case, when the set **S2** is selected, the risk level is reduced for all production tools. However, in this case, the risk levels are much higher than the risk levels of tools M1 and M2 compared to the first case where **S1** is selected. Hence the following question: *Is it better to select and inspect a set of lots that strongly reduces the risk level of one or two production tools, or to select and inspect a set of lots that reduces only a little the risk level of all production tools?* To answer this question, a GSI has been developed to give a weight or score to each set of lots S depending on control parameters and inspection capacity. The set of lots S can be empty (Table 6.1) or not (Table 6.2).

The GSI is computed for different sets of lots and not for each lot separately. These sets of lots correspond to different possible combinations of lots to be inspected. Let us consider the example in Figure 6.3. In the queue, there are 4 lots $\{L1, L2, L3, L4\}$ selected and waiting to be inspected. A lot LX arrives in front of the defectivity inspection and we need to decide whether the lot LX must be sampled or not.

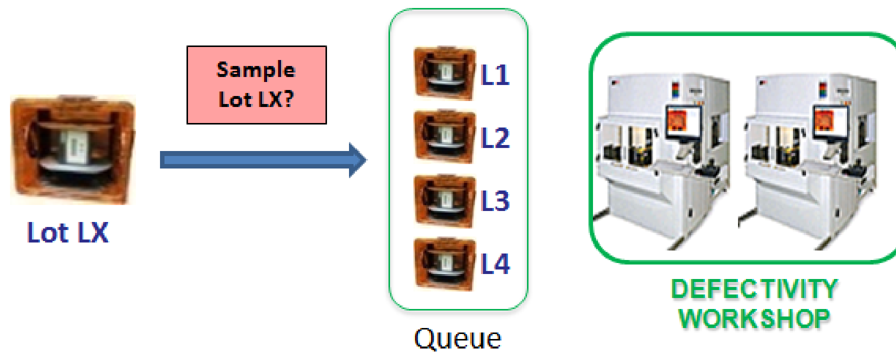


Figure 6.3: GSI combinations.

In the case illustrated in Figure 6.3, the queue is full. This means that sampling lot LX will lead to skipping another lot waiting in the queue. Let us denote by $S_{initial}$ the set of lots already waiting for a defectivity inspection, i.e. $S_{initial} = \{L1, L2, L3, L4\}$. To decide whether or not LX must be sampled, we need to analyze

five different combinations with the GSI:

1. $GSI(S_{initial}) = GSI(\{L1, L2, L3, L4\})$
2. $GSI(\{S_{initial} \setminus L1\} \cup \{LX\}) = GSI(\{LX, L2, L3, L4\})$
3. $GSI(\{S_{initial} \setminus L2\} \cup \{LX\}) = GSI(\{L1, LX, L3, L4\})$
4. $GSI(\{S_{initial} \setminus L3\} \cup \{LX\}) = GSI(\{L1, L2, LX, L4\})$
5. $GSI(\{S_{initial} \setminus L4\} \cup \{LX\}) = GSI(\{L1, L2, L3, LX\})$

The set with the smallest GSI is selected. For example, if the third combination gives the smallest GSI, then it is better to inspect the set of lots $\{L1, LX, L3, L4\}$, i.e. sampling LX and skipping $L2$. If the first combination gives the smallest GSI, then it is better to not sample LX, i.e. to inspect the set of lots $\{L1, L2, L3, L4\}$ already waiting in the queue.

The same type of combinations is computed for scheduling lots on inspection tools. However, our approach is different than the one used when sampling/skipping lots, in which the set of lots with the smallest GSI is selected. When scheduling lots on inspection tools, the priority of a lot L, denoted by **LSI(L)** (**L**ot **S**cheduling **I**ndicator), is defined by the difference between $GSI(S_{initial} \setminus \{L\})$ and $GSI(S_{initial})$. $LSI(L)$ is always positive since, by definition, $GSI(S_{initial} \setminus \{L\}) \geq GSI(S_{initial})$. The principle is to evaluate the impact of L in $S_{initial}$ by determining how much would be lost in terms of GSI if L was not measured. The larger $LSI(L)$, the greater the priority of lot L on inspection tools. Let us consider the example in Figure 6.3 with four lots $\{L1, L2, L3, L4\}$ waiting to be inspected. To define the priority of these lots on inspection tools, five combinations of lots are evaluated. These five combinations are obtained by successively removing one by one one lot from the initial set of lots $S_{initial}$, i.e.:

1. $GSI(S_{initial}) = GSI(\{L1, L2, L3, L4\})$
2. $LSI(L1) = GSI(S_{initial} \setminus \{L1\}) - GSI(S_{initial}) = GSI(\{L2, L3, L4\}) - GSI(\{L1, L2, L3, L4\})$

3. $LSI(L2) = GSI(S_{initial} \setminus \{L2\}) - GSI(S_{initial}) = GSI(\{L1, L3, L4\}) - GSI(\{L1, L2, L3, L4\})$
4. $LSI(L3) = GSI(S_{initial} \setminus \{L3\}) - GSI(S_{initial}) = GSI(\{L1, L2, L4\}) - GSI(\{L1, L2, L3, L4\})$
5. $LSI(L4) = GSI(S_{initial} \setminus \{L4\}) - GSI(S_{initial}) = GSI(\{L1, L2, L3\}) - GSI(\{L1, L2, L3, L4\})$

The larger $LSI(L)$ of L ($LSI(L1)$, $LSI(L2)$, $LSI(L3)$, or $LSI(L4)$), the greater the priority of L on inspection tools. This means that the larger the difference with the initial GSI ($GSI(S_{initial})$), the more not inspecting the lot degrades the GSI. In the example illustrated above, if $LSI(L1) > LSI(L2) > LSI(L3) > LSI(L4)$, the priority on inspection tools is: $L1$, $L2$, $L3$, and finally $L4$. If $LSI(L2) > LSI(L4) > LSI(L1) > LSI(L3)$, then the priority is $L2$, $L4$, $L1$, and finally $L3$.

To be more general, a risk array (same type for several tools and/or several risk types) is assigned to each lot [20]. This array contains the new value of each risk (or of the risk reduction) if the lot is inspected. Let us consider the following notations:

- R: Number of risks,
- WL_r: Warning Limit for risk r,
- IL_r: Inhibit Limit for risk r,
- RV_r: Current risk value for risk r,
- $G_{r,l}$: Gain on risk r if lot l is inspected,
- $NRV_{r,l}$: New risk value if lot l is inspected, i.e. $NRV_{r,l} = RV_r - G_{r,l}$.
- $NRV_r(S)$: New risk value if lots in set S are inspected. The new risk value if lots in set S are inspected is calculated as follows:

$$NRV_r(S) = Min_{l \in S} NRV_{r,l}.$$

For defectivity controls, the risk RV_r corresponds to the *Wafer at Risk* (WAR) for production tool r. The WAR is the number of wafers processed on tool r since

the process of the latest lot inspected in defectivity. This can be seen as the number of wafers which have been processed on a tool r since the latest good defectivity control. In this case, the gain $G_{r,l}$ is the WAR *reduction* of the tool r if lot l is inspected. Two control parameters are defined: *Warning Limit* and *Inhibit Limit*. The *Warning Limit* WL_r corresponds to the value of the WAR beyond which the situation starts to become critical in term of control. The *Inhibit Limit* IL_r is the maximum number of wafers that can be ran between two defectivity inspections for the considered tool. Exceeding this limit for the WAR may lead to stopping the production tool.

Using parameters described above, two GSI formulas (GSI-1 and GSI-2) computing two different scores have been proposed. These two GSI formulas are used in different **GSI algorithms** (Section 6.4) for sampling, skipping, and scheduling lots dynamically.

6.3.1 GSI 1

The first formula of the GSI aims at selecting sets that contain lots that help to reduce risk values that are closer to their *Inhibit Limits* IL_r. The *Inhibit Limit* IL_r represents the maximum risk value that the company tolerates to ensure that, when a problem occurs, the potential loss will not exceed this limit. The focus is thus put on the ratio NRV_r/IL_r and the goal is to increase the priority of lots with significant gain, i.e. lots for which NRV_r/IL_r is very small (parameter $1/\beta$), and decrease the priority of lots for which NRV_r/IL_r is close or higher than 1 (parameter α).

$$GSI(S) = \sum_{r=1}^{R} \left[\left(\frac{NRV_r(S)}{IL_r} \right)^{1/\beta} + \left(\frac{NRV_r(S)}{IL_r} \right)^{\alpha} \right]$$

The two parameters α ($\geqslant 1$) and β ($\geqslant 1$) in the GSI formula are thus used to put more or less emphasis on getting as far as possible from the *Inhibit Limit* for which the current risk value is closer. Figure 6.4 shows the evolution of the GSI depending on the ratio between the new risk value NRV_r if lots in the set S are inspected and the *Inhibit Limit* IL_r, with α set to 6 and β to 2. These two parameter values are

based on numerical experiments presented in Section 6.5. They ($\alpha = 6$ **and** $\beta = 2$) guarantee the expected trend of the curve (Figure 6.4). However, their values may vary depending on the factory dynamics.

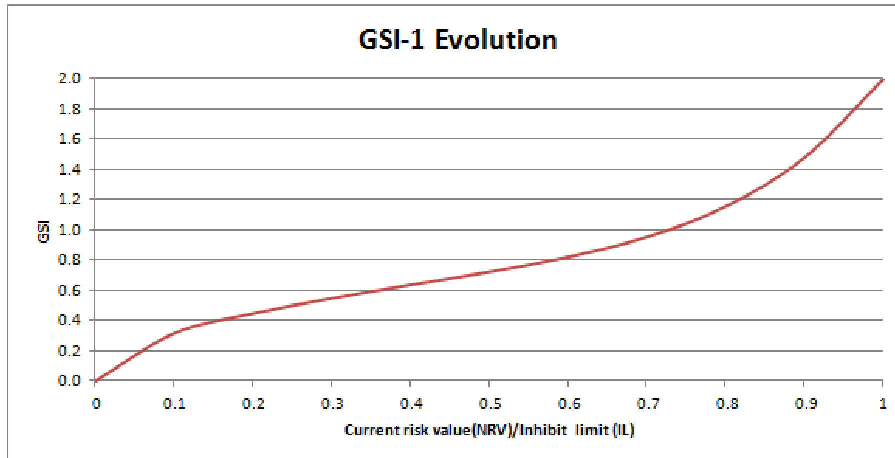


Figure 6.4: GSI-1 Evolution.

Between 0 and 0.5 (see X-axis), the shape of the curve is mostly driven by the parameter β ($1/\beta$). Beyond 0.5, the parameter α penalizes the fact that NRV_r is close to the *Inhibit Limit*. For example, if $NRV_r/IL_r = 0.1$, NRV_r is ten times smaller than IL_r. In other words, by selecting and inspecting a given set S of lots, the new risk value (NRV_r) is much smaller than the *Inhibit Limit*. Since the objective is to stay as much as possible below the *Inhibit Limit*, the set S of lots needs to be prioritized. This is why, in Figure 6.4, a smaller GSI (=0.3) is associated to the ratio $NRV_r/IL_r = 0.1$. Similarly, $NRV_r/IL_r = 0.9$ means that, when inspecting a set S of lots, NRV_r is very close to IL_r. This is not interesting, since it corresponds to a significant GSI (=1.5) that reduces the priority of selecting such a set of lots. In Figure 6.4, note that, when the ratio NRV_r/IL_r is very small (i.e. the new risk value is very far from the *Inhibit Limit*), the priority increases with a small GSI. When the ratio NRV_r/IL_r increases, the priority is decreased with a significant GSI. Let us consider the example in Table 6.3 already introduced in the previous section.

Production tools	NRV_r	IL_r
$r1$	50	500
$r2$	10	500
$r3$	450	500
$r4$	450	500

(a) Set of lots **S1** selected.

Production tools	NRV_r	IL_r
$r1$	200	500
$r2$	200	500
$r3$	200	500
$r4$	200	500

(b) Set of lots **S2** selected.

Table 6.3: Example1 - Evaluating two different set of lots **S1** and **S2** with the GSI.

We want to select a set of lots ($S1$ or $S2$) to inspect in order to reach the best possible state in production. The *Inhibit Limit* is set to 500 for all production tools. Inspecting the set of lots $S1$ will strongly reduce the risk level of two production tools ($r1$ and $r2$) while keeping the risk level of two other tools ($r3$ and $r4$) close to their *Inhibit Limits*. Inspecting the set of lots $S2$ will reduce the risk level of all production tools while ensuring to stay well below the *Inhibit Limit*. Without any computation, we can see that it would interesting to select and inspect the lots in set $S2$ and keep all risk levels far from the *Inhibit Limit*. This decision can be verified with the GSI formula using $\alpha = 6$ and $\beta = 2$:

- Selecting $S1$:

$$GSI(S1) = \left[\left(\frac{50}{500}\right)^{1/2} + \left(\frac{50}{500}\right)^{6}\right] + \left[\left(\frac{10}{500}\right)^{1/2} + \left(\frac{10}{500}\right)^{6}\right] + \left[\left(\frac{450}{500}\right)^{1/2} + \left(\frac{450}{500}\right)^{6}\right] + \left[\left(\frac{450}{500}\right)^{1/2} + \left(\frac{450}{500}\right)^{6}\right] = 3.42$$

- Selecting $S2$:

$$GSI(S2) = \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] + \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] + \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] + \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] = \mathbf{2.55}$$

The GSI associated to $S2$ is smaller than the one associated to $S1$. Selecting and inspecting the lots in set $S2$ ensures the best possible resulting state in terms

of risk. Let us consider another example (Table 6.4) where inspecting two different sets of lots $S3$ and $S4$ allows staying below the *Inhibit Limit* but we need to select the best set.

Production tools	NRV_r	IL_r
$r1$	350	500
$r2$	300	500
$r3$	400	500
$r4$	300	500

(a) Set of lots **S3** selected.

Production tools	NRV_r	IL_r
$r1$	350	500
$r2$	350	500
$r3$	350	500
$r4$	350	500

(b) Set of lots **S4** selected.

Table 6.4: Example2 - Evaluating two different sets of lots **S3** and **S4** using the GSI.

The resulting states (Table 6.4a and Table 6.4b) are very similar and it is not easy to identify the best set. Using the GSI formula, we have:

- Selecting $S3$ gives:

$$GSI(S3) = \left[\left(\frac{350}{500}\right)^{1/2} + \left(\frac{350}{500}\right)^{6}\right] + \left[\left(\frac{300}{500}\right)^{1/2} + \left(\frac{400}{500}\right)^{6}\right] + \left[\left(\frac{400}{500}\right)^{1/2} + \left(\frac{400}{500}\right)^{6}\right] + \left[\left(\frac{300}{500}\right)^{1/2} + \left(\frac{300}{500}\right)^{6}\right] = \mathbf{3.75}$$

- Selecting $S4$ gives:

$$GSI(S4) = \left[\left(\frac{350}{500}\right)^{1/2} + \left(\frac{350}{500}\right)^{6}\right] + \left[\left(\frac{350}{500}\right)^{1/2} + \left(\frac{350}{500}\right)^{6}\right] + \left[\left(\frac{350}{500}\right)^{1/2} + \left(\frac{350}{500}\right)^{6}\right] + \left[\left(\frac{350}{500}\right)^{1/2} + \left(\frac{350}{500}\right)^{6}\right] = 3.81$$

Therefore, it is better to select and inspect the lots in $S3$ since it has the smallest GSI. In order to understand why $S3$ provides the smallest GSI, let us separately analyze the impact of parameters β and α using two different cases.

6.3.1.1 Impact of parameter β

Parameter β plays an important role when we need to analyze what happens when NRV_r is well below IL_r. Let us consider the example in Table 6.5.

Production tools	NRV_r	IL_r
$r1$	200	500
$r2$	200	500
$r3$	200	500
$r4$	200	500

(a) Set of lots **S5** selected.

Production tools	NRV_r	IL_r
$r1$	10	500
$r2$	280	500
$r3$	280	500
$r4$	280	500

(b) Set of lots **S6** selected.

Table 6.5: Example3 - Evaluating two different sets of lots **S5** and **S6** using the GSI.

There are two different sets of lots $S5$ and $S6$ where the maximum value of NRV_r for each set is well below IL_r $(= 500)$. In this case, it could be interesting to compare the total gain that each set brings, i.e.:

- $r_1(S5) - r_1(S6) = 200 - 10 = +190$
- $r_2(S5) - r_2(S6) = 200 - 280 = -80$
- $r_3(S5) - r_3(S6) = 200 - 280 = -80$
- $r_4(S5) - r_4(S6) = 200 - 280 = -80$

This means that selecting and inspecting the lots in set $S5$ will bring a gain of $+190 - 80 - 80 - 80 = -50$ compared to the lots in set $S6$. In other words, inspecting the lots in $S5$ will reduce the total risk by **50** more than inspecting the lots in $S6$. Therefore, the best choice would be to select $S5$ because of the gain in risk reduction. However, as the objective is to stay well below the *Inhibit Limit*, β (in the GSI formula) will prioritize the set of lots where the risk level of one production tool can be strongly reduced since, in both cases, we stay below and far from the *Inhibit Limit*. This can be verified using the GSI formula:

- Selecting $S5$ gives:

$$GSI(S5) = \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] + \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] + \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] + \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] = 2.55$$

- Selecting $S6$ gives:

$$GSI(S6) = \left[\left(\frac{10}{500}\right)^{1/2} + \left(\frac{10}{500}\right)^{6}\right] + \left[\left(\frac{280}{500}\right)^{1/2} + \left(\frac{280}{500}\right)^{6}\right] + \left[\left(\frac{280}{500}\right)^{1/2} + \left(\frac{280}{500}\right)^{6}\right] + \left[\left(\frac{280}{500}\right)^{1/2} + \left(\frac{280}{500}\right)^{6}\right] = \mathbf{2.48}$$

$S6$ is prioritized because of the significant risk reduction on production tool r_1.

6.3.1.2 Impact of parameter α

Contrary to β, α penalizes cases where NRV_r is close to IL_r values. Let us consider the example in Table 6.6.

Production tools	NRV_r	IL_r
$r1$	200	500
$r2$	200	500
$r3$	450	500
$r4$	200	500

(a) Set of lots **S7** selected.

Production tools	NRV_r	IL_r
$r1$	300	500
$r2$	300	500
$r3$	300	500
$r4$	300	500

(b) Set of lots **S8** selected.

Table 6.6: Example4 - Evaluating two different sets of lots **S7** and **S8** using the GSI.

As for parameter β, we can compare the total gain of selecting lots in $S7$ versus selecting lots in $S8$:

- $r_1(S7) - r_1(S8) = 200 - 300 = -100$

- $r_2(S7) - r_2(S8) = 200 - 300 = -100$
- $r_3(S7) - r_3(S8) = 450 - 300 = +150$
- $r_4(S7) - r_4(S8) = 200 - 300 = -100$

We obtain: $-100 - 100 + 150 - 100 = -150$. This means that selecting and inspecting the lots in set $S7$ will help reducing the global risk of -150 more than by selecting and inspecting the lots set $S8$. In this case, the best decision is to select and inspect the lots in set $S7$. However, when looking to the NRV_r values if $S7$ is selected (Table 6.6a), we can see that, for r_3, NRV_{r_3} (= 450) is *very* close to IL_{r_3} (=500). Using the GSI formula, this is strongly penalized with α (NRV_r very close to IL_r). Therefore, the GSI formula indicates that $S8$ is the best set of lots to inspect even if $S7$ brings the largest gain in term of risk reduction:

- Selecting $S7$ gives:

$$GSI(S7) = \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] + \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] + \left[\left(\frac{450}{500}\right)^{1/2} + \left(\frac{450}{500}\right)^{6}\right] + \left[\left(\frac{200}{500}\right)^{1/2} + \left(\frac{200}{500}\right)^{6}\right] = 3.39$$

- Selecting $S8$ gives:

$$GSI(S8) = \left[\left(\frac{300}{500}\right)^{1/2} + \left(\frac{300}{500}\right)^{6}\right] + \left[\left(\frac{300}{500}\right)^{1/2} + \left(\frac{300}{500}\right)^{6}\right] + \left[\left(\frac{300}{500}\right)^{1/2} + \left(\frac{300}{500}\right)^{6}\right] + \left[\left(\frac{300}{500}\right)^{1/2} + \left(\frac{300}{500}\right)^{6}\right] = \mathbf{3.29}$$

Hence, using the *GSI* formula, it is always possible to identify and select the best set of lots to inspect. However, in this first *GSI* formula, the *Warning Limit* that represents a limit above which the situation starts to become critical is not taken into account. Depending on the criticality of some processing steps, or the cycle time between process and inspection tools, some lots may be prioritized to avoid reaching the *Inhibit Limit*. Missing to consider the *Warning Limit* may lead to situations where the *Inhibit Limit* will be reached because of delayed actions. This is why a second *GSI* formula integrating the *Warning Limit* was proposed.

6.3.2 GSI 2

The GSI-2 formula is based on the GSI-1 formula but integrates the *Warning Limit* that represents a threshold above which the situation starts to become critical in term of control. The objective is to select sets of lots that allow staying well below the *Inhibit Limit* and if possible below the *Warning Limit*. Parameters α and β, as in the GSI-1 formula, are used to give more or less priority depending on sets of lots S to analyze. The score is given by:

$$GSI(S) = \sum_{r=1}^{R} \left[\left(Min \left(1, \frac{\frac{NRV_r}{IL_r}}{\frac{WL_r}{IL_r}} \right) \right)^{1/\beta} + \left(Max \left(0, \frac{\frac{NRV_r}{IL_r} - \frac{WL_r}{IL_r}}{1 - \frac{WL_r}{IL_r}} \right) \right)^{\alpha} \right]$$

Or

$$GSI(S) = \sum_{r=1}^{R} \left[\left(Min \left(1, \frac{NRV_r}{WL_r} \right) \right)^{1/\beta} + \left(Max \left(0, \frac{NRV_r - WL_r}{IL_r - WL_r} \right) \right)^{\alpha} \right]$$

With this new formula, the aim is not only to stay well below the *Inhibit Limit* but also penalize sets of lots where risk values (NRV_r) are larger than *Warning Limits*. When $NRV_r < WL_r$, the GSI is given by:

$$GSI(S) = \sum_{r=1}^{R} \left(\frac{NRV_r}{WL_r} \right)^{1/\beta}$$

Parameter β increases the priority of sets of lots that allow staying below WL_r. The smaller NRV_r/WL_r, the smaller the associated GSI (Figure 6.5).

When $NRV_r \geqslant WL_r$, the associated set of lots is strongly penalized with parameter α. In this case, the GSI is given by:

$$GSI(S) = 1 + \sum_{r=1}^{R} \left(\frac{\frac{NRV_r}{IL_r} - \frac{WL_r}{IL_r}}{1 - \frac{WL_r}{IL_r}} \right)^{\alpha} = 1 + \sum_{r=1}^{R} \left(\frac{NRV_r - WL_r}{IL_r - WL_r} \right)^{\alpha}$$

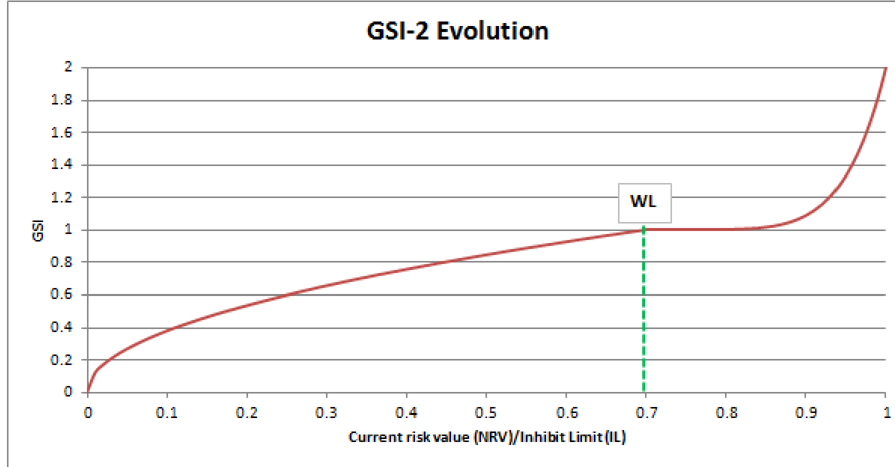


Figure 6.5: GSI-2 Evolution.

Figure 6.5 shows the evolution of this new GSI formula when $WL_r = 0.7 * IL_r$. Contrary to the GSI-1 formula, sets of lots where NRV_r are close to IL_r are strongly penalized with a very large value of the GSI.

Let us consider the example in Table 6.7. The same example has been evaluated with the GSI-1 formula (see Table 6.6). The best set was $S8$ because, with $S7$, the new risk value (NRV_r) associated to tool r_3 is very close to the *Inhibit Limit* IL_r.

Tools	NRV_r	WL_r	IL_r
$r1$	200	350	500
$r2$	200	350	500
$r3$	450	350	500
$r4$	200	350	500

(a) Set of lots **S9** selected.

Tools	NRV_r	WL_r	IL_r
$r1$	300	350	500
$r2$	300	350	500
$r3$	300	350	500
$r4$	300	350	500

(b) Set of lots **S10** selected.

Table 6.7: Example5 - Evaluating two different sets of lots **S7** and **S8** using the GSI.

Using the GSI-2 formula and defining *Warning Limits* based on the *Inhibit Limits* ($WL_r = 0.7 * IL_r = 0.7 * 500 = 350$), we obtain (with $\alpha = 6$ and $\beta = 2$):

- Selecting $S9$ gives:

$$GSI_2(S9) = \left[\left(Min \left(1, \frac{200}{350} \right) \right)^{1/2} + \left(Max \left(0, \frac{200-350}{500-350} \right) \right)^6 \right] +$$
$$\left[\left(Min \left(1, \frac{200}{350} \right) \right)^{1/2} + \left(Max \left(0, \frac{200-350}{500-350} \right) \right)^6 \right] +$$
$$\left[\left(Min \left(1, \frac{450}{350} \right) \right)^{1/2} + \left(Max \left(0, \frac{450-350}{500-350} \right) \right)^6 \right] +$$
$$\left[\left(Min \left(1, \frac{200}{350} \right) \right)^{1/2} + \left(Max \left(0, \frac{200-350}{500-350} \right) \right)^6 \right] = \mathbf{3.36}$$

- Selecting $S10$ gives:

$$GSI_2(S10) = \left[\left(Min \left(1, \frac{300}{350} \right) \right)^{1/2} + \left(Max \left(0, \frac{300-350}{500-350} \right) \right)^6 \right] +$$
$$\left[\left(Min \left(1, \frac{300}{350} \right) \right)^{1/2} + \left(Max \left(0, \frac{300-350}{500-350} \right) \right)^6 \right] +$$
$$\left[\left(Min \left(1, \frac{300}{350} \right) \right)^{1/2} + \left(Max \left(0, \frac{300-350}{500-350} \right) \right)^6 \right] +$$
$$\left[\left(Min \left(1, \frac{300}{350} \right) \right)^{1/2} + \left(Max \left(0, \frac{300-350}{500-350} \right) \right)^6 \right] = 3.70$$

Contrary to the GSI-1 formula, the GSI-2 selects $S9$ as the best set of lots for inspection. The set of lots $S10$ where NRV_r values are close to WL_r ($r_1 = r_2 = r_3 = r_4 = 300$ and $WL_r = 350$) are penalized. Depending on the GSI formula, the set of lots that is selected is different:

- GSI-1 formula $\Rightarrow S8$. States where NRV_r is close to IL_r are strongly penalized.

- GSI-2 formula $\Rightarrow S7$. States where NRV_r is close to WL_r and IL_r are penalized.

When looking at Table 6.7, the solutions provided by the two GSI formulas (GSI-1 and GSI-2) can be discussed. If the values of parameters α and β are not

the same for the two GSI formulas, only using the GSI formulas to select lots is not enough. Therefore, there is a need to consider additional parameters to ensure efficient sampling, skipping, and scheduling of lots on inspection tools. For example, increasing the priority of some lots based on their processing history, or defining threshold values below which a lot that has been sampled cannot be skipped.

To consider additional parameters, two different GSI sampling algorithms are proposed. They are based on the GSI formulas (GSI-1 and GSI-2) and production constraints such as prioritizing lots that minimize the number of *Inhibit Limits* that are not satisfied, or not sampling lots that bring less than a given percentage of gain on the GSI. The next section describes these two algorithms and section 6.5.2 presents the performance of each algorithm based on simulations.

6.4 GSI sampling algorithms

GSI sampling algorithms are based on *GSI formulas*, *Warning Limits*, *Inhibit Limits*, and some *thresholds*. *Thresholds* help mastering the cycle time of lots by avoiding sampling lots that may be skipped later because of the arrival of new lots containing more information. The objectives are twofold:

1. Sample, skip, and dynamically schedule lots on inspection tools while ensuring an optimal use of the inspection capacity.
2. Minimize the risks on the entire production while ensuring a maximum risk level below the *Inhibit Limits*.

In this section, we first define the different types of *thresholds* and then present the GSI sampling algorithms that are evaluated through simulations (Section 6.5.2).

6.4.1 Threshold definitions

A threshold can be defined as a limit above or below which an action may be taken. For dynamically sampling, skipping, and scheduling lots, three different thresholds are defined:

1. **Minimum threshold** (T_{Min}) = Minimum gain required for a lot to enter the inspection queue when the latter is *empty*.
2. **Maximum threshold** (T_{Max}) = Minimum gain required for a lot to enter the inspection queue when the latter is *full*.
3. **Metrology threshold** (T_{Metro}) = Minimum gain required for a lot to remain in the inspection queue after completing the inspection of another lot.

T_{Min} and T_{Max} are used in the entrance of the inspection queue to help deciding whether or not a lot should be sampled. T_{Metro} is used for lots already in the inspection queue. The three thresholds (T_{Min}, T_{Max}, T_{Metro}) are based on the GSI, i.e. the gain of a lot l is always evaluated within a set S of lots (section 6.3). This gain for a lot l in a set S is given by:

$$Gain(l) = 1 - \frac{GSI(S \cup \{l\})}{GSI(S)} \in [0, 1] \, .$$

$Gain(l)$ is strictly positive since inspecting an additional lot cannot worsen the GSI.

- **Minimum threshold (T_{Min}) and Maximum threshold (T_{Max}).** T_{Min} and T_{Max} are fixed values that never vary. When the inspection queue is *empty*, T_{Min} is used and, when the queue is full, T_{Max} is used. When the inspection queue is partially filled, the threshold used is proportional to the size of the inspection queue. The threshold is defined with the following formula:

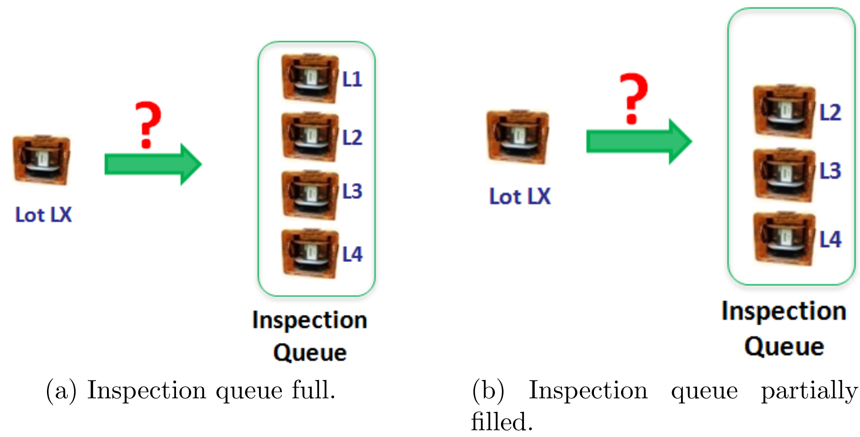
$$Threshold = T_{Min} + \left[\frac{NBQ}{SQ} * (T_{Max} - T_{Min}) \right]$$

where NBQ is the number of lots in the inspection queue and SQ the inspection queue size (i.e. capacity). All thresholds are given in percentages. Let us consider Figure 6.6 and Figure 6.7. There are three possible cases: The inspection queue is full (Figure 6.6a), partially filled (Figure 6.6b), or empty (Figure 6.7). For these three cases, a decision must be taken regarding the sampling of a lot LX that arrives in front of the inspection step. Two steps are performed: Compute the GSI and verify if the gain associated to LX satisfies the threshold limits.

Example: $T_{Min} = 5\%$, $T_{Max} = 20\%$.

A. Figure 6.6a → The queue is full. In this case, sampling LX leads to skipping another lot in the inspection queue. Therefore, we need to analyze all possible combinations and select the best one. This means evaluating:

$$\begin{aligned} GSI_1 &= GSI(\{L1, L2, L3, L4\}) \\ GSI_2 &= GSI(\{LX, L2, L3, L4\}) \\ GSI_3 &= GSI(\{L1, LX, L3, L4\}) \\ GSI_4 &= GSI(\{L1, L2, LX, L4\}) \\ GSI_5 &= GSI(\{L1, L2, L3, LX\}) \end{aligned}$$



(a) Inspection queue full. (b) Inspection queue partially filled.

Figure 6.6: Inspection queue not empty.

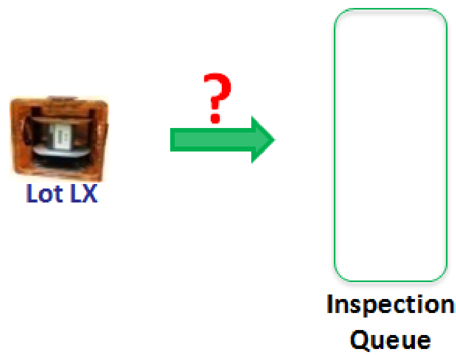


Figure 6.7: Inspection queue empty.

GSI_1 is the reference, i.e. the score associated to the set of lots $\{L1, L2, L3, L4\}$ already in the inspection queue. If GSI_2, GSI_3, GSI_4, or GSI_5 is lower than GSI_1, then sampling LX and skipping another lot in the queue is valuable. For example, if $GSI_3 < GSI_1$, sampling LX and skipping $L2$ is interesting. But, before performing such a decision, we need to verify that the gain associated to sampling LX satisfies the threshold. Since the inspection queue is full (Figure 6.6a), the maximal threshold $T_{Max} = 20\%$ is used and it is necessary to verify that:

$$Gain(LX) = \left[1 - \frac{GSI_3}{GSI_1}\right] * 100\% \geqslant 20\%$$

If $Gain(LX) \geqslant 20\%$, then sampling LX and skipping $L2$ improves the GSI score by at least 20%.

B. Figure 6.6b → The queue is partially filled. Three lots are waiting in the inspection queue and the queue size is 4. It means that there is an *available* place and LX can be sampled. However, as we want to optimally use all inspection tools, we want to ensure that inspecting an additional lot improves the situation (in term of risk) enough. Therefore, we verify that:

$$Gain(LX) = 1 - \frac{GSI(\{L1, L2, L3, LX\})}{GSI(\{L1, L2, L3\})} * 100\% \geqslant$$
$$T_{Min} + \frac{NBQ}{SQ} * (T_{Max} - T_{Min})$$

i.e.

$$Gain(LX) = \left[1 - \frac{GSI(\{L1, L2, L3, LX\})}{GSI(\{L1, L2, L3\})}\right] * 100\% \geqslant 5 + (3/4) * (20 - 5)\%$$

$$\Rightarrow Gain(LX) = \left[1 - \frac{GSI(\{L1, L2, L3, LX\})}{GSI(\{L1, L2, L3\})}\right] * 100\% \geqslant 16.25\%$$

If $Gain(LX) \geqslant 16.25\%$, then LX is sampled and added to the inspection queue, otherwise, LX will not be sampled.

C. Figure 6.7 → The queue is empty. There are 4 available places into the queue but we need to verify that inspecting LX improves the situation enough. The case where LX is inspected ($GSI(\{LX\})$ is compared to the case where no lot is inspected ($GSI(\varnothing)$), i.e.:

$$Gain(LX) = 1 - \frac{GSI(\{LX\})}{GSI(\varnothing)} * 100\% \geqslant T_{Min}$$

$$\Rightarrow Gain(LX) = 1 - \frac{GSI(\{LX\})}{GSI(\varnothing)} * 100\% \geqslant 5\%$$

If $Gain(LX) \geqslant 5\%$, then LX is sampled and added to the inspection queue.

- **Metrology threshold** (T_{Metro}). T_{Metro} is a fixed value in percentage that helps mastering the number of lots skipped after the inspection of other lots. Each time a lot is inspected, the gain associated to lots in the inspection queue is modified. Depending on the time spent in the inspection queue, the gain of some lots may strongly decrease and, thus, it may be interesting to skip these lots to avoid increasing their cycle time for nothing. Let us consider the example in Figure 6.8. Lot $L1$ has been inspected and there is a need to define whether or not a lot (LX, or $L3$, or $L4$) in the queue should be skipped. To skip a lot l after an inspection, the gain having l in a set S of lots must be lower than T_{Metro}. This means:

$$Gain(l \text{ in } S) = \frac{GSI(S \setminus \{l\})}{GSI(S)} - 1 < T_{Metro}.$$

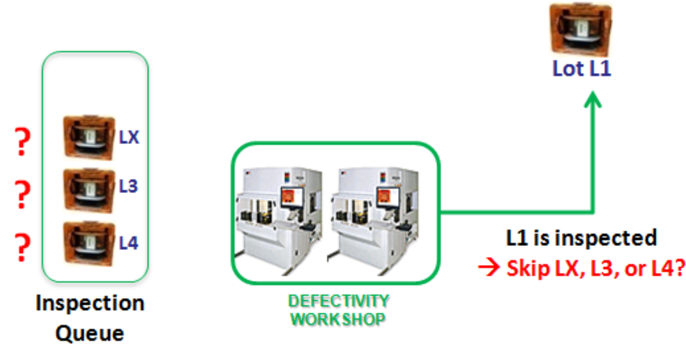


Figure 6.8: Threshold metrology (T_{Metro}) - Skipping lots after inspection.

For example, if $T_{Metro} = 10\%$ and $L1$ has just been inspected, to decide whether a lot should be removed from the inspection queue, we need to assess the gain of each lot in the queue i.e. in the set $S = \{LX, L3, L4\}$. This means evaluating:

$$Gain(LX \text{ in } S) = \frac{GSI(\{L3, L4\})}{GSI(\{LX, L3, L4\})} - 1$$

$$Gain(L3 \text{ in } S) = \frac{GSI(\{LX, L4\})}{GSI(\{LX, L3, L4\})} - 1$$

$$Gain(L4 \text{ in } S) = \frac{GSI(\{LX, L3\})}{GSI(\{LX, L3, L4\})} - 1$$

If there is a lot l such that $Gain(l \text{ in } S) < T_{Metro}$, then l will be skipped, removed from the inspection queue and directed to its next processing operation. For example, if $Gain(LX \text{ in } S) < 10\%$, then lot LX will be skipped.

These three *thresholds* (T_{Max}, T_{Min}, T_{Metro}) were progressively introduced based on simulations results (Section 6.5). The *threshold* T_{Max} was introduced to reduce the number of lots skipped. By using the GSI formulas to sample lots, we observed that some lots were sampled because of their gains but never inspected because of the arrival of new lots bringing more information. Consequences were the increasing of the cycle times of those lots that were stopped at inspection steps without being inspected. By defining a minimum gain (T_{Max}) to satisfy before sampling a lot, we could reduce the number of sampled lots and thus the number of skipped lots. T_{Min} was defined to ensure that, even if the inspection queue is empty, sampling a lot will always improve the *situation* within production.

T_{Metro} was defined to master the number of skipped lots due the reduction of the gain after each inspection. This threshold was introduced to dissociate the threshold in the entrance of the queue (T_{Max}) and the threshold required for lots to stay in the queue after each inspection. Increasing the threshold in the entrance of the queue leads to the reduction of the number of sampled lots, and thus the number of skipped lots, while increasing the threshold required for each lot to stay in the queue leads to increasing the number of skipped lots because of the incompressible waiting time in front of inspection tools. By separating the two thresholds (T_{Max} and T_{Metro}), we could master both the number of sampled lots and skipped lots.

In the next sections, we present the different GSI sampling algorithms that have been developed based on *GSI formulas*, *Warning Limits*, *Inhibit Limits*, and the

different *thresholds*.

6.4.2 GSI sampling algorithm 1 (GSI-SA-1)

Sampling a lot l is based on how much is gained when adding l to the set of lots $S_{initial}$ already in the inspection queue. We compare the number of inhibit limits that are violated, the number of warning limits that are violated and the GSI of the sets of lots obtained by adding l in $S_{initial}$, i.e. $S_{initial} \cup \{l\}$, or by removing l' from $S_{initial}$ and adding l, i.e. $S_{initial} \setminus \{l'\} \cup \{l\}$. Let us consider the following notations:

- $S_{initial}$: Set of lots already in the inspection queue,
- NBQ: Number of lots in $S_{initial}$ ($NBQ = |S_{initial}|$), i.e. number of lots already in the inspection queue,
- SQ: Size of the inspection queue,
- $NbIL(S)$: Number of Inhibit Limits that are violated if the set of lots S is selected for inspection,
- $NbWL(S)$: Number of Warning Limits that are violated if the set of lots S is selected for inspection.

The first GSI sampling algorithm (GSI-SA-1) determines the best set of lots $S^{\star}$ and uses the GSI-1 formula (section 6.3.1), i.e.:

$$GSI(S) = \sum_{r=1}^{R} \left[\left(\frac{NRV_r(S)}{IL_r} \right)^{1/\beta} + \left(\frac{NRV_r(S)}{IL_r} \right)^{\alpha} \right]$$

If the number of lots already in the inspection queue is strictly smaller than the size of the inspection queue, i.e. $NBQ < SQ$, then only adding l in $S_{initial}$ is evaluated and compared to not adding l. Otherwise, i.e. $NBQ = SQ$, all combinations associated to removing each $l' \in S_{initial}$ from $S_{initial}$ and adding l in $S_{initial}$ are evaluated.

In the description below, SS denotes a set of sets of lots.

GSI-SA-1 – Selecting the best set of lots $S^{\star}$ using IL, WL and GSI-1

1: **Initialization**: $S^{\star} = S_{initial}$
2: **If** $NBQ = SQ$ **then**
3: $SS = \emptyset$
4: **For** each lot $l' \in S_{initial}$
5: $SS = SS \cup \{S_{initial} \setminus \{l'\} \cup \{l\}\}$
6: **End for**
7: **ElseIf** $NBQ < SQ$ **then**
8: $SS = \{S_{initial} \cup \{l\}\}$
9: **End if**
10: **For** each set of lots $S \in SS$
11: **If** $NbIL(S) < NbIL(S^{\star})$ **then**
12: $S^{\star} = S$
13: **ElseIf** $NbIL(S) = NbIL(S^{\star})$ and $NbWL(S) < NbWL(S^{\star})$ **then**
14: $S^{\star} = S$
15: **ElseIf** $NbIL(S) = NbIL(S^{\star})$ and $NbWL(S) = NbWL(S^{\star})$ **then**
/ Only gains that satisfy threshold values are accepted. */*
16: **If** $GSI(S) < GSI(S^{\star})$ and
$[1 - GSI(S)/GSI(S_{initial})] \geqslant \left[T_{Min} + \frac{NBQ}{SQ} * (T_{Max} - T_{Min})\right]$ **then**
17: $S^{\star} = S$
18: **End If**
19: **End If**
20: **End for**

6.4.3 GSI sampling algorithm 2 (GSI-SA-2)

Contrary to the first algorithm, the second GSI sampling algorithm ((GSI-SA-2) uses the GSI-2 formula (section 6.3.2). Since Warning Limits are already included in the GSI-2 formula, this second algorithm does not start by ranking solutions in a lexicographical order, i.e. verifying the number of Inhibit Limit and Warning Limit that are violated. The selection of the best set of lots is directly performed by using the GSI-2 formula and threshold limits:

$$GSI(S) = \sum_{r=1}^{R} \left[\left(Min \left(1, \frac{\frac{NRV_r}{IL_r}}{\frac{WL_r}{IL_r}} \right) \right)^{1/\beta} + \left(Max \left(0, \frac{\frac{NRV_r}{IL_r} - \frac{WL_r}{IL_r}}{1 - \frac{WL_r}{IL_r}} \right) \right)^{\alpha} \right]$$

Warning Limits and Inhibit Limits are no longer thresholds to avoid in this second algorithm.

GSI-SA-2 – Selecting the best set of lots $S^\star$ using GSI-2

1: **Initialization**: $S^\star = S_{initial}$
2: **If** $NBQ = SQ$ **then**
3: $SS = \emptyset$
4: **For** each lot $l' \in S_{initial}$
5: $SS = SS \cup \{S_{initial} \setminus \{l'\} \cup \{l\}\}$
6: **End for**
7: **ElseIf** $NBQ < SQ$ **then**
8: $SS = \{S_{initial} \cup \{l\}\}$
9: **End if**
10: **For** each set of lots $S \in SS$
11: **If** $GSI(S) < GSI(S^\star)$ and $[1 - GSI(S)/GSI(S_{initial})] \geqslant \left[T_{Min} + \frac{NBQ}{SQ} * (T_{Max} - T_{Min})\right]$ **then**
12: $S^\star = S$
13: **End if**
14: **End for**

These two GSI sampling algorithms could have been evaluated using a fab-wide simulation as in **Gissrau and Rose** [27]. However, we wanted to use historical data that were available and focus on the sampling mechanisms. Hence, we used a simulator called **S5** (**S**mart **S**ampling **S**kipping **S**cheduling **S**imulator) [104] developed by the EMSE within the framework of the European project IMPROVE, to test the GSI sampling algorithms. **S5** also helped to evaluate the sampling algorithms on datasets from various European semiconductor fabs.

In the next section we discuss the performances of each sampling algorithm and analyze the impact of input parameters in the GSI sampling performances.

6.5 Numerical experiments

Simulations presented in this section were performed on actual data from the 300-mm site of STMicroelectronics in Crolles, France. We used six weeks of historical data. With 8 different technologies and 244 production tools, the time required to simulate a fab-wide sampling policy is 7 minutes (4 minutes for simulation and 3 minutes to generate statistics). The characteristics of the computer are: 2.53GHz, 4GB of RAM, and Windows 7 as the operating system. For all simulations, we defined the Warning and Inhibit Limits to 1000 and 2000 respectively for all production tools. The impact of these Warning and Inhibit Limits are evaluated and computed for each production tool in the next chapter (Chapter 7).

6.5.1 S5 simulator

The **S5** (**S**mart **S**ampling **S**kipping **S**cheduling **S**imulator) simulator was developed within the framework of the European project IMPROVE [104]. It uses historical data to simulate various sampling policies (see Appendix C.1 for further details). For each sampling policy that is simulated, the simulator provides several statistics and we use some of these statistics to assess the performance of GSI sampling algorithms. Among these statistics, we use the following indicators:

1. **Number of lots that are sampled**: The number of lots that are selected for measurement and placed in the metrology queue.

2. **Number of lots that are measured**: The number of lots that are processed on metrology tools.

3. **Number of lots that are skipped**: The number of lots that are removed from the metrology queue, i.e. the number of lots that are sampled but not measured.

4. **Number of lots that are skipped (entry queue)**: The number of lots that are removed from the metrology queue due to the arrival of new lots having more information.

5. **Number of lots that are skipped (metrology)**: The number of lots that are removed from the metrology queue due to the measurement of other lots.

6. **Medium WAR (average)**: The sum of the WAR of all production tools divided by the number of tools. It is equal to $\sum_{j=1}^{NbTools} \frac{WAR_j}{NbTools}$ where WAR_j is the WAR for production tool j, and $NbTools$ the number of production tools.

7. **Maximum WAR (average)**: The sum of the maximum WAR of all production tools divided by the number of tools. It is equal to $\sum_{j=1}^{NbTools} \frac{MaximumWAR_j}{NbTools}$.

8. **Number of wafers above Warning Limit**: The number of wafers that are processed on production tools when WAR is above the Warning Limit.

9. **Number of wafers above Inhibit Limit**: The number of wafers that are processed on production tools when WAR is above the Inhibit Limit.

6.5.2 Evaluating GSI sampling algorithms

To evaluate the performances of the GSI sampling algorithms (section 6.4) versus Fab sampling, we define and consider the same parameter values for the two GSI sampling algorithms. We first aim at evaluating and quantifying the performances of each GSI sampling algorithm, and then analyzing the impact of different parameters (α, β, T_{Max}, T_{Min}, T_{Metro}) in the sampling policy performances. We define:

- $\alpha = 6$ and $\beta = 2$.
- $T_{Min} = 0\%$.
- $T_{Max} = 4\%$.
- $T_{Metro} = 0\%$.
- $WL = 1000$ and $IL = 2000$ for all production tools.

- Number of metrology tools = 2 and queue size = 4. In fact, in our historical data, 13 metrology tools are used. However, these metrology tools are not used at 100% because of maintenance, engineering actions, or qualifications. To compare the GSI and Fab sampling policies, we need to have the same number of measurements. This is why we adjust the number of metrology tools as well as the measure time. By defining 2 metrology tools used at 100% with a measure time of X, we ensure the saturation of our metrology capacity. The queue size is set to be twice the number of metrology tools, i.e. two lots in front of each metrology tool. This is a choice that can be discussed depending on the factory dynamics or the historical data.

Table 6.8 shows the experimental results for the four cases: "Fab sampling", "All sampling", "GSI algo-1", "GSI algo-2". **Fab sampling** corresponds to the sampling that was actually performed in production. **All sampling** corresponds to measuring all lots. It gives indication on theoretical performances if all lots were measured, i.e. infinite capacity. **GSI algo-1** and **GSI algo-2** correspond to the sampling obtained with the two GSI sampling algorithms (Section 6.4.2 and Section 6.4.3). All results are normalized based on **Fab sampling** results.

Note that, whatever the sampling policy, the four following performances indicators are improved compared to **Fab sampling**: Medium WAR (average), Maximum WAR (average), Number of wafers above WL, Number of wafers above IL.

The case of **All sampling** shows that, measuring all lots does not ensures zero risk. Hence the importance of optimally using the metrology capacity. The two GSI sampling algorithms (**GSI algo-1** and **GSI algo-2**) provide better results compared to **Fab sampling** for the *same* measurement capacity (see the number of measured lots). **GSI algo-2** provides the minimum "Medium and Maximum WAR" whereas **GSI algo-1** ensures the minimum "Number of wafers above WL and IL". However, the differences are not so significant and can thus be discussed depending on the production environment. When exceeding IL may lead to stopping production tools, then **GSI algo-1** is preferable. When the primary goal is to minimize risk, i.e. the "Medium and Maximum WAR", then **GSI algo-2** is *more* suitable.

Indicators	**Fab sampling**	**All sampling**	**GSI algo-1**	**GSI algo-2**
Number of sampled lots	**A**	14.21*A	8.10*A	1.25*A
Number of measured lots	**A**	14.21*A	0.98*A	0.98*A
Number of skipped lots	**0**	0	7.12*A	0.27*A
Number of skipped lots (entry queue)	**0**	0	4.65*A	0.01*A
Number of skipped lots (metrology)	**0**	0	2.47*A	0.26*A
Medium WAR (average)	**B**	0.11*B	0.36*B	0.35*B
Maximum WAR (average)	**C**	0.16*C	0.45*C	0.44*C
Number of wafers above WL	**D**	0.10*D	0.79*D	0.87*D
Number of wafers above IL	**E**	0.06*E	0.49*E	0.59*E

Table 6.8: Evaluating the GSI sampling algorithms.

Looking at the number of lots that are sampled and the number of lots that are skipped, **GSI algo-2** outperforms **GSI algo-1**. For the same measurement capacity and approximatively the same performances in term of risk reduction (Medium and Maximum WAR), **GSI algo-2** samples only 1.25*A lots whereas **GSI algo-1** samples 6 times more lots (8.10*A). Consequences are the significant number of lots that are skipped (**GSI algo-1** skips 26 times more lots than **GSI algo-2**) since the measurement capacity is constant. This may impact the cycle times of lots that are sampled but never measured, i.e. skipped. The same for the "Number of skipped lots (entry queue)" and "Number of skipped lots (metrology)".

Considering all the performance indicators, the two GSI sampling algorithms outperforms Fab sampling. With the same number of measurements, the GSI sam-

pling algorithms help in strongly reducing both the risk, i.e. the WAR (Medium and Maximum), and the number of wafers above WL and IL. However, no conclusion can be taken regarding the GSI sampling algorithm that provides the best performances. Depending on the production environment (automated or not) or the management priorities (stopping production tools once IL is exceeded or minimizing the overall risk, i.e. Medium and Maximum WAR), a sampling algorithm (**GSI algo-1** or **GSI algo-2**) may be more suitable than the other.

In the next section, we analyze the impact of the five following parameters that are used in the GSI sampling algorithms: α, β, T_{Max}, T_{Min}, T_{Metro}. We aim at understanding their real impact in the GSI sampling performances, verifying the expected behavior of the sampling algorithms, and identifying values that ensure good performances. **We choose GSI algo-2 because of the reduced number of skipped lots**. However, this is just a choice, and the impact of parameters may vary depending on the GSI sampling algorithm or the production constraints.

6.5.3 Analyzing the impact of GSI parameters

We successively and separately vary all the parameters. We first analyze the impact of parameters α and β before analyzing the impact of *threshold* parameters (T_{Max}, T_{Min}, and T_{Metro}). α and β are directly used in the GSI formulas to compute scores that are associated to sets of lots, whereas T_{Max}, T_{Min}, and T_{Metro} are used in the GSI sampling algorithms[5] to manage the filling of metrology queues, i.e. the number of lots that are sampled and skipped.

Each parameter is separately analyzed, i.e. when we vary a parameter, we keep all the other parameters constant. This choice can be discussed since the impact of a parameter may be linked to the fixed values of the other parameters. However, simulating all possible combinations is not possible because of the time of each simulation (7 minutes). Moreover, since our primary goal is to verify that the GSI sampling algorithms have the expected behavior whatever the parameter values,

[5]GSI sampling algorithms are a *combination* of GSI formulas, threshold values, and additional constraints linked to metrology capacity and production environment.

analyzing parameters one after another help us understand and assess the robustness of the GSI sampling algorithms.

We focus on evaluating different situations with different parameter values rather than trying to find the parameters values that provide the optimal performances of the GSI sampling algorithms. We aim at identifying values that lead to a kind of instability, and thus reduce the *space* of possible values that each parameter can take. Our choice is also motivated by the fact the GSI sampling algorithms are to be used in different manufacturing environments with different constraints and priorities. The parameter values might not be the same in all situations. Hence our focus on analyzing the impact of parameters, identifying abnormalities, i.e. parameter values that make the GSI sampling algorithms unstable, and discussing the set of values that each parameter should take.

6.5.3.1 Analyzing the impact of parameters α and β

Parameters α and β are used in the GSI formulas to put more or less emphasis on getting as far as possible from the Inhibit Limit for which the current risk value is closer. α is used to penalize sets of lots for which the resulting risk values are closer to the Inhibit Limits. β prioritizes sets of lots fro which the associated risk values are far from the Inhibit Limits. The goal is stay as far away as possible below Inhibit Limit while minimizing the overall risk in the entire production (Section 6.3).

We use the S5 simulator [104] and the following performance indicators to assess the impacts of α and β:

1. Number of lots that are sampled.
2. Number of lots that are measured.
3. Number of lots that are skipped.
4. Medium WAR (average).
5. Maximum WAR (average).

We set $T_{Max} = T_{Min} = T_{Metro} = 0\%$. We start by analyzing the impact of α when $\beta = 1$. Then, we vary β for different values of α. Results show that, to ensure

good performances with the GSI sampling algorithms, α must be lower than 13 and $\beta \in [2, 10]$. **For the case of the 300-mm fab of STMicroelectronics, the values of $\alpha = 6$ and $\beta = 2$ provide satisfactory results.**

a) Impact of parameter α. To analyze the impact of α, we vary its value between 1 and 100 and consider the following parameters:

- $\beta = 1$
- $T_{Max} = T_{Min} = T_{Metro} = 0\%$
- $Warning\ Limit = 1000$
- $Inhibit\ Limit = 2000$

Figure 6.9 shows that α **impacts the number of measured lots**. Note that, if $\alpha > 13$, the number of measured lots decreases, i.e. metrology tools are no longer fully used. This means that there are either lots that are not sampled when the metrology queue is not full, or lots that are not measured when there are available capacity on metrology tools. This can be explained by the fact that, when $\alpha > 13$, the GSI score ($(NRV/IL)^{\alpha} + ...$) becomes very large and thus, differences between sets of lots (GSI scores) are not so significant in term of gains. Hence the reduced number of measured lots.

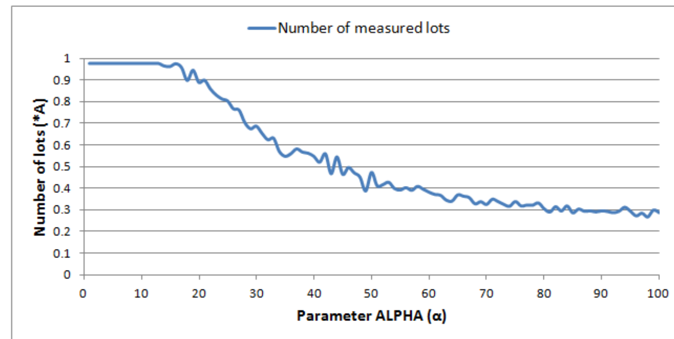


Figure 6.9: Impact of α on the number of measured lots.

Since metrology tools should be fully used, we only consider values of $\alpha \in [1, 12]$ and analyze the impact on the other performance indicators. Figure 6.10 and Figure 6.11 show the impact of α ($\in$ [1,12]) on the WAR values and on the number of sampled/skipped lots respectively.

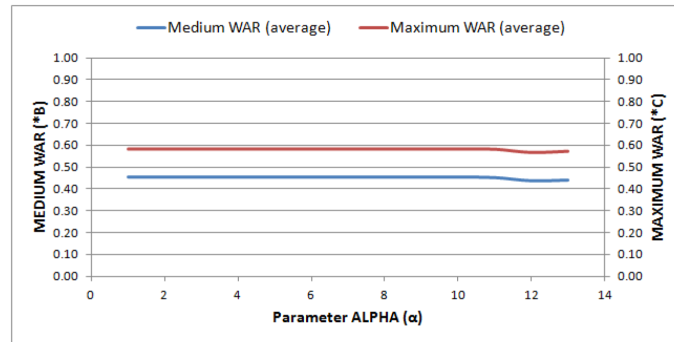


Figure 6.10: Impact of α on the Medium WAR and Maximum WAR.

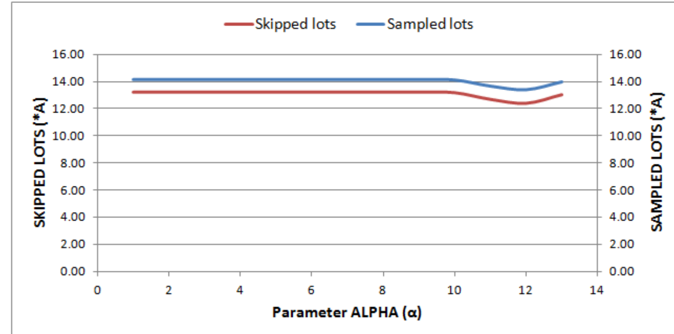


Figure 6.11: Impact of α on the number of sampled lots and skipped lots.

Note that the impact of α is negligible. The WAR values as well as the number of lots that are sampled and skipped are not really impacted by α. The GSI sampling algorithm ensures the measurement of the best possible sets of lots. Therefore, to expect good performances from the GSI sampling algorithms, α should not be too large. We thus only keep and consider $\alpha \in$ [1,12] in the remaining simulations.

b) Impact of parameter β. To analyze the impact of β, we consider the following parameters:

- $\alpha \in [1,12] \rightarrow \{1,2,4,6,8,10,12\}$
- $T_{Max} = T_{Min} = T_{Metro} = 0\%$
- *Warning Limit* = 1000
- *Inhibit Limit* = 2000

Parameter β is used in the GSI formula (Section 6.3.2) as $1/\beta$ (... + $(\text{NRV/IL})^{1/\beta}$). This implies that, the higher β, the lower its impact in the GSI score. For example, if $\beta = 20$, it implies that $1/\beta = 0.05$, i.e. $(\text{NRV/IL})^{1/\beta}$ # 1.

For β to have an impact in the GSI formula, we only consider values of β below 10. Then, by varying β between 1 and 10 for $\alpha = \{1, 2, 4, 6, 8, 10, 12\}$, we analyze different performance indicators (Appendix C.2 - See the average value of each performance indicator).

Results show that, when α increases, whatever the value of β, the average Medium WAR and Maximum WAR tend to increase whereas the number of lots that are skipped decreases. As one of the primary goal is to minimize the risk in the entire production without increasing the cycle time of lots (i.e. number of skipped lots), we need to find a trade-off between the number of lots that are skipped and the Medium and Maximum WAR. We thus choose the medium value of α between 1 and 12, i.e. $\alpha = 6$.

To identify the value of β that ensures good performances for the GSI sampling algorithms, we analyze our performances indicators when $\alpha = 6$. Table 6.9 shows the impact of β when it varies between 2 and 10[6].

Note that[7] the differences in the Medium and Maximum WAR can be neglected. Therefore, we focus on the number of lots that are skipped. Results show that the

[6]We do not consider the value of $\beta = 1$ since, if $\beta = 1$, it does not impact the GSI formula (Section 6.3.2).

[7]All results are normalized based on Fab sampling. **A** is the number of lots that are sampled, B the Medium WAR (average), and C the Maximum WAR(average).

Values of β with $\alpha = 6$	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
2	5.34*A	0.98*A	4.35*A	0.24*B	0.28*C
3	5.58*A	0.98*A	4.60*A	0.24*B	0.28*C
4	5.76*A	0.98*A	4.78*A	0.24*B	0.28*C
5	5.93*A	0.98*A	4.95*A	0.24*B	0.28*C
6	6.05*A	0.98*A	5.07*A	0.24*B	0.28*C
7	6.23*A	0.98*A	5.25*A	0.24*B	0.28*C
8	6.26*A	0.98*A	5.28*A	0.24*B	0.28*C
9	6.37*A	0.98*A	5.38*A	0.25*B	0.28*C
10	6.58*A	0.98*A	5.60*A	0.24*B	0.28*C

Table 6.9: Impact of β when $\alpha = 6$.

value of $\beta = 2$ ensures a good trade-off between the Medium WAR, the Maximum WAR, and the number of lots that are skipped. **Hence the choice of $\alpha = 6$ and $\beta = 2$ for the case of the 300mm fab of STMicroelectronics**. If these values of α and β may vary depending on the set of data or the production environment, selecting $\alpha < 12$ and $\beta \in [2,10]$ seem to ensure optimized sampling policies with the GSI sampling algorithms.

6.5.3.2 Analyzing the impact of *threshold* parameters

Threshold parameters (T_{Max}, T_{Min}, and T_{Metro}) have been introduced in the GSI sampling algorithms to master the number of sampled lots and skipped lots. The aim is to ensure the measurement of the best possible lots for a fixed metrology capacity (Section 6.4). T_{Max} and T_{Min} are used at the entrance of the metrology queue, whereas T_{Metro} is used each time a measurement is completed. T_{Max} is the minimum gain required for a lot to enter the metrology queue when it is full. T_{Min} is the minimum gain required for a lot to enter the metrology queue when it is empty. And T_{Metro} is the minimum gain required for a lot to remain in the metrology queue after the measurement of another lot is completed.

These three *threshold* parameters have direct impacts on the lots that are sampled, skipped, measured, and thus on the sampling policy performances. For example, if the threshold in the entrance of the queue (T_{Max}) is very high, only few lots will be sampled because of the higher gain required to enter the queue. Consequences can be the increase of the number of lots processed above Inhibit Limits and thus an increase of the potential loss if a problem occurs in production.

To analyze the impact of the three threshold parameters, we use the S5 simulator (Appendix C.1) and analyze the following indicators:

1. Number of lots that are sampled.
2. Number of lots that are measured.
3. Number of lots that are skipped.
4. Number of lots that are skipped (entry queue).
5. Number of lots that are skipped (metrology).
6. Medium WAR (average).
7. Maximum WAR (average).
8. Number of wafers above Warning Limits.
9. Number of wafers above Inhibit Limits.

10. Average time spent in the queue before measurement.

11. Average time in the queue before skip (entry queue).

12. Average time in the queue before skip (metrology).

Results indicate that, T_{Min} **impacts the number of measured lots** leading to reducing the saturation of the metrology tools. T_{Max} **impacts the number of lots that are skipped in the entrance of the queue** leading to an increase of the number of lots processed above Inhibit Limits, and T_{Metro} **impacts the number of lots that are removed from metrology queue** leading to an increase of the cycle time of lots, i.e. the number of lots that are sampled but never measured. **For the case of the 300-mm fab of STMicroelectronics, the best trade-off seems to be $T_{Max} = 1\%$, $T_{Min} = 0\%$, and $T_{Metro} = 0\%$.**

a) Impact of parameter T_{Max}. Table 6.10 shows the impact of T_{Max} on the GSI sampling performances. Let us focus on the first three indicators: The number of sampled lots, measured lots, and skipped lots. When T_{Max} increases (0%, 1%, 2%, 10%, 20%), the number of lots that are sampled decreases leading to a reduction in the number of lots that are skipped. The number of measured lots remains constant ($0.98 * A$). The higher the value of T_{Max}, the lower the number of skipped lots.

The number of skipped lots is the sum of the number of skipped lots at the entrance of the queue and the number of skipped lots after a measurement is completed (see indicators 4 and 5). T_{Max} mainly impacts the number of skipped lots at the entrance of the queue (indicator 4). Note that, when $T_{Max} \geq 10$, no lots are skipped. This means that the gain required for each lot to enter the queue becomes so large that, once a lot enters the queue, it is very difficult or impossible to remove it from the queue, or find another lot that brings a gain which is large enough.

	T_{Max}	0%	1%	2%	10%	20%
1	Number of sampled lots	5.34*A	1.12*A	1.05*A	1.00*A	1.00*A
2	Number of measured lots	0.98*A	0.98*A	0.98*A	0.98*A	0.98*A
3	Number of skipped lots	4.35*A	0.14*A	0.07*A	0.02*A	0.02*A
4	**Number of skipped lots (entry queue)**	**4.33*A**	**0.09*A**	**0.01*A**	**0**	**0**
5	Number of skipped lots (metrology)	0.02*A	0.05*A	0.06*A	0.02*A	0.02*A
6	Medium WAR (average)	0.24*B	0.31*B	0.32*B	0.36*B	0.37*B
7	Maximum WAR (average)	0.28*C	0.40*C	0.42*C	0.47*C	0.48*C
8	Number of wafers above WL	0.67*D	0.82*D	0.83*D	0.88*D	0.89*D
9	Number of wafers above IL	0.20*E	0.47*E	0.51*E	0.63*E	0.65*E
10	Average time spent in the queue before measurement	**X**	1.07*X	0.98*X	0.83*X	0.81*X
11	Average time in the queue before skip (entry queue)	**Y**	4.15*Y	8.83*Y	–	–
12	Average time in the queue before skip (metrology)	**Z**	1.64*Z	1.87*Z	1.02*Z	1.09*Z

Table 6.10: Impact of $T_{Max} \in [0,20\%]$.

Consequences are the increasing of the WAR values (Medium and Maximum WAR), and the number of wafers processed above Warning Limits and Inhibit Limits (see indicators 6, 7, 8, and 9). As the gain becomes very large to enter the queue, fewer lots are sampled. However, as the metrology capacity must be fully used, the rest of capacity is *wasted* by the measurement of lots that do not satisfy the required gain.

The average time spent in the queue before measurement (indicator 10) decreases since there is little waiting for lots in front of metrology tools because of the reduced number of lots that are sampled. Lots that are sampled are quickly measured. The time spent in the queue before skip (indicator 11) increases because lots that enter the queue have large gains, and it is very difficult to find other lots with higher gains. For $T_{Max} = 10\%$ and $T_{Max} = 20\%$, we do not report time since no lot is skipped. The time spent in the queue before skip after a measurement is completed (indicator 12) is impacted by T_{Max} but no conclusion can be taken at this stage since this last indicator is mainly mastered with T_{Metro}.

Between 0 and 1%, the number of skipped lots in the entrance of the queue (indicator 4) is reduced by more than 97% (4.35*A → 0.14*A). To understand what happens between 0 and 1%, we performed simulations by varying T_{Max} between 0 and 1% (see Table 6.11 and Table 6.12). The same kind of observations are made as in Table 6.10, i.e. reduction of the number of skipped lots, increase of the WAR values, increase of the number of wafers above WL and IL, reduction of the time spent in the queue before measurement, and increase of the time before skip.

	T_{Max}	0%	0.1%	0.2%	0.3%	0.4%	0.5%
1	Number of sampled lots	5.34*A	1.84*A	1.55*A	1.42*A	1.34*A	1.27*A
2	Number of measured lots	0.98*A	0.98*A	0.98*A	0.98*A	0.98*A	0.98*A
3	Number of skipped lots	4.35*A	0.86*A	0.57*A	0.44*A	0.36*A	0.29*A
4	**Number of skipped lots (entry queue)**	**4.33*A**	**0.81*A**	**0.51*A**	**0.39*A**	**0.31*A**	**0.23*A**
5	Number of skipped lots (metrology)	0.02*A	0.05*A	0.06*A	0.05*A	0.05*A	0.06*A
6	Medium WAR (average)	0.24*B	0.28*B	0.29*B	0.29*B	0.29*B	0.30*B
7	Maximum WAR (average)	0.28*C	0.36*C	0.37*C	0.37*C	0.37*C	0.39*C
8	Number of wafers above WL	0.67*D	0.76*D	0.77*D	0.78*D	0.80*D	0.79*D
9	Number of wafers above IL	0.20*E	0.36*E	0.39*E	0.41*E	0.42*E	0.43*E
10	Average time spent in the queue before measurement	**X**	1.13*X	1.08*X	1.08*X	1.07*X	1.06*X
11	Average time in the queue before skip (entry queue)	**Y**	1.58*Y	2.10*Y	2.28*Y	2.64*Y	2.79*Y
12	Average time in the queue before skip (metrology)	**Z**	1.76*Z	1.53*Z	1.61*Z	1.70*Z	1.99*Z

Table 6.11: Impact of $T_{Max} \in [0,0.5\%]$.

By analyzing results reported in Table 6.11 and Table 6.12, we note that the reduction in the number of lots that are skipped is rather constant between 0.1 and

	T_{Max}	0.5%	0.6%	0.7%	0.8%	0.9%	1%
1	Number of sampled lots	1.27*A	1.23*A	1.20*A	1.17*A	1.16*A	1.12*A
2	Number of measured lots	0.98*A	0.98*A	0.98*A	0.98*A	0.98*A	0.98*A
3	Number of skipped lots	0.29*A	0.25*A	0.22*A	0.19*A	0.18*A	0.14*A
4	**Number of skipped lots (entry queue)**	**0.23*A**	**0.19*A**	**0.17*A**	**0.13*A**	**0.12*A**	**0.09*A**
5	Number of skipped lots (metrology)	0.06*A	0.06*A	0.05*A	0.06*A	0.06*A	0.05*A
6	Medium WAR (average)	0.30*B	0.30*B	0.31*B	0.31*B	0.31*B	0.31*B
7	Maximum WAR (average)	0.39*C	0.39*C	0.40*C	0.40*C	0.40*C	0.40*C
8	Number of wafers above WL	0.79*D	0.80*D	0.81*D	0.82*D	0.81*D	0.82*D
9	Number of wafers above IL	0.43*E	0.44*E	0.46*E	0.47*E	0.47*E	0.47*E
10	Average time spent in the queue before measurement	1.06*X	1.07*X	1.03*X	1.04*X	1.01*X	1.07*X
11	Average time in the queue before skip (entry queue)	2.79*Y	3.04*Y	3.27*Y	3.40*Y	3.98*Y	4.15*Y
12	Average time in the queue before skip (metrology)	1.99*Z	1.62*Z	1.80*Z	1.77*Z	1.65*Z	1.64*Z

Table 6.12: Impact of $T_{Max} \in [0.5,1\%]$.

1%. However, between 0 and 0.1%, there is a significant reduction of the number of skipped lots (4.33*A → 0.81*A). We thus performed additional simulations between 0 and 0.1% (Table 6.13) to understand when T_{Max} starts impacting the number of skipped lots. Results in Table 6.13 show that, as soon as there is a minimal gain ($T_{Max} = 0.00005\%$) to satisfy before entering the queue, the number of lots can be strongly reduced without impacting too much the other indicators.

This value of T_{Max} will be strongly linked to the production environment. For example, in a production environment where lots are manually transported by operators in front of metrology tools, skipping a lot may be very expensive because of the time required to first transport the lot in front of metrology tools before transporting it to the next process operation. In this case, having a high T_{Max} value may be very interesting to avoid skipping lots that enter the metrology queue without impacting the sampling performances through the GSI sampling algorithm. **For the case of the 300mm fab of STMicroelectronics where transportation is automated, a value of $T_{Max} = 1\%$ has been identified as a good trade-off between the number of skipped lots and the other performance indicators.**

	T_{Max}	0%	0.00005%	0.0005%	0.005%	0.05%	0.1%
1	Number of sampled lots	5.34*A	5.11*A	4.24*A	3.02*A	2.07*A	1.84*A
2	Number of measured lots	0.98*A	0.98*A	0.98*A	0.98*A	0.98*A	0.98*A
3	Number of skipped lots	4.35*A	4.13*A	3.25*A	2.04*A	1.09*A	0.86*A
4	**Number of skipped lots (entry queue)**	**4.33*A**	**4.11*A**	**3.21*A**	**1.96*A**	**1.03*A**	**0.81*A**
5	Number of skipped lots (metrology)	0.02*A	0.02*A	0.04*A	0.08*A	0.06*A	0.05*A
6	Medium WAR (average)	0.24*B	0.24*B	0.24*B	0.24*B	0.27*B	0.28*B
7	Maximum WAR (average)	0.28*C	0.28*C	0.28*C	0.28*C	0.33*C	0.36*C
8	Number of wafers above WL	0.67*D	0.68*D	0.68*D	0.67*D	0.75*D	0.76*D
9	Number of wafers above IL	0.20*E	0.21*E	0.20*E	0.21*E	0.32*E	0.36*E
10	Average time spent in the queue before measurement	**X**	1.04*X	1.10*X	1.17*X	1.17*X	1.13*X
11	Average time in the queue before skip (entry queue)	**Y**	1.04*Y	1.25*Y	1.37*Y	1.49*Y	1.58*Y
12	Average time in the queue before skip (metrology)	**Z**	1.15*Z	1.51*Z	1.66*Z	1.64*Z	1.76*Z

Table 6.13: Impact of $T_{Max} \in [0,0.1\%]$.

b) Impact of parameter T_{Min}. T_{Min} has been introduced in the GSI sampling algorithms to master the number of lots that enter the metrology queue when the latter is empty. The aim is to ensure that, whatever the situation in production, or the size of the queue, measuring a lot always improves the situation.

Table 6.14 and Table 6.15 show the impact of T_{Min} when $T_{Max} = 1\%$. Note that increasing T_{Min} leads to reducing the number of measured lots (indicator 2). As larger and larger gains are required before entering the queue even when it is empty, metrology tools are no longer fully used. Consequences are the increase of the Medium WAR (indicator 6), the Maximum WAR (indicator 7), the number of wafers above WL (indicator 8), and the number of wafers above IL (indicator 9). However, differences are not so significant since the GSI sampling algorithm tries to minimize the overall risk by selecting the best possible lots. The average time before measurement (indicator 10) decreases because of the reduced number of lots that are sampled and measured. The time before skip (at the entrance of the queue and after a measurement is completed) decreases or increases depending on the number of lots that are skipped (see indicators 3 and 4). This is because of the value of $T_{Max} = 1\%$ that strongly reduces the number of skipped lots. As T_{Max} and T_{Min} are used in combination with the size of the queue[8], there is not direct link between

[8]When the queue is full, T_{Max} is used. When the queue is empty, T_{Min} is used. When the queue

the number of skipped lots and the value of T_{Min}.

	T_{Min}	0%	0.1%	0.2%	0.3%	0.4%	0.5%
1	Number of sampled lots	1.12*A	0.75*A	0.70*A	0.61*A	0.61*A	0.58*A
2	**Number of measured lots**	**0.98*A**	**0.62*A**	**0.56*A**	**0.50*A**	**0.48*A**	**0.46*A**
3	Number of skipped lots	0.14*A	0.13*A	0.14*A	0.11*A	0.13*A	0.12*A
4	Number of skipped lots (entry queue)	0.09*A	0.10*A	0.10*A	0.08*A	0.10*A	0.09*A
5	Number of skipped lots (metrology)	0.05*A	0.03*A	0.04*A	0.03*A	0.03*A	0.03*A
6	Medium WAR (average)	0.31*B	0.35*B	0.38*B	0.40*B	0.42*B	0.42*B
7	Maximum WAR (average)	0.40*C	0.44*C	0.48*C	0.51*C	0.53*C	0.53*C
8	Number of wafers above WL	0.82*D	0.92*D	0.96*D	0.99*D	1.02*D	1.03*D
9	Number of wafers above IL	0.47*E	0.60*E	0.67*E	0.72*E	0.76*E	0.80*E
10	Average time spent in the queue before measurement	**X**	0.74*X	0.69*X	0.68*X	0.68*X	0.67*X
11	Average time in the queue before skip (entry queue)	**Y**	1.02*Y	0.87*Y	0.99*Y	0.89*Y	0.83*Y
12	Average time in the queue before skip (metrology)	**Z**	1.1*Z	1.16*Z	0.85*Z	0.94*Z	1.01*Z

Table 6.14: Impact of $T_{Min} \in$ [0,0.5%].

To ensure the optimal use of metrology capacity, T_{Min} must be as small as possible, the number of skipped lots being mainly mastered by the value of T_{Max}. Nevertheless, T_{Min} can be very important in the case of an unavailability of a metrology tool that leads to a reduction of the metrology capacity.

For the case of the 300mm Fab of STMicroelectronics, the value of $T_{Min} = 0\%$ seems to be the most effective.

is partially filled, the threshold used is given by $Threshold = T_{Min} + \left[\frac{NBQ}{SQ} * (T_{Max} - T_{Min})\right]$ where NBQ is the number of lots in the metrology queue and SQ the metrology queue size (i.e. capacity).

	T_{Min}	0.5%	0.6%	0.7%	0.8%	0.9%	1%
1	Number of sampled lots	0.58*A	0.56*A	0.54*A	0.53*A	0.52*A	0.47*A
2	**Number of measured lots**	**0.46*A**	**0.45*A**	**0.44*A**	**0.41*A**	**0.41*A**	**0.38*A**
3	Number of skipped lots	0.12*A	0.11*A	0.10*A	0.12*A	0.11*A	0.08*A
4	Number of skipped lots (entry queue)	0.09*A	0.08*A	0.08*A	0.10*A	0.08*A	0.06*A
5	Number of skipped lots (metrology)	0.03*A	0.03*A	0.02*A	0.02*A	0.03*A	0.02*A
6	Medium WAR (average)	0.42*B	0.44*B	0.45*B	0.46*B	0.47*B	0.47*B
7	Maximum WAR (average)	0.53*C	0.55*C	0.56*C	0.57*C	0.58*C	0.58*C
8	Number of wafers above WL	1.03*D	1.05*D	1.07*D	1.08*D	1.08*D	1.10*D
9	Number of wafers above IL	0.80*E	0.83*E	0.85*E	0.88*E	0.90*E	0.92*E
10	Average time spent in the queue before measurement	0.67*X	0.64*X	0.60*X	0.61*X	0.66*X	0.64*X
11	Average time in the queue before skip (entry queue)	0.83*Y	1.04*Y	1.02*Y	0.70*Y	0.93*Y	0.98*Y
12	Average time in the queue before skip (metrology)	1.01*Z	0.89*Z	1.12*Z	0.94*Z	1.07*Z	0.98*Z

Table 6.15: Impact of $T_{Min} \in [0.5,1\%]$.

c) Impact of parameter T_{Metro}. T_{Metro} is the minimum gain that must be satisfied by lots to remain in the metrology queue each time a measurement is completed. The aim is to optimally use the metrology capacity by avoiding keeping in the queue lots that are covered by other lots (i.e. lots that bring less information than other lots). Indeed, our studies focused on defectivity controls where a control operation on lots may cover or provide information on several production tools. As lots are processed on different tools before arriving in front of metrology tools for control, the risk of having lots that bring approximatively the same information is increased with the number of products that are run concurrently. This is the case in the 300mm fab of STMicroelectronics where more than 200 products are run in production. This is why T_{Metro} was introduced.

To understand its impact on the GSI sampling performances, we performed simulations by varying T_{Metro} between 0 and 1%. T_{Min} and T_{Max} are set to 0% and 1% respectively. Table 6.16 and Table 6.17 show that T_{Metro} impacts the number of skipped lots after measurement (indicator 5). Increasing T_{Metro} leads to increasing the number of lots that are skipped (metrology). As larger gain is required to stay in the queue after each measurement, the number of lots that are skipped increases.

	T_{Metro}	0%	0.1%	0.2%	0.3%	0.4%	0.5%
1	Number of sampled lots	1.12*A	1.26*A	1.28*A	1.31*A	1.35*A	1.40*A
2	Number of measured lots	0.98*A	0.98*A	0.98*A	0.98*A	0.98*A	0.98*A
3	Number of skipped lots	0.14*A	0.27*A	0.30*A	0.33*A	0.37*A	0.42*A
4	Number of skipped lots (entry queue)	0.09*A	0.08*A	0.08*A	0.08*A	0.09*A	0.09*A
5	**Number of skipped lots (metrology)**	**0.05*A**	**0.19*A**	**0.22*A**	**0.25*A**	**0.28*A**	**0.33*A**
6	Medium WAR (average)	0.31*B	0.31*B	0.31*B	0.31*B	0.30*B	0.31*B
7	Maximum WAR (average)	0.40*C	0.40*C	0.39*C	0.40*C	0.40*C	0.40*C
8	Number of wafers above WL	0.82*D	0.80*D	0.80*D	0.81*D	0.80*D	0.81*D
9	Number of wafers above IL	0.47*E	0.46*E	0.46*E	0.46*E	0.46*E	0.47*E
10	Average time spent in the queue before measurement	**X**	0.84*X	0.82*X	0.79*X	0.76*X	0.72*X
11	Average time in the queue before skip (entry queue)	**Y**	0.97*Y	0.92*Y	0.90*Y	0.83*Y	0.76*Y
12	Average time in the queue before skip (metrology)	**Z**	0.76*Z	0.73*Z	0.70*Z	0.70*Z	0.70*Z

Table 6.16: Impact of $T_{Metro} \in [0,0.5\%]$.

	T_{Metro}	0.5%	0.6%	0.7%	0.8%	0.9%	1%
1	Number of sampled lots	1.40*A	1.40*A	1.43*A	1.45*A	1.49*A	*A
2	Number of measured lots	0.98*A	0.98*A	0.98*A	0.98*A	0.98*A	*A
3	Number of skipped lots	0.41*A	0.42*A	0.45*A	0.47*A	0.51*A	*A
4	Number of skipped lots (entry queue)	0.09*A	0.06*A	0.07*A	0.07*A	0.08*A	*A
5	**Number of skipped lots (metrology)**	**0.33*A**	**0.36*A**	**0.38*A**	**0.40*A**	**0.43*A**	***A**
6	Medium WAR (average)	0.31*B	0.31*B	0.31*B	0.30*B	0.30*B	*B
7	Maximum WAR (average)	0.40*C	0.39*C	0.40*C	0.39*C	0.39*C	*C
8	Number of wafers above WL	0.81*D	0.81*D	0.81*D	0.80*D	0.80*D	*D
9	Number of wafers above IL	0.47*E	0.46*E	0.47*E	0.45*E	0.45*E	*E
10	Time spent in the queue before measurement	0.72*X	0.73*X	0.70*X	0.70*X	0.68*X	0.68*X
11	Time spent in the queue before skip (entry queue)	0.76*Y	0.75*Y	0.60*Y	0.59*Y	0.53*Y	0.44*Y
12	Time spent in the queue before skip (metrology)	0.70*Z	0.65*Z	0.64*Z	0.63*Z	0.62*Z	0.63*Z

Table 6.17: Impact of $T_{Metro} \in [0.5,1\%]$.

However, contrary to T_{Min} and T_{Max} where the other performance indicators were impacted, T_{Metro} *does not* impact the Medium WAR, the Maximum WAR, the number of wafers above WL, and the number of wafers above IL. The time spent in the queue before measurement and skip are even improved. The problem is in the number of lots that are skipped. As skipping too many lots may lead to increasing the cycle time of some lots, a trade-off has to be found on the number of skipped

lots and the other performance indicators.

As discussed for the other indicators, the value of T_{Metro} is also linked to the production environment where the cost incurred by the skipping mechanism will not be the same in an automated manufacturing environment or not. **For the case of the 300mm fab of STMicroelectronics,** $T_{Metro} = 0\%$ **has been identified as the best trade-off because of the reduced number of skipped lots**.

6.5.3.3 Discussions and perspectives

The different simulations performed in this section helped us to assess the robustness of the GSI sampling algorithms that always ensure the minimum risk value within production whatever the parameter values. By successively varying the different parameters (α, β, T_{Max}, T_{Min}, T_{Metro}), results showed that sampling performances can be strongly improved, but that a trade-off is necessary between the different production objectives. There are no fixed values that can ensure *perfect* performances but the choice depends on the production strategy or priorities. Nevertheless, there are some values for which the GSI sampling algorithms may not be as efficient and thus these values must be carefully chosen. These values mainly concern α that must be lower than 13 and β lower than 10. The values of the other parameters are strongly linked to the production environment.

Two main points are perspectives for further research. The first point concerns the choice of the threshold parameters, and the second point is the anticipation of the arrival of lots.

1. **Threshold values**. In the simulations presented and discussed in this section, the value of *threshold* parameters (T_{Max}, T_{Min}, T_{Metro}) were defined in percentage (%). The problem of such an approach is that the gain that brings a lot strongly depends on the number of tools in production. In the case a lot covers two tools in a production environment with more than 300 tools, the gain of the lot (in %) will not be significant. If the production environment is only made of 50 tools, the gain of the lot will be significant, and so the priority on metrology tools. The sampling strategy or selection of lots will not be the same in the two situations. In the first case, the risk is to have too many tools exceeding their IL because of *under-estimated gains*. To ensure correct evaluations of gains brought by each lot, the threshold values should be expressed as a difference between a given situation in production and the new situation if the lot is measured. For example, the number of tools for which the lot will help avoid reaching or exceeding the IL. This approach was implemented and provides better results. However, it has not been done within the framework of this thesis and results are not presented in this manuscript.

2. **Anticipating the arrival of lots**. By defining different threshold values (T_{Max}, T_{Min}, T_{Metro}), we aimed at mastering the number of sampled and skipped lots without impacting the sampling performances. However, through simulations, we observed that if it is possible to strongly reduce the number of sampled/skipped lots without impacting the sampling performances, there is a trade-off between sampling lots with gains on the GSI and keeping lots already sampled and waiting in the metrology queue. To improve the sampling policy, the sampling of lots could be anticipated, i.e. not only sampling lots when they arrive at a metrology step, but also lots that are still being processed and will soon arrive at a metrology step. Some lots could be accelerated or some special actions taken. This will avoid sampling a lot that will be skipped later.

   Anticipating the arrival of lots might be modeled as a scheduling problem of jobs with release dates and with multiple objectives, where minimizing the risk and the waiting times of lots should be balanced. The resulting scheduling problem could be solved using a multi-objective approach such as the one proposed in **Dugardin *et al.*** [24]. This is an original scheduling problem in semiconductor manufacturing, not mentioned for example in **Mőnch *et al.*** [54].

6.6 Conclusion

In this chapter, we presented smart sampling policies based on two GSI sampling algorithms. These algorithms are used to dynamically sample, skip, and schedule lots on metrology tools. They are based on an indicator called GSI and on some *threshold* values. The GSI gives a weight to set of lots to be selected for inspection, and the *threshold* values are used to manage the filling of metrology queues with the aim of mastering both the number of sampled lots and skipped lots. The evaluation of the GSI sampling algorithms were performed through simulations that indicate a risk reduction of more than 70% compared to Fab sampling. The two GSI sampling algorithms outperform Fab sampling and the performance of each algorithm is linked to the production environement and management priorities.

In the next chapter, a Mixed-Integer Linear Programming model is proposed to optimize the WL and IL parameters that are used in the GSI sampling algorithms.

Chapter 7

Optimizing Smart Sampling Policies

This chapter[1] *introduces three versions of a Mixed-Integer Linear Programming (MILP) model that we developed to compute the values of two parameters: Warning Limit (WL) and Inhibit Limit (IL). These two parameters are used in the sampling algorithms introduced in the previous chapter. They represent the level of the risks that may be expected by a company depending on the available metrology capacity. By varying these values in a sampling policy, results indicate that the average risk level can be strongly impacted. By using the values of Warning Limit and Inhibit Limit obtained with our MILP model, results show an overall risk reduction without additional metrology capacity.*

7.1 *Introduction*

7.2 *Warning Limit and Inhibit Limit*

7.3 *Analyzing the impact of the Warning Limit and Inhibit Limit*

7.4 *Mixed-Integer Linear Programming model 1*

7.5 *Mixed-Integer Linear Programming model 2*

7.6 *Mixed-Integer Linear Programming model 3*

7.7 *Numerical experiments*

7.8 *Conclusion*

[1] Part of this chapter is submitted for publication in the **International Journal of Production Economics** [65].

7.1 Introduction

The efficiency of an algorithm is directly linked to the quality of input parameters. If these parameters are not optimally set, the strategy, mechanism, or algorithm can lead to poor results. In the previous chapter, we proposed sampling algorithms that use two main parameters (WL and IL) to dynamically select the best sets of lots to measure. In this chapter, we assess the impact of these two parameters on the efficiency of the algorithms, and propose a MILP model to optimize them.

The chapter is structured as follows. Section 7.2 introduces and defines WL and IL. Section 7.3 analyzes, through simulations, the impact of WL and IL on a sampling policy. In section 7.4, we present the MILP model we developed to optimally compute the values of WL and IL. Section 7.5 and section 7.6 present two improved versions of our MILP model that integrates additional constraints linked to the production environment. Section 7.7 is devoted to numerical experiments. We assess the performance of the results of the MILP model on the sampling algorithm introduced in the previous chapter. Section 7.8 concludes the chapter and gives perspectives for further works.

7.2 Warning Limit and Inhibit Limit

WL is the limit above which the situation starts to become critical in term of control. IL is the limit above which production tools might be stopped if a control is not performed. In term of wafers, IL represents the maximum number of wafers that can be run between two controls, and WL is the number of wafers that needs to be run on a production tool before increasing the priority of the tool for a control. IL is the maximum risk that can be tolerated, and WL is an alarm that helps avoiding to reach and exceed IL. WL and IL are defined per production tool and the objective is twofold:

- Dynamically sample, skip, and schedule lots on metrology tools while ensuring a maximum risk level lower than IL.

- Reduce the number of measurements when the risk level is lower than WL, and increase the priority of measurable lots when the risk level is closer to IL.

Let us consider the example in Figure 7.1. We have the evolution of the W@R for a given production tool. When lots are processed on the production tool, the W@R is increased of the number of wafers contained in the lot (Equipment W@R). When a control is performed, the W@R is decreased of the number of wafers processed on the production tool since the last control and before the process of the lot that has been measured (W@R reduction). WL and IL are set to A and $2*A$ respectively. They help identifying and avoiding the two cases of *Lack-of-control* and *Too-many-controls*. There is *Lack-of-control* when the value of the W@R exceeds the value of IL, and *Too-many-controls* when controls are performed whereas the value of the W@R is very far from IL.

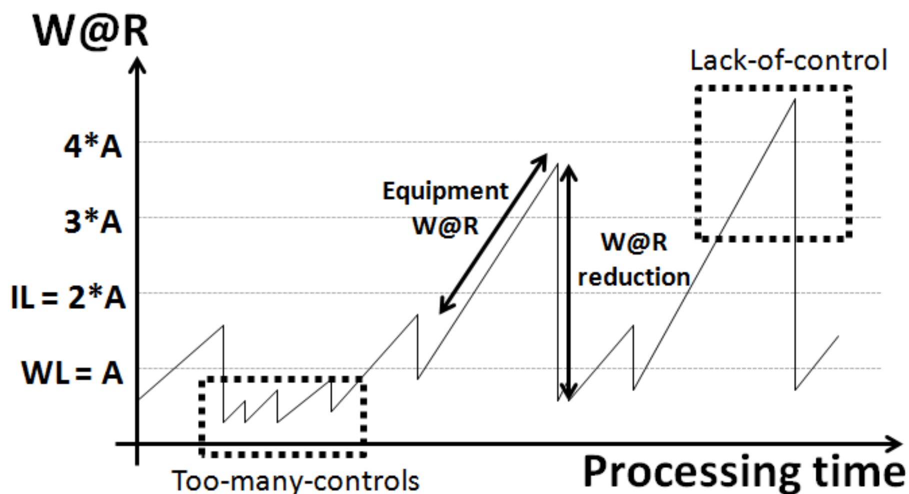


Figure 7.1: Evolution of the W@R on a production tool.

Defining WL and IL helps ensuring optimized sampling plan policies. However, the problem is that these values are most of the time set based on the experience of engineers or on historical data analysis. Since the sampling algorithm has to consider these parameters to prioritize lots on metrology tools, it is clear that if these values are over- or under-estimated, the efficiency of the sampling algorithm

may be impacted. For example, if these values are under-estimated, the risk may be to prioritize too many lots to avoid exceeding the value of IL. However, as the metrology capacity is limited, consequences will be the increasing of the cycle time of these lots that will be sample but never measured. To deeply analyze the role of WL and IL, and understand how they may impact a sampling policy, we perform several sampling policy simulations using different values of WL and IL. The next section presents and discusses the different results obtained.

7.3 Analyzing the impact of the Warning Limit and Inhibit Limit

We run simulations using six weeks of historical data[2]. With different values of WL and IL, we simulate a sampling policy that uses the first GSI sampling algorithm (Section 6.4.2) with the S5 simulator (Appendix C). For all the production tools, we start by defining WL and IL to 1000 and 2000 respectively. Then, by varying these values, we analyze the following indicators:

- Number of lots that are sampled,
- Number of lots that are measured,
- Number of lots that are skipped,
- Average medium W@R,
- Average maximum W@R.

Table 7.1 presents experimental results when WL and IL are varied together. Note that, depending on the values of WL and IL, the results of the performance indicators are different. Let us focus on the first three indicators: Number of sampled lots, number of measured lots, and number of skipped lots. These three indicators correspond to A, B, and C when $WL = 1000$ and $IL = 2000$. Considering these latter values as a reference, let us analyze what happens in the case of over- or under-estimation.

When WL and IL are over-estimated (+5%, +10%, +20%, etc.), the number of lots that are sampled (0.94*A, 0.90*A, 0.79*A, etc.) decreases. This is because the maximum risk that is tolerated becomes very large. As long as WL and IL are not reached, the situation is supposed to be under control, and thus only few lots are sampled. This reduction of the number of sampled lots leads to a reduction of the number of skipped lots (0.91*C, 0.85*C, 0.68*C, etc.) since the metrology capacity

[2]Part of this section has been communicated to the 4th **International Conference on Industrial Engineering and Systems Management** [63].

remains constant (**B**). The GSI sampling algorithm ensures that the metrology capacity is always used even if the maximum risk that is tolerated is very far from the actual risk. Consequences are the waste of metrology capacity due to the measurement of lots that do not bring any added value. This can be seen in the values of the Medium (D, D, 0.99*D, 1.01*D, D, 1.02*D, **1.02*D**) and Maximum W@R (E, 0.99*E, E, 1.01*E, 1.01*E, 1.01*E, **1.03*E**) that tend to increase. Note that if the variations of the Medium and Maximum W@R seem to be negligible, the impact can be significant when dynamically sampling lots. Indeed, in our experiments, we use 6 weeks of historical data. This means that the Medium and Maximum W@R reported in Table 7.1 are based on 6 weeks of historical data. As the GSI sampling algorithm aims at minmizing the overall risk within the production, this explains why we do not have too much variations of the Medium and Maximum W@R.

Variation of the WL and IL values	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
-90% ⇒ (100, 200)	1.38*A	B	1.59*C	1.50*D	1.62*E
-80% ⇒ (200, 400)	1.41*A	B	1.63*C	1.36*D	1.47*E
-60% ⇒ (400, 800)	1.39*A	B	1.59*C	1.21*D	1.28*E
-40% ⇒ (600,1200)	1.27*A	B	1.41*C	1.03*D	1.07*E
-20% ⇒ (800,1600)	1.21*A	B	1.33*C	1.02*D	1.05*E
-10% ⇒ (900,1800)	1.14*A	B	1.21*C	1.02*D	1.02*E
-5% ⇒ (950,1900)	1.09*A	B	1.14*C	1.02*D	1.02*E
(WL,IL) = (1000,2000)	**A**	**B**	**C**	**D**	**E**
+5% ⇒ (1050,2100)	0.94*A	B	0.91*C	D	0.99*E
+10% ⇒ (1100,2200)	0.90*A	B	0.85*C	D	E
+20% ⇒ (1200,2400)	0.79*A	B	0.68*C	0.99*D	0.99*E
+40% ⇒ (1400,2800)	0.68*A	B	0.50*C	1.01*D	1.01*E
+60% ⇒ (1600,3200)	0.61*A	B	0.40*C	D	1.01*E
+80% ⇒ (1800,3600)	0.58*A	B	0.36*C	1.02*D	1.01*E
+90% ⇒ (1900,3800)	0.55*A	B	0.31*C	1.02*D	1.03*E

Table 7.1: Impact of the WL and IL values on the sampling plan policy.

When WL and IL are under-estimated (-5%, -10%, -20%, etc.), the situation corresponds to an over-estimation of the metrology capacity. Note that the number of sampled lots (1.09*A, 1.14*A, 1.21*A, etc.) increases. This is because the maximum risk that is tolerated becomes smaller and smaller. Hence the necessity to sample more and more lots in order to reduce the risk level. However, as the metrology capacity is limited (**B**), the more lots are sampled, the more the number

of skipped lots (1.14*C, 1.21*C, 1.33*C, etc.). There is thus a negative impact on the cycle time for these lots that are sampled but never measured. Moreover, the overall risk is increased because of the inability of the sampling algorithm to take relevant decisions (see the Medium and Maximum W@R).

In order to better understand the impact of WL as well as IL, we vary them separately. In Table 7.2, we vary WL (first column) while keeping IL constant. In Table 7.3, we vary IL (first column) while keeping WL constant.

Variation of the **WL** value	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
-90% of **WL** ⇒ (**100**,2000)	1.32*A	B	1.49*C	1.08*D	1.10*E
-80% of **WL** ⇒ (**200**,2000)	1.33*A	B	1.50*C	1.05*D	1.08*E
-60% of **WL** ⇒ (**400**,2000)	1.29*A	B	1.44*C	1.02*D	1.05*E
-40% of **WL** ⇒ (**600**,2000)	1.20*A	B	1.31*C	1.01*D	1.03*E
-20% of **WL** ⇒ (**800**,2000)	1.10*A	B	1.15*C	D	1.01*E
-10% of **WL** ⇒ (**900**,2000)	1.05*A	B	1.08*C	D	E
-5% of **WL** ⇒ (**950**,2000)	1.04*A	B	1.06*C	D	1.01*E
(WL,IL) = (1000,2000)	**A**	**B**	**C**	**D**	**E**
+5% of **WL** ⇒ (**1050**,2000)	0.96*A	B	0.94*C	D	E
+10% of **WL** ⇒ (**1100**,2000)	0.96*A	B	0.93*C	D	0.99*E
+20% of **WL** ⇒ (**1200**,2000)	0.90*A	B	0.84*C	1.01*D	1.01*E
+40% of **WL** ⇒ (**1400**,2000)	0.78*A	B	0.66*C	1.02*D	E
+60% of **WL** ⇒ (**1600**,2000)	0.65*A	B	0.46*C	1.01*D	1.01*E
+80% of **WL** ⇒ (**1800**,2000)	0.59*A	B	0.36*C	1.04*D	1.02*E
+90% of **WL** ⇒ (**1900**,2000)	0.55*A	B	0.30*C	1.05*D	1.03*E

Table 7.2: Impact of the WL value on the sampling plan policy.

Results reported in Table 7.2 shows that the value of WL may impact the performance of the sampling policy. The same kinds of observations as in Table 7.1 can be made regarding the over- or under-estimation of WL, i.e. impact on cycle time, instability, and inefficiency of the sampling algorithm. However, contrary to Table 7.1 where WL and IL are varied together, Table 7.2 shows that the value of WL does not impact too much the Medium and Maximum W@R. This because WL is just a alarm that indicates that the situation starts to become critical. As long as the value of IL is far from the actual risk, the sampling algorithm is still able to take relevant decisions regarding the best lots to measure for reducing the risk. The

main problem is in the cycle time of lots that are sampled because of the WL, but never measured because of the IL.

In Table 7.3, we fix WL and analyze the impact of IL. We do not vary IL below -40% since WL must be lower than IL. Varying IL below -40% would lead to situations where the value of WL will be higher than the value IL, which does not make sense.

Variation of the **IL** value	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
-40% of **IL** ⇒ (1000,**1200**)	1.01*A	B	1.02*C	1.04*D	1.08*E
-20% of **IL** ⇒ (1000,**1600**)	1.13*A	B	1.21*C	1.01*D	1.03*E
-10% of **IL** ⇒ (1000,**1800**)	1.08*A	B	1.13*C	1.01*D	1.01*E
-5% of **IL** ⇒ (1000,**1900**)	1.06*A	B	1.09*C	1.02*D	1.02*E
(WL,IL) = (1000,2000)	**A**	**B**	**C**	**D**	**E**
+5% of **IL** ⇒ (1000,**2100**)	0.99*A	B	0.98*C	D	E
+10% of **IL** ⇒ (1000,**2200**)	0.96*A	B	0.93*C	D	E
+20% of **IL** ⇒ (1000,**2400**)	0.93*A	B	0.90*C	D	E
+40% of **IL** ⇒ (1000,**2800**)	0.86*A	B	0.79*C	0.99*D	1.01*E
+60% of **IL** ⇒ (1000,**3200**)	0.86*A	B	0.79*C	0.99*D	1.02*E
+80% of **IL** ⇒ (1000,**3600**)	0.85*A	B	0.77*C	0.99*D	1.03*E
+90% of **IL** ⇒ (1000,**3800**)	0.85*A	B	0.77*C	D	1.03*E

Table 7.3: Impact of the IL value on the sampling plan policy.

As in Table 7.1 and Table 7.2, over- or under-estimating the value of IL has an impact on all the performance indicators, i.e. the number of lots sampled, skipped, Medium W@R, and Maximum W@R. However, results reported in Table 7.3 show that if the value of IL is very close to the value of WL, the whole system becomes instable. This is the case when the value of IL is under-estimated of -40% ($WL = 1000$ and $IL = 1200$). The number of sampled lots decreases instead of increasing, and the Medium and Maximum W@R are strongly increased compared to variations between -5% and -20%.

Through all of these experiments on varying WL and IL, note that if the values of WL and IL are not optimally set, the entire sampling strategy can be impacted and the resulting risk level can be significant. With the aim of minimizing the risk

level on entire fab while ensuring an optimal use of metrology tools, we propose in the next section a MILP model to compute optimally the two values of WL and IL for each production tool.

7.4 Mixed-Integer Linear Programming model 1

The MILP model presented in this section aims at minimizing the maximum exposure within the production, i.e. minimizing the maximum risk level that can be incurred when processing a lot on a production tool[3]. The approach consists in determining IL values by allocating controls to production tools such that the maximum risk is minimized. The values of WL are deduced from the IL values.

Parameters:

- E_t: Exposure for tool t (i.e. the financial cost for each wafer processed on a production tool t).
- V_t: Production volume on tool t.
- Pm_t: Time of a measurement to validate production tool t.
- K_{MAX}: Maximum number of measurements for any production tool.
- $CAPA$: Total capacity (given in time) for measurement.
- M: Number of production tools.

Variables:

- IL_t: Inhibit Limit of production tool t.
- d_t^k: Binary variable that is equal to 1 if the number of measurements for production tool t is k, 0 otherwise.
- E_{MAX}: Maximum exposure.

The MILP model is as follows:

$$Minimize\, E_{MAX} \tag{7.1}$$

[3]Part of this section has been communicated to the **13ème congrès de la société Française de Recherche Opérationnelle et d'Aide à la DÉcision (ROADEF)** [66].

Subject to:

$$E_{MAX} \geq E_t * IL_t \quad \forall t \in \{1 \dots M\}. \tag{7.2}$$

$$IL_t \geq \sum_{k=1}^{K_{MAX}} \frac{V_t}{k} * d_t^k \quad \forall t \in \{1 \dots M\}. \tag{7.3}$$

$$\sum_{k=1}^{K_{MAX}} d_t^k = 1 \quad \forall t \in \{1 \dots M\}. \tag{7.4}$$

$$\sum_{t=1}^{M} \sum_{k=1}^{K_{MAX}} Pm_t * k * d_t^k \leq CAPA. \tag{7.5}$$

$$IL_t \geq 0 \quad \forall t \in \{1 \dots M\}. \tag{7.6}$$

$$d_t^k \in \{0,1\} \quad \forall t \in \{1 \dots M\}, \quad \forall k \in \{1 \dots K_{MAX}\}. \tag{7.7}$$

$$E_{MAX} \geq 0. \tag{7.8}$$

Constraints 7.2 define the maximum exposure among all production tools, which is minimized in the objective function. Constraints 7.3 express that the IL of production tool t (IL_t) is larger than or equal to the production volume on t divided by the selected number of measurements for t. Constraints 7.4 specify the number of measurements for the production tool t, i.e. that one and only one variable must be equal to 1. Constraint 7.5 ensures that the measurement capacity is satisfied.

This MILP model is evaluated in Section 7.7.1. Results indicate an optimized sampling plan policy without additional metrology capacity. However, this first MILP version presents some limitations that have been pointed out during simulations. **The delay or traveling time between production and metrology tools has not been taken into account**. Indeed, depending on the process type or production state, a control operation can be performed either directly after the process operation, five hours later, or sometimes one or two days after the process. Moreover, the organization of the clean room is such that the distance between production and metrology tools is not always the same. The availability or qualification of production tools as well as metrology tools can also increase or reduce the traveling time between tools. Hence the necessity of defining a kind of *average delay* between tools. This delay can be expressed as an average time required before obtaining the result of a control operation, or as an average number of wafers that are processed on a production tool between the time a decision is taken to

perform a control, and the time the control is actually performed on a metrology tool. The next section presents a second version of our MILP model that integrates this average delay between production and metrology tools.

7.5 Mixed-Integer Linear Programming model 2

The MILP model 2 is an evolution of the MILP model 1 where the main modification concerns the delay or traveling between production and metrology tools. This delay is defined as WD_t and corresponds to the number of wafers that are processed on the production tool t between the end of a process operation on t and the control of this process operation on a metrology tool. It is expressed as $WD_t = TH_t * CT_t$, where:

- TH_t = Throughput of the production tool t.
- CT_t = Average cycle time of lots that are processed on the production tool t between the end of the process operation on t and the end of the control operation on a metrology tool.

The new version of the MILP model is as follows:

$$Minimize\, E_{MAX} \tag{7.9}$$

Subject to:

$$E_{MAX} \geq E_t * IL_t \quad \forall t \in \{1 \ldots M\}. \tag{7.10}$$

$$IL_t \geq \sum_{k=1}^{K_{MAX}} \frac{V_t}{k} * d_t^k + WD_t \quad \forall t \in \{1 \ldots M\}. \tag{7.11}$$

$$\sum_{k=1}^{K_{MAX}} d_t^k = 1 \quad \forall t \in \{1 \ldots M\}. \tag{7.12}$$

$$\sum_{t=1}^{M} \sum_{k=1}^{K_{MAX}} Pm_t * k * d_t^k \leq CAPA. \tag{7.13}$$

$$IL_t \geq 0 \quad \forall t \in \{1 \ldots M\}. \tag{7.14}$$

$$d_t^k \in \{0,1\} \quad \forall t \in \{1 \ldots M\}, \quad \forall k \in \{1 \ldots K_{MAX}\}. \tag{7.15}$$

$$E_{MAX} \geq 0. \tag{7.16}$$

The difference with the MILP model 1 is in the constraints 7.11 that express that the IL of production tool t (IL_t) is larger than or equal to the production volume on t divided by the selected number of measurements for t, plus the average delay necessary to control the production tool t.

This new version of our MILP model is evaluated in Section 7.7.2. Results show an improved sampling policy compared to the first version of our MILP model. However, as the first version, this second version of our MILP model also presents some limitations that have been highlighted when coming to the industrial implementation. The model does not include metrology tools qualifications and capabilities.

All control operations cannot be performed by all metrology tools. To each metrology tool is associated a group of control operations. These control operations are defined based on the capabilities of the metrology tools and, the time of a control operation is linked to the metrology tool. This means that the metrology capacity is not consumed in the same way depending on the metrology tool or the set of metrology tools to be used. The same for the delay or traveling time between production and metrology tools. Depending on the metrology tool availability or qualifications, the time between process and control operations is not the same. Hence the necessity of defining different group of metrology tools, with different capacity, and different delays between tools.

The next section presents a third version of the MILP model that integrates these new parameters.

7.6 Mixed-Integer Linear Programming model 3

The MILP model 3 is an evolution of the MILP model 2 and integrates the additional following parameters:

- D: Number of groups (one per capability) of metrology tools.
- $CAPA^d$: Total capacity (given in time) for measurement for the group of metrology tools with capability d.
- $Pm_{t,d}$: Time of a measurement on a metrology tool with capability d to validate production tool t.
- $WD_{t,d}$: Average delay between a process operation (performed on a production tool t) and the control operation performed on a metrology tool that belongs to the group of tools with capability d.

The binary variable d_t^k becomes $d_{t,d}^k$ that is equal to 1 if the number of measurements on metrology tools with capability d to validate the production tool t is k, 0 otherwise.

Model 3 is as follows:

$$Minimize\, E_{MAX} \tag{7.17}$$

Subject to:

$$E_{MAX} \geq E_t * IL_t \qquad \forall t \in \{1 \dots M\}. \tag{7.18}$$

$$IL_t \geq \frac{V_t}{\sum_{d=1}^{D} \sum_{k=0}^{K_{MAX}} (k * d_{t,d}^k)} + \frac{\sum_{d=1}^{D} \sum_{k=0}^{K_{MAX}} (WD_{t,d} * k * d_{t,d}^k)}{\sum_{d=1}^{D} \sum_{k=0}^{K_{MAX}} (k * d_{t,d}^k)} \qquad \forall t \in \{1 \dots M\}. \tag{7.19}$$

$$\sum_{k=0}^{K_{MAX}} d_{t,d}^k = 1 \qquad \forall t \in \{1 \dots M\}, \ \forall d \in \{1 \dots D\}. \tag{7.20}$$

$$\sum_{t=1}^{M} \sum_{k=0}^{K_{MAX}} Pm_{t,d} * k * d_{t,d}^k \leq CAPA^d \qquad \forall d \in \{1 \dots D\}. \tag{7.21}$$

$$IL_t \geq 0 \qquad \forall t \in \{1 \dots M\}. \tag{7.22}$$

$$d_{t,d}^k \in \{0,1\} \qquad \forall t \in \{1 \dots M\}, \ \forall k \in \{1 \dots K_{MAX}\}, \ \forall d \in \{1 \dots D\}. \tag{7.23}$$

$$E_{MAX} \geq 0. \tag{7.24}$$

Constraints 7.19 express that the IL of production tool t (IL_t) is larger than or equal to the production volume on t divided by the selected number of measurements for production tool t on all the D groups of metrology tools, plus the sum of delays necessary to control production tool t on the right group of metrology tools. Constraints 7.20 specify the number of measurements for production tool t on all the D groups of metrology tools, i.e. that one and only one variable $d_{t,d}^k$ must be equal to 1 for each production tool t and each group of metrology tools. Constraint 7.21 ensures that the measurement capacity in each group d of metrology tools is satisfied.

This new model that integrates additional industrial constraints is no longer linear. The problem comes from **Constraints 7.19** and especially from the term $\sum_{d=1}^{D} \sum_{k=0}^{K_{MAX}} (k * d_{t,d}^k)$ in the denominator. To linearize our model, we define new variable dt_t^k such that:

$$dt_t^k = \frac{1}{\sum_{d=1}^{D} d_{t,d}^k}. \tag{7.25}$$

where $dt_t^k = 1$ if the number of measurements on the production tool t is k, and 0 otherwise.

Constraints 7.19 can be rewritten as:

$$IL_t \geq V_t * \sum_{k=0}^{K_{MAX}} \frac{1}{k} * \frac{1}{\sum_{d=1}^{D} d_{t,d}^k} + \sum_{d=1}^{D} \sum_{k=0}^{K_{MAX}} (WD_{t,d} * k * d_{t,d}^k) * \sum_{k=0}^{K_{MAX}} \frac{1}{k} * \frac{1}{\sum_{d=1}^{D} d_{t,d}^k} \quad \forall t \in \{1 \dots M\}. \tag{7.26}$$

By replacing 7.25 in 7.26, we obtain:

$$IL_t \geq V_t * \sum_{k=1}^{K_{MAX}} \frac{1}{k} * dt_t^k + \sum_{d=1}^{D} \sum_{k=0}^{K_{MAX}} (WD_{t,d} * k * d_{t,d}^k) * \sum_{k=1}^{K_{MAX}} \frac{1}{k} * dt_t^k \quad \forall t \in \{1 \dots M\}. \tag{7.27}$$

With

$$\sum_{k=1}^{K_{MAX}} dt_t^k = 1 \quad \forall t \in \{1 \dots M\}. \tag{7.28}$$

$$dt_t^k \in \{0,1\} \quad \forall t \in \{1 \dots M\}, \quad \forall k \in \{1 \dots K_{MAX}\}. \tag{7.29}$$

The model is still **non linear** because of the term $d_{t,d}^k$ in Constraints 7.27. We again introduce a new variable $dd_{t,d}^{k,k_1}$ such that:

$$dd_{t,d}^{k,k_1} = d_{t,d}^{k_1} * dt_t^k \tag{7.30}$$

where $dd_{t,d}^{k,k_1} = 1$ if the number of measurements to validate production tool t is equal to k_1 on group d of metrology tools, and k on all metrology tools in all groups of tools, and 0 otherwise.

Constraints 7.27 can be rewritten as:

$$IL_t \geq V_t * \sum_{k=1}^{K_{MAX}} \frac{1}{k} * dt_t^k + \sum_{d=1}^{D} WD_{t,d} * \sum_{k=0}^{K_{MAX}} (k * d_{t,d}^k) * \sum_{k=1}^{K_{MAX}} \frac{1}{k} * dt_t^k \quad \forall t \in \{1 \dots M\}. \tag{7.31}$$

By defining parameter k_1 such that $k_1 \leq k$, Constraints 7.31 become:

$$IL_t \geq V_t * \sum_{k=1}^{K_{MAX}} \frac{1}{k} * dt_t^k + \sum_{d=1}^{D} WD_{t,d} * \sum_{k_1=0; k_1 \leq k}^{K_{MAX}} (k_1 * d_{t,d}^{k_1}) * \sum_{k=1}^{K_{MAX}} \frac{1}{k} * dt_t^k \quad \forall t \in \{1 \dots M\}. \tag{7.32}$$

Or

$$IL_t \geq V_t * \frac{1}{k} * dt_t^k + \sum_{d=1}^{D} WD_{t,d} * \sum_{k_1=0}^{k} (k_1 * d_{t,d}^{k_1}) * \frac{1}{k} * dt_t^k \quad \forall t \in \{1 \dots M\}, \forall k \in \{1 \dots K_{MAX}\}. \tag{7.33}$$

Or

$$IL_t \geq V_t * \frac{1}{k} * dt_t^k + \sum_{d=1}^{D} \sum_{k_1=0}^{k} \frac{k_1 * WD_{t,d}}{k} * (d_{t,d}^{k_1} * dt_t^k) \quad \forall t \in \{1 \dots M\}, \forall k \in \{1 \dots K_{MAX}\}. \tag{7.34}$$

By replacing 7.30 in 7.34, we obtain:

$$IL_t \geq V_t * \frac{1}{k} * dt_t^k + \sum_{d=1}^{D} \sum_{k_1=0}^{k} \frac{k_1 * WD_{t,d}}{k} * dd_{t,d}^{k,k_1} \quad \forall t \in \{1 \dots M\}, \forall k \in \{1 \dots K_{MAX}\}. \tag{7.35}$$

With

$$\sum_{k=1}^{K_{MAX}} dt_t^k = \sum_{d=1}^{D} \sum_{k=1}^{K_{MAX}} \sum_{k_1=0}^{k} dd_{t,d}^{k,k_1} \quad \forall t \in \{1 \dots M\}. \tag{7.36}$$

$$\sum_{d=1}^{D} \sum_{k=1}^{K_{MAX}} \sum_{k_1=1}^{k} dd_{t,d}^{k,k_1} = 1 \quad \forall t \in \{1 \dots M\}. \tag{7.37}$$

$$dd_{t,d}^{k,k_1} \in \{0,1\} \quad \forall t \in \{1 \dots M\}, \quad \forall d \in \{1 \dots D\}, \quad \forall t \in \{k_1 \dots k\}, \quad \forall k \in \{1 \dots K_{MAX}\}. \tag{7.38}$$

Therefore, the final version of MILP model 3 is:

$$Minimize\, E_{MAX} \tag{7.39}$$

Subject to:

$$E_{MAX} \geq E_t * IL_t \qquad \forall t \in \{1 \dots M\}. \tag{7.40}$$

$$IL_t \geq V_t * \sum_{k=1}^{K_{MAX}} \frac{1}{k} * dt_t^k + \sum_{d=1}^{D} \sum_{k=1}^{K_{MAX}} \sum_{k_1=0}^{k} \frac{k_1 * WD_{t,d} * dd_{t,d}^{k,k_1}}{k} \qquad \forall t \in \{1 \dots M\}. \tag{7.41}$$

$$\sum_{k=1}^{K_{MAX}} dt_t^k = 1 \qquad \forall t \in \{1 \dots M\}. \tag{7.42}$$

$$\sum_{d=1}^{D} \sum_{k=1}^{K_{MAX}} \sum_{k_1=0}^{k} dd_{t,d}^{k,k_1} = 1 \qquad \forall t \in \{1 \dots M\}. \tag{7.43}$$

$$\sum_{k=1}^{K_{MAX}} dt_t^k = \sum_{d=1}^{D} \sum_{k=1}^{K_{MAX}} \sum_{k_1=0}^{k} dd_{t,d}^{k,k_1} \qquad \forall t \in \{1 \dots M\}. \tag{7.44}$$

$$\sum_{t=1}^{M} \sum_{k=1}^{K_{MAX}} Pm_{t,d} * k * d_{t,d}^k \leq CAPA^d \qquad \forall d \in \{1 \dots D\}. \tag{7.45}$$

$$IL_t \geq 0 \qquad \forall t \in \{1 \dots M\}. \tag{7.46}$$

$$dt_t^k \in \{0, 1\} \qquad \forall t \in \{1 \dots M\}, \; \forall k \in \{1 \dots K_{MAX}\}. \tag{7.47}$$

$$dd_{t,d}^{k,k_1} \in \{0, 1\} \qquad \forall t \in \{1 \dots M\}, \; \forall d \in \{1 \dots D\}, \; \forall k_1 \in \{1 \dots k\}, \; \forall k \in \{1 \dots K_{MAX}\}. \tag{7.48}$$

$$E_{MAX} \geq 0. \tag{7.49}$$

This model, which seems to be most suitable for industrial implementation, can no longer be solved with a standard solver because of its complexity. It is thus necessary to develop dedicated methods to solve the model. This has not been done within the framework of this thesis. Nevertheless, a first approach and perspective could be to relax some constraints or develop specific heuristic.

7.7 Numerical experiments

We ran simulations using six weeks of historical data from the site of STMicroelectronics in Crolles, France. The different versions of our MILP model are solved with the commercial solver MP-XPRESS. Each version provides IL values (per production tool) that we use to deduce the values of WL ($WL = 0.5 * IL$). Then, with the S5 prototype (Appendix C), we simulate GSI sampling policies (Section 6.4.2) by using the WL and IL obtained with the MILP model. To assess the efficiency of the values obtained with our model, we perform comparisons with a sampling policy where WL and IL are set to 1000 and 2000 for all production tools [4]. We use the following indicators:

- Number of sampled lots,
- Number of skipped lots,
- Number of measured lots,
- Average Medium W@R,
- Average Maximum W@R,
- Number of wafers above WL,
- Number of wafers above IL.

7.7.1 Evaluating MILP model 1

The first version of the MILP model was run on a computer of 2.8GHz, 12GB of RAM, and with Windows 7 as the operating system. With 578,769 variables and 733 constraints, the computational time to obtain the optimal solution is 24,766 seconds.

Table 7.4 and Table 7.5 present the GSI sampling policy performances when the WL and IL computed with the MILP model are used. We ran the model with different values of the exposure. We aimed at analyzing and quantifying the impact of

[4] The values of WL = 1000 and IL = 2000 are representative of the production.

the tool criticality on the sampling policy. In Table 7.4, the performance indicators are related to the values of WL and IL obtained for an exposure value which is equal to 1 for all of the tools. In Table 7.5, the exposure is defined per workshop and per tool. The different values of exposure are based on historical data where we use the Medium W@R. The larger the Medium W@R, the lower the exposure value.

Table 7.4 presents a comparison of the sampling performances when WL = 1000 / IL = 2000, and when WL and IL are obtained with the MILP model for an exposure value of 1. Note that, with the same number of measured lots (C), all the performance indicators (except the Maximum W@R) are improved with the WL and IL computed with the MILP model. This means that, the GSI sampling algorithm is able to take relevant decisions regarding the lot to sample, skip, or measure. Only few lots are sampled and skipped for an improved Medium W@R value. The case of the Maximum W@R is explained by the production volume and exposure. Indeed, all of the production tools do not produce or process the same quantity of wafers. Some tools process twice or three times more wafers than others. As we define the same exposure value for all tools, the situation is such that controls are allocated in the *same* way for all tools. However, as the metrology capacity is limited, only a fixed number of lots can be measured and, thus, the overall risk level cannot be minimized for all tools. This is why there is an increased value for the average Maximum W@R.

Performance indicators	**WL = 1000 and IL = 2000**	**MILP-1 values** (exposure = 1)
Number of sampled lots	A	0.72*A
Number of skipped lots	B	0.56*B
Number of measured lots	C	C
Average Medium W@R	D	0.99*D
Average Maximum W@R	E	1.01*E
Number of wafers above WL	F	0.62*F
Number of wafers above IL	G	0.43*G

Table 7.4: Evaluating the WL and IL obtained with the MILP model 1 (Exposure=1 for all production tools).

The number of lots above WL and IL is also reduced but we still have lots pro-

cessed on production tools above IL ($0.43 * G$). This means that the GSI sampling algorithm cannot satisfy the WL and IL optimized by the MILP model. Therefore, although the sampling policy can be improved, defining the same exposure value for all production tools does not provide WL and IL that can ensure optimal sampling decisions with the GSI sampling algorithm. There is a need to adjust this value of exposure depending on the tool or set of tools.

Table 7.5 shows the sampling performances obtained for an exposure defined per workshop and per tool. Note that, when the exposure is defined per workshop, the number of lots above IL is divided by more than two ($0.18 * G$). When the exposure is defined per tool, this number is almost zero ($0.02 * G$). This means that the exposure must be defined per tool, and that the WL and IL provided by our MILP model help ensuring optimized sampling decisions. No lot is processed above IL, and the number of lots that are sampled and skipped is reduced. However, the values of the Medium and Maximum W@R are increased compared to the case where the exposure is set to 1 (Table 7.4). This can be explained as follows.

Performance indicators	**WL = 1000 and IL = 2000**	**MILP-1 (Exposure per workshop)**	**MILP-1 (Exposure per tool)**
Number of sampled lots	A	0.66*A	0.70*A
Number of skipped lots	B	0.47*B	0.53*B
Number of measured lots	C	C	C
Average Medium W@R	D	1.20*D	1.65*D
Average Maximum W@R	E	1.19*E	1.57*E
Number of wafers above WL	F	0.43*F	0.28*F
Number of wafers above IL	G	0.18*G	0.02*G

Table 7.5: Evaluating the WL and IL obtained with the MILP model 1 for different values of the exposure.

When the exposure is set to 1, the risk is minimized in the same way for all tools. This is why the Medium and Maximum W@R are decreasing. However, the problem of defining the same exposure for all tools is that a very bad or less critical tool may relax all the other tools. On the contrary, by defining an exposure per tool, we minimize the overall risk taking into account the criticality of production tools. The

overall risk is thus increased because of less critical tools for which we tolerate a high level of risk. This increased level of risk (Medium and Maximum W@R) may also be impacted by the distance between tools. As the delay or traveling time between process and metrology tools may vary depending on the process operation to be performed, or the process tool to be used, missing to consider such a parameter in the model may also explain the increased level of the Medium and Maximum W@R. This is why we proposed a second version of the MILP model to integrate the delay between tools. The next section presents numerical experiments related to this new version of the MILP.

7.7.2 Evaluating the MILP model 2

This second version of the MILP model was run on a computer of 2.8GHz, 12GB of RAM, and with Windows 7 as the operating system. With 578,769 variables and 733 constraints, the computational time to obtain the optimal solution is 24,766 seconds.

We ran the MILP model by defining an exposure per tool and an average delay between tools. This delay is expressed as an average number of wafers. Table 7.6 presents the GSI sampling performances obtained with the WL and IL computed with this new version of the MILP model. We compare the case where the delay is defined per workshop and the case where delay is defined per tool. Note that, for both cases, the number of lots above IL is equal to zero. This shows that the values or WL and IL provided by the second MILP model are optimized values since the GSI sampling algorithm is able to satisfy these limits. Moreover, the Medium and Maximum W@R are reduced compared to the first model where delays between tools are not taken into account (Table 7.5). The GSI sampling algorithm is able to select a reduced number of lots while minimizing the overall risk.

Using the WL and IL computed with the MILP model helps ensuring optimized GSI sampling decisions. However, the model presents some limitations that need to be highlighted. First, the model is based on historical data. The values of WL

Performance indicators	WL = 1000 and IL = 2000	MILP-2: Delay per workshop	MILP-2: Delay per tool
Number of sampled lots	A	0.44*A	0.46*A
Number of skipped lots	B	0.13*B	0.16*B
Number of measured lots	C	C	C
Average Medium W@R	D	0.96*D	1.15*D
Average Maximum W@R	E	1.04*E	1.23*E
Number of wafers above WL	F	0.02*F	0.02*F
Number of wafers above IL	G	0	0

Table 7.6: Evaluating the WL and IL obtained with MILP model 2 (delay defined per workshop and per tool).

and IL are thus only valid for some specific periods, i.e. when the mix of products or fab loading does not change too much. In the case of a significant change of the production volume for example, there will be a necessity to compute again these values. Second, the model computes WL and IL based on the actual control plan. It gives indication on the level of the risk (IL) that may be expected by the company, but do not provide solutions to reduce this expected level of risk. The GSI sampling algorithm tries to minimize the risk on the entire fab, but the mechanism is based on WL and IL that give indication on what can be expected. If WL and IL values are arbitrarily set, the GSI sampling policy will lead to poor results because of incoherent information. Therefore, the only way to improve or reduce the values of WL and IL computed with our MILP model is to work directly on the control plan, and try to define optimal positions of control operations. This is done within the framework of the PhD thesis of B. BETTAYEB [10] [11] and G. RODRIGUEZ-VERJAN [81].

The works of B. BETTAYEB mostly focus on allocating controls on production tools with the aim of minimizing the global exposure. Depending on the available metrology capacity and the mix of products, the goal is to define, for each production tools the number of control operations that can be performed using some predefined criteria. **The works of G. RODRIGUEZ-VERJAN** mostly focus on defining the right positions of control operations throughout production. The goal is to minimize the average delay or traveling time between control operations and thus reduce the W@R in production. The two works mainly address the modeling and definition of an optimal control plan whereas in my thesis I am interested in computing WL and IL using a predefined control plan.

7.8 Conclusion

In this chapter, we presented three versions of a MILP model we developed to optimize the values of two parameters: WL and IL. These two parameters are key in the GSI sampling policies introduced in Chapter 6. We firstly discussed the impact of these values on the sampling policy performances, and then analyzed their added value through simulations. By simulating sampling policies with different values of WL and IL, results indicate a direct impact on the sampling policy performances. The average risk level can be increased and cycle time of lots impacted. If WL and IL are arbitrarily set, the sampling mechanism becomes instable. By using the WL and IL obtained with our MILP model in a GSI sampling policy, results show an overall risk reduction on the entire Fab without additional metrology capacity.

However, the MILP model we propose is based on historical data and, whenever there is a significant change in production (e.g. production volume or fab loading), it will be necessary to update the different values. Moreover, when modeling all the parameters that may interact in a real-time industrial implementation, the model complexity has strongly increased. Therefore, one of the main perspectives is in working in the source of the problem instead of focusing on the consequences, i.e. identifying and defining optimized positions of control operations instead of computing the expected levels of risk based on a static control plan.

Chapter 8

Industrial Developments and Implementations

This chapter[1] *gives a general overview of specific solutions that have been proposed within the framework of this thesis: Prototypes and financial metrics. These specific solutions have supported the industrialization of the main concepts proposed and developed in the thesis.* ***This is one of the strengths of this thesis****.*

8.1 *Introduction*
8.2 *CMP-WAR Prototype*
8.3 *Excursion Management Prototype*
8.4 *Financial Metrics*
8.5 *Conclusion*

[1]Part of this section is accepted for publication in the proceedings of the 8th **International Conference on Modeling and Analysis of Semiconductor Manufacturing (included in the 2012 Winter Simulation Conference)** [62].

8.1 Introduction

The company culture, production constraints, and resource management in a high-mix manufacturing environment are such that, providing, developing, or deploying a new concept or algorithm requires getting in touch with everybody. On the one hand, even if a new software solution can contribute to increase yield or reduce cycle time, it will also often impact other activities, workload of engineers, or existing tools with consequences on product prices. On the other hand, the main objective of a company is to go forward by developing and finding new strategies to stay competitive in the market. This implies that each innovation within the company must be easily adopted and understood, otherwise it may be rejected. This is why, within the framework of this thesis, we focused on interacting as much as possible with experts in the company, and on validating our algorithms and solutions through simulations and prototypes that could be understood by everybody.

Section 8.2 presents a general overview of the CMP prototype introduced in Chapter 5. In section 8.3, we describe the prototype that has been developed for an optimized management of excursions using the IPC mechanism. Section 8.4 is devoted to the financial metrics we proposed to assess the return on investment of the dynamic sampling algorithms introduced in Chapter 6.

8.2 CMP-WAR Prototype

The CMP-WAR prototype described in Chapter 5 has been implemented using data coming from different databases. Figure 8.1 shows an overview of the different databases that are used and, Figure 8.2 the final user interface. Excel-VBA and SQL languages were used for the main developments.

The prototype has been deployed for in-line use in the CMP and defectivity workshops. During the evaluation phase, the prototype was used only twice a day by the engineering team. Figure 8.3 shows the first evaluation performed two weeks after the prototype was deployed. Note that the number of lots processed on production tools with a risk indicator larger than 0.33 (0.33 is the maximum risk allowed by

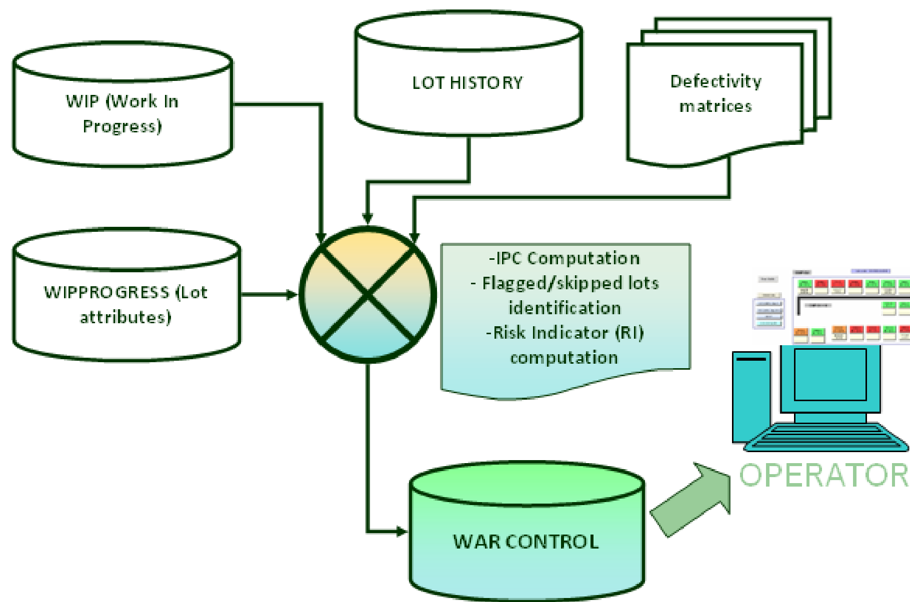


Figure 8.1: General schema – Databases.

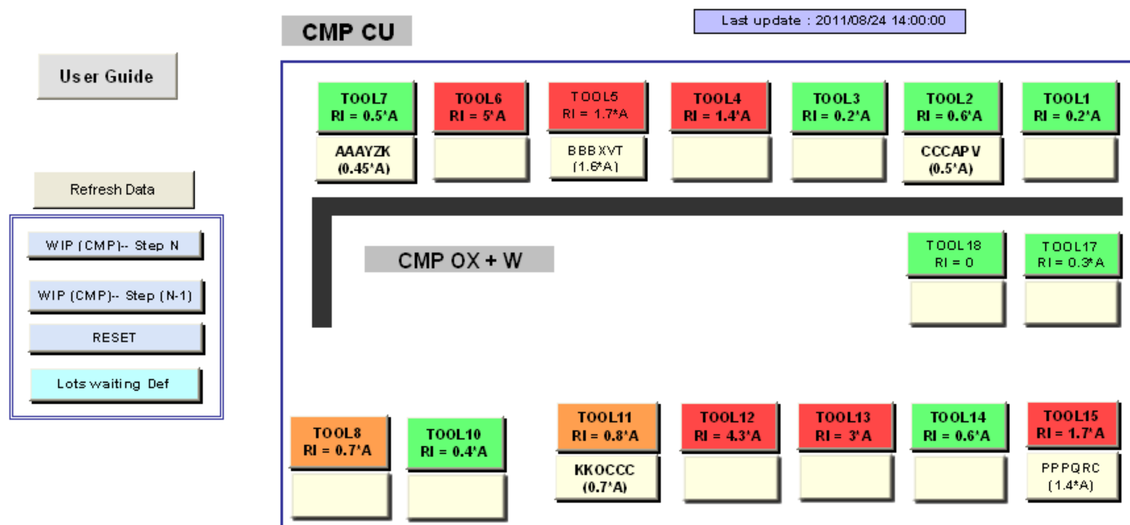


Figure 8.2: Overview of the CMP WAR prototype.

the company) was reduced by more than 65%. **0.33** represents the maximum risk allowed by the company.

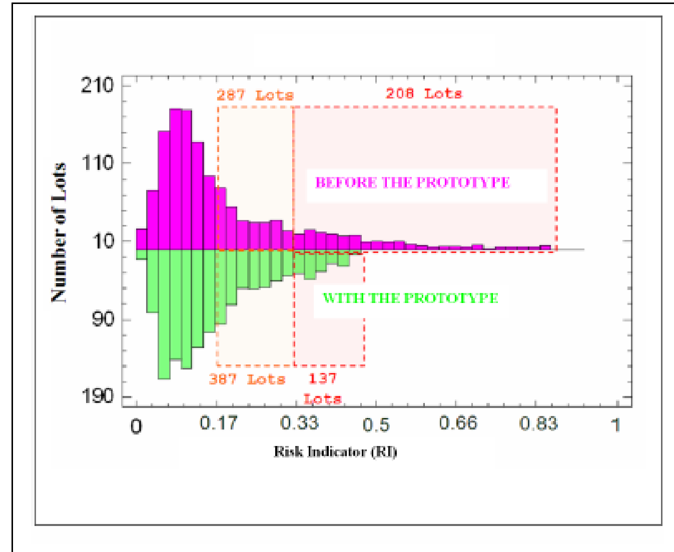


Figure 8.3: Results on global Risk Indicator (RI) reduction.

Figure 8.4 provides another evaluation of the prototype on several weeks. Note the strong impact of the prototype on the risk reduction. Note also that, during the holidays where the number of qualified operators is reduced, the risk significantly increases until the prototype was used again.

These encouraging results led to additional analysis and observations within the company. **After 6 months of evaluations and analysis, the decision was taken to industrialize the solution**.

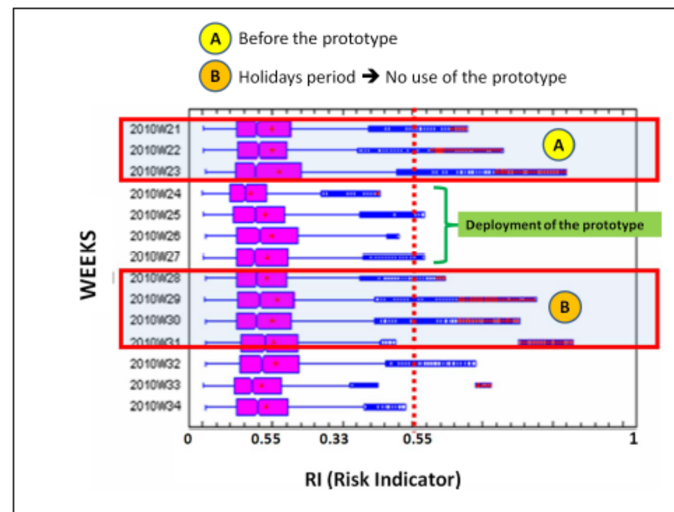


Figure 8.4: Impact of the prototype on the overall risk.

8.3 Excursion Management Prototype

This second prototype has been developed based on the same types of data than the CMP-WAR prototype. Figure 8.5 shows the final user interface that has been deployed in the defectivity workshop. By entering the name of a lot for which an excursion is detected, the prototype provides the set of tools that can be removed from the initial scope of analysis. This information is computed in real time using the IPC mechanism (see Chapter 5).

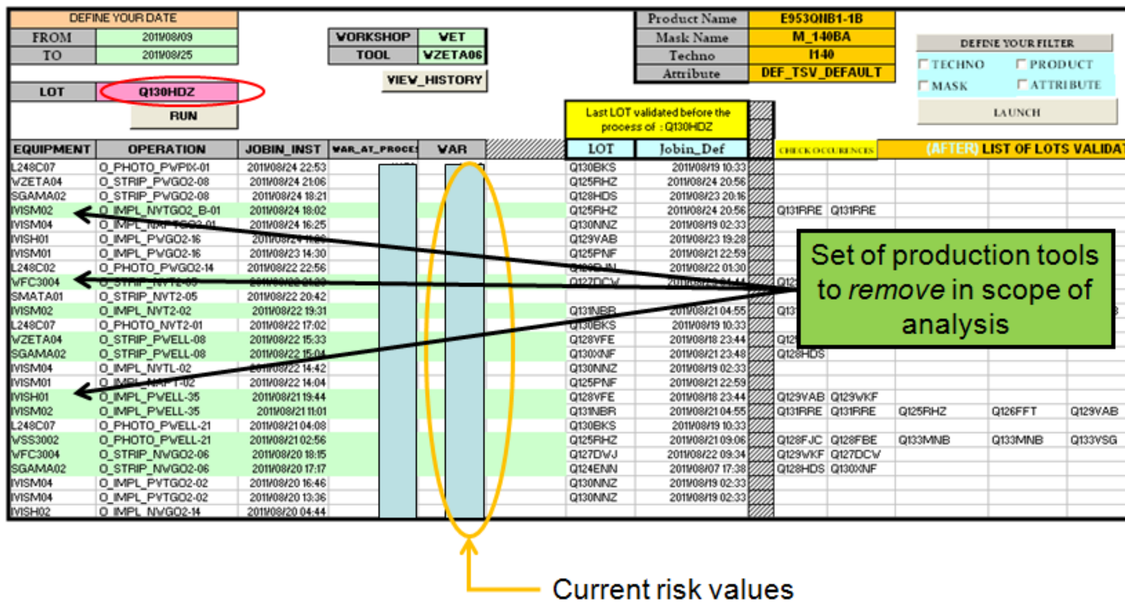


Figure 8.5: Overview of the Excursion Management prototype.

8.4 Financial Metrics

To assess the added value of the dynamic sampling algorithms introduced in Chapter 6, we proposed three different financial metrics, among which one was deemed to be more suitable for defectivity controls. However, depending on the fab or the type of the risk that is addressed, one metric may be more suitable than another. The three metrics are:

1. **The number of metrology tools that can be saved by using a GSI sampling algorithm**. The idea is to compute the number of metrology tools required with the GSI algorithm to obtain performance indicators that are as good as in current fab sampling. However, as our goal is not to reduce the number of measured lots but to reduce the number of measurements without added value, this first metric was not judged to be the most suitable.

2. **The downtime costs incurred when Inhibit Limits are exceeded**. The idea is to consider that a financial cost is incurred when a production tool is stopped because an Inhibit Limit is exceeded. The resulting downtime is a non

productive time which costs money. We assume that the production tool stays down until the $W@R$ becomes smaller than its corresponding Inhibit Limit. However, the problem of this second metric is that, in practice, a production tool will not always be stopped when its Inhibit Limit is exceeded. Therefore, this second metric was not chosen.

3. **The risk related costs incurred when Inhibit Limits are exceeded**. The idea is to consider that the risk of losing wafers is increased when the $W@R$ of a production tool is above its Inhibit Limit. Based on a probability of failure, it is possible to calculate how much money could be saved. **This last metric was deemed to be the most practical**.

Using the third metric, we introduce the following notations to compute the gains of the GSI sampling algorithms:

- $NW^t_{s,IL}$: Number of wafers above the Inhibit Limit (IL) with the fab sampling (*static sampling*) for production tool t.
- $NW^t_{d,IL}$: Number of wafers above the Inhibit Limit (IL) with the GSI sampling (*dynamic sampling*) for production tool t.
- P^t_w : Cost of a wafer above the Inhibit Limit (IL) for production tool t.
- G^t_w : Gain in term of term of risk reduction (number of wafers) for production tool t.
- PG_e : Potential financial gain in term of money (€). This value represents the global amount of money potentially saved with the GSI sampling.

Therefore, for a given production tool t, the gain in term of risk reduction (G^t_w) is given by:

$$G^t_w = \max(0, NW^t_{s,IL} - NW^t_{d,IL})$$

The potential financial gain (PG_e) is:

$$PG_e = \sum_t G^t_w * P^t_w * P_{loss}$$

where P_{loss} is the probability of losing wafers when the $W@R$ of a production tool is above its Inhibit Limit.

By considering the results in Table 6.8 (see Section 6.5.2), assuming a probability of losing wafers $P_{loss} = 1/2000$, and an average wafer cost of 1500 €, the potential financial gains are:

PG_e = (9,517,277 - 7,759,743)*(1500) * (1/2000) = 1,318,150 €.

This significant potential gain, combined with the prototype results, was one of the main drivers behind the industrialization of the GSI sampling algorithms.

8.5 Conclusion

In this chapter, we summarized the main industrial solutions that have been proposed within the framework of this thesis. We presented the two main prototypes that have supported the industrialization of the IPC mechanism, and financial metrics that helped in assessing the added value of the GSI sampling algorithms before a fab-wide industrialization. Several other prototypes have been developed and deployed in the fab, but we did not describe them in this manuscript since they are more or less based on the two main prototypes and algorithms described in the previous chapters.

General Conclusion and Perspectives

General Conclusion

Not everything that is faced can be changed. But nothing can be changed until it is faced[2]. In this thesis, we faced the problem of implementing dynamic control plans in semiconductor manufacturing. We analyzed the complexity of designing control plans, developed novel algorithms, and provided smart solutions to support the change from static to dynamic control plans. Our different algorithms have been validated through simulations and prototypes, before being industrialized within the 300mm fab of STMicroelectronics in Crolles, France. Some of the solutions proposed in this thesis have been used in other sites of STMicroelectronics. The question remains whether they can be extended to other types of industries or activities.

One of the important contribution of this thesis lies in the industrial implementation of dynamic control plans in a high-mix environment. Indeed, in such an environment, the complexity is such that, if some critical parameters (product types, tool specificities, production constraints, etc.) are not appropriately taken into account, a dynamic control plan can lead to very poor results and worsen the situation in production. This is why most of the solutions proposed in the literature are usually impracticable for an industrial deployment. The required investment, the resource management, or the huge amount of data to handle in real time are factors that lead companies to prefer static control plans whereas dynamic con-

[2] **James Arthur Baldwin**.

trol approaches have been shown to be more suitable. This thesis has offered new approaches and solutions to support industrial implementation of dynamic control plans, showing that it is the only way for modern companies to stay competitive by increasing the yield without impacting the cycle time.

This thesis was conducted within the framework of a joint collaboration between industry and academics. We thus started our research by modeling and understanding the various control plan approaches within the 300mm fab of STMicroelectronics. We focused on defectivity controls and especially on sampling techniques that aim at finding a trade-off between yield and cycle time. We stated our working hypothesis on the added value of controls. Observations helped us to understand the main drawbacks of static sampling that often lead to several cases of over- and lack-of-controls.

With the aim of generalizing our problem related to static sampling, we performed a literature review to classify our problem and analyze previously proposed solutions. We noted that dynamic sampling are more suitable for modern semiconductor plants, but that the efficiency of each solution or approach is directly linked to the production environment.

Once our problem was clearly understood, generalized, and classified, we proposed dynamic sampling algorithms that we validated through simulations and prototypes. To support industrial implementation of these dynamic sampling algorithms (GSI algorithms), we developed the IPC (Permanent Index per Context) indicator to handle a large amount of data with little CPU effort. The combination of both the IPC and the GSI sampling algorithms led to the industrial implementation of dynamic control plans whose potential gains was estimated to more than $1,000,000$ € and a return on investment of less than 6 months.

Perspectives

This thesis on dynamic control plans opened perspectives for further works. These perspectives concern **the optimized management of excursions** using

the IPC mechanism, and the implementation of **predictive sampling**.

Our first perspective is related to the **management of excursions**. By implementing dynamic sampling policies, lots or wafers are no longer stopped at all control steps. There is thus no quantification of defects whenever a problem occurs throughout production. The amount of data to analyze in order to contain the excursion strongly increases. As the IPC is efficient to handle a very large amount of data, we propose to formalize the problem using the concept of dominating sets and, based on the IPC, select the lot that covers the maximum number of lots in production. This will help in quickly reducing both the material at risk and the scope of analysis.

Figure 8.6 gives an overview of the concept of dominating sets. The aim is to contain as quickly as possible the excursion by selecting and measuring lots that release the uncertainty on other lots. For example, selecting **LB** help to release the uncertainty on lots **LJ** and **LK**, i.e. LB was processed after LJ and LK on the same production tools and/or with the same context[3]. Selecting **LH** help to release the uncertainty on the set of lots {LA, LB, LC, LD, LE, LF, LG, LT, LS}, and selecting LA provides information on the set {LJ, LK, LM, LN, LP, LQ, LO, LU, LS, LT, LW, LV}. As **LA** covers the largest number of lots, the priority would be given to measuring **LA**. If the control of LA does not reveal any problem, then the uncertainty can be *released* on the set of lots covered by **LA**. If a problem is detected on LA, then the set of lots covered by LA could be quickly stopped and potentially saved with rework operations.

The second perspective involves the implementation of **predictive sampling** [67]. In dynamic sampling, lots are dynamically selected without taking into account the *arrival of future lots*. There is no guarantee that a lot that has been sampled will be measured since new lots bringing much more information may arrive in the *near future*. We therefore think that it would be interesting to analyze sets of lots by considering lots that are known to arrive in a near future. This could be done by defining a time horizon where lots containing more information will be prioritized,

[3] A context can be a recipe, a technology, a process operation, etc.

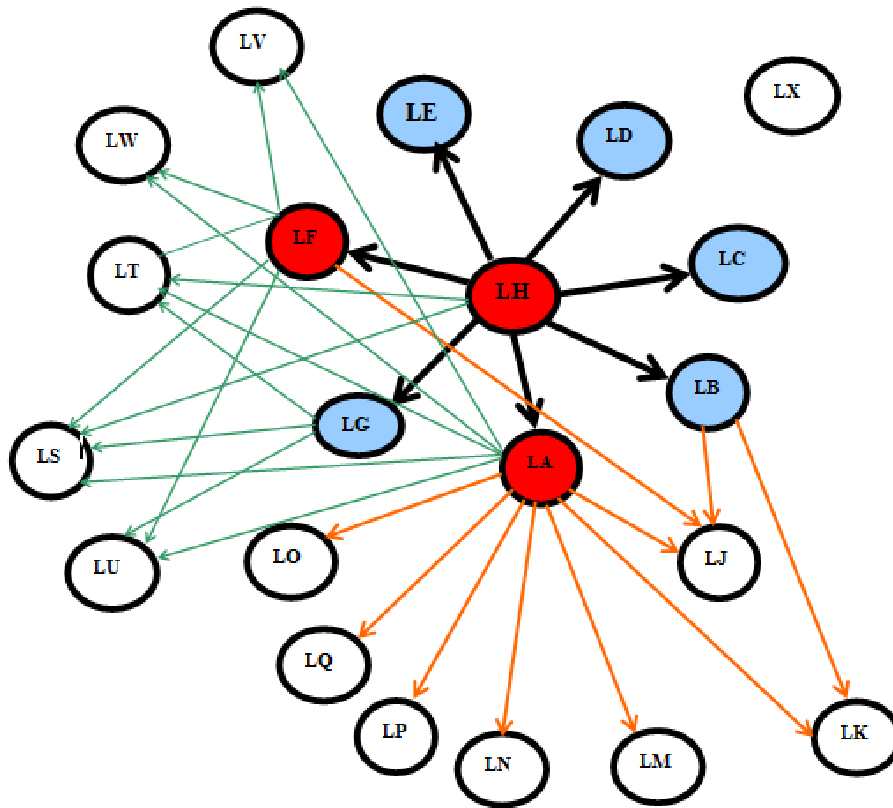


Figure 8.6: Concept of dominating sets.

and others directly sent to the next process operation.

This anticipation of the arrival of lots can be modeled as a scheduling problem of jobs with release dates and with multiple objectives, where minimizing the risk and the waiting times of lots should be balanced. The resulting scheduling problem could be solved using a multi-objective approach such as the one proposed in **Dugardin *et al.*** [24]. This is an original scheduling problem in semiconductor manufacturing, not mentioned for example in **Mőnch *et al.*** [54].

Appendix A

A.1 Glossary

APC	**Advanced Process Control** - A set of four control techniques (FDC, R2R, SPC, VM) used for controlling processes and machines.
Breach of control	A place in the control plan where a control operation might be added or removed.
Bottleneck	A place in the production chain where the capacity is limited such that the capacity is reduced in the whole production chain.
CMP	**Chemical Metal Polishing** - A work area where wafers are mechanically and chemically polished.
Cycle time	The time a wafer or a lot stays in a work area or the entire fab.

Defectivity	A term used to describe particles or defects generated on wafers during the production. This is mainly due to the mechanical parts of production tools, and the size of ICs is such that every particle can be critical.
Exposure	Financial cost incurred when processing a lot on a production tool.
Excursion	Deviation in process or product specifications.
Fab	A semiconductor **fab**rication plant - The factory where integrated circuits are produced on silicon wafers.
FDC	**Fault Detection and Classification** - A technique of monitoring statistically process variations by analyzing process equipment parameters.
FOUP	**Front Opening Unified Pod** - A box that contains 25 wafers.
GSI	**Global Sampling Indicator** - A score that helps selecting the best set of lots to sample, measure, or skip.
IC	**Integrated Circuit** - An electronic circuit built on a single piece of substrate (typically silicon).
IL	**Inhibit Limit** - A limit above which the production may be stopped if no control is performed. In term of wafers, IL represents the maximum number of wafers that should be run on a production tool between two controls.
IMPROVE	**Implementing Manufacturing science solutions to increase equiPment pROductiVity and fab pErformance** - A 42-month European project that focuses on the development of virtual metrology, predictive equipment behavior, and dynamic control plans.

IPC ***Index Permanent par Contexte*** **or Permanent Index per context** - A counter increased each time a context is verified. The context can be defined by chamber, by recipe or recipe type level (e.g. photoresist), by technology, by product, or by any combination of the previous element.

Lot A group of 25 wafers placed in a FOUP.

Measurable lot A lot for which the product is measurable i.e. a lot that contains a product for which a recipe exists and has been created on a metrology tool.

Metrology In this thesis, the term *Metrology* as well as *Inspection* are related to control operations performed on metrology or inspection tools.

Qualification The definition and approval of different recipes that can be used on a process tool.

Recipe A set of data required for an equipment to physically treat a wafer or a lot.

SPC **Statistical Process Control** - A technique based on statistical methods to analyze process stability.

S5 **Smart Sampling Skipping Scheduling Simulator** - A simulator that simulates several sampling policies using historical data. It is implemented in Excel VBA.

Risk In this thesis, the term risk is related to the material at risk, i.e. the potential loss if a problem occurs in production.

R2R **Run-to-Run** - A closed-loop control solution to correct for process deviation from the desired target.

Throughput time	The production speed (of a recipe) on a tool.
TRIZ	**Teoriya Resheniya Izobretatelskikh Zadatch**. In English, it is defined as **T**heory of **I**nventive **P**roblem **S**olving (**TIPS**) approach [4] developed in 1946 by Genrich S. Altshuller for solving technical problems.
Toolset	A group of tools in a workshop that can perform the same or similar kinds of recipes.
VM	**Virtual Metrology** - A technique for predicting measurements based on previous metrology measurements and equipment outputs.
Wafer	A thin circular plate on which the integrated circuits are produced.
WAR	**Wafer at Risk** - The number of wafers processed on a production tool since the last control.
WIP	**Work-In-Progress** or **Work-In-Process** - The set of lots that are awaiting to be processed.
WL	**Warning Limit** - A limit above which a situation starts to become critical in term of control. In term of wafers, WL represents the number of wafers that needs to be run on a production tool before increasing the priority of the tool for a control.
Workshop	A set of tools that are used for conducting a certain production step.

Appendix B

B.1 Clean room - ISO standard Classification

Classification Numbers (N)	Maximum concentration limits (particles/cm^3 of air) for particles equal to and larger than the considered sizes shown below					
	0.1μm	0.2μm	0.3μm	0.5μm	1μm	5.0μm
ISO1	10	2				
ISO2	100	24	10	4		
ISO3	1 000	237	102	35	8	
ISO4	10 000	2 370	1 020	352	83	
ISO5	100 000	23 700	10 200	3 520	832	29
ISO6	1 000 000	237 000	102 000	35 200	8 320	293
ISO7				352 000	83 200	2 930
ISO8				3 520 000	832 000	29 300
ISO9				35 200 000	8 320 000	293 000

Table B.1: Clean room - ISO Standard Classification [97].

Appendix C

C.1 S5 prototype

The prototype **S5** (**S**mart **S**ampling **S**kipping **S**cheduling **S**imulator) was implemented in Excel VBA by the EMSE. Figure C.1, Figure C.2, and Figure C.3 show different user interfaces for defining the parameters of simulations.

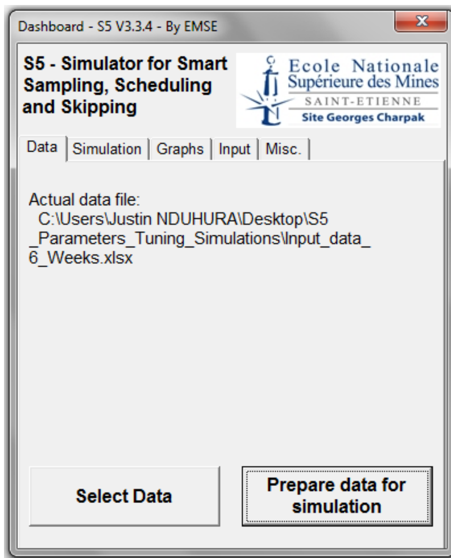
- Figure C.1 shows the user interfaces for selecting data and a type of simulation. The version of the prototype used in this thesis offers the possibility to simulate six different sampling policies (Figure C.1b):

    1. **Without sampling**, i.e. a sampling policy where no lot is inspected. This is not really a sampling policy but the aim is to evaluate the maximum risk that can be achieved within the production if no lot is sampled for inspection.
    2. **All sampling**, i.e. a sampling policy where all lots are inspected. The aim is to compare other policies to the ideal case.
    3. **Threshold sampling**, i.e. a sampling policy where lots are sampled only when a given risk threshold (Warning Limit) is reached.
    4. **GSI sampling**, i.e. a sampling policy using the first GSI sampling algorithm (Section 6.4.2).
    5. **Fab sampling**, i.e. the actual fab sampling policy (based on historical data).

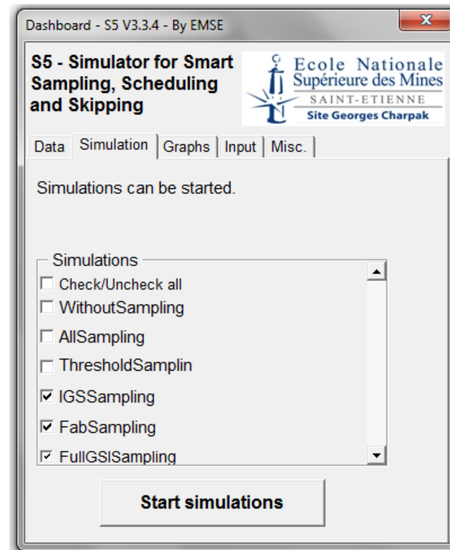
6. **Full GSI sampling**, i.e. a sampling policy using the second GSI sampling algorithm (Section 6.4.3).

- Figure C.2 shows the user interfaces for generating graphs after each simulation, and for defining the values of parameters (α, β, T_{Max}, T_{Min}, T_{Metro}, *Measure time*, *Warning Limit*, *Inhibit Limit*, *Number of metrology tools*, and *Inspection queue size*).

- Figure C.3 shows the user interface for initializing data or for generating more statistics.

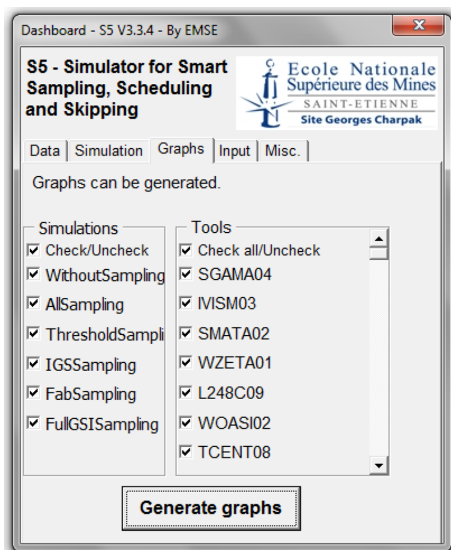


(a) S5 - Module DATA INPUT.

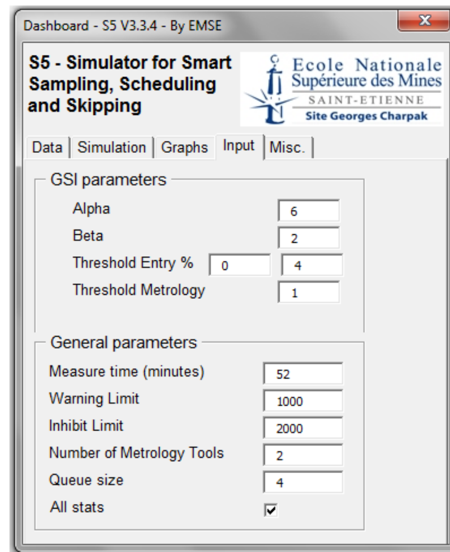


(b) S5 - Module SIMULATIONS.

Figure C.1: S5 interfaces (Modules DATA INPUT and SIMULATIONS).



(a) S5 - Module GRAPHS.



(b) S5 - Module PARAMETERS.

Figure C.2: S5 interfaces (Modules GRAPHS and PARAMETERS).

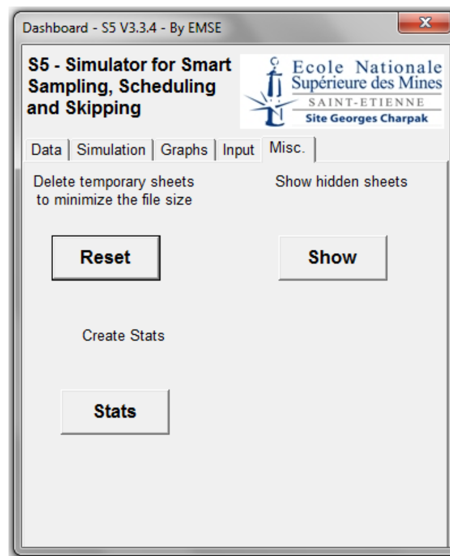


Figure C.3: S5 (Module RESET).

C.1.1 Input data

Three main types of historical data are necessary to simulate a sampling policy:

1. **Process input data** (Figure C.4) that contains historical data of production tools.
2. **Measurement input data** (Figure C.5) that contains historical data of inspection tools (Defectivity tools).
3. **Defectivity models** (Figure C.6) that defines the set of process operations that can be validated by a control operation, and the set of control operations that can validate a process operation.

PROCESS TOOL	JOBIN_INST	LOT	LOT_QTY_OUT	OPERATION_NAME	TECHNO_GROUP
SGAMA04	2/26/2011 5:00	Q104DHT	25	O_STRIP_NLDDGO2-07	C055
SMATA02	2/26/2011 5:00	Q105DSA	25	O_ETCH_GO2-02	C055
L248C09	2/26/2011 5:00	Q103BNJ	24	O_PHOTO_PSD-13	C055
L248C09	2/26/2011 5:01	Q105NPC	25	O_PHOTO_PLDDGO2-04	C055
WFC3002	2/26/2011 5:01	Q104BEK	25	O_ANN_WELL-14	I140
L248C09	2/26/2011 5:01	Q106SNB	25	O_PHOTO_NVT2-08	C065
TCENT10	2/26/2011 5:03	Q103CKH	25	O_RTP_SALICIDE-02	C055
WFC3003	2/26/2011 5:03	Q106SXF	25	O_ETCH_ON-11	C055
L193C04	2/26/2011 5:03	Q052CHS.04	12	O_PHOTO_VIA2-08	I140
IVISL03	2/26/2011 5:05	Q046BRJ.04	15	O_IMPL2_NPOCKCELL-01	C028
SMATA04	2/26/2011 5:05	Q053IMV	25	O_ETCH_PSD-01	C045
WSS3002	2/26/2011 5:05	Q052BMC	22	O_CMP_W-35	C055

Figure C.4: Process input data.

METROLOGY TOOLS	STEP_OUT_INST	LOT
F236002	2/26/2011 6:38	Q053ERG
F236001	2/26/2011 6:43	Q105CWR
FCOMP02	2/26/2011 7:03	Q101VED
F236002	2/26/2011 7:18	Q105AVE
F236001	2/26/2011 7:30	Q102JKM
FCOMP02	2/26/2011 8:09	Q103DFR
F236001	2/26/2011 8:30	Q053JXH
F280003	2/26/2011 8:32	Q051MVE
FCOMP02	2/26/2011 8:47	Q105AEK
FES3201	2/26/2011 9:13	Q050YBK
FCOMP01	2/26/2011 9:36	Q051PNE
F236001	2/26/2011 9:39	Q051XAP

Figure C.5: Measurement input data.

GENERIC OPERATION	DEF_OPER				
O_OXID_PAD	O_PHOTO_ACTIVE				
O_DEP_NIT_ACTIVE	O_PHOTO_ACTIVE				
O_DEP_HMASK_ACTIVE	O_PHOTO_ACTIVE				
O_PHOTO_ACTIVE	O_PHOTO_ACTIVE	O_ETCH_ACTIVE	O_ETCH_ON	O_OXID_SACOX	O_STRIP_PWGO2
O_PHOTO_ACTIVE	O_PHOTO_ACTIVE				
O_ETCH_ACTIVE	O_ETCH_ACTIVE	O_ETCH_ON	O_OXID_SACOX	O_STRIP_PWGO2	
O_ETCH_ACTIVE	O_ETCH_ACTIVE				
O_OXID_LINER	O_DEP_GAPFILL_STI				
O_OXID_LEB	O_DEP_GAPFILL_STI				
O_DEP_GAPFILL_STI	O_DEP_GAPFILL_STI	O_ETCH_ON	O_OXID_SACOX	O_STRIP_PWGO2	
O_DEP_GAPFILL_STI	O_DEP_GAPFILL_STI				
O_ANN_STI					
O_CMP_STI	O_ETCH_ON	O_OXID_SACOX	O_STRIP_PWGO2		

Figure C.6: Defectivity models.

C.1.2 Output results - Statistics

The **S5** prototype provides several indicators that are used for further analysis. Among these parameters, we have:

1. Number of lots that are sampled, i.e. the number of lots that are chosen to be inspected and which are placed in the inspection queue.

2. Number of sampled lots (directly inspected), i.e. the number of lots that do not wait in the inspection queue. Once sampled, they are directly inspected.

3. Number of sampled lots (directly queued), i.e. the number of lots that are sampled when the inspection queue is not full.

4. Number of sampled lots (exchanged), i.e. the number of lots that are sampled when the inspection queue is full. These lots are exchanged with some lots already waiting in the inspection queue.

5. Number of lots that are inspected i.e. the number of lots that are actually inspected or the number of lots processed on an inspection tool.

6. Number of lots that are skipped, i.e. the number of lots removed from the inspection queue or the number of lots that are sampled but not inspected.

7. Number of lots that are skipped (entry queue), i.e. the number of lots that are skipped due to the arrival of new lots bringing more information.

8. Number of lots that are skipped (metrology), i.e. the number of lots that are skipped due to the inspection of other lots.

9. Medium WAR (average) i.e. the sum of the WAR for all production tools divided by the number of tools, i.e. $\sum_{j=1}^{NbTools} \frac{WAR_j}{NbTools}$ where WAR_j is the WAR for the production tool j.

10. Maximum WAR (average), i.e. the sum of the maximum WAR of all production tools divided by the number of tools, i.e. $\sum_{j=1}^{Nbtools} \frac{MaximumWAR_j}{NbTools}$.

11. Number of lots above Warning Limit.

12. Number of wafers above Warning Limit.

13. Time spent above the Warning Limit. It corresponds to the sum of the times that all lots spent above the Warning Limit.

14. Number of lots above the Inhibit Limit.
15. Number of wafers above Inhibit Limit.
16. Time spent above the Inhibit Limit. It corresponds to the sum of the times that all lots spent above the Inhibit Limit.
17. The average time spent by lots in the inspection queue before being inspected.
18. The maximum time spent by lots in the inspection queue before being inspected.
19. The average time spent by lots in the inspection queue before being skipped.
20. The maximum time spent by lots in the inspection queue before being skipped.
21. The set of control operations used to reduce the WAR during the simulation.
22. The number of times a control operation was used during the simulation.
23. The number of control operations used during the simulation.

C.2 Impact of parameter β

Values of β with $\alpha = 1$	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
1	5.15*A	0.98*A	4.17*A	0.24*B	0.28*C
2	5.34*A	0.98*A	4.35*A	0.23*B	0.28*C
3	5.59*A	0.98*A	4.61*A	0.23*B	0.28*C
4	5.7*A	0.98*A	4.71*A	0.23*B	0.28*C
5	5.87*A	0.98*A	4.89*A	0.23*B	0.28*C
6	6*A	0.98*A	5.02*A	0.23*B	0.28*C
7	6.15*A	0.98*A	5.17*A	0.23*B	0.29*C
8	6.27*A	0.98*A	5.28*A	0.24*B	0.29*C
9	6.33*A	0.98*A	5.35*A	0.23*B	0.28*C
10	6.43*A	0.98*A	5.45*A	0.23*B	0.28*C
MIN	5.15*A	0.98*A	4.17*A	0.23*B	0.28*C
MAX	6.43*A	0.98*A	5.45*A	0.24*B	0.29*C
AVERAGE	5.88*A	0.98*A	4.9*A	0.23*B	0.28*C

Table C.1: Impact of β when $\alpha = 1$.

Values of β with $\alpha = 2$	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
1	5.22*A	0.98*A	4.24*A	0.24*B	0.27*C
2	5.34*A	0.98*A	4.36*A	0.23*B	0.27*C
3	5.47*A	0.98*A	4.48*A	0.23*B	0.27*C
4	5.71*A	0.98*A	4.73*A	0.23*B	0.28*C
5	5.9*A	0.98*A	4.92*A	0.23*B	0.27*C
6	6.01*A	0.98*A	5.02*A	0.23*B	0.28*C
7	6.15*A	0.98*A	5.17*A	0.23*B	0.28*C
8	6.26*A	0.98*A	5.28*A	0.23*B	0.28*C
9	6.27*A	0.98*A	5.29*A	0.24*B	0.28*C
10	6.4*A	0.98*A	5.42*A	0.23*B	0.28*C
MIN	5.22*A	0.98*A	4.24*A	0.23*B	0.27*C
MAX	6.4*A	0.98*A	5.42*A	0.24*B	0.28*c
AVERAGE	5.87*A	0.98*A	4.89*A	0.23*B	0.28*C

Table C.2: Impact of β when $\alpha = 2$.

Values of β with $\alpha = 4$	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
1	5.21*A	0.98*A	4.23*A	0.24*B	0.28*C
2	5.31*A	0.98*A	4.33*A	0.24*B	0.28*C
3	5.53*A	0.98*A	4.54*A	0.24*B	0.28*C
4	5.68*A	0.98*A	4.7*A	0.24*B	0.28*C
5	5.9*A	0.98*A	4.91*A	0.24*B	0.28*C
6	6.07*A	0.98*A	5.09*A	0.24*B	0.28*C
7	6.25*A	0.98*A	5.27*A	0.24*B	0.28*C
8	6.27*A	0.98*A	5.29*A	0.24*B	0.28*C
9	6.35*A	0.98*A	5.37*A	0.24*B	0.28*C
10	6.46*A	0.98*A	5.48*A	0.24*B	0.28*C
MIN	5.21*A	0.98*A	4.23*A	0.24*B	0.28*C
MAX	6.46*A	0.98*A	5.48*A	0.24*B	0.28*C
AVERAGE	5.9*A	0.98*A	4.92*A	0.24*B	0.28*C

Table C.3: Impact of β when $\alpha = 4$.

Values of β with $\alpha = 6$	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
1	5.21*A	0.98*A	4.23*A	0.25*B	0.28*C
2	5.34*A	0.98*A	4.35*A	0.24*B	0.28*C
3	5.58*A	0.98*A	4.6*A	0.24*B	0.28*C
4	5.76*A	0.98*A	4.78*A	0.24*B	0.28*C
5	5.93*A	0.98*A	4.95*A	0.24*B	0.28*C
6	6.05*A	0.98*A	5.07*A	0.24*B	0.28*C
7	6.23*A	0.98*A	5.25*A	0.24*B	0.28*C
8	6.26*A	0.98*A	5.28*A	0.24*B	0.28*C
9	6.37*A	0.98*A	5.38*A	0.25*B	0.28*C
10	6.58*A	0.98*A	5.6*A	0.24*B	0.28*C
MIN	5.21*A	0.98*A	4.23*A	0.24*B	0.28*C
MAX	6.58*A	0.98*A	5.6*A	0.25*B	0.28*C
AVERAGE	5.93*A	0.98*A	4.95*A	0.24*B	0.28*C

Table C.4: Impact of β when $\alpha = 6$.

Values of β with $\alpha = 8$	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
1	5.19*A	0.98*A	4.21*A	0.24*B	0.28*C
2	5.26*A	0.98*A	4.28*A	0.24*B	0.28*C
3	5.57*A	0.98*A	4.59*A	0.24*B	0.28*C
4	5.82*A	0.98*A	4.83*A	0.24*B	0.28*C
5	5.93*A	0.98*A	4.95*A	0.24*B	0.28*C
6	6.07*A	0.98*A	5.09*A	0.24*B	0.28*C
7	6.21*A	0.98*A	5.22*A	0.24*B	0.28*C
8	6.24*A	0.98*A	5.26*A	0.25*B	0.29*C
9	6.39*A	0.98*A	5.41*A	0.24*B	0.28*C
10	6.46*A	0.98*A	5.48*A	0.24*B	0.28*C
MIN	5.19*A	0.98*A	4.21*A	0.24*B	0.28*C
MAX	6.46*A	0.98*A	5.48*A	0.25*B	0.29*C
AVERAGE	5.91*A	0.98*A	4.93*A	0.24*B	0.28*C

Table C.5: Impact of β when $\alpha = 8$.

Values of β with $\alpha = 10$	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
1	5.17*A	0.98*A	4.19*A	0.25*B	0.28*C
2	5.34*A	0.98*A	4.36*A	0.24*B	0.28*C
3	5.51*A	0.98*A	4.53*A	0.24*B	0.28*C
4	5.77*A	0.98*A	4.79*A	0.24*B	0.29*C
5	5.9*A	0.98*A	4.92*A	0.24*B	0.28*C
6	6.12*A	0.98*A	5.14*A	0.25*B	0.29*C
7	6.25*A	0.98*A	5.27*A	0.25*B	0.28*C
8	6.27*A	0.98*A	5.29*A	0.25*B	0.28*C
9	6.39*A	0.98*A	5.41*A	0.25*B	0.29*C
10	6.5*A	0.98*A	5.51*A	0.25*B	0.28*C
MIN	5.17*A	0.98*A	4.19*A	0.24*B	0.28*C
MAX	6.5*A	0.98*A	5.51*A	0.25*B	0.29*C
AVERAGE	5.92*A	0.98*A	4.94*A	0.25*B	0.28*C

Table C.6: Impact of β when $\alpha = 10$.

Values of β with $\alpha = 12$	Number of sampled lots	Number of measured lots	Number of skipped lots	Average Medium W@R	Average Maximum W@R
1	5.19*A	0.98*A	4.21*A	0.24*B	0.27*C
2	5.39*A	0.98*A	4.41*A	0.24*B	0.28*C
3	5.52*A	0.98*A	4.54*A	0.24*B	0.28*C
4	5.71*A	0.98*A	4.73*A	0.25*B	0.29*C
5	5.95*A	0.98*A	4.97*A	0.25*B	0.29*C
6	6.05*A	0.98*A	5.06*A	0.25*B	0.29*C
7	6.18*A	0.98*A	5.2*A	0.25*B	0.29*C
8	6.3*A	0.98*A	5.31*A	0.25*B	0.29*C
9	6.45*A	0.98*A	5.47*A	0.25*B	0.29*C
10	6.41*A	0.98*A	5.43*A	0.25*B	0.28*C
MIN	5.19*A	0.98*A	4.21*A	0.24*B	0.27*C
MAX	6.45*A	0.98*A	5.47*A	0.25*B	0.29*C
AVERAGE	5.91*A	0.98*A	4.93*A	0.25*B	0.28*C

Table C.7: Impact of β when $\alpha = 12$.

Bibliography

[1] Inline Automated Defect Classification: a Novel Approach to Defect Management. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2005), pp. 43–48.

[2] Alegret, C. *Développement de méthodes génériques de corrélation entre les mesures électriques et physiques des composants et les étapes élémentaires de fabrication.* PhD thesis, University of Joseph Fourier, Grenoble, France, 2006.

[3] Allebe, C., Govaerts, B., Van Kerschaver, E., De Wolf, S., and Szlufcik, J. Control Charts and Efficient Sampling Methodologies in the Field of Photovoltaics. In *IEEE Photovoltaic Specialists Conference* (2002), pp. 387–390.

[4] Altshuller, G. The Innovation Algorithm: TRIZ, Systematic Innovation, and Technical Creativity. Tech. rep., Worcester, MA: Technical Innovation Center, Inc., 1999.

[5] Babikian, R., and Engelhard, C. Statistical Methods for Measurement Reduction in Semiconductor Manufacturing. pp. 212–215.

[6] Bai, P. Advancing Moore's Law: Challenges and Opportunities. In *International Conference on ASIC* (2009).

[7] Bassetto, S. *Contribution la Quantification et l'Amélioration des Moyens de Production.* PhD thesis, Ecole Nationale Supérieure d'Arts et Métiers, Metz, France, 2005.

[8] Bassetto, S., and Siadat, A. Operational Methods for Improving Manufacturing Control Plans: Case Study in a Semiconductor Industry. *Journal of Intelligent Manufacturing 20*, 1 (2009), 55–65.

[9] Bean, J. W. Variation reduction in a wafer fabrication line through inspection optimization. Master's thesis, Massachusets Institute of Technology, USA, 1997.

[10] Bettayeb, B., Bassetto, S., Vialletelle, P., and Tollenaere, M. Quality and Exposure Control in Semiconductor Manufacturing. Part I: Modelling. *International Journal of Production Research*.

[11] Bettayeb, B., Bassetto, S., Vialletelle, P., and Tollenaere, M. Quality and Exposure Control in Semiconductor Manufacturing. Part II: Evaluation. *International Journal of Production Research*.

[12] Boussetta, A., and Cross, A. J. Adaptive Sampling Methodology for In-Line Defect Inspection. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2005), pp. 25–31.

[13] Bunday, B., Rijpers, B., Banke, B., Archie, C., Peterson, I. B., Ukraintsev, V., Hingst, T., and Asano, M. Impact of Sampling on Uncertainty: Semiconductor Dimensional Metrology Applications. In *Metrology, Inspection, and Process Control for Microlithography* (2008), vol. 6922, pp. 1–22.

[14] Bush, H. *Towards the Re-Use of Knowledge in Dynamic Industrialization Processes: Case Study in the Microelectronic Domain.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2006.

[15] Chang, Y.-J., Kang, Y., Hsu, C.-L., Chang, C.-T., and Chan, T. Y. Virtual Metrology Technique for Semiconductor Manufacturing. In *International Joint Conference on Neural Networks* (2006), pp. 5289–5293.

[16] Chen, A., Hsueh, S., and Blue, J. Optimum Sampling for Track PEB CD Integrated Metrology. In *IEEE International Conference on Automation Science and Engineering* (2009), pp. 439–442.

[17] Chien, C.-F., Chang, K.-H., and Chen, C.-P. Sampling Strategy and Model to Measure and Compensate the Overlay Errors. In *Metrology, Inspection, and Process Control for Microlithography* (2001), vol. 4344, pp. 245–256.

[18] Chien, C.-F., Hsu, S.-C., Peng, S., and Wu, C.-H. A Cost-Based Heuristic for Statistically Determining Sampling Frequency in a Wafer Fab. In *Semiconductor Manufacturing Technology Workshop* (2000), pp. 217 –229.

[19] Cronshagen, A. Analysis of a Cumulative Results Sampling Plan for Use with Sampling Tables using Zero Acceptance Numbers. *Transactions of the IRE Professional Group on Reliability and Quality Control 4* (1954), 14–31.

[20] Dauzère-Pérès, S., Rouveyrol, J., Yugma, C., and Vialletelle, P. A Smart Sampling Algorithm to Minimize Risk Dynamically. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2010), pp. 307–310.

[21] Dennison, C. Developing Effective Inspection Systems and Strategies for Monitoring CMP Processes. *MICRO 16*, 2 (1998), 31–41.

[22] Ducotey, G., Couvrat, A., Audran, V., Pepper, D., and Couturier, L. InLine Methodology for Defectivity Analysis from Dark Field Wafer Inspection to Defect Root Cause Analysis using FIB Cut. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2008), pp. 138–141.

[23] Dudding, B. P. Sampling Inspection and Quality Control. *Journal of the Institution of Production Engineers 21*, 1 (1942), 13–34.

[24] Dugardin, F., Yalaoui, F., and Amodeo, L. New Multi-Objective Method to Solve Reentrant Hybrid Flow Shop Scheduling Problem. *European Journal of Operational Research 203*, 1 (2010), 22–31.

[25] Elliott, R. C., Nurani, R. K., Gudmundsson, D., Preil, M., Nasongkhla, R., and Shanthikumar, J. G. Critical Dimension Sample Planning for Sub-0.25 Micron Processes. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (1999), pp. 139–142.

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

[27] Gissrau, M., and Rose, O. A Detailed Model for a High-Mix Low-Volume Asic Fab. In *Modeling and Analysis of Semiconductor Manufacturing Conference (included in the Winter Simulation Conference)* (2011), pp. 1948–1958.

[28] Good, R. P., Pabst, D., and Stirton, J. B. Compensating for the Initialization and Sampling of EWMA Run-to-Run Controlled Processes. *IEEE Transactions on Semiconductor Manufacturing 23*, 2 (2010), 168–177.

[29] Good, R. P., and Purdy, M. A. An MILP Approach to Wafer Sampling and Selection. *IEEE Transactions on Semiconductor Manufacturing 20*, 4 (2007), 400–407.

[30] Guldi, R. In-Line Defect Reduction from a Historical Perspective and Its Implications for Future Integrated Circuit Manufacturing. *IEEE Transactions on Semiconductor Manufacturing 17*, 4 (2004), 629–640.

[31] Guldi, R., Shaw, J. B., Ritchison, J., Oestreich, S., Davis, K., and Fiordalice, R. Characterization of Copper Voids in Dual Damascene Processes. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2002), pp. 351–355.

[32] Holfeld, A., Barlović, R., and Good, P. A Fab-Wide APC Sampling Application. *IEEE Transactions on Semiconductor Manufacturing 20*, 4 (2007), 393–399.

[33] Hsu, J. I. S. A Cost Model for Skip-Lot Destructive Sampling. *IEEE Transactions on Reliability 26*, 1 (1977), 70–72.

[34] Hyung Joo Lee, B. *Advanced Process Control and Optimal Sampling in Semiconductor Manufacturing*. PhD thesis, The University of Texas at Austin, 2008.

[35] Jansen, S., Florence, G., Perry, A., and Fox, S. Utilizing Design Layout Information to Improve Efficiency of SEM Defect Review Sampling. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2008), pp. 69–71.

[36] Johnzén, C. *Modeling and Optimizing Flexible Capacity Allocation in Semiconductor Manufacturing*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne, Gardanne, France, 2009.

[37] Johnzén, C., Dauzère-Pérès, S., and Vialletelle, P. Flexibility Measures for Qualification Management in Wafer Fabs. *Production Planning and Control 22*, 1 (2011), 81–90.

[38] Kaga, Y., Sato, Y., Yamada, Y., Yamazaki, Y., Aoki, M., Harukawa, R., and Chang, E. Integrated Defect Sampling Method by Using Design Attribute for High Sensitivity Inspection in 45nm Production Environment. In *IEEE International Symposium on Semiconductor Manufacturing* (2008), pp. 379–381.

[39] Khan, A. A., Moyne, J. R., and Tilbury, D. M. An Approach for Factory-Wide Control Utilizing Virtual Metrology. *IEEE Transactions on Semiconductor Manufacturing 20*, 4 (2007), 364–375.

[40] Kiba, J.-E. *Simulation et Optimisation du Transport Automatisé dans la Fabrication des Semi-Conducteurs*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne, Gardanne, France, 2010.

[41] Kuo, W. W., Akella, R., and Fletcher, D. Adaptive Sampling for Effective Multi-Layer Defect Monitoring. In *IEEE International Symposium on Semiconductor Manufacturing* (1997), pp. 289–293.

[42] Kuo, W. W., Wang, A.-H., Akella, R., and Nurani, R. K. A Combined Adaptive Sampling Strategy with Limited Inspection Capacity. In *IEEE International Symposium on Semiconductor Manufacturing* (1996), pp. 235–238.

[43] Kwang, C. L., and Chin, O. E. A Novel Push-Pull Sampling Methodology for Test Production in Semiconductor Manufacturing Industries. In *IEEE/CPMT International Symposium on Electronic Manufacturing Technology* (2008), pp. 1–3.

[44] Langford, R. E., Hsu, G., and Sun, C. The Identification and Analysis of Systematic Yield Loss. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2000), pp. 92–95.

[45] Lazaroff, D., Bakker, D., and Granath, D. R. Evaluating Different Sampling Techniques for Process Control Using Automated Patterned Wafer Inspection Systems. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (1991), pp. 92–97.

[46] Le Denmat, J., Charbois, V., Luche, M. C., Kerrien, G., Couturier, L., Karsenti, L., and Geshel, M. Tracking of Design Related Defects Hidden in the Random Defectivity in a Production Environment. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2009), pp. 5–13.

[47] Lee, J. H. Artificial Intelligence-Based Sampling Planning System for Dynamic Manufacturing Process. *Expert Systems with Applications 22*, 2 (2002), 117–133.

[48] Lee, J. H., You, S. J., and Park, S. C. A New Intelligent SOFM-Based Sampling Plan for Advanced Process Control. *Expert Systems with Applications 20*, 2 (2001), 133–151.

[49] Lee, J. H., Yu, S. J., and Park, S. C. Design of Intelligent Data Sampling Methodology Based on Data Mining. *IEEE Transactions on Robotics and Automation 17*, 5 (2001), 637–649.

[50] Lensing, K., and Stirton, B. Perspectives on Integrated Metrology and Wafer-Level Control. In *IEEE International Symposium on Semiconductor Manufacturing* (2007), pp. 1–5.

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

[52] McIntyre, M., Nurani, R. K., and Akella, R. Key Considerations in the Development of Defect Sampling Methodologies. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (1996), pp. 81–85.

[53] Mili, A. *Vers des méthodes fiables de contrôle des procédés par la maîtrise du risque*. PhD thesis, Institut Polytechnique de Grenoble, Grenoble, France, 2009.

[54] Mönch, L., Fowler, J. W., Dauzère-Pérès, S., Mason, S. J., and Rose, O. A Survey of Problems, Solutions Techniques, and Future Challenges in Scheduling Semiconductor Manufacturing Operations. *Journal of Scheduling 14*, 6 (2011), 583–599.

[55] Montgomery, D. *Introduction to Statistical Quality Control*. John Wiley & Sons, 2005. 5th Edition.

[56] Moon, B., McNames, J., Whitefield, B., Rudolph, P., and Zola, J. Wafer Sampling by Regression for Systematic Wafer Variation Detection. In *Data Analysis and Modeling for Process Control* (2005), vol. 5755, pp. 212–221.

[57] Mouli, C. Adaptive Sampling Technology - the next step to factory efficiency. *EuroAsia Semiconductor Magazine* (2005), 1–3.

[58] Mouli, C., and Scott, M. J. Adaptive Metrology Sampling Techniques Enabling Higher Precision in Variability Detection and Control. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2007), pp. 12–17.

[59] Nduhura Munga, J., Dauzère-Pérès, S., Vialletelle, P., and Yugma, C. A Novel Approach to Minimize the Number of Controls in the

Defectivity Area. In *13th Scientific and Technical Meeting of ARCSIS* (2010), pp. 1–2.

[60] Nduhura Munga, J., Dauzère-Pérès, S., Vialletelle, P., and Yugma, C. Dynamic Management of Controls in Semiconductor Manufacturing. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2011), pp. 18–23.

[61] Nduhura Munga, J., Dauzère-Pérès, S., Vialletelle, P., and Yugma, C. Optimized Management of Excursions in Semiconductor Manufacturing. In *Modeling and Analysis of Semiconductor Manufacturing Conference (included in the Winter Simulation Conference)* (2011), pp. 2105–2112.

[62] Nduhura Munga, J., Dauzère-Pérès, S., Vialletelle, P., and Yugma, C. Industrial Implementation of a Dynamic Sampling Algorithm in Semiconductor Manufacturing: Approach and Challenges. In *Modeling and Analysis of Semiconductor Manufacturing Conference (included in the Winter Simulation Conference)* (2012), pp. 1–9.

[63] Nduhura Munga, J., Dauzère-Pérès, S., Yugma, C., and Vialletelle, P. A Mathematical Programming Approach for Determining Control plans in Semiconductor Manufacturing. In *International Conference on Industrial Engineering and Systems Management* (2011), pp. 1–9.

[64] Nduhura Munga, J., Dauzère-Pérès, S., Yugma, C., and Vialletelle, P. A Fab Wide Indicator to Support Inductrial Implementation of Dynamic Control Plans in Semiconductor Manufacturing. *International Journal of Production Research* (2012), 1–20.

[65] Nduhura Munga, J., Dauzère-Pérès, S., Yugma, C., and Vialletelle, P. A Mathematical Programming Approach for Optimizing Control Plans in Semiconductor Manufacturing. *International Journal of Production Economics* (2012), 1–18.

[66] Nduhura Munga, J., Dauzère-Pérès, S., Yugma, C., and Vialletelle, P. Optmisation des Contrôles dans la Fabrication des Semiconduc-

teurs. In *13ème congrès de la société Française de Recherche Opérationnelle et d'Aide à la DÉcision (ROADEF)* (2012), pp. 1–2.

[67] Nduhura Munga, J., Rodriguez-Verjan, G., Dauzère-Pérès, S., Yugma, C., Vialletelle, P., and Pinaton, J. A Literature Review on Sampling Techniques in Semiconductor Manufacturing. *IEEE Transactions on Semiconductor Manufacturing* (2012), 1–8.

[68] Nurani, R. K., Akella, R., and Strojwas, A. J. In-Line Defect Sampling Methodology in Yield Management: an Integrated Framework. *IEEE Transactions on Semiconductor Manufacturing 9*, 4 (1996), 506–517.

[69] Nurani, R. K., Akella, R., Strojwas, A. J., and Wallace, R. Role of In-Line Defect Sampling Methodology in Yield Management. In *International Symposium onSemiconductor Manufacturing* (1995), pp. 243–247.

[70] Nurani, R. K., Akella, R., Strojwas, A. J., Wallace, R., McIntyre, M. G., Shields, J., and Emami, I. Development of an Optimal Sampling Strategy for Wafer Inspection. In *IEEE International Symposium on Semiconductor Manufacturing* (1994), pp. 143–146.

[71] Nurani, R. K., and Shantikumar, J. G. Process Control for Items Produced in Lots with Inter and Intra Lot Variations. *International Journal of Industrial Engineering 7*, 1 (2000), 57–66.

[72] Nurani, R. K., Strojwas, A. J., Maly, W. P., Ouyang, C., Shindo, W., Akella, R., McIntyre, M. G., and Derrett, J. In-Line Yield Prediction Methodologies Using Patterned Wafer Inspection Information. *IEEE Transactions on Semiconductor Manufacturing 11*, 1 (1998), 40–47.

[73] O'mara, W. C., Herring, R. B., and Hunt, L. P. *Handbook of Semiconductor Silicon Technology.* Noyes Publications, 1990.

[74] Ossame, A. A., and Travis, H. P. Industrial Clean Rooms: Architectural Engineering Considerations. *Journal of Architectural Engineering 1*, 1 (1995), 37–52.

[75] Paccard, C. *Développement d'Outils Statistiques pour la Mise en Place des Boucles de Régulation.* PhD thesis, University of Toulouse III - Paul Sabatier, France, 2008.

[76] Pesotchinsky, L. Problems Associated with Quality Control Sampling in Modern IC Manufacturing. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology 10*, 1 (1987), 107–110.

[77] Prabhu, S., Montgomery, D. C., and Runger, G. C. A Combined Adaptive Sample Size and Sampling Interval X-bar Control Scheme. *Journal of Quality Technology 26*, 3 (1994), 164–176.

[78] Purdy, M. Dynamic, Weight-Based Sampling Algorithm. In *IEEE International Symposium on Semiconductor Manufacturing* (2007), pp. 1–4.

[79] Purdy, M., Nicksic, C., and Lensing, K. Method for Efficiently Managing Metrology Queues. In *IEEE International Symposium on Semiconductor Manufacturing* (2005), pp. 71–74.

[80] Purdy, M. A. Dynamic metrology sampling methods, and system for performing same, 2005. Patent, United States.

[81] Rodriguez-Verjan, G., Dauzère Pérès, S., and Pinaton, J. Impact of Control Plan Design on Tool Risk Management: A Simulation Study in Semiconductor Manufacturing. In *Modeling and Analysis of Semiconductor Manufacturing Conference (included in the Winter Simulation Conference)* (2011), pp. 1918–1925.

[82] Sahnoun, M., Bassetto, S., Bastoini, S., and Vialletelle, P. Optimisation of the Process Control in a Semiconductor Company: Model and Case Study of Defectivity Sampling. *International Journal of Production Research 49*, 13 (2011), 3873–3890.

[83] Sahnoun, M., Vialletelle, P., Bassetto, S., Bastoini, S., and Tollenaere, M. Optimizing Return on Inspection Trough Defectivity Smart Sampling. In *IEEE International Symposium on Semiconductor Manufacturing* (2010), pp. 1–4.

[84] Scanlan, B. Defect Inspection Sampling plans-which one is right for me? In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (1998), pp. 103–108.

[85] Schellenberger, M., Roeder, G., chsner, R., Schpka, U., and Kasko, I. Advanced Process Control Lessons Learned from Semiconductor Manufacturing. *International Journal of Photovoltaics 9* (2010), 79–87.

[86] Shanthikumar, J. G. Effects of Capture Rate and Its Repeatability on Optimal Sampling Requirements in Semiconductor Manufacturing. In *IEEE International Symposium on Semiconductor Manufacturing* (2007), pp. 1–6.

[87] Shumaker, J., Phillips, A., and Lauderdale, M. Intelligent Sample Test Using Cost Based Methodologies. In *IEEE International Symposium on Semiconductor Manufacturing* (2003), pp. 439–442.

[88] Song-bor, L., Ta-Yung, R. L., Janson, L., and Yu-Ching, C. A Capacity-Dependence Dynamic Sampling Strategy. In *IEEE International Symposium on Semiconductor Manufacturing* (2003), pp. 312–314.

[89] Stokowski, S., and Vaez-Iravani, M. Wafer Inspection Technology Challenges for ULSI Manufacturing. In *International Conference on Characterization and Metrology for ULSI Technology;* (1998), vol. 449, pp. 405–415.

[90] Stratton, R., and Mann, D. andOtterson, P. The Theory of Inventive Problem-Solving (TRIZ) and Systematic Innovation - A Missing Link In Engineering Education? *TRIZ Journal* (2000), 1–14.

[91] Su, A.-J., Yu, C.-C., and Ogunnaike, B. A. On the Interaction between Measurement Strategy and Control Performance in Semiconductor Manufacturing. *Journal of Process Control 18*, 3-4 (2008), 266–276.

[92] Sullivan, D. B., Conrad, E. W., and Smyth, J. S. Overlay Metrology Sampling Capability Analysis and Implementation in Manufacturing. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2004), pp. 208–212.

[93] Sun, S., and Johnson, K. 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.

[94] Tomlinson, W., Nurani, R. K., Burns, M., and Shanthikumar, J. G. Development of Cost Effective Sampling Strategy for In-Line Monitoring. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (1997), pp. 8–12.

[95] Veetil, V., Sylvester, D., Blaauw, D., Shah, S., and Rochel, S. Efficient Smart Sampling Based Full-Chip Leakage Analysis for Intra-Die Variation Considering State Dependence. In *IEEE Design Automation Conference* (2009), pp. 154–159.

[96] Western Electric Company, I. *Statistical Quality Control Handbook*. Mack Printing Company, 1958. 2nd Edition.

[97] Whyte, W. *Cleanroom Technology: Fundamentals of Design, Testing and Operation*. Wiley, 2010. 2nd Edition.

[98] Williams, R., Gudmundsson, D., Monahan, K., and Shanthikumar, J. G. Optimized Sample Planning for Wafer Defect Inspection. In *IEEE International Symposium on Semiconductor Manufacturing* (1999), pp. 43–46.

[99] Williams, R., Gudmundsson, D., Nurani, R., Stoller, M., Chatterjee, A., Seshadri, S., and Shanthikumar, J. G. Challenging the Paradigm of Monitor Reduction to Achieve Lower Product Costs. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (1999), pp. 420–425.

[100] Wootton, P., Saville, B., Lutz, A., and Oakley, J. Review Sample Shaping Through the Simultaneous Use of Multiple Classification Technologies in IMPACT ADC. In *IEEE/SEMI Advanced Semiconductor Manufacturing Conference* (2001), pp. 207 –217.

[101] Wright, P. K. *21st Century Manufacturing.* Prentice-Hall, 2001. 1st Edition.

[102] Wu, C.-W., and Pearn, W. L. A Variables Sampling Plan Based on Cpmk for Product Acceptance Determination. *European Journal of Operational Research 184*, 2 (2006), 549–560.

[103] Xuemei, C., Preil, M. E., Le gaff dussable, M., and Maenhoudt, M. Optimizing Overlay Sampling for Higher Yields. *Semiconductor International 26*, 4 (2003), 56–60.

[104] Yugma, C., Dauzère-Pérès, S., Rouveyrol, J., Vialletelle, P., Pinaton, J., and Relliaud, C. A Smart Sampling Scheduling and Skipping Skipping Simulator and its Evaluation on Real Data Sets. In *Modeling and Analysis of Semiconductor Manufacturing Conference (included in the Winter Simulation Conference)* (2011), pp. 1908–1917.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2012 EMSE 0663

Justin NDUHURA MUNGA

IMPLEMENTING AND OPTIMIZING DYNAMIC CONTROL PLANS IN SEMICONDUCTOR MANUFACTURING

Speciality: Industrial Engineering

Keywords: Dynamic control, sampling, excursion, defectivity, optimization.

Abstract:

In this thesis, we have worked on the problem of implementing dynamic control plans in a high-mix semiconductor environment. We focused on the trade-off between yield and cycle time, the minimization of the number of controls without added value, and the optimization of the use of inspection capacity. We started our works by formalizing and generalizing the problem through a literature review. Then, we proposed three main solutions to industrially implement dynamic control plan policies. The first solution we proposed is based on an indicator that enables a very large amount of data to be handled and several risk types assessed with little CPU. The second solution is based on smart sampling algorithms we developed to enable the dynamic selection of the best products or lots to control in real time. And the third solution is a mixed-integer linear programming model we developed to optimize the key parameters that are used in the smart sampling algorithms.

The originality of this thesis lies in the industrial implementation of the general solutions we proposed. All solutions have been industrially validated and some of the solutions have been extended to other sites of the company. Several perspectives have been highlighted and offer numerous avenues for further works.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : **2012 EMSE 0663**

Justin NDUHURA MUNGA

MISE EN ŒUVRE ET OPTIMISATION DES PLANS DE CONTRÔLE DYNAMIQUE DANS LA FABRICATION DES SEMI-CONDUCTEURS

Spécialité : **Génie Industriel**

Mots Clefs : contrôles dynamiques, échantillonnage, excursion, défectivité, optimisation.

Résumé :

Dans cette thèse, nous avons travaillé sur le problème de la mise œuvre des plans de contrôle dynamique au sein d'un environnement semiconducteur multi-produits. Nous nous sommes focalisés sur le compromis entre le rendement et le temps de cycle, la réduction du nombre de contrôles sans valeur ajoutée, et l'optimisation de l'utilisation de la capacité de contrôle. Nous avons commencé par formaliser et généraliser le problème au travers d'une révue de la littérature. Ensuite, nous avons proposé trois principales solutions pour supporter l'implémentation industrielle des plans de contrôle dynamique. La première solution que nous avons proposée est basée sur un indicateur qui permet le traitement d'un très grand volume de données et l'évaluation de plusieurs types de risques avec une très faible consommation des ressources informatiques. La deuxième solution est basée sur des algorithmes d'échantillonnage intelligents que nous avons développés pour permettre le choix en dynamique des meilleurs produits ou lots à contrôler. Et la troisième solution est un programme linéaire mixte en nombres entiers que nous avons développé pour optimiser les paramètres clés qui sont utilisés dans les algorithmes d'échantillonnage dynamique.

L'originalité des travaux de cette thèse se trouve dans l'industrialisation des différentes solutions que nous avons proposées. Toutes les solutions ont été validées industriellement et certaines solutions ont été étendues à d'autres sites de la compagnie. Plusieurs perspectives ont été identifiées et offrent ainsi de nombreuses pistes de recherche.