



HAL
open science

Approche holistique du contrôle du focus en photolithographie 193nm immersion pour les niveaux critiques en 28nm et 14nm FD-SOI

Jean-Gabriel Simiz

► **To cite this version:**

Jean-Gabriel Simiz. Approche holistique du contrôle du focus en photolithographie 193nm immersion pour les niveaux critiques en 28nm et 14nm FD-SOI. Micro et nanotechnologies/Microélectronique. Université de Lyon, 2016. Français. NNT: . tel-01419388v1

HAL Id: tel-01419388

<https://ujm.hal.science/tel-01419388v1>

Submitted on 19 Dec 2016 (v1), last revised 7 Jan 2019 (v2)

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.



N°d'ordre

(créé ultérieurement)

NNT : xxx

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
STMicroelectronics Crolles 2 SAS

Ecole Doctorale N° 488
Sciences, Ingénierie, Santé

Spécialité de doctorat :
Discipline : Microélectronique

Soutenue publiquement le 24/11/2016, par :
Jean-Gabriel SIMIZ

**Approche holistique du contrôle du focus en
photolithographie 193nm immersion pour les
niveaux critiques en 28nm et 14nm FD-SOI**

Devant le jury composé de :

Panagiota MORFOULI, Professeur, Grenoble INP, IMEP-LAHC, Présidente de jury

Jean Jacques SIMON, Professeur, Aix Marseille Université, IM2NP, Rapporteur

Jean-Hervé TORTAI, HDR, Université Grenoble Alpes, LTM CNRS/UJF, Rapporteur

Bertrand LE-GRATIET, Docteur, STMicroelectronics, Encadrant

Yves JOURLIN, Professeur, Université de Lyon, Laboratoire Hubert Curien UMR CNRS,
Directeur de thèse

Jan-Willem GEMMINK, ASML, Invité

Thomas KÄMPFE, Université de Lyon, Laboratoire Hubert Curien UMR CNRS, Invité

A la mémoire d'Alexandre Tishchenko,

SOMMAIRE

SOMMAIRE	7
REMERCIEMENTS	11
TABLE DES FIGURES	13
LISTE DES TABLEAUX	19
GLOSSAIRE	21
0 INTRODUCTION	25
1 L'ECOSYSTEME INDUSTRIEL DE LA MICROELECTRONIQUE	31
1.1 La microélectronique et l'industrie des semi-conducteurs	31
1.1.1 9 décades de densité de transistors en 50 ans !	31
1.1.2 L'industrie du semi-conducteur	33
1.1.3 La loi de Moore et le « More Than Moore »	35
1.2 La fabrication des circuits intégrés	39
1.2.1 Intégration	39
1.2.1.1 Les technologies CMOS	39
1.2.1.2 Le contrôle de procédé	42
1.2.1.3 Technologies dérivées	42
1.2.2 Quelques indicateurs clés de la fabrication	43
1.3 Conclusion	44
2 LA PHOTOLITHOGRAPHIE	45
2.1 Le procédé photolithographique	45
2.2 Les machines	46
2.3 La diffraction et l'image aérienne	48
2.3.1 La formation de l'image aérienne	48
2.3.2 La « computational lithography »	53
2.3.2.1 Les illuminations	53
2.3.2.2 Les OPC	54
2.4 La lithographie dans la Loi de Moore	55
2.5 La fenêtre de procédé de l'exposition	57

2.5.1	Fenêtre de procédé.....	57
2.5.2	Décalage de focus.....	59
2.5.3	Décalage de dose	60
2.5.4	Quelques exemples.....	60
2.6	Les challenges de la thèse.....	61
3	ETUDE SUR LE BUDGET ET LE CONTROLE DU FOCUS	63
3.1	Analyse et décomposition du budget focus	64
3.1.1	La multi-wafer matrice de focus.....	64
3.1.2	Imagerie vs. Focus mesuré sur plaquette.....	66
3.1.2.1	Un unique focus optimum par motif	66
3.1.2.2	Reconstruction de la Bossung	67
3.2	Variabilité focus	70
3.2.1	Les différents niveaux de variabilité	70
3.2.2	Les sources de variabilité	72
3.2.2.1	Les motifs	72
3.2.2.2	Les effets masque	74
3.2.2.3	Les effets du scanner	75
3.2.2.4	Wafer	77
3.3	Les effets du produit.....	78
3.3.1	Design.....	79
3.3.2	Architecture et assemblage.....	80
3.3.3	Modulation de la réflectivité	81
3.3.4	Modulation de la topographie.....	82
3.4	Focus vs. Topographie	83
3.4.1	Correction de la topographie pendant l'exposition.....	83
3.4.2	Décalage du focus d'exposition provoqué par la topographie non corrigeable.....	87
3.5	Conséquences sur la fenêtre de procédé.....	91
3.6	Conclusion.....	94
4	TOPOGRAPHIE INTRA-CHAMP ET MODELISATION.....	97

4.1	Les mesures de topographie	97
4.1.1	Leveling.....	97
4.1.2	Le Wyko.....	98
4.1.3	Le WaferSight PWG.....	102
4.1.4	Comparaison des différentes méthodes de mesure.....	104
4.2	La régression PLS	107
4.2.1	Introduction	107
4.2.2	L’algorithme PLS	108
4.2.3	Les indicateurs de performance.....	110
4.3	Méthodologie et résultats	111
4.3.1	La genèse de l’idée fondatrice.....	112
4.3.2	Le principe de l’idée fondatrice.....	114
4.3.3	Choix du niveau de l’étude.....	117
4.3.4	Extraction de densités de design sur le GDS.....	118
4.3.4.1	Les différents types de densité	118
4.3.4.2	Manipulation des niveaux de design	119
4.3.4.2.1	Reconstruction du design	119
4.3.4.2.2	Combinaison de niveaux	121
4.3.5	Le modèle de leveling en 14FD-SOI Contact	121
4.3.6	Le modèle de topographie haute fréquence.....	123
4.3.6.1	Modèle simple	124
4.3.6.1.1	Puces CMOS	124
4.3.6.1.2	Puces non CMOS	127
4.3.6.2	Modèle combiné.....	130
4.3.7	Analyse des VIP	132
4.3.8	Application à un masque de production en 28nm FD-SOI.....	133
4.4	Conclusion.....	134
5	APPLICATIONS ET PERSPECTIVES D’OPTIMISATIONS.....	137
5.1	Définition des zones d'intérêts.....	137

5.1.1	Optimisation du leveling	139
5.1.2	Optimisation du design.....	141
5.1.2.1	Dummies	141
5.1.2.2	« Leveling aware assembly »	143
5.1.3	Optimisation de la métrologie	143
5.2	Pattern Fidelity Check	144
6	CONCLUSION GENERALE	147
	REFERENCES.....	149
	ANNEXE: LISTE DES PAPIERS PUBLIES	157
	RESUME.....	197
	ABSTRACT	198

REMERCIEMENTS

Avant de rentrer dans les détails scientifiques de ce travail de thèse, j'aimerais remercier toutes les personnes qui, de près ou de loin voire même sans le savoir, m'ont soutenu et aidé pendant ces trois années de long et parfois douloureux labeur.

Merci à Alexandre Tishchenko, directeur de cette thèse, qui nous a quittés récemment et que l'on regrette tous, sans qui cette thèse n'aurait pas eu lieu.

Merci au Laboratoire Hubert Curien et en particulier à Yves Jourlin et Thomas Kämpfe qui ont assuré l'intérim et ont suivi mes travaux tout au long de cette thèse.

Merci à Bertrand Le-Gratiet, mon chef à ST, qui a été mon plus grand soutien dans mon travail pendant ces trois ans. Merci pour ton aide et aussi pour ta motivation et ton courage en tant que seul correcteur et relecteur de cette thèse lors de sa rédaction. J'espère que ton premier thésard ne t'a pas dégouté d'en prendre d'autre plus tard (à priori ça va) ! Petit conseil, ne reprend plus de thésard qui fasse de la Via Ferrata...

Merci à la société STMicroelectronics Crolles2 SAS de m'avoir accueilli pour l'ensemble de ma thèse et en particulier à l'équipe de Litho R&D (trop nombreux pour tous les nommer) qui a eu le plaisir de me subir pendant 3 ans (et de me mettre dans un coin derrière un poteau pour être plus tranquille). Merci aussi à l'équipe OPC que j'ai rejoint après un petit jour de chômage à la fin de la thèse. Merci enfin à Julien Ducoté pour son aide sur les mesures CDSEM et pour la visite guidée de la côte californienne pendant les conférences.

Merci à tous les gens qui m'ont aidé pour les mesures de topographie. Viorel Balan et Yorrick Exbrayat de l'équipe CMP du CEA-LETI et Florent Dettoni de la métro ST pour les mesures sur le Wyko ; Jaydeep Sinha, Dieter Muller, Gavin Simpson et toute l'équipe de développement du PWG chez KLA-Tencor à Milpitas (CA, USA) qui ont fait mes mesures pour moi et ont « failli » m'accueillir pendant une journée sur leur site.

Merci aux ingénieurs d'ASML et de Brion basés un peu partout (Veldhoven au Pays-Bas, Bernin et Crolles en France, Santa Clara et San José en Californie) avec lesquels j'ai beaucoup travaillé et qui m'ont beaucoup appris de par leurs expériences et leurs expertises respectives. Dans le désordre : Tanbir Hasan, Raphaël La-Greca, Christopher Prentice (A really special thank to you!), Wim Tel, Frank Staals, Laurent Dépré, et j'en oublie pour sûr ! Un très grand merci tout particulier à Jan-Willem Gemmink qui non content d'avoir supporté mon travail coté ASML a accepté de rejoindre mon jury de thèse (donc même si tu parles français) : DANK JE WEL!

Merci à quelques rencontres en conf : Philippe Hurat de chez Cadence, les gens de Toshiba, de KLA, d'AMAT, d'ASML, de l'IMEC... avec qui j'ai eu des discussions qui m'ont donné quelques bonnes idées ! ou juste des discussions forts agréables !

Merci à deux chercheurs en particulier : Victor Benno Meyer-Rochow & Jozsef Gal ! Non, je ne les connais pas, ils ne bossent pas en microélectronique et encore moins en optique. Ce sont des ornithologues mais leur travail¹ m'a donné le sourire et réalisé que la recherche ne doit pas forcément (bon, peut-être un peu quand même) se prendre au sérieux.

¹ Meyer-Rochow, V. B.; Gal, J. (2003). "Pressures produced when penguins pooh? calculations on avian defaecation". *Polar Biology*. **27**: 56–58 → Lauréat d'un IgNobel en Dynamique des Fluides promo 2005 !

Je tiens à remercier maintenant toutes les personnes qui en dehors du boulot ont participé à mon soutien psychologique. Liste non-exhaustive :

- A mes deux maisonnées par lesquelles je suis passé depuis 3 ans : merci pour votre prière !
- A Marie-Caroline, Isaure, P'tit Mat', Eric, Marie-Claire, Ajith, Fred, Claire, Angela, Elsa, Elodie, Axel, PEP etc : merci pour les soirées films, les apéros improvisés, les jeux de société, les randos et les scéances de grimpe encore plus à l'arrache.
- Tommy, Tuan, Ajith, Damien, et tous les autres thésards ou alternants ou ingénieurs de ST avec qui j'ai fait une au moins des activités suivantes : spéléo, rando, escalade, canyoning, alpinisme, bar, resto. MERCI les gars !
- A PEP, un coloc prof de maths qui ne sait pas plus que moi si c'est lui qui m'a subi pendant 2 ans ou l'inverse : n'oublie pas, la division par 0 est toujours autorisé à l'appart !
- et à tous les autres que j'oublie bien sûr, vous êtes trop nombreux sur St Jo et à ST et à Phelma et au CEA et dans la Communauté de l'Emmanuel et dans ...

Enfin MERCI à ma famille qui n'a rien compris et ne comprendra toujours rien à ma thèse et mon boulot : c'est ça d'être (presque) le seul à faire des Sciences durs dans une famille d'universitaires en Sciences Humaines.

Bon, place à la science !

TABLE DES FIGURES

<i>Figure 0-1 : Les différentes sources de données et leur utilisation dans l'étude du contrôle du focus. (GDS: Design layout).....</i>	<i>27</i>
<i>Figure 1-1 : Densité de transistors par unité de surface de 1947 à 2015 pour des circuits intégrés depuis leur invention jusqu'à la technologie 14nm. Les dimensions de gravure sont notées (Source : Wikipédia).....</i>	<i>32</i>
<i>Figure 1-2 : L'impact économique de l'industrie du semi-conducteur. Celle-ci contribue à l'échelle de 2700 Milliards de dollars US dans le Produit Intérieur Brut (PIB) mondial en 2012 [6].....</i>	<i>34</i>
<i>Figure 1-3 : Emplois directs par type de business de l'industrie microélectronique en 2012. Au total, on comptait 1,3 millions d'emplois directs dans le monde entiers en 2012 [6].....</i>	<i>35</i>
<i>Figure 1-4 : Coût d'un millions de transistors par nœuds technologiques (source: Handel Jones, IBS).....</i>	<i>37</i>
<i>Figure 1-5 : Les deux chemins du « More Moore » (miniaturisation) et du « More Than Moore » (diversification) chez STMicroelectronics.....</i>	<i>38</i>
<i>Figure 1-6 : Exemple de route d'un produit en technologie 28nm FD-SOI avec 49 masques et 9 niveaux d'interconnexions. Cette route compte 885 étapes de fabrication (dont 126 nettoyages, 473 étapes de métrologie et 237 procédés de fabrication) réparties en 262 opérations et 105 briques.....</i>	<i>40</i>
<i>Figure 1-7 : Fabrication du premier niveau d'interconnexion en 28nm FD-SOI. Il faut 4 briques, 11 opérations et 46 étapes de fabrication (dont 27 de métrologie) pour ce faire.</i>	<i>41</i>
<i>Figure 1-8 : Vue en coupe d'une puce électronique à 5 niveaux d'interconnexions.....</i>	<i>41</i>
<i>Figure 2-1 : Principe de base de la lithographie. Le wafer (substrat) est recouvert avec l'empilement de lithographie, est exposé puis développé. Suite au procédé, le design du masque aura été transféré comme image développée sur le substrat.....</i>	<i>46</i>
<i>Figure 2-2 : Comparaison de plusieurs technologies de photolithographie. Le masque et le wafer sont représentés à l'échelle.....</i>	<i>47</i>
<i>Figure 2-3 : A gauche, scanner de lithographie ASML TWINSCAN NXT:1950i utilisé pour le travail de thèse (Source : ASML).....</i>	<i>48</i>
<i>Figure 2-4 : Représentation de la diffraction d'un faisceau lumineux au travers d'une fente et de sa figure de diffraction.....</i>	<i>49</i>
<i>Figure 2-5 : Evolution de la longueur d'onde et de l'ouverture numérique avec la réduction des dimensions des composants.....</i>	<i>49</i>
<i>Figure 2-6 : Réduire de la longueur d'onde d'exposition permet de s'affranchir d'une diffraction trop importante sans augmenter plus la taille des lentilles.</i>	<i>50</i>
<i>Figure 2-7 : Augmentation de l'ouverture numérique pour une même longueur d'onde. En augmentant la taille des lentilles de projection, plus d'ordres de diffraction sont capturés et on gagne en contraste sur le wafer.....</i>	<i>50</i>
<i>Figure 2-8 : Explication du passage à la lithographie à immersion. Les angles β sont tous 4 fois plus grands que les angles α correspondants en raison de la projection 4X.....</i>	<i>51</i>
<i>Figure 2-9 : Comparaison des images aérienne et du NILS lorsque l'image est construite en focus ou avec un défocus.</i>	<i>53</i>
<i>Figure 2-10 : Principe de l'illumination annulaire. Une illumination de type annulaire permet de capturer partiellement (en rouge) le premier ordre de diffraction sans modifier le NA au prix d'une perte de contraste.</i>	<i>54</i>

<i>Figure 2-11 : Comparaison des images aérienne et du NILS lorsque l'image est construite avec l'ordre 1 complètement ou partiellement capturé par les lentilles de projection.....</i>	<i>54</i>
<i>Figure 2-12: Design, masque et image dans la résine avec et sans OPC. (Source: [14]).....</i>	<i>55</i>
<i>Figure 2-13 : Evolution du facteur k1 dans le temps. En blanc, on peut voir les limites imposées par la réduction du facteur. En bleu, sont exposées quelques solutions technologiques. Les acronymes sont définis ci-dessous. (Source : B. Le-Gratiet)</i>	<i>56</i>
<i>Figure 2-14 : Schéma d'une FEM. A gauche, cartographie des conditions dans chacun des champs (dose et focus) et à droite images MEB correspondantes de la tranchée isolée dans les conditions de la FEM. Les colorisations en vert, jaune, orange et rouge correspondent à la dimension du motif par rapport aux spécifications dimensionnelles du procédé.....</i>	<i>57</i>
<i>Figure 2-15 : Courbes de Bossung du motif dense (P90) et de la tranchée isolée en BEOL 28nm FD-SOI. En haut à droite de chaque graphique se trouvent les images SEM post-lithographie du réseau dense et de la tranchée isolée.</i>	<i>58</i>
<i>Figure 2-16 : Fenêtre de procédé du motif isolée en 28nm BEOL. Le focus optimum (BF pour Best Focus) est de -45nm et la dose optimale (BD pour Best Dose) est 18.6mJ/cm². La profondeur de champ (ou DOF pour Depth of Focus) est de l'ordre de 60nm et la latitude d'exposition (EL) est de ±9% de variation de dose environ.</i>	<i>59</i>
<i>Figure 2-17 : Fenêtres de procédé d'un niveau d'interconnexion en 28nm déterminées par trois méthodes différentes sur le même wafer FEM. Sur la fenêtre CD, la zone verte pâle est la fenêtre de procédé alors que sur les autres, il s'agit de la zone en fond blanc.</i>	<i>61</i>
<i>Figure 2-18 : Réduction de la fenêtre de procédé et effets optique et topologique.....</i>	<i>61</i>
<i>Figure 3-1 : Comparaison du budget focus pour deux situations. Celle de gauche est viable alors que celle de droite est critique.</i>	<i>64</i>
<i>Figure 3-2 : Plan de mesure de la Multi-wafer FEM</i>	<i>66</i>
<i>Figure 3-3 : Comparaison des cartes d'uniformité focus intra-wafer avec la méthode classique et la méthode optimisée.</i>	<i>67</i>
<i>Figure 3-4 : Mesure dimensionnelle CDSEM du P188 en Contact 14nm FD-SOI en fonction du focus de centrage d'exposition de la multi-wafer FEM. Chaque couleur représente un wafer différent. Le graphique de droite représente les Bossung du motif en trois positions sur la plaquette.</i>	<i>67</i>
<i>Figure 3-5 : Bossung réelle du P188 (à droite) extraite des mesures brutes (à gauche) à partir de la méthode précédemment expliquée. La courbe de gauche représente le CD en fonction du « set focus » et celle de droite en fonction du « get focus ».</i>	<i>70</i>
<i>Figure 3-6 : Cartographie de l'uniformité focus sur un wafer 14FD-SOI au niveau Contact (à droite). Cette carte correspond la répartition spatiale de la distribution en focus du wafer 25, en vert foncé sur le graphique de la Figure 3.5.....</i>	<i>70</i>
<i>Figure 3-7 : Illustration des différents niveaux de variabilité avec la production de 10 millions d'unités d'une puce de 7x7mm² (Source : B. Le-Gratiet)</i>	<i>71</i>
<i>Figure 3-8: Images MEB des deux mires dense et isolée ainsi que des 8 Hot Spots étudiés</i>	<i>73</i>
<i>Figure 3-9 : Focus optimaux et profondeurs de champ de plusieurs motifs en BEOL 28nm (à gauche), carte de variation de profondeur de champ (à droite en haut) et de focus (à droite en bas) sur une puce. Les données du graphique de gauche sont des mesures sur silicium et à droite des simulations (source : Tachyon, ASML-Brion).</i>	

<i>Le focus optimal est repéré par le triangle orange et la profondeur de champ par la barre bleue de part et d'autre de cette valeur. L'uDOF est entouré par les pointillés jaunes.</i>	74
<i>Figure 3-10 : Cartographies intra-plaque du test Single Shot Focal (SSF) à gauche, du focus mesuré dans le produit au milieu, et corrélation point à point entre les deux, à droite.</i>	77
<i>Figure 3-11 : Impact de la topographie du substrat sur la valeur du focus d'exposition sur le wafer.</i>	77
<i>Figure 3-12 : En orange et bleu clair : Focus optimums et profondeur de champ par position dans le champ pour trois motifs en BEOL 28nm FD-SOI.</i>	78
<i>Figure 3-13 : Chaque zone d'une puce possède une fonctionnalité spécifique et un design adapté à cette fonctionnalité.</i>	80
<i>Figure 3-14 : Un masque de photolithographie. La partie centrale contient le design du circuit à transférer sur le wafer entouré du cadre. La pellicule est située au-dessus du design pour le protéger. (Source : Wikipédia, Peelliden)</i>	81
<i>Figure 3-15 : Deux types de produits différents. A gauche, un SLR 28nm FD-SOI et à droite un MPW 14nm FD-SOI (à l'échelle)</i>	81
<i>Figure 3-16 : Les étapes de la CMP (Source : Cadence Design Systems)</i>	82
<i>Figure 3-17 : Topographie créée par la cohabitation sur un même design d'une mémoire flash et de circuits logiques (schéma simplifié)</i>	83
<i>Figure 3-18 : Capacités de correction de topographie pendant l'exposition en fonction du type de machine</i>	84
<i>Figure 3-19 : Comparaison des modes de correction de la topographie par le leveling en fonction de la machine. La zone rouge visible dans la première ligne du tableau est la surface sur laquelle le contrôle du focus est appliqué à chaque instant de l'exposition.</i>	85
<i>Figure 3-20 : Illustration des causes de l'erreur de mesure du capteur optique de niveau. A gauche, la réflexion de la lumière incidente est perturbée par l'empilement sur le wafer. A droite, l'effet Goos-Hänchen cause un décalage de la lumière réfléchi par rapport à son point d'incidence. Sur le wafer, les deux effets s'ajoutent. ..</i>	86
<i>Figure 3-21 : L'exposition en « focus meander »</i>	87
<i>Figure 3-22 : A gauche, image SEM du motif mesuré sur FEM le CDSEM. Au milieu, la position des 12 occurrences de la structure mesurée dans le bloc étudiée. A droite, une vue en 3D de la topographie mesurée sur le bloc (mesures Wyko, cf. Chap. 4.1). Les hauteurs dans la zone intéressantes (en bleu et vert) varient entre -15nm et -50nm.</i>	88
<i>Figure 3-23 : L'impact de la topographie locale non corrigeable par le scanner sur le focus optimum mesuré sur silicium. (DOF = Depth Of Focus c.à.d. profondeur de champ et uDOF = Usable DOF i.e. profondeur de champ effective)</i>	88
<i>Figure 3-24 : Corrélation entre la topographie et l'écart au focus optimal pour un unique motif en 14nm FD-SOI</i>	89
<i>Figure 3-25 : Corrélation entre la topographie et l'écart au focus optimal pour plusieurs motifs en 14nm FD-SOI (en bleu et vert) et comparaison avec la distribution de topographie dans le champ.</i>	90
<i>Figure 3-26 : Corrélation entre la topographie et l'écart au focus optimal pour plusieurs motifs critiques présentant un décalage de focus optique important en 28nm</i>	90
<i>Figure 3-27 : Corrélation entre la topographie et l'écart au focus optimal pour plusieurs motifs dans différentes technologies (intégrations, dimensions et procédés différents).</i>	91

<i>Figure 3-28 : Densité de défauts en fonction de la dose et du focus pour un motif isolée en 28nm BEOL extraits à partir de la fenêtre de procédé du motif.....</i>	<i>92</i>
<i>Figure 3-29 : Densité de défauts par wafer en fonction du focus pour 4 motifs plus ou moins critiques en 28nm FD-SOI.....</i>	<i>93</i>
<i>Figure 3-30 : Représentation graphique de la convolution de la fenêtre de procédé théorique de P1 (Image aérienne), de ses occurrences dans le produit (design) et de la topographie local (mesures).....</i>	<i>94</i>
<i>Figure 4-1 : Le procédé de leveling du scanner (NCE = non-correctable error soit les résiduels de topographie non corrigéable par le scanner).....</i>	<i>98</i>
<i>Figure 4-2 : La méthode de mesure par assemblage d'image sur le Wyko</i>	<i>98</i>
<i>Figure 4-3 : Mise à niveau des données Wyko dans Gwyddion.....</i>	<i>100</i>
<i>Figure 4-4 : Rotation et troncature des données</i>	<i>101</i>
<i>Figure 4-5 : Mesures Wyko champ complet des produits en 14nm FD-SOI et en 28nm. L'échelle est en nanomètres de topographie.....</i>	<i>101</i>
<i>Figure 4-6 : Distribution de topographie sur un champ en 14nm FD-SOI Contact en 28nm BEOL.....</i>	<i>102</i>
<i>Figure 4-7 : A gauche, courbures des wafers en Contact 14nm FD-SOI à l'étape de lithographie. La mesure du wafer POR pour Process of record (i.e. wafers standards tri-couche) est représentée en 3D à droite. (Source : KLA Tencor).....</i>	<i>104</i>
<i>Figure 4-8 : Défocus estimé à l'échelle d'une plaquette entière. (Source : KLA Tencor)</i>	<i>104</i>
<i>Figure 4-9 : Cartographies des mesures de topographie avec le scanner (à gauche), le PWG (au milieu) et le Wyko (à droite).....</i>	<i>104</i>
<i>Figure 4-10 : Comparaison des mesures Wyko et scanner (Source : ASML).....</i>	<i>105</i>
<i>Figure 4-11 : Corrélation entre le Wyko et le PWG. A gauche, la corrélation sur l'ensemble des points de mesures et, à droite, sur la distribution sans les valeurs extrêmes que ne peut pas résoudre le PWG.</i>	<i>105</i>
<i>Figure 4-12 : Algorithme de la PLS adapté au logiciel SIMPCAP.....</i>	<i>107</i>
<i>Figure 4-13 : Graphique de VIP fourni par le logiciel SIMCAP. Les barres vertes correspondent à un VIP > 1, les grises à un VIP < 0.8 et en orange entre 0.8 et 1.</i>	<i>111</i>
<i>Figure 4-14 : Schéma explicatif de ce que représente le périmètre de Grille sur Active. Il s'agit du périmètre des polygones rouges.....</i>	<i>113</i>
<i>Figure 4-15 : Corrélation originelle à partir de laquelle la méthodologie a été développée</i>	<i>114</i>
<i>Figure 4-16 : Les quatre phases de la construction du modèle PLS de topographie.....</i>	<i>114</i>
<i>Figure 4-17 : Méthodologie pour la modélisation de la topographie par analyse PLS [32]</i>	<i>115</i>
<i>Figure 4-18 : Etapes déjà réalisées de l'intégration (cf. Figure 1-6 p.38) et empilement présent sur le wafer au moment de l'exposition du premier niveau de BEOL en 28nm.....</i>	<i>116</i>
<i>Figure 4-19 : Coefficients de corrélation du modèle de PLS du leveling pour chaque niveau de lithographie 193nm immersion du FEOL, MEOL et pour la première ligne métallique en 14nm FD-SOI.</i>	<i>117</i>
<i>Figure 4-20 : Extrait d'un design sur le logiciel Calibre de chez Mentor Graphics. Le design complet est composé de plusieurs niveaux superposés les uns aux autres.....</i>	<i>118</i>
<i>Figure 4-21 ; Définition des deux types de densités extraites du GDS</i>	<i>119</i>
<i>Figure 4-22 : Principes de combinaisons de niveaux de design entre eux</i>	<i>120</i>
<i>Figure 4-23 : Définition des niveaux d'épitaxies Source / Drain.....</i>	<i>121</i>

<i>Figure 4-24 : Comparaison de données du leveling du scanner au niveau Contact 14nm FD-SOI et de l'agencement spatial du produit sur le masque. La partie en jaune sur le produit est la zone mesurée par le capteur.</i>	122
<i>Figure 4-25 : Performances respectives des différents modèles du leveling [32]</i>	123
<i>Figure 4-26 : Quelques ré-échantillonnages des mesures Wyko de la puce de calibrage du modèle (PROLIGHT) avec Gwyddion.</i>	125
<i>Figure 4-27 : Performances des modèles issus du balayage de différentes échelles spatiales sur le 14nm FD-SOI</i>	125
<i>Figure 4-28 : Pente de la corrélation entre le modèle et les mesures en fonction de la taille du pixel</i>	126
<i>Figure 4-29 : Comparaison des coefficients et pentes de corrélation pour les modèles avant (« complet ») et après (« léger ») tri des paramètres en fonction de leurs VIP respectifs à plusieurs échelles spatiales.</i>	126
<i>Figure 4-30 : Temps d'extraction des densités de design avec Calibre en fonction de la surface de la puce. A gauche, pour différentes tailles de pixels et à droite, en modifiant le nombre de composantes pour la PLS.</i>	127
<i>Figure 4-31 : Les différentes puces du masque 14nm FD-SOI utilisées pour la construction du modèle (échelle X10) et leurs positions respectives dans le champ du MPW. (1) Prolight, (2) FE_RX, (3) SIMS, (4) Promo, (5) ESDRF.</i>	128
<i>Figure 4-32 : Comparaison de la topographie mesurée et de la topographie modélisée dans les zones de fort dishing.</i>	129
<i>Figure 4-33 : Cartographies de la topographie mesurée avec le Wyko sur le PROMO en 14nm FD-SOI et de la topographie prédite par les modèles PLS à 2.5, 50 et 100μm de résolution spatiale pour la même puce.</i>	130
<i>Figure 4-34 : Cartographie des mesures et du modèle PLS pour le produit 28nm BEOL et corrélation entre le modèle et les mesures.</i>	133
<i>Figure 5-1 : Les différentes optimisations proposées à la suite de la thèse.</i>	137
<i>Figure 5-2 : Risque lié à la cohabitation d'une forte topographie et d'une zone critique en focus.</i>	138
<i>Figure 5-3 : Schéma de principe de la détérioration du focus dans une zone critique par correction d'une zone non-critique voisine.</i>	138
<i>Figure 5-4 : Schéma de principe de l'optimisation du leveling</i>	140
<i>Figure 5-5 : Application des deux profils de correction scanner sur les mesures Wyko intra-champ en Contact 14nm FD-SOI. Les cartographies obtenues sont les erreurs non-corrigeables haute fréquence. La différence (échelle différente) donne le delta entre les deux corrections. Plus la valeur du delta est élevée et plus la version optimisée est performante pour cette position.</i>	140
<i>Figure 5-6 : Nombre de positions (pixel de 2.5μm) en spécification focus (écart au focus optimal inférieur à 15nm pour les zones critiques et inférieur à 25nm pour les zones non-critiques) pour les profils de correction optimisé et non-optimisé.</i>	140
<i>Figure 5-7 : Principe de l'émulation de wafer.</i>	144

LISTE DES TABLEAUX

<i>Tableau 1-1 : Quelques exemples de produits électroniques au fur et à mesure de la loi de Moore</i>	<i>31</i>
<i>Tableau 1-2 : Classement des 20 premières sociétés IDM, Fonderies et Fabless pour 2016 [Source : IC Insight]</i>	<i>35</i>
<i>Tableau 3-1 : Tableau récapitulatif de la matrice de focus multi-wafer. BF (Best Focus) désigne la valeur optimale de focus utilisée en production.</i>	<i>65</i>
<i>Tableau 3-2 : Liste des effets impactant le focus, classés par sources et mécanismes.....</i>	<i>72</i>
<i>Tableau 3-3 : Conditions expérimentales du test SSF.....</i>	<i>76</i>
<i>Tableau 3-4 : Images SEM de quatre motifs en conditions optimales de procédé et hors focus</i>	<i>92</i>
<i>Tableau 4-1 : Paramètres de mesures Wyko réalisées sur les wafer 28 et 14nm FD-SOI.....</i>	<i>99</i>
<i>Tableau 4-2 : Wafers envoyés chez KLA Tencor pour les mesures sur le WaferSight PWG</i>	<i>103</i>
<i>Tableau 4-3 : Tableau des paramètres et des notations mathématiques. La troisième colonne fait le lien entre les notations mathématiques et le cas étudié de la modélisation de topographie à partir du design.</i>	<i>108</i>
<i>Tableau 4-4 : Comparaison des avantages et inconvénients des deux méthodes de modélisation haute fréquence de la topographie avec la régression PLS.....</i>	<i>124</i>
<i>Tableau 4-5 : Résultats R² des différents modèles</i>	<i>129</i>
<i>Tableau 4-6 : Coefficients et résultats des modèles combinés pondéré et non pondéré</i>	<i>131</i>
<i>Tableau 4-7 : Classement des niveaux de design et des types de densité les plus influents sur le modèle du Contact en 14nm FD-SOI.....</i>	<i>132</i>
<i>Tableau 4-8 : Classement des niveaux de design et des types de densité les plus influents sur le modèle en BEOL 28nm.....</i>	<i>134</i>
<i>Tableau 5-1 : Impact des différents niveaux de masques sur la topographie à plusieurs échelles spatiales</i>	<i>142</i>

GLOSSAIRE

Terme	Définition
AGILE	Le capteur AGILE (pour Air Gauge Improved LEveling) est un capteur pneumatique qui permet de corriger les erreurs de la mesure optique de topographie dans le scanner.
BARC	Couche polymère anti-réfléctive déposée sous la résine photosensible pour éliminer les effets de réflexion parasites.
BaseLiner®	Routine de calibration du focus pour compenser les dérives du scanner
BEOL	Back-End Of Line. Correspond à la fabrication des lignes d'interconnexions métalliques entre les transistors dans une puce.
BF	Le Best Focus ou focus optimum est la position du plan image lors de l'exposition d'un motif.
CD	La dimension critique (critical dimension) est la plus petite dimension présente sur un design. Par extension, on utilise le terme pour toute dimension de motif.
CDSEM	Le CDSEM est un microscope électronique à balayage automatisé pour la mesure de CD pour les contrôles en ligne.
Cluster	Ensemble de deux machines composé d'un scanner et d'une piste permettant de réaliser l'ensemble des étapes du procédé de photolithographie.
CMOS	Complementary Metal Oxide Semiconductor. Il s'agit du cœur des technologies de la microélectronique consistant en l'association de deux types de transistors pour réaliser des fonctions logiques.
D/T/QP	Double / Triple / Quadruple Patterning. Techniques d'expositions multiples permettant de poursuivre la loi de Moore quand la diffraction devient trop importante.
DOF	Depth Of Focus ou Profondeur de champ. Il s'agit du volume au-dessus et en dessous du plan focal dans lequel l'image aérienne est nette.
DTCO	Design Technology Co-Optimisation. technique d'optimisation conjointe du design et des procédés pour limiter les risques de perte de rendement.
Far BEOL	Dernière partie de la fabrication des puces après le BEOL. Cela correspond à la réalisation des plots sur lequel les connectiques extérieurs à la puce sont branchées.

FD-SOI	Fully Depleted Silicon On Insulator. Technologie dans laquelle les transistors sont fabriqués sur un substrat composé d'une fine couche de silicium sur un box d'isolation en oxyde. L'intérêt de la technologie est de diminuer les pertes de courant dans le substrat.
FEM	Focus Exposure Matrix. Méthode de détermination des conditions optimales d'exposition par matriçage de plusieurs conditions sur le même wafer.
FEOL	Front-End Of Line. Fabrication des transistors et des résistances sur le substrat. Le FEOL est suivi du MEOL puis du BEOL.
GDS	Grid Design System. Format de fichier de design qui contient l'ensemble du plan du circuit électronique d'une puce. Par extension, le GDS désigne le plan de la puce même si celui-ci est enregistré dans un autre format de fichier.
HDFM	Hyper Dense Focus map. Méthode de détermination des variabilités du focus d'exposition à l'échelle d'un wafer complet développé au cours de cette thèse.
Image aérienne	Image optique de l'objet composée des intensités lumineuses en chaque point
Image développée	Image transférée dans la résine photosensible après développement de celle-ci
Image latente	Image transférée dans la résine photosensible par transformation photochimique de celle-ci suite à l'exposition du masque
k_1	Facteur traduisant la complexité du procédé de lithographie. Plus k_1 est petit est plus le procédé est difficile à mettre en œuvre.
Leveling	Mise à niveau en français. Le procédé de leveling a lieu dans le scanner de lithographie avant l'exposition et permet de compenser la topographie de la plaquette pour diminuer le risque d'erreur de positionnement en focus.
Masque	Objet sur lequel le plan de la puce électronique est dessiné par des motifs laissant ou non passer la lumière. L'information contenue sur le masque est projetée sur le wafer pendant l'exposition de la plaquette.
MEOL	Middle-End Of Line. Fabrication des niveaux de transitions entre les transistors (FEOL) et les interconnexions métalliques (BEOL) qui les connectent entre eux.
MPW	Multi Project Wafer. Type de masque contenant plusieurs contributions différentes (prototypes, structures et motifs de développement).
NA	Numerical Aperture. Capacité d'un système optique à capturer des ordres de diffraction élevés.

OPC	Optical Proximity Correction. Méthodes de compensation des effets de diffraction permettant de réaliser des masques avec lesquels l'image transférée dans la résine n'est pas déformée et reste fidèle au design.
PLS	Régression statistique par la méthode des moindres carrés. La PLS est ici utilisée pour la calibration d'un modèle de prédiction de la topographie à partir du design de la puce.
PW	Process Window ou fenêtre de procédé. Ensemble de conditions de procédé autour du point de fonctionnement idéal pour lesquels l'image transférée dans la résine est de bonne qualité.
RET	Reticle Enhancement Techniques. Techniques d'amélioration du masque complémentaires à l'OPC et permettant d'augmenter le contraste de l'image aérienne.
Scanner	Machine permettant de projeter l'image du masque sur le wafer en scannant celui-ci une centaine de fois par plaquette.
Sigma inner et Sigma outer	Taux de remplissage de la pupille d'entrée du scanner par le faisceau laser.
Slit	Sur le scanner, la slit correspond à la zone qui est exposée à chaque instant. La slit fait 6mm par 26mm.
SLR	Single Layout Reticle. Type de masque ne contenant qu'un seul produit matricé plusieurs fois dans le champ.
Slurry	Mélange physico-chimique utilisée pour les étapes de polissage de la plaque de silicium au cours de la fabrication des puces. Le slurry est composé d'un réactif chimique et de billes abrasives.
SSF	Single Shot Focal. Routine de détermination des erreurs focus du scanner au niveau du wafer.
Track	Machine servant aux dépôts, recuits et développements des résines photosensibles.

0 INTRODUCTION

Depuis l'évènement de la photolithographie à immersion introduite sur les nœuds technologique CMOS 45nm, la notion de lithographie holistique a pris de plus en plus d'ampleur. Ce travail s'inscrit dans ce type d'approche, aussi est-il opportun d'introduire dès à présent cette notion. Ce terme a été introduit par le sud-africain Smuts dans *Holism and evolution*, pour décrire le caractère symbiotique de l'évolution : l'holisme est « *la tendance dans la nature à constituer des ensembles qui sont supérieurs à la somme de leurs parties, au travers de l'évolution créatrice* »². L'holisme peut s'appliquer à de nombreux domaines (philosophie, neurologie, sociologie,...) et peut se définir comme suit :

Holistique, adj., du grec *holos* > entier

Qui relève de l'holisme, qui s'intéresse à son objet comme constituant un tout. (cnrtl.fr³)

Le Larousse en ligne définit l'holisme comme une « *doctrine philosophique défendue notamment par Duhem, et selon laquelle ce n'est jamais un énoncé scientifique isolé, mais le corps tout entier de la science qui affronte le verdict de l'expérience.* »

Le CNRTL (Centre National de Ressources Textuelles et Lexicales) ajoute que l'holisme « *[considère] les phénomènes comme des totalités*⁴ ».

Dans le cas d'un domaine scientifique, une approche holistique consiste à aborder un sujet non pas uniquement par lui-même dans sa manière la plus évidente mais en faisant apparaître des corrélations entre des sources de données très diverses voire parfois considérées comme indépendantes les unes des autres. L'approche holistique d'un problème industriel peut s'apparenter à du « data mining » (exploration de données en français) dans laquelle le sujet de l'étude est confronté à de très nombreuses données afin de mettre en évidence des liens de cause à effet. Une approche holistique en microélectronique consiste à considérer chaque étape de fabrication, chaque composant du circuit, chaque propriété comme étant dépendant de l'ensemble de la puce. Les facteurs ayant une influence importante sur une étape en particulier sont multiples et il est possible de les diviser en plusieurs catégories : le design, l'intégration, les procédés, les machines, les matériaux. Cette thèse se consacre en particulier à une vision holistique du contrôle du focus en lithographie optique à immersion.

La fabrication des circuits intégrés nécessite plusieurs centaines d'étapes parmi lesquelles on compte des dépôts de matériaux, des recuits, des nettoyages, des implantations ioniques (dopage des semi-

² Smuts, Jan. '*Holism and Evolution*'. Londres: Macmillan & Co Ltd, 1926, 362 p.

³ <http://www.cnrtl.fr/definition/holistique> (27 avril 2106)

⁴ Sumpf-Hug. 1973

conducteurs) et des étapes de structuration (« patterning ») comme la photolithographie, la gravure et le polissage

La photolithographie précède généralement la gravure et les implantations car elle permet de définir les zones à graver ou à doper. Le procédé photolithographique est étroitement lié au design de la puce, lequel va être transféré optiquement par l'exposition de l'image d'un photomasque sur un film photosensible déposé sur le silicium. Avec le prolongement de la loi de Moore⁵, les limites optiques de la photolithographie sont désormais atteintes. La diffraction de la lumière est telle que seule une partie de l'information du masque parvient sur le wafer (« low k1 lithography » ou lithographie faible k1), limitant le contraste de l'image. Le retard de l'EUV (Extrême UV) conduit à l'utilisation de double exposition pour réaliser des éléments du circuit précédemment fait en une seule fois. La double exposition impose un contrôle plus grand des variabilités sur la plaquette. Avec la complexification de l'intégration tant d'un point de vue chimique (matériaux de moins en moins conventionnels, accroissement du nombre d'éléments chimiques différents sur une puce, ...) que design de circuit (antennes Radio Fréquence ou Wifi, co-intégration Analogique-Logique, mémoires embarquées, capteurs divers, ...), les effets croisés sont plus nombreux et viennent dégrader les marges de procédés.

Bien que la loi de Moore soit toujours d'actualité, il est à noter que depuis la technologie 28nm, réduire les dimensions ne permet plus de réduire le coût d'un transistor. Ainsi, le nœud 28nm est attendu pour rester très longtemps ce qui imposera d'y apporter de nombreuses améliorations, pour lesquelles un meilleur contrôle des variabilités est indispensable. Le 28nm est aussi pour cette même raison un nœud privilégié pour le développement de dérivatifs dans le futur.

Dans ce contexte, il devient nécessaire de chercher des corrélations entre des paramètres très variés, soit par leur source soit par leur nature. Ceux-ci étaient auparavant traités et corrigés indépendamment les uns des autres. La mise en évidence de ces corrélations permettant à terme de prévoir les défauts et de sécuriser le rendement en production.

Cette thèse s'inscrit intégralement dans cette approche holistique multi-source du contrôle de procédé et ceci dans le cadre spécifique de la photolithographie et du contrôle du focus. Le schéma ci-dessous (Figure 0-1) en est une illustration. Il y est représenté les sources de données qui ont été exploitées et valorisées pour montrer les paramètres importants dans l'apparition de variabilité dans le focus d'exposition du wafer et réussir à en prévoir une partie. Ces données sont issues du design, de la machine et du silicium.

⁵ La loi de Moore régit la miniaturisation des composants électroniques. Cf. Chap. 1.1

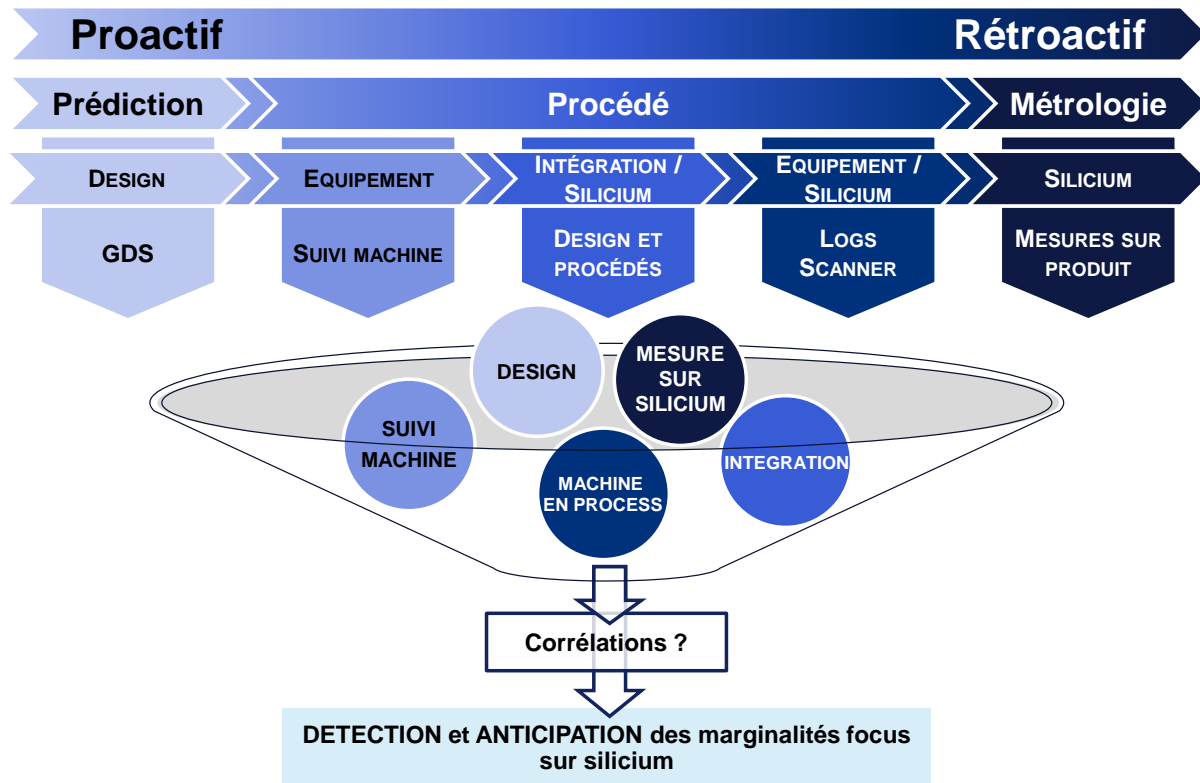


Figure 0-1 : Les différentes sources de données et leur utilisation dans l'étude du contrôle du focus. (GDS: Design layout)

Du design de la puce et de l'intégration, il est possible de récupérer des données sur le fichier GDS (Graphic Database System) qui contient l'intégralité du plan de la puce, niveau par niveau. Combiné à l'intégration qui est le choix de l'enchaînement des étapes de fabrication et des matériaux, on obtient la puce telle qu'elle est voulue par le designer.

Coté machine, le scanner de lithographie mesure systématiquement toutes les plaquettes de production pour corriger les distorsions de 2D (déformation X/Y) et 3D (topographie) des wafers afin de projeter une image avec la meilleure netteté et le meilleur alignement possible pendant l'exposition. A chaque plaque, une quantité importante de données (> 12MB) dont l'analyse permet de fournir beaucoup d'informations avec un poids statistique significatif. Ces données ne sont que très rarement utilisées en dehors de l'analyse de problèmes de production. Pour limiter les dérives dans la fabrication, les machines sont suivies de manière régulière pour les calibrer par rapport à elles-mêmes (monitoring). Ces routines qui existent déjà (Baseliner, Single Shot Focal, FOCAL, ...) sont réalisées sur des wafers de silicium vierge spécifiques appelés NPW (Non-Product Wafers). Il est possible d'évaluer la composante machine de la variation de focus que l'on a sur produit. Ces données seront complétées par des mesures silicium du produit.

Sur les lots de production, des mesures du circuit et des motifs de suivi de production après exposition vont permettre de déterminer le focus optimum d'exposition en chaque point de mesure. Une cartographie de la répartition du focus sur le wafer est alors obtenue. Des mesures de topographie ont

été réalisées sur les mêmes lots avant exposition. Des corrélations avec le design, la métrologie du scanner ainsi que les mesures post exposition sont attendues.

Le focus en photolithographie est une grandeur relative du positionnement vertical du wafer par rapport au plan image de projection dont la valeur dépend de (i) la calibration de la machine et de ses limites (ii) de la forme et de la taille du motif à imprimer (iii) de la nature du substrat. Un motif hors focus sera imprimé flou et conduira à la formation de défauts locaux dont leur gravité dépendra de leur impact sur le fonctionnement de la puce. Les marges d'erreurs sont très réduites (<120nm) et s'amenuisent avec la miniaturisation et le développement de technologies dérivées. Le contrôle du focus est extrêmement critique pour sécuriser la production et assurer un bon rendement.

Le présent manuscrit a pour objectif de montrer la complexité à laquelle nous sommes confrontés lorsque nous devons gagner les nanomètres permettant la sécurisation des risques liés au focus.

La première partie présente le contexte économique et industriel de la microélectronique, lequel joue un rôle primordiale dans la définition des challenges. Il est important de comprendre l'écosystème macroéconomique car l'ensemble de ce travail de thèse a été menée sur le site de R&D et de production de Crolles de STMicroelectronics et en grande partie sur des lots de productions. A chaque avancée technologiques, les interactions entre les différent éléments de la fabrication (designs, procédés, matériaux, machines) deviennent de plus en plus importantes, ce qui resserrent encore les marges de variabilité. Cela est d'autant plus vrai pour les technologies dérivées qui par nature ajoutent une complexité spécifique à leurs designs et options.

La deuxième partie porte une attention particulière sur la photolithographie, facteur clé de la course à la miniaturisation. Nous présenterons d'une part le cadre historique et économique de la lithographie (acteurs, loi de Moore, ...) avant de développer des considérations plus techniques. Le procédé tel que nous le trouvons en production chez STMicroelectronics sera décrit et détaillé. Enfin, les défis de la résolution de l'image dans la résine photosensible, de pertes de contraste, de multiplicité des motifs à imprimer seront discutés en termes de profondeur de champ disponible. Nous terminerons cette partie en rappelant les challenges du travail de thèse.

La troisième partie est consacrée à la mise en évidence des effets et sources de variabilité focus que l'on rencontre lors de l'exposition d'une plaquette. Pour cela, les méthodes de la multi-wafers FEM (Focus Exposure Matrix) et de la cartographie dense de focus (HDFM ou « hyper dense focus map ») développées au cours de ce projet seront présentées. Ces deux méthodes d'analyses mettent en œuvre la vision multi-source et holistique introduite précédemment. L'analyse des résultats de ces tests montrera les sources de variabilité, leur part dans le budget focus total ainsi que les solutions déjà existantes de minimisation et de contrôle. Nous terminerons en soulignant l'effet de modulation du design macroscopique de la puce sur ces effets.

La quatrième partie se concentre sur un effet particulier : celui du design sur la topographie et le focus. Nous montrerons qu'il est possible de prédire les défocalisations sur produit à partir du design du circuit. La méthode de régression multilinéaire PLS (Partial Least Square ou méthode des moindres carrés partiels) est l'outil qui a été privilégié dans la création d'un modèle prédictif. La méthodologie intégrale de cette modélisation sera donnée. Les informations annexes (importance de chaque composante, tri et classement des paramètres) que nous fournit la PLS seront discutées à la suite. Elles permettront d'expliquer et d'anticiper les conséquences en termes de probabilité de défauts dans chaque zone de la puce en fonction du modèle et du procédé de lithographie.

Le cinquième et dernier chapitre de cette thèse est consacré à la proposition de plusieurs solutions permettant d'anticiper et/ou minimiser des effets de focus. La définition de zones d'intérêts en fonction de l'utilisation conjointe de données machine, design, modèle et mesures et leur utilisation seront présentées. Une méthode d'optimisation de la correction de topographie par le scanner qui est désormais en discussions chez ASML suite à ce travail et à l'évaluation de cette méthode sera décrite. Enfin, nous présenterons la méthode de l'émulation de cartographie de focus (« emulated focus map »), méthode holistique permettant de visualiser le focus sur plaquette sans avoir recours à des campagnes de mesure coûteuses en temps et en wafers.

CHAPITRE 1

1 L'ECOSYSTEME INDUSTRIEL DE LA MICROELECTRONIQUE

1.1 LA MICROELECTRONIQUE ET L'INDUSTRIE DES SEMI-CONDUCTEURS

1.1.1 9 décades de densité de transistors en 50 ans !

L'année 2015 a vu l'anniversaire des cinquante ans de la loi de Moore, cette loi qui a régi l'ensemble de la miniaturisation et des développements en électronique depuis l'invention de la puce électronique. Cette course à la réduction des dimensions et à l'augmentation de la densité de transistors a conduit au développement de puces électroniques de plus en plus puissantes et intégrant de plus en plus d'options servant des applications de plus en plus compliquées. Ces avancées technologiques ont permis le avènement de toute l'électronique utilisée quotidiennement depuis les premiers ordinateurs de la taille d'une salle jusqu'au smartphone tenant dans la main.

Le tableau 1-1 rapporte les caractéristiques de produits d'électronique à différentes époques afin de mieux appréhender l'impact des évolutions technologiques sur leurs performances.




Nom du produit	Apollo Guidance Computer	IBM PC 5150	Samsung Galaxy S5
Accessibilité	Ordinateur de bord de la mission lunaire Apollo 11	Premier IBM Personal Computer	Smartphone accessible au grand public
Année	1969	1981	2014
Nœud technologique (nm)	> 10 000	3 000 (÷3)	28 (÷100)
RAM	64KB	256KB (x4)	2GB (x10 ⁴)
ROM	Aucune	40KB	32GB (x10 ⁶)
Fréquence d'horloge (MHz)	0.043	4.77 (x100)	2500 (x500)
Poids (kg)	35	17 (÷2)	0.145 (÷100)
Prix (en équivalent dollar 2015)	~ 150 000	7 822 (÷20)	600 (÷10)
Image			

Tableau 1-1 : Quelques exemples de produits électroniques au fur et à mesure de la loi de Moore

L'ensemble du développement de la microélectronique repose sur l'invention d'un composant, le transistor qui du fait de sa taille, de son intégrabilité en une seule puce pour permettre de réaliser des fonctions complexes et de sa fiabilité a rapidement remplacé les tubes à vide des premiers ordinateurs.

C'est après environ 100 ans de développement de la physique du solide [1] [2] et grâce aux besoins (radio et radar) de la Seconde Guerre Mondiale, qu'en 1947 John Bardeen, William Shockley et Walter Brattain, travaillant chez Bell Labs, inventent le transistor [3]. Celui-ci n'est pas encore basé sur du Silicium mais est en Germanium. Suivent plusieurs autres innovations permettant la création de d'autres dispositifs [4] [5].

En 1958, Jack Kilby (Texas Instruments) invente le premier circuit intégré. Il s'agit simplement d'un unique transistor connecté à des résistances sur un substrat de 100mm², soit une densité de 1 transistor par cm². Un an plus tard, en 1959, Robert Noyce de Fairchild Semiconductor fabrique lui aussi un circuit intégré composé de deux transistors bipolaires sur un même substrat. C'est le début de l'intégration de l'électronique et l'invention de la puce, composant regroupant plusieurs transistors sur le même support, ce qui permet un gain de place mais aussi de vitesse de calcul.

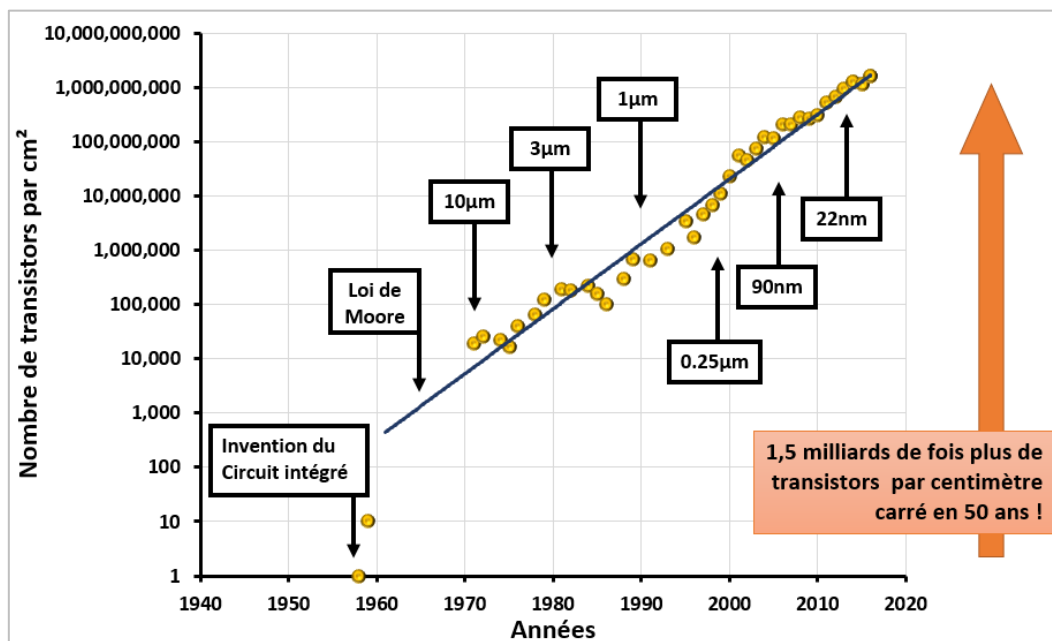


Figure 1-1 : Densité de transistors par unité de surface de 1947 à 2015 pour des circuits intégrés depuis leur invention jusqu'à la technologie 14nm. Les dimensions de gravure sont notées (Source : Wikipédia)

Il faudra attendre plus d'une décennie pour voir apparaître en 1971 le premier microprocesseur de l'histoire : l'Intel 4004. Il est gravé en 10µm et contient 2300 transistors planaires sur 12mm², soient environ 19 000 transistors par cm². 45 ans plus tard, en 2016, Intel produit le 22-core Xeon Broadwell-E5 gravé en 14nm FinFet avec vingt-deux cœurs et 7 200 000 000 transistors sur 456 mm², soient plus de 1.5 milliards de transistors par cm². Le record du nombre de transistors sur une puce est battu en 2015 par Oracle et TSMC (Taiwan Semiconductor Manufacturing Company) : le Sparc M7

avec ses 32 cœurs, ses 256 threads et ses 10 Milliards de transistors en 20nm FinFET, soient 1.5 milliards de transistors par cm² sur une puce de plus de 600mm² !

Ainsi entre 1958 et le premier circuit intégré et 2015, le nombre de transistors sur une même puce est passé de 1 à 10 000 000 000, soit une augmentation de densité de transistor par centimètre carré d'un facteur de 9 décades ! En prenant en référence le premier microprocesseur d'INTEL en 1971 et ses 20000 transistors par cm², nous retrouvons précisément la loi de Moore qui prédit un doublement de nombre de transistor par cm² tous les 18 mois. La Loi de Moore est représentée sur le graphique de la Figure 1-1.

1.1.2 L'industrie du semi-conducteur

L'industrie du semi-conducteur est un des secteurs les plus innovants du dernier siècle. En 50 ans, elle a révolutionné la vie de milliards d'êtres humains. En science, elle a fourni la puissance de calcul et les outils permettant la conquête spatiale ou la recherche du cœur de la matière. Dans la vie quotidienne, elle est à l'origine du développement des télécoms et de nombreux loisirs comme les jeux vidéo par exemple. Elle a aussi radicalement modifié l'industrie en introduisant la robotisation et l'automatisation de nombreuses usines. On parle désormais du programme Industrie 4.0 lequel est ancré sur l'utilisation des nouveaux moyens de communication, de contrôle et d'asservissement pour augmenter la productivité. Enfin le développement de la microélectronique a vu la naissance d'un nouvel écosystème macroéconomique et industriel de pointe qui est désormais très important tant d'un point de vue création de valeur (cf. Figure 1-2) que de celui du nombre d'emplois directs et indirects (cf. Figure 1-3).

Il existe plusieurs types d'acteurs dans l'écosystème de la fabrication des puces électroniques. On distingue les suivants : IDM, « Fabless », Fonderies, OSAT et les fournisseurs EDA/IP.

Les sociétés IDM (Integrated Device Manufacturers, ou Fabricants de Composants Intégrés) couvrent l'ensemble du travail de design, de fabrication, d'emballage, de test et de vente de leurs produits semi-conducteurs en interne. On peut nommer Intel, Texas Instruments et Micron aux USA, Samsung et SKHynix en Corée du Sud, Toshiba et Renesas au Japon, et STMicroelectronics, Infineon et NXP pour l'Europe. Ce sont ces entreprises qui dominent l'écosystème industriel des semi-conducteurs avec $\frac{2}{3}$ du chiffre d'affaire généré (soit \$1350 milliards) et plus de 60% de emplois directs en 2012.

Les « Fabless » (ou « sans unités de production ») dessinent leurs propres puces et les vendent à leurs clients mais sous-traitent la fabrication, l'emballage et les tests à un tiers. Qualcomm, Broadcom ou Nvidia sont parmi les sociétés phares de cette catégorie et ont récemment intégré le TOP10 et détrôné nombre de sociétés IDM.

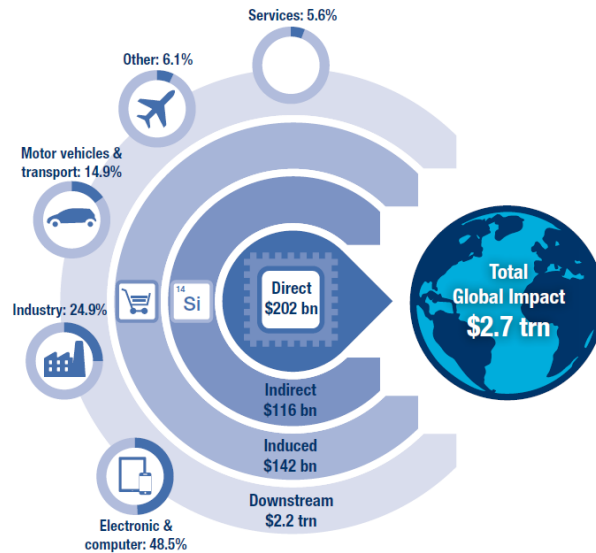


Figure 1-2 : L'impact économique de l'industrie du semi-conducteur. Celle-ci contribue à l'échelle de 2700 Milliards de dollars US dans le Produit Intérieur Brut (PIB) mondial en 2012 [6].

Les fonderies sont des sociétés qui sont spécialisées dans la fabrication des puces dessinées par les « fabless » et dans la sous-traitance pour les IDM. Les plus notables sont TSMC (Taiwan), UMC (Taiwan), Samsung (Corée du Sud), Global Foundries (US) et SMIC (Chine). Taiwan en particulier regroupait près de 57% des emplois mondiaux en fonderies en 2012.

Les sociétés d'EDA, ou Electronic Design Automation (Automatisation des Designs Electroniques), supportent les fabricants et les fonderies en créant des logiciels permettant de dessiner et produire les puces. On retrouve par exemple Cadence, Mentor Graphics, Synopsys ou Brion. Les fournisseurs IP (Intellectual Property ou Propriété Intellectuelle) développent les designs de cellules qui vont servir à dessiner les circuits complets ensuite. On peut citer ARM Holdings par exemple.

Les OSAT (Outsourced Semiconductor Assembly and Test, ou Sous-traitant d'Assemblage et de Test) sont des sociétés spécialisées dans l'emballage et les tests de puces électroniques travaillant pour les Fabless et certaines IDM's.

Les deux figures suivantes (Figure 1-3 et Tableau 1-2) présentent le nombre d'emplois directs générés par l'industrie du semi-conducteur au niveau mondial et le classement prédictif des compagnies pour l'année 2016.

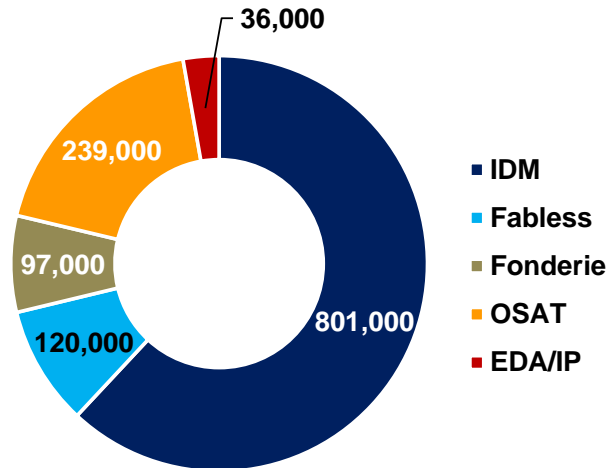


Figure 1-3 : Emplois directs par type de business de l'industrie microélectronique en 2012. Au total, on comptait 1,3 millions d'emplois directs dans le monde entiers en 2012 [6].

1Q16 Top 20 Semiconductor Sales Leaders (\$M, Including Foundries)

1Q16 Rank	1Q15 Rank	Company	Headquarters	1Q15 Tot Semi	1Q16 Tot Semi	1Q16/1Q15 % Change
1	1	Intel*	U.S.	12,067	13,115	9%
2	2	Samsung	South Korea	9,336	9,340	0%
3	3	TSMC (1)	Taiwan	6,995	6,122	-12%
4	7	Broadcom Ltd. (2)*	Singapore	3,679	3,550	-4%
5	4	Qualcomm (2)	U.S.	4,434	3,337	-25%
6	5	SK Hynix	South Korea	4,380	3,063	-30%
7	6	Micron	U.S.	4,061	2,930	-28%
8	8	TI	U.S.	2,940	2,804	-5%
9	10	Toshiba	Japan	2,619	2,446	-7%
10	9	NXP*	Europe	2,636	2,224	-16%
11	12	Infineon	Europe	1,666	1,776	7%
12	13	MediaTek (2)	Taiwan	1,506	1,691	12%
13	11	ST	Europe	1,700	1,601	-6%
14	14	Renesas	Japan	1,470	1,415	-4%
15	17	Apple (2)**	U.S.	1,260	1,390	10%
16	15	GlobalFoundries (1)*	U.S.	1,436	1,360	-5%
17	20	Nvidia (2)	U.S.	1,118	1,285	15%
18	16	Sony	Japan	1,272	1,125	-12%
19	18	UMC (1)	Taiwan	1,140	1,034	-9%
20	21	AMD (2)	U.S.	1,030	832	-19%
—	—	Top 20 Total	—	66,745	62,440	-6%

(1) Pure-play foundry

(2) Fabless supplier

* Includes Intel/Altera, Avago/Broadcom, NXP/Freescale, and GlobalFoundries/IBM sales for 1Q15 and 1Q16.

**Custom processors for internal use made by TSMC and Samsung foundry services.

Source: Companies, IC Insights' Strategic Reviews Database

Tableau 1-2 : Classement des 20 premières sociétés IDM, Fonderies et Fabless pour 2016 [Source : IC Insight]

1.1.3 La loi de Moore et le « More Than Moore »

Le 19 Avril 1965, alors que les dernières avancées permettent d'intégrer jusqu'à 64 transistors sur une puce, Gordon Moore énonce la fameuse « Loi de Moore » dans un article désormais célèbre, *Cramming more components onto integrated circuits* [7], que je cite littéralement et en anglais :

« The complexity for minimum component costs has increased at a rate of roughly a factor of two per year. Certainly over the short term this rate can be expected to continue, if not

to increase. Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years.

G. Moore, 1965 »⁶

La conjecture de Moore, basée empiriquement sur une interpolation du nombre de transistors par puces, et révisée par Gordon Moore lui-même en 1975 (il faudra 2 ans et non plus 18 mois pour doubler la densité) s'avèrera incroyablement exacte : entre 1971 et 2001, le cycle moyen sera de 1,96 ans ! Jusqu'en 2015, la loi de Moore s'avère juste mais on remarque depuis quelque temps que ce n'est plus exactement vrai. En effet les dimensions deviennent si petites (Intel et TSMC ont développés en 2014 leur technologie 16nm) que les effets quantiques prennent des proportions importantes dans les performances des circuits. La conséquence est que pour diminuer encore les dimensions, le circuit intégré va se complexifier au niveau de l'intégration sur silicium, du design et de l'architecture, ce qui provoque un allongement notable du temps de développement des technologies.

La miniaturisation est le premier challenge de la microélectronique. Cette effort de recherche a permis d'augmenter toujours un peu plus les performances des puces en gardant un rapport performance sur coût raisonnable (la miniaturisation a pour l'instant, au moins jusqu'aux technologies 28nm, permit de réduire le coût unitaire du transistor [8] [9]). A chaque augmentation de la densité de transistor sur une même surface, la fréquence de fonctionnement augmente (avec la taille des composants qui diminue), les possibilités de calculs complexes augmentent (avec le nombre de transistors), la consommation électrique par composant diminue, ... De même, chaque nouvelle innovation va nécessiter des avancées dans des domaines très variés. On trouve la chimie (avec les procédés de gravure humide ou de purification des matériaux), en sciences de matériaux (propriétés électriques, chimiques, mécaniques, couches minces), en techniques du vide, en plasmas (avec les gravures sèches et certains dépôts), en optique (laser et lentilles avec la lithographie optique), en informatique (logiciels de support au développement et programmation des puces), ... Chaque nœud technologique va nécessiter une dizaine d'années de développement en laboratoire et entre 2 et 3 ans de développement en industrie avant son industrialisation et son entrée sur le marché.

La dynamique de miniaturisation a aussi permis à l'industrie et à la recherche microélectronique de s'auto-entretenir et de croître en initiant un cercle vertueux dans lequel chaque avancée technologique va permettre de générer des profits permettant le développement des nœuds suivants mais aussi fournir la puissance de calcul nécessaire à la réalisation des technologies lui succédant (capacités de simulation

⁶ NdA : « La complexité des coûts minimum des composants a augmenté à un rythme d'environ un facteur de deux par an. Certes, sur le court terme ce taux devrait rester le même, voire augmenter. À long terme, le taux d'augmentation est un peu plus incertain, bien qu'il n'y ait aucune raison de croire qu'il ne restera presque constant pendant au moins 10 ans. »

et de modélisation de plus en plus pointues, CAO-DAO⁷, automatisation de la chaîne de fabrication, contrôle, suivi et correction des procédés en direct, ...).

Le principe du « More Moore » consiste à continuer de plus en plus dans cette voie et de poursuivre la loi de Moore. Cependant, fin 2015, l'ITRS a officiellement annoncé la fin de la loi de Moore [10]. Si les dimensions peuvent encore être diminuées jusqu'au nœud 3.25nm d'après Samsung, le rythme se ralentit fortement en raison de la complexification des intégrations que nécessite la continuation de la miniaturisation. La technologie CMOS, phare de la microélectronique, est remplacée par d'autres solutions. On parle désormais de « Beyond CMOS » (au-delà du CMOS) pour définir ces solutions technologiques. Le FD-SOI (Fully Depleted Silicon on Insulator, ou Sicium sur isolant totalement déplété) et le FinFET (Fin-shaped Field Effect Transistor, ou Transistor « aileron ») sont deux solutions actuellement en production dans le monde. Du côté des mémoires, ce sont les intégrations 3D des technologies NAND qui ont permis d'augmenter les densités de stockage des puces. Chez STMicroelectronics, la technologie FD-SOI a été choisie car elle permet aussi une intégration basse consommation ciblant les marchés de la voiture autonome et de l'internet des objets. Ce travail s'intègre à l'optimisation de la fabrication de cette technologie en 28nm et 14nm. Le nœud 28nm est particulièrement intéressant car c'est celui pour lequel le coût unitaire du transistor est le plus bas et donc le meilleur candidat pour le développement de technologies dérivées (cf. Figure 1-4).

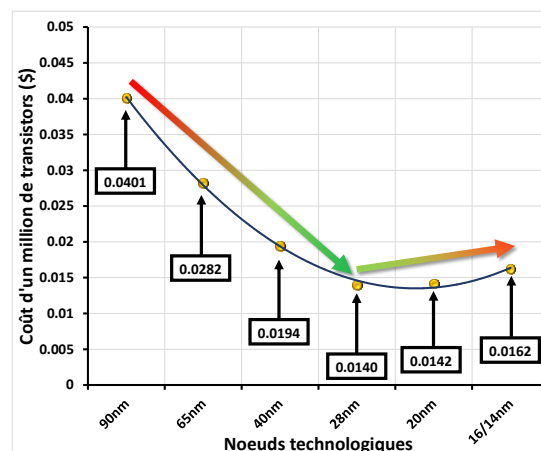


Figure 1-4 : Coût d'un millions de transistors par nœuds technologiques (source: Handel Jones, IBS)

Cette mutation de l'industrie microélectronique favorise le « More Than Moore » qui consiste en une diversification des produits pour des applications spécifiques. On trouve les technologies Radio Fréquence, les Imageurs, les MEMS⁸, les capteurs. Ces systèmes peuvent aussi être miniaturisés mais ils ne suivent pas la loi de Moore. Les technologies CMOS peuvent aussi être concernées par le « More

⁷ Conception Assistée par Ordinateur – Design Assisté par Ordinateur

⁸ Micro Electro-Mechanical Sytems, ou Systèmes électromécaniques micrométriques. On trouve par exemple les systèmes microfluidiques, les actuateurs ou encore divers capteurs (pression, mouvement, ...).

« More than Moore » dans le sens où il est possible de faire la co-intégration pour inclure ces capacités dans des puces logiques classiques.

En ce qui concerne les CPU⁹ et la logique, l'avènement des produits mobiles et des objets connectés conduit les développeurs à designer des produits « Low Power » (i.e. basse consommation) qui nécessitent de trouver le compromis optimum entre une grande puissance de calcul (par exemple pour permettre l'accès à des contenus en Haute Définition sur un smartphone) et une grande autonomie de fonctionnement. Pour cela, de nouvelles architectures et intégrations sont nécessaires. De ce point de vue le FD-SOI est très intéressant car il permet de diminuer fortement la consommation par transistor tout en gardant une grande puissance de calcul. En complément du cœur CMOS, la stratégie de ST (cf. Figure 1-5) s'inscrit comme une stratégie de diversification et de développement de dérivatifs, basés sur les technologies MOS qui les supportent.

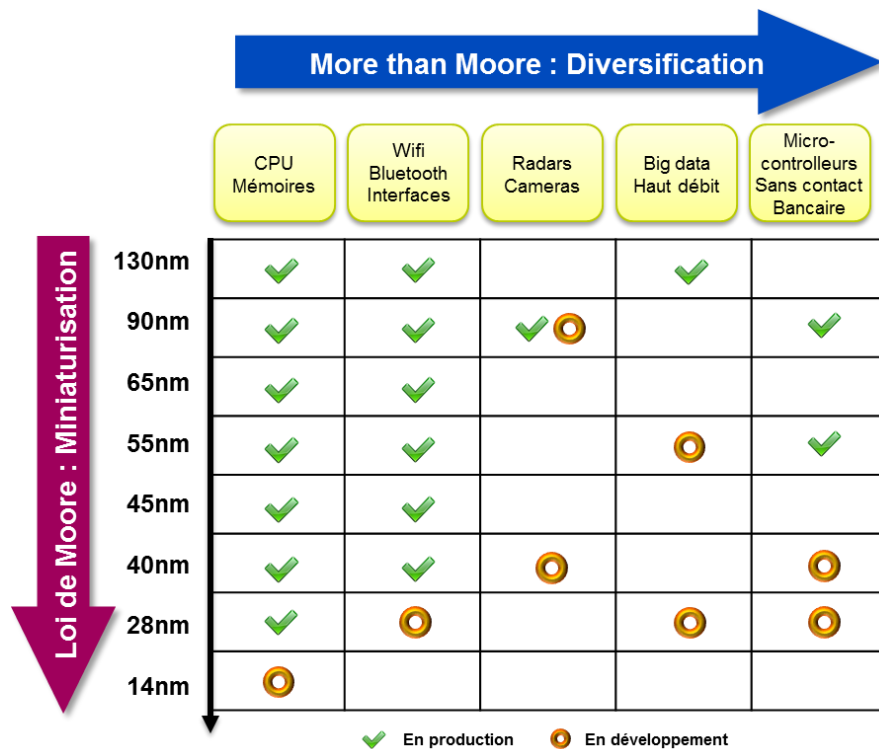


Figure 1-5 : Les deux chemins du « More Moore » (miniaturisation) et du « More Than Moore » (diversification) chez STMicroelectronics.

L'ajout de fonctions analogiques ou d'antennes Radio Fréquence permet le développement de puces autorisant une connexion sans fil (Wifi, Bluetooth) et de systèmes de traitement de données. Parmi celles-ci, on compte des intégrations complexes comme le BiCMOS (Bipolar CMOS) qui intègre sur la même puce des transistors MOS et bipolaires pour gérer des signaux analogiques. Ces solutions sont complémentaires à la photonique qui permet le transfert d'information entre une fibre optique et un circuit digital par exemple, la puce BiCMOS servant alors à la conversion du signal. Les applications

⁹ CPU ou Central Processing Unit désigne les microprocesseurs électroniques.

visées sont le traitement de données très haut débit (plus de 400GHz de fréquence d'horloge) dans les centres de stockage de données.

Une autre voie en croissance est le secteur de l' « embedded memory » (mémoire embarquée) qui intègre sur une même puce un circuit logique et de la mémoire flash programmable. Ces technologies permettent la fabrication des puces sécurisées pour les passeports, les cartes bancaires, les smartphones. Elles peuvent aussi intégrer des antennes RF pour permettre le paiement sans contact par exemple. Le travail qui suit pourra être une grande valeur ajoutée pour le développement de ces technologies en raison des spécificités de design que la technologie Flash nécessite.

Les imageurs sont un domaine en pleine croissance et sont de plus en plus présents dans les applications de la vie de tous les jours (caméras de smartphone, radars intégrés sur les véhicules ...).

Enfin, l'intégration 3D est aussi une voie en plein développement. Il est possible de coller et de superposer plusieurs puces de technologies différentes (mémoire, CPU, capteurs, ...) entre elles, d'intégrer plusieurs niveaux de transistors (comme c'est le cas pour les mémoires NAND 3D chez Samsung ou Toshiba).

1.2 LA FABRICATION DES CIRCUITS INTEGRES

1.2.1 Intégration

1.2.1.1 Les technologies CMOS

En microélectronique, l'intégration consiste à définir les étapes de fabrication, leur enchaînement, les matériaux, les masques, ... afin de passer d'un design électrique du circuit, dessiné d'un point de vue fonctionnel, à un produit manufacturable. Dans cette partie, nous verrons les principales étapes de fabrication d'une puce 28nm FD-SOI sur le wafer de silicium. Les étapes de découpage des puces, de connections externes et d'emballage puis de tests qui viennent après les étapes du Front-End Manufacturing (fabrication des composants) ne seront pas décrites ci-dessous.

Dans la fabrication des puces, on distingue plusieurs grands blocs : le FEOL (Front-End of Line ou Début de la ligne), le MEOL (Middle-End Of Line ou Milieu de la ligne), le BEOL (Back-End of Line ou Fin de la ligne) et le FarBEOL qui vient après le BEOL. Ces blocs sont définis par le morceau du circuit qui est fabriqué mais aussi par les matériaux utilisés.

- Le FEOL consiste en la fabrication du transistor en lui-même et les matériaux sont majoritairement des oxydes, nitrures et semi-conducteurs.

- Le MEOL consiste en la transition entre le FEOL et le BEOL. Il est constitué de niveaux de contacts permettant de connecter les différentes électrodes des transistors aux lignes d'interconnexions du BEOL.
- Le BEOL consiste en la fabrication des lignes d'interconnexion entre les transistors et correspond aussi à l'introduction de matériaux métalliques dans la puce.
- Le Far BEOL clôt la fabrication de la puce avec le dépôt d'une couche qui permet de protéger la puce et la fabrication des pads métalliques qui permettent de connecter la puce à son boîtier lors de l'emballage.

Chacun de ces blocs est composé de plusieurs briques correspondant à la fabrication de chacune des parties des composants et des interconnexions. Chaque brique est composée de plusieurs opérations elles-mêmes composées de plusieurs étapes de procédés (lithographies, gravures, implantations, polissages, dépôts, remplissages), de mesures et de nettoyages. Les deux figures 1-6 et 1-7 montrent une route complète et le détail de la fabrication d'un niveau de métal en 28nm FD-SOI.

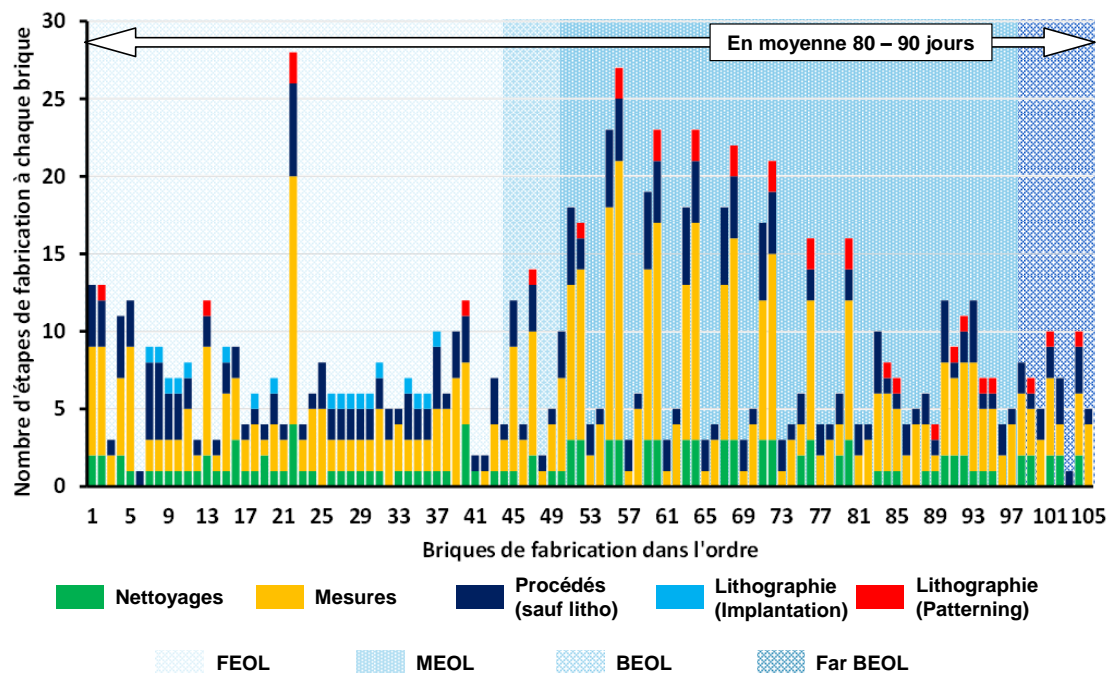


Figure 1-6 : Exemple de route d'un produit en technologie 28nm FD-SOI avec 49 masques et 9 niveaux d'interconnexions. Cette route compte 885 étapes de fabrication (dont 126 nettoyages, 473 étapes de métrologie et 237 procédés de fabrication) réparties en 262 opérations et 105 briques

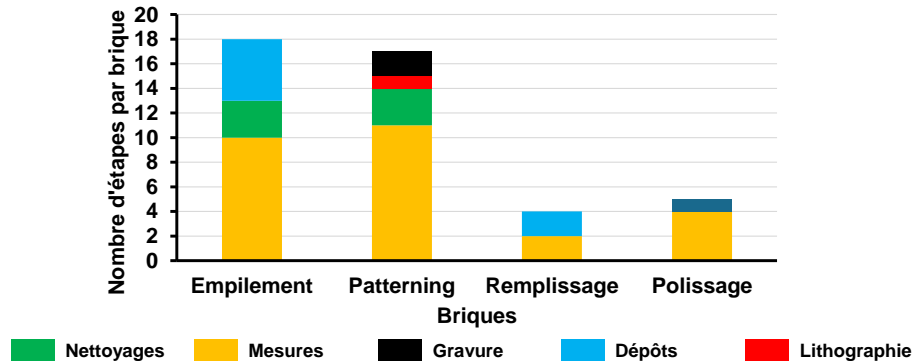


Figure 1-7 : Fabrication du premier niveau d'interconnexion en 28nm FD-SOI. Il faut 4 briques, 11 opérations et 46 étapes de fabrication (dont 27 de métrologie) pour ce faire.

Une très grande partie des étapes de fabrication est constituée de mesures et de contrôle en ligne. Dans le cas de l'exemple choisi, la métrologie représente plus de la moitié des étapes de fabrication.

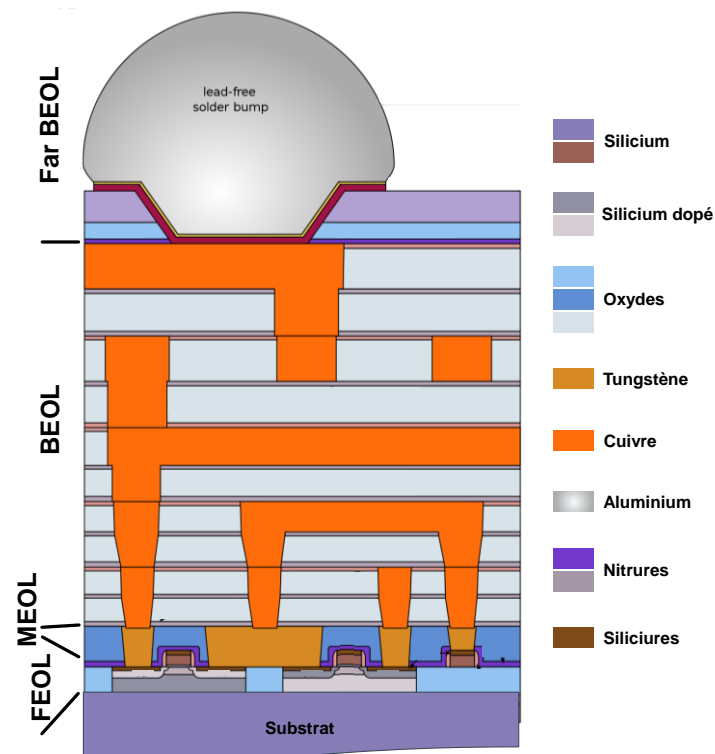


Figure 1-8 : Vue en coupe d'une puce électronique à 5 niveaux d'interconnexions

Sur la vue en coupe de la Figure 1-8, on peut reconnaître les transistors (FEOL), les contacts (MEOL), les interconnexions (BEOL) et la protection et le pad (Far BEOL). En revanche, de nombreux autres matériaux que ceux visibles sur le schéma sont utilisés pour la réalisation de la puce. C'est le cas par exemple des couches sacrificielles nécessaires à l'intégration mais qui n'apparaissent pas dans le design du composant, comme les résines en photolithographie ou les masques durs nécessaires à la gravure et la CMP.

1.2.1.2 Le contrôle de procédé

Lors de la création d'un procédé, son point de fonctionnement et la fenêtre de procédé sont déterminés. La fenêtre de procédé définit la zone de variation des principaux paramètres du procédé pour laquelle les performances ne sont pas détériorées. Elle est utilisée pour définir les spécifications de contrôle en ligne. Pour sécuriser le rendement et limiter le nombre de défauts, il est nécessaire de s'assurer que le procédé reste dans les spécifications et ne dérive pas. Pour cela, un plan de contrôle est mis en place. Il est basé sur des mesures du silicium en ligne au niveau des étapes critiques. Ces mesures sont utilisées comme indicateurs de fluctuations du procédé, préviennent des non-uniformités qui vont créer de la variabilité dans les puces produites et peuvent alimenter des boucles de régulation automatique pour corriger en direct les dérives.

En photolithographie, on trouve des mesures d'épaisseur, de dimension et de désalignement. Des non-uniformités d'épaisseur des matériaux de l'empilement de lithographie (BARC + Résine, pour plus de détails voir le Chap. 2) induisent des variabilités dans la qualité de l'impression de l'image dans la résine. Les dimensions des motifs imprimés sont liées aux performances électriques du composant comme par exemple la résistance d'une connexion. Le désalignement entre deux niveaux de masque est lui aussi très critique. Un désalignement peut causer un court-circuit ou empêcher le contact entre deux matériaux qui devraient être connectés.

Dans le cadre de ce travail, des mesures CDSEM ont été réalisées. Il s'agit de mesures dimensionnelles des motifs. Le CDSEM (Critical Dimension Scattering Electron Microscope, ou Microscope électronique à balayage de mesures de dimension critique) est un appareil de métrologie qui va réaliser une image MEB du motif et le mesurer automatiquement selon la recette de mesure choisie. Le CDSEM est utilisé en ligne pour contrôler tous les lots de production au cours de leur fabrication.

De même, les machines sont contrôlées et calibrées plus ou moins régulièrement selon la criticité du paramètre. Des routines et des tests normalisés permettent de contrôler les paramètres séparément. On trouve des tests de contrôle de l'alignement, du focus, de la qualité du faisceau laser, de la défektivité des masques, de la propreté des supports. Les tests permettant de contrôler le focus de la machine seront décrits plus en détail dans le Chap. 3.1.3 avec les méthodes de contrôle de la variabilité.

1.2.1.3 Technologies dérivées

Les deux technologies dérivées les plus proches du cœur CMOS sont les mémoires embarquées et le BiCMOS. Dans ces deux cas, des transistors non MOS sont intégrés sur le même substrat pour fonctionner avec le cœur CMOS de la puce (bipolaires pour le BiCMOS et Flash dans le cas des mémoires embarquées). Leurs architectures sont très différentes de la logique et il faut ajouter de nombreux niveaux de masque pour les réaliser ainsi que d'autres matériaux. De plus, la simple présence

de ces dispositifs spécifiques va impacter le design même de la puce et créer des sources de variabilité qui n'existaient pas sans elles ou alors dont l'impact était moindre.

Les imageurs, si leurs dimensions ne sont pas aussi petites que les précédents dérivatifs, nécessitent l'ajout de niveaux spécifiques mais aussi de résines particulières pour la formation de lentilles sur les pixels.

Il existe enfin aussi d'autres technologies comme les MEMS ou le discret dont les challenges de miniaturisation sont aussi très spécifiques mais dont il ne sera pas question dans ce manuscrit.

1.2.2 Quelques indicateurs clés de la fabrication

Dans un environnement industriel comme celui de STMicroelectronics, il est nécessaire de définir des indicateurs de qualité qui permettent de suivre la production. Parmi ceux-là, on trouve le temps de cycle et le rendement.

Le « turn rate » ou temps de cycle peut se définir comme le temps nécessaire pour réaliser un certain nombre d'étapes de fabrication. Il s'agit d'un indicateur de la vitesse de production, qui s'exprime en nombre de « moves » (ou déplacement, manœuvre) par jour, c'est-à-dire en nombre d'opérations par jour. Le but étant évidemment de l'augmenter pour accélérer le rythme de production et produire plus.

Le rendement est, comme dans toutes les industries, un indicateur très important de la qualité de la production. Il est défini comme le ratio des produits propres à la vente sur l'ensemble des produits fabriqués.

La sécurisation du rendement passe par un suivi et un contrôle des facteurs de risque et des procédés. Cela permet un suivi des variabilités et des défauts susceptibles de moduler le fonctionnement de la puce électronique.

A mesure de la diminution des dimensions et la complexification de l'intégration, la sécurisation du rendement nécessite de plus en plus de contrôle. La tendance est à l'augmentation du nombre d'étapes de métrologie et de contrôles en ligne des produits pour sécuriser la production. En raison de la vitesse de production et de l'intégration des composants, un défaut n'est pas forcément détectable tout de suite, mettant alors une grande partie de la production en risque car le temps de détecter le problème, d'autres lots auront passé l'étape de fabrication défectueuse. L'intégration empêche de plus de pouvoir corriger ces erreurs dans un grand nombre de cas [11]. La détection d'un défaut doit se faire le plus tôt possible pour éviter de mettre trop de lots en risque, ce qui impose plus de contrôle et de mesures en ligne. Cela va à l'encontre de la volonté de diminuer le temps de cycle. L'ingénieur va chercher à trouver des moyens pour éviter des mesures trop prenantes en temps. La simulation, les modèles et les mesures croisées sont les solutions usuelles. Une vision globale voire holistique [12] peut ici s'avérer très intéressante dans le sens où la mesure d'un seul paramètre permettra de remonter à plusieurs autres par

corrélations croisées par exemple. Ce travail de thèse proposera en fin de manuscrit des solutions innovantes permettant de caractériser le procédé et de le suivre sans avoir à mesurer directement les wafers.

1.3 CONCLUSION

Le premier chapitre était dédié à une analyse de la situation actuelle de l'industrie microélectronique en termes de mutations et de challenges de fabrication. Avec plusieurs trillions de dollars générés par an, plusieurs millions d'emplois directs et indirects et une quasi-omniprésence dans notre vie de tous les jours, la microélectronique constitue le cœur de la Troisième révolution industrielle et permet le développement de la Quatrième qui commence. Il n'est pas exagéré de dire que le transistor est sans doute la plus « grande » invention du XX^{ème} siècle.

Son importance industrielle, scientifique et socio-économique a tiré l'innovation pendant plus de 50 ans, résultant en un suivi strict de la loi de Moore, en de nombreuses découvertes dans des domaines variés de la physique et de la chimie et en le développement de l'informatique. A chaque réduction des dimensions du transistor, les challenges proposés sont de plus en plus difficiles et intéressants car ils poussent les technologies, les machines, les matériaux dans leurs retranchements.

Les derniers nœuds technologiques voient l'émergence de relations croisées de plus en plus étroites entre les différentes parties du design, les procédés, les matériaux, les machines qui sont autant de paramètres qui peuvent devenir des détracteurs du rendement. Le contrôle des indicateurs de performances (métrologie, défektivité, dérives des procédés,...) doit alors être plus strict pour assurer un rendement correct. Il devient alors absolument nécessaire d'avoir une vision globale de l'intégration pour pouvoir appliquer un contrôle adapté des procédés : une approche dite holistique.

CHAPITRE 2

2 LA PHOTOLITHOGRAPHIE

La photolithographie est « l'ensemble des opérations permettant de transférer une image vers un substrat. » (Wikipédia). Elle est utilisée en microélectronique en raison de sa relative simplicité de mise en œuvre et de sa reproductibilité.

Ce procédé repose sur l'exposition à la lumière d'une résine photosensible au travers d'un masque optique afin de transférer le design de la puce électronique sur le substrat de silicium. Le développement de la résine avec une chimie adaptée permet alors de découvrir certaines zones préférentiellement à d'autres [13]. L'étape de lithographie est suivie d'une gravure (permettant de retirer de la matière au niveau des zones découvertes), d'une implantation ionique (dopage des transistors) ou de la croissance d'un matériau dans l'espace libéré. Dans le cadre de cette thèse, le niveau contact et le premier niveau d'interconnexion ont été au centre de l'étude. Dans les deux cas, la lithographie a été suivie d'une étape de gravure.

Dans ce chapitre, je présenterai tout d'abord le procédé et les machines utilisées pour la réalisation de l'étape de lithographie. Dans un deuxième temps, au travers d'une explication de la formation de l'image pendant le procédé, les principaux challenges de la lithographie et le passage de la photolithographie conventionnelle à la « computational lithography » puis à la lithographie holistique seront expliqués. Nous terminerons en exposant les challenges propres à ce sujet de thèse.

2.1 LE PROCEDE PHOTOLITHOGRAPHIQUE

Le procédé de photolithographie en microélectronique englobe toutes les procédés entre le dépôt de la résine, et de l'anti-réfléctif et/ou du planarisant si besoin, et le développement de la résine après son exposition à un laser UV. La photolithographie permet une structuration spatiale de la puce en définissant des zones qui doivent recevoir un traitement particulier (gravure, implantation, croissance épitaxiale) et ainsi définir les différents éléments du circuit (isolations, grille des transistors, contacts, interconnexions métalliques). Généralement, l'étape de photolithographie suit le dépôt de certains matériaux (oxyde, métal ou semi-conducteur).

La photolithographie repose sur la propriété photochimique de certaines résines, lesquelles possèdent un seuil de sensibilité à l'intensité lumineuse au-delà duquel leur solubilité est modifiée. La résine, si elle est exposée, devient plus soluble dans le développeur et non soluble si elle n'est pas exposée suite

au recuit qui suit l'exposition dans le cas d'une résine positive. L'inverse se produit dans le cas d'une résine négative. La projection de l'image d'un masque optique sur ces résines permet de la structurer de telle manière que certaines zones soient exposées suffisamment pour déclencher la réaction photochimique alors que d'autres n'auront pas été modifiées photochimiquement. L'utilisation du développeur permet de découvrir certaines zones, créant des motifs dans la résine. Dans le cas de l'industrie microélectronique, le masque contient le plan de la puce électronique. Ce procédé permet ainsi de transférer de manière répétable le design de la puce sur le wafer en utilisant de nombreuses fois le même masque. On parle de photorépétition.

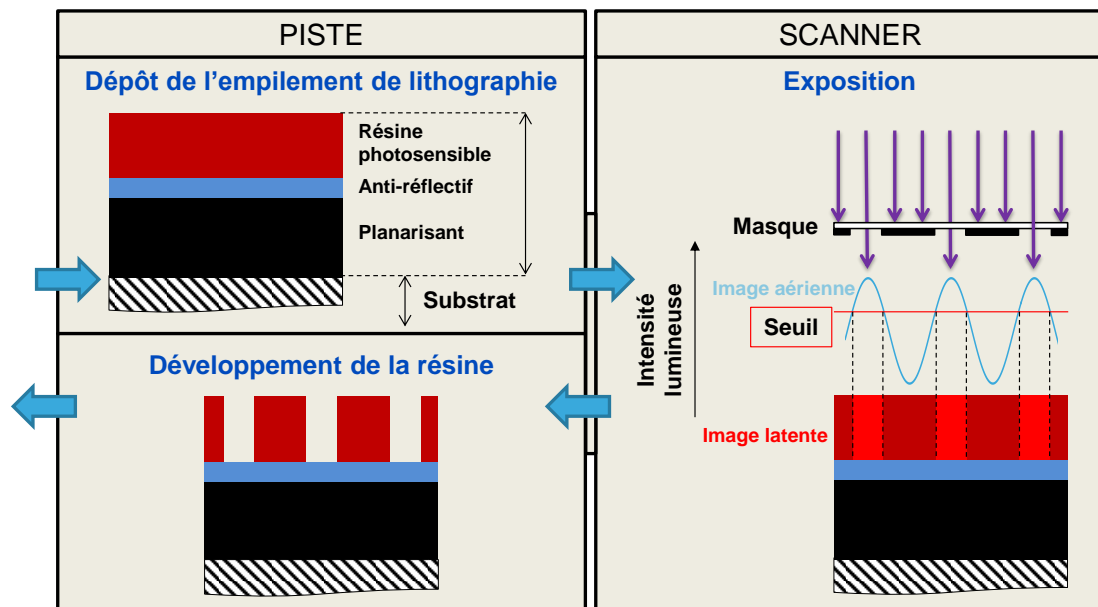


Figure 2-1 : Principe de base de la lithographie. Le wafer (substrat) est recouvert avec l'empilement de lithographie, est exposé puis développé. Suite au procédé, le design du masque aura été transféré comme image développée sur le substrat.

Lors du procédé de lithographie (schématisé en Figure 2-1), on distingue trois types d'images qui sont formées successivement sur le substrat. La projection du masque forme l'image aérienne au niveau du plan image du système optique. En positionnant le substrat recouvert de la résine photosensible au niveau du plan image, celle-ci va être exposée avec le maximum de contraste. Une image latente est alors formée dans la résine photosensible. Elle correspond à la différenciation chimique entre les zones exposées et celles qui n'ont pas reçu de lumière. Les zones exposées et non exposées sont définies par le masque. A ce stade, l'image aérienne du masque de lithographie est transférée dans la résine. La résine doit ensuite être développée. On obtient alors l'image développée. Elle est constituée de parties du substrat qui sont encore recouvertes de résine et d'autres qui sont découvertes avec le substrat à nu.

2.2 LES MACHINES

Le cluster est un ensemble de deux machines reliées entre elles permettant de réaliser toutes les étapes de dépôts de couches organiques, de recuits, d'exposition et de développement. Il est constitué de

scanner qui permet l'exposition et de la piste dans laquelle sont réalisées toutes les autres étapes, c'est-à-dire les dépôts par centrifugation, les recuits et le développement. Dans la suite de cette partie, le scanner est présenté plus en détail.

La Figure 2-2 donne un bref aperçu de l'évolution des machines de lithographie au fil des années.

- Au début de la microélectronique, on utilisait une lithographie par contact, avec le masque posé directement sur le wafer puis une lithographie par proximité où il y a un espace contrôlé entre le masque et le wafer.
- Est arrivée ensuite la lithographie par projection où le masque est séparé du wafer par un système optique. Dans tous ces cas, le masque avait la même taille que le wafer et l'ensemble de toutes les puces de la plaque étaient exposés en même temps.
- Au milieu des années 80, sont arrivés les steppers ou Step-and-Repeat. Le masque ne contient alors qu'un nombre limité de puces et il est exposé de nombreuses fois sur le wafer, champ par champ. Le masque est exposé en une fois pour chaque champ et reste fixe.
- Enfin, les scanners ou Step-and-Scan, introduits dans la deuxième moitié des années 90, utilisent le même type de masque en 4X (c'est-à-dire 4 fois plus grand que le design sur le wafer) que les steppers les plus avancés mais celui-ci est en mouvement ainsi que le wafer mais dans le sens inverse de ce dernier et 4 fois plus vite. Chaque champ est exposé séparément.

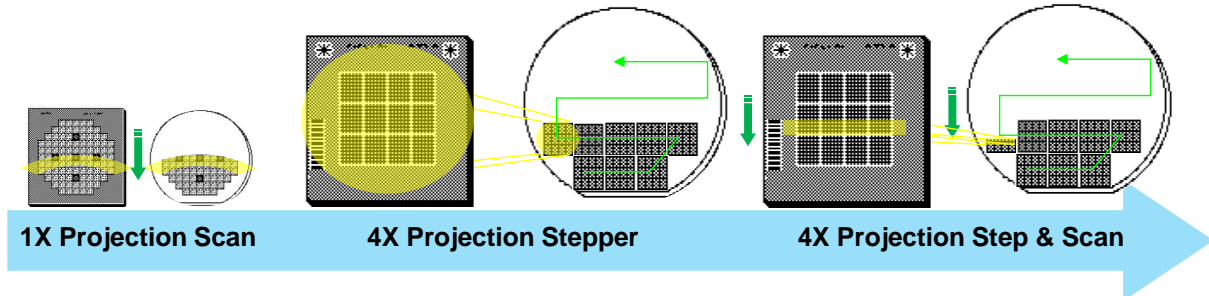


Figure 2-2 : Comparaison de plusieurs technologies de photolithographie. Le masque et le wafer sont représentés à l'échelle.

Pour les derniers nœuds technologiques (90nm et en dessous), la machine d'exposition est un scanner acceptant des wafers de 300mm de diamètre. Dans le cadre de cette étude, un TWINSCAN NXT:1950i d'ASML a été utilisé. Il s'agit d'un scanner ArFi, i.e. 193nm à immersion. Cette machine utilise un laser Excimer au Fluorure d'argon. Le laser est ensuite guidé jusqu'au masque puis du masque vers la plaquette de silicium. Le 1950i est un scanner à immersion, c'est-à-dire qu'un ménisque d'eau est généré entre la colonne optique et le wafer. Ce système permet d'atteindre une ouverture numérique de 1.35. Enfin, les scanners « Twinscan » ont la particularité d'avoir deux chucks permettant de mesurer un wafer pendant l'exposition d'un autre afin de garder un débit de plaquette à l'heure compatible avec une unité de production. La formation de l'image et les solutions techniques qui ont été développées pour pallier aux problèmes de résolution induits par la miniaturisation seront discutés dans le Chap 2.4. Le scanner utilise une projection 4X.

La Figure 2-3 montre le scanner utilisé pendant cette thèse et son système optique pendant l'exposition d'un wafer.

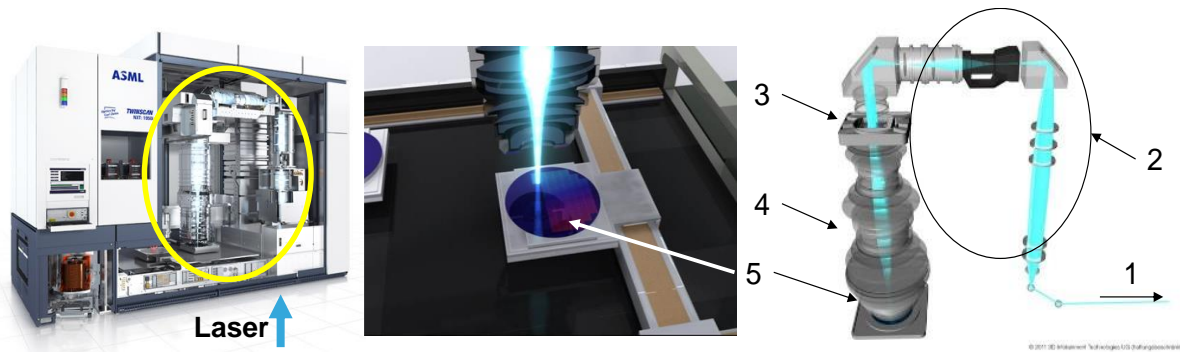


Figure 2-3 : A gauche, scanner de lithographie ASML TWINSCAN NXT:1950i utilisé pour le travail de thèse (Source : ASML) A droite, schéma optique d'un scanner de lithographie à immersion. (1) Source Laser (non visible sur le dessin), (2) Illuminateur, (3) Masque, (4) Lentilles d'exposition en projection 4X, (5) Wafer en court d'exposition. Le ménisque d'eau entre la colonne optique et le wafer n'est pas représenté ici. (Source : 3dit.de)
Au milieu, zoom autour du wafer, le système d'immersion n'est pas visible ici. (Source : 3dit.de)

2.3 LA DIFFRACTION ET L'IMAGE AERIENNE

2.3.1 La formation de l'image aérienne

L'exposition de la résine repose sur la formation d'une image aérienne nette et non déformée du masque dans la résine. Il s'agit du profil d'intensité au niveau du plan image résultant de la diffraction de la lumière source par le masque. Au passage au travers d'une fente ou d'un réseau et si la largeur de cette fente ou le pas de ce réseau avoisine la longueur d'onde du photon incident, la lumière va diffracter. On observe au niveau du plan de Fourier la formation d'une figure de diffraction. Chaque ordre de diffraction va se propager dans l'espace avec un angle θ suivant la loi de Bragg :

$$\sin \theta = \frac{n\lambda}{p} \quad \text{avec} \quad \begin{cases} n \text{ l'ordre de diffraction} \\ \lambda \text{ la longueur d'onde} \\ \theta \text{ l'angle de diffraction} \\ p \text{ le pas du réseau} \end{cases} \quad (1)$$

C'est l'interférence positive de ces différents ordres de diffraction au niveau du plan focal qui va permettre la formation de l'image. L'ordre 0 apporte uniquement de l'intensité lumineuse alors que les autres ordres apportent du contraste dans l'image. Il est donc nécessaire de capturer au moins un ordre de diffraction d'ordre supérieur à l'ordre 0 pour former une image comme le montre la Figure 2-4. Plus le nombre d'ordres de diffraction utilisé pour la reconstitution de l'image est grand et plus cette image sera nette. Dans le cadre des derniers nœuds technologiques en microélectronique, la diffraction est très grande et seul le premier ordre, voire une partie seulement, peut être capturé par le système optique. Le contraste ne sera donc pas très grand et il devient alors nécessaire de corriger cette diffraction pour avoir une image correcte sur le wafer.

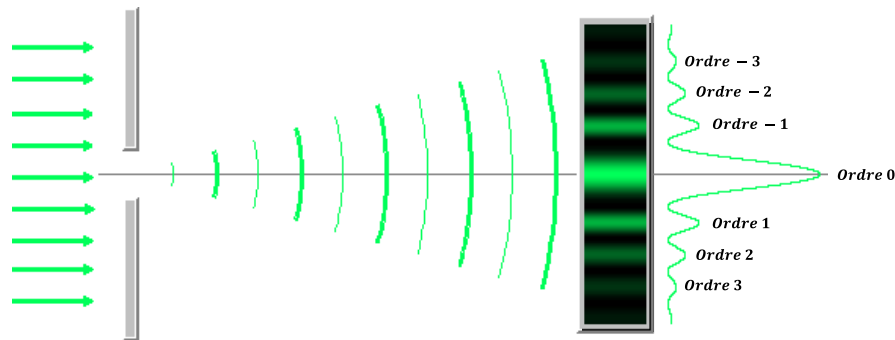


Figure 2-4 : Représentation de la diffraction d'un faisceau lumineux au travers d'une fente et de sa figure de diffraction

Le challenge principal de la lithographie est de supporter la loi de Moore en imprimant des motifs de plus en plus petits sur le wafer de silicium. La résolution, c'est-à-dire la capacité à définir des motifs de petite taille suffisamment rapprochés les uns des autres, est le paramètre qui doit être optimisée pour permettre de telles performances. Rayleigh a montré que la résolution d'un système optique s'exprimait comme suit :

$$CD = 0.61 * \frac{\lambda}{NA} \quad (2)$$

où $NA = n \cdot \sin \alpha$ est l'ouverture numérique du système permettant de capturer les ordres de diffraction jusqu'à un angle de diffraction α et λ la longueur d'onde du faisceau lumineux. Le CD ou Critical Dimension (dimension critique) est la taille du motif que l'on veut imprimer sur le silicium.

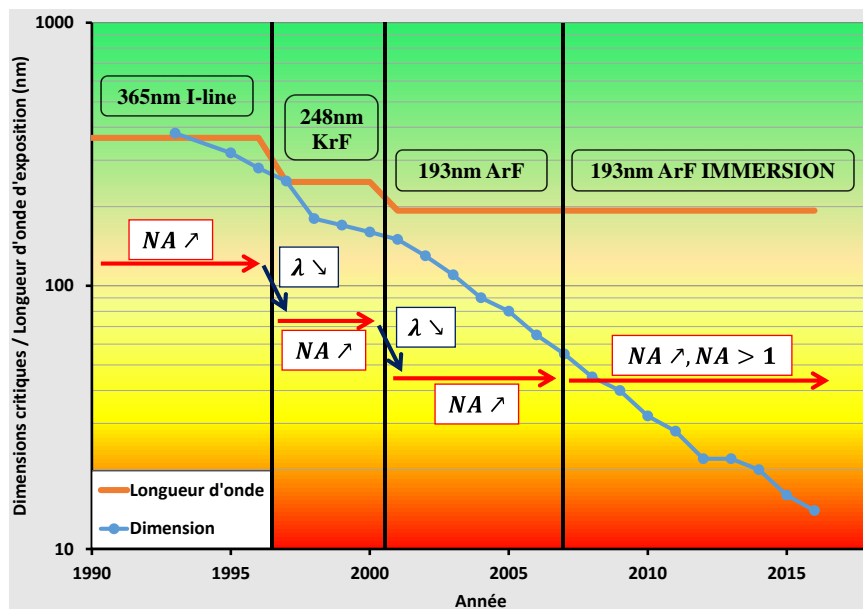


Figure 2-5 : Evolution de la longueur d'onde et de l'ouverture numérique avec la réduction des dimensions des composants

L'équation (2) donne deux solutions pour imprimer des motifs plus petits. Soit on diminue λ soit on augmente NA . La figure 2-5 montre l'évolution des dimensions des composants en parallèle de l'évolution de certains paramètres clés de la lithographie : la longueur d'onde d'exposition et l'ouverture numérique. Concernant la longueur d'onde, il s'agit de modifier la lampe ou le laser utilisé pour l'exposition. Sur le graphique, on passe des lampes au mercure pour la lithographie i-line (365nm) à

l'utilisation de laser Excimer pour les lithographies KrF (Fluorure de krypton, 248nm) et ArF (Fluorure d'argon, 193nm). Entre ces avancées, c'est l'ouverture numérique du système qui a été optimisée pour compenser l'augmentation de l'angle de diffraction avec la réduction des dimensions du masque. Les conséquences de la réduction de la longueur d'onde et de l'augmentation de l'ouverture numérique sont exposées en Figures 2-6 et 2-7.

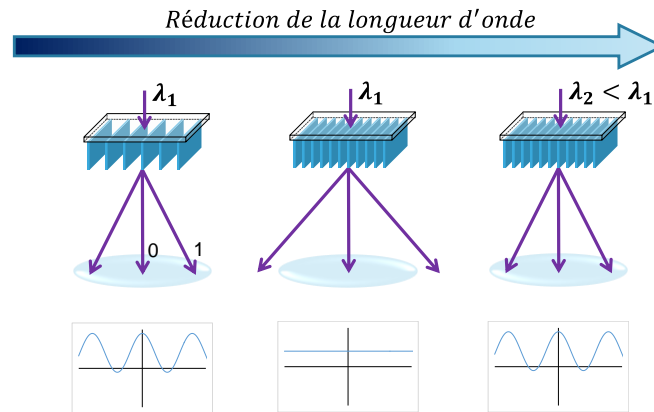


Figure 2-6 : Réduire de la longueur d'exposition permet de s'affranchir d'une diffraction trop importante sans augmenter plus la taille des lentilles.

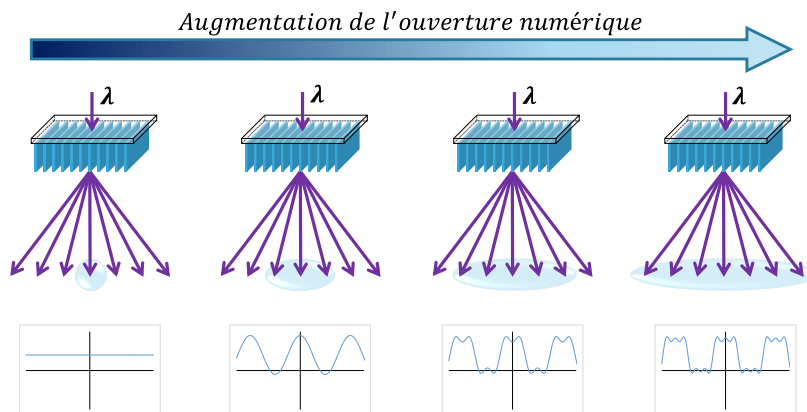


Figure 2-7 : Augmentation de l'ouverture numérique pour une même longueur d'onde. En augmentant la taille des lentilles de projection, plus d'ordres de diffraction sont capturés et on gagne en contraste sur le wafer.

Le passage à la lithographie à immersion permet de continuer d'augmenter l'ouverture numérique au-dessus de sa limite de 1 en changeant le milieu pour avoir un indice de réfraction plus élevé. Il est en effet possible de réaliser des lentilles dont l'ouverture numérique est supérieure ou égale à 1 mais leur utilisation conjointe avec la projection 4X cause une réflexion totale des faisceaux diffractés au niveau de la dernière lentille avant le wafer et aucune information n'est transmise à la résine sur le wafer.

Dans le cas de la lithographie à immersion, c'est l'eau qui a été choisie comme milieu alternatif à l'air. Un ménisque d'eau est généré entre le wafer et la colonne optique pendant l'exposition. L'eau a un indice de réfraction plus élevé que celui de l'air pour les UV, passant de 1 dans l'air à 1,44 dans l'eau. Cela permet d'atteindre un NA de 1,35 alors qu'il était limité à 0,95 pour les systèmes sans immersion. Ce phénomène est illustré par la Figure 2-8.

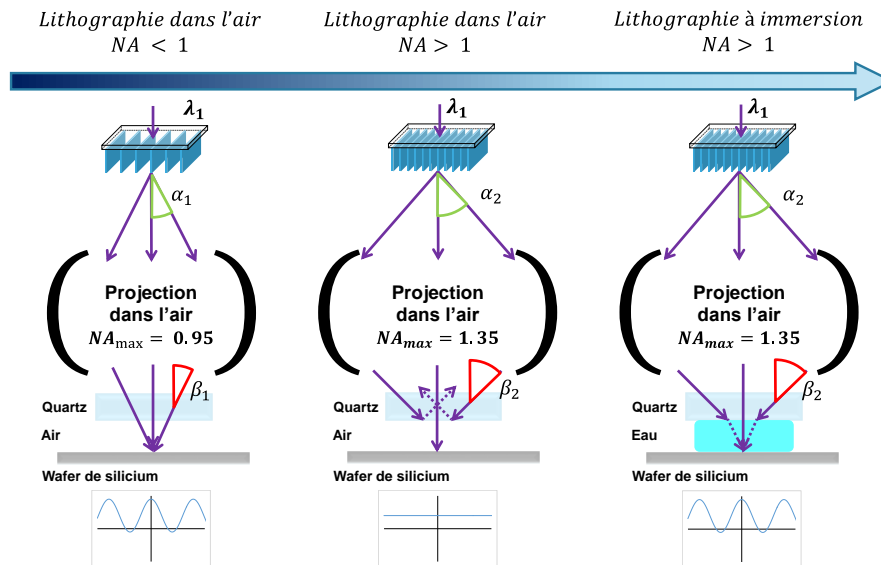


Figure 2-8 : Explication du passage à la lithographie à immersion. Les angles β sont tous 4 fois plus grands que les angles α correspondants en raison de la projection 4X.

Si on reprend l'équation (2), on remarquera que la constante 0.61 ne permet pas de décrire réellement les performances du procédé, particulièrement pour les dernières technologies. L'équation devient alors :

$$CD = k_1 * \frac{\lambda}{NA} \quad (3)$$

Où on introduit k_1 ($0 < k_1 < 1$) qui est une constante dépendant des conditions de procédé, du type de masque utilisé, de la machine, de l'illumination. Cette constante permet de quantifier la complexité de mise en œuvre du procédé lithographie. Plus k_1 est petit et plus il est difficile et complexe d'imprimer le CD résultant. La limite basse de k_1 par un réseau de lignes/espaces est d'environ 0.25 soit un pas de 78nm avec $NA = 1.35$ et $\lambda = 193nm$. Dans la pratique, en deçà de $k_1 = 0.3$, la double exposition devient une option quasi obligatoire. Pour un $k_1 \leq 0.61$, on se trouve dans le régime du « two-beam imaging » ou lithographie à deux faisceaux dans lequel seuls les ordres 0 et 1 de diffraction sont capturés et utilisés pour la reconstitution de l'image aérienne, voire dans le cas du « partial beam imaging » ou lithographie en faisceaux partiels où seul l'ordre 0 et une partie de l'ordre 1 servent à la création de l'image aérienne.

La réduction du k_1 implique un contrôle plus difficile, non seulement des dimensions mais aussi des conditions optimales de procédé comme la dose ou le focus. La dose correspond à l'intensité lumineuse que le wafer reçoit par unité de surface, exprimée en $mJ.cm^{-2}$, et le focus à la position verticale du wafer par rapport à l'image aérienne du masque dans le plan focal.

Concernant le focus, un paramètre important à connaître est la profondeur de champ, c'est la dire la distance autour du plan focal dans laquelle l'image est nette. Plus on augmente l'ouverture numérique

d'un système, plus la résolution est grande mais en contrepartie, la profondeur de champ est réduite. Cette profondeur de champ est la distance maximum entre deux rayons d'ordre différents à laquelle ils peuvent interférer positivement pour former une image autour du point focal. La profondeur de champ se déduit assez facilement du critère de Rayleigh.

La différence de phase $\Delta\phi$ entre les ordres de diffraction au niveau du plan image est induite par la différence de chemin optique parcouru et peut s'exprimer de la manière suivante :

$$\Delta\phi = \delta(1 - \cos \alpha) \cong \frac{1}{2} \delta \cdot \sin^2 \alpha \quad (4)$$

avec $\Delta\phi$ étant le décalage de phase, δ la valeur du défocus, α l'angle de diffraction et en appliquant le cas des petits angles, i.e. $1 - \cos \alpha \cong \sin^2 \alpha$.

En appliquant le critère de Rayleigh qui impose un angle $\alpha_{\max} = \frac{\pi}{2}$, angle à partir duquel les ordres 0 et 1 de diffraction n'interfèrent plus, empêchant toute formation d'image, le décalage de phase maximum autorisé est :

$$\Delta\phi_{\max} < \frac{\lambda}{4} \quad (5)$$

La profondeur de champ est égale à 2 fois la distance δ_{\max} qui est le défocus maximal autorisé pour avoir une image aérienne correctement formée au niveau du wafer.

$$DOF = 2\delta_{\max} < \frac{1}{2} \frac{\lambda}{(1 - \cos \alpha)} = \frac{1}{2} \frac{\lambda}{\sin^2 \alpha} \quad (6)$$

$$\Leftrightarrow DOF < \frac{n^2}{2} * \frac{\lambda}{NA^2} \quad (7)$$

$$\Leftrightarrow DOF = k_2 \frac{\lambda}{NA^2} \text{ avec } k_2 < 1 \quad (8)$$

La réduction des dimensions imposant une diminution de la longueur d'onde et une augmentation de l'ouverture numérique, la profondeur de champ diminue avec le CD. Le positionnement du wafer dans le plan focal devient alors de plus en plus compliqué car la marge de positionnement est très réduite. La constante k_2 , comme k_1 pour la résolution, traduit la complexité du procédé et diminue avec k_1 .

Une exposition hors focus crée une image latente floue sur le wafer en raison d'une perte de contraste dans l'image aérienne. La perte de résolution qui en résulte est due aux interférences constructives entre les différents ordres de diffraction qui diminuent en intensité avec le défocus. Les effets du focus sur l'image sont traités avec ceux de la dose dans la partie sur la fenêtre de procédé. Lorsque le wafer est exposé au focus optimal, le NILS (Near Image Log Slope ou Pente logarithmique de l'image proche)

est maximal et il diminue avec le défocus. Il s'agit d'une mesure du contraste. Le NILS est la pente de l'image aérienne au niveau des bords du motif. La Figure 2-9 compare les images aériennes et les NILS obtenus pour deux expositions au focus optimum et avec un défocus non nul.

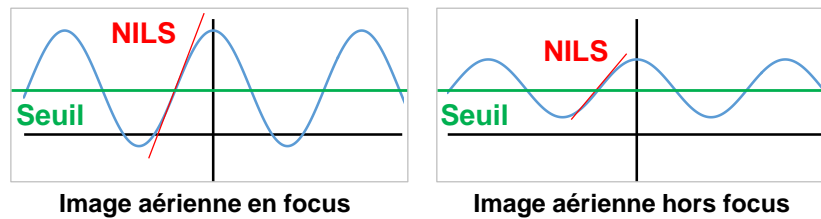


Figure 2-9 : Comparaison des images aériennes et du NILS lorsque l'image est construite en focus ou avec un défocus.

2.3.2 La « computational lithography »

Avec la complexification de la formation de l'image aérienne et la réduction de la constante k_1 , il est devenu nécessaire d'implémenter d'autres solutions afin de limiter la déformation des motifs dans la résine. La « computational lithography » ou lithographie assistée par ordinateur offre la possibilité d'optimiser l'illumination, le masque et le design via l'utilisation de modèles optiques du scanner et de la résine. Les illuminations sont optimisées en fonction du design ainsi que le masque qui sera modifié pour permettre l'impression du design de la puce sur le wafer. Les techniques d'amélioration de la résolution (ou RET pour Resolution Enhancement Techniques) sont décrites dans la partie suivante et consiste à prévoir les déformations de l'image aérienne et de l'image latente afin d'optimiser le masque et l'illumination.

2.3.2.1 Les illuminations

Une première solution est de modifier le mode d'illumination. Le principe est fondé sur l'illumination hors axe. Au lieu d'illuminer le wafer avec un rayon lumineux normal au plan du masque, il est possible de capturer l'un des deux faisceaux diffractés au premier ordre en changeant l'angle d'incidence du laser sur le masque. Il y aura alors interférence constructive dans le plan image et formation de l'image aérienne. Il est en réalité impossible d'illuminer réellement hors axe dans un scanner de lithographie. Le procédé consiste à modifier la forme du faisceau laser en passant d'une illumination conventionnelle ponctuelle à des formes plus complexes (annulaire, dipôle, multipôle, quad, freeform). L'illumination hors axe se traduit ici par un cône de lumière incidente. Ces illuminations sont utilisées dans le cas de la « two-beam lithography » (ou lithographie à deux faisceaux) qui apparaît avec les dernières technologies. Les motifs du masque sont alors du même ordre de grandeur que la longueur d'onde du laser et la diffraction est très forte au point que le système optique ne capture plus que l'ordre 0 de diffraction. La formation de l'image aérienne est alors impossible. L'illumination complexe a pour but de capturer au moins partiellement un ordre élevé de diffraction ($n \geq 1$), généralement l'ordre 1. La Figure 2-10 donne l'exemple de l'illumination annulaire.

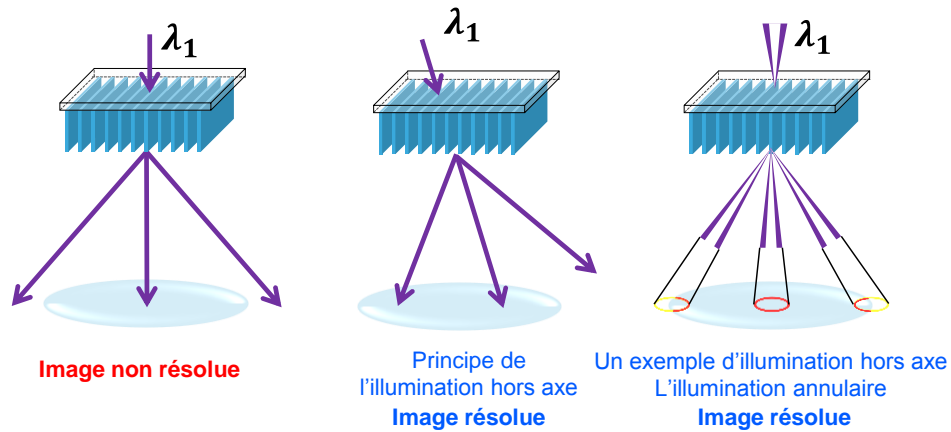


Figure 2-10 : Principe de l'illumination annulaire. Une illumination de type annulaire permet de capturer partiellement (en rouge) le premier ordre de diffraction sans modifier le NA au prix d'une perte de contraste.

Cette méthode, si elle permet de pousser les performances de la machine plus en avant et d'imprimer des motifs de plus en plus petits, a un désavantage majeur : la quantité d'information reçue pour la formation de l'image aérienne est limitée si on compare à une illumination conventionnelle ponctuelle. Plus la forme est complexe et plus la partie capturée de l'ordre 1 de diffraction est réduite. La conséquence est une perte de contraste dans l'image aérienne (cf. Figure 2-11). En reprenant le fonctionnement des résines photosensibles, on conviendra que la formation de l'image latente est compromise si le contraste n'est pas suffisant.

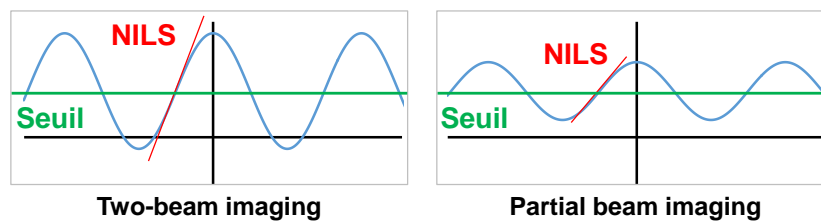


Figure 2-11 : Comparaison des images aérienne et du NILS lorsque l'image est construite avec l'ordre 1 complètement ou partiellement capturé par les lentilles de projection.

2.3.2.2 Les OPC

Avec la réduction des dimensions, apparaissent des déformations optiques de l'image aérienne provoquées par des effets de proximité. L'image du masque transférée dans la résine apparaît avec des coins arrondis, des motifs de dimensions trop petites ou encore des pincements ou des ponts de résine indésirés.

Les OPC (Optical Proximity Corrections ou corrections des effets de proximité) permettent d'assurer la fidélité de forme et de dimension entre les motifs du design original et leur impression dans la résine en anticipant les distorsions de l'image aérienne. Les modèles OPC calculent alors le design des motifs sur le masque permettant l'impression dans la résine de motifs approchant le plus possible le motif souhaité par le designer. Cela se fait par élargissements locaux du design, insertion de sérifs ou de motifs non-résolus d'assistance à la formation de l'image, voire par calcul d'un design complètement différent mais qui rendra le dessin voulu sur le wafer.

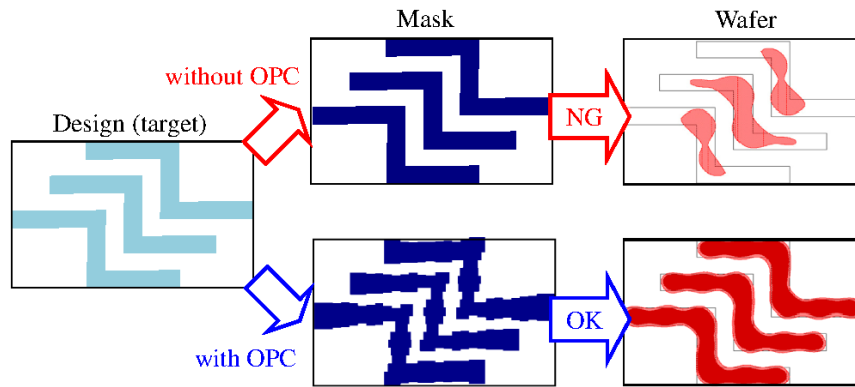


Figure 2-12: Design, masque et image dans la résine avec et sans OPC. (Source: [14])

2.4 LA LITHOGRAPHIE DANS LA LOI DE MOORE

La réduction des dimensions a levé des challenges très importants en termes de résolution et de qualité des images aériennes et latentes. Cette partie a pour but de revenir sur l'évolution de la lithographie au fur et à mesure des avancées technologiques en microélectronique et de présenter les évolutions qui ont conduit la lithographie des procédés conventionnels à la lithographie assistée par ordinateur et enfin à la lithographie holistique.

La Figure 2-13 donne un aperçu de l'évolution de la constante k_1 , dont la valeur caractérise la complexité du procédé au cours du temps et en parallèle des technologies de la microélectronique. Les améliorations nécessaires à la réalisation de chacune de ces technologies sont répertoriées au niveau du k_1 pour chaque avancée.

Dans la Figure 2-13, DP/TP/QP signifient respectivement Double, Triple et Quadruple Patterning, c'est-à-dire que des expositions multiples sont nécessaires pour pallier à l'impossibilité de résoudre les motifs. RET signifie Resolution Enhancement Techniques ou techniques d'amélioration de la résolution. OPC (Optical proximity correction) définit les méthodes de correction des effets de proximité, il s'agit de l'une de techniques de RET. DTCO (Design-Technology Co-Optimization) désigne les méthodes d'optimisation du design du circuit pour permettre aux procédés de fabrication de réaliser les circuits. La lithographie EUV pour Extrême UV utilise un laser avec une longueur d'onde de 13.5nm^{10} .

Ce travail de thèse a été réalisé sur le Back-End du 28nm FD-SOI et sur le Contact du 14FD-SOI. Dans le premier cas, le k_1 est de 0.31. Dans le second, on descend à une valeur de 0.23 ce qui implique donc une double exposition en lithographie 193nm immersion.

¹⁰ En cours de développement avec une mise en production envisagée pour les nœuds 7 et 5nm.

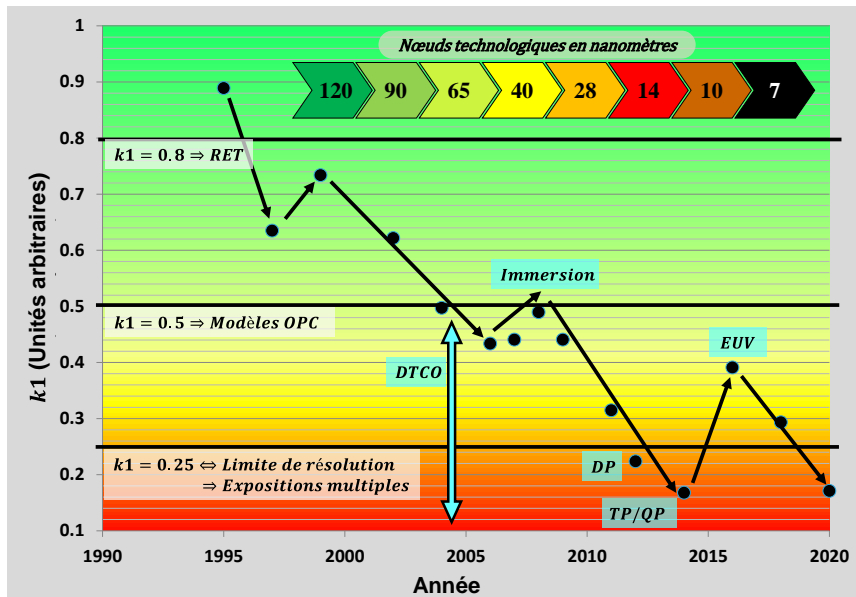


Figure 2-13 : Evolution du facteur k_1 dans le temps. En blanc, on peut voir les limites imposées par la réduction du facteur. En bleu, sont exposées quelques solutions technologiques. Les acronymes sont définis ci-dessous. (Source : B. Le-Gratiet)

La Figure 2-13 présente les options nécessaires à l'amélioration de la résolution dont le nombre et la complexité augmente quand k_1 diminue. Les paramètres à prendre en compte sont de plus en plus nombreux : la diffraction est plus importante, les interactions multi-niveaux sont de plus en plus nombreuses (pour l'alignement par exemple), les intégrations complexes introduisent des nouveaux niveaux de masque comme dans le cas des expositions multiples ou des matériaux inconnus, les machines ont besoin de plus de calibration et de contrôle, ... La lithographie assistée par ordinateur est devenue nécessaire depuis le passage à un $k_1 < 0.8$ car il a fallu introduire des illuminations complexes et des améliorations des masques. Personnellement, je considère que la lithographie holistique [15] [16] a réellement commencé avec l'apparition des modèles OPC et de l'immersion ce qui correspond à un $k_1 < 0.5$. Il est devenu nécessaire de prendre en compte de plus en plus de paramètres pour sécuriser la formation de l'image dans la résine. L'apparition du DTCO (Design Technology Co-Optimization ou Co-optimisation du design et des procédés, [17] [18] [19] [20]) en est un bon exemple.

Le principe du DTCO repose sur le fait que les OPC ne peuvent pas tout corriger. En particulier, certains modes d'illumination favorisent grandement certains motifs par rapport à d'autres. Il devient alors impossible de faire cohabiter sur le même design des motifs que l'on peut imprimer et des motifs que l'on ne va pas réussir à imprimer fidèlement dans la résine. Ces « hotspots » ou motifs critiques apparaissent quand leurs conditions optimales de procédé ne sont pas les mêmes que celles des autres motifs du design. Le DTCO permet d'optimiser le design afin de favoriser la fabricabilité de la puce et n'est pas réservé à la lithographie. Parmi les solutions que le DTCO apporte, on peut citer entre autres la direction préférentielle des motifs dans le design, l'interdiction de certains pas dans les réseaux ou la création de structures dites « dummies » ou motifs factices qui n'ont aucune utilité dans le fonctionnement électrique de la puce mais permettent d'homogénéiser le design, ...

Si on considère que les OPC, le RET, le DTCO sont à mettre en œuvre ensemble et qu'il faut y ajouter la métrologie de contrôle en ligne, les interactions entre les matériaux, l'utilisation de plusieurs machines pour réaliser chacune des étapes de fabrication et d'autres sources de variabilité, une approche holistique s'avère absolument nécessaire. Des corrélations peuvent alors émerger, permettant la mise en évidence, le contrôle et la prédiction de la variabilité sur le wafer.

2.5 LA FENETRE DE PROCEDE DE L'EXPOSITION

2.5.1 Fenêtre de procédé

Lors du développement d'un procédé de lithographie, il est nécessaire de choisir les conditions de procédé. Celles-ci sont nombreuses et on peut citer entre autre les matériaux, les conditions de dépôts de la résine, les températures des recuits, les épaisseurs de matériaux, la dose et le focus d'exposition, le mode d'illumination. Dans cette partie, nous nous intéresserons à la détermination de la dose et du focus d'exposition.

Pour cela, on expose une FEM ou Focus Exposure Matrix (Matrice Dose/Focus). Il s'agit d'un wafer de production sur lequel on va exposer plusieurs conditions croisées de dose et de focus. Chaque champ va donc être exposé de manière différente de son voisin comme suit (Figure 2-14).

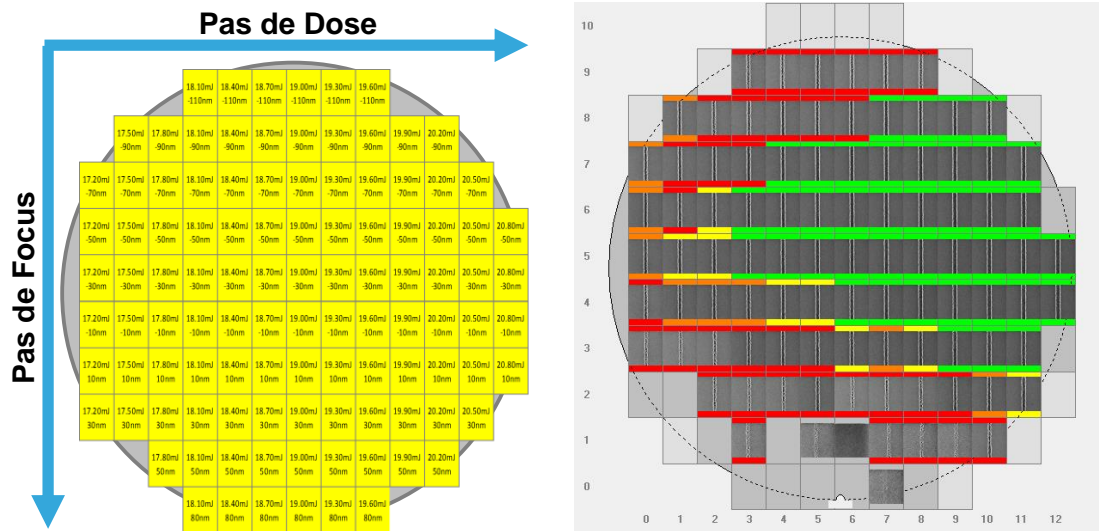


Figure 2-14 : Schéma d'une FEM. A gauche, cartographie des conditions dans chacun des champs (dose et focus) et à droite images MEB correspondantes de la tranchée isolée dans les conditions de la FEM. Les colorisations en vert, jaune, orange et rouge correspondent à la dimension du motif par rapport aux spécifications dimensionnelles du procédé.¹¹

On mesure ensuite des motifs de référence sur le wafer à l'aide d'un microscope électronique à balayage, ou CDSEM (Critical Dimension Scattering Electro Microscope), optimisé pour la mesure automatisée

¹¹Vert = ± 8% / Jaune = ± 12% / Orange = ± 16% / Rouge = ± 20% ou motif non integer.

de motifs de résine. Pour chaque condition (*Dose, Focus*), on obtient le dimensionnel associé sur le wafer. On trace alors la courbe de Bossung du motif pour déterminer le focus et la dose optimal. Un exemple est montré en Figure 2-15.

Chacune des courbes sur ces graphiques correspond à une variation iso-dose du CD en fonction du focus. On cherche le point de fonctionnement du procédé. Pour le focus optimum, cela revient à chercher la valeur de focus pour laquelle la variation en CD est la plus faible. Cela correspond à la valeur de focus à laquelle la dérivé du CD en fonction du focus est nulle. Pour la dose, il faut trouver la dose qui permet d'obtenir le CD voulu à la valeur de focus optimum.

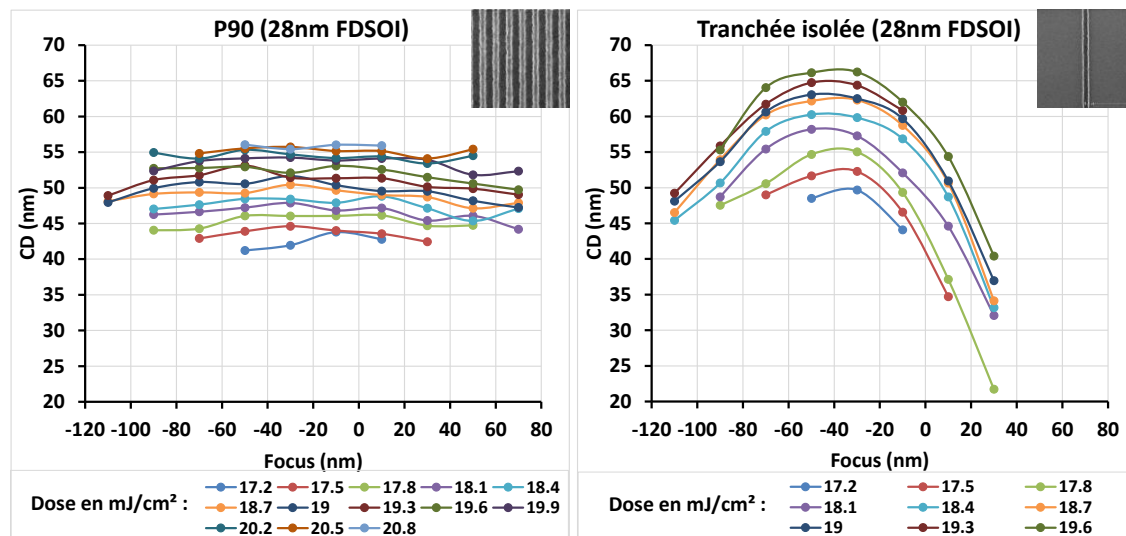


Figure 2-15 : Courbes de Bossung du motif dense (P90) et de la tranchée isolée en BEOL 28nm FD-SOI. En haut à droite de chaque graphique se trouvent les images SEM post-lithographie du réseau dense et de la tranchée isolée.

Dans ce cas, la dose optimale est de 18.4 mJ/cm^2 et le focus optimal est de -45 nm . On remarque que le motif isolé est beaucoup plus sensible au focus que le motif dense avec 2 à 4 nm de variation CD pour 200 nm de variation de focus contre 25 nm de variation CD pour un défocus de 75 nm pour le motif isolé. Le P90 est sensible à la dose avec 3.6 nm/mJ de variation CD en fonction de la dose. Le motif isolé varie de 6.8 nm/mJ .

Pour déterminer les conditions optimales d'exposition, un ensemble de courbes de tendances est calculé pour suivre au mieux les variations de dimensions mesurées à l'aide du SEM. Ces courbes de tendances suivent l'équation suivante et ont toutes les mêmes valeurs pour les coefficients de régression a, b, c, d, e et f . Il s'agit là de l'une des modélisations possible du CD en fonction du couple de conditions (*Dose, Focus*).

$$CD = a + b.Focus + c.Focus^2 + d.Dose + e.Dose.Focus + f.Dose.Focus^2 \quad (9)$$

La figure 2-16 représente les mesures de la tranchée isolée et le set de courbes de tendance correspondant. La fenêtre de procédé de l'exposition est extraite des courbes de tendance iso dose. En déterminant les couples (*Dose, Focus*) résolvant le système, on peut tracer les courbes iso-CD de la

dose en fonction du focus pour les valeurs maximale (USL ou Upper Specification Limit) et minimale (LSL ou Lower Specification Limit) de CD permettant d'atteindre les performances électriques désirées.

L'espace entre ces deux courbes est constitué de tous les couples (*Dose, Focus*) conduisant au bon dimensionnel du motif. Il contient la fenêtre de procédé du motif. Celle-ci donne la marge d'erreur de dose, ou latitude d'exposition (Exposure latitude EL), et la marge de focus, ou profondeur de champ (Depth of focus DOF). Les conditions optimales de procédé sont au centre de la fenêtre de procédé. Cette fenêtre est représentée par une ellipse illustrant les effets combinés de la dose et du focus en bord de fenêtre de procédé et peut être expliquée par l'analyse de FEM électriques ou de PWQ. Les FEM électriques sont des matrices d'exposition classiques qui suivent ensuite le reste des étapes de fabrications jusqu'aux tests électriques. La fenêtre de procédé conduisant aux performances requises au bon fonctionnement de la puce est extraite de ces mesures. La PWQ (Process Window Qualification ou qualification de la fenêtre de procédé) est une méthode permettant de relier des conditions d'exposition similaires à la FEM et les densités de défauts dans la puce. On expose le wafer sous plusieurs conditions et une inspection de défauts est réalisée. Le problème de la densité de défauts est décrit plus en détail comme exemple de conséquence de la variabilité dans le Chap. 3.5.

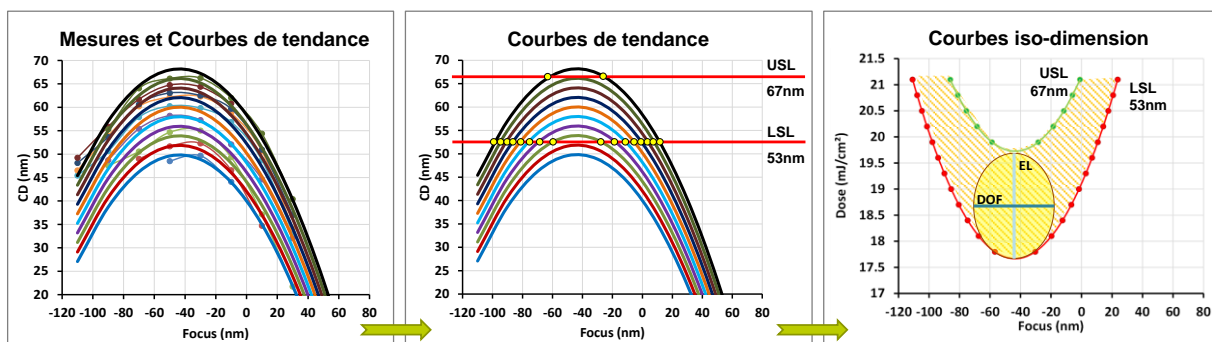


Figure 2-16 : Fenêtre de procédé du motif isolée en 28nm BEOL. Le focus optimum (BF pour Best Focus) est de -45nm et la dose optimale (BD pour Best Dose) est $18.6\text{mJ}/\text{cm}^2$. La profondeur de champ (ou DOF pour Depth of Focus) est de l'ordre de 60nm et la latitude d'exposition (EL) est de $\pm 9\%$ de variation de dose environ.

Dans la suite de ce manuscrit, nous ne nous intéresserons qu'aux effets de focus. Pour cela, soit les pas de dose ne seront pas exposés sur le wafer et seules des variations de focus seront réalisées, soit la variation à la dose sera normalisée. L'effet du focus est pris comme étant un effet quadratique et celui de la dose un effet linéaire sur la variation de CD.

2.5.2 Décalage de focus

Sur un masque, on trouve toute sorte de motifs différents dont la forme, la taille, la densité dépend de leur fonctionnalité dans la puce complète. En raison de leurs formes variées et des effets de masques 3D (cf. Chap. 3.2.2.2), ces motifs ne diffractent pas exactement de la même manière. Ainsi, la reconstitution de l'image aérienne de chaque motif ne se fait pas dans le même plan focal. La profondeur de champ de

chaque motif n'est pas non plus la même, car en changeant la diffraction du motif, l'angle de diffraction n'est pas identique, ce qui modifie l'ouverture numérique et donc aussi la DOF.

2.5.3 Décalage de dose

Comme pour le focus, la dose d'exposition présente des variabilités au sein de la plaque et du champ. On peut citer entre autres les effets suivants :

- Les effets « CD vs. Pitch » (ou dimension critique par rapport au pas du réseau) pour lesquels le NILS va varier fortement avec le motif, imposant une dose optimale plus élevée pour les configurations pour lesquels le NILS est faible afin de compenser la perte de contraste.
- Les effets d'uniformité du masque qui sont issus des variations de CD sur le masque suite à sa fabrication. Le même motif présentant alors des variations de CD sur le masque aura besoin de conditions de dose différentes pour être imprimé de manière identique sur le wafer.
- Les effets de réflectivité du wafer qui va causer des différences d'intensité de l'image aérienne en fonction des motifs déjà présents sur le wafer.
- Le biais masque (mask bias [21]) est un effet de dimensionnement de l'image aérienne qui impose un dimensionnement spécifique des motifs en fonction de leur densité (CD vs Pitch) sur le masque afin de les imprimer avec le bon dimensionnel sur le wafer.

2.5.4 Quelques exemples

La fenêtre de procédé (*Dose, Focus*) peut être déterminée de nombreuses façons : dimensionnel des motifs, performances électriques du circuit, présence ou non de défauts d'impression. La figure 2-17 permet la comparaison des différentes fenêtres de procédé déterminées de manières diverses sur le même wafer FEM. Les conditions de la FEM sont les suivantes :

- variation de dose entre $19.3\text{mJ}/\text{cm}^2$ et $21.3\text{mJ}/\text{cm}^2$ avec des pas de $0.2\text{mJ}/\text{cm}^2$.
- variation de focus entre -80nm et $+55\text{nm}$ avec des pas de 15nm .

On remarque que la fenêtre de procédé CD déterminée avec la mesure CDSEM de la structure de référence de métrologie est bien plus large que les fenêtres de procédé électrique et définitivité. Cela s'explique par l'analyse de la variation d'un seul motif avec la mesure CDSEM alors que les tests électriques et les inspections de défauts prennent l'ensemble de la puce en considération.

Au final, c'est la fenêtre de procédé électrique qui est celle recherchée mais si celle-ci pourra être de nouveau modulée par les tests de fiabilité des puces.

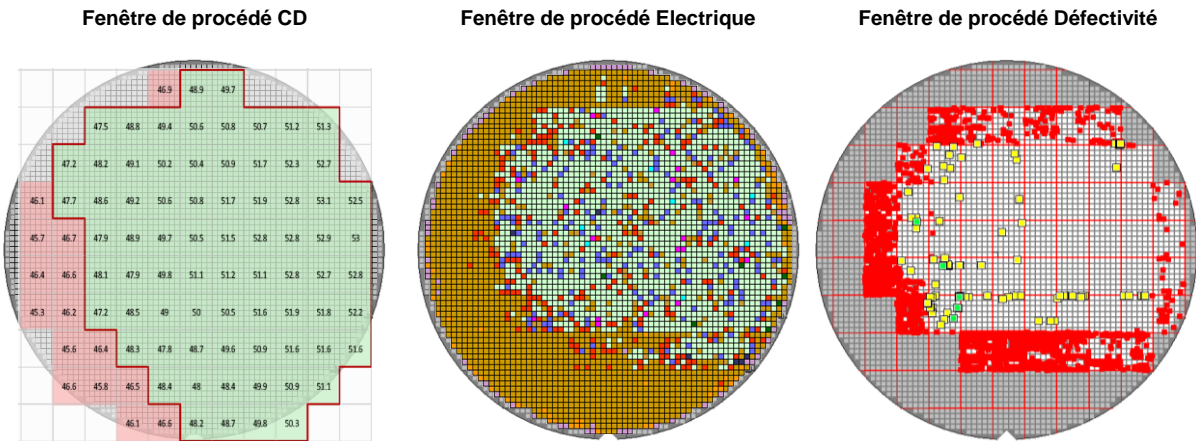


Figure 2-17 : Fenêtres de procédé d'un niveau d'interconnexion en 28nm déterminées par trois méthodes différentes sur le même wafer FEM. Sur la fenêtre CD, la zone verte pâle est la fenêtre de procédé alors que sur les autres, il s'agit de la zone en fond blanc.

2.6 LES CHALLENGES DE LA THESE

Ce sujet de thèse s'inscrit dans la logique de lithographie holistique. Le contrôle du focus d'exposition est en effet régi par de nombreux paramètres très différents qui interviennent en parallèle. Afin de pouvoir appréhender le problème sous tous ses angles, il est nécessaire de pouvoir récupérer des données issues de différentes sources : monitoring de la machine, contrôle en ligne du procédé de lithographie, mesures de différents paramètres, intégration et design du circuit, et ceci sur 8 ordres de grandeur spatiales allant du nanomètre à la dizaine de centimètres.

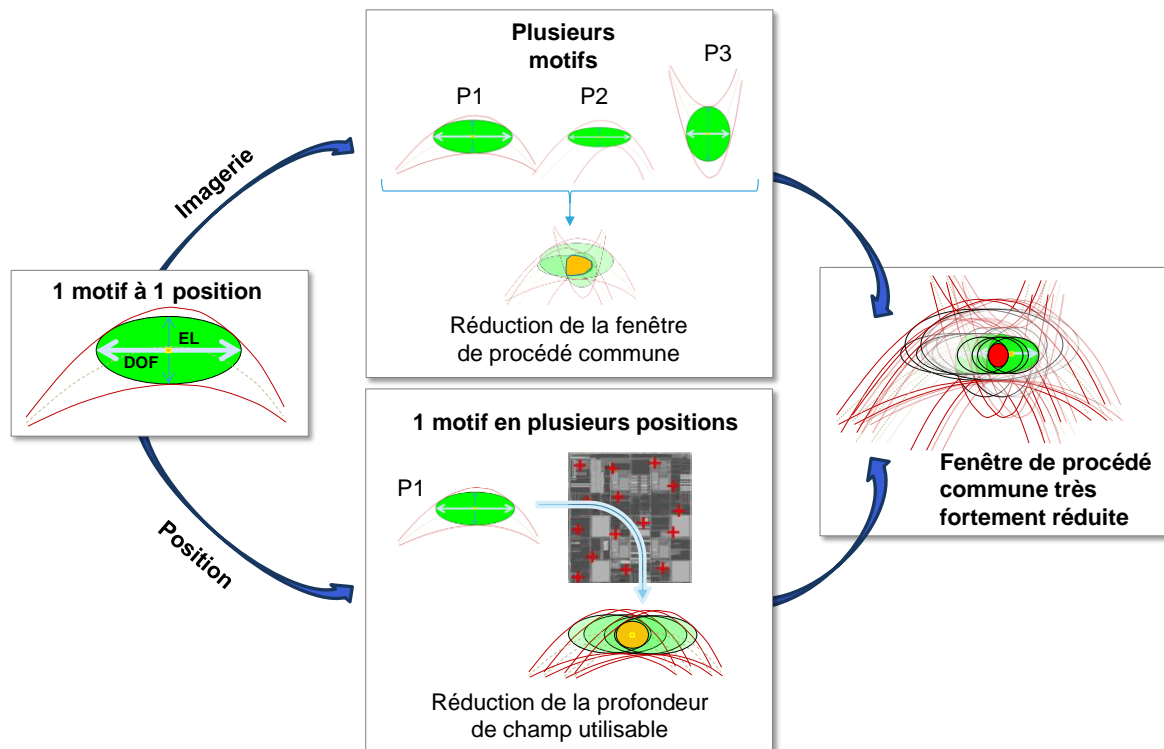


Figure 2-18 : Réduction de la fenêtre de procédé et effets optique et topologique.

Cette thèse a pour objectif d'appréhender l'ensemble de ces sources de données, de trouver des corrélations entre elles et d'en extraire les informations permettant de mettre en évidence l'influence de chaque source de variabilité afin d'optimiser le contrôle du focus. Au travers de cette étude, il a été montré l'intérêt d'une telle approche multi-sources dans le cadre de la sécurisation du rendement. L'ensemble du travail réalisé autour de la topographie intra-puce comme sujet de cette étude avait pour but de quantifier l'influence de ce paramètre et de trouver des solutions pour l'inclure dans une optimisation globale du procédé mais aussi de servir de démonstration du concept d'approche holistique et de sa validation.

Concernant la topographie, l'objectif était de démontrer son influence sur le focus et la mise en œuvre d'une méthode de prédiction de celle-ci ainsi que les opportunités et solutions proposées en termes d'optimisation du suivi de production en lithographie, de l'exposition et des règles de dessin qui seront présentées dans la suite vont toutes être orientées dans ce sens.

CHAPITRE 3

3 ETUDE SUR LE BUDGET ET LE CONTROLE DU FOCUS

Cette partie développe les challenges liés à la focalisation de l'image sur la résine lors de l'exposition du masque (contenant le design de la puce) sur le wafer de silicium.

Comme il a été expliqué précédemment, la sensibilité au focus et le focus optimum d'exposition sont dépendants de la forme du motif que l'on veut imprimer sur le wafer. Cependant, d'autres effets, issus de la machine, du masque, du wafer ou bien liés au design même de la puce, vont aussi être des éléments perturbateurs du focus. Sachant que plus les dimensions diminuent et plus la profondeur de champ disponible pour le procédé est faible, cette variabilité du focus pendant l'exposition est de plus en plus susceptible de provoquer des distorsions d'images locales pouvant être à l'origine de défauts majeurs (non fonctionnalité de la puce).

Le budget focus [22] consiste en un seuil de variabilité minimum qu'une configuration design/procédé/machine peut tolérer. Maintenir la configuration dans une situation viable consiste à s'assurer que la profondeur de champ du procédé ne descend pas en dessous de ce seuil. Ce budget, exprimé en écart-type 3σ , considère les performances du scanner, l'intégration, la métrologie. Il doit être comparé à la profondeur de champ disponible calculée par simulation optique de la formation de l'image aérienne, ou mesuré par FEM électrique, PWQ, ...

La figure 3-1 présente deux situations différentes de budget focus. Dans la première, la profondeur de champ disponible reste supérieure au budget. Dans le second cas, la profondeur de champ disponible est plus faible que le budget en raison d'un nœud technologique plus avancé et d'une imagerie plus critique malgré une diminution du budget focus en raison d'améliorations en termes d'intégration et de contrôle des variabilités. L'objectif est de se trouver dans la première situation. En effet, la profondeur de champ doit être supérieure au budget focus afin de sécuriser le rendement. Un budget supérieur à la DOF dénote une situation où l'impact de rendement lié au focus est systématique sur le wafer.

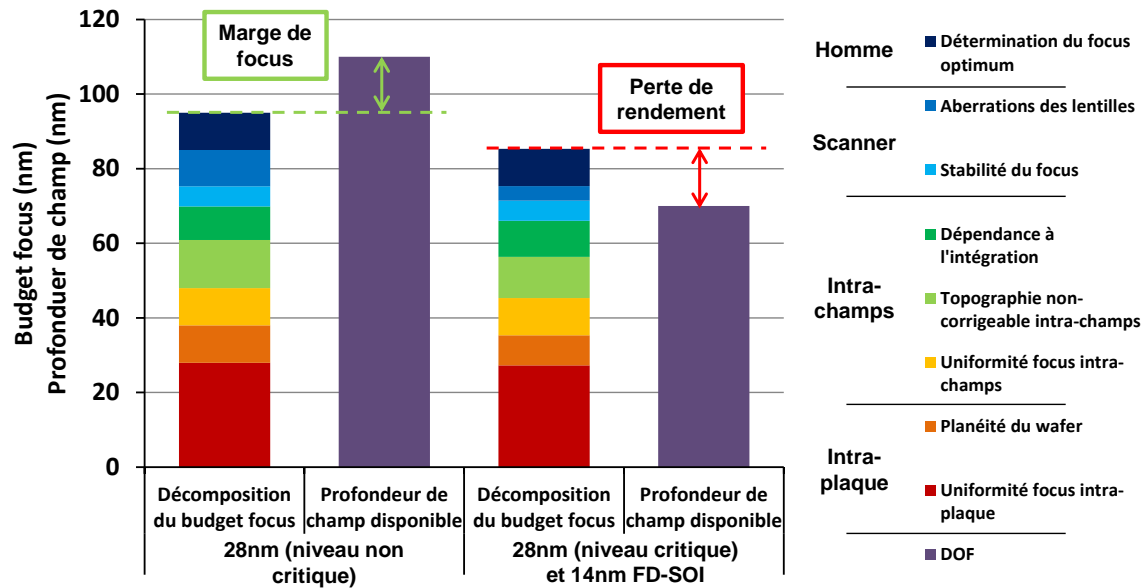


Figure 3-1 : Comparaison du budget focus pour deux situations. Celle de gauche est viable alors que celle de droite est critique.

Les budgets focus pour le BEOL 28nm et le Contact 14FD-SOI ont été établis dans le cadre du programme ECLIPSE entre STMicroelectronics et ASML. Le programme ECLIPSE, sur lequel la quasi-intégralité du travail réalisé dans cette thèse s'est greffée, consiste en un programme d'amélioration continue des procédés, des contrôles de lots et de machines, des modèles de régulations de production et de la machine. La figure 3-1 a servi de base pour la thèse. Le travail développé dans la suite de ce manuscrit cherchera à mettre en évidence les sources de variabilité à l'origine de ce budget et particulièrement au niveau de l'intra-champ, qui correspond à environ 50% du budget total (en vert foncé, vert clair et jaune sur le graphique).

Il y a deux manières d'améliorer le budget focus, soit en améliorant la machine (diminution du budget focus), soit en améliorant l'imagerie (augmentation de la DOF).

Dans la suite de cette partie, à partir de mesures de focus optimum (voir Chap. 2.5) une décomposition des différentes sources de variabilité indépendantes de la formation purement optique de l'image sera faite.

3.1 ANALYSE ET DECOMPOSITION DU BUDGET FOCUS

3.1.1 La multi-wafer matrice de focus

Dans la partie précédente, la FEM (Focus Exposure Matrix ou Matrice d'exposition) a été présentée. Elle est réalisée sur une unique plaquette de silicium pour déterminer la fenêtre de procédé en dose et en focus pour un produit donné. La FEM présente l'avantage de ne nécessiter qu'une seule plaquette qui, après mesures, permet d'obtenir des données de sensibilité en dose et en focus à l'échelle d'un champ complet. La contrepartie est qu'il est impossible de donner les variations sur une plaquette

entière. Pour ce faire, la méthodologie de la cartographie ultra-dense de focus (Hyper dense focus map ou HDFM) a été développée au cours de cette thèse.

Afin d'obtenir des variations de focus sur produit à l'échelle d'un wafer entier, il est nécessaire de modifier la méthodologie de matricage de conditions d'exposition. Pour cela, la multi-wafer FEM a été mise en place. Il s'agit de réaliser le même type de manipulations que pour une FEM mais en utilisant plusieurs wafers au lieu d'un seul. Chaque plaquette est exposée à une condition différente des autres. Dans cette étude, ce sont les variabilités focus qui nous intéressaient, aussi seuls des pas de focus ont été réalisés sur les wafers de l'expérience. Il est tout à fait possible d'imaginer faire de même avec la dose.

L'exemple suivant a été réalisé au niveau Contact en 14nm FD-SOI sur 7 wafers avec des pas de focus de 15nm autour du focus optimum, tous exposés à la dose optimum utilisée en production. Cette expérience a permis de développer la méthodologie de l'HDFM. La matrice d'expérience est donnée en tableau 3-1.

Numéro du wafer	Focus d'exposition	Décalage à l'optimum
19	-45nm	BF + 45nm
20	-120nm	BF - 30nm
21	-90nm	BF
22	-60nm	BF + 30nm
23	-105nm	BF - 15nm
24	-135nm	BF - 45nm
25	-75nm	BF + 15nm

Tableau 3-1 : Tableau récapitulatif de la matrice de focus multi-wafer. BF (Best Focus) désigne la valeur optimale de focus utilisée en production.

Les plaquettes sont exposées dans l'ordre donné dans le tableau. Les expositions n'ont volontairement pas été faites dans un ordre croissant ou décroissant de focus pour éviter l'accumulation d'erreurs de positionnement de la plaquette. Ces erreurs « servo » dues aux moteurs qui contrôlent le positionnement de la plaquette pendant l'exposition (cf. Chap. 3.2.2.3 « Les effets du scanner » et Chap. 3.4.1 « Focus et topographie » avec le paragraphe sur le « leveling ») sont inévitables et dépendent du sens dans lequel le chuck se déplace. Le mélange des focus d'exposition permet de compenser une partie de ces erreurs.

Après exposition, environ 5000 points de mesures par plaquette (soit près de 35000 points au total) sont mesurés avec un CDSEM CG5000 de Hitachi. Le motif de référence sélectionné pour l'étude est un réseau semi-dense de lignes de 45nm de large avec un pas de 188nm. Il s'agit du motif de contrôle en ligne du procédé et il est sensible au focus. Cela correspond à environ 60h de mesure et, par conséquent cette méthode reste purement applicable à des travaux de recherche. Le plan de mesures est donné en Figure 3-2.

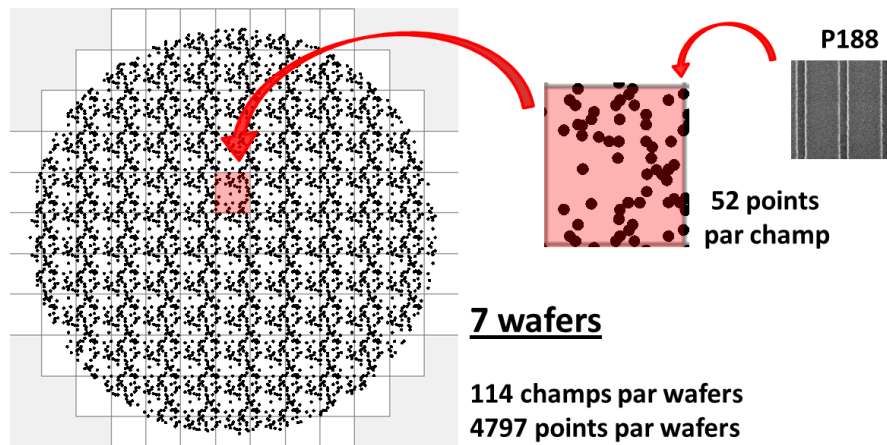


Figure 3-2 : Plan de mesure de la Multi-wafer FEM

Pour chaque position sur la plaque, nous disposons d'une mesure par wafer correspondant à une valeur de focus choisie pour l'exposition. La Bossung iso-dose du motif pour chaque position intra-wafer est tracée. La même chose peut être faite pour l'intra-champs en moyennant les valeurs pour chaque position intra-champ sur le wafer. Il est ainsi possible de déterminer le focus optimum pour chacune des positions mesurées sur le wafer. N'utilisant que des pas de focus dans cette expérience, le CD peut être considéré comme variant de manière quadratique par rapport au focus. Pour chaque position intra-wafer, le focus optimum se situe au point de dérivée nulle sur la courbe. Ma méthode présentée lors d'une revue de programme a fortement intéressé ASML qui avec ses outils propres a proposé une méthode optimisée qui aujourd'hui fait partie de l'outil PFC (cf. Chap. 5.2 « *Pattern Fidelity Check* »).

3.1.2 Imagerie vs. Focus mesuré sur plaquette

3.1.2.1 Un unique focus optimum par motif

On observe que les mesures montrent des variations de ce focus pour différentes positions au sein du champ mais aussi des Bossung qui paraissent être différentes. Partant de l'hypothèse qu'il n'existe qu'une unique Bossung par motif, dépendant uniquement de la forme du motif et de la formation de l'image aérienne associée, les courbes extraites des mesures devraient être les mêmes pour chacun de ces points de mesure. La valeur de la sensibilité au focus en particulier (a dans l'équation de régression Eq. 10) a été fixée lors des interpolations des mesures.

$$CD = a * Focus^2 + b * Focus + c \quad (10)$$

La figure 3-3 ci-dessous donne les cartographies de focus obtenues pour la méthode simple que nous avons développée et la méthode optimisée de détermination du focus optimal par interpolation quadratique en fonction du focus des courbes de Bossung. On remarque que la forme de l'empreinte focus du procédé sur la plaquette est la même dans les deux cas. L'intérêt de la méthode optimisée est principalement de diminuer le bruit de détermination de la valeur du focus optimum pour chaque position.

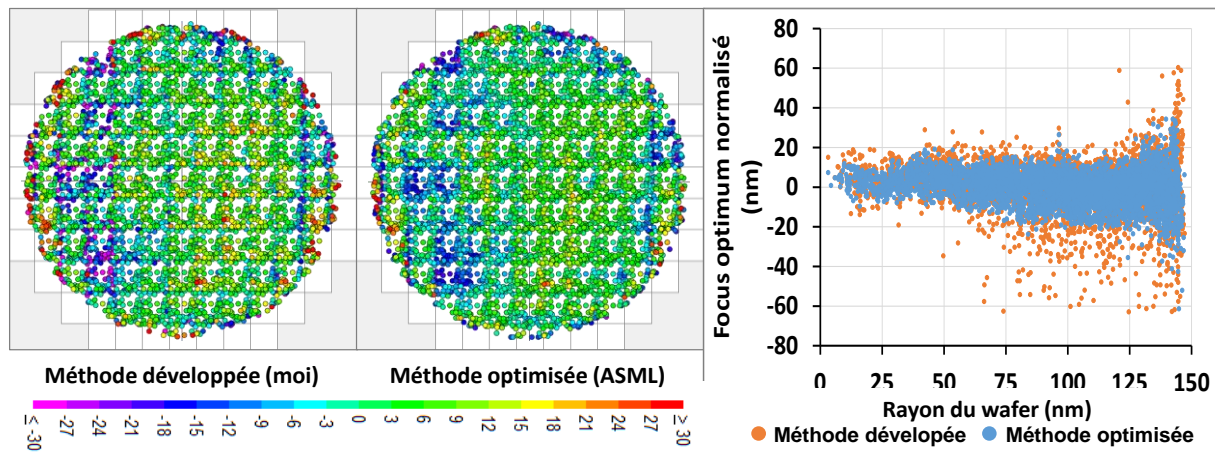


Figure 3-3 : Comparaison des cartes d'uniformité focus intra-wafer avec la méthode classique et la méthode optimisée.

3.1.2.2 Reconstruction de la Bossung

En représentant sur le même graphique les valeurs de CD pour chacune des mire et à chaque focus choisi pour l'exposition de la multi-wafer FEM, la Figure 3-4 est obtenue.

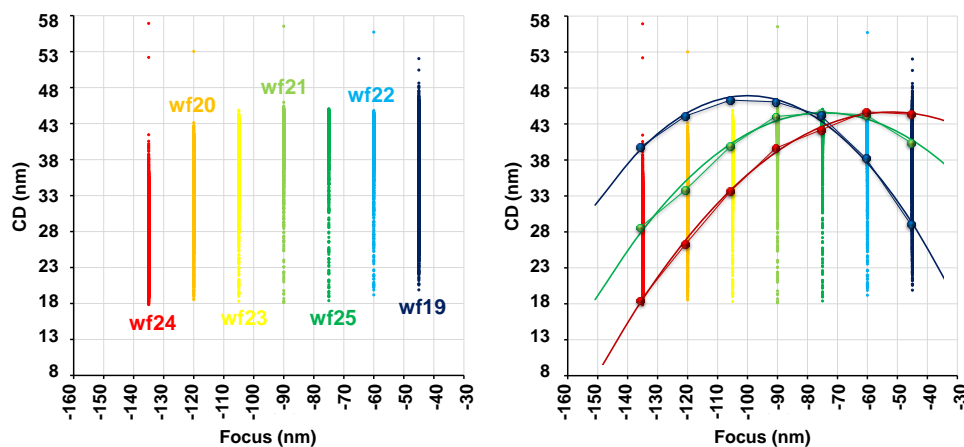
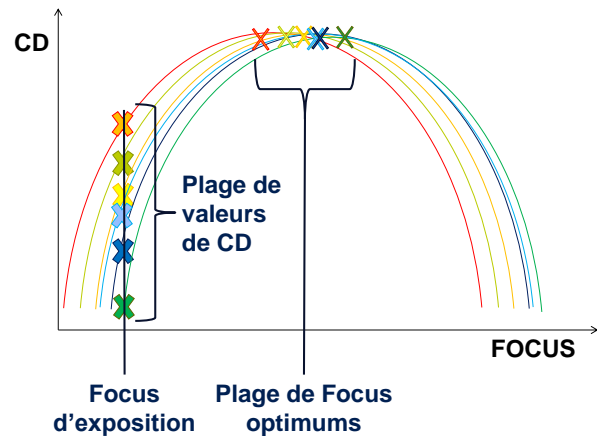


Figure 3-4 : Mesure dimensionnelle CDSEM du P188 en Contact 14nm FD-SOI en fonction du focus de centrage d'exposition de la multi-wafer FEM. Chaque couleur représente un wafer différent. Le graphique de droite représente les Bossung du motif en trois positions sur la plaquette.

Les valeurs de CD couvrent sur chaque plaquette une même plage de valeurs entre 18 et 45 nm en moyenne. Le graphique est en réalité composé d'une superposition de nombreuses courbes représentant le CD en fonction du focus, soit une par position intra-plaquette mesurée. Sont représentées sur le graphique quelques-unes d'entre elles à titre d'illustration et pour une meilleure compréhension. Une reconstitution d'une seule et unique Bossung pour ce motif s'obtient en réalignant les Bossung par position les unes par rapport aux autres en prenant une seule et unique valeur de focus optimum pour chacune d'entre elle. La méthode est décrite dans les prochaines pages.

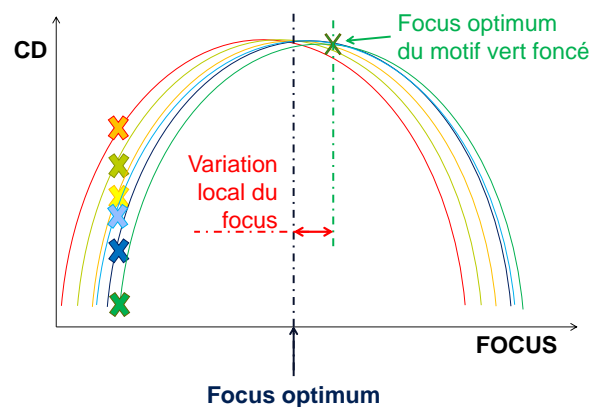
(a) Situation initiale

Avant la reconstruction de la Bossung, chaque valeur de CD mesurée au microscope électronique est associée au focus qui a été programmé sur le scanner pour l'exposition. Pour chaque pas de focus, la plage de CD est assez étendue, de l'ordre de 15nm 3σ dans le cas de cette étude. Pour chaque position, l'interpolation quadratique de la Bossung iso-dose est calculée et la valeur de focus au point de dérivée nulle est extraite avec la méthode optimisée. La précision de détermination du focus optimal a été estimée à 3nm 3σ .



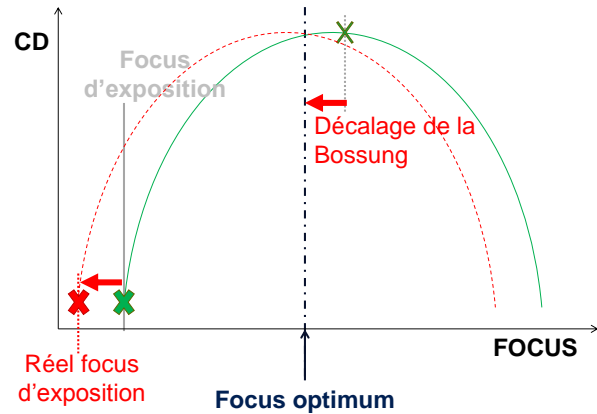
(b) Calcul de l'écart au focus moyen

Sur chaque Bossung, la valeur du focus optimum est normalisée par rapport à la valeur moyenne du focus sur l'échantillon, pouvant être considéré du fait de la statistique (près de 5000 points de mesure) comme le focus optimum réelle d'exposition du motif. Cette normalisation permet de calculer le défocus local de chaque position du champ. L'ensemble de ces valeurs permet de déterminer la variabilité normalisée du focus intra-wafer.

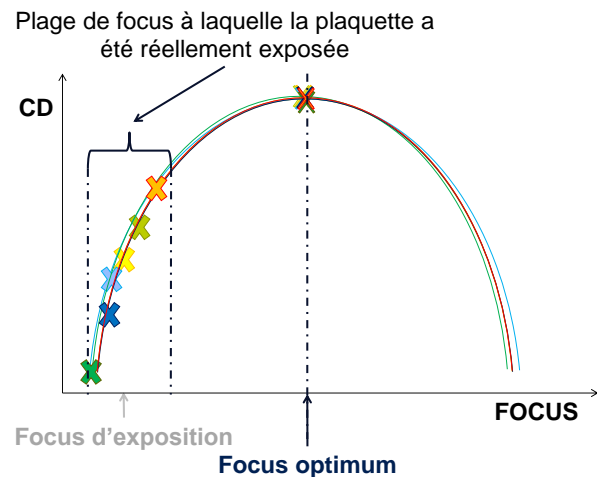


(c) Centrage des Bossung sur le focus optimum**moyen**

Chaque point de mesure est repositionné de son écart au focus moyen pour être recentré sur cette valeur de focus. Ainsi pour chaque valeur de CD, il est possible d'associer la valeur réelle du focus auquel le point a été exposé. Les opérations décrites en (b) et en (c) sont répétées pour tous les points de mesures de l'expérience.

**(d) Bossung unique**

Après avoir réajusté toutes les Bossung par centrage sur le focus optimum moyen, une unique Bossung est obtenue. Autour de chaque focus théorique d'exposition de la matrice d'expérience, s'observe désormais une distribution non uniforme de focus correspondant aux valeurs réelles auxquelles le wafer a été localement exposé.



La courbe de Bossung obtenue (Figure 3-5) présente pour chaque wafer non plus une variation de CD pour un seul focus mais une variation de CD correspondant à une variation intra-plaquette du focus d'exposition. On démontre ainsi que chaque wafer est en réalité exposé non pas à un focus unique choisi par le lithographe lors de la détermination des conditions de procédés optimales pour le niveau de masque en question mais à une distribution de focus autour d'une valeur optimale. La répartition du focus sur la plaquette est de $27.7\text{nm } 3\sigma$ autour de la valeur moyenne de focus. La cartographie de la Figure 3-6 représente cette distribution spatiale des variabilités focus sur une plaquette.

La variabilité en CD pour un focus donnée n'est néanmoins pas entièrement supprimée. Il s'agit d'autres contributions de variabilités non liées au focus comme des non-uniformités en dose, des erreurs de gravure du masque qui se transfèrent sur la plaque, ... qui ne font pas partie de cette étude.

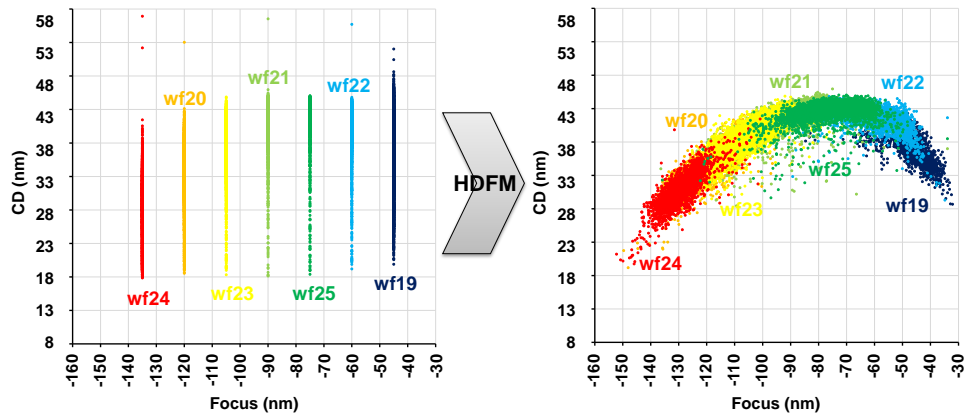


Figure 3-5 : Bossung réelle du P188 (à droite) extraite des mesures brutes (à gauche) à partir de la méthode précédemment expliquée. La courbe de gauche représente le CD en fonction du « set focus » et celle de droite en fonction du « get focus ».

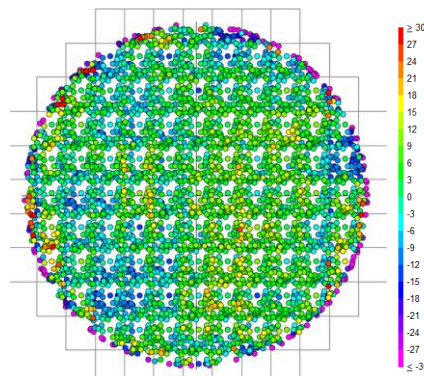


Figure 3-6 : Cartographie de l'uniformité focus sur un wafer 14FD-SOI au niveau Contact (à droite). Cette carte correspond la répartition spatiale de la distribution en focus du wafer 25, en vert foncé sur le graphique de la Figure 3.5.

Chaque point du wafer est donc en réalité exposé à une valeur de focus différente alors que les conditions de procédé sélectionnées sur le scanner correspondent à une valeur unique de focus : il s'agit du focus sélectionné (Set Focus) et du focus obtenu sur wafer (Get Focus). Cette variabilité se traduit par des non-uniformités dans le fonctionnement de la puce ou entre plusieurs puces sensément identiques. La question est maintenant de déterminer l'origine de ces 27.7nm 3σ de distribution de focus.

3.2 VARIABILITE FOCUS

3.2.1 Les différents niveaux de variabilité

La variabilité peut être définie comme l'ensemble des fluctuations des propriétés et des performances visées dans la fabrication ou le fonctionnement des circuits imprimés. Elle correspond simplement à une dispersion autour d'un point optimum. On peut les classer de diverses manières selon trois axes : (1) le « moment auquel l'effet se produit », (2) le mode d'apparition et (3) l'échelle ou répartition spatiale.

- Variabilité procédé et variabilité de fonctionnement

Les variations qui se présentent lors du fonctionnement du circuit fini sont environnementales (température, alimentation, ...) alors que celles qui apparaissent lors de la fabrication de la puce sont

physiques (désalignement, uniformité de dépôt, effets de densité en gravure, ...). Ici, nous ne traiterons que de ce deuxième cas.

- *Variabilité systématique et variabilité stochastique*

Les effets systématiques sont modélisables et donc prédictifs alors que la variabilité stochastique survient de manière imprévisible. Dans cette étude, nous nous intéresserons principalement aux effets systématiques.

- *Variabilité puce à puce et variabilité intra-puce*

Ce classement concerne l'échelle spatiale à laquelle l'effet se produit et a une influence. La première échelle de variabilité, puce à puce, concerne des variations pouvant être détecté entre des puces en théorie identique (à l'échelle d'un wafer, d'un lot ou entre plusieurs lots de production). Les effets intra-champ correspondent à des fluctuations d'un paramètre au sein même d'une puce entre des composants théoriquement identiques sur l'ensemble du produit.

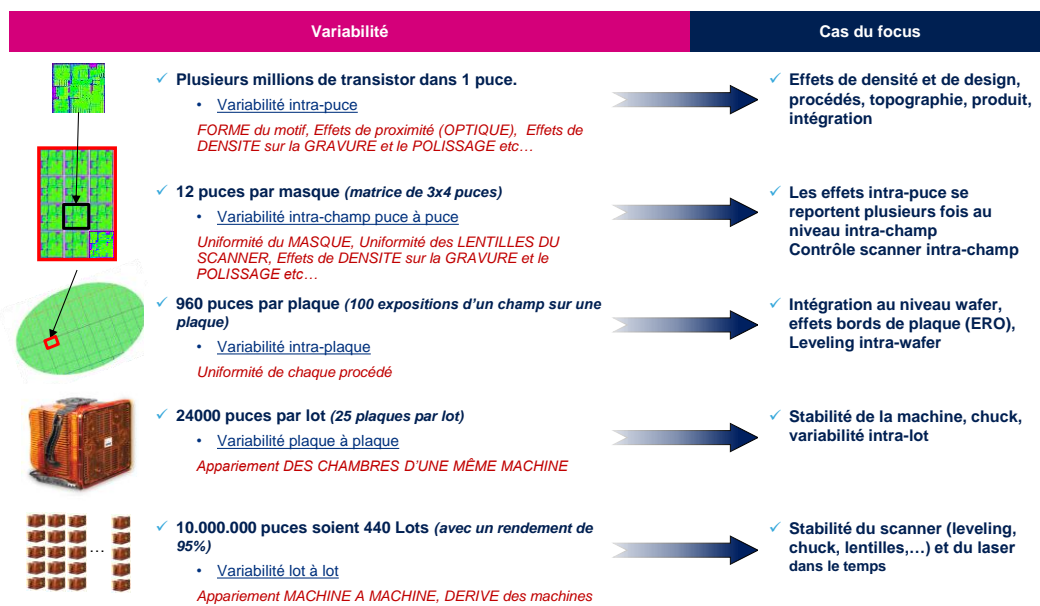


Figure 3-7 : Illustration des différents niveaux de variabilité avec la production de 10 millions d'unités d'une puce de 7x7mm² (Source : B. Le-Gratiet)

La figure 3-7 donne un aperçu non exhaustif d'effets procédé à différentes échelles sur un exemple de production de 10⁷ pièces identiques d'une puce électronique¹².

Dans la suite de cette étude, seuls les challenges de contrôle de la variabilité focus seront développés.

¹² L'acronyme ERO signifie Edge Roll-Off. Il s'agit d'une déformation mécanique de grande amplitude (plus de 150nm) en bord de plaque en raison des tensions / compressions de la plaquette suite aux différents procédés.

3.2.2 Les sources de variabilité

Il existe différents mécanismes de variabilités focus sur le wafer. Le Tableau 3-2 classe la majeure partie d'entre eux en deux catégories : effets optiques et effets topologiques.

Source	Effets optiques	Effets topologiques
Masque	✓ Masque 3D [23]	✓ Support quartz [24] [25]
Scanner : Lentilles	✓ Aberrations [26], coma, échauffement [27] [28] [29]	✗
Scanner : Laser	✓ Largeur de bande [30] [31]	✗
Scanner : Chuck	✗	✓ Planéité, maintien de la plaque [32]
Scanner : Capteur de niveau	✓ Erreur de mesure [33]	✓ AGILE [34]
Scanner : Wafer table	✗	✓ Moteurs [32]
Wafer	✓ Effets d'empilement sur le capteur de niveau [33] [32] et sur l'image aérienne [35]	✓ Intégration, topographie, produit [36] [37] [22] [38], Edge Roll-Off [39]
Design	✓ Effets de diffraction sur l'image aérienne Modulation de la réflectivité du wafer [35]	✓ Modulation de la topographie du wafer [40] [17] [41]

Tableau 3-2 : Liste des effets impactant le focus, classés par sources et mécanismes

3.2.2.1 Les motifs

La première source de variabilité focus est l'offset de focus entre deux motifs différents en raison de leurs imageries différentes, expliqué en Chap. 2, avec l'exemple du biais iso-dense. Pour le corriger, des techniques d'amélioration de la résolution (RET) sont appliquées au design et le masque est fabriqué avec ces corrections pour permettre une impression optimale des motifs sur le wafer. Les méthodes RET ont pour principe de trouver les meilleures conditions d'illumination et de déterminer les transformations de motifs nécessaires à l'impression d'un motif. Quand les corrections à apporter deviennent trop importantes voire impossibles, des restrictions de règles de dessins sont nécessaires pour éviter au maximum la cohabitation au sein du même design de motifs dont les fenêtres de procédé ne sont plus compatibles [19].

Pour mettre en évidence cet effet, deux FEMs (cf. Chap. 2.5 « La fenêtre de procédé ») ont été exposées sur deux wafers identiques de silicium brut. Utiliser des wafers de silicium brut permet de s'affranchir des effets d'empilement et des effets du produit décrits dans les chapitres suivants de ce Chap. 3. Deux wafers ont été exposés pour limiter l'influence du scanner qui sera réduit par le calcul de la moyenne des mesures. Le niveau sélectionné est un niveau métal 28nm FD-SOI avec des pas de dose de 20mJ/cm² et des pas de focus de 20nm. Au niveau des blocs de métrologie, des motifs très différents ont été sélectionnés et mis en place par Anna SZUCS, une doctorante en OPC avant le début de ma thèse et sont toujours disponibles [42] [43]. On trouve le motif de monitoring en ligne utilisé en production qui consiste en un réseau de lignes parallèles avec un pas de 90nm nommé P90, une tranchée isolée (ISO)

et 8 motifs dont la fenêtre de procédé est très réduite nommés Hot Spots 1 à 8 (HS1 à 8). La figure 3-8 présente les images MEB des dix mires de l'étude. Les mesures ont été réalisées dans un seul bloc de métrologie au centre du champ.

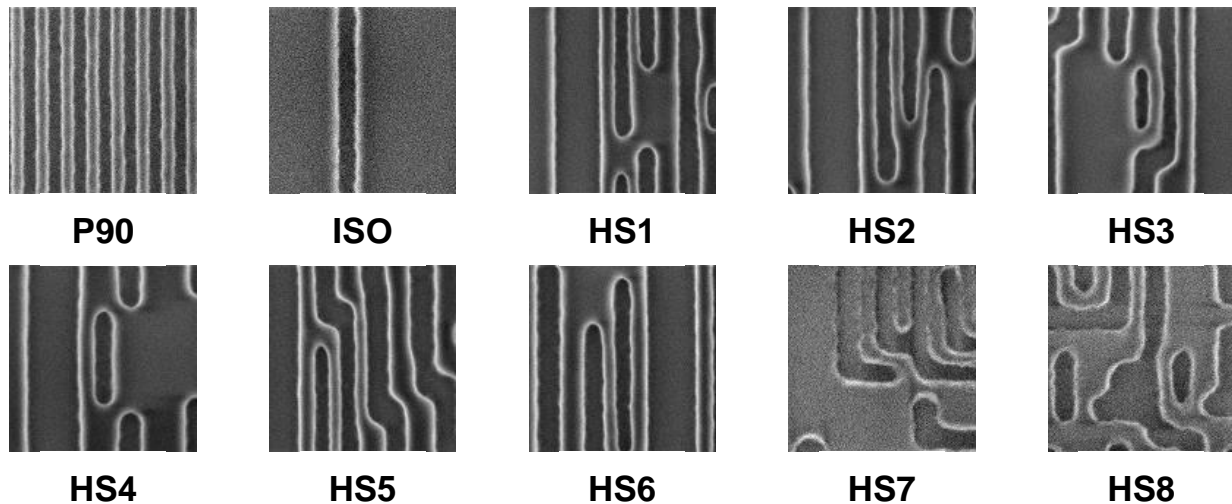


Figure 3-8: Images MEB des deux mires dense et isolée ainsi que des 8 Hot Spots étudiés

Pour chaque motif, le focus optimal et la dose optimale ont été calculés en utilisant la fonction donnée en Chap. 2.5 puis les effets de la dose ont été normalisés pour ne mettre en évidence que l'influence du focus. Le graphique en Figure 3-9 présente les fenêtres de procédé en focus calculées pour ces 10 motifs différents, c'est-à-dire le focus optimum et la profondeur de champ. Le focus optimal présente une variabilité de l'ordre de 25nm entre les différents motifs mesurés. Les profondeurs de champ varient entre 70nm pour les plus faibles et 285nm pour le P90. Ce résultat corrobore le fait que le motif dense est moins sensible au focus que des motifs plus isolés. Le graphique montre aussi que chaque motif présente un focus optimum d'exposition différent.

La cohabitation de ces différents motifs permet de calculer l'uDOF (Usable DOF ou Profondeur de champ utilisable) pour cette position. L'uDOF correspond à la profondeur de champ commune à tous les motifs. Elle est ici de 65nm.

Dans la même figure, sont données les cartographies de focus et de DOF calculés par LMC (Lithographic Manufacturability Check ou Vérification de faisabilité du procédé lithographique). Pour cela, l'image latente dans la résine est calculée par simulation optique sur un wafer idéalement plat et la position verticale du plan image ainsi que la profondeur de champ en chaque point du champ est extraite. Contrairement aux mesures qui permettent de caractériser les conditions de focus en une position donnée pour un motif donné, les cartographies LMC donnent la variabilité motif à motif pour un champ complet.

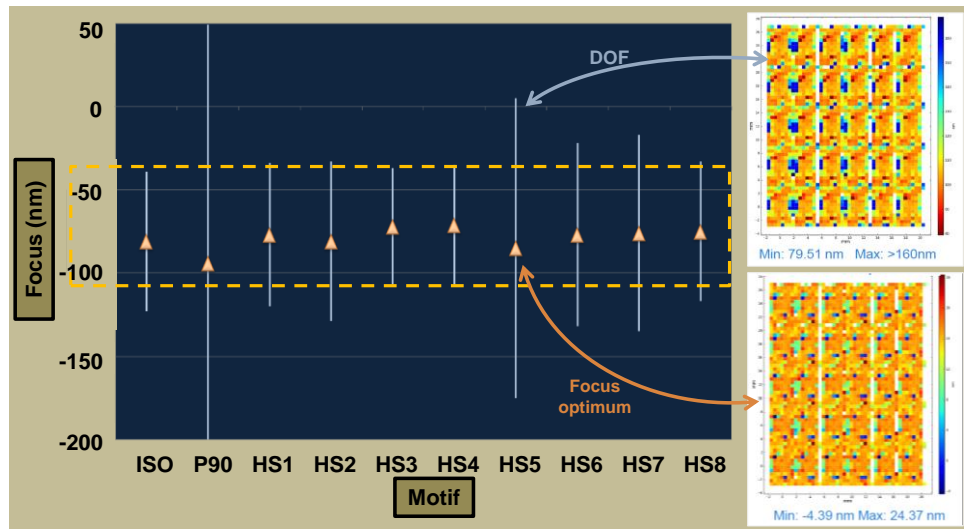


Figure 3-9 : Focus optimaux et profondeurs de champ de plusieurs motifs en BEOL 28nm (à gauche), carte de variation de profondeur de champ (à droite en haut) et de focus (à droite en bas) sur une puce. Les données du graphique de gauche sont des mesures sur silicium et à droite des simulations (source : Tachyon, ASML-Brion). Le focus optimal est repéré par le triangle orange et la profondeur de champ par la barre bleue de part et d'autre de cette valeur. L'uDOF est entouré par les pointillés jaunes.

3.2.2.2 Les effets masque

Le masque est aussi un objet fabriqué par lithographie et gravure, et est donc sujet à des variabilités locales. Sa fabrication est contrôlée par des spécifications plus ou moins serrées selon le grade de masque.

Des structures de dimensions identiques au niveau du design auront des CD différents sur le masque. Cette erreur se transférera avec un facteur, le MEEF ou Mask Error Enhancement Factor (facteur d'augmentation de l'erreur masque) sur la plaquette pendant l'exposition. Ce paramètre impacte majoritairement la dose d'exposition.

D'un point de vue purement focus, le masque présente plusieurs sources de variabilité. Tout d'abord, le substrat de quartz, soit la partie du masque laissant passer la lumière, n'est pas uniforme. Sa composition n'est pas identique sur toute sa surface en raison de la fabrication du matériau. Son épaisseur peut varier localement en raison de possibles sur-gravures des motifs du masque lors de sa fabrication. Ces non-uniformités peuvent causer un décalage de phase de la lumière, provoquant un décalage de focus au niveau du wafer.

Pendant l'exposition, le masque est verrouillé sur son support. Comme pour le wafer (cf. Chap. 3.4.1 « Correction de la topographie par le scanner »), cela se traduit par une déformation de l'objet. Le masque subit donc des contraintes mécaniques qui le bombent et le tordent. Le scanner a la capacité de corriger en partie cette déformation du masque via une option nommée RSC (Reticle Shape Correction ou correction de la forme du réticule) qui va mesurer la distorsion du masque et imposer une correction en X et en Y similaire à celle du leveling côté wafer.

Lors de la fabrication du masque, la gravure des motifs dans le matériau opaque (Chrome, MoSi, ... selon la technologie de masque choisie) n'est pas parfaite. Le profil de gravure en particulier influence le front d'onde et peut le décaler localement, provoquant alors un offset local du focus. Le profil 3D du masque (M3D) a toujours eu des influences sur la diffraction mais ce n'est que récemment que ces effets prennent une telle importance. Désormais, il devient de plus en plus nécessaire de prendre les effets M3D en compte. Cela est fait du côté OPC. [23]

3.2.2.3 Les effets du scanner

Le scanner de lithographie lui-même a des limites en termes de contrôle du focus. Son influence va se voir de manière systématique sur tous les lots de production.

Tout d'abord le chuck du scanner n'est pas plat. La correction de topographie réalisée par le scanner présentée en Chap. 3.4.1 corrige en partie cet effet. Cependant, on remarque des différences entre les deux chucks du TWINSCAN, qui peuvent causer des variabilités focus du type « 1 plaque sur 2 ». Le mouvement du chuck pendant l'exposition subit des vibrations provoquées par les moteurs qui permettent ce mouvement. Les dernières générations de machine utilisent des moteurs magnétiques pour éviter au maximum ces vibrations en supprimant toute liaison mécanique entre le chuck et le reste de la machine.

De plus la mesure de topographie par le capteur de niveau n'est pas exacte. En effet, le capteur optique est sujet à des effets d'empilement sur le wafer. Cet effet est décrit plus en détail en Chap. 3.4.1.

Le laser DUV n'est pas parfait. La longueur d'onde n'est pas exactement 193nm mais varie autour de cette valeur selon une gaussienne, nommée la largeur de bande qui est de 300fm. Cette variation est suivie en ligne et constitue un indicateur de la qualité du laser. Une variation légère de cette longueur d'onde va influencer la diffraction au niveau du masque. Comme ce masque est optimisé en termes de matériau et de motifs pour un laser à 193nm, un décalage du front d'onde va avoir lieu causant un défocus sur le wafer [44]. P. Alagna [45] a montré qu'une variation de la largeur de bande du laser de l'ordre de 100fm pouvait causer jusqu'à 30nm de perte de profondeur de champ sur certains motifs. Le contrôle du laser est donc de plus en plus serré [46]. Cet effet peut être vu à toutes les échelles de variabilité.

Les lentilles aussi peuvent avoir une influence sur le focus. Leur échauffement au fur et à mesure de l'exposition va modifier légèrement leurs performances optiques. Cet effet est un effet visible majoritairement sur un même lot entre les premières et dernières plaques exposées. Cet effet est bien décrit dans le papier de Y. Cui [47].

Dans le cadre de ce travail et dans le but de caractériser la part de la variabilité focus originaire de la machine, un test SSF¹³ a été réalisé sur des wafers de silicium vierge en même temps que le test de la multi-wafer FEM expliquée précédemment. Une cartographie de variabilité focus à l'échelle d'un wafer complet est le résultat de ce test.

Dans ce travail, le test a été exécuté dans les conditions décrites dans le tableau 3-3 ci-dessous. L'empilement de lithographie utilisé n'est pas le même que celui nécessaire à l'exposition du Contact en 14nm FD-SOI. Il s'agit d'un empilement optimisé et adapté au masque FOCAL. Les conditions d'illumination sont aussi adaptées au test et ne correspondent pas à l'illumination des wafers du test de la multi-wafer FEM.

Machine	Illumination	NA	Sigma outer	Sigma inner
NXT::1950i	Quasar, DOE 12, Polarisation XY	1.35	0.94	0.79

Corrections scanner	AGILE	Baseliner Focus
Utilisées	Oui	Non

Wafers	Empilement de lithographie
Silicium vierge (x2)	- BARC 80nm - Résine

Tableau 3-3 : Conditions expérimentales du test SSF

Le test a été réalisé avec le capteur AGILE mais sans utiliser les corrections Baseliner Focus¹⁴. C'est-à-dire que cette analyse contient tous les effets issus du scanner en termes de variabilité focus et qu'une part de ceux-ci sont corrigibles. Le test n'a été réalisé que sur un seul des deux chucks. La comparaison de ces erreurs focus avec les mesures de l'« hyper dense focus map » (exposée sur le même chuck, avec AGILE et sans BaseLiner Focus) précédemment présentées offre la possibilité de quantifier la part de variabilité focus originaire de la machine en elle-même.

La figure 3-10 donne les deux cartographies du SSF et de l'HDFM ainsi que la corrélation entre les deux. Environ 50% de la variabilité focus est issue du scanner. La courbe de corrélation montre que la plage de valeurs est la même car la pente est de 1 et que la forme est légèrement différente en raison de la largeur du nuage de point [32]. Cela confirme que la signature du chuck et du wafer stage se répercute sur le budget focus.

¹³ Le masque SSF est constitué de marques spécifiques nommées marques FOCAL que l'on trouve au nombre de 13x19 sur un champ. Après exposition, ces marques sont mesurées sur l'ensemble du wafer dans le scanner par le système d'alignement. Par design, les marques FOCAL traduisent un défocus d'exposition par un désalignement de la marque par rapport à son positionnement théorique sur le wafer.

¹⁴ Le test Baseliner Focus permet de dresser une cartographie de la forme du chuck du scanner afin de contrôler la dérive en focus du scanner et de calibrer régulièrement la machine via une sous recette corrective du positionnement verticale de la plaquette.

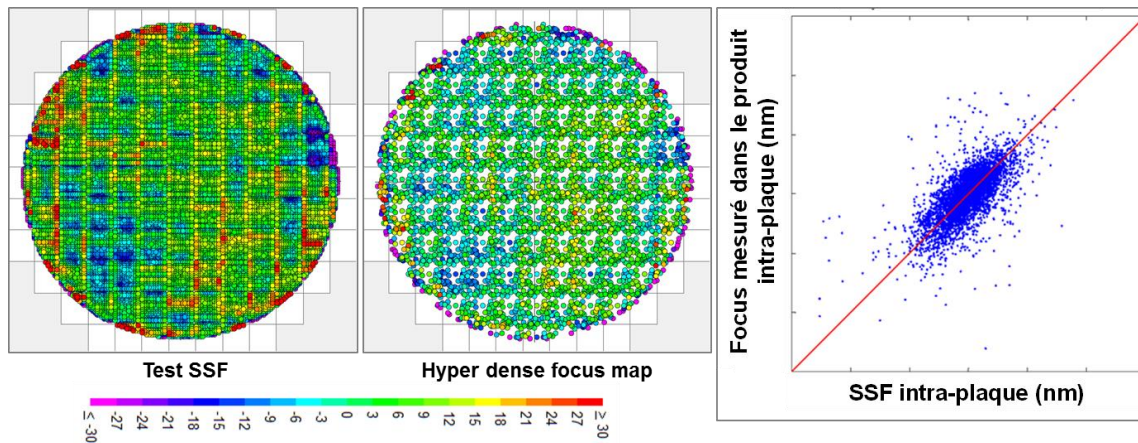


Figure 3-10 : Cartographies intra-plaque du test Single Shot Focal (SSF) à gauche, du focus mesuré dans le produit au milieu, et corrélation point à point entre les deux, à droite.

3.2.2.4 Wafer

Le wafer est une plaque de silicium monocristallin qui fait $775 \pm 25\mu\text{m}$ d'épaisseur et 300mm de diamètre. En raison de sa forme, le wafer présente une courbure dont la cause est l'équilibre mécanique de la plaquette. La couche d'oxyde enterrée des wafers SOI crée un champ de contraintes causant une courbure plus importante que pour des wafers de silicium classiques. Avec la succession de dépôts, gravures, polissages, cette forme va varier car chaque nouveau matériau et nouvelle structuration spatiale du wafer va ajouter ou relâcher des contraintes mécaniques. Le wafer n'est donc pas plat et sa courbure, de l'ordre de $45\mu\text{m}$ d'amplitude (cf. mesures PWG, Chap. 4.1.3), est bien plus grande que la profondeur de champ disponible pour l'impression de l'image aérienne dans la résine. Or une topographie locale va créer un décalage local du focus d'exposition. Cette topographie est corrigée en partie par le scanner pendant l'exposition (cf. Partie 3.4 « Focus et topographie »). Le schéma de la Figure 3-11 montre comment une topographie de surface sur le wafer va décaler le focus local d'exposition. Une topographie positive créant un défocus vers les valeurs négatives et une topographie négative un défocus vers les focus positifs.

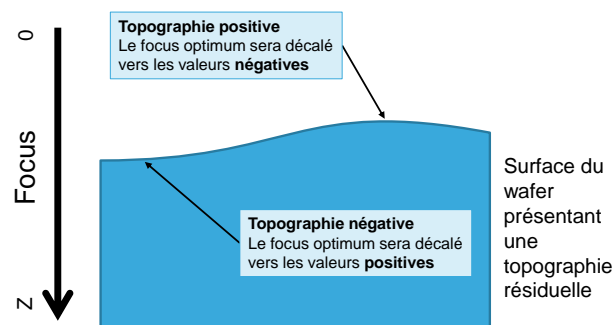


Figure 3-11 : Impact de la topographie du substrat sur la valeur du focus d'exposition sur le wafer.

Le plan focal dans lequel se forme l'image aérienne est situé à la position à laquelle les interférences constructives entre les différents ordres de diffraction permettent la formation d'une image aérienne nette. Les réflexions des rayons lumineux sur le wafer interfèrent aussi avec les rayons incidents,

pouvant alors décaler la position du plan focal et modifier la valeur du focus optimum. Sur un circuit, chaque partie du circuit (logique, RAM¹⁵, composants analogiques) est différente ce qui va créer des variations dans la réflectivité du substrat (par exemple entre l'oxyde de silicium qui est transparent et le cuivre ou le tungstène qui sont eux très réfléchifs). Pour corriger la réflectivité du wafer, on dépose un anti-réfléctif sur la plaquette avant la résine.

Cet anti-réfléctif n'est pas toujours suffisant et il faut parfois tenir compte de l'empilement qui est sous la résine dans les simulations OPC afin de les corriger. Ces effets et les solutions qui existent pour y pallier ont été étudiés par J-C Michel [48] [35].

3.3 LES EFFETS DU PRODUIT

Les effets du masque, les non-uniformités de réflectivité et la topographie de surface sont des effets spatiaux systématiques que l'on retrouve de manière répétée dans tous les champs d'exposition sur le wafer. Ainsi, pour une structure donnée en diverses positions dans le champ ou la puce, le focus optimal d'exposition n'est pas forcément le même. La figure 3-12 ci-dessous donne la variation du focus optimum pour le P90, la tranchée isolée et le HS1 (cf Chap. 3.1.4.1) en BEOL 28nm FD-SOI pour 50 positions dans le champ.

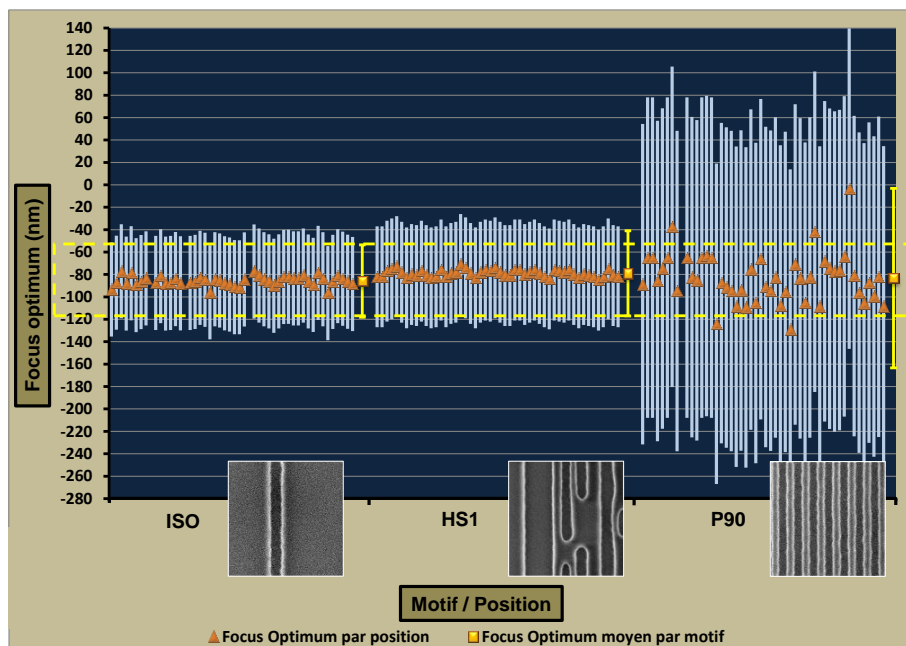


Figure 3-12 : En orange et bleu clair : Focus optimums et profondeur de champ par position dans le champ pour trois motifs en BEOL 28nm FD-SOI

En Jaune : Focus optimums moyens et profondeur de champ utilisable par motif (moyenne sur un champ complet)

Cadre jaune pointillé : uDOF des 150 motifs mesurés

Pour chaque motif et à chaque position, le focus optimum déterminé par FEM est différent mais la profondeur de champ reste la même (85nm pour la tranchée isolée, 90nm pour le HS1 et 285nm pour le

¹⁵ Random Access Memory

P90). On observe une variation 3σ du focus optimum de 12nm pour la ligne isolée, de 9.5nm pour le HS1 et de 65nm pour le P90. Ce dernier étant quasiment insensible au focus, la détermination du focus optimum de ce motif est très bruitée car l'interpolation polynomiale sur les mesures est alors plus dépendante du bruit de la mesure que la variation dimensionnelle du motif avec le focus. Ce décalage de focus optimum entre différentes occurrences du même motif dans le champ conduit à la diminution de la profondeur de champ utilisable (uDOF pour usable DOF) par motif (cf. Chap. 3.1.4.1). Celle-ci est de 65nm pour l'isolé, 75nm pour le HS1 et 160nm pour le P90, soit entre 16 et 44% de réduction par rapport aux DOF respectives de chaque motif. L'uDOF totale de tous les motifs mesurés ici est déterminé comme la DOF commune entre les uDOF de chaque motif. Elle est de 61nm soit 6% de moins que l'uDOF de l'isolé qui était déjà la plus faible de trois.

La modulation des effets précédemment décrits par l'agencement spatiale de la puce électronique fait l'objet de la prochaine partie de ce manuscrit.

3.3.1 Design

Un design est un ensemble de motifs que l'on transfère sur la plaque et qui compose le circuit. La première partie du Chap. 3 de ce manuscrit a montré que des réseaux de lignes denses et des lignes isolées n'avaient pas la même fenêtre de procédé, même si on pouvait les trouver sur le même design. Cependant, un design n'est pas exclusivement composé de réseaux de lignes parallèles, ceux-ci étant généralement utilisés pour le suivi en ligne et le contrôle métrologique en lithographie et en gravure. On trouve de nombreux motifs différents permettant de connecter les transistors entre eux, des antennes, des structures de remplissage pour uniformiser le design, les marques d'alignement pour le scanner et la métrologie, ... L'ensemble de ces motifs présentent des formes variables pouvant être parfois assez compliquées (des U, des zigzags, ...). De même que les réseaux denses et les structures isolées ont des imageries différentes, ces motifs complexes ont eux aussi des conditions optimales d'exposition particulières. Dans certains cas, les conditions optimales sont très différentes du reste du circuit. Ces motifs sont alors nommés « hot spots » (ou points chauds) et sont par nature très critiques en termes d'imagerie. C'est-à-dire que :

- Soit leur point de fonctionnement (*Dose, Focus*) est très différent du reste du design.
- Soit leur fenêtre de procédé (*EL, DOF*) est très réduite.

Pour limiter de tels effets, les OPC ne sont pas toujours suffisants. Les motifs trop critiques sont alors traités par les règles de dessin. Celles-ci définissent les formes autorisées ou non sur un circuit afin d'assurer que la puce reste manufacturable.

Les règles de dessin ne se limitent pas à la lithographie mais permettent une plus grande robustesse du design vis-à-vis des contraintes imposées par les procédés de fabrication.

Dans le cas du Moore Than Moore, certaines options sont ajoutées aux puces (Wifi, RF, co-intégration,... cf. Chap. 1). Entre une partie analogique, une mémoire et le cœur logique de la même puce électronique, les règles de dessin sont les mêmes. Il faut cependant faire attention à ce que ces structures (dont le rôle dans le fonctionnement du circuit impose des tailles, des formes et des densités différentes) n'interfèrent pas négativement sur leur voisinage lors de la fabrication.

3.3.2 Architecture et assemblage

L'architecture de la puce consiste à assembler l'ensemble des blocs fonctionnels du composant (Mémoire, logique, analogique, capteurs, ...) et de les connecter entre eux. Cette étape crée des hétérogénéités de design au sein de la puce elle-même. Elle est réalisée par les designers.

La Figure 3-13 donne un exemple de produit en 28nm et souligne les différents blocs fonctionnels de la puce.

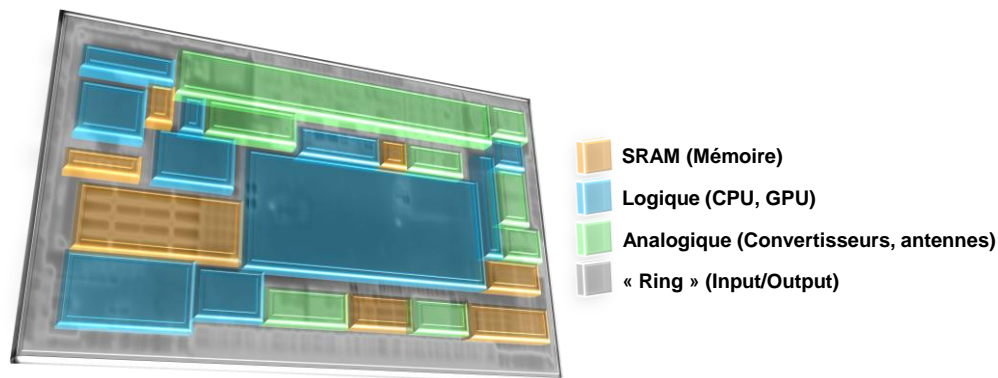


Figure 3-13 : Chaque zone d'une puce possède une fonctionnalité spécifique et un design adapté à cette fonctionnalité.

L'assemblage permet de créer les masques avant corrections RET. Une équipe nommée RAT (Reticle Assembly Team ou Equipe d'assemblage du réticule) est dédiée à cette tâche. Pour chaque produit, il faut mettre le plus de puces possibles sur la surface du champ d'exposition en gardant de l'espace pour les structures non fonctionnelles nécessaires à la fabrication (mires de métrologie, marques d'alignement) qui sont insérées dans les chemins de découpe entre chaque puce. Le cadre est aussi créé à ce moment. Il se trouve autour du champ et contient des mires d'alignement pour le masque, les mires pour RSC (cf. 3.1.4.3), le code-barres du masque,...

Suite à l'assemblage, les OPC sont calculées sur le champ d'exposition et le plan du masque ainsi défini est envoyé au fournisseur de masque (maskshop) qui va le fabriquer. Une pellicule est ajoutée pendant sa fabrication pour protéger le champ de contaminations particulières pendant l'usage du masque.

La Figure 3-14 représente un masque de lithographie.

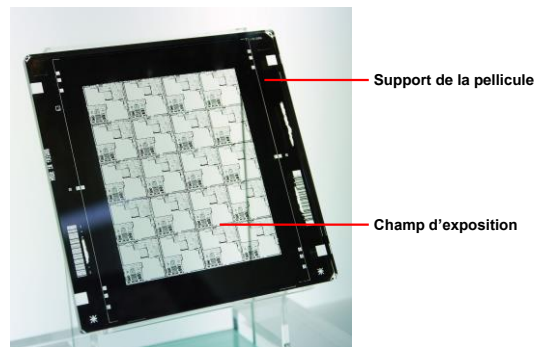


Figure 3-14 : Un masque de photolithographie. La partie centrale contient le design du circuit à transférer sur le wafer entouré du cadre. La pellicule est située au-dessus du design pour le protéger. (Source : Wikipédia, Peellden)

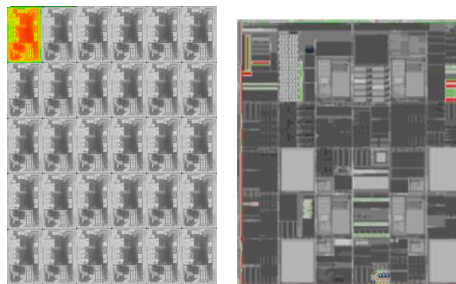


Figure 3-15 : Deux types de produits différents. A gauche, un SLR 28nm FD-SOI et à droite un MPW 14nm FD-SOI (à l'échelle).

On peut distinguer deux grands types de produits d'un point de vue assemblage : les SLR et les MPW, illustrés en Figure 3-15. Le SLR, pour Single Layout Reticle, est un masque contenant un unique produit matricé de nombreuses fois sur le champ. Il est utilisé en production. Le MPW, pour Multi-project Wafer, contient plusieurs produits différents sur le même masque. Il est utilisé en développement ou pour faire du prototypage. L'intérêt du MPW est de diviser les coûts entre plusieurs clients en ne commandant qu'un masque. Le MPW est par construction beaucoup moins uniforme d'un point de vue design qu'un SLR dans le sens où sur un MPW des puces différentes sont assemblées sur le même masque.

3.3.3 Modulation de la réflectivité

Selon les matériaux présents sur le wafer, le focus et la dose d'exposition nécessaires à l'impression d'un motif peut varier. Cela est dû à la réflectivité du substrat. Si d'un niveau de masque à un autre le matériau présent en surface est différent et va impacter le focus et la dose de manière uniforme sur le wafer, dès l'instant où des étapes de structuration spatiale ont lieu (lithographies, gravures, implantations et remplissages), il est possible de distinguer localement plusieurs empilements de matériaux différents en fonction de la position dans la puce. Ces variations locales sont dépendantes du design. On aura par exemple en BEOL une différence notable d'empilement de matériaux entre les lignes métalliques et le diélectrique d'isolation. Ces matériaux différents causent une modulation de la réflectivité du substrat lors de la même exposition.

3.3.4 Modulation de la topographie

Dans le Chap. 3.1.4.4, il a été montré que le wafer n'était pas plat et qu'il présentait une certaine topographie.

La première composante est à l'échelle du wafer et constitue la plus grande partie de cette non-planéité de la plaque en terme d'amplitude soit environ $50\mu\text{m}$. La courbure de la plaque lorsque celle-ci est maintenue sur le chuck du scanner est moindre (1 à $2\mu\text{m}$) et cette composante est corrigeable par le scanner qui peut en effet suivre la surface du wafer à l'échelle millimétrique (cf. Chap. 4.1.1).

La seconde est à l'échelle intra-champ. Elle est issue par exemple du dishing de la plaquette en CMP (polissage de la plaquette, cf. Chap. 1.2.1 « *Intégration* ») mais aussi à des contraintes mécaniques localisées provoquées par la cohabitation de nombreux matériaux différents dans une même puce ou encore aux enchaînements dépôt-gravure que subit le wafer. Cette topographie est fortement corrélée au design du produit pour deux raisons.

La première, c'est que le dishing est dépendant de la taille des motifs et de la répartition spatiale d'un matériau par rapport à un autre. Le dishing consiste en un creusement plus important du substrat pendant les étapes de polissage en certaines zones. Il est issu de la différence de vitesse de gravure de chaque matériau présent sur le wafer au moment du polissage. Celle-ci dépend de la chimie, du type de « slurry », de la pression, de la vitesse de rotation de wafer et des caractéristiques chimiques et mécaniques des matériaux. En BEOL, où la CMP est un procédé utilisé entre chaque niveau de métal, une topographie locale se crée entre les motifs de cuivre et ceux en SiO_2 dont le module d'Young est beaucoup plus élevé.

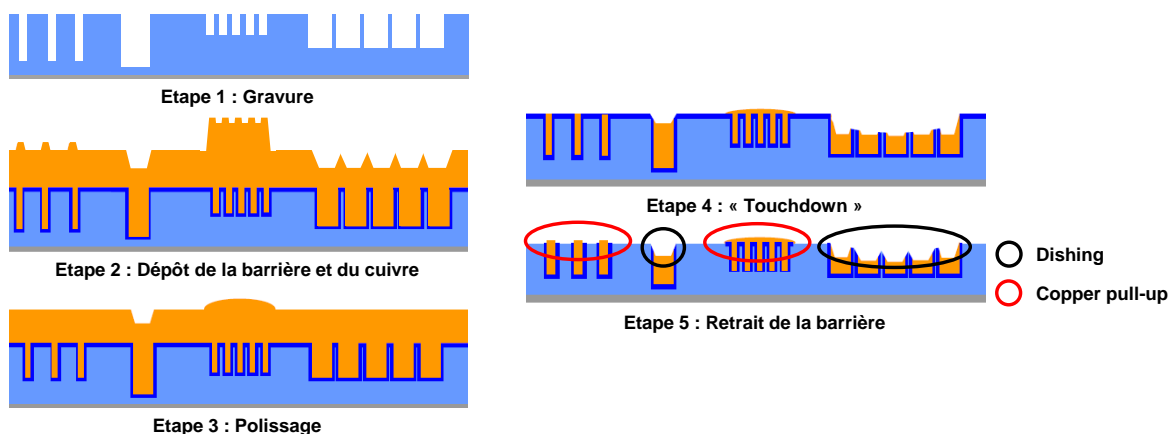


Figure 3-16 : Les étapes de la CMP (Source : Cadence Design Systems)

La figure 3-16 montre l'enchaînement des étapes conduisant à une topographie locale provoquée par le design du circuit. L'exemple correspond à la fabrication de lignes d'interconnexions métalliques en BEOL. On remarque tout d'abord un effet de la densité et de la taille des motifs sur la gravure sur la

profondeur de gravure. Le remplissage des lignes avec du cuivre par dépôt électrochimique présente déjà une topographie certaine. Les trois étapes suivantes de la figure constituent le procédé de CMP.

La seconde raison est le design qui a aussi une forte influence sur la formation de la topographie dans le cas de technologies dérivatives comme les mémoires embarquées. Dans ce cas, c'est la manière dont la technologie elle-même est construite qui a une influence. Les mémoires embarquées font cohabiter sur le même design des zones mémoires flash et des zones logiques. Par construction, la mémoire flash nécessite une double grille alors que la logique n'a besoin que d'une seule, générant ainsi une topographie importante comme illustré par la Figure 3-17.

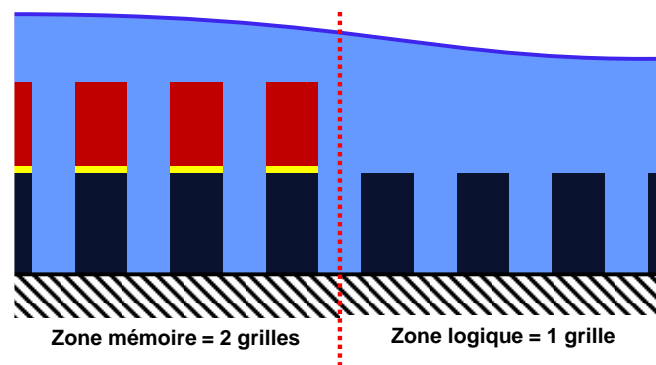


Figure 3-17 : Topographie créée par la cohabitation sur un même design d'une mémoire flash et de circuits logiques (schéma simplifié)

Les challenges de contrôle de l'impact de la topographie sur le focus sont discutés et mis en évidence dans la suite.

3.4 FOCUS VS. TOPOGRAPHIE

3.4.1 Correction de la topographie pendant l'exposition

La problématique de la topologie du wafer n'est pas récente [49] et des systèmes de plus en plus performants ont été mis en place pour la corriger pendant l'exposition. On appelle ce système le « leveling » (mise à niveau) [50]. Pour ce faire, le wafer est en mouvement pendant l'exposition. Le « leveling » consiste à amener mécaniquement la surface du wafer le plus proche possible du plan focal pour que chaque point de la plaquette dans la slit d'exposition se trouve à l'intérieur de la profondeur de champ. Avec la réduction des dimensions, la profondeur de champ disponible a été fortement diminuée et le maintien de la surface du wafer dans le plan image est devenu de plus en plus critique et complexe. Le passage des mask aligners (machines qui permettaient l'exposition d'un wafer entier avec un masque de taille identique) aux steppers step-and-repeat (qui expose chaque champ indépendamment) puis aux scanners (qui scannent le wafer champ par champ) a permis un meilleur suivi de la surface du wafer pour garder le plan image aligné dans la résine. Les solutions techniques permettant un « leveling » performant sont détaillées dans les Figures 3-18 et 3-19.

La première solution technique qui a permis de réduire très fortement la topographie du wafer est le chuck de la machine d'exposition. Celui-ci maintient le wafer par aspiration ce qui le colle à plat sur son support.

Le mask aligner permet de pivoter le wafer autour des axes (x, y) et de le positionner en z . En revanche, comme le wafer est exposé en une seule fois (cf. Partie II « *La photolithographie* »), il n'est pas possible de corriger le positionnement vertical du wafer puce par puce.

Le stepper « Step-and-repeat » permet quant à lui de positionner au mieux le wafer champ par champ en le pivotant autour des axes x et y et en le montant ou le descendant selon l'axe z . Il est donc possible de corriger beaucoup plus la topologie du wafer et en particulier en suivant la forme globale bombée du wafer. Cependant, lorsque les profondeurs de champ deviennent plus faibles, le « leveling » intra-champ devient à son tour très critique. C'est l'une des raisons qui a poussé au développement des scanners « Step-and-Scan ».

Lors de l'exposition sur un scanner, le scan du champ est réalisé selon l'axe y . Selon x , le wafer ne peut encore une fois qu'être pivoté autour de l'axe y car il ne peut être plié, autorisant une correction au premier ordre. Grâce au scan, ce pivot est dynamique et va donc pouvoir être modifié pour suivre au mieux la topologie du champ exposé. Dans l'autre direction, selon y , le scan autorise un suivi plus précis de la surface du wafer. Le chuck va pouvoir monter, descendre et pivoter autour de l'axe x afin de corriger les bosses et les vallées que présente le wafer selon un polynôme d'ordre élevé.

Solution technique	Amélioration des machines				Amplitude des résiduels	% en focus (DOF = 60nm)
	Wafer	Wafer aligners	Wafer steppers	Scanners		
Wafer posé à plat	✓	✗	✗	✗	$> 50\mu m$	$< 2\%$
Maintien par vide	✗	✓	✓	✓	$0.5 \text{ à } 2\mu m$	$\cong 15\%$
Correction à l'échelle du wafer du 1 ^{er} ordre	✗	✓	✓	✓	$\cong 0,3\mu m$	$\cong 90\%$
Correction à l'échelle du wafer par un polynôme d'ordre élevé et du champ au 1 ^{er} ordre	✗	✗	✓	✓	$0.1 \text{ à } 0.2\mu m$	$> 90\%$
Correction à l'échelle du wafer et du champ par un polynôme d'ordre élevé	✗	✗	✗	✓	$\leq 0.1\mu m \text{ iW}$ $\leq 0.04\mu m \text{ iF}$	$> 95\%$

Figure 3-18 : Capacités de correction de topographie pendant l'exposition en fonction du type de machine

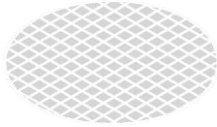
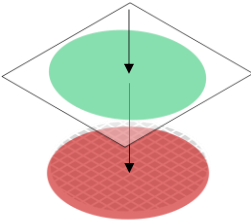
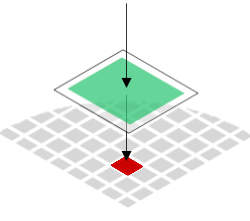
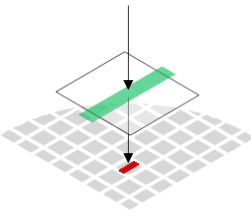
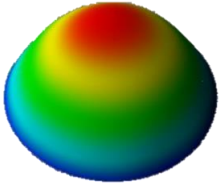
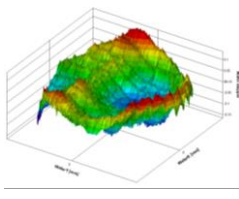
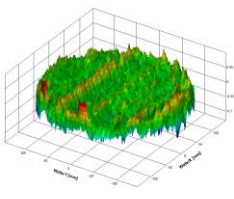
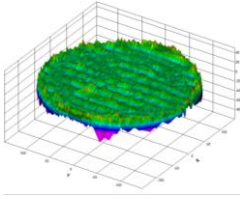
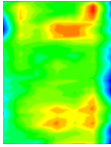
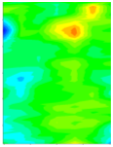
Machine	WAFER	WAFER ALIGNER	STEPPER	SCANNER
Type d'exposition				
Surface de correction	Aucune correction	Wafer complet $300^2 \cdot \pi \approx 283\ 000\ \text{mm}^2$	Champ complet $20 \times 20 = 400\ \text{mm}^2$	Taille de la slit $6 \times 26 = 156\ \text{mm}^2$
Résiduels Intra-wafer				
Résiduels Intra-champ	Aucune correction	Aucune correction		

Figure 3-19 : Comparaison des modes de correction de la topographie par le leveling en fonction de la machine. La zone rouge visible dans la première ligne du tableau est la surface sur laquelle le contrôle du focus est appliqué à chaque instant de l'exposition.

Une partie de cette topographie n'est pas corrigable. En effet, si les variations de topologie de surface sont trop importantes, le mouvement mécanique ne permettra pas de suivre parfaitement la surface. Ce résiduel de correction de topographie est appelé NCE pour Non-Correctable Error ou erreur non-corrigeable.

Le scanner génère un très grand nombre de données par wafer lors de la métrologie interne et l'exposition de la plaquette, soit environ 40 à 45MB de données par wafer. En configuration classique, ces fichiers ne sont pas enregistrés car ils surchargent très rapidement la mémoire interne de la machine. Dans ce travail, le scanner a été systématiquement passé en mode « Extended Data Collection » (collection de données étendue) pour permettre l'enregistrement des fichiers journaux. Cette option nous intéresse particulièrement car elle contient une option « Extended Leveling » qui enregistre les données de la mesure et de la correction de la topographie par le scanner sur la plaquette sans arrondis des valeurs. La suite du paragraphe décrit le fonctionnement du système de leveling du scanner 193nm à immersion ASML TWINSKAN NXT::1950i utilisée pour l'ensemble de l'étude. Les cartographies de wafer de la Figure 3-19 sont extraites de l'exposition d'un lot de production de référence de 25 plaquettes en Contact 14nm FD-SOI.

Pour pouvoir réaliser sa correction de topographie, le scanner possède deux capteurs de niveau.

Le premier est un capteur optique nommé « Level Sensor » (capteur de niveau, noté LS). Ce capteur fonctionne par interférométrie par triangulation et est composé de 7 spots lumineux de 2.8mm par

2.5mm qui balayent l'ensemble du wafer. En raison de la longueur d'onde utilisée par le LS qui se situe dans le domaine du visible, une erreur de lecture de la topographie de surface apparait. Cette erreur est due à l'effet Goos – Hänchen [33] [51]. Il s'agit d'un décalage latéral de la lumière lors de la réflexion de celle-ci par la surface. Ce décalage dépend du matériau et de l'empilement présent sur le substrat. La lumière visible n'est que très peu absorbée par les matériaux du substrat et se reflète sur des couches de matériaux structurés situées sous l'empilement de lithographie. La surface paraît alors se situer plus basse qu'elle ne l'est réellement. Ce phénomène porte le nom d'ASD pour « Apparent Surface Depression » ou dépression de la surface apparente et est illustré par la Figure 3-20. Si la topographie mesurée n'est pas correcte, la correction de celle-ci par le positionnement de la plaquette dans le plan image du masque ne correspond pas à ce qui doit effectivement être corrigé. Soit on sur-corrigea certaines zones qui n'en aurait pas besoin soit au contraire on négligera un autre partie de la puce qui mériterait de l'être.

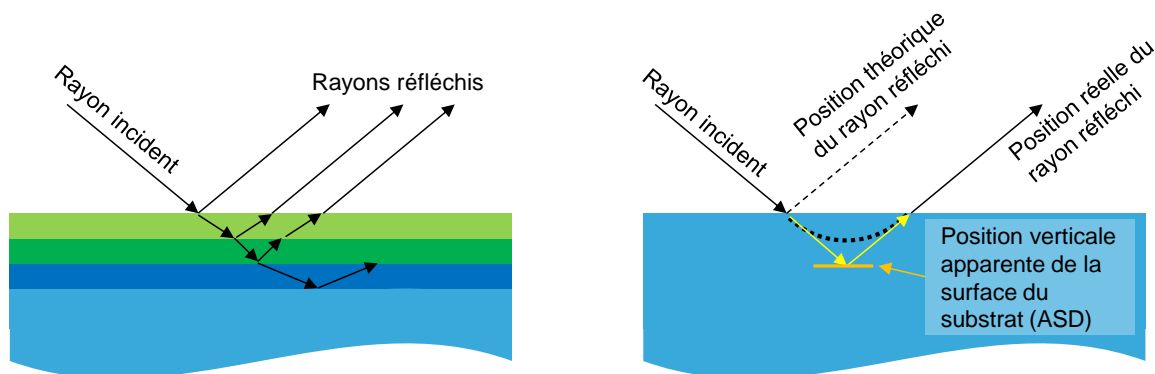


Figure 3-20 : Illustration des causes de l'erreur de mesure du capteur optique de niveau. A gauche, la réflexion de la lumière incidente est perturbée par l'empilement sur le wafer. A droite, l'effet Goos-Hänchen cause un décalage de la lumière réfléchie par rapport à son point d'incidence. Sur le wafer, les deux effets s'ajoutent.

Pour s'assurer que la mesure donne une valeur correcte de la topographie du wafer, un second capteur pneumatique (non impacté par ces effets), nommé AGILE [52] pour « Air Gauge Improved Leveling » (ou mise à niveau améliorée par capteur pneumatique), a été installé dans les scanners de lithographie ASML. AGILE mesure une différence de pression entre deux jets d'air : une référence et le capteur. La dépression mesurée est liée à la distance verticale entre la plaque de référence et le wafer. Ce capteur mécanique n'est pas du tout impacté par l'empilement du substrat mais la mesure est beaucoup plus lente que celle du capteur optique. Pour des raisons de débit de plaquettes exposées, il n'est donc pas possible de substituer le capteur optique par le capteur pneumatique. AGILE est utilisé pour corriger la valeur de la mesure du LS. La mesure se passe en deux étapes simultanées. Le LS mesure l'intégralité de la plaque alors que AGILE mesure 9 champs sur la centaine que comporte un wafer. La partie intra-champ de la mesure du LS est extraite et la différence entre la moyenne des mesures intra-champ du LS et la moyenne des mesures AGILE est calculée. Cette grandeur, appelée PDO pour « Process Dependency Offset », ou décalage issu de la dépendance au procédé, est ensuite appliquée comme correctif à chaque champ du wafer. C'est à partir de cette cartographie corrigée de la topographie du

wafer que le scanner calcule le mouvement du chuck permettant de positionner la plaquette au mieux pendant l'exposition.

$$LS = \text{Topographie réelle} + \text{Dépendance au procédé} \quad (11)$$

$$LS(\text{AGILE}_{corrected}) = LS - PDO \approx \text{Topographie réelle} \quad (12)$$

Une autre manière de s'affranchir des effets de réflectivité d'empilement consiste à modifier la méthode de mesure avec le LS, ce qui est le cas pour les dernières machines de lithographie qui utilisent un laser UV moins sensible à l'empilement de matériaux du substrat et un angle d'incidence plus élevé pour mesurer la topographie de surface.

3.4.2 Décalage du focus d'exposition provoqué par la topographie non corrigéable

Dans le scanner, le focus est une mesure de la position verticale du wafer par rapport à une distance idéale du réticule au chuck, calibré par des tests réguliers du suivi de dérive de la machine (le test BaseLiner® dont il a été question dans le Chap. 3.2.2). La valeur du focus est plus grande lorsque la distance entre le réticule et le wafer augmente, c'est-à-dire quand la surface du wafer est située plus bas dans le référentiel du scanner. Suite au leveling de la plaquette par le scanner, le résiduel non-corrigeable (cf. paragraphe précédent) crée un défocus local qui dépend du design macroscopique de la puce.

La suite de cette partie présente la comparaison entre des mesures de focus sur un réseau dense de chaînes de contacts en 14nm FD-SOI sur un wafer avec topographie – c'est-à-dire un wafer ayant subi l'ensemble des étapes de fabrication –, un wafer sans topographie – c'est-à-dire un wafer de silicium vierge – et les mesures de cette topographie [36].

Pour cette expérience, les deux wafers ont été exposés selon un protocole appelé le « Focus Meander » (FMEAND, ou méandre de focus). Contrairement à la FEM qui va croiser les conditions de dose et en focus, la FMEAND ne permet d'exposer que des pas de focus. Chaque champ est ainsi exposé à une valeur de focus différente selon le schéma de la Figure 3-21. Cela permet de couvrir pendant l'expérience un plus grand nombre de conditions de focus que pour une FEM en faisant des pas plus petits. Ici, les pas sont de 10nm et couvrent 220nm de focus.

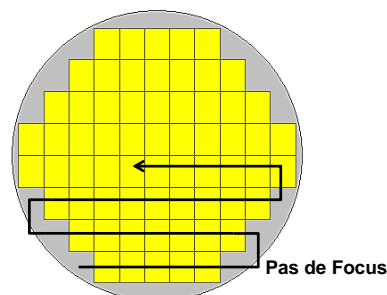


Figure 3-21 : L'exposition en « focus meander »

La figure 3-22 donne les conditions de l'analyse. Un bloc en particulier sur le produit de développement du 14nm FD-SOI présente un très fort dishing d'environ 45nm de profondeur sur une surface de 1.4mm par 0.8mm. Les dimensions du bloc ne permettent pas une correction efficace de la topographie locale par le scanner car sa taille est trop petite par rapport à la profondeur du dishing. Remarquons qu'un tel dishing est non usuel et est ici uniquement dû à une structure de test nécessaire pour le développement de la technologie et non d'une structure que l'on trouvera ensuite dans un produit commercialisé. Cependant, il offre un exemple parfait pour mettre en évidence les effets de la topographie. Ce bloc consiste en une très longue chaîne de contacts en maillon à une forte densité et surtout tous absolument identiques. Une simulation optique – qui n'a pas été faite – donnerait un focus optimum identique pour l'intégralité du bloc, en raison de la régularité du design. Ainsi, si on néglige l'influence de la non-uniformité du masque que l'on peut considérer comme très faible au vue la dimension spatiale du bloc, tout décalage de focus possiblement mesuré devrait intégralement être causé par la topologie locale.

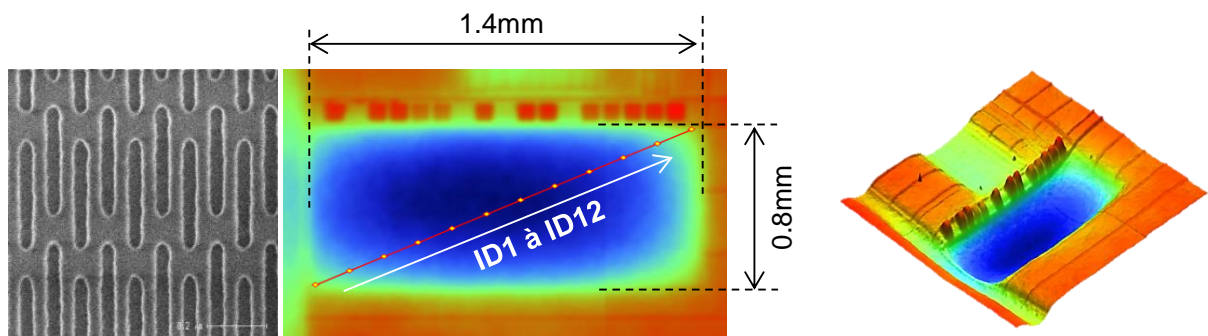


Figure 3-22 : A gauche, image SEM du motif mesuré sur FEM le CDSEM. Au milieu, la position des 12 occurrences de la structure mesurée dans le bloc étudiée. A droite, une vue en 3D de la topographie mesurée sur le bloc (mesures Wyko, cf. Chap. 4.1). Les hauteurs dans la zone intéressantes (en bleu et vert) varient entre -15nm et -50nm.

Les graphiques de la figure 3-23 présente les mesures de focus optimum de la structure en chainons réalisées dans le bloc présentant un fort « dishing ». L'erreur de détermination du focus est de $\pm 3\text{nm } 3\sigma$ et pour la mesure de topographie, elle est de $\pm 2.5\text{nm } 3\sigma$.

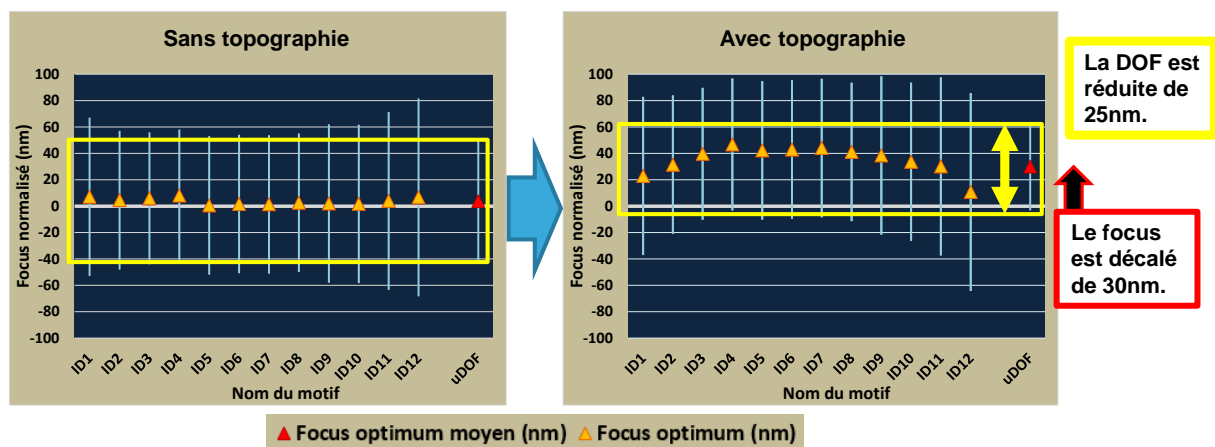


Figure 3-23 : L'impact de la topographie locale non corrigée par le scanner sur le focus optimum mesuré sur silicium. (DOF = Depth Of Focus c.à.d. profondeur de champ et uDOF = Usable DOF i.e. profondeur de champ effective)

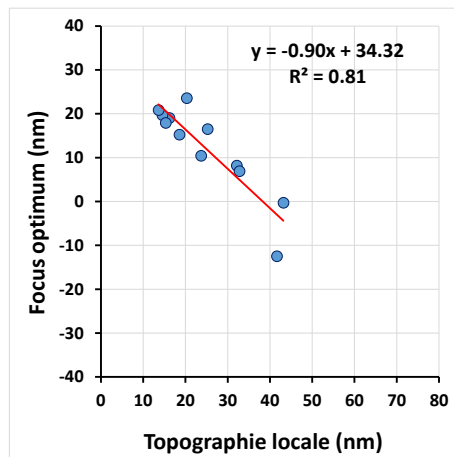


Figure 3-24 : Corrélation entre la topographie et l'écart au focus optimal pour un unique motif en 14nm FD-SOI

En traçant le focus optimum obtenu sur la plaque avec topographie en fonction de cette topographie locale (Figure 3-24), on obtient une droite de pente proche de -1 avec un très bon coefficient de corrélation.

La pente de -1 s'explique par la manière dont le focus est exprimé dans le scanner qui fait que le focus diminue en valeur absolue quand on s'approche du masque, par exemple au niveau d'une topographie positive.

En ajoutant les focus optimums d'autres motifs situés dans des zones de topographie différentes, il est possible d'améliorer la corrélation (cf. Figure 3-25). Les motifs ont été choisis très proches en design et en densité pour considérer leur image aérienne très proche de celles des maillons précédents et leur position dans la puce n'est pas très éloignée de celle des maillons pour minimiser les effets de CD du masque. Ainsi la majeure partie des variations de focus sur ces motifs est due à la topographie locale.

On peut voir sur la figure 3-25 que les mesures de focus optimum réalisées sur les différents motifs correspondent à des positions dans le champ pour lesquelles la topographie locale varie avec une amplitude de 60nm environ ce qui cause 60nm de décalage de focus optimum pour les motifs mesurés. Une grande partie de ces mesures a été réalisée sur des blocs atypiques du produit MPW du 14FD-SOI et la distribution de topographie pour les hauteurs inférieures à 40nm correspond à environ 1.7% du champ. Aussi une telle amplitude de variation de focus n'est pas attendue sur un produit manufacturé de type SLR. Si on ne prend en compte que les parties du champ dont le design correspond à une puce CMOS, la topographie ne varie qu'entre 40 et 75nm de hauteur, ce qui correspond à 97% de la surface du champ. Le focus va en revanche pouvoir varier de 25nm en raison de la topographie dans les zones dans lesquelles le design n'est pas atypique pour un même motif.

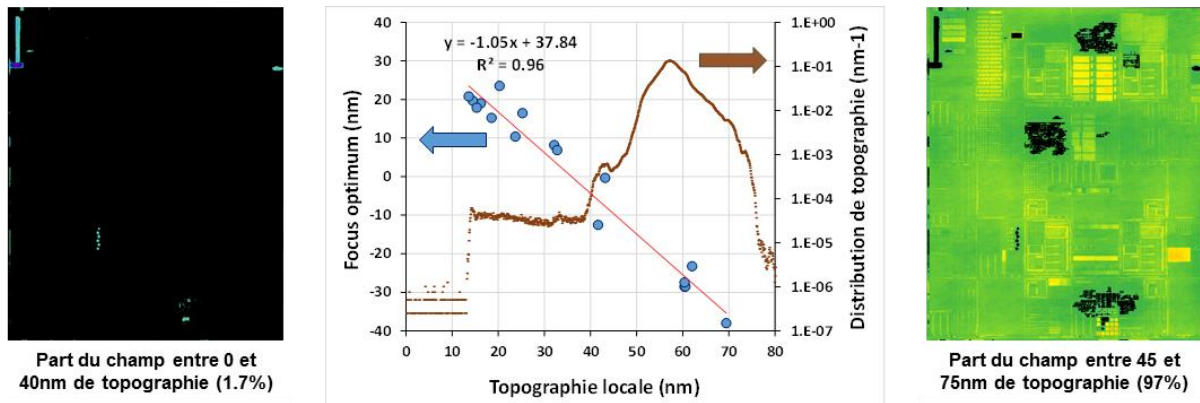


Figure 3-25 : Corrélation entre la topographie et l'écart au focus optimal pour plusieurs motifs en 14nm FD-SOI (en bleu et vert) et comparaison avec la distribution de topographie dans le champ.

Lors de l'étude des « hot spots » sur le 28nm BEOL (cf. Chap. 3.2.2.1), plusieurs métriques ont été mises en parallèles. Tout d'abord, les simulations LMC ont permis la détermination des focus optimums de 7 motifs différents à partir de la formation de l'image aérienne avec un modèle de masque 3D. Coté silicium, des mesures ont été faites avec la méthode de l'« hyper dense focus map » pour déterminer les focus optimums réellement observés sur plaquette à l'échelle d'un wafer complet sur les mêmes motifs. Enfin le champ a été mesuré avec le Wyko et les valeurs locales de la topographie aux positions des motifs mesurés avec le CDSEM ont été extraites.

La corrélation entre le focus obtenu par simulation et par mesure CDSEM est très mauvaise. Ainsi, malgré le décalage de focus induit par les effets de masque 3D, le silicium ne répond pas du tout de la même manière (cf. Figure 3-26 graphique de gauche).

En comparant la topographie locale avec l'écart entre la valeur attendue et la valeur mesurée de focus, la corrélation est très bonne, ce qui montre bien que le décalage supplémentaire observé est bien dû à la topologie de la plaquette (cf. Figure 3-26 graphique de droite). De plus la courbe obtenue contient les informations de focus de plusieurs motifs différents ce qui montre que la topographie a un effet général qui s'applique à tout le champ indépendamment de l'image aérienne, ce qui est logique puisqu'elle ne dépend que de l'état du wafer à l'instant de l'exposition.

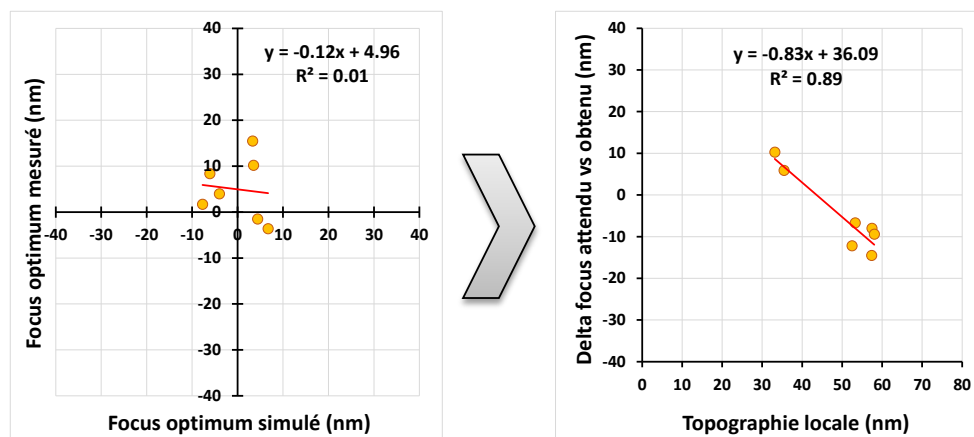


Figure 3-26 : Corrélation entre la topographie et l'écart au focus optimal pour plusieurs motifs critiques présentant un décalage de focus optique important en 28nm

Enfin, en mettant sur le même graph les focus optimums indépendamment de la nature du motif pour le 14nm FD-SOI au niveau Contact et pour le 28nm BEOL en fonction de la topographie mesurée à la même position, on obtient de nouveau une droite de pente -1 avec un $R^2 > 0.9$. Cela corrobore l'effet de la topographie, qui ne dépend donc que du wafer et des capacités de correction du scanner et non du motif (8 motifs différents), du niveau de masque (Contact et BEOL), de la technologie (14nm FD-SOI et 28nm), de l'intégration (empilement de lithographie tri-couche pour le Contact 14FD-SOI et empilement BARC + Résine pour le BEOL 28nm), du type de résine (négative en 14FD-SOI et positive en 28nm).

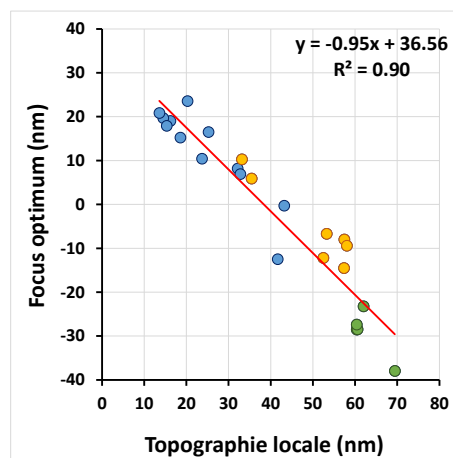


Figure 3-27 : Corrélation entre la topographie et l'écart au focus optimal pour plusieurs motifs dans différentes technologies (intégrations, dimensions et procédés différents).

3.5 CONSEQUENCES SUR LA FENETRE DE PROCEDURE

Sur un produit manufacturé, les effets masque, scanner, topographie, intégration s'accumulent à la fois au niveau intra-champ et inter-champ. La courbe de Bossung d'un motif permet de déterminer sa fenêtre de procédé (cf. Chap. 2). Cette fenêtre est elliptique en raison de l'influence conjuguée de la dose et du focus. Il est possible de la calculer en faisant des simulations optiques de la formation de l'image dans la résine. Il est aussi possible de remonter à celle-ci via des mesures sur Silicium. L'analyse de la fenêtre de procédé offre la possibilité d'en extraire la densité de probabilité de défaut du motif en fonction du focus. Ici un défaut est défini comme la présence d'un pont de résine parasite ou le pincement d'une ligne de résine, c'est-à-dire les variations extrêmes du dimensionnel. La présence de résidus de résine au fond de la tranchée est aussi prise comme un défaut d'impression. Malgré tout, de manière général, la fenêtre de procédé est avant tout un problème de contrôle dimensionnel. C'est-à-dire que l'on peut avoir une ligne de résine sans défaut pour des conditions de procédé données sans que celle-ci se soit imprimée à la dimension désirée. Lesdites conditions sont alors hors de la fenêtre de procédé. Quelques exemples de défauts sont donnés dans le tableau 3-4 :

Motif	Exposition hors focus Défocal négatif	Exposition au focus optimum	Exposition hors focus Défocal positif
HS1			
HS3			
ISO			
P90			

Tableau 3-4 : Images SEM de quatre motifs en conditions optimales de procédé et hors focus

Les graphiques des Figures 3-28 et 3-29 ont été tracés à partir des mesures de la FEM présentée en Chap. 2.5 « La fenêtre de procédé » et des mesures réalisées sur la multi-wafer FEM du Chap. 3.1 « Le budget focus ». La combinaison de ces deux densités de probabilité de défauts en fonction de la dose et du focus respectivement montre bien la forme elliptique de la fenêtre de procédé. Les deux graphs ci-dessus sont des mesures et ainsi ne contiennent pas que la variabilité du motif avec la dose et le focus.

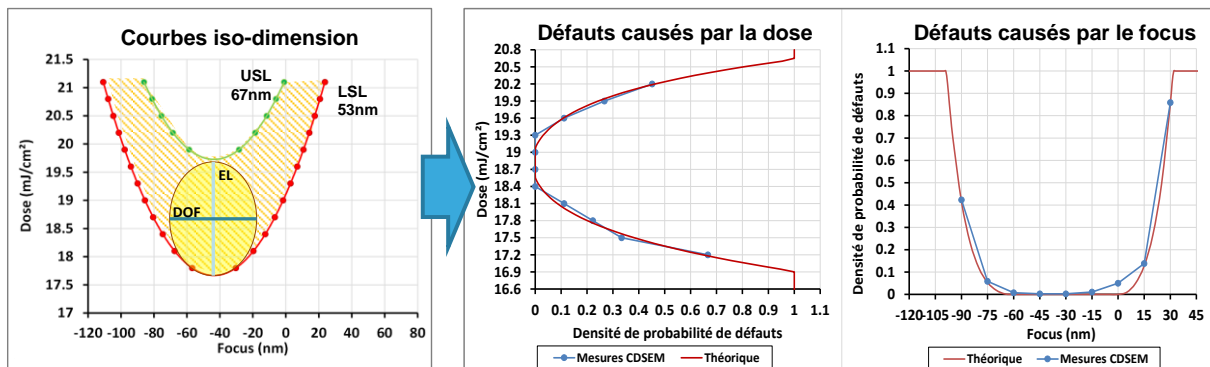


Figure 3-28 : Densité de défauts en fonction de la dose et du focus pour un motif isolée en 28nm BEOL extraits à partir de la fenêtre de procédé du motif

Les impacts de la topographie et de l'imagerie sur l'impression dans la résine du motif P1 en chacune de ses occurrences dans la puce sont une convolution de la probabilité d'occurrence de défauts dans l'image aérienne, de la topographie de surface du wafer à la position à laquelle se trouve le motif et de l'environnement de ce motif. Cette convolution est en elle-même une approche holistique de la fenêtre de procédé car elle permet de considérer des mécanismes de variabilité différents.

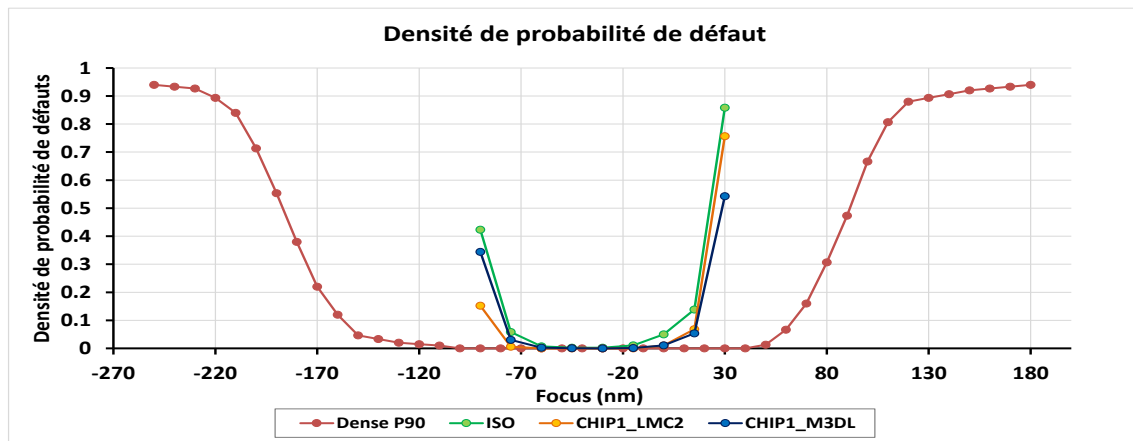


Figure 3-29 : Densité de défauts par wafer en fonction du focus pour 4 motifs plus ou moins critiques en 28nm FD-SOI.

La convolution permettant de calculer les distributions de défauts de manière théorique est donnée ci-dessous :

$$\rho(\text{défaut}) = \rho(\text{défaut de P1 imagerie}) * \rho(\text{Topographie au niveau de P1}) \quad (13)$$

$$= f(\text{focus}) * \text{Topographie}_{\text{locale}}(\text{position}_i) \quad (14)$$

$$= f(\text{focus}, \text{Topographie}_{\text{locale}}) \quad (15)$$

En ajoutant la convolution avec les occurrences spatiales du motif P1, il est alors possible d'évaluer la densité de probabilité de défauts du motif P1 pour un produit, une intégration et un niveau de masque donnés. En rajoutant les cartes de focus optimum et de profondeur de champ calculées sur la puce complète par simulation LMC (Lithographic Manufacturability Check ou Vérification de Faisabilité Lithographique), on obtient la carte de prédiction de défaut lithographique intra-champ. Il est possible d'étendre cette prédiction à un wafer complet en ajoutant une convolution avec la carte d'uniformité focus inter-champ. Celle-ci est obtenue par la méthode de l'Hyper Dense Focus Map décrite dans la Chap. 3.1.

Il est possible de faire de même pour les variations de dose. La fenêtre de procédé s'obtient en convolant les deux pour obtenir un graphique représentant $EL = f(\text{Focus})$. La fenêtre de procédé est représentée soit à l'aide de ce graphique soit de nouveau sous la forme d'une courbe de Bossung, c'est-à-dire avec un graphique Focus-Dose-CD.

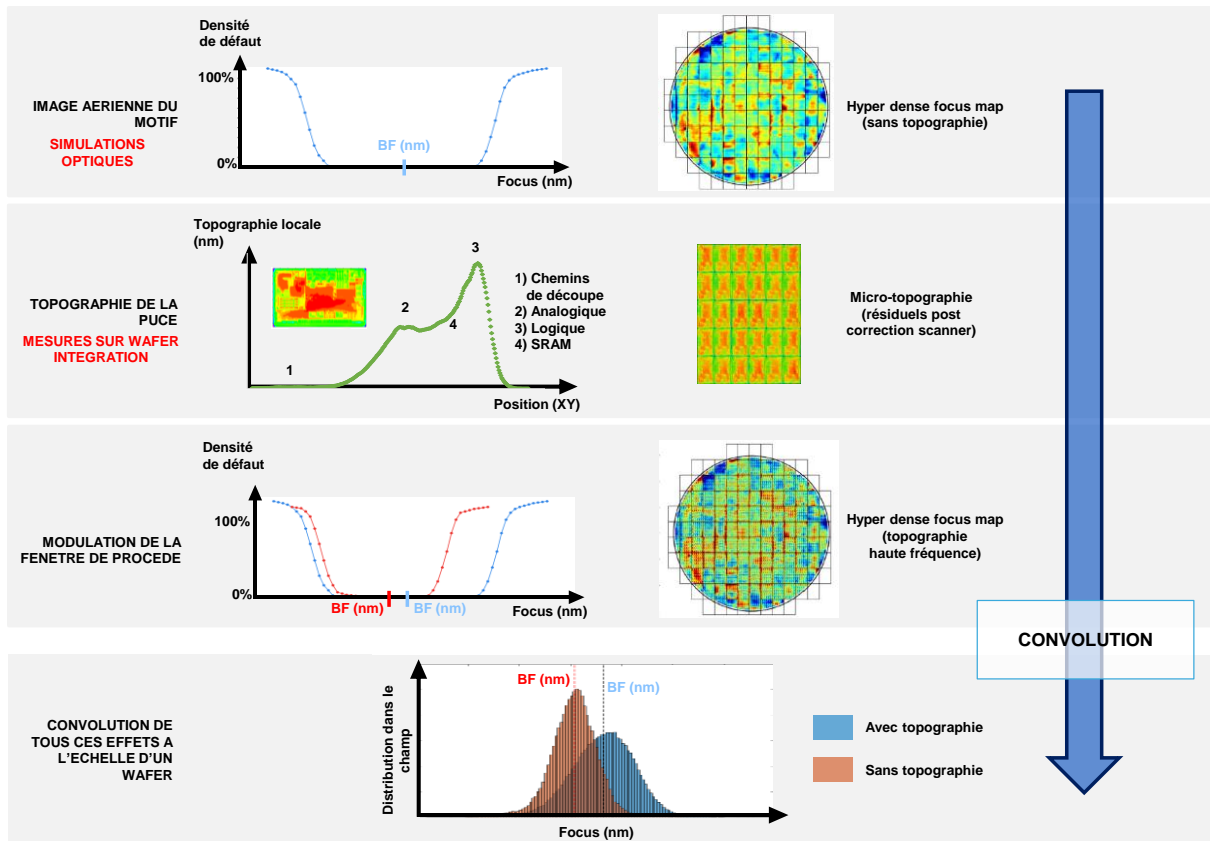


Figure 3-30 : Représentation graphique de la convolution de la fenêtre de procédé théorique de PI (Image aérienne), de ses occurrences dans le produit (design) et de la topographie local (mesures).

3.6 CONCLUSION

Le chapitre 3 a présenté les sources et les mécanismes à l'origine de la variabilité du focus d'exposition pendant le procédé photolithographique. Que ce soit le wafer, le scanner, le masque, l'intégration ou le design, des effets topologiques et optiques s'additionnent et impactent le focus.

Parmi ces sources de variabilités, les effets du produit en lui-même ont été étudiés plus en détail. Le design et l'architecture de la puce créent des hétérogénéités au sein même de celle-ci. A l'échelle d'un champ d'exposition, l'assemblage du masque cause des différences similaires.

Ces hétérogénéités spatiales sont la source de modulations de l'image aérienne et des propriétés physiques du wafer. Chaque partie de la puce a une fonctionnalité propre nécessitant un design spécifique. Un motif correspondant à une image aérienne particulière, ces différenciations de design causent des non-uniformités de l'image aérienne, perturbant le focus.

La réflectivité du substrat suit la répartition spatiale des matériaux sur le wafer. L'impact est optique, modifiant l'image aérienne du masque en fonction de l'environnement local de la position du motif sur le wafer.

Enfin, l'agencement spatial des matériaux dans la puce module la topographie locale du wafer. La part non-corrigeable par le scanner de cette topologie de surface cause un décalage local de la valeur du focus d'exposition égal la valeur de cette topographie, et cela indépendamment de l'image aérienne du motif. A l'échelle des capacités de mesure et de correction du scanner (c'est-à-dire à l'échelle millimétrique), la topographie intra-champ est évalué comme responsable de près de 50% des erreurs de focus au sein d'une puce (soit environ $20\text{nm } 3\sigma$) mais la partie haute fréquence de la topographie est beaucoup plus importante.

Caractériser cette topographie à plusieurs échelles de résolution spatiale est donc une nécessité pour rendre compte de l'impact réel de ce paramètre sur la qualité de la lithographie. A partir de cette analyse, deux pistes sont envisageables : la prédiction (cf. Chap. 4) et la correction (cf. Chap. 5).

CHAPITRE 4

4 TOPOGRAPHIE INTRA-CHAMP ET MODELISATION

4.1 LES MESURES DE TOPOGRAPHIE

Dans le Chap. 3.2, nous avons décrit comment une topographie de surface pouvait être créée sur le wafer. Cette topographie est dépendante de la répartition de la matière sur la puce et donc par extension du design même de la puce. Du fait de cette dépendance, il est envisageable de modéliser et prédire la topologie locale du wafer à partir du design du circuit.

Les mesures de topographie sur produit nécessaires pour la calibration du modèle ont été faites sur plusieurs équipements de métrologie différents. Ces mesures ont permis de mettre en évidence la topographie intra-champ à plusieurs échelles spatiales. Les fréquences d'échantillonnage vont du micron au millimètre selon les outils de mesure. L'analyse des résultats, en particulier les images et les études de distribution de la topographie a été réalisée avec le logiciel Gwyddion [53].

4.1.1 Leveling

Le leveling consiste en :

- une mesure de la topographie de surface du wafer avant l'exposition dans le scanner
- la correction mécanique de cette topographie via le calcul d'un mouvement optimisé du chuck afin d'amener le wafer dans le plan image du masque à chaque instant de l'exposition.

Le scanner mesure la topographie à une échelle latérale millimétrique et ne peut la corriger qu'à cette échelle. La manière dont cette mesure et la correction sont réalisées dans le scanner est décrite en Chap. 3.4.1 « *Focus vs. Topographie – Correction de la topographie pendant l'exposition* ».

Pour simuler de manière prédictive et robuste la topographie mesurée par le scanner, il est important de s'assurer de la stabilité de cette mesure dans le temps. Pour cela, deux méthodes sont possibles. Tout d'abord, les spécifications d'ASML en termes de performances et de stabilité de la mesure de topographie par les capteurs optique et pneumatique peuvent être utilisées. Le budget focus (cf. Figure 3-1) donne une variation 3σ de 5nm et de 10nm pour le capteur AGILE et pour la dépendance à l'empilement du capteur optique respectivement. Ces valeurs sont calculées comme la variation maximale à laquelle on peut s'attendre à l'échelle d'un wafer complet. La deuxième méthode consiste

à utiliser directement les données intra-champ de leveling de lots de production sur la machine. Les fichiers journaux de plusieurs lots de production ont été récupérés pendant une période de temps de plusieurs semaines. Les mesures sont très similaires d'un lot à l'autre (de l'ordre de $5\text{nm } 3\sigma$), au sein d'un même lot (entre 3 à $5\text{nm } 3\sigma$) et d'un champ à l'autre sur une même plaquette (inférieure à $2.5\text{nm } 3\sigma$).

La Figure 4-1 rappelle le fonctionnement du « leveling ».

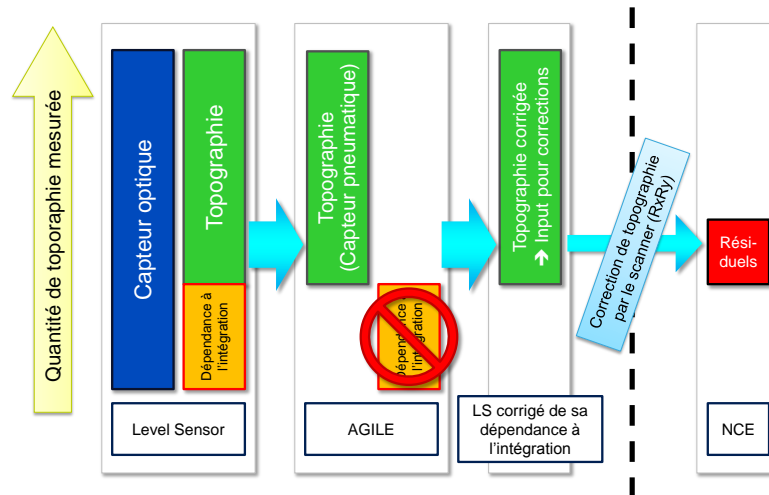


Figure 4-1 : Le procédé de leveling du scanner (NCE = non-correctable error soit les résiduels de topographie non correctable par le scanner)

4.1.2 Le Wyko

Pour la mesure de topographie haute fréquence, le Veeco NT9300 Wyko est parfaitement adapté [54] [55]. Les mesures ont été réalisées au CEA-LETI à Grenoble. Il s'agit d'un interféromètre de Michelson en mode de décalage de phase dont la résolution latérale atteint $0.1\mu\text{m}$ et la résolution verticale 1nm . Dans cette étude, la mesure a été réalisée avec un pixel de quelques micromètres de côté (cf. Tableau 4-1). Cet interféromètre travaille à une longueur d'onde de 545nm . Il offre la possibilité de mesurer une zone de 2.4mm par 1.8mm en une seule fois et par assemblage automatique de plusieurs images, il est possible de mesurer des surfaces de plusieurs dizaines voire centaines de centimètres carrés (cf. Figure 4-2).

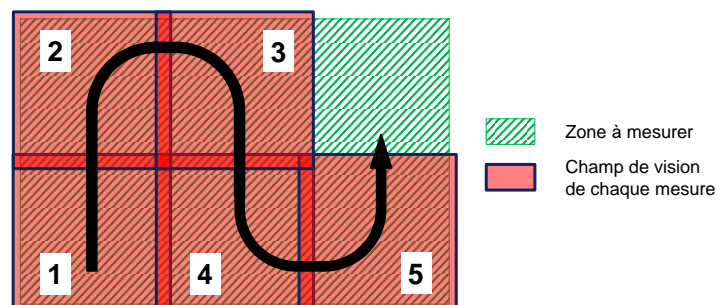


Figure 4-2 : La méthode de mesure par assemblage d'image sur le Wyko

La quantité de donnée générée par la mesure Wyko est très importante. Lors de la mesure du champ sur le Contact 14nm FD-SOI, près de 55 millions de points ont été mesurés. Sur le BEOL 28nm FD-SOI, la mesure a été réduite à 12 millions de points. La mesure est assez complexe à réaliser car l'alignement de la plaquette est manuel et la quantité de points de mesure conduit à une complexification de l'assemblage automatique de l'image. Pour permettre la mesure d'une très grande surface, il est donc nécessaire de soit diminuer la résolution soit diminuer le recouvrement des différents scans de mesure. Dans chaque cas cela revient à diminuer la qualité des mesures. Cependant, comme le montrera la suite, une mesure à la plus haute résolution n'est pas forcément nécessaire pour permettre la création d'un modèle de topographie performant. La mesure dure environ 40 minutes pour un champ complet sur le wafer. Le champ mesuré se situe à mi-rayon de la plaquette.

Les conditions de mesures sont données dans le Tableau 4-1 :

	14nm FD-SOI Contact	28nm FD-SOI BEOL
Taille du champ	22.9 × 29.4	22.3 × 29.9
Nombre de scans Wyko (en X et en Y)	11 × 19	12 × 21
Recouvrement des scans	15%	25%
Résolution latéral (µm)	3.687	7.375
Nombre de pixels (en X et en Y)	6834 × 8001	3065 × 4136

Tableau 4-1 : Paramètres de mesures Wyko réalisées sur les wafer 28 et 14nm FD-SOI

Comme pour le capteur optique du scanner, il existe un risque d'erreur de mesure du fait de la réflectivité de l'empilement des matériaux sur le wafer (les couches transparentes à la lumière blanche sont particulièrement dérangeantes). C'est pourquoi il est conseillé de déposer une couche réfléchive en surface. En l'occurrence, on dépose 30nm de Tantale, très dense optiquement, sur les échantillons. Le dépôt est conforme pour éviter de gommer la topographie et a spécifiquement été développé pour les besoins de cette mesure [56]. Cette méthode impose aussi de n'utiliser que des wafers sans empilement de lithographie (BARC + Résine ou Tri-couche présentées en Chap. 2) car il est constitué de couches polymères, ce qui empêche la réalisation du dépôt de Tantale à haute température. Aussi, toutes les mesures Wyko sont réalisées juste après dépôt du diélectrique et du masque dur. Du fait de ce dépôt de Tantale, la mesure est destructive.

Certaines plaques ont été envoyées avec l'empilement polymère de lithographie usuel – et donc sans Tantale – et ont été mesurées de la même manière, confirmant que les interférences optiques sont trop fortes et les mesures inutilisables.

Les données brutes doivent être traitées et corrigées avant de pouvoir être utilisées. La contribution du chuck et des erreurs de déplacement de la plaque pendant la mesure ainsi que les erreurs d'assemblage d'image et la précision de la mesure doivent être vérifiées. Dans le cas de ce travail, des champs complets

(env. 26mm x 33mm) ont été mesurés. La précision de la mesure a été évaluée par F. Dettoni [56] comme étant de 1nm 3σ par point de mesure et de l'ordre de 5% de l'amplitude totale pour une mesure d'une puce entière avec un minimum de 1nm et un maximum de 10nm après traitement de recouvrement des différentes images et soustraction des effets de support.

Lors de la mesure, le logiciel du Wyko assemble les images entre elles pour former l'image complète de la surface mesurée. Les données brutes sont extraites et leur traitement est réalisé à l'aide du logiciel libre Gwyddion 2.39.

La première étape consiste à soustraire la forme globale de cette surface qui est d'un ordre polynomial élevé et contient la forme globale du wafer à l'échelle du champ mesuré d'une part et la forme du chuck du Wyko d'autre part. Elle correspond à une topographie fictive de plusieurs microns d'amplitude. Cette solution n'est pas la forme optimale de mise à niveau des données Wyko selon F. Dettoni [57] [58]. Le logiciel développé pendant sa thèse offre des solutions de leveling adaptées à la machine. Cependant, cette solution a été sélectionnée afin de garder la main sur le traitement des données et permettre une uniformisation des méthodes avec les ingénieurs d'ASML qui ont aidé à l'analyse. Une couche supplémentaire de traitement des données a été réalisée par ASML qui a simulé le leveling du scanner sur le set complet de données Wyko. La Figure 4-3 montre les mesures Wyko avant et après leveling des données brutes.

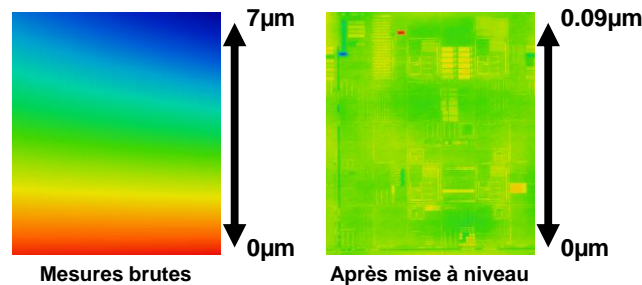


Figure 4-3 : Mise à niveau des données Wyko dans Gwyddion

Suite au leveling, on procède à une rotation (cf. Figure 4-4). Cette étape est nécessaire dans le sens où de légers désalignements peuvent apparaître entre les différentes prises d'image, provoquant une rotation globale de l'image. Les chemins de découpe servent de référence pour cette étape car ils sont tous soit horizontaux soit verticaux. Généralement, une rotation entre 0.5 et 2° est suffisante. La grille restant la même par rapport à l'image, une interpolation des données est réalisée. L'image est ensuite tronquée pour supprimer un artefact de la rotation. Des points non existants préalablement ont dû être créés pour combler les coins de l'image et une valeur arbitraire leur a été assignée. Il est donc préférable de mesurer un peu plus que la zone d'intérêt afin de ne pas perdre une partie des données en bord de matrice.

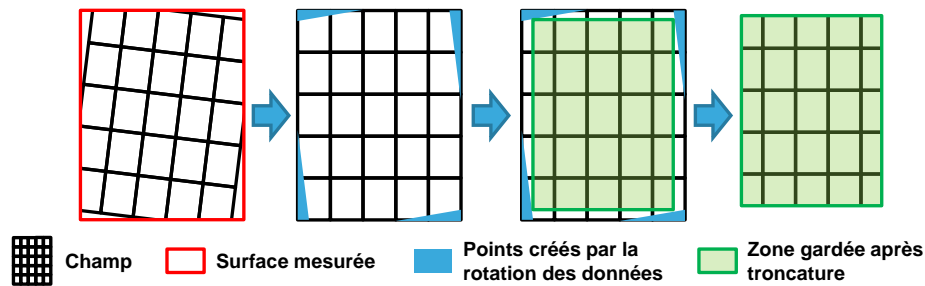


Figure 4-4 : Rotation et troncature des données

Le logiciel Gwyddion offre de nombreux outils de traitements en plus de l'interface graphique de visualisation et de correction de données. Entre autres, un outil d'analyse statistique très complet permet une étude globale du set de donnée et a été largement utilisé pour étudier la distribution de topographie sur la puce.

La dernière étape est de ré-échantillonner les mesures à plusieurs tailles de pixels sur la grille d'extraction des données de design (cf. Chap. 4.3.4). La fonction « Scale » du logiciel permet de changer la résolution de l'image et de ré-échantillonner les mesures. L'utilisateur choisit le nombre de pixel qu'il souhaite en X et en Y et la méthode d'interpolation pour recalculer les valeurs sur la nouvelle grille. Une interpolation de Schaum du 4^{ème} ordre [53] a été sélectionnée. Cette interpolation est celle qui modifiait le moins la distribution de valeur de topographie dans les mesures tout en gardant la continuité de la topologie de surface que l'on a sur le wafer. En revanche, cette méthode a un effet de filtre passe-haut.

Les mesures réalisées sur les plaquettes 14 et 28nm, respectivement au niveau Contact et en BEOL, montrent des résultats très différents (cf. Figures 4-5 et 4-6). Cela est majoritairement dû au fait que le masque pour le produit 14nm est un MPW (Multi Project Wafer) qui regroupent des motifs et des structures très diverses n'étant pas forcément des designs fonctionnels que l'on retrouvera sur un produit manufacturé comme celui présent sur le masque de production de 28nm qui a été sélectionné pour l'étude.

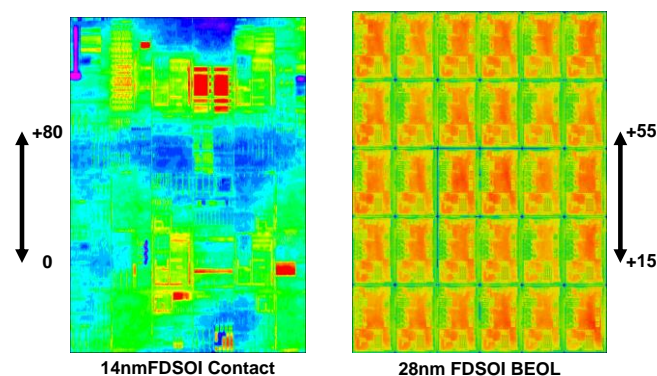


Figure 4-5 : Mesures Wyko champ complet des produits en 14nm FD-SOI et en 28nm. L'échelle est en nanomètres de topographie..

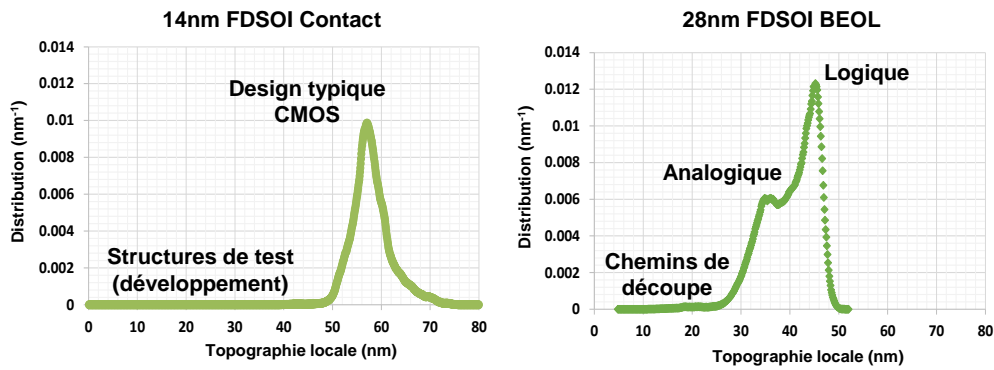


Figure 4-6 : Distribution de topographie sur un champ en 14nm FD-SOI Contact et 28nm BEOL

Comme pour les mesures de topographie du scanner, il est nécessaire de s'assurer de la stabilité de la topographie dans le temps. Tout d'abord, les mesures par le scanner sont stables ce qui donne un premier indice. Ensuite, des mesures ont été réalisées sur les mêmes plaquettes à l'aide d'une autre machine. Ces mesures sont décrites dans le chapitre suivant.

4.1.3 Le WaferSight PWG

Le WaferSight PWG (pour Patterned Wafer Geometry ou géométrie du wafer structuré) de KLA Tencor [59] permet quant à lui de mesurer un wafer complet dans le même temps de mesure que le Wyko pour un champ unique au prix d'une résolution latérale de l'ordre de 100 à 150 μ m, ce qui correspond à plusieurs millions de points de mesure par wafer.

Le WaferSight PWG est un double interféromètre de Fizeau travaillant à une longueur d'onde de 635nm permettant de mesurer simultanément les faces avant et arrière de la plaquette de silicium. Celle-ci est tenue verticalement pendant la mesure. A partir de ces mesures, il est possible de déterminer de nombreuses informations différentes pour le wafer comme les contraintes mécaniques dans la plaque ou la topographie. Dans le cadre de cette étude, la NanoTopographie et le SFQX sont les deux métriques qui nous intéressent car ce sont celles qui correspondent à des défocus lors de l'exposition.

- La **NT (nanotopographie)** qui est la topographie haute fréquence obtenue par filtrage des basses fréquences après combinaison des mesures face avant et face arrière sur un chuck virtuel proche de celui du scanner. Il s'agit de la topographie du wafer non « chucké » dans le scanner.
- Le **SFQR (Site Front Least sQuares Range)** et le **SFQX (Site Flatness Quality Residuals)** qui sont respectivement une estimation de la topographie du wafer une fois posé dans le scanner et une estimation de la partie non-corrigeable par le scanner.

Comme pour le Wyko et le scanner, la lumière utilisée est du domaine du visible ce qui risque de créer des erreurs lors de la mesure. Il est donc conseillé de recouvrir les plaquettes d'une couche métallique réfléchissante [60]. La mesure est donc destructive si l'étape choisie pour la mesure n'offre pas une surface uniformément métallique en production. Les wafers mesurés sont les mêmes que ceux qui ont été mesuré avec le Wyko. Il y a donc des wafers sans l'empilement de lithographie et avec 30nm de Tantale

conforme et des wafers avec l'empilement de production (Résine + anti-réfléctif) sans Tantale. Le détail de l'empilement présent sur le wafer a été fourni à KLA Tencor, ce qui a permis de modéliser la réponse du wafer à la mesure. Les mesures ont été faites par KLA Tencor sur leur site de Milpitas en Californie dans le cadre d'une démonstration de leur machine pour STMicroelectronics.

Les wafers mesurés dans le cadre de la démonstration sont au nombre de 9 et sont pour la plupart les mêmes que ceux qui ont été mesuré sur le Wyko au LETI et sur le scanner. Le détail des wafers mesurés est donné par le Tableau 4-2. Pour les mesures de topographie NT avec le PWG, l'incertitude de mesures est de 3nm 3s.

Wafer	Tri-couche	Tantale 30nm	Sans tri-couche ni tantale
3		X	
4	X		
5	X		
6			X
7			X
9	X		
10		X	
11	X		
12			X

Tableau 4-2 : Wafers envoyés chez KLA Tencor pour les mesures sur le WaferSight PWG

A l'échelle du wafer complet, la courbure est très grande, jusqu'à plus de 100 μ m pour les wafers recouverts de Tantale, car le dépôt du métal crée des contraintes tensives à l'échelle du wafer, et de l'ordre de 60 à 70 μ m pour les wafers standards (avec et sans tri-couche) qui, grâce à la modélisation de la réponse de l'empilement à la mesure, présentent des résultats exploitables. Ces résultats sont donnés en Figure 4-7. Le wafer est collé par aspiration sur le chuck du scanner pendant l'exposition, ce qui réduit la topographie que la machine doit corriger à 1 à 2 μ m à l'échelle du wafer complet comme décrit dans la partie sur le leveling du scanner (cf. Chap. 3.3.4 et 3.4.1).

En retirant la forme globale de la courbure puis en décomposant le résultat en une partie intra-plaque et une partie intra-champ, on extrait la nanotopographie et le SFQX du wafer. Le graphique de la Figure 4-8 présente les valeurs du SFQX 3 σ par champ chacun des 9 wafers de l'étude. On remarque que les wafers standards de lithographie, comme on pouvait l'espérer, présente une topographie non corrigeable par le scanner moindre que les autres avec 50nm de défocus non corrigeable contre 70nm pour les wafers recouverts de Tantale. Cela signifie que la tri-couche aplanie effectivement la surface du wafer à l'échelle de la centaine de microns. Le modèle de défocus de KLA Tencor n'est pas parfait et reste une estimation des capacités du scanner [61].

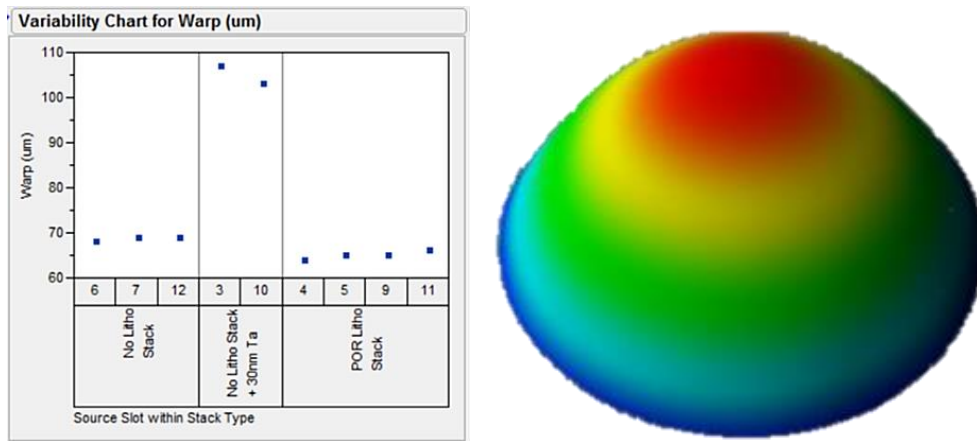


Figure 4-7 : A gauche, courbures des wafers en Contact 14nm FD-SOI à l'étape de lithographie. La mesure du wafer POR¹⁶ pour Process of record (i.e. wafers standards tri-couche) est représentée en 3D à droite. (Source : KLA Tencor)

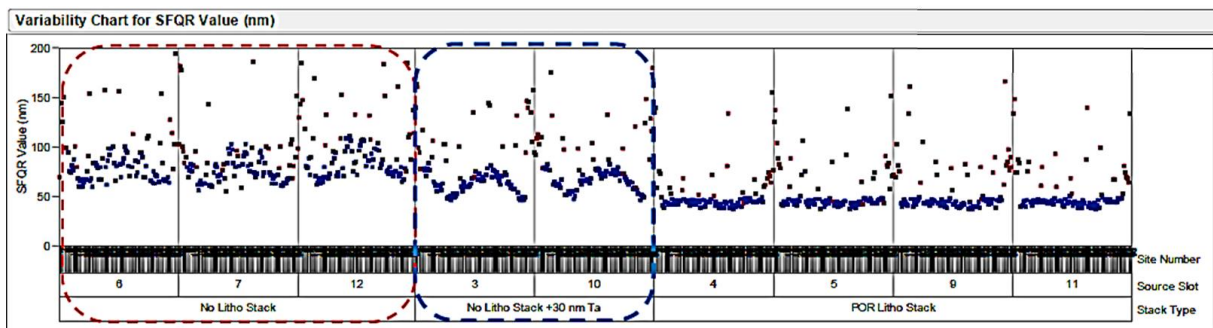


Figure 4-8 : Défocus estimé à l'échelle d'une plaquette entière. (Source : KLA Tencor)

4.1.4 Comparaison des différentes méthodes de mesure

Afin de s'assurer que chacune de ces méthodes de mesure permet d'obtenir les mêmes résultats, une étude comparative des topographies intra-champ a été menée. Elle a été faite en deux fois. Tout d'abord, avec ASML, le capteur de niveau du scanner a été comparé avec le Wyko. Ensuite, dans le cadre de la démonstration réalisée avec KLA Tencor, les mesures Wyko ont été mises en parallèle aux données PWG. Les mesures de topographie intra-champ avec ces trois machines sont donnés par la Figure 4-9.

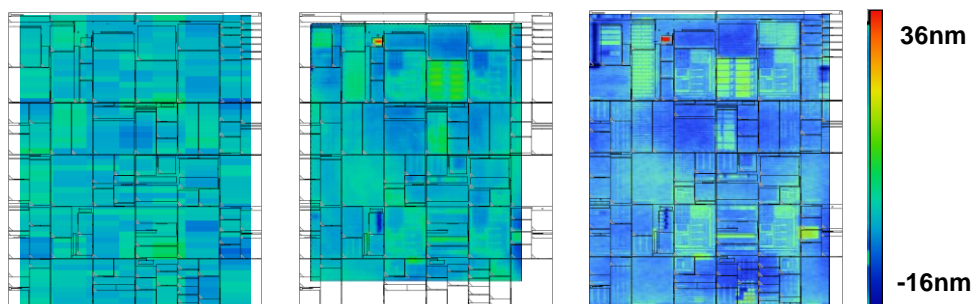


Figure 4-9 : Cartographies des mesures de topographie avec le scanner (à gauche), le PWG (au milieu) et le Wyko (à droite)

¹⁶ Le procédé POR (process of record) est le procédé de référence que les wafers subissent en production pour la réalisation d'une étape de fabrication. Ici, il s'agit de l'empilement standard tri-couche de lithographie. Les wafers POR mesurés sont dans le même état qu'un wafer qui serait exposé dans le scanner.

Pour comparer le Wyko au capteur du scanner comme aux résultats PWG, il est nécessaire d'utiliser Gwyddion pour changer la grille et recalculer les mesures Wyko avec un échantillonnage plus réduit. Cette analyse a été faite par ASML (cf. Figure 4-10)

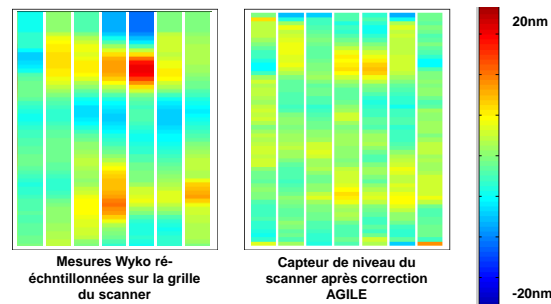


Figure 4-10 : Comparaison des mesures Wyko et scanner (Source : ASML)

En comparant les cartographies obtenues avec le capteur de niveau du scanner (avec correction par le capteur pneumatique) et le ré-échantillonnage des mesures Wyko, les résultats sont très similaires à l'exception de l'amplitude qui est plus faible pour la mesure dans le scanner

La comparaison entre Wyko et PWG donne aussi de très bons résultats. En ne prenant que les valeurs appartenant à la distribution autour de la moyenne de topographie plus ou moins 3σ , la pente est de 1 et le R^2 de l'ordre de 0.7. En prenant la distribution complète, on remarque que le R^2 reste bon mais la pente diminue. Cela est très certainement dû à la différence de résolution entre le Wyko et le PWG. En effet, la résolution plus faible de 2 ordres de grandeur du PWG (une centaine de microns pour le PWG par rapport à quelques microns pour le Wyko) aura plus de mal à détecter de hautes collines ou de profondes vallées sur le wafer si les dimensions latérales de celles-ci sont très réduites. La topographie très haute fréquence est donc sous-estimée par le PWG d'environ 40%.

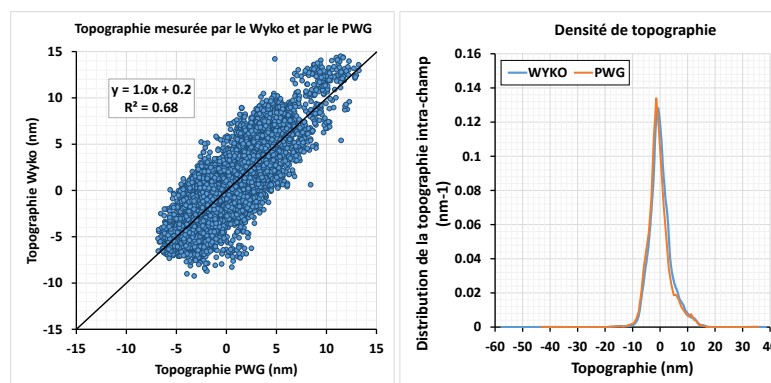


Figure 4-11 : Corrélation entre le Wyko et le PWG. A gauche, la corrélation sur l'ensemble des points de mesures et, à droite, sur la distribution sans les valeurs extrêmes que ne peut pas résoudre le PWG.

Les mesures PWG ont montré que le trilayer aplanie le wafer d'une vingtaine de nanomètre à l'échelle de la centaine de microns. Il est important de noter que le modèle sera dans la suite construit sur les données Wyko et donc à partir de mesures réalisées sur des plaques recouvertes de Tantale. Il est donc possible que le défocus dérivé d'un tel modèle soit légèrement surestimé si on regarde les valeurs à cette échelle.

Une sous-estimation des valeurs est aussi causée par la taille du spot lumineux et la résolution de la mesure. Ainsi, l'amplitude mesurée par le scanner est plus faible que celle mesurée par le PWG elle-même plus faible que celle mesurée par le Wyko.

Cependant, des mesures ont été réalisées chez ASML à Veldhoven (Pays-Bas) à l'aide de leur dernière génération de capteur de niveau. Ces mesures ont montré que les wafers recouverts de Tantale et les wafers recouverts de l'empilement de lithographie ont, à l'échelle du millimètre, une topographie intra-champ équivalente.

De plus, comme montré dans le Chap. 3.4, la corrélation entre les mesures de focus et la topographie mesurée avec le Wyko est très bonne ce qui signifie que la planarisation de la plaquette que détecte le PWG lors de sa mesure pleine plaque de la surface n'a pas une grande influence à l'échelle du micromètre.

4.2 LA REGRESSION PLS

4.2.1 Introduction

Développée par Herman Wold [62] dans les années 1970, la méthode de régression PLS ou Partial Least Square (moindres carrés partiels) a prouvé son intérêt pour l'industrie chimique [63] [64] et pétrolière, en géologie et plus récemment pour le contrôle de procédé. Dans chacune de ces situations les résultats obtenus dépendent de très nombreux paramètres et il est souvent difficile de déterminer le ou lesquels ont une influence et si oui dans quelle mesure. La régression PLS offre une solution d'analyse multivariée qui permet de répondre à la problématique de détermination des facteurs influents sur un paramètre choisi et propose aussi un classement de ceux-ci par ordre d'importance. En pétrochimie, elle a permis entre autres l'optimisation des carburants automobiles ; en géologie, de déterminer les conditions de formation de telle ou telle roche ; et en contrôle de procédé, de limiter les échantillonnages et ainsi éviter des mesures trop nombreuses qui impactent la productivité.

Dans cette étude, le logiciel SIMCAP+ v11.0 a été utilisé pour construire les modèles PLS. Le logiciel, édité par Umetrics et développé par L. Eriksson et S. Wold [65], permet de faire une analyse PLS robuste en quelques minutes.

L'établissement d'un modèle PLS avec SIMCAP suit la méthode décrite dans la figure 4-12. Seule l'étape de préparation de données se fait en dehors du logiciel. Les blocs de données X et Y sont sélectionnés manuellement directement dans l'interface de SIMCAP. Les trois étapes de la résolution de la PCA (Analyse en composantes principales, cf. Chap. 4.2.2), de la prédiction et de son amélioration sont toutes réalisées en intégralité par le logiciel et les calculs sont résolus automatiquement. La préparation des données est développée dans les parties suivantes sur les mesures de topographie (fichiers de calibration du modèle i.e. bloc Y) et les extractions de GDS (données d'entrée pour le calcul i.e. bloc X) respectivement (cf. Chap. 4.1 et 4.3).

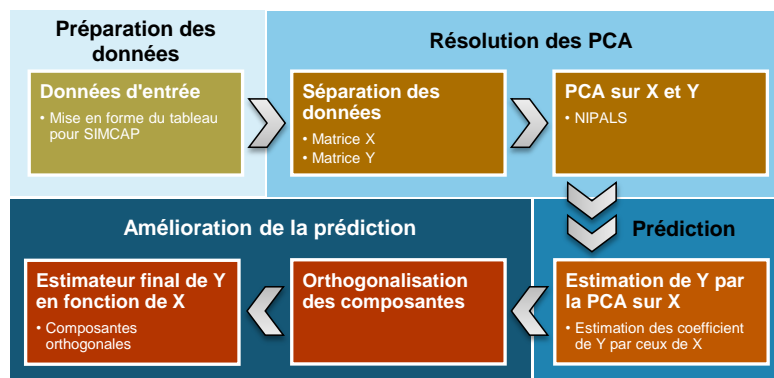


Figure 4-12 : Algorithme de la PLS adapté au logiciel SIMCAP

4.2.2 L'algorithme PLS

Dans cette partie, la méthode PLS sera décrite de manière simplifiée afin d'en comprendre le principe et le fonctionnement [66] [67] [68] [69]. Les indicateurs de qualité que SIMCAP+ détermine en plus des coefficients de régression seront aussi présentés. Tout d'abord, il convient néanmoins de définir quelques termes et paramètres qui seront nécessaires à la compréhension du manuscrit (cf. tableau 4-3).

Symbole	Description	Paramètre
$\ \cdot \ $	Norme euclidienne	
n	Nombre d'observables de l'échantillon de calibrage	Nb de points (topographie)
m	Nombre de paramètres d'entrée (vecteurs de X)	Nb de fichiers de densité de design
p	Nombre de variables de sortie (vecteurs de Y)	Nb de fichiers de topographie
h	Indice de numérotation des composantes	
i	Indice de numérotation des observables	
X	Matrice de paramètres d'entrée (taille $n \times m$)	Matrice des densités de design
Y	Matrice de variables de sortie (taille $n \times p$)	Topographie mesurée
Y_{pred}	Matrice de Y prédite par le modèle (taille $n \times p$)	Topographie prédite
E_h	Résidus de la PCA à h composantes sur X (taille $n \times m$)	
F	Résidus de la PCA à h composantes sur Y (taille $n \times p$)	
err	Seuil de précision de la régression	
T	Matrice des coefficients PCA du bloc X (taille $n \times a$)	
P'	Matrice de poids du bloc X (taille $a \times m$)	
U	Matrice des coefficients PCA du bloc Y (taille $n \times a$)	
Q'	Matrice de poids du bloc Y (taille $a \times p$)	
\hat{u}_h	Estimateur de u_h en fonction de t_h	
b_h	Coefficient de régression PLS pour le paramètre h	Coefficients du modèle (<i>Topographie = B * Densités</i>)
w_h	Pondération du paramètre h (forme matricielle W)	
\hat{t}_h	Estimateur de t_h pour la prédiction de Y en fonction de X	
$RESS_h$	Somme des résidus carrés entre Y et Y_{pred} pour les h premières composantes	Critères de sélection de la composante pour le modèle
$PRESS_h$	$RESS_h$ des h premières composantes sans l'observable i , $i < h$	
Q_h^2	Paramètre de validation croisée de la composante h dans SIMCAP+ 11.0	Performance attendue en prédiction
VIP	Variable Importance in the Projection	Critère de sélection d'un fichier de densité pour le modèle

Tableau 4-3 : Tableau des paramètres et des notations mathématiques. La troisième colonne fait le lien entre les notations mathématiques et le cas étudié de la modélisation de topographie à partir du design.

La régression PLS repose sur l'utilisation de l'algorithme NIPALS (Nonlinear Iterative Partial Least Squares, ou algorithme itératif non-linéaire de moindres carrés partiels), servant à la détermination des composantes dans l'analyse PCA (Principal Component Analysis, ou Analyse en Composantes Principales) [70] [71]. La PCA consiste à trouver les composantes principales p'_h de X . Celles-ci permettent d'estimer à l'influence de chaque paramètre indépendamment des autres et contrairement à la méthode classique des vecteurs propres, les composantes principales sont ici calculées les unes après les autres par ordre d'importance. Les vecteurs résultats sont les mêmes et cette méthode permet de s'arrêter quand les vecteurs propres suivants ne décrivent plus assez le système pour être utiles. L'algorithme NIPALS répète itérativement la suite :

$$E_h = E_{h-1} - t_h p'_h \quad \text{pour } h \in [1, n] \text{ jusqu'à ce que } \|E_h\| < \text{err} \quad (16)$$

Où *err* est la précision avec laquelle on veut décrire le système et en prenant $E_0 = X$.

Pour la PLS, l'algorithme NIPALS est appliqué au bloc des données d'entrée X , soit les paramètres du modèle, et à celui des données de sortie Y , soit les données de calibration du paramètre que l'on veut modéliser. De cette manière, les composantes principales de X et Y sont dérivées de l'algorithme.

$$\begin{cases} X = TP' + E = \sum t_h p'_h + E \\ Y = UQ' + F^* = \sum u_h q'_h + F \end{cases} \text{ soit } \begin{cases} \text{Densités} = TP' + E \\ \text{Topographie} = UQ' + F \end{cases} \Leftrightarrow \begin{cases} PCA_{\text{Densités}} \\ PCA_{\text{Topographie}} \end{cases} \quad (17)$$

La finalité est de mettre en évidence une relation entre X et Y . En exprimant u_h en fonction de t_h , il est possible d'en extraire une fonction linéaire de l'estimateur de u_h nommé \hat{u}_h .

$$\hat{u}_h = b_h t_h \text{ soit } PCA_{\text{Topographie}} = b_h * PCA_{\text{Densités}} \quad (18)$$

Ainsi, Y peut s'estimer en fonction des composantes principales de X :

$$Y = \sum b_h t_h q'_h + F = TBQ' + F \text{ avec } \|F\| \text{ minimale} \quad (19)$$

$$\text{Topographie} = f(PCA_{\text{Densités}}) \quad (20)$$

Les composantes ne sont pas orthogonales car l'algorithme NIPALS a été légèrement modifié pour le calcul de la PLS. L'équation de Y en fonction des composantes de X est la suivante après orthogonalisation.

$$Y = \sum b_h w_h q'_h + F = WBQ' + F \text{ avec } \|F\| \text{ minimale} \quad (21)$$

Cette expression correspond à la description et à la modélisation des valeurs du bloc Y par les composantes principales du bloc X de paramètres d'entrée. L'inconvénient de cette notation est qu'on est obligé d'avoir les composantes principales de Y pour calculer l'estimation de Y par X . Or ce que l'on veut, c'est calculer un bloc Y inconnu à partir des paramètres X .

Pour cela, il faut pouvoir exprimer directement le bloc Y en fonction de X . On calcule alors les coefficients de régression b_{PLS} à partir des poids w_h et des composantes principales du bloc Y de calibration :

$$B_{PLS} = W^* Q' \quad (22)$$

$$\Rightarrow Y_{pred} = X B_{PLS} + F \text{ avec } \|F\| \text{ minimale} \quad (23)$$

$$\Rightarrow \text{Topographie}_{prédite} = \text{Coefficients}_{PLS} \times \text{Densités} + \text{Précision du modèle} \quad (24)$$

Comme pour la PCA, il est possible de s'arrêter quand $\|F_h\| < err$, ce qui permet de ne garder que les composantes principales de X permettant de décrire correctement Y. Cependant, augmenter le nombre de composantes va diminuer la valeur de $\|F\|$ et donc améliorer la description de Y par X. Cela ne signifie pas forcément une meilleur prédictibilité de Y par X et peut même dans certains conduire à une détérioration de celle-ci. La prédiction se détériore lorsque les composantes supplémentaires ajoutent un poids à des artefacts, en modélisant du bruit de mesures par exemple.

Au final, on obtient un modèle du type :

$$Topographie_{prédite} = \sum Coefficient_{PLS}(densité_{niveauN}) * Densité_{niveauN} \quad (25)$$

4.2.3 Les indicateurs de performance

Pour éviter de se retrouver dans la situation de la prédiction de bruit de mesure, il est nécessaire d'avoir un indicateur pour juger de la pertinence de chaque composante calculée par l'algorithme. La somme des résidus carrés de la prédiction (Prediction Residual Sum of Squares ou *PRESS*) permet cela. En minimisant le *PRESS*, il est possible d'obtenir le modèle le plus prédictif possible tout en restant précis.

Le logiciel SIMCAP+ calcule un paramètre noté Q_h^2 , qui est un indicateur de validation croisée permettant de sélectionner automatiquement le nombre de composante. L'indicateur Q_h^2 est calculé comme suit :

$$Q_h^2 = 1 - \frac{PRESS_h}{RESS_{h-1}} \quad (26)$$

$RESS_{h-1}$ est calculé avec toutes les observables et les $h - 1$ premières composantes du modèle uniquement alors que $PRESS_h$ est calculé en retirant successivement toutes les observables i et avec les h premières composantes du modèle. On peut juger de la pertinence de la composante h dans la prédiction de l'observable i retirée du calcul, le critère pour garder la composante h étant :

$$PRESS_h \leq 0.95 \times RESS_{h-1} \Leftrightarrow Q_h^2 \geq 0.05 \quad (27)$$

Si $PRESS_h$ est trop proche de $RESS_{h-1}$, alors l'influence de la composante h sur la prédiction des observables i est négligeable. Celle-ci est rejetée pour ne pas complexifier le modèle et risquer de perdre en prédictibilité. La sélection ou non d'une composante supplémentaire va influencer la valeur des coefficients de régression car chacune va préciser un peu plus chaque coefficient de la régression.

En ajoutant tous les Q_h^2 , on obtient un autre indicateur important :

$$Q_{cum}^2 = \sum Q_h^2 \quad (28)$$

Cet indicateur donne une estimation de la qualité totale de prédiction que l'on peut attendre du modèle PLS à h composantes ainsi construit. Pour plus de simplicité, on notera : $Q_{cum}^2 = Q^2$.

L'utilisateur peut aussi sélectionner manuellement le nombre de composantes qui lui sont nécessaire pour améliorer son modèle. Pour cela, il faut calculer toutes les composantes et regarder les Q_h^2 de chacune d'entre elles. SIMCAP arrête automatiquement le calcul des composantes suivantes dès qu'il trouve la composante h qui ne vérifie par le critère de sélection. Malgré tout, il est tout à fait possible que l'une au moins des composantes suivantes soit importante dans le modèle et satisfasse le critère.

Enfin, SIMCAP fournit un dernier indicateur qui sera largement utilisé dans la suite de l'étude. Ce paramètre est le *VIP* [72] pour Variable Importance in the Projection, ou Importance de la variable dans la projection. Le *VIP* caractérise l'importance relative de chacun des m paramètres d'entrée sur laquelle l'algorithme NIPALS a été réalisé dans le modèle. Il permet de faire un tri entre les paramètres nécessaires au modèle et les paramètres superflus. Un *VIP* supérieur à 1 impose l'utilisation du paramètre car son influence est non-négligeable. Une valeur inférieure à 0.8 caractérise un paramètre négligeable que l'on peut exclure du modèle sans impacter les performances. Pour les valeurs de *VIP* comprises entre 0.8 et 1, l'utilisateur décide de conserver ou non le paramètre. Le retrait des paramètres est manuel quelle que soit la valeur de *VIP* calculée par le logiciel.

Ici, le *VIP* est utilisé pour trier les fichiers de densités de design qui permettent d'avoir un modèle performant de ceux qui ne sont pas utiles. La Figure 4-13 donne un exemple de graphique de *VIP*.

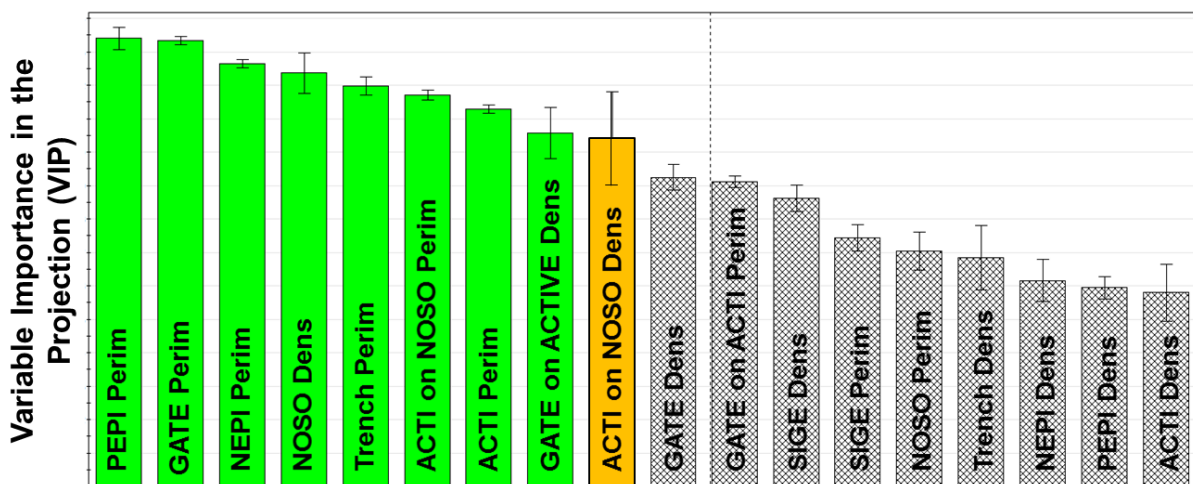


Figure 4-13 : Graphique de *VIP* fourni par le logiciel SIMCAP. Les barres vertes correspondent à un *VIP* > 1, les grises à un *VIP* < 0.8 et en orange entre 0.8 et 1.

4.3 METHODOLOGIE ET RESULTATS

Dans le chapitre 3, il a été montré que le focus était décalé par la partie non-corrigeable de la topographie intra-champ. Dans cette même partie du manuscrit, il a été expliqué comment la topographie intra-champ est dépendante du design, à la fois à l'échelle des motifs, à l'échelle des zones de la puce et à celle

macroscopique de la taille des blocs (mémoire, logique, analogique, ...) et du champ complet. Dans la suite, le travail se porte sur trois échelles de topographie : la topographie basse fréquence (échelle millimétrique), moyenne fréquence (échelle de la centaine de microns) et haute fréquence (échelle micrométrique).

Cette topographie est une donnée physique absolue, au contraire du focus qui est une valeur arbitraire dépendant de la calibration en cours du scanner et sujette à des modifications suite au suivi des dérives de la machine. Ainsi, le défocus provoqué par la topographie est-il lui aussi une valeur absolue. Il existe déjà dans la littérature [73] et dans l'industrie des moyens de prédire cette topographie (comme le logiciel CMP Predictor de chez Cadence Design Systems ou Calibre CMPAnalyzer édité par MentorGraphics). Il est aussi possible de corriger le design dans une certaine mesure en introduisant des règles de dessin [74] [75]. Dans toutes ces méthodes, on utilise des simulations physiques rigoureuses ou approchées qui vont permettre de modéliser à la fois les matériaux utilisés et le procédé de CMP pour obtenir une cartographie de la topographie attendue dans une puce à l'étape choisie.

Dans le cadre de cette étude, le but était de trouver une méthode simple et rapide à mettre en œuvre permettant de prédire la topographie et définir des zones à risque lesquelles pourront servir à optimiser le contrôle du focus de manière diverses.

La topologie intra-puce est très dépendante au design, à sa densité et à la taille des motifs. Ainsi, même une analyse visuelle des cartographies de densité locale de design par rapport à la topographie mesurée permet de conclure qu'une large part de celle-ci est issue de cette densité locale de motif. Aussi, la méthode développée ci-après montrera comment relier le design à la topographie avec un modèle empirique.

Dans la suite, la construction du modèle depuis les prémices de l'étude jusqu'au modèle de prédiction de la topographie haute fréquence est présentée.

4.3.1 La genèse de l'idée fondatrice

L'idée d'utiliser les densités de masque pour modéliser le leveling du scanner puis la topographie haute fréquence vient de la comparaison d'un fichier de densité périmétrique de Grille sur Active (c'est-à-dire la densité de recouvrement des zones actives par la grille des transistors, illustrée par le schéma de la Figure 4-14) extraite pour des raisons électriques sur un design en 28nm FD-SOI avec les données du fichier journal du scanner.

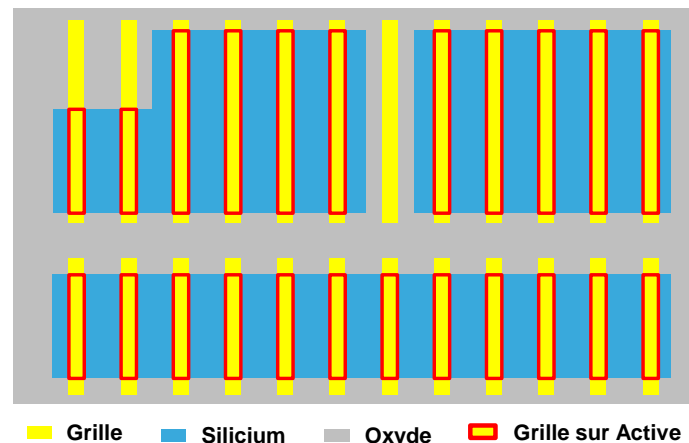


Figure 4-14 : Schéma explicatif de ce que représente le périmètre de Grille sur Active. Il s'agit du périmètre des polygones rouges.

Visuellement, la corrélation apparaît nettement et les résultats de la corrélation point à point ont confirmé l'intuition. Ces résultats ont fourni une opportunité quant à la manière de modéliser la topographie.

Un traitement particulier permet de ré-échantillonner les données design sur la grille de mesure de la topographie par le scanner. Le but est d'obtenir la topographie lue par le Level Sensor du scanner si on lui présentait une topographie intra-champ égale à la densité locale de Grille sur Active. Une simulation des corrections scanner sur l'échantillonnage scanner de la densité de design a ensuite été réalisée. En effet, les deux sets de données à comparer ne contiennent pas exactement le même type d'information. D'une part, le fichier extrait du journal du scanner contient les résiduels après correction mécanique de la topographie de surface. D'autre part, la densité devrait être corrélée, à priori de manière assez importante, avec la topographie avant correction. La cartographie de résiduels non-corrigeables obtenue par application de la correction scanner sur la densité est purement fictive et n'a pour seul but que de pouvoir comparer des données qui représentent « la même chose », au moins au même niveau de traitement. Pour cela, l'effet de la densité de Grille sur Active sur la topographie a été considérée comme étant exactement égale à un facteur 1, ce qui est faux en réalité.

La figure 4-15 ci-dessous présente la méthodologie de cette première analyse et le résultat de la corrélation point à point.

Le R^2 de 0.54 montre clairement une corrélation entre la densité périmétrique du design de la grille sur active et la topographie mesurée et corrigée par le scanner. Cette topographie étant le résultat de la superposition de nombreux niveaux de masques et matériaux, il semble probable qu'une combinaison des densités de plusieurs niveaux de design précédant l'exposition en lithographie permettrait de modéliser la topographie à ce niveau. En raison de la multiplicité des densités de design (échelle, combinaison de niveau entre eux, type de densité ; cf. Chap. 4.3.4), un outil de régression multi-variables est nécessaire pour pouvoir trouver rapidement et de manière robuste les niveaux de design qui sont nécessaires à la construction et au calcul du modèle. Parmi les solutions existantes (Machine learning, analyse multi variées,..), la régression PLS a été sélectionnée en raison de sa simplicité de mise en œuvre.

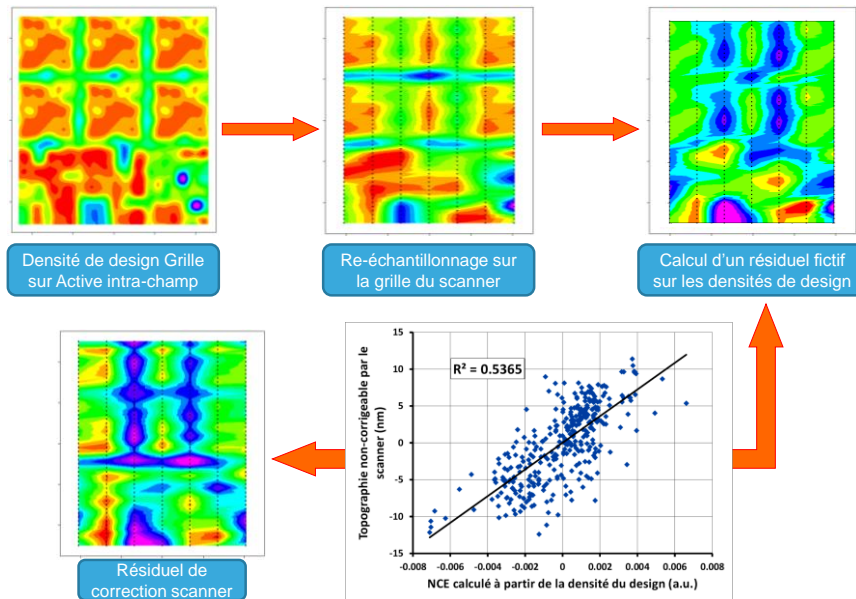


Figure 4-15 : Corrélation originelle à partir de laquelle la méthodologie a été développée

L'analyse précédente a été menée sur les résiduels de correction de la topographie par le scanner de lithographie. Ceci demande beaucoup de manipulation de données pour calculer le NCE fictif sur les densités de design avant de pouvoir construire le modèle. Ainsi, dans la suite, la mesure de topographie par le scanner (par le capteur optique tout d'abord puis la correction apportée par l'utilisation d'AGILE, cf. Chap. 3.4.1 « Correction de la topographie par le scanner ») sera l'objet de la modélisation. Il suffira ensuite d'appliquer un modèle de la correction du scanner sur le résultat de la modélisation pour obtenir l'erreur résiduelle.

4.3.2 Le principe de l'idée fondatrice

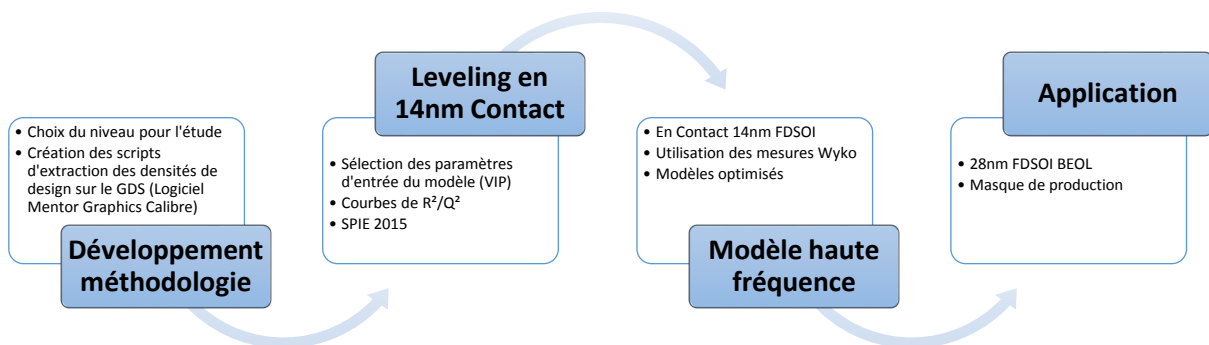


Figure 4-16 : Les quatre phases de la construction du modèle PLS de topographie

La création du modèle de topographie s'est fait en 4 grandes phases, illustrées par le Figure 4-16 ci-dessus.

La première consistait à sélectionner le meilleur candidat pour permettre la création d'un modèle performant et pertinent dans le sens où l'existence du modèle permettrait de prévoir des défocus pour un procédé critique (cf. Chap. 5). Cette phase a aussi permis de prendre en main les scripts SVRF d'extraction de densité au niveau du design de la puce (cf. Chap. 4.3.4).

La deuxième phase a permis de modéliser les données de leveling du scanner de lithographie. Le but était de confirmer la possibilité d'utiliser la méthode PLS comme méthode robuste de prédiction de la topographie (cf. Chap. 4.3.5).

Enfin, en troisième phase, un modèle de topographie haute fréquence a été développé à partir des mesures Wyko (cf. Chap. 4.3.6). Le modèle « haute fréquence » a été optimisé puis une comparaison avec le modèle de leveling et le modèle « basse fréquence » est proposée (cf. Chap. 4.3.7).

La quatrième et dernière phase a consisté à appliquer cette méthodologie dans le cadre d'une étude plus large en partenariat avec la société ASML sur le Back-End Of Line du 28nm FD-SOI (cf. Chap. 4.3.8 et Chap. 5).

Dans chacune de ces phases, des modèles PLS de plus en plus performants ont été construits. Leur construction suit dans chaque cas le même schéma. La Figure 4-17 décrit cette méthode qui permet de relier les données design et les données de topographie mesurées sur silicium.

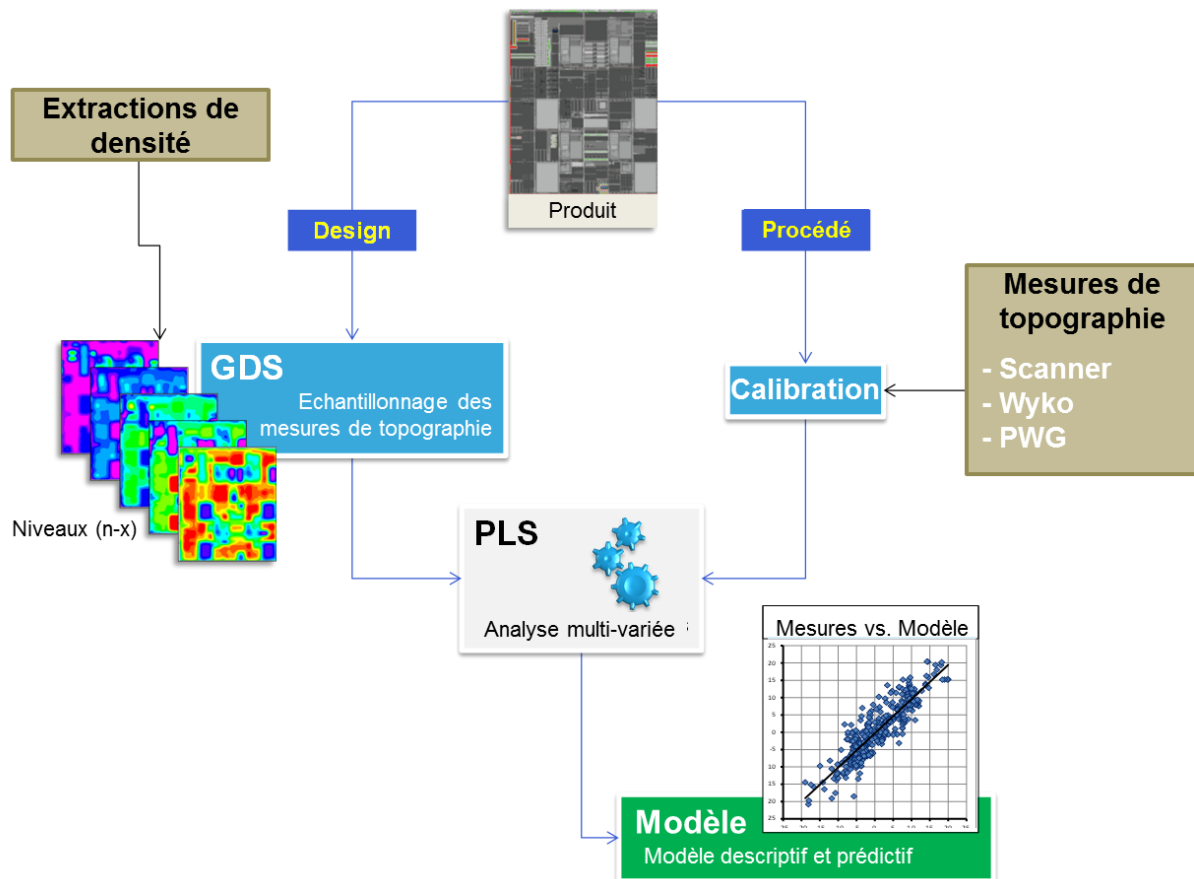


Figure 4-17 : Méthodologie pour la modélisation de la topographie par analyse PLS [32]

Tout d'abord, il faut mesurer les wafers de calibration du modèle sur le Wyko, le PWG ou le scanner selon la résolution spatiale et le type de topographie que l'on veut modéliser. Ensuite, coté GDS, on extrait les densités du design en adaptant la taille de pixel à la mesure de topographie. En réalité, il sera

nécessaire de faire coïncider la grille d'échantillonnage des densités et la grille des mesures pour obtenir un résultat correct car la méthode PLS consiste en une analyse point à point des données.

Le choix de niveaux de design sur lesquels réaliser les extractions se fait grâce à la connaissance de l'intégration et de l'enchaînement des procédés de fabrication. Quand il arrive à l'étape choisie pour l'étude, le wafer sera passé par un certain nombre d'étapes de fabrications lesquelles auront eu plus ou moins d'influence sur la construction d'une topographie sur la plaquette. La connaissance de ces procédés permet de trier les étapes susceptibles d'impacter la topologie de surface de celles qui n'auront aucune influence.

Le principe va ici être expliqué à partir de la route du produit 28nm qui avait été prise comme exemple dans le Chap. 1. Avant la réalisation du premier niveau d'interconnexion, (cf. schéma de droite dans la figure 4-18), l'ensemble du Front-End et du Middle-End ont été fabriqués.

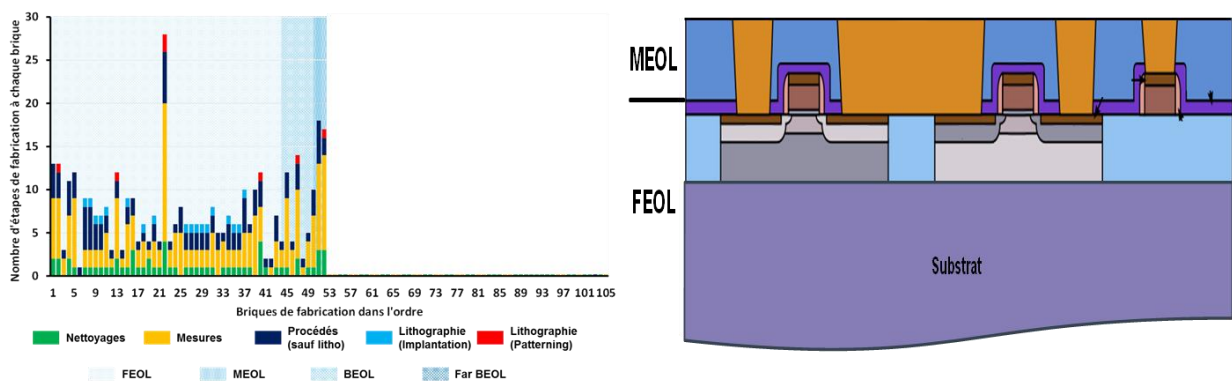


Figure 4-18 : Etapes déjà réalisées de l'intégration (cf. Figure 1-6 p.38) et empilement présent sur le wafer au moment de l'exposition du premier niveau de BEOL en 28nm.

Le wafer a vu une succession de dépôts, lithographies, gravures, implantations, traitements thermiques, polissages. Il apparaît que les traitements thermiques et les implantations ne vont pas jouer un grand rôle dans la construction de la topographie du wafer car il s'agit de procédés qui modifient la composition et la microstructure des matériaux préalablement déposés sur la plaquette. La lithographie, la gravure, la CMP (polissages) et les dépôts quant à eux vont ajouter et retirer de la matière et permettre une structuration spatiale des différents matériaux les uns par rapport aux autres. Ainsi la surface du wafer est recouverte de manière non uniforme par différents matériaux dont la répartition spatiale conduit à la formation d'une certaine topographie. La partie intra-puce de celle-ci sera dépendante de la manière dont les matériaux sont arrangés les uns par rapport aux autres, c'est-à-dire du design de la puce (cf. Chap. 3.2.4 « Modulation de la topographie »). Les critères de sélection des niveaux et les niveaux sur lesquels les extractions ont été réalisés sont décrit en Chap. 4.3.4 « Extraction de densités de design sur les GDS ».

Après mise en forme du fichier de travail pour le logiciel SIMCAP+, on sélectionne les données d'entrée (paramètres sur lesquels sont calculés le modèle) et de sortie (set de calibration du modèle). Le calcul de la régression permet d'obtenir les coefficients nécessaires à la prédiction de la topographie.

4.3.3 Choix du niveau de l'étude

La première étape a été de sélectionner le meilleur candidat pour montrer l'intérêt de la régression PLS pour la modélisation de la topologie intra-champ et éprouver la méthodologie. Cela a été fait avec le masque de développement du 14nm FD-SOI. Le niveau de masque sélectionné devait remplir deux conditions : il devait s'agir d'un niveau critique en termes de contrôle du focus et les performances que l'on pouvait attendre du modèle devaient être les plus élevées possibles. Une étude a été menée sur des plaquettes en 14nm FD-SOI depuis le début de son intégration jusqu'à l'exposition de la première ligne de métal. Les données de leveling du scanner ont été extraites des fichiers journaux de la machines pour plusieurs lots à toutes les étapes critiques de lithographie 193nm à immersion. Les densités surfaciques et périmétriques du design ont été extraites sur les niveaux déjà présents sur le wafer au moment de l'étape de lithographie sans aucune optimisation. Seuls les niveaux de masque suivis d'une gravure ou d'une épitaxie ont été sélectionnés car ils peuvent potentiellement avoir une influence sur la topographie du wafer (cf. Chap. 4.3.2).

Le graph de la figure 4-18 donne les résultats en R^2 de la corrélation entre le modèle PLS calculé à partir des densités de design des niveaux précédents le niveau étudié et les mesures de topographie réalisées par le scanner.

Le niveau Contact (qui correspond à l'exposition des masques CNTA et CNTP) est le meilleur candidat pour l'étude car il est celui dont la mesure de topographie par le scanner est la plus facilement modélisable. De plus, une étude d'uniformité de focus intra-champ sur une FEM (cf. Chap. 2-5 « Fenêtre de procédé ») montre que le niveau CNTA est aussi le niveau dont la criticité en focus est la plus élevée avec seulement une cinquantaine de nanomètres de profondeur de champ disponible.

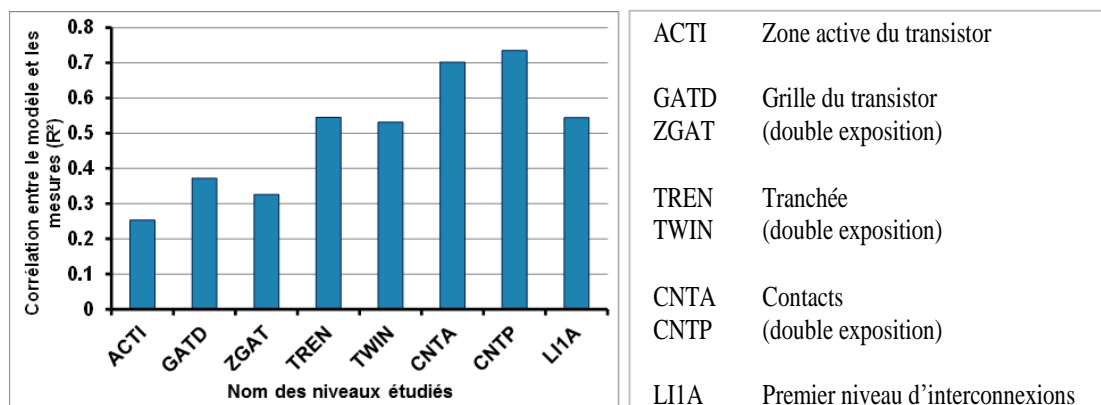


Figure 4-19 : Coefficients de corrélation du modèle de PLS du leveling pour chaque niveau de lithographie 193nm immersion du FEOL, MEOL et pour la première ligne métallique en 14nm FD-SOI.

Dans les résultats de la figure 4-19, on remarque aussi que les binômes GATD-ZGAT, TREN-TWIN et CNTA-CNTP réagissent de manière identique en termes de performance du modèle PLS de la mesure de topographie par le capteur optique du scanner. Ce résultat est parfaitement normal et était attendu car il s'agit ici des masques de la double exposition des niveaux Grille, Tranchée et Contact respectivement.

Aucun procédé susceptible de modifier la topographie n'ayant lieu entre les deux expositions, la topologie du wafer reste donc la même. On attend un résultat similaire à la corrélation trouvée pour LI1A pour le niveau LI1B (non étudié) car ce sont les deux masques complémentaires nécessaires à la réalisation en double exposition du premier niveau d'interconnexion métallique.

Pour ces raisons, le niveau de masque CNTA a été sélectionné pour la suite de cette étude.

4.3.4 Extraction de densités de design sur le GDS

La méthode de prédiction de la topographie sélectionnée dans cette étude est l'utilisation d'un modèle empirique déterminé par régression PLS à partir du design de la puce. Pour cela, nous allons extraire des informations de densité locale du design directement sur le GDS¹⁷ qui est le fichier contenant l'ensemble du design de la puce, niveau par niveau. La Figure 4-20 donne un aperçu d'un design.

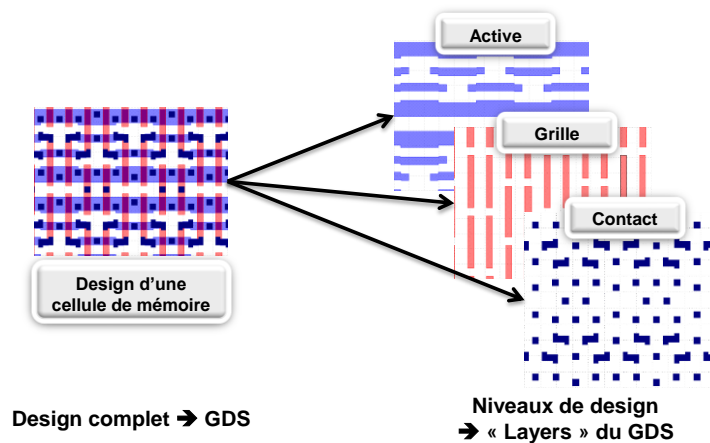


Figure 4-20 : Extrait d'un design sur le logiciel Calibre de chez Mentor Graphics. Le design complet est composé de plusieurs niveaux superposés les uns aux autres

Les extractions de densité sur le GDS ont été réalisées avec le logiciel Calibre édité par Mentor Graphics. Ce logiciel sert à créer le design du circuit, à vérifier les règles de dessin et permet aussi de faire des OPC. La partie vérification de règle de dessin de Calibre utilise le langage SVRF (Standard Verification Rule Format). Les scripts SVRF permettent de réaliser de nombreuses opérations directement sur les GDS.

4.3.4.1 Les différents types de densité

Il existe deux types de densités facilement extractibles avec Calibre. Il s'agit de la densité surfacique et de la densité périmétrique du design [76]. Les deux types de densités sont détaillés dans la suite et illustrés par la Figure 4-21.

¹⁷Le format GDS, ou Grid Design System, est un format de fichier utilisé couramment pour le design des puces électroniques.

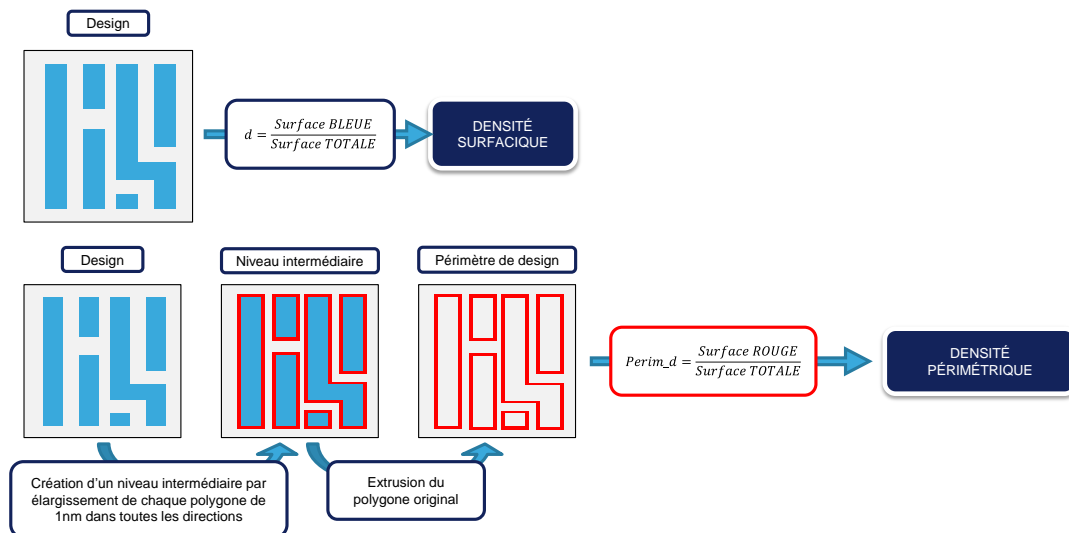


Figure 4-21 ; Définition des deux types de densités extraites du GDS

La densité surfacique locale est simplement le rapport de la surface de motif sur la surface totale du pixel choisi. Elle donne une information sur la quantité d'un matériau par rapport à l'autre sur une surface donnée.

La densité périmétrique est la densité de paroi de motif sur une surface donnée. Les deux types de densité ont un impact sur la répartition spatiale des matériaux. Elle est liée à la taille des motifs. A densité surfacique constante, plus la densité périmétrique est élevée et plus il y a aura de motifs sur une surface donnée et donc plus ces motifs seront petits. Dans le calcul de la densité périmétrique, le redimensionnement de 1nm permet d'obtenir une densité (rapport d'une surface sur une surface) sans pour autant modifier la valeur recherchée qui est celle du périmètre du motif.

4.3.4.2 Manipulation des niveaux de design

Pour estimer la topographie ou à la dépendance du capteur optique du scanner, il est nécessaire que les informations extraites du GDS représentent l'empilement tel qu'on le trouve sur le wafer. Or certains niveaux physiques de matériaux sont des combinaisons de plusieurs niveaux dessinés du GDS. C'est le cas des niveaux en double exposition ou les niveaux d'épitaxie. La connaissance du design et de l'intégration permet de retrouver les niveaux tels qu'ils sont physiquement présents sur le wafer. Le script SVRF propose des opérations de combinaison logique des niveaux de masque du GDS qui permettent de reconstruire ces niveaux et d'extraire la densité directement sur ceux-ci.

4.3.4.2.1 Reconstruction du design

En 14nm FD-SOI, on trouve de nombreux niveaux nécessitant une double exposition et qui apparaissent donc en deux niveaux de design dans le GDS. C'est le cas des tranchées d'isolation entre les transistors (ou Active, noté ZACT), de la grille du transistor (noté ZGAT), de la tranchée (TRENCH), du contact (noté CNT) et des premiers niveaux d'interconnexion métallique (LIN1, 2, 3...). La tranchée est un niveau spécifique au 14nm qui permet de connecter le contact à la source et au drain du transistor. Ce

niveau est nécessaire en raison des dimensions et de la densité de transistor et permet de limiter les risques de court-circuit entre le contact et la grille pendant la gravure.

On peut donc dériver les niveaux suivants en recombinaison des deux masques de lithographie nécessaire à leur réalisation :

$$ZGAT = GATE \text{ NOT } GATD \quad (29)$$

$$ZACT = ACTI \text{ NOT } ACTD \quad (30)$$

$$TRENCH = TREN \text{ NOT } TWIN \quad (31)$$

$$CNT = CNTA \text{ OR } CNTP \quad (32)$$

$$LIN1 = LI1A \text{ OR } LI1B \quad (33)$$

La manière dont ces niveaux se combinent entre eux est illustré dans le tableau de la Figure 4-22.

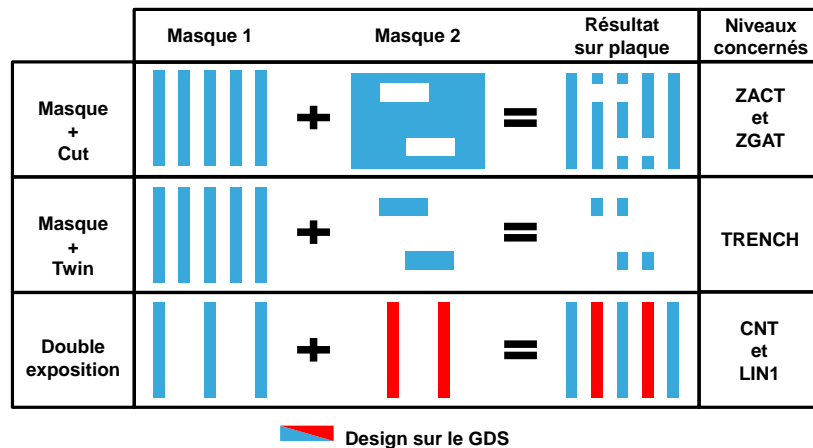


Figure 4-22 : Principes de combinaisons de niveaux de design entre eux

Les niveaux ZACT (zone active du transistor) et ZGAT (grille du transistor) sont construits à partir d'un niveau de base (ACTI et GATE respectivement) et d'un niveau « cut » (ACTD et GATD respectivement) qui permet de définir des zones de coupure de lignes d'active ou de polysilicium de grille pour former plusieurs motifs. Le niveau TWIN de la tranchée est l'inverse d'un « cut » vis-à-vis du niveau TREN. Le contact (CNT) et la ligne de métal 1 (LIN1) sont des niveaux de double exposition plus classiques dans lesquels les deux motifs de la première et de la deuxième exposition sont gardés intégralement pour former le niveau complet.

Deux autres niveaux doivent aussi être recalculés à partir du design avant de pouvoir extraire les densités surfacique et périmétrique qui les caractérisent. Il s'agit des épitaxies de la source et du drain pour les deux types de transistors, NMOS et PMOS. Faire croître par épitaxie la source et le drain est une spécificité du FD-SOI. En effet, la couche d'isolant enterrée dans le substrat interdit la création de ces deux électrodes dans le substrat comme pour les technologies CMOS traditionnelles. Les lithographies

des épitaxies (PEPI et NEPI pour les PMOS et NMOS respectivement) définissent des zones autour de la grille dans laquelle faire croître sélectivement la source et le drain sur les zones actives du transistor uniquement sans croître sur le nitrure qui protège la grille. La zone épitaxiée à la fin du procédé est cette ouverture définie par la lithographie moins la surface de grille qui a aussi été découverte par les P/NEPI comme illustré en Figure 4-23.

$$P/NEPI = P/NEPI \text{ NOT ZGAT} \quad (34)$$

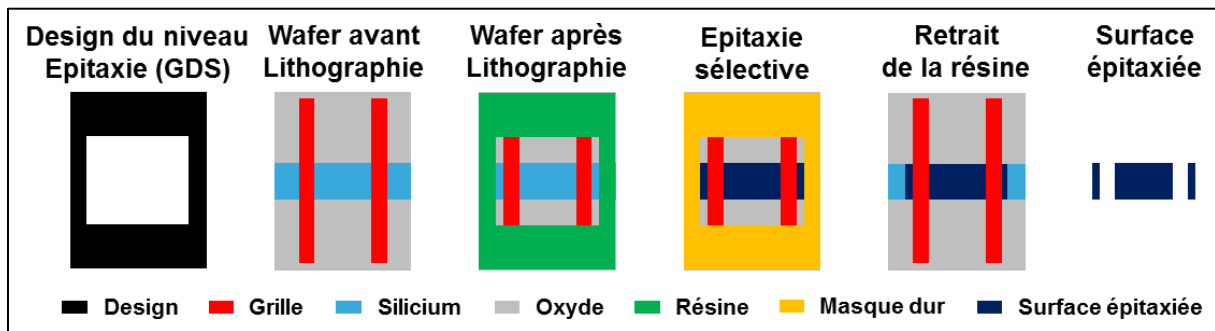


Figure 4-23 : Définition des niveaux d'épitaxies Source / Drain

4.3.4.2.2 Combinaison de niveaux

Toutes les combinaisons de niveaux de design décrites ci-dessus permettent de reconstituer les différentes couches présentes sur le wafer au moment de l'exposition d'un niveau. Il est aussi possible de combiner entre eux des niveaux pour obtenir des motifs non réellement existants sur le wafer et pouvant eux aussi avoir une influence. Il s'agit ici majoritairement d'effets combinés. Le recouvrement de l'active par la grille a été calculé entre autres (comme dans l'exemple du Chap. 4.3.1). Les niveaux de ce type sont notés de la manière suivante : Niveau1_Niveau2.

4.3.5 Le modèle de leveling en 14FD-SOI Contact

Lors de la mesure et la correction de la topographie par le scanner, trois fichiers journaux nous intéressent. Le premier est la mesure brute de la topographie intra-champ par le capteur optique en lumière visible. Le second est la correction de cette mesure calculée à partir de la mesure avec le capteur pneumatique AGILE, appelé PDO pour Process Dependency Offset, traduisant la dépendance à la réflectivité du wafer du capteur en lumière visible. Le troisième set de données est le résiduel non-corrigeable. A partir de la mesure par le capteur optique et de la correction de celle-ci par AGILE, la topographie intra-champ réellement vue par le scanner peut être calculée. La Figure 4-24 donne une comparaison visuelle des données générées par le scanner pendant l'exposition et du masque correspondant.

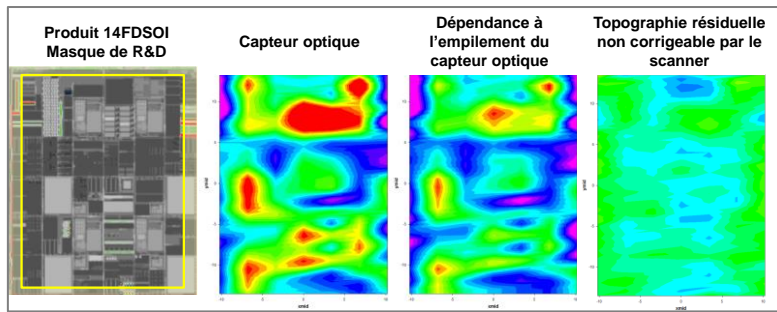


Figure 4-24 : Comparaison de données du leveling du scanner au niveau Contact 14nm FD-SOI et de l'agencement spatial du produit sur le masque. La partie en jaune sur le produit est la zone mesurée par le capteur.

Pour le calcul de ces modèles, les niveaux réels – c'est-à-dire en prenant compte des doubles expositions par exemple – sont calculés et leurs densités sont extraites sur la grille du scanner. Deux itérations sont nécessaires pour construire les modèles. La première consiste en trier les densités utiles au modèle et la deuxième permet d'aboutir au modèle en lui-même.

Pour la première itération, le *VIP* (cf. Chap. 4.1 « *La régression PLS* ») est la métrique qui est utilisée pour caractériser l'influence de chaque paramètre sur le modèle. L'analyse des résultats confirme que l'utilisation de niveaux réels tels que présents sur le wafer est pertinente comparée à l'utilisation des niveaux de masque. Une analyse plus précise des résultats de la métrique est proposée au Chap. 4.3.7 où l'impact de chaque niveau sera comparé entre le modèle du leveling et celui de la topographie haute fréquence (cf. Chap. 4.3.6).

La figure 4-25 donne les performances du modèle PLS pour les données scanner suivantes :

- Moyenne intra-champ de la mesure par le capteur optique
- Moyenne intra-champ de l'erreur de mesure dépendante de l'empilement faite par le capteur optique (déterminé avec le capteur pneumatique Agile)
- Moyenne intra-champ du capteur après correction par Agile
- Moyenne intra-champ du résiduel non corrigé par le scanner

Le modèle décrit très bien la mesure faite par le capteur optique ($R^2=0.79$) mais aussi la dépendance de ce capteur à l'empilement déjà présent sur le wafer ($R^2=0.72$). A une moindre mesure, la topographie à une fréquence millimétrique mesurée par l'utilisation conjointe du capteur optique et du capteur pneumatique est dépendante à environ 50% de la densité du design à cette même échelle. Il est possible de tirer deux conclusions de ce résultat.

Tout d'abord, à l'échelle millimétrique, le design aura plus d'influence sur l'interférence des rayons lumineux du capteur optique en lumière visible, et donc sur la qualité de la mesure en elle-même, que sur la topographie réellement présente dans le champ. Il est intéressant de remarquer que même si le R^2 est bien plus faible dans le cas de la topographie corrigée par Agile que pour le capteur optique seul, l'analyse de la cartographie des valeurs prédites en comparaison avec la mesure montre que la forme

générale est la même. Les zones présentant une topographie fortement négative ou positive (i.e. haute colline ou profonde vallée) sont quand même détectées par le modèle, cependant l'amplitude de la topographie est largement sous-estimée.

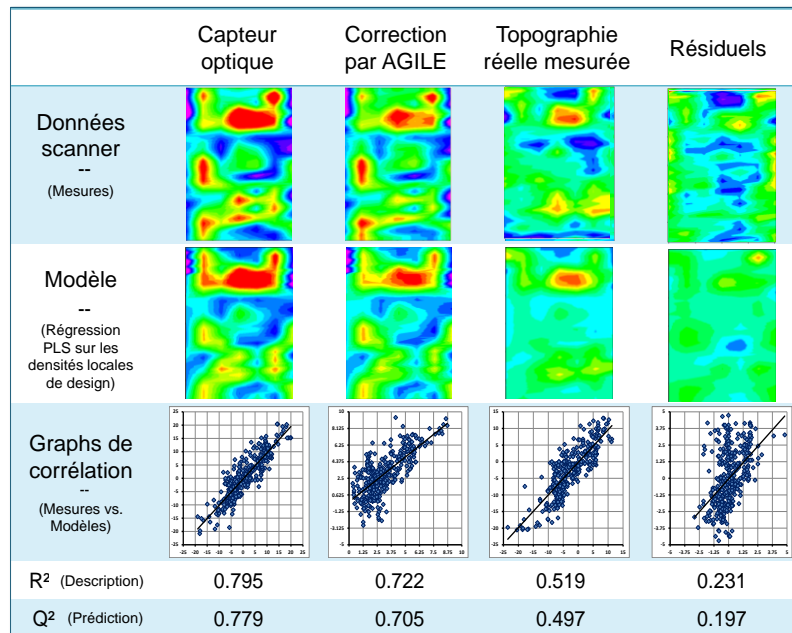


Figure 4-25 : Performances respectives des différents modèles du leveling [32]

Enfin, la topographie résiduelle après correction par le scanner n'est pas du tout dépendante du design à cette échelle de résolution spatiale. Ce résultat était attendu dans le sens où ces données contiennent en plus les capacités de correction du scanner qui sont indépendantes de l'intégration. Ce résiduel est parfaitement calculable à partir du modèle de topographie réelle mesurée. Il suffit pour cela de soustraire la surface que le scanner peut suivre mécaniquement en bougeant le chuck pendant l'exposition.

Avec ce premier modèle, il a été montré qu'il était possible d'utiliser la méthode développée pour obtenir un modèle de la topographie intra-champ performant. Cependant, même si le scanner ne peut pas intégralement corriger cette topographie intra-champ basse fréquence, la mesure ne prend pas du tout en compte la topographie haute fréquence et gomme une bonne partie des variations de surface. Il paraît alors intéressant de pouvoir modéliser la topographie à une échelle plus proche du design.

4.3.6 Le modèle de topographie haute fréquence

En CMP, les distances moyennes pour les effets spatiaux sont de l'ordre de 100 à 200 μm c'est-à-dire que les variations de topologie de surface vont s'établir à cette échelle. Avec les mesures Wyko et PWG, nous avons pu collecter des données à des échelles équivalentes voire inférieures. En corrélant les densités de design avec les données de topographie haute fréquence, il devrait être possible de tenir compte de ces effets de CMP moyenne fréquence. Comme précisé plus haut, les mesures Wyko et PWG donnent des résultats équivalents. Le pixel offert par le Wyko étant plus petit, ce sont ces mesures qui

vont être utilisées pour la suite car elles offrent plus de souplesse sur le choix de la taille du pixel permettant la meilleure prédictibilité.

Dans cette étude, plusieurs approches ont été abordées telles que décrites dans le tableau 4-4.

Méthode	Principe	Intérêts	Inconvénients
Modèle simple	Balayage de plusieurs tailles de pixels entre 2.5 et 100 μ m	Choix de la taille de pixel pour avoir le meilleur modèle simple	Risque de négliger des effets à courte ou longue distance selon le pixel
Modèle combiné	Combinaison linéaire haute fréquence de plusieurs modèles simples à différentes résolutions latérales	Modèle plus complet	Nécessité de faire 2 à 4 régressions PLS

Tableau 4-4 : Comparaison des avantages et inconvénients des deux méthodes de modélisation haute fréquence de la topographie avec la régression PLS

Pour chacun de ces modèles, les mêmes niveaux réels que ceux qui ont été déterminés comme influents et importants dans le modèle des données de correction de topographie par le scanner ont été utilisés. Seuls les niveaux d'épitaxie des source/drain ont été ajoutés au nombre de paramètre d'entrée. La densité du design est extraite du GDS directement à la taille de pixel étudiée et le logiciel Gwyddion [53] est utilisé pour interpoler la topographie intra-champ sur la nouvelle grille. Dans chaque cas, l'algorithme choisi est une interpolation de Schaum du 4^{ème} ordre. Les mesures de topographie réalisées sur le Wyko du LETI sont celles qui sont prises comme référence pour créer le modèle. En premier lieu, un modèle a été créé sur les puces du masque de R&D du 14nm FD-SOI dont le design est similaire à celui d'un produit client (cf. Chap. 4.3.6.1.1). Les puces et structures de test non standards ont été étudiés ensuite (cf. Chap. 4.3.6.1.2). Enfin le principe modèle combiné sera abordé (cf. Chap. 4.3.6.2).

4.3.6.1 Modèle simple

4.3.6.1.1 Puces CMOS

L'objectif ici est de développer un modèle qui puisse être applicable en production. Le modèle a été construit en priorité sur les puces CMOS.

Dans un MPW (cf. Chap. 3.2.2 « *Assemblage* ») sont assemblées de nombreuses puces différentes, certaines étant des structures de test pour supporter le développement technologique et d'autre étant des puces fonctionnelles, soit pour le développement soit pour le prototypage. Après analyse du MPW en 14nm FD-SOI, deux puces de développement ont un design proche de puces client. L'une d'entre elles (cf. Figure 4-26) nommée PROLIGHT (d'une surface de 3.2mm²) a été sélectionnée pour construire le modèle et celui-ci a été validé sur l'autre puce de test appelée PROMO (d'une surface de 20mm²).

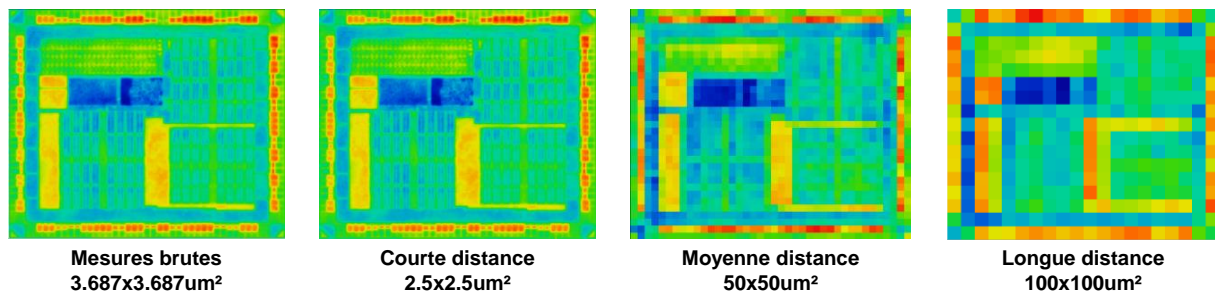


Figure 4-26 : Quelques ré-échantillonnages des mesures Wyko de la puce de calibrage du modèle (PROLIGHT) avec Gwyddion

Douze tailles de pixels différentes ont été testées afin de trouver la résolution à laquelle la topographie était la plus facilement et correctement prédite. Pour chacune, la régression PLS a été calculée sur les données de calibration et les coefficients ont été appliqués aux valeurs de densité de la puce de validation. Les résultats en termes de performances de prédiction sont tracés par taille de pixel dans le graphique de la Figure 4-27.

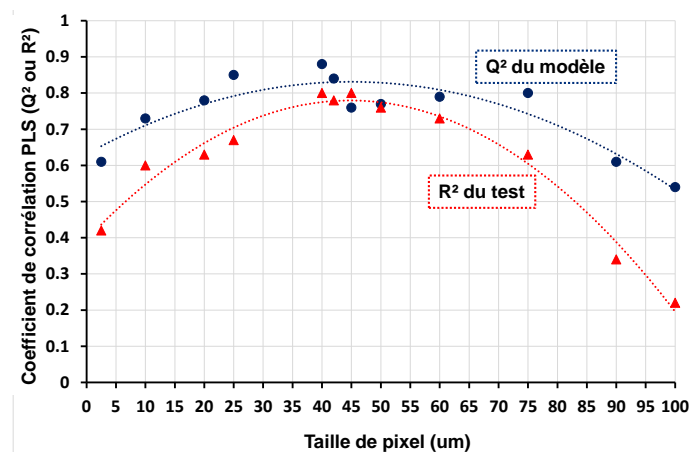


Figure 4-27 : Performances des modèles issus du balayage de différentes échelles spatiales sur le 14nm FD-SOI

Deux résultats se démarquent. Tout d'abord, le R^2 obtenu lors du test du modèle sur les données de validation est toujours plus bas que le Q^2 (cf. Chap. 4.1 « La régression PLS ») qui donne la performance maximale attendue du modèle. Ensuite, à la fois le R^2 et le Q^2 suivent une variation quadratique en fonction de la taille du pixel et passent tous deux par un maximum pour un pixel de $45\mu\text{m}$, qui est donc l'optimum de taille de pixel d'un point de vue corrélation entre la prédiction et les mesures.

Pour juger de la qualité de la prédiction, il est nécessaire de regarder d'autres paramètres. La pente entre les valeurs prédites et les mesures fournit une indication sur la capacité du modèle à évaluer l'impact de chaque empilement local sur la création de topographie. Le rapport entre l'amplitude des valeurs prédites et l'amplitude des mesures, indépendamment de la pente, traduit la capacité du modèle à prédire les valeurs extrêmes de topographie. Le centrage des valeurs est aussi important.

De manière générale, le centrage des valeurs calculées avec le modèle est bon car la méthode PLS fonctionne par normalisation.

Concernant la pente entre le modèle et la mesure (cf. Figure 4-28), aucun des modèles ne donne une pente de 1 alors que lors de la création du modèle sur le set de calibration où les coefficients sont calculés pour avoir une pente de 1. Cet écart de prédiction est très certainement dû à des effets de topographie qui apparaissent à une échelle autre que celle du modèle. On remarque aussi que contrairement aux R^2/Q^2 , la pente varie de manière non régulière selon la taille de pixel.

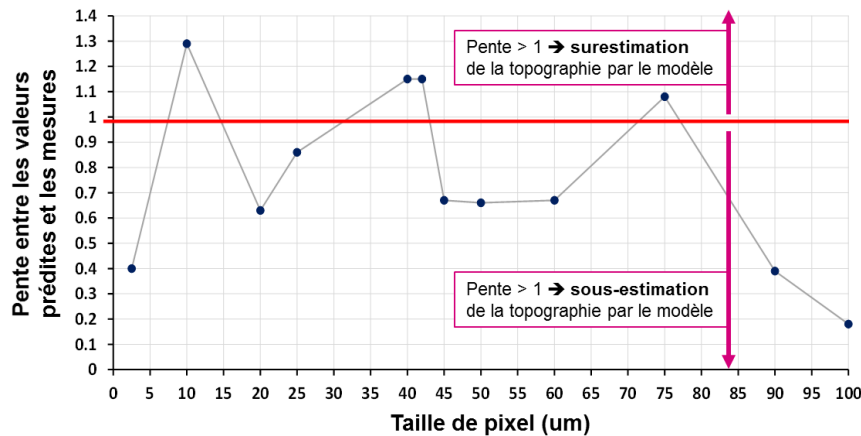


Figure 4-28 : Pente de la corrélation entre le modèle et les mesures en fonction de la taille du pixel

Comme il a déjà été discuté dans le Chap. 4.3.5 « Le modèle de leveling en 14nm FD-SOI Contact », l'utilisation du VIP permet de ne sélectionner que les paramètres d'entrée dont l'impact est suffisant pour être significatif. L'autre avantage est de pouvoir diminuer le temps d'extraction des densités depuis le GDS. Un modèle est plus robuste dans le cas où la suppression des paramètres négligeables ne détériore pas les performances de prédiction.

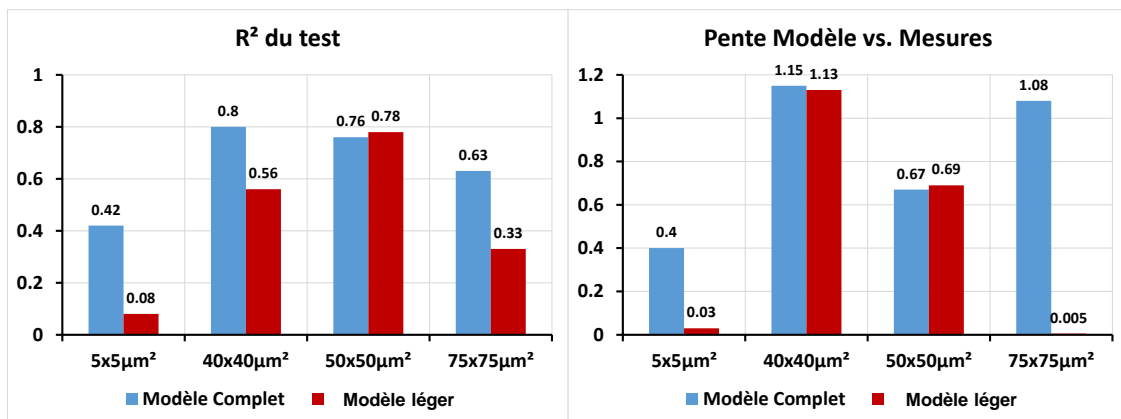


Figure 4-29 : Comparaison des coefficients et pentes de corrélation pour les modèles avant (« complet ») et après (« léger ») tri des paramètres en fonction de leurs VIP respectifs à plusieurs échelles spatiales.

Dans chacun des cas de la figure 4-29, le tri des paramètres à l'aide du VIP a permis de réduire de 18 à 8 le nombre de composantes du modèle. Le modèle dit « complet » est le modèle avec 18 composantes et le modèle « léger » est construits sur les 8 composantes les plus importantes après analyse des VIP. Seul le modèle basé sur un pixel de 50x50um² ne voit ni le R² du test ni la pente être modifiée par la réduction du nombre de composantes. Il s'agit donc du modèle le plus stable et le plus intéressant en vue d'une réduction du temps de calcul.

En analysant les temps nécessaires à l'extraction de la densité avec Calibre, on remarque deux choses en fixant les conditions d'extraction (nombre de processeurs et de threads, quantité de RAM disponible et interdiction de swapper la mémoire). La réduction du nombre de composantes nécessaires à l'élaboration du modèle de 18 à 8 diminue de 30% le temps de calcul pour une surface de puce donnée et la variation est logarithmique en fonction de la taille de la puce. Dans les deux cas, les temps de calcul sont les mêmes quelle que soit la taille du pixel d'extraction. Ces résultats sont détaillés en Figure 4-30.

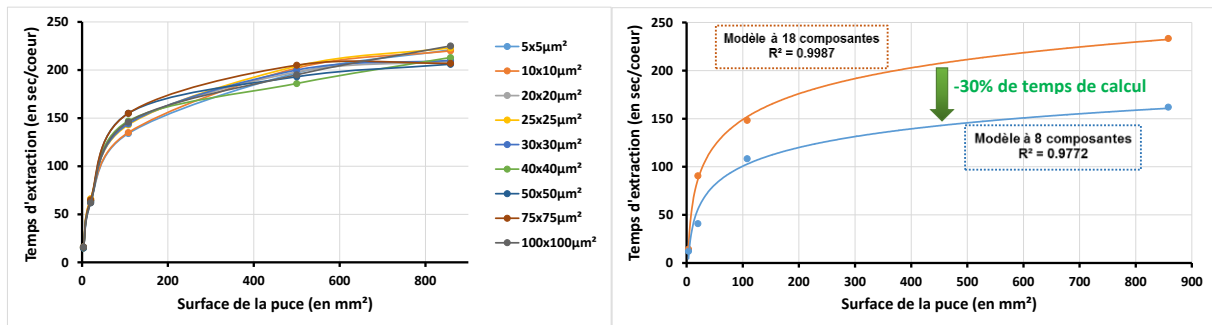


Figure 4-30 : Temps d'extraction des densités de design avec Calibre en fonction de la surface de la puce. A gauche, pour différentes tailles de pixels et à droite, en modifiant le nombre de composantes pour la PLS.

Dans la suite, un pixel de 50µm avec un modèle PLS à 8 composantes a été choisi. Ce choix se justifie pour plusieurs raisons. Tout d'abord, les résultats ci-dessus montrent une plus grande stabilité du modèle en prédiction avec moins de variabilité des valeurs en optimisant le temps de calcul. De plus, quand on analyse l'écart entre le Q^2 calculé par la méthode PLS et le R^2 de la validation, c'est dans ce cas que l'on a le moins de changement entre ce qui est attendu et ce qu'on obtient en réalité. Ce modèle est donc plus robuste.

Le pixel de 50µm est aussi celui qui permet de résoudre le mieux les effets visibles sur le wafer. Les effets CMP locaux se traduisent généralement à une échelle de 100 à 200µm environ. Avec un modèle à 100µm de résolution, on ne représentera ces effets qu'à l'aide de quelques pixels alors que le modèle à 50µm permet d'utiliser jusqu'à 16 pixels pour représenter la même zone. Un pixel trop gros conduit aussi, lors du calcul des valeurs de topographie sur la nouvelle grille à l'aide de Gwyddion, à une sous-estimation de l'amplitude totale de cette topographie.

D'un point de vue pratique, il est aussi plus simple de faire coïncider les grilles d'échantillonnage des mesures de topographie et des extractions de densités avec un pixel de 50µm. En particulier, le ré-échantillonnage des mesures Wyko à l'aide de Gwyddion crée un léger décalage de centrage des pixels qui dépend de la taille de ceux-ci. Le décalage est moindre en proportion pour les pixels de 50µm que pour ceux de 45µm par exemple, pour lequel le modèle est aussi très performant.

4.3.6.1.2 Puces non CMOS

Sur le masque de développement du 14nm FD-SOI, il y a de nombreuses autres puces dont le design n'est pas de type CMOS (comme des composantes analogiques par exemple) voire totalement inhabituel

pour des structures spécifiques nécessaires au développement de certains procédés et à des caractérisations physique, chimique ou électrique. Il est intéressant de savoir quelle est la limite du modèle en termes d'applicabilité sur un masque. Dans cette partie, 5 puces différentes du masque ont été choisies. Les critères de sélections étaient d'avoir des puces dont la fonctionnalité ou l'objectif pour le développement soient différents mais aussi que la topographie mesurée au niveau de ces puces avec le Wyko présente une grande variation par rapport au reste du masque.

Les puces sélectionnées sont les suivantes et les mesures Wyko desdites puces sont données en Figure 4-31 :

- PROLIGHT : petite puce CMOS de 3.2mm² sélectionnée pour la calibration du modèle dans la partie précédente
- PROMO : puce CMOS de 20mm² utilisée précédemment pour la validation du modèle PROLIGHT
- FE_RX : structure de test pour la CMP avec des variations de densité croisées entre les différents niveaux subissant un procédé de CMP
- ESDRF : structure de test de protection ESD (Electrostatic Discharge ou décharge électrostatique) présentant un design analogique
- SIMS : structure design pour les mesures SIMS (Secondary ion mass spectrometry ou Spectromètre de masse à ionisation secondaire) composé de larges pavés sans motifs (300µm de coté) dans lequel les niveaux de masque ne sont pas forcément tous représentés.

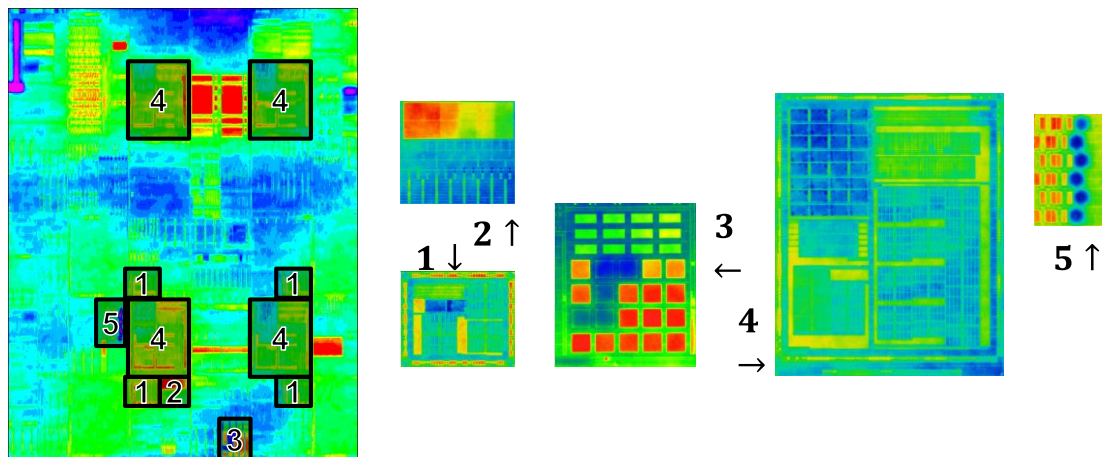


Figure 4-31 : Les différentes puces du masque 14nm FD-SOI utilisées pour la construction du modèle (échelle X10) et leurs positions respectives dans le champ du MPW. (1) Prolight, (2) FE_RX, (3) SIMS, (4) Promo, (5) ESDRF.

Des modèles PLS à 8 composantes avec un pixel de 50x50µm² ont été construits sur le PROLIGHT, le SIMS et l'ESDRF puis testés sur chacune des cinq puces. Le tableau 4-5 présente les R² obtenus entre l'application du modèle et les mesures de référence pour chaque cas.

Méthode	Puce DOE		Puce CMOS		Puce Analogique
	SIMS	FE_RX	Prolight	Promo	ESDRF
Modèle SIMS	0.89	0.89	0.04	<0.01	0.30
Modèle Prolight	0.11	0.01	0.70	0.75	0.26
Modèle ESDRF	0.07	0.09	0.05	0.06	0.54

Tableau 4-5 : Résultats R^2 des différents modèles

On remarque que le modèle construit sur le Prolight ne s'applique correctement que sur les puces CMOS, que le modèle ESDRF ne s'applique qu'au design analogique et que le modèle SIMS n'est performant que pour les puces « DOE » (Design of Experiment ou Plan d'expérience) au sein desquelles plusieurs situations croisées sont représentées sans pour autant correspondre à des cas réellement existants sur un produit client.

Dans le cas de la puce ESD, le coefficient de corrélation est moins performant si on utilise le modèle PLS basé sur les données ESDRF que les autres modèles appliqués à leurs propres données de calibrage. Cela vient d'une des limites du modèle. En effet, le modèle PLS sur les densités de design ne prend en compte que la combinaison locale de densité entre les différents niveaux de l'empilement. Cela est dû à la méthode PLS qui construit un modèle à partir des corrélations point à point entre les densités de design et les mesures de topographie de calibration. Aucune donnée mécanique ou chimique spécifique aux matériaux et aux procédés utilisés (particulièrement la CMP dans ce cas) n'est prise en compte ici. Ainsi, des effets comme le dishing (cf. Chap. 3.3.4 « Modulation de la topographie ») ne peuvent pas être pris en compte dans le modèle. Ces effets apparaissent en effet dans des zones dans lesquelles la densité est uniforme, ce qui pour le modèle ne donnera qu'une seule valeur de topographie prédite.

Cette limite de la méthode PLS est facilement illustrée en appliquant le modèle Prolight avec un pixel de $50 \times 50 \mu\text{m}^2$ (pour les designs CMOS) sur une chaîne de contacts. Cette structure en maillons a déjà été utilisée précédemment pour montrer l'impact de la topographie sur le focus d'exposition (cf. Chap. 3.4). Sur la zone sélectionnée pour l'application du modèle (cf. Figure 4-32), se trouve à gauche les maillons sujets à un fort dishing (de l'ordre de 45nm) et l'environnement immédiat à droite qui est relativement plat est pris comme référence.

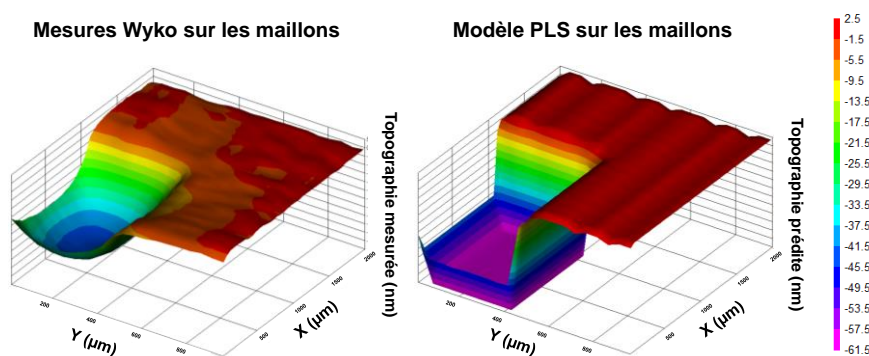


Figure 4-32 : Comparaison de la topographie mesurée et de la topographie modélisée dans les zones de fort dishing

Le modèle surestime grandement la topographie dans la zone des maillons avec une vallée très abrupte plongeant à -64nm par rapport à la zone de référence soit une vallée profonde de près de 20nm de plus que la valeur fournie par la mesure. De plus, le modèle calculant point par point la topographie sans être influencée par le voisinage, la topographie prédite présente une marche de 60nm de topographie entre l'environnement des maillons et la structure de test alors que le dishing réellement présent sur la plaquette est plutôt parabolique avec une pente plus douce.

Cette limite du modèle peut être contournée en ajoutant une deuxième couche au modèle. Le modèle PLS serait alors calculé comme précédemment sur les densités du design puis un lissage serait appliqué par-dessus pour mieux représenter le dishing. Il s'agit ici d'une suggestion qui n'a pas été étudiée [77] faute de temps mais aussi car le modèle simple permet déjà d'obtenir de très bonnes performances en termes de prédiction.

4.3.6.2 Modèle combiné

Le modèle à $50 \times 50 \mu\text{m}^2$ de résolution fonctionne plutôt bien mais certains effets ne sont pas pris en compte. C'est le cas de effets de design locaux très haute fréquence et les effets longue distance.

La PLS ne peut pas prendre en compte en même temps plusieurs échelles de résolution différentes car la méthode impose un calcul point à point. La solution est de combiner plusieurs modèles à différentes échelles pour tenter d'améliorer la prédictibilité.

Pour cette analyse, les modèles à $2.5 \times 2.5 \mu\text{m}^2$, $50 \times 50 \mu\text{m}^2$ et $100 \times 100 \mu\text{m}^2$ construits sur le Prolight ont été sélectionnés et combinés entre eux sur la grille la plus fine, c'est-à-dire la grille à $2.5 \mu\text{m}$. Le test a été réalisé sur le PROMO. La Figure 4-33 donne la mesure de la topographie du PROMO avec le Wyko ainsi que les résultats de l'application des trois modèles PLS. Deux méthodes ont été explorées. La première consiste à simplement faire la moyenne des valeurs prédites par les trois modèles pour chaque position. La deuxième consiste à utiliser la méthode PLS sur les trois sets de valeurs modélisées et la mesure de topographie très haute fréquence pour déterminer la meilleure pondération pour chaque modèle.

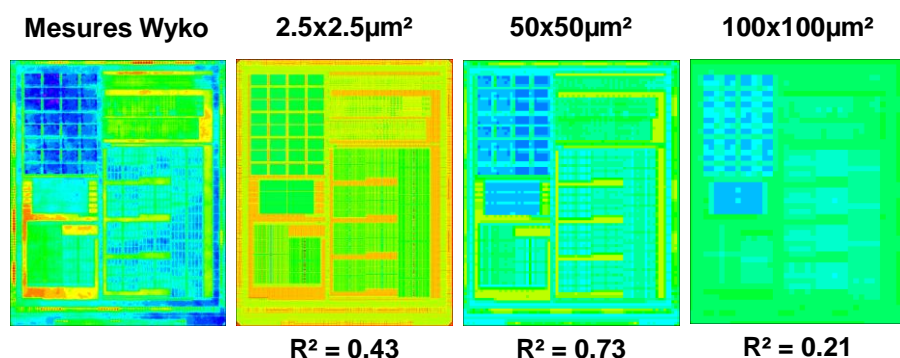


Figure 4-33 : Cartographies de la topographie mesurée avec le Wyko sur le PROMO en 14nm FD-SOI et de la topographie prédite par les modèles PLS à 2.5 , 50 et $100 \mu\text{m}$ de résolution spatiale pour la même puce

Les conditions de combinaison des modèles entre eux et leurs résultats sont présentés dans le Tableau 4-6.

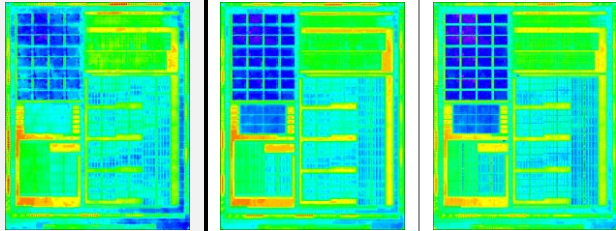
		Modèle non pondéré	Modèle pondéré
Coefficients de pondération	PLS2.5	1/3	0.37
	PLS50	1/3	0.59
	PLS100	1/3	0.04
R ² du test		0.76	0.80
			
Mesures Wyko			

Tableau 4-6 : Coefficients et résultats des modèles combinés pondéré et non pondéré

Le coefficient de la PLS à 100 μ m est très faible (0.04) par rapport à ceux des autres PLS (> 0.35). Pour gagner en temps de calcul et simplifier le modèle, une combinaison des modèles courte et moyenne distance a été calculée en négligeant les effets à très longues distances. Ainsi, un coefficient de 0.4 pour la PLS 2.5 et un coefficient de 0.6 pour la PLS à 50 μ m ont été déterminés comme l'idéal de la combinaison simplifiée. Le résultat reste très similaire à la PLS pondéré à trois composantes puisque que la pente est de environ 1 et le R² du test de 0.79.

Il est donc possible de représenter en même temps les effets à courte et moyenne distance en combinant deux modèles à 2.5 et 50 μ m de taille de pixel pour gagner en résolution et en performance de prédiction. Ici, le gain est de 5% de prédictibilité au prix d'un plus grand temps d'extraction (extraction de la densité à deux ou trois échelles différentes) et de trois (ou quatre) PLS au lieu d'une seule par rapport au modèle simple à une seule fréquence spatiale.

Le modèle combiné a été testé sur la chaîne de contact sujette à un fort dishing (cf. Chap. 4.3.6.1.2) mais il n'a pas permis de retrouver la topographie du wafer via le modèle. Le modèle combiné est donc plus performant que le modèle simple mais présente les mêmes défauts. On pouvait s'y attendre car une zone de dishing est une zone dans laquelle le design est homogène ce qui pour le modèle PLS, qu'il soit simple ou combiné, correspond à une zone de topographie homogène.

Dans la suite, le modèle simple (c'est-à-dire sans combinaison de plusieurs pixels) sera pris comme référence en raison du rapport plus intéressant qu'il présente entre le temps d'extraction des densités du GDS vis-à-vis de la performance du modèle.

4.3.7 Analyse des VIP

Lors de la création des modèles pour le leveling, i.e. topographie basse fréquence, et les mesures Wyko, i.e. topographie haute fréquence, le *VIP* (Variable Importance in the Projection, cf. Chap. 4.2.3) a permis de faire un tri entre les extractions du GDS utiles pour un modèle PLS performant et celles qui n'apportaient rien d'un point de vue prédictibilité du modèle. Dans cette partie du chapitre, les *VIP* des différentes extractions vont être comparées niveaux par niveaux afin de tirer les impacts de chacun d'entre eux en terme de topographie sur le wafer et de réflectivité pour le capteur optique du scanner. Pour le modèle du leveling, la mesure par le capteur optique, l'erreur de cette mesure et la topographie après correction de cette mesure par le capteur pneumatique sont ici considérées. Pour le modèle haute fréquence, le modèle simple à 50 μ m de résolution spatiale est celui qui est étudié.

Comme il a déjà été précisé dans la partie sur la création des modèles, les niveaux d'épitaxie P/NEPI étaient disponibles pour le modèle de topographie haute fréquence mais pas pour le modèle du leveling du scanner. Ainsi, s'il est possible de comparer les effets des autres niveaux entre eux, aucune donnée de comparaison n'est disponible pour les épitaxies.

Les résultats de l'analyse *VIP* sont donnés dans le Tableau 4-7.

Machine	Scanner		Scanner		Scanner		Scanner		Wyko	
Résolution du modèle	2.5mm		2.5mm		2.5mm		2.5mm		50 μ m	
Ordre d'importance	Capteur optique		Erreur de mesure		Topographie basse fréquence		Résiduels du leveling		Topographie haute fréquence	
1	NOSO	P	NOSO	P	NOSO	D	NOSO	P	PEPI	P
2	GATE	D	GATE on ACTIVE	D	NOSO	P	GATE on ACTIVE	P	GATE	P
3	NOSO	D	GATE	D	TRENCH	D	NOSO	D	NEPI	P
4	GATE on ACTIVE	D	ACTI	D	GATE	D	GATE on ACTIVE	D	NOSO	D
5	ACTI	D	GATE on ACTIVE	P	ACTI	P	GATE	D	TRENCH	P
6	GATE on ACTIVE	P	NOSO	D	ACTIVE on NOSO	P	ACTI	D	ACTIVE on NOSO	P
7					GATE on ACTIVE	D	ACTIVE on NOSO	D	ACTI	P
8									GATE on ACTIVE	D
9									ACTIVE on NOSO	D

Tableau 4-7 : Classement des niveaux de design et des types de densité les plus influents sur le modèle du Contact en 14nm FD-SOI

Tout d'abord, on remarque que les densités surfaciques de design (D) ont plus d'influence sur les mesures de topographie basse fréquence et les densités périmétriques (P) sont plus importantes pour la topographie haute fréquence. Cela peut s'expliquer par un effet de zone. Sur une mesure comme celle du scanner qui moyenne la topographie sur plusieurs millimètres carrés par point de mesure, c'est la

densité moyenne sur une large zone qui aura une influence. Sur une mesure locale, les transitions d'un empilement à l'autre seront plus visibles par la mesure qui moyennera moins la topographie.

Le niveau de masque NOSO apparaît quasiment systématiquement comme l'un des quatre paramètres les plus importants que ce soit pour la topographie ou pour les modèles de l'erreur de mesure. Ce niveau est spécifique de la technologie FD-SOI. Il s'agit de créer des composants au sein du design pour lesquels il n'y aura pas de box d'oxyde enterré alors que le reste des composants du circuit auront ce box. NOSO signifie No-SOI (pour « pas SOI »). Il est le tout premier niveau de masque manufacturé et par construction, la création du NOSO crée deux substrats complètement différents sur une même plaquette. Son influence très importante sur l'erreur de mesure optique de la topographie par le scanner est alors évidente puisque ces deux substrats ne vont pas réfléchir la lumière de la même manière. Pour réaliser ce NOSO, le box, déjà présent sur la plaquette de silicium avant le début de l'intégration, est gravé, ce qui crée des creux de plusieurs dizaines de nanomètres dans le wafer. Pour éviter que cette importante topographie locale ne détériore fortement l'intégration, les zones NOSO sont remplies par épitaxie mais une marche de quelques nanomètres reste visible en haut de ces zones, créant une topographie initiale qui va se transférer sur l'ensemble de la fabrication de la puce. C'est pourquoi le NOSO a une influence sur la topographie basse et haute fréquence mais que son effet est alors plus léger.

4.3.8 Application à un masque de production en 28nm FD-SOI

Le produit 28nm qui a été mesuré avec le Wyko a lui aussi été modélisé avec la méthode PLS. Le niveau étudié est le premier niveau d'interconnexions. Le produit sélectionné n'est pas un produit FD-SOI (il n'y a donc pas de box oxyde sur le substrat ni de niveau NOSO ni d'épitaxies source / drain). Les niveaux qui peuvent influencer la topographie sont moins nombreux : les tranchées d'isolation, la grille, le contact.

Les résultats sont très similaires à ceux obtenus sur le 14nm FD-SOI comme le montre la Figure 4-34.

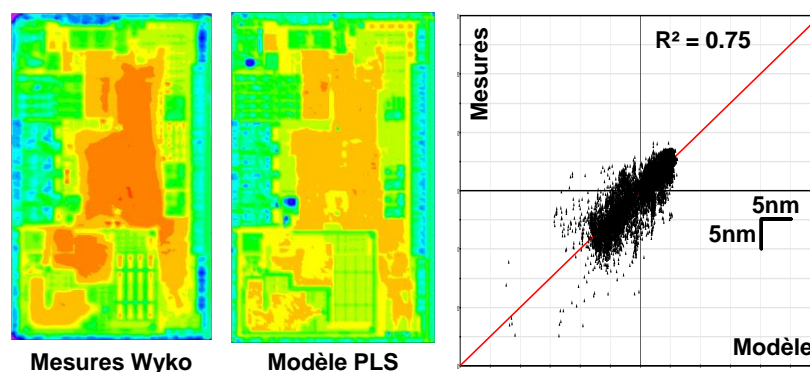


Figure 4-34 : Cartographie des mesures et du modèle PLS pour le produit 28nm BEOL et corrélation entre le modèle et les mesures

En ce qui concerne l'importance de chaque niveau de masque sur la construction de la topographie, l'analyse des *VIP* ne donne seulement que trois densités extraites qui auraient une influence notable.

Ces niveaux sont détaillés en Figure 4-35. Toutes ces trois extractions de densité locale sont de la densité périmétrique, ce qui est consistant avec ce qui a été observé sur le modèle du 14nm FD-SOI. L'étude ayant montré que le périmètre avait plus d'influence sur la topographie haute fréquence que la densité surfacique.

Machine	Wyko	
Résolution du modèle	50 μ m	
Ordre d'importance	Topographie haute fréquence	
1	GATE on ACTIVE	P
2	GATE	P
3	ACTIVE	P

Tableau 4-8 : Classement des niveaux de design et des types de densité les plus influents sur le modèle en BEOL 28nm

L'exercice n'a malheureusement pas pu être réalisé sur un produit 28nm FD-SOI avec un niveau NOSO et des épitaxies source/drain mais on peut s'attendre à une forte influence de ces niveaux de design en interpolant les résultats et les mécanismes observés avec le produit 14nm FD-SOI. Le NOSO devrait avoir un effet particulièrement élevé car contrairement au 14nm FD-SOI, l'intégration en 28nm FD-SOI ne propose pas d'étape de remplissage des zones NOSO qui permettent de limiter la hauteur de la marche entre substrat SOI et substrat No-SOI.

4.4 CONCLUSION

Ce chapitre a présenté les mesures de la topographie qui ont été réalisées sur des plaquettes des plaquettes de silicium et leur analyse ainsi qu'une méthode statistique de modélisation de la topographie.

La topographie mesurée au niveau Contact sur la technologie 14nm FD-SOI présente une amplitude de 90nm environ pour le champ complet. Localement les puces de test dont le design est similaire à un produit CMOS présentent une topographie de 30nm d'amplitude non corrigeable par le scanner. Pour le BEOL 28nm, la topographie intra-champ mesurée avant exposition est de 40nm d'amplitude dont 30nm de variations intra-puce.

La méthode PLS s'est avérée être très adaptée et performante pour modéliser de manière empirique la topographie intra-puce avant exposition. Sur les deux technologies sur laquelle cette méthode a été testée, les capacités de prédiction du modèle ont été quantifiées à un R^2 pouvant atteindre 0.80 et une pente proche de 1 entre le modèle et les mesures de validation.

Cette méthode permet aussi, à travers plusieurs régressions différentes, de prédire la topographie à plusieurs résolutions spatiales, notamment à 2.5 μ m (haute fréquence), 50 μ m (moyenne fréquence) et à

l'échelle du millimètre (basse fréquence) qui est l'échelle à laquelle le scanner mesure et corrige cette topographie.

La méthode développée ici ne permet pas de tester plusieurs conditions de procédé pour faire une première sélection des chimies, pads et designs les plus adaptés pour les étapes d'aplanissement contrairement à d'autres outils de modélisation de la topographie. En revanche, pour un procédé unique déjà robuste, elle est parfaitement adaptée et permet de prédire la réaction du design à ce procédé. De plus, contrairement aux logiciels de simulation de CMP existants dans le commerce, la topographie que l'on modélise par PLS n'est pas uniquement la topographie post-CMP mais la topographie à n'importe quelle étape de l'intégration du moment que l'on puisse mesurer la plaquette pour calibrer le modèle.

A partir de la mesure ou de la prédiction de la topographie, il est possible d'alimenter les besoins du contrôle de procédé, des designers, de la métrologie, du scanner ou encore des outils de prédiction de défauts afin de co-optimiser procédés et design des puces en vue d'un meilleur contrôle du focus. Plusieurs optimisations sont proposées dans le Chap. 5 de ce manuscrit.

CHAPITRE 5

5 APPLICATIONS ET PERSPECTIVES D'OPTIMISATIONS

Le travail réalisé au cours de cette thèse ouvre la porte à de nombreuses possibilités d'optimisation et de travail futur dans le domaine de la réduction de la variabilité focus et du contrôle du focus. La figure 5-1 résume les différentes possibilités qui seront développées brièvement dans la suite.

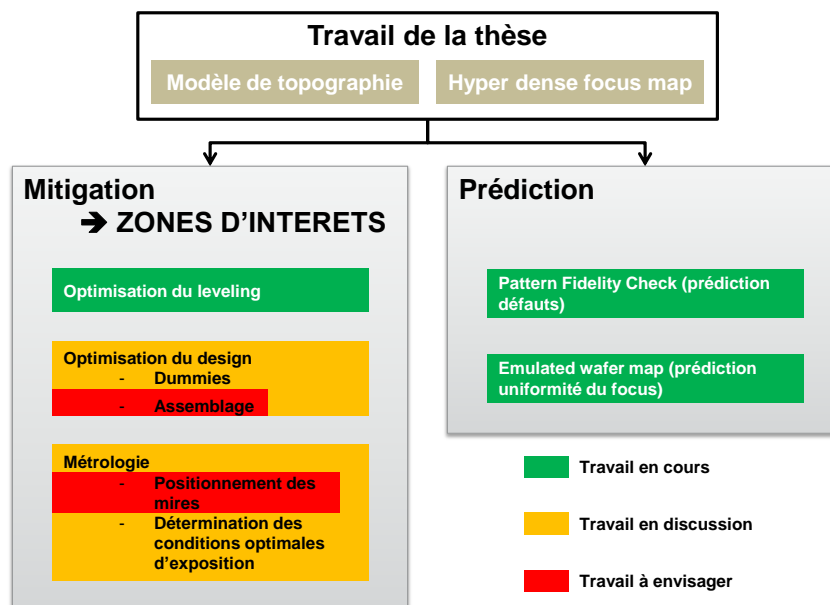


Figure 5-1 : Les différentes optimisations proposées à la suite de la thèse

5.1 DEFINITION DES ZONES D'INTERETS

L'architecture du circuit et l'assemblage (cf. Chap. 3.2) créent un design hétérogène au niveau de la puce et du masque respectivement. D'un point de vue lithographique, cela revient à faire cohabiter des zones sensibles au focus, présentant ou non une topographie, avec des réflectivités différentes, ... sur le même champ d'exposition. Si on retrouve par exemple un motif dont la profondeur de champ d'impression est très faible dans une zone dont la topographie est très différente du reste de la puce et qui ne serait pas corrigable par le scanner, on a un pire cas dans lequel on peut s'attendre à avoir des pertes de rendement systématiques en raison de défauts d'impression en photolithographie (cf. Figure 5-2).

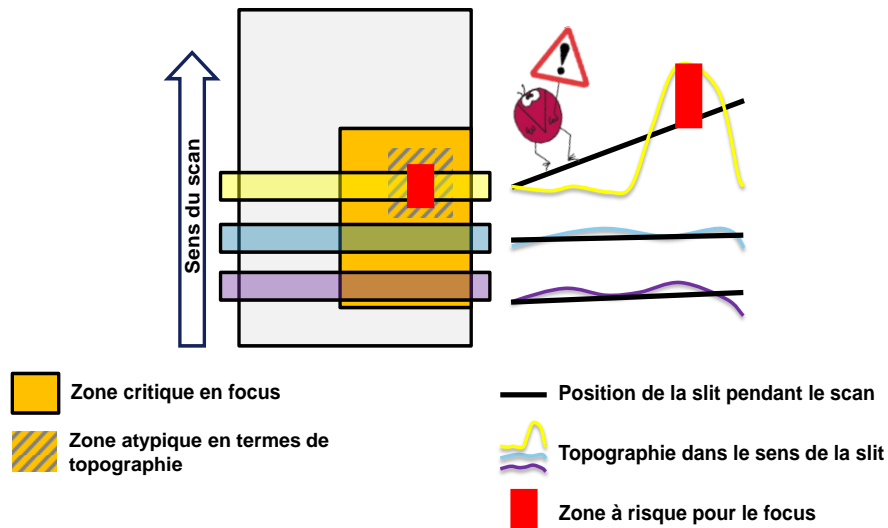


Figure 5-2 : Risque lié à la cohabitation d'une forte topographie et d'une zone critique en focus

Au contraire, on peut trouver sur un design une zone qui présente une forte topographie que le scanner va corriger dans une zone qui n'est pas du tout critique en terme de focus et qui pourrait sans risque être exposée hors conditions optimales. Cette correction de la zone non-critique peut causer une exposition hors focus dans une zone plus critique située dans son voisinage (cf. Figure 5-3).

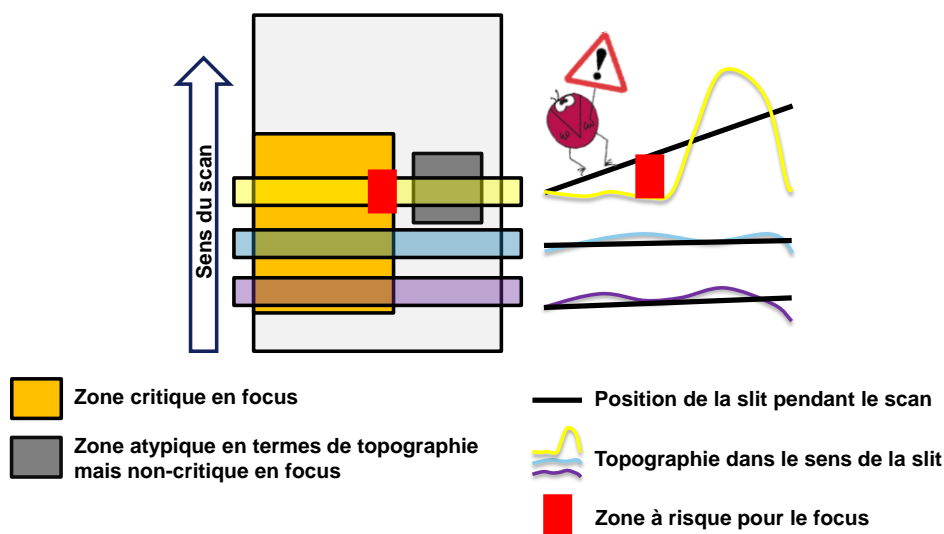


Figure 5-3 : Schéma de principe de la détérioration du focus dans une zone critique par correction d'une zone non-critique voisine

La définition des zones d'intérêts doit donc se faire à partir de la connaissance :

- Du design, qui permet de définir des zones d'importance dans le fonctionnement de la puce.
- De l'imagerie. Pour cela des simulations optiques de la réponse du design à la lithographie ou des mesures sur silicium permettent de dresser une cartographie de focus et de profondeur de champ.
- De la topographie, issue de mesures ou d'un modèle.

Ainsi, les zones d'intérêts peuvent être définies à posteriori sur un produit déjà en fabrication, auquel cas elles permettront d'améliorer le procédé et le contrôle en ligne (cf. Chap. 5.1.1 et 5.1.3), ou bien

avant la commande des masques. Dans ce cas, le design et l'assemblage pourraient être optimisés en fonction des besoins du produit (cf. Chap. 5.1.2).

5.1.1 Optimisation du leveling

L'une de manière de corriger les effets du produit sur l'uniformité du focus en intra-champ et en intra-puce est d'utiliser de manière optimisée les capacités de la mesure et de la correction de topographie par le scanner.

Celui-ci est capable de mesurer l'intégralité du wafer avec le capteur en lumière visible et de corriger les erreurs de mesures à l'aide du capteur pneumatique AGILE (cf. Chap. 4.1.1) et une version du capteur optique en lumière UV est disponible sur les dernières générations de scanner avec plus de spots et une sensibilité moindre à la réflectivité du wafer. Ainsi la mesure est d'ores et déjà optimisée.

Les capacités de correction restent quant à elles limitées par l'impossibilité de réaliser certains mouvements lors du déplacement du chuck, comme plier le wafer dans le sens de la slit par exemple. Une topographie locale, même si elle est mieux mesurée par le scanner, n'est donc pas forcément mieux corrigée par celui-ci pendant l'exposition.

L'idée de l'optimisation proposée ici est de définir des zones d'intérêts qui vont permettre d'ajouter une pondération au calcul de la correction par le scanner pour corriger les zones de la puce qui en ont le plus besoin.

Dans cette étude, trois types de zones ont été définis :

- les zones d'intérêt, critiques en profondeur de champ et en topographie
- les zones à ignorer, dont la topographie est atypique mais qui ne sont pas critiques en focus
- les zones non-spécifiques, qui regroupent le reste du masque

Pour le calcul de la correction par le scanner, des pondérations différentes ont été affectées à chaque zone. Une pondération élevée pour les zones d'intérêt, une pondération basse pour les zones non-spécifique et une pondération nulle pour les zones à ignorer. Ces facteurs sont appliqués à la topographie mesurée par le scanner (capteur optique corrigé par le capteur pneumatique dans notre cas) et permettent de favoriser ou d'ignorer certains points de mesure lors du calcul du leveling.

La figure 5-4 donne le schéma de principe de cette méthode. Le graphique à gauche de la figure est une coupe de la topographie à une position selon Y dans le champ. Le profil optimisé ainsi calculé peut être utilisé comme sous-recette de correction en ligne du leveling par produit. Le wafer serait alors mesuré normalement dans le scanner mais la sous-recette attachée au procédé fournirait les pondérations d'optimisation du calcul du mouvement mécanique du chuck pendant l'exposition.

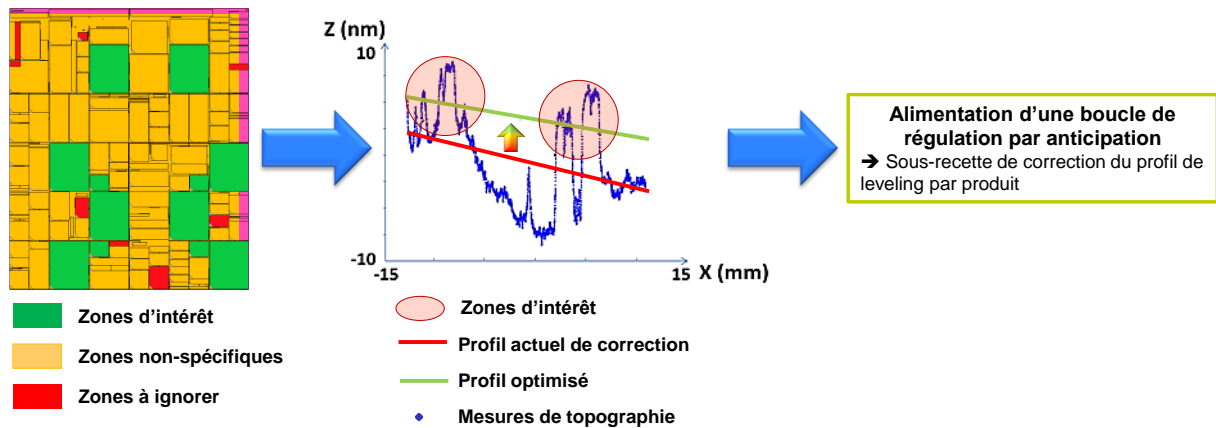


Figure 5-4 : Schéma de principe de l'optimisation du leveling

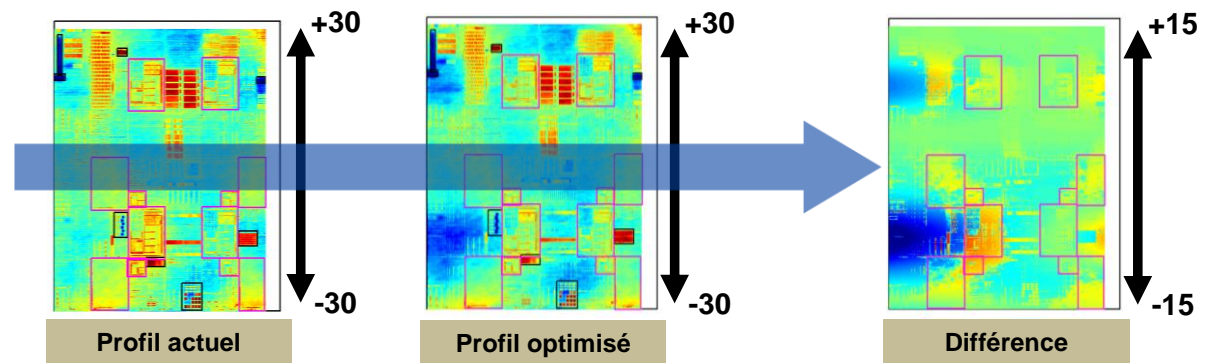


Figure 5-5 : Application des deux profils de correction scanner sur les mesures Wyko intra-champ en Contact 14nm FD-SOI. Les cartographies obtenues sont les erreurs non-corrigeables haute fréquence. La différence (échelle différente) donne le delta entre les deux corrections. Plus la valeur du delta est élevée et plus la version optimisée est performante pour cette position.

Les zones encadrées dans la figure 5-5 (ci-dessus) sont les zones d'intérêts dont on souhaitait optimiser la correction de topographie. On remarque que ces zones sont mieux corrigées avec le profil optimisé, jusqu'à 15nm de plus de correction au maximum et 9nm 3σ de correction en plus dans les zones d'intérêts. Il s'agit cependant d'un compromis entre une meilleure correction dans les zones privilégiées et une perte de performance dans les zones moins importantes qui peuvent subir une dégradation du même ordre de grandeur. Il convient donc de vérifier que l'optimisation des aires critiques du design ne met pas en danger le reste de la puce.

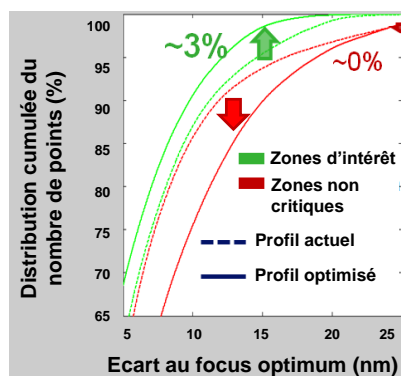


Figure 5-6 : Nombre de positions (pixel de 2.5µm) en spécification focus (écart au focus optimal inférieur à 15nm pour les zones critiques et inférieur à 25nm pour les zones non-critiques) pour les profils de correction optimisé et non-optimisé.

Dans l'exemple ci-dessus, le compromis a été choisi afin qu'aucune zone dite non-spécifique ne soit dégradée au-delà de 25nm de défocus, soit 25nm de topographie non-corrigeable (NCE). Cela a permis de gagner 3% de points en plus dans les zones critiques avec un NCE moindre que 15nm et éviter totalement qu'une zone critique ne présente un défocus de plus de 25nm (cf. Figure 5-6).

Cette optimisation, dont l'intérêt a été démontré par la simulation, est particulièrement intéressante pour des MPW ou des puces de grande taille. En effet, les hétérogénéités de design vont alors être visibles à l'échelle du champ, concerné par le leveling, et non seulement à l'échelle de la puce. Pour des produits matricés (SLR) de petites tailles, le champ est plus homogène en raison de la répétition du produit. La pondération aura alors moins d'influence sur le résultat final.

5.1.2 Optimisation du design

Il est aussi envisageable d'adapter le design lui-même pour limiter la création de topographie et/ou faciliter le leveling du scanner. Ces solutions font partie du DTCO (Design Technology Co-Optimization, cf. Chap. 3.2.1). Contrairement à l'optimisation précédente, il s'agit de propositions d'étude qui n'ont pas encore été démarrées.

5.1.2.1 Dummies

La topographie se forme sur le wafer de plusieurs manières.

Le wafer se bombe suite à des dépôts compressifs ou extensifs et à des gravures ou polissages qui vont relâcher les contraintes subies par la plaquette.

Le design de la puce peut aussi être à l'origine d'une topographie locale comme il a été montré dans le Chap. 4 sur le modèle. Cette topographie vient de l'hétérogénéité du design entre les différentes zones fonctionnelles de la puce.

Pour homogénéiser la puce, et en particulier la densité du design, des structures « dummies » sont créées et remplissent les zones de faible densité. Ces structures jouent aussi un rôle pour certains traitements thermiques. Cependant, la densité seule de ce remplissage n'est pas suffisante pour s'assurer d'une bonne homogénéisation de la topographie à l'échelle de la puce. Le design même des structures dummies doit être le plus proche possible du design du circuit pour que la réponse de la structure aux procédés soit identique à la réponse du reste de la puce. Cela peut être montré par les mesures de topographie réalisées sur le Contact 14nm FD-SOI et sur le BEOL 28nm FD-SOI.

Sur le 14nm FD-SOI, on peut remarquer que la densité de design de la chaîne de contact présentant un fort dishing (cf. Chap. 3.4.2) est nulle pour tous les niveaux précédents le contact. C'est-à-dire qu'aucun motif (ni design ni dummies) n'est présent dans ce bloc pour tous les niveaux précédents le contact. Ainsi, malgré une densité équivalente de contacts entre cette structure de test et un circuit logique

fonctionnelle qui permettra d'obtenir une image aérienne équivalente, l'impression du motif sera fortement impacté par le dishing provoqué par l'absence de structures dummies de remplissage. Les tests électriques réalisés sur cette chaîne (contacts-vias-ligne) après remplissage et CMP du premier niveau d'interconnexion seront alors biaisés par la non-typicité de la structure en design et en topographie. Le procédé risque d'être centré sur les résultats de ce test et donc décalé par rapport aux conditions optimales que l'on attend sur circuit.

Sur le même produit, les puces PROLIGHT et PROMO présentent des différences de topographie entre les zones de dummies de la mémoire et celles de la logique. Cela vient d'un design de dummies qui n'est pas uniforme. Dans la logique, les dummies ressemblent fortement à des transistors réglementaires (sauf qu'ils ne seront ni fonctionnels ni connectés au reste du circuit) et vont réagir comme le circuit aux procédés de fabrication. Les zones de remplissage présentent alors une topographie très proche de celle qu'on observe dans les cellules logiques. En revanche, les dummies de la SRAM sont des motifs très larges par rapport au reste du circuit. On observe un facteur 20 dans les dimensions des structures de remplissage. Cela conduit à une variation de topographie bien plus élevée dans ces zones que dans le reste du circuit.

Si l'absence de dummies conduit évidemment à des non-uniformités de topographie et de procédés au sein même de la puce, le mauvais design de celle-ci peut provoquer des effets tout à fait similaires. Un travail devrait démarrer prochainement chez STMicroelectronics sur le design optimal de ces structures.

Niveau de design	Erreur de mesure (scanner)	Topographie basse fréquence	Topographie haute fréquence	Impact du niveau de design
NOSO	X			Création de deux substrats très différents (avec et sans le box SOI)
		X	X	Marche entre les zones NOSO et les zones SOI (gravure du box et remplissage de la tranchée)
ACTIVE	X			Substrat Silicium sur box en zones Active et Oxyde en zone de tranchées d'isolation
		X	X	Marche en bord des tranchées après remplissage par un Oxyde
ACTIVE on NOSO		X	X	Cumul des marches des zones NOSO et ACTIVE
GATE	X			Empilement de grille métallique
		X	X	Forme de la grille et dépôts des espaceurs autour de celle-ci
GATE on ACTIVE	X			Différents substrats (grille sur oxyde, grille sur SOI ou No-SOI)
TRENCH		X	X	Gravure, remplissage et CMP juste avant l'exposition du Contact
PEPI			X	Epitaxie en escaliers au niveau des zones actives
NEPI			X	

Tableau 5-1 : Impact des différents niveaux de masques sur la topographie à plusieurs échelles spatiales

L'analyse des *VIP* qui accompagne la création du modèle PLS (cf. Chap. 4.3.7) permet de déterminer les niveaux de masques les plus influents et d'en déduire leur effet exact sur la topographie et la mesure de leveling dans le scanner. Le tableau 5-1 ci-dessous donne un aperçu des effets de la densité local de chacun des niveaux de masque utilisés dans la construction du modèle en 14nm FD-SOI. La connaissance et la compréhension des mécanismes en jeu est primordiale pour optimiser le design au mieux. Non seulement, le *VIP* permet de cibler les principaux détracteurs et donc de sélectionner les priorités en termes d'optimisation mais la compréhension du mécanisme donne le sens dans lequel cette optimisation doit être réalisée pour minimiser les risques.

5.1.2.2 « Leveling aware assembly »

Lors de l'architecture de la puce (cf. Chap. 3.2.2), le but est de réussir à optimiser les fonctionnalités de la puce (consommation électrique, temps de réponse, fréquence de fonctionnement, ...). Il peut paraître alors compliqué d'ajouter un niveau de complexité supplémentaire qui consisterait à positionner les différentes parties de la puce les unes par rapport aux autres en fonction de la topographie de chacun de ces blocs pour permettre un leveling plus performant.

En revanche, lors de l'assemblage, il est envisageable de modifier l'agencement du masque pour permettre un leveling optimisé. C'est particulièrement le cas pour les MPW qui regroupent sur un seul masque des puces de développement, des prototypes, des structures de test et de caractérisation très diverses. En effet, il a été montré que la présence de certains de ces blocs atypiques peut perturber le leveling et ainsi détériorer le focus sur des zones plus critiques au profit de zones moins critiques.

Cette proposition est à envisager comme une optimisation supplémentaire en plus de la pondération du leveling et non comme une solution alternative, la modification de l'assemblage laissant moins de marge de manœuvre. L'idée serait d'éviter autant que faire se peut la cohabitation d'une structure atypique et d'une structure critique dans le sens de la slit. On pourrait par exemple imaginer de positionner toutes les structures atypiques au même endroit afin de n'être impacté que sur une partie du champ que l'on sait atypique.

5.1.3 Optimisation de la métrologie

La connaissance de zones d'intérêts à risque peut aussi permettre d'améliorer le suivi en ligne des procédés lors de la fabrication de puces. On peut tout d'abord sélectionner, parmi l'ensemble des mires disponibles dans le champ, les meilleures candidates pour un contrôle plus performant. Celles-ci sont évidemment les mires qui se situent dans les zones les plus critiques du champ (profondeur de champ réduite, risque de forte topographie locale, motifs dont la fenêtre de procédé est faible comparée au reste du design, ...). Il est aussi possible de positionner selon les besoin les mires sur le masque aux meilleurs endroits dès l'assemblage.

De plus, lors de l'exposition d'une FEM pour déterminer les conditions optimales de dose et de focus (cf. Chap. 2.5) ou encore lors d'une PWQ (Process Window Qualification), la méthode de l'« Hyper dense focus map » (cf. Chap. 3.1) ou simplement la connaissance de la topographie locale (déterminée par des mesures ou un modèle) permet d'utiliser le focus réel d'exposition – qui n'est pas le même que le focus que l'on a sélectionné sur la machine – pour déterminer plus précisément le focus optimum et la profondeur de champ que l'on a sur le produit.

5.2 PATTERN FIDELITY CHECK

Le projet Pattern Fidelity Check (Vérification de la fidélité des motifs) est un programme de développement commun entre ST et ASML. Le principe est de prévoir les défauts d'impression avant la commande des masques dans l'objectif de simplifier le suivi des défauts de transfert de motif (lithographie et gravure) sur produit et d'augmenter le taux de détection de la mesure.

Tout au long de cette thèse, le travail réalisé a été intégré en partie dans la méthodologie de ce programme. En particulier, le développement de l'« hyper dense focus map » et sa comparaison avec les données du scanner (cf. Chap. 3.1) ont débouché sur la méthode de l'émulation de wafer.

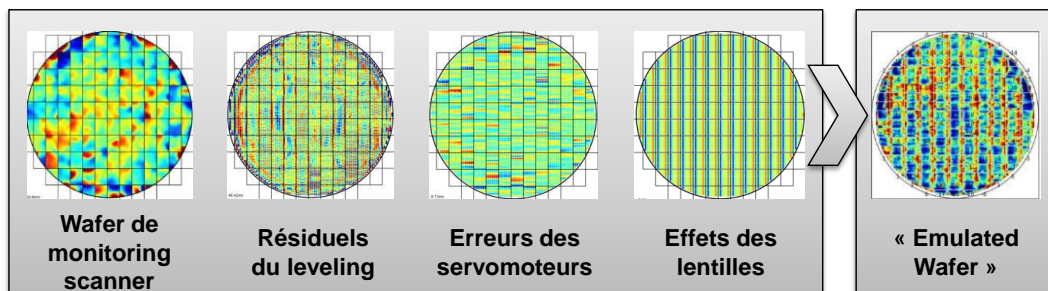


Figure 5-7 : Principe de l'émulation de wafer

L'émulation de wafer [78] consiste à prévoir la cartographie d'uniformité focus sur un wafer complet. Elle prend en compte à la fois les effets du scanner et du produit à la résolution des mesures du scanner. Des données extraites de wafers de monitoring de la machine ainsi que des rapports de procédé extraits sur lots de production permettent de calculer une cartographie du focus sur le wafer sans avoir à mesurer le silicium avec le CDSEM. La Figure 5-7 donne la méthodologie.

Les tests réalisés sur des wafers de production montrent jusqu'à un R^2 de 0.8 et une pente de 1 entre le wafer émulée et les mesures sur produit du focus. En ajoutant une cartographie intra-puce de la profondeur de champ et la position du plan focal par rapport au wafer (simulations LMC, cf. Chap. 3.2.2.1 et 3.5), il est possible d'établir une cartographie intra-plaque des risques de défauts. Les capacités de prédiction de ces défauts sont décrites dans les papiers suivants [79] [80] [81].

L'avantage de cette solution est qu'elle permet aussi d'imaginer un suivi et un contrôle automatique du focus pendant la production du wafer. En effet, les données utilisées par le modèle sont toutes générées régulièrement en ligne lors de l'exposition des plaquettes.

Cette méthode possède cependant un léger désavantage : il est nécessaire d'avoir déjà des fichiers journaux de l'exposition disponibles pour le produit que l'on veut optimiser (résiduels de leveling, erreurs servos et effets des lentilles). Si on veut utiliser la cartographie émulée de focus pour la prédiction de défauts avant même d'exposer un seul wafer, il devient alors impossible de le faire.

Une solution serait d'utiliser le test SSF (cf. Chap 3.1) pour déterminer la part des erreurs focus causées par la machine et d'y ajouter les simulations optiques des fenêtres de procédé. Cependant le test SSF est assez lourd à mettre en place et, comme pour la solution précédente, seule une cartographie basse fréquence du focus est obtenue.

On peut aussi ajouter la topographie intra-puce modélisée par régression PLS à la cartographie de focus établie par le test de monitoring du scanner. Il suffit ensuite d'y appliquer une simulation du leveling du scanner pour obtenir la cartographie émulée de focus. En le combinant avec les données issues des simulations LMC calculées séparément, on obtient alors une cartographie haute fréquence des non-uniformités focus pendant le procédé lithographique.

Une autre méthode consisterait à calculer le LMC en ne considérant non plus le substrat comme plat, ce qui le cas actuellement, mais en utilisant le modèle de topographie comme substrat.

L'ensemble de la méthodologie PFC a été développée et appliquée d'abord en 14nm FD-SOI puis en BEOL 28nm. Entre autres résultats de prédiction de défauts qui montrent que la méthode fonctionne, il a pu être montré que l'impact de la topographie sur le focus est indépendant du motif, de l'intégration et de la technologie.

6 CONCLUSION GENERALE

Ces travaux de thèse ont porté sur l'étude multi-source du contrôle du focus en photolithographie. Le contexte actuel de la lithographie « low k_1 » limite grandement la marge de procédé lors de l'exposition des plaquettes de silicium dans le scanner de lithographie. En effet, la complexité des motifs du design ne sont plus forcément transférables correctement sur le wafer, leur image aérienne étant dégradée par la diffraction trop importante de la lumière. Il ne s'agit cependant pas de la seule raison de la perte de profondeur de champ utilisable en lithographie.

La première partie de cette étude a consisté à mettre en évidence et à caractériser un grand nombre de sources et de mécanismes à l'origine des non-uniformités de focus que l'on observe sur la plaquette. A partir de cette étude, une approche holistique du contrôle du focus s'est avérée nécessaire.

La méthode de l'« hyper dense focus map » a été développée afin d'obtenir une empreinte du procédé à l'échelle d'un wafer complet en termes de focus. Elle a permis de chiffrer les impacts respectifs de plusieurs mécanismes : imagerie pure, erreurs causées par les limites du scanner, influence de l'intégration. Au total, il faut compter environ $25\text{nm } 3\sigma$ de variation intra-champ pour un produit SLR.

La topographie en particulier a été déterminée comme étant un détracteur important du focus. Issue des effets combinés de l'intégration, des procédés, des matériaux, du design et de la machine, la partie non-corrigeable de cette topographie participe en une part non négligeable des erreurs de positionnement de l'image aérienne dans la résine. Son impact a été évalué à 20% des erreurs de focus à l'échelle d'un wafer et près de 50% à l'échelle d'une puce à l'échelle des capacités de correction du scanner. En réalité, si on prend en compte les échelles spatiales non corrigeable (jusqu'au micromètre), cette part est beaucoup plus élevée. Cet effet topologique a un impact mécanique direct sur le focus. Aussi la connaissance de cette topographie permet-elle de mieux anticiper les capacités du procédé. Après analyse des mesures et leur comparaison avec le design, la corrélation entre les deux a montré qu'une prédiction de ces données est possible.

La deuxième partie du travail a permis la création d'un modèle de cette topographie. La méthode de régression multilinéaire PLS, pour Partial Least Square, a offert une méthode simple de création d'un modèle prédictif robuste. Une combinaison linéaire des densités locales du design, directement extraites du GDS de la puce, permet d'obtenir la topographie intra-puce attendue au niveau de lithographie désirée. Ce modèle empirique donne des résultats très satisfaisants à plusieurs échelles de résolution spatiale. A l'échelle millimétrique, c'est-à-dire à l'échelle à laquelle peut mesurer et corriger le scanner, les capacités de prédiction montrent un R^2 de 0.76 pour une pente proche de 1 entre les mesures et le modèle. Pour la topographie haute fréquence, les résultats du modèle sont légèrement meilleurs avec un R^2 de 0.80 et une pente proche de 1.

Il a été démontré que l'on peut alors déterminer une cartographie prédictive du focus à l'échelle d'un wafer complet en ajoutant les valeurs de monitoring de la planéité du chuck du scanner à la topographie haute fréquence issue du modèle puis en appliquant un modèle de leveling du scanner. Cette méthode, appelé Emulated Wafer Map, ouvre la voie à de nombreuses possibilités d'amélioration des procédés, de la métrologie, du contrôle en ligne et de la sécurisation du rendement. On peut citer entre autres les solutions suivantes :

- l'optimisation du design afin de limiter la formation de cette topologie de surface à l'échelle d'une puce ;
- le choix du positionnement de mires de suivi de production ;
- l'amélioration des méthodes de détermination des conditions optimales d'exposition (FEM, PWQ) en utilisant la méthode du Set/Get Focus de l' « hyper dense focus map » ;
- l'amélioration des OPC et le calcul du masque en fonction du décalage de focus induit par la topographie de la plaquette ;
- l'optimisation du leveling en fonction du design et de la topographie haute fréquence non visible par le scanner ;

Le modèle PLS ouvre la voie à l'utilisation de modèles de topographie simples et performants. Il est possible d'améliorer les performances de ce modèle en ajoutant un lissage dépendant des distances de planarisations des procédés de CMP et des dépôts. Ce type d'amélioration permettrait de prendre en compte la densité effective du design sur le procédé et de modéliser les effets de dishing en particulier. Il sera alors aussi possible d'adapter les extractions de design en fonction des procédés utilisés.

REFERENCES

- [1] M. Faraday, "On Conducting Power Generally," in *Experimental Researches in Electricity*, 1839, pp. 168-170.
- [2] E. Hall, "On a new action of the magnet on electric currents," *American Journal of Mathematics* Vol.2 (3), p. 287–292, 1879.
- [3] W. Brattain and J. Bardeen, "Three-electrode circuit element utilizing semiconductive materials". Patent US Patent 2524035, 1948.
- [4] H. F. Mataré, "Nouveau système cristallin à plusieurs électrodes réalisant des relais d'effets électroniques". Patent Brevet d'invention FR 1.010.427, 1948.
- [5] W. Shockley, "The theory of p-n junctions in semiconductors and p-n junction transistors," *Bell System Technical Journal*, vol. 28 , no. 3, pp. 435-489, 1949.
- [6] GSA, "Enabling the Hyperconnected Age," Global Semiconductor Alliance & Oxford Economics, 8 Septembre 2016.
- [7] G. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, p. 114 et suivantes, 1965.
- [8] ITRS, "ITRS Roadmap," ITRS, 2001 à 2013.
- [9] W. Arden, M. Brillouët, P. Cogez, M. Graef, B. Huizing and R. Mahnkopf, "More Than Moore White Paper," ITRS.
- [10] P. Bright, "Moore's law really is dead this time," arstechnica, 11 Février 2016. [Online]. Available: <http://arstechnica.com/information-technology/2016/02/moores-law-really-is-dead-this-time/>. [Accessed 08 Septembre 2016].
- [11] A. Landzberg, in *Microelectronics Manufacturing Diagnostics Handbook*, Springer Science & Business Media, 2012, pp. 37-38.
- [12] B. Le-Gratiet, M. Gatefait, J. Ducotè, J. Decaunes, A. Lam, B. Beraud, M. Mikolajczak, A. Pelletier, B. Orlando, F. Sundermann, A. Ostrovsky and C. Lapeyre, "How holistic process control translates into high mix logic fab APC?," in *Proc. SPIE 9231, 30th European Mask and Lithography Conference*, 2014.

- [13] H. J. Levinson, *Principles of Lithography - Third Edition*, Bellingham (WA): Society of Photo-Optical Instrumentation Engineers, 2010.
- [14] T. Matsunawa, B. Yu and D. Z. Pan, "Optical proximity correction with hierarchical Bayes model," *J. Micro/Nanolith. MEMS MOEMS*, vol. 15, no. 2, Mars 2016.
- [15] M. v. d. Brink, "Holistic lithography and metrology's importance in driving patterning fidelity," in *Proc. SPIE 9778, Metrology, Inspection, and Process Control for Microlithography XXX*, 2016.
- [16] J. Mulkens, P. Hinnen, M. Kubis, A. Padiy and J. Benschop, "Holistic optimization architecture enabling sub-14-nm projection lithography," *J. Micro/Nanolith. MEMS MOEMS.*, vol. 13, no. 1, 2014.
- [17] C. Chiang and J. Kawa, *DFM & Yield for nano-scale CMOS*, Springer, 2007.
- [18] A. Basalinski, *Design for Manufacturability: from 1D to 4D for 90-22nm technology nodes*, Springer, 2014.
- [19] J. Tirapu-Azpiroz, A. E. Rosenbluth and T. Brunner, "Layout optimization method to equalize the best-focus position of different patterns," *J. Micro/Nanolith. MEMS MOEMS.*, vol. 13, no. 2, 2014.
- [20] D. Yang, C. Gan, P. R. Chidambaram, G. Nallapadi, J. Zhu, S. C. Song, J. Xu and G. Yeap, "Technology-design-manufacturing co-optimization for advanced mobile SoCs," in *Design-Process-Technology Co-optimization for Manufacturability VIII*, 2014.
- [21] C. A. Mack and P. M. Kaufman, "Mask bias in submicron optical lithography," *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 6, no. 6, p. 2213, 1988.
- [22] R. Seltmann, "28nm node process optimization: A Lithographic Centric View," in *30th European Mask and Lithography Conference*, 2014.
- [23] N. Abidine, F. Sundermann, E. Yesilada, V. Farys, F. Huguennet, A.-M. Armeanu, I. Bork, M. Chomat, P. Buck and I. Schanen, "Accurate mask model implementation in optical proximity correction model for 14-nm nodes and beyond," *J. Micro/Nanolith. MEMS MOEMS*, vol. 15, no. 2, 2016.
- [24] N. Abidine and al., "Mask Blank Optimization through rigorous EMF approach from IDM perspective, for 28 nm node and beyond," in *Photomask Japan 22*, Yokohama, 2015.

- [25] A. Bouma, J. Miyazaki, M. v. Veen and J. Finders, "Impact of mask absorber and quartz over-etch on mask 3D induced best focus shifts," in *Proc. SPIE 9231, 30th European Mask and Lithography Conference*, 2014.
- [26] T. A. Brunner, "Impact of lens aberrations on optical lithography," *IBM Journal of Research and Development*, vol. 41, no. 1/2, pp. 57-67, 1997.
- [27] S. Halle, M. Crouse, A. Jiang, Y. v. Dommelen, T. Brunner and al., "Lens heating challenges for negative tone develop layers with freeform illumination: a comparative study of experimental vs. simulated results," in *Proc. SPIE 8326, Optical Microlithography XXV*, 2012.
- [28] L. Subramany, W. J. Chung, P. Samudrala, H. Gao, N. Aung and al., "Analysis of wafer heating in 14nm DUV layers," in *Proc. SPIE 9778, Metrology, Inspection, and Process Control for Microlithography XXX*, 2016.
- [29] T. A. Brunner, "Impact of lens aberrations on optical lithography," *IBM Journal of Research and Development*, vol. 41, no. 57, 1997.
- [30] P. Alagna, O. Zurita, V. Timoshkov, P. Wong, G. Rechtsteiner, J. Baselmans and J. Mailfert, "Optimum ArFi laser bandwidth for 10nm node logic imaging performance," in *Proc. SPIE 9426, Optical Microlithography XXVIII*, 2015.
- [31] K. Yoshimochi, T. Tamura, T. Kuribayashi, T. Uchiyama, N. Farrar, T. Oga and J. Bonafede, "32nm node device laser-bandwidth OPE sensitivity and process matching," in *Optical Microlithography XXII*, 2009.
- [32] J.-G. Simiz, T. Hasan, F. Staals, B. Le-Gratiet, P. Gilgenkrantz, A. Villaret, F. Pasqualini, W. T. Tel, C. Prentice and A. Tishchenko, "Predictability and impact of product layout induced topology on across-field focus control," in *Metrology, Inspection, and Process Control for Microlithography XXIX*, 2015.
- [33] F. Goos and H. Hänchen, "Ein neuer und fundamentaler Versuch zur Totalreflexion," *Ann. Phys.*, vol. 436, no. 7-8, pp. 333-346, 1947.
- [34] M. Gatefait, A. Lam, B. Le-Gratiet, M. Mikolajczak, V. Morin, N. Chojnowski, Z. Kocsis, I. Smith, J. Decaunes, A. Ostrovsky and C. Monget, "AGILE integration into APC for high mix logic fab," in *31st European Mask and Lithography Conference*, 2015.

- [35] J.-C. Le-Denmat, C. Martinelli, E. Sungauer, J.-C. Michel, E. Yesilada and F. Robert, "Wafer sub-layer impact in OPC/ORC models for 2x nm node implant layers," in *Optical Microlithography XXVI*, 2013.
- [36] J.-G. Simiz, T. Hasan, F. Staals, B. Le-Gratiet, W. T. Tel, C. Prentice and A. Tishchenko, "Product layout induced topography effects on intrafield levelling," in *31st European Mask and Lithography Conference*, 2015.
- [37] H. Lee, J. Lee, S. Kim, C. Lee, S. Han, M. Kim, W. Kwon, S.-K. Park, S. Veeraraghavan, J. Kim, A. Awasthi, J. Byeon, D. Mueller and J. Sinha, "Improvement of depth of focus control using wafer geometry," in *Metrology, Inspection, and Process Control for Microlithography XXIX*, 2015.
- [38] U. Katakamsetty, H. Colin, S. Yeo, P. Valerio, Y. Qing, Q. S. Fong, N. S. Aravind, R. Matthias and S. Roberto, "Scanner correction capabilities aware CMP / Lithography hotspot analysis," in *Design-Process-Technology Co-optimization for Manufacturability VIII*, 2014.
- [39] M. Gatefait, B. Le-Gratiet, C. Prentice and T. Hasan, "An evaluation of edge roll off on 28nm FDSOI (fully depleted silicon on insulator) product," in *Metrology, Inspection, and Process Control for Microlithography XXX*, 2016.
- [40] P. Ong, L. Economikos, D. H. Hong, M. Chae, R. Qong, S. Grunow, D. Dipaola, S. Siddiqi, B. Liegl, S. Ponoht, W.-T. Tseng, A. Ticknor, R. Fang, D. Kulkarni, M. Lagus, G. Matusiewicz, M. Angyal and D. Watts, "Design Influence on CMP-Induced Topography at Chip and Wafer Scales over Multiple Levels," in *International Conference on Planarization/CMP Technology*, 2006.
- [41] J.-G. Simiz, T. Hasan, F. Staals, B. Le-Gratiet, W. T. Tel, C. Prentice, J.-W. Gemmink, A. Tishchenko and Y. Jourlin, "Verification and application of multi-source focus quantification," in *Design-Process-Technology Co-optimization for Manufacturability X*, 2016.
- [42] A. Szucs, J. Planchot, V. Farys, E. Yesilada, C. Alleaume, L. Depre, R. Dover, C. Gourgon, M. Besacier, A. Nachtwein and P. Rusu, "Best focus shift mitigation for extending the depth of focus," in *Proc. SPIE 8683, Optical Microlithography XXVI*, 2013.
- [43] A. Szucs, J. Planchot, V. Farys, E. Yesilada, L. Depre, S. Kapasi, C. Gourgon, M. Besacier, O. Mouraille and F. Driessen, "Advanced OPC Mask-3D and Resist-3D modeling," in *Proc. SPIE 9052, Optical Microlithography XXVII*, 2014.
- [44] W. Conley, S. Hsieh, P. Alagna, Y. Hou and P. Martinez, "Impact of bandwidth on contrast sensitive structures for low k1 lithography," in *Optical Microlithography XXVIII*, 2015.

- [45] P. Alagna, O. Zurita, V. Timoshkov, P. Wong, G. Rechtsteiner, J. Baselmans, J. Mailfert, W. Conley and S. Hsieh, "Advanced process characterization of a 10nm Metal 1 Logic layer using light source modulation and monitoring," in *31st European Mask and Lithography Conference*, 2015.
- [46] W. Coney and al., "Process Improvements with Lower Bandwidth Light Sources and the Impact of Reduced Bandwidth Variation," in *32nd European Mask and Lithography Conference*, 2016.
- [47] Y. Cui, "Fine-tune lens-heating-induced focus drift with different process and illumination settings," in *Optical Microlithography XIV*, 2001.
- [48] J.-C. Michel, *Caractérisation et modélisation des effets d'empilement des couches minces sous la résine photosensible pendant le procédé de photolithographie optique*, Université de Saint Etienne: Thèse de doctorat, 2014.
- [49] M. Pike, S. Holmes, B. Leigl, M. Lagus, S. Greco and D. Coleman, "Effects of intra chip topography in back end of line processes on focus leveling control and process window degradation with high NA exposures," in *IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, 2004.
- [50] P. Teunissen, R. Queens, P. Dekkers-Rog and A. Van-Well, "Level sensor for lithographic apparatus". Patent US7265364 B2, 10 June 2004.
- [51] A. W. Snyder and J. D. Love, "Goos-Hänchen shift," *Applied Optics*, vol. 15, no. 1, p. 236, 1976.
- [52] T. J. Wiltshire, B. R. Liegl, E. M. Hwang and M. R. Lucksinger, "Application of automated topography focus corrections for volume manufacturing," in *Metrology, Inspection, and Process Control for Microlithography XXIV*, 2010.
- [53] P. Klapetek, D. Necas and C. Anderson, *Gwyddion User Guide*, 2014.
- [54] *Wyko NT9300 Optical Profiling System*, Veeco Instruments Inc., 2006.
- [55] F. Dettoni, M. Rivoire, S. Gaillard, O. Hinsinger, F. Bertin and C. Beitia, "High Resolution Nanotopography Characterization at Die Scale of 28nm FDSOI CMOS Front-end CMP Processes," *Microelectronic Engineering*, vol. 113, pp. 105-108, 2014.
- [56] F. Dettoni, *Développement de procédés de mesure spatialement résolue de la nano-topographie sur des distances centimétriques : application au polissage mécano-chimique*, Université de Grenoble: Thèse de doctorat, 2013.

- [57] F. Dettoni, Y. Morand, S. Gaillard, O. Hinsinger and M. Rivoire, "Interferometry: a direct die level characterization technique," in *International Conference on Planarization/CMP Technology*, 2012.
- [58] N. Ruiz, F. Dettoni, M. Rivoire, V. Balan and C. Beitia, "In-Die High Resolution Nanotopography Data, Impact in the CMP Process Monitoring for Advanced Nodes," in *International Conference on Frontiers of Characterization and Metrology for Nanoelectronics*, 2015.
- [59] J. F. Valley, N. Poduje, J. Sinha, N. Judell, J. Wu, M. Boonman, S. Tempelaars, Y. v. Dommelen, H. Kattouw, J. Hauschild, W. Hughes, A. Grabbe and L. Stanton, "Approaching new metrics for wafer flatness: an investigation of the lithographic consequences of wafer non-flatness," in *Metrology, Inspection, and Process Control for Microlithography XVIII*, 2004.
- [60] T. A. Brunner, Y. Zhou, C. W. Wong, B. Morgenfeld, G. Leino and S. Mahajan, "Patterned wafer geometry (PWG) metrology for improving process-induced overlay and focus problems," in *Optical Microlithography XXIX*, 2016.
- [61] B. J. Morgenfeld, T. A. Brunner, K. Numm, D. Stoll, N. Jing, H. Lin, P. Vukkadala, P. Herrera, R. Ramkhalawon and J. Sinha, "Monitoring Process-induced Focus Errors Using High-resolution Flatness Metrology," in *26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2015.
- [62] H. Wold, "Path Models with Latent Variables: The NIPALS Approach," *Quantitative Sociology*, pp. 307-357, 1975.
- [63] S. Wold, H. Martens and H. Wold, "The Multivariate Calibration Problem in Chemistry Solved by the PLS Method," *Matrix Pencils Lecture Notes in Mathematics*, pp. 286-293, 1983.
- [64] S. Wold, P. Geladi, K. Esbensen and J. Öhman, "Multiway Principal Components and PLS-Analysis," *Journal of Chemometrics*, vol. 1, pp. 41-56, 1987.
- [65] L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström and S. Wold, *Multi and Mega-Variate Data Analysis*, Umea: Umetrics Academy UMETRICS AB, 2006.
- [66] P. Geladi and B. R. Kowalski, "Partial Least-Square Regression: A Tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1-17, 1986.
- [67] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," *Subspace, Latent Structure and Feature Selection Lecture Notes in Computer Science*, pp. 34-51, 2006.

- [68] S. Wold, M. Sjöström and L. Eriksson, "PLS-Regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, pp. 109-130, 2001.
- [69] A. Höskuldsson, "PLS regression methods," *Journal of Chemometrics*, vol. 2, pp. 211-228, 1988.
- [70] B. Vandeginste, C. Sielhorst and M. Gerritsen, "The NIPALS Algorithm for the Calculation of the Principal Components of a Matrix," *Trends in Analytical Chemistry*, vol. 7, no. 8, pp. 286-287, 1988.
- [71] Y. Miyashita, T. Itozawa, H. Katsumi and S.-I. Sasaki, "Comments on the NIPALS Algorithm," *Journal of Chemometrics*, vol. 4, no. 1, pp. 97-100, 1990.
- [72] L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström and S. Wold, "4.6.3 The variable influence on projection, VIP, parameter & Appendix II: Partial Least Square Modeling - Variable Influence: VIP," in *Multi- and Megavariate Data Analysis*, Umea, Umetrcis Academy UMETRICS AB, 2006, p. 85 & 390.
- [73] C. Chiang and J. Kawa, "4.4 A Full Chip Simulation Algorithm," in *Design fo Manufacturability and Yield for Nano-Scale CMOS*, Dordrecht, Springer, 2007, pp. 112-121.
- [74] C. Chiang and J. Kawa, "4.6 Dummy Filling - 4.8 Application: Cu CMP Model Based Filling," in *Design for Manufacturability and Yield for Nano-Scale CMOS*, Dordrecht, Springer, 2007, pp. 124-150.
- [75] C. Hui, X. B. Wang, H. Huang, U. Katakamsetty, L. Economikos, M. Fayaz, S. Greco, X. Hua, S. Jayathi, C.-M. Yuan, S. Li, V. Mehrotra, K. H. Chen, T. Gbondo-Tugbawa and T. Smith, "Hotspot detection and design recommendation using silicon calibrated CMP model," in *Design for Manufacturability through Design-Process Integration III*, 2009.
- [76] *Standard Verification Rule Format Manual*, Mentor Graphics Corp., 1996-2014, pp. 166-192.
- [77] S. Babu, W. Fan and D. Boning, "Part1 Section 6. Multiscale modeling of chemical mechanical planarization - Fig. 6-8," in *Advances in Chemical Mechanical Planarization (CMP)*, Cambridge (UK), Woodhead Publishing, Elsevier Ltd., 2016, pp. 137-168.
- [78] W.T.Tel, B.Segers, R.Anunciado, B. Le-Gratiet, T.Hasan, C.Prentice, J.-G. Simiz and A. Lakcher, "Efficient hybrid metrology for focus, CD, and overlay," in *SPIE Advanced Lithography (submitted)*, 2017.

- [79] S. Hunsche, M. Jochemsen, V. Jain, X. Zhou, F. Chen, V. Vellanki, C. Spence, S. Halder, D. v. d. Heuvel and V. Truffert, "A new paradigm for in-line detection and control of patterning defects," in *Metrology, Inspection, and Process Control for Microlithography XXIX*, 2015.
- [80] P. Fanton, R. L. Greca, V. Jain, C. Prentice, J.-G. Simiz, S. Hunsche, B. Le-Gratiet and L. Depre, "Process window optimizer for pattern based defect prediction on 28nm metal layer," in *Metrology, Inspection, and Process Control for Microlithography XXX*, 2016.
- [81] P. Fanton, J.-G. Simiz, A. Lakcher, B. Le-Gratiet, C. Prentice, T. Hasan, R. L. Greca, L. Depre and S. Hunsche, "Advanced in-production hotspot prediction and monitoring with microtopography," in *SPIE Advanced Lithography (submitted)*, 2017.

ANNEXE: LISTE DES PAPIERS PUBLIES

- **J.-G. Simiz** ; T. Hasan ; F. Staals ; B. Le-Gratiet ; P. Gilgenkrantz ; A. Villaret ; F. Pasqualini ; W. T. Tel ; C. Prentice ; A. Tishchenko ;
" Predictability and impact of product layout induced topology on across-field focus control ", Proc. SPIE 9424, Metrology, Inspection, and Process Control for Microlithography XXIX, 94241C (March 19, 2015); doi:10.1117/12.2085283;

- **J.-G. Simiz** ; T. Hasan ; F. Staals ; B. Le-Gratiet ; W. T. Tel ; C. Prentice ; A. Tishchenko ;
" Product layout induced topography effects on intrafield levelling ", Proc. SPIE 9661, 31st European Mask and Lithography Conference, 96610R (September 4, 2015); doi:10.1117/12.2194079;

- **J.-G. Simiz** ; T. Hasan ; F. Staals ; B. Le-Gratiet ; W. T. Tel ; C. Prentice ; J.-W. Gemmink ; A. Tishchenko ; Y. Jourlin,
" Verification and application of multi-source focus quantification ", Proc. SPIE 9781, Design-Process-Technology Co-optimization for Manufacturability X, 97810S (March 16, 2016); doi:10.1117/12.2219143

- P. Fanton ; R. La Greca ; V. Jain ; C. Prentice ; **J.-G. Simiz** ; S. Hunsche ; B. Le-Gratiet ; L. Depre ;
" Process window optimizer for pattern based defect prediction on 28nm metal layer ", Proc. SPIE 9778, Metrology, Inspection, and Process Control for Microlithography XXX, 97782O (March 8, 2016); doi:10.1117/12.2220295;

Predictability and impact of product layout induced topology on across-field focus control

J-G. Simiz^{1,2}, T. Hasan³, F. Staals³, B. Le-Gratiet¹, P. Gilgenkrantz¹
A. Villaret¹, F. Pasqualini¹, W.T. Tel³, C. Prentice⁴, A. Tishchenko²

¹STMicroelectronics, 850 rue Jean Monnet, F-38926 Crolles Cedex, France

²LaHC CNRS-UMR 5516, 18 Rue Professeur Benoît Lauras, F-42000 Saint-Étienne, France

³ASML, De Run 6501, 5504DR Veldhoven, the Netherlands

⁴ASML SARL, 459 chemin des Fontaines, F-38190 Bernin, France

Abstract:

With continuing dimension shrinkage using the TWINSCAN NXT:1950i scanner on the 28nm node and beyond, the imaging depth of focus (DOF) becomes more critical. Focus budget breakdown studies [Ref 1, 5] show that even though the intrafield component stays the same this becomes a larger relative percentage of the overall DOF. Process induced topography along with reduced Process Window can lead to yield limitations and defectivity issues on the wafer. To improve focus margin, a study has been started to determine if some correlations between scanner levelling performance, product layout and topography can be observed. Both topography and levelling intrafield fingerprints show a large systematic component that seems to be product related. In particular, scanner levelling measurement maps present a lot of similarities with the layout of the product. The present paper investigates the possibility to model the level sensor's measured height as a function of layer design densities or perimeter data of the product. As one component of the systematics from the level sensor measurements is process induced topography due to previous deposition, etching and CMP, several layer density parameters were extracted from the GDS's. These were combined through a multiple variable analysis (PLS: Partial Least Square regression) to determine the weighting of each layer and each parameter. Current work shows very promising results using this methodology, with description quality up to 0.8 R² and expected prediction quality up to 0.78 Q². Since product layout drives some intrafield focus component it is also important to be able to assess intrafield focus uniformity from post processing. This has been done through a hyper dense focus map experiment which is presented in this paper.

Keywords: depth of focus, intrafield, scanner levelling, topography, scanner, product design layout effect, PLS regression analysis

INTRODUCTION

For low k1 lithography, the scanner control capabilities are very similar to the process window of critical features so that process margin is drastically reduced. In early phase of technology development TWINSCAN NXT:1950i focus budget breakdown shows values that are sometimes larger than the depth of focus of some critical patterns within the chip. In this scope, bringing process to maturity requires actions to retrieve appropriate process margin to secure the yield. Some of these actions will address random dynamic behaviour of the scanner while others will focus on systematic fingerprint and their mitigation.

The depth of focus can be enhanced or controlled using a number of different methods.

- Process window enhancement (Mask3D/Resist3D aware OPC, SMO, etc.)
- Feature to feature Best Focus shift mitigation through mask enhancement (mask blank)
- Precision and accuracy gains for Best Focus determination using new methods

- In-line monitoring
- Topography compensation (Design for Manufacturing, other processes)
- Levelling control (TWISCN NXT:1970Ci, UV-LS, AGILE2)
- Interfield focus control (BaseLiner Focus, Field Width Optimized Levelling, Chuck Deformation Map updating, Imaging Optimizer)
- Layout (or hotspot) aware scanner control(Process Window Optimizer)

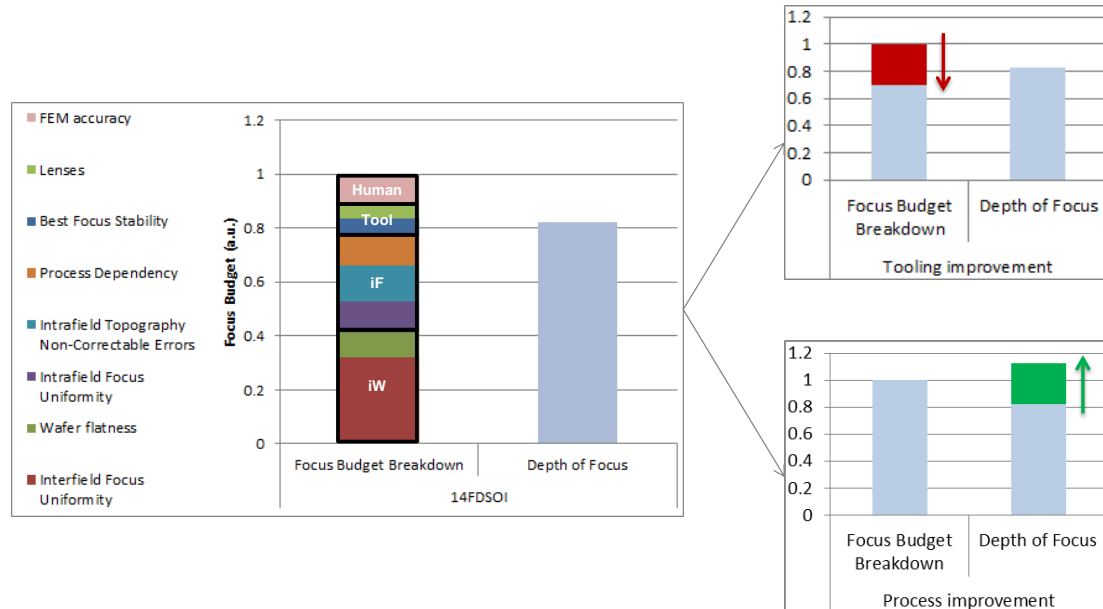


Figure 1: Scanner focus budget breakdown vs. process DOF of the studied reference pattern for the 14FD-SOI technology

Figure 1 shows the current focus budget breakdown for 14FD-SOI versus the depth of focus of the process at Contact layer [Ref 1]. Intrafield and interfield focus components count together for more than $2/3$ of the total budget for both 28 and 14FD-SOI technologies.

This paper will look into the topological and levelling component of the focus budget.

I – SCANNER LEVELLING FLOW

During the lithographic exposure processes, it is important to ensure that the mask image is correctly focused on the dies in the wafer. For this correct positioning of the wafer, levelling is used in the lithographic tools. Levelling can be referred to as the process of measuring the 'vertical' position of the wafer by a Level Sensor (LS) and using this information to keep the wafer in (best) focus during exposure. Level Sensor is an optical interferometric sensor that measures the surface height using optical triangulation method. From the level sensor measurement, a 'wafer map' containing the surface height of the wafer placed on a wafer stage is created.

Figure 2 shows a wafer map. A wafer map can be divided into multiple areas corresponding to dies or fields. After subtracting the global shape from the wafer map, an average height map of the fields are then calculated and shown in figure 2 (right). The latter is representative of the intra field topography measured by the level sensor (LS).

Level sensors used for the wafer map measurements can be subject to process dependency [Ref 6]. Process dependency is a form of measurement error in which level sensor provide differing results depending on how the measured wafer has previously been processed. For example, a level sensor may provide a particular height measurement for a wafer including a silicon substrate coated with a single layer of resist, and may provide a different height measurement for a wafer including a silicon substrate coated with several layers of resist, even if both wafers' surface are at the same actual height.

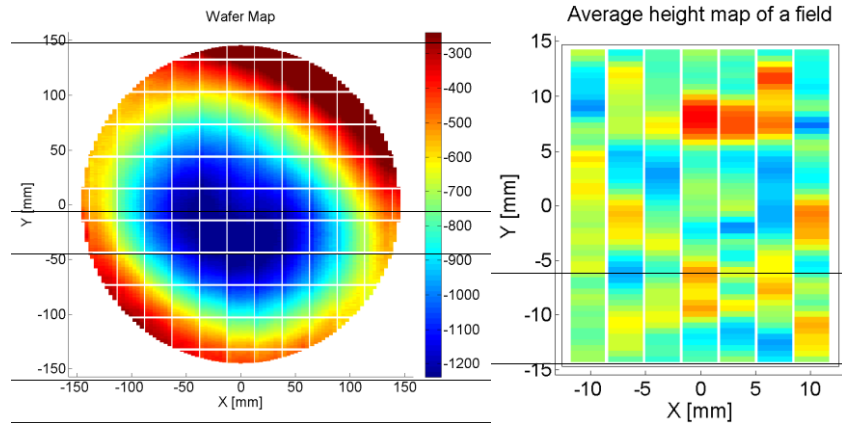


Figure 2: (Left) Level sensor measured wafer surface height map and (Right) average height map of a field (die) of the wafer.

One error caused by process dependency is referred to as process dependent apparent surface depression, and is understood to be caused by an optical effect known as the Goos-Haenchen shift [Ref 7]. The Goos-Haenchen shift is a lateral translation of light along a reflecting surface (in this case the resist) during reflection. The shift is dependent upon the material and layer structure of the substrate and is due to the visible light used by the Level Sensor. The newest scanner generation has a level sensor using UV light suffering less from this effect.

$$LS = \text{Actual topography} + \text{Process dependency} \quad (1)$$

As a result of process dependency, a die may not be correctly located in the focal plane of the projection lens. To overcome this, process dependency corrections are applied to the level sensor measured wafer map. ASML's commercial products 'Air Gauge (AG)' and 'AGILE' (Air Gauge Improved Levelling)" is used to do this correction. So, the process dependency corrected intra field height map is more accurate measurement of the intra field topography.

$$LS_{AGcorrected} = LS - \text{Process Dependency Correction} \approx \text{Actual Topography} \quad (2)$$

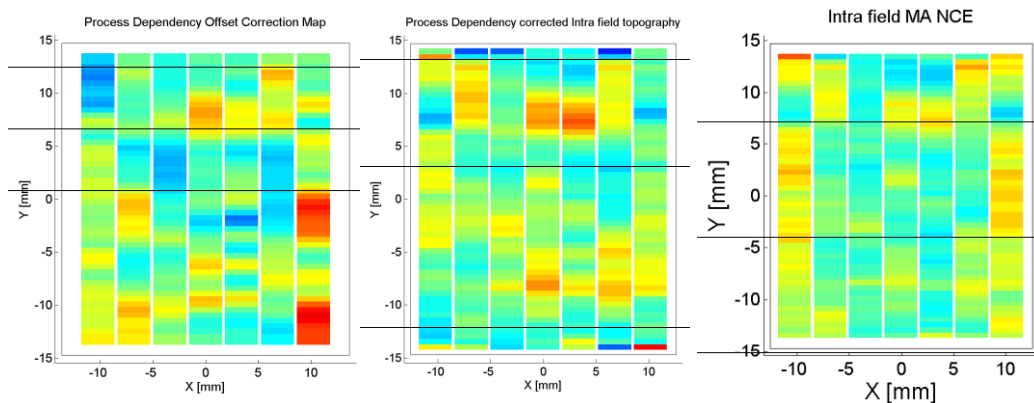


Figure 3: (Left) Process dependency correction, (Middle) process dependency corrected intra field topography map and (Right) intra field levelling MA NCE

The wafer map is used to derive setpoints to actuate the wafer stage to position the dies in focus at the location where the projection takes place. Due to the finite size and rectangular nature of the exposure slit, some of the surface topography cannot be leveled completely. This non-level-able residual is called non-correctable errors (NCE). During a scanned exposure, the non-correctable errors change continuously as the slit is scanned over a particular position on the wafer. In the latter case, the average value of the non-correctable errors over exposure time, defines the average defocus that this position experiences. This average value is termed as the moving average of the non-correctable errors (MA-NCE). Topography changes that are smoother than the slit dimensions can be leveled by adjusting the stage height and tilt angle accordingly. If the topography varies within the slit

dimensions, the height changes cannot be leveled effectively. Hence, a large part of the highly variant topography is non correctable and is often present in the MA NCE map.

II – METHODOLOGY / CONCEPT

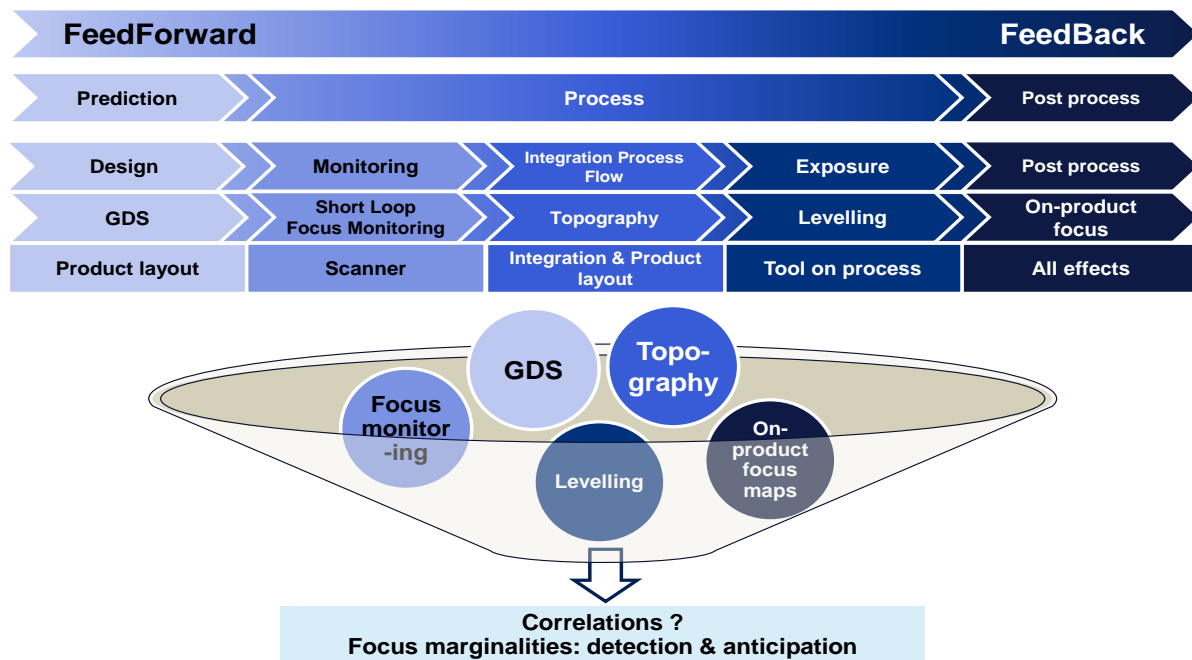


Figure 4: Different data sources and their use in the focus correlation study (GDS: Design layout, Short loop focus monitoring: Single Shot Focal test, On-product focus: Focus Uniformity Map “Bossung top best focus”)

The study that is initiated in this paper consists of a multi-source data analysis. Inputs include scanner log files, GDS’s, on-product and bare-wafer focus and topography measurements. They are processed together in a multivariate analysis tooling (Partial Least square regression software) in order to generate a prediction model. Figure 4 above summarizes the different sources of data and their availability in the manufacturing flow.

1. GDS extracted data

It is possible to generate layer density and perimeter files directly from the product GDS. Density and perimeter data are extracted at both a millimetre and micrometre scale to match the scanner reports, the focus uniformity maps and the topography measurements layouts. Density maps correspond to spatial density of patterns, i.e. to the density of top surface of patterns for a given area. The extracted perimeter data is the density of sidewall of patterns for a given area. Both density and perimeter influence the material distribution, which has some impact on the optical modulation that the level sensor is suffering from (Level Sensor Process Dependency), and process induced topography interfield and intrafield (previous CMP and deposition steps). This data can be used to generate a high density map of wafer topography, which may be used as one of the inputs for Process Window Optimizer.

2. Scanner reports

These logs are automatically generated for every lot that is processed on the scanner. The levelling part of the report contains information on the wafer topology at a millimetre resolution for every wafer in the lot. Data that can be extracted from these files are the following:

- Optical Level Sensor raw maps
- Process Dependency corrected LS maps
- Process Dependency maps of the LS
- Non-correctable errors (NCE) maps

3. Focus uniformity mappings

Interfield and intrafield focus maps are measured by CDSEM at 52 locations across field on product exposures. This was done by measuring CD on 7 product wafers with 15nm focus steps between each wafer. A Bossung was then re-constructed at each measurement location and the Bossung top was taken as the best focus position. These exposures were done on a TWINSCAN NXT:1950i with AGILE enabled and Baseline Focus disabled. An inline alternative to this technique is on-product intrafield and interfield focus measurements using ASML's uDBF targets.

4. Topography measurements

Intrafield and interfield topography measurements can be performed at micrometre resolution using multiple different offline optical measurement tools [Ref 4]. This gives high frequency topography maps that the scanner cannot measure due to the resolution of the level sensor and AirGauge sensor. With this extra information it will be possible to perform a 1 time verification of the results from the GDS density design analysis and use the data for optimizing scanner focus control. Note that for this paper, these data are not yet fully available and analysed.

5. Short loop focus monitoring

The Single Shot Focal test was exposed with 13x19 FOCAL marks per field on a bare silicon wafer with the same field layout as the product wafers. This was measured using the scanners alignment sensor to give the interfield scanner focus fingerprint. The test was executed on a TWINSCAN NXT:1950i with AGILE enabled and without Baseline Focus correction (which is used to correct for the scanner fingerprint in production). This means that the SSF fingerprint contains correctable errors as well as scanner non-correctable errors.

6. Inputs summary

INPUT N°	INPUT NAME	INTRAFIELD	INTERFIELD	TIMING	USE
1	GDS			No Silicon needed	FeedForward
2	Scanner reports			Silicon / Lots	FeedBack
3	Focus maps			Silicon / Lots	FeedBack
4	Topography	On-going	Planned	Silicon / Lots	FeedForward & FeedBack
5	SSF			Silicon / Monitoring	FeedForward

7. Multivariate analysis

The Partial Least Square (PLS) regression analysis was chosen to investigate the correlations that exist between each of the previous datasets. PLS is a linear multivariate method.

GDS is the most predictive file that is available since it exists before any silicon is processed in the waferfab. It will be utilized as the x-input of each of the following analysis.

Scanner levelling data, focus maps and topography measurement were used as y-values in the regression tooling. Figure 5 gives a schematic view of the analysis process.

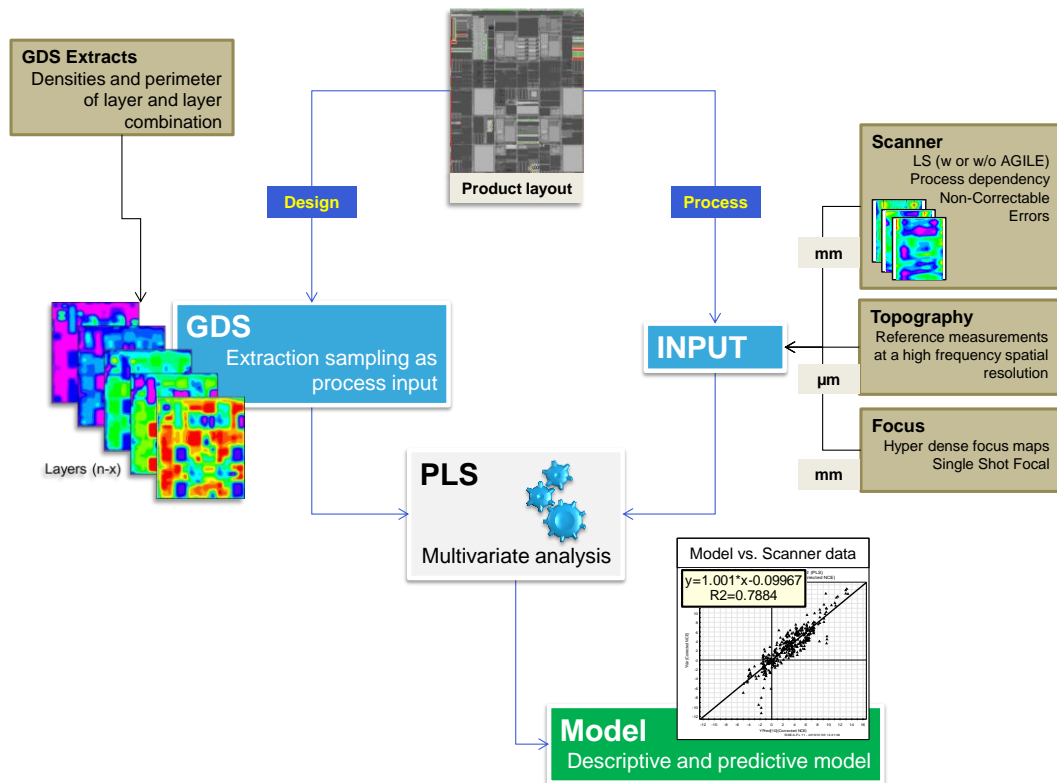


Figure 5: Schematic view of the process through PLS regression analysis

Two correlation parameters are calculated by the PLS analysis tooling.

- R^2 or goodness of fit of the model. It gives the description quality of the model.
- Q^2 or goodness of prediction. It is related to the capability of prediction that can be expected from the model.

The best model is the one with the best couple (R^2 , Q^2).

Here the PLS is used to answer to the following question:

- Can scanner levelling data and topography be anticipated from design data? How does it relate to focus?

III – GDS DESIGN DENSITY ANALYSIS INTRA-FIELD LEVELLING PREDICTABILITY

The scanner levelling measurements are related to the product layout as described in chapter I.

If the levelling capabilities of the scanner could be predicted, it would be possible to anticipate the existence and the extent of the uncorrected parts of the chip.

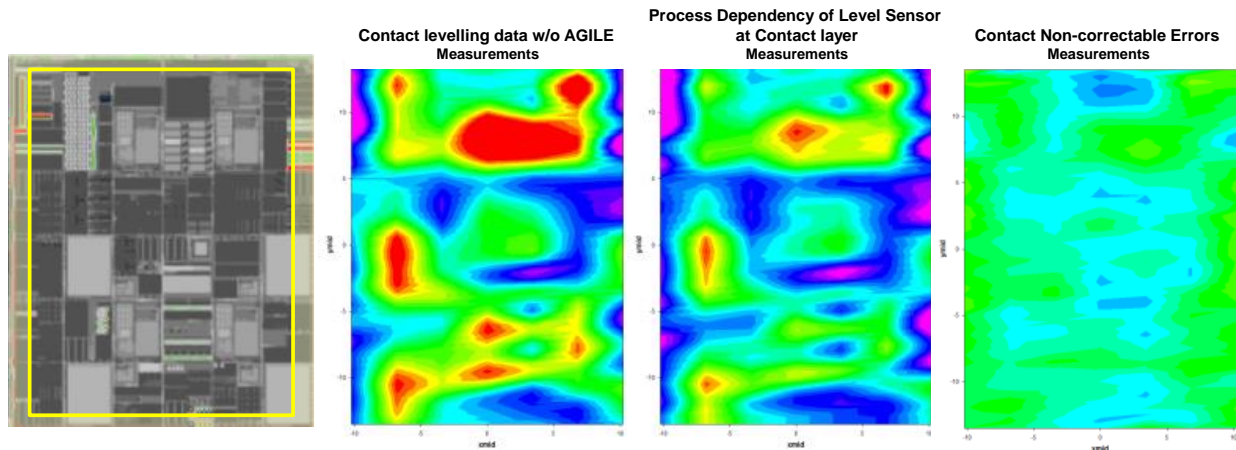


Figure 6: Scanner levelling reports vs. Product layout (the yellow box is the part of the chip that is being measured by the levelling system of the scanner). Visually, the shape of the levelling fingerprints seems to be related to the product layout.

The methodology consists of reconstructing what is on the wafer before the exposure step using all the previous layer design data from GDS's. Extractions were performed for every previous patterning layer. These may be single patterning, double patterning or the convolution of two layers that are expected to contribute to topography together. It can also be a perimeter density or a spatial density maps. All underlying layers were investigated but only the ones that showed correlation were kept.

As Figure 6 suggests, it seems that a correlation exists between the product layout and the levelling data of the scanner. The optimal underlying GDS density selection for achieving best possible correlation is shown in figure 7 for three 14FD-SOI layers.

Due to the fact that it was expected to have more topography, the Contact layer has been specifically studied into more details. GDS extracted densities were used as X parameters for the model building and the 14FD-SOI Contact layer scanner reports were set as the output of the modelling. Running the simulation several times, it was possible to select the best combination of GDS data and the layers with the most influence on topography, level sensor performances, levelling non-correctable errors. The same GDS inputs were used in all cases.

These results are presented in Figure 8. It shows the reconstructed intrafield levelling fingerprint using the modelled data versus the measurements as well as correlation plots and coefficients (R^2 and Q^2).

	X-inputs:	Y-inputs:		
	GDS densities	Scanner levelling reports for layer N		
	PREVIOUS LAYERS	TRENCH	CONTACT	METAL 1
Single layers input Density and Perimeter	<i>Layer N-7</i>	X	X	X
	<i>Layer N-6</i>	X	X	X
Convolutions of layers (double patterning, stack effects...) Density and Perimeter	<i>Convolution 1</i>	X	X	X
	<i>Convolution 2</i>	X	X	X
	<i>Convolution 3</i>	X	X	X
	<i>Convolution 4</i>		X	X
	<i>Convolution 5</i>			X

Figure 7: Underlying layers GDS densities used for correlating Trench, Contact and Metal 1 layers

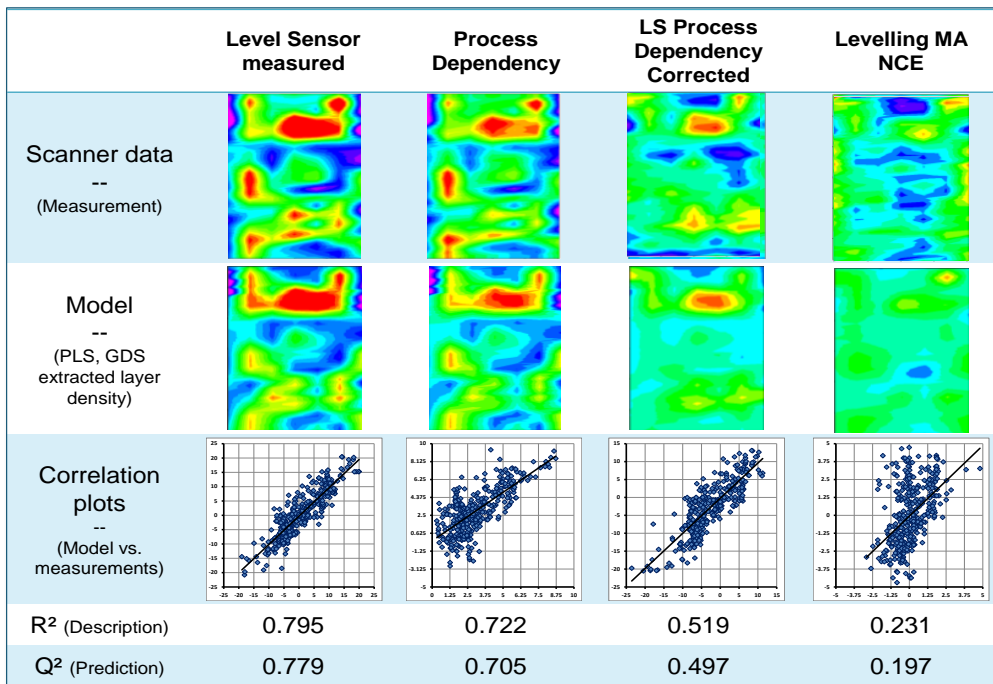


Figure 8: Measured and predicted levelling fingerprints; Correlation plots; description correlation coefficient R² and expected prediction capabilities Q² of the model for intrafield levelling data at Contact layer

1. Level Sensor without AGILE

For the level sensor alone, the model has a R² of 0.80 and a Q² of 0.78. This shows that the level sensor measurements have a high correlation to the stack densities. The next step is to check if this reading can be calculated beforehand using another maskset. The level sensor measurement is giving both topography measurements and a certain amount of optical modulation induced reading error. The two parts are explained in sections III.2 and III.3.

2. Process Dependency

The delta map between the LS and the LS with AGILE measurement contains process dependency of the LS. It is an indirect measurement of the optical modulation of the stack.

With the model, about 72% of this reading error can be explained and around 71% could be predicted. As shown in Figure 9, SOI (Silicon on Insulator) parts of the chip measure a lower topography compared to the rest of the field which contains some No SOI, i.e. Bulk, zones. This can be explained simply by the presence of a reflective Silica layer in the SOI part. This layer explains about 50% of the reading errors.

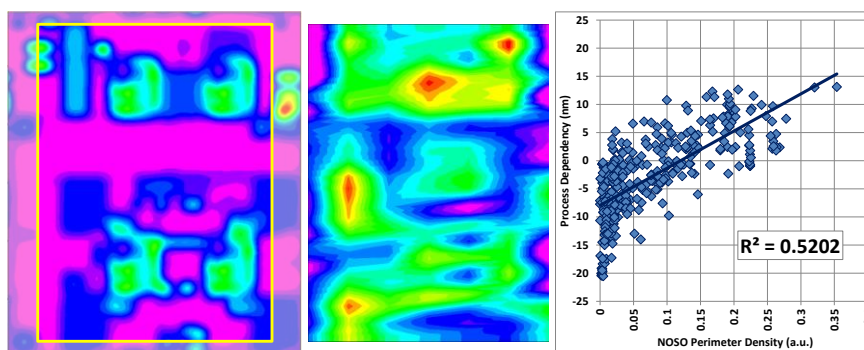


Figure 9: SOI/Bulk intrafield layer density (purple is SOI and blue and green are Bulk substrates, at the left) vs. scanner measured process dependency (middle) and correlation plot between the two (right).

Once again, the absence of trilayer planarization effects causes some error. Other processes such as CMP or doping could be added to the model in order to increase the correlation coefficients since these layers create different stacks that have different effects on the optical LS and may have an influence topography build-up.

3. Level Sensor Process Dependency corrected

Using the AGILE sensor, it is possible to extract the wafer topography without the process dependency error that may be present in the optical LS measurements.

The model shows a $0.52R^2$ correlation to product layout and a prediction of around 50% can be expected applying the model to GDS extracted data. The correlation is lower than for the LS alone but this was expected since the material distribution induced optical modulation is not taken into account. The model also only considers the patterned material that is present on the wafer and the planarizing effects of the litho stack (trilayer) were not added to the modelling parameters. As expected, it is the previous layer (named Trench) that has the largest influence on the wafer topography at the Contact layer.

4. Levelling Non-Correctable Errors

Non-correctable errors do not show a strong correlation to the GDS density model ($R^2 = 0.23$, $Q^2 = 0.20$). This is expected since the non-correctable errors include the scanner levelling model (in which the exposure slit is fitted through the data) and the GDS design does not.

As a conclusion to part III, this model has been built on a specific maskset and has yet to be checked with another dataset. The model will be applied on the GDS densities of the new product and compared with real-life measurements (FEM, Levelling reports, topography) on silicon wafers.

IV – 7 WAFER FEM ANALYSIS

In order to qualify on-product total focus uniformity performance, 7 product wafers were exposed with focus offsets in a range of ± 45 nm around best focus. Per full field 52 identical product features were measured by CD-SEM. For each of the almost 5000 points on the wafer the local defocus is fitted, as shown in Figure 10. A plot of the measured CD values as a function of focus setpoint plus local defocus shows the fit quality as the ‘bandwidth’ of points around the fitted Bossung shape; see Figure 11. The remaining bandwidth is the result of inevitable scanner wafer-to-wafer variation and SEM noise.

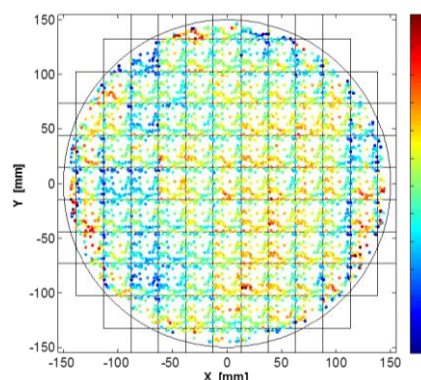


Figure 10: Defocus map obtained from multi wafer FEM; the estimated accuracy per point is around 3 nm 3s

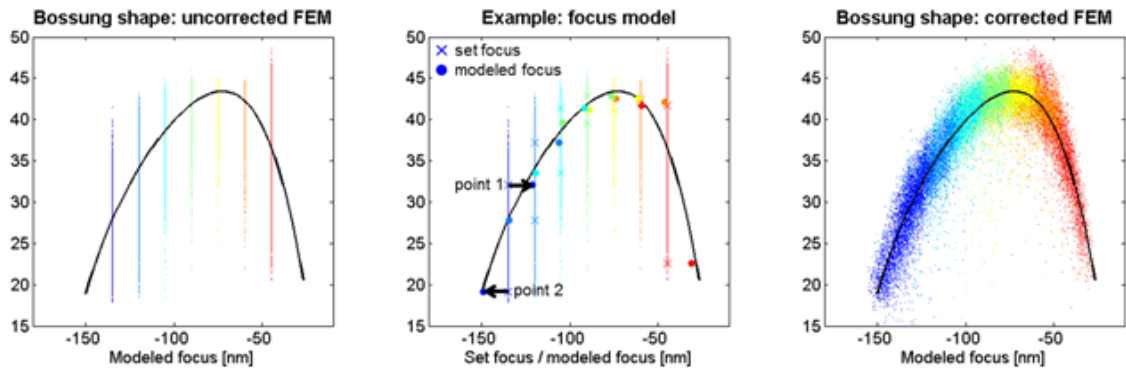


Figure 11: Bossung plot with corrected focus axis indicates regression quality. In the center graph the focus axis correction is shown for two points on the wafer. All 7 wafers (focus setpoints) belonging to that wafer point get shifted by the same modelled defocus.

Figure 12 shows the average field defocus map. All fields not too far to the edge are used for averaging. Levelling MAz residual correction is applied.

In-die on-product focus measurement and control could be used to correct for the linear wedge (tilt over Y) component. For correction of the average X curvature, lens tuning is needed. Locally (between neighbouring points) no large focus offsets are seen, which could be because all measured points are in sub-scribes, where similar topography is expected.

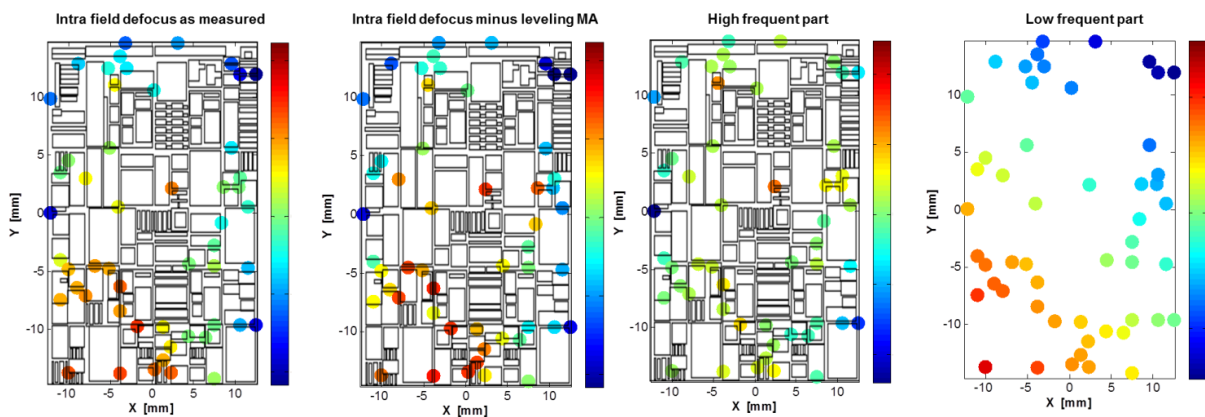


Figure 12: Measured intra field defocus map with and without leveling MA correction; the latter is also split into a low order shape and a spatially high frequent part (to be compared to topography measurements); the estimated accuracy per point is around $2 \text{ nm } 3\sigma$

In Figure 13 we see that the on-product inter field defocus map contains global and more local components. The difference map contains a clear global shape. Spatially intermediate frequencies like field offsets are caught by both FEM and SSF. These frequencies do catch our eye, and that's why the resemblance between FEM and SSF seems so good. Actually around 50% of the on-product defocus variance measured by FEM is caught by the SSF measurement.

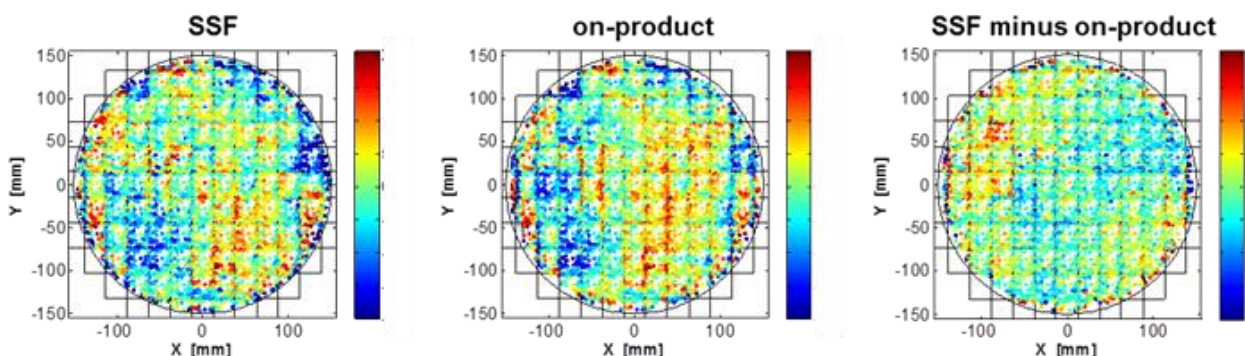


Figure 13: About 50% of the inter field focus variance seen by FEM is also seen by Single Shot Focal (SSF)

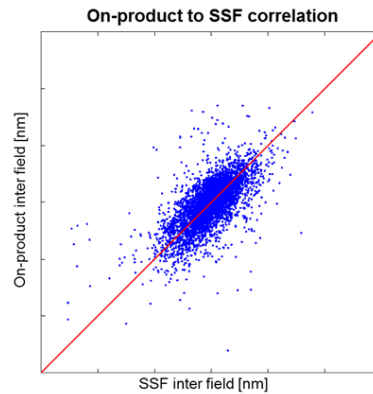


Figure 14: A correlation plot between FEM and SSF shows there is a shape difference ('cloud width'), but not scaling error ('cloud on red 1:1 line')

CONCLUSION

Reticle design GDS maps show high ($\sim 0.7 R^2$) correlation to, predominantly, level sensor raw data and level sensor process dependency. The next step would be to verify this correlation on another mask set. If correlation is seen this could be used to feedforward focus and levelling optimization, improve FEM strategies and die layout optimization. To lesser extent a correlation ($\sim 0.5 R^2$) is seen to wafer topography measured by the scanner level sensor. Further work is ongoing to improve the correlation by measuring micrometre scale topography maps using offline optical tooling. This correlation, also, needs to be verified on another mask set. A strong visual correlation is seen between SSF and on-product focus fingerprints and about 50% of the variance seen by 7-wafer FEM are also seen by Single Shot Focal (SSF). This indicates that SSF can be used to give a good representation of interfield focus uniformity seen at wafer level. The intrafield on-product focus is dominated by a lower order fingerprint. This can be further corrected by the scanner using wafer tilts. The higher order topography effects are not dominant in this measurement because of the choice of measurement features. Further work can be done by measuring features within the product and the sub-subscribes together.

This paper has shown correlations between level sensor and reticle design GDS. The next step will be to investigate correlation between on-product focus and topography measurements, and close the circle by confirming the correlation with reticle design GDS maps.

The layout dependent local topography investigation in this paper is also linked to focus induced defectivity. The defect detection and control of this is covered in the paper and talk: A new paradigm for inline detection and control of patterning defects [8].

REFERENCES:

- [1] Seltmann, R., "28nm node process optimization: A Lithographic Centric View", EMLC 30, Proc. of SPIE 9231, 2014
- [2] Katakamsetty, U., Colin, H. et al., "Scanner correction capabilities aware CMP / Lithography hotspot analysis", Proc. of SPIE Vol. 9053, 2014
- [3] Colin, H., Bin, W. X., et al., "Hotspot Detection and Design Recommendation Using Silicon Calibrated CMP Model", Proc. of SPIE 7275, 2009
- [4] Brunner, T., Menon, V. et al., "Characterization and mitigation of overlay error on silicon wafers with non-uniform stress", Proc. of SPIE Vol. 9052, 2014
- [5] Jang, J. H., Park, T. et al., "Focus control budget analysis for critical layers of flash devices", Proc. of SPIE Vol. 9050, 2014
- [6] ASML Netherlands B.V., "Level sensor for lithographic apparatus", US7265364 B2, Jun 10, 2004
- [7] Goos, F. and Hänchen, H., "Ein neuer und fundamentaler Versuch zur Totalreflexion", Ann. Phys. (436) 7–8, 333–346 (1947)
- [8] Hunsche, S., Jochemsen, M., et al., "A new paradigm for inline detection and control of patterning defects", Proc. Of SPIE Vol, 9424-47

Product layout induced topography effects on intrafield levelling

J-G. Simiz^{1,4}, T. Hasan², F. Staals², B. Le-Gratiet¹, W.T. Tel², C. Prentice³, A. Tishchenko⁴

¹STMicroelectronics, 850 rue Jean Monnet, F-38926 Crolles Cedex, France

²ASML, De Run 6501, 5504DR Veldhoven, the Netherlands

³ASML SARL, 459 chemin des Fontaines, F-38190 Bernin, France

⁴LaHC CNRS-UMR 5516, 18 Rue du Professeur Benoît Laurus, F-42000 Saint-Étienne, France

Abstract:

With continuing dimension shrinkage using the TWINSCAN NXT:1950i scanner on the 28nm node and beyond, the imaging depth of focus (DOF) becomes more critical. Focus budget breakdown studies [Ref 2, 5] show that even though the intrafield component stays the same, it becomes a larger relative percentage of the overall DOF. Process induced topography along with reduced Process Window can lead to yield limitations and defectivity issues on the wafer. In a previous paper, the feasibility of anticipating the scanner levelling measurements (Level Sensor, Agile and Topography) has been shown [1]. This model, built using a multiple variable analysis (PLS: Partial Least Square regression) and GDS densities at different layers showed prediction capabilities of the scanner topography readings up to 0.78 Q^2 (the equivalent of R^2 for expected prediction). Using this model, care areas can be defined as parts of the field that cannot be seen nor corrected by the scanner, which can lead to local DOF shrinkage and printing issues. This paper will investigate the link between the care areas and the intrafield focus that can be seen at the wafer level, using offline topography measurements as a reference. Some improvements made on the model are also presented.

KEYWORDS:

depth of focus, intrafield, scanner levelling, topography, GDS, product design layout effect, PLS regression analysis, optical lithography

INTRODUCTION

For 193 nm immersion lithography focus control is limited by the topography measurement accuracy combined with the scanner correction capability. For the critical features on the layer investigated, the depth of focus is of the same order as the scanner correction capability. This is, in part, driven by high frequency topography effects that cannot be handled fully by the scanner's wafer levelling & focussing systems. Product layout induced topography is an important factor that combined with tight focus control and low DOF values can lead to local yield loss.

In this paper, topography correlation to on-product focus was investigated with the ultimate goal of linking GDS to topography and focus. A new way of using topography data and product layout knowledge is, also, presented by using this information to determine optimized weighting factors during scanner levelling. This work was done on the Contact layer on the 14FD-SOI development shuttle.

I – METHODOLOGY / CONCEPT

In a previous paper [1], the possibility of modelling scanner levelling using design densities was investigated. This could be used to test and optimize the intrafield levelling friendliness before any silicon is exposed on the tool. This multi-source data analysis uses scanner log files, GDS's, on-product and bare-wafer focus. They are processed together in a multivariate analysis tooling (Partial Least square regression software) in order to generate the model.

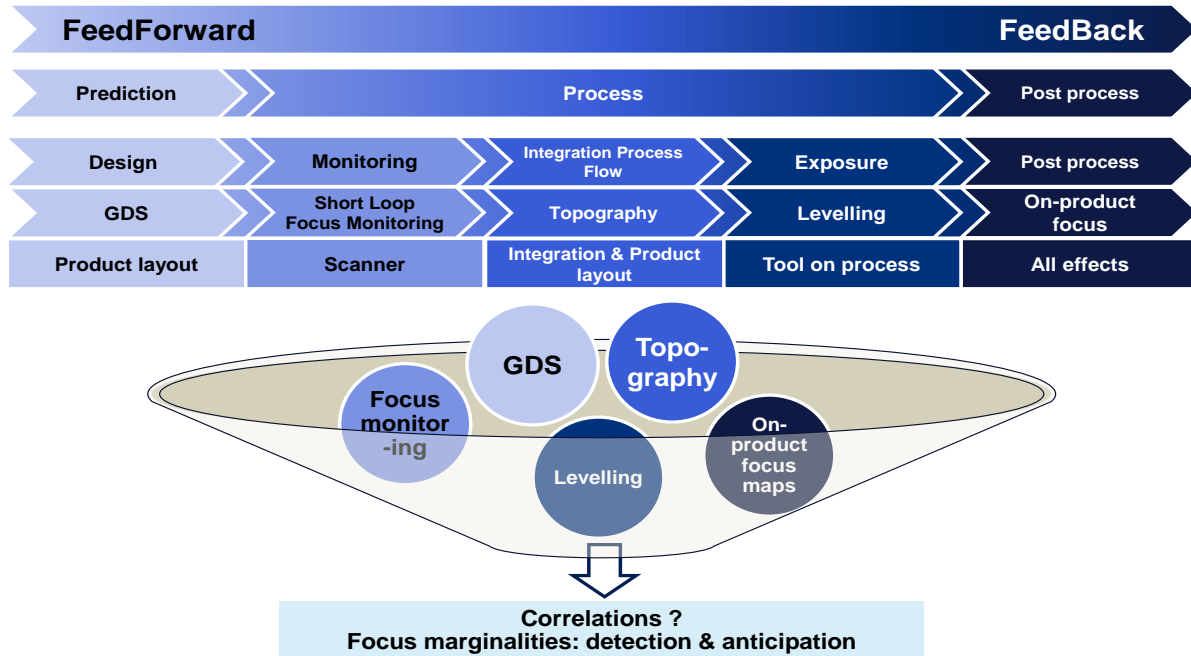


Figure 1: Different data sources and their use in the focus correlation study (GDS: Design layout, Short loop focus monitoring: Single Shot Focal test, On-product focus: Focus Uniformity Map “Bossung top best focus”) [1]

INPUT N°	INPUT NAME	SPIE 2015	EMLC 2015	GOAL
1	GDS	✓		✓
2	Scanner reports	✓		✓
3	On-product focus maps	✓	✓	✓
4	Topography		✓	✓
5	Bare-wafer focus maps	✓		✓
USE		Prediction	Correction	Link the two

Figure 2: Inputs summary and availability

II – BEST FOCUS vs. TOPOGRAPHY ANALYSIS

Local topography effect will affect on-product focus on the wafer. In order to correct for these height variations, the scanner performs a levelling optimization. This involves successively measuring the topography on the wafer and then mechanically correcting for it, by moving the stage during the exposure, to keep the wafer within focus.

However, this system cannot correct for high frequency topography variations. And the areas where the topography changes are extreme can lead to defocus. In the case of the 14FD-SOI development shuttle, several care areas were defined using data extracted from offline topography measurements. Most of these areas are not expected to be present on a product but are necessary for the development of a technology. The topography measurements were done a Veeco WYKO NT9300 tool in LETI without litho stack [7]. The measurements allow a mapping of one field with pixels of a few μm^2 . Figure 3 shows the mapping obtained for one field as well as some areas were due to high spatial frequency topography locally some high defocus is expected.

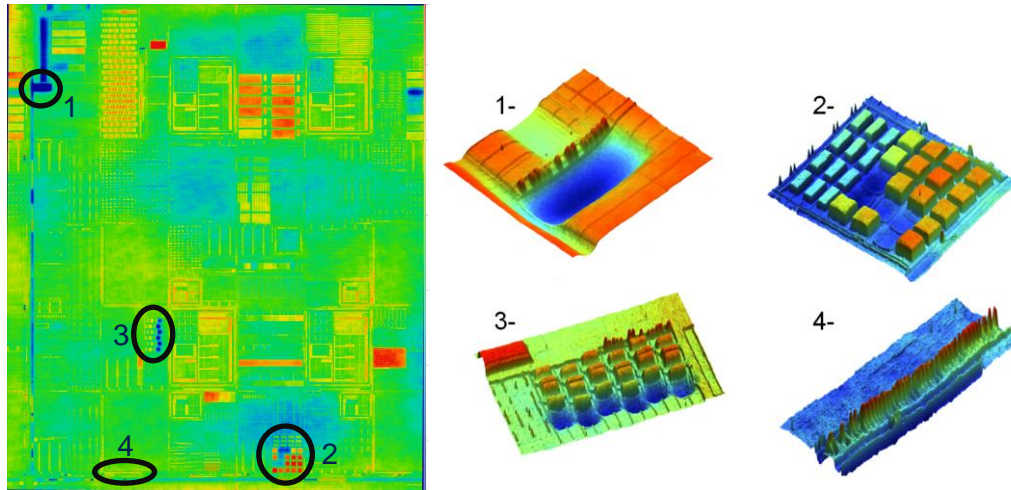


Figure 3: Wyko measurements of one field of 14FD-SOI chip at Contact layer without Litho stack. Zooming is done on some defocus care areas defined for the 14FD-SOI Contact layer

Care area 1 on the shuttle has the worse – but known atypical - local topography. We measured in this area one pattern on a focus matrix wafer for Best Focus determination and analysis. The results of the analysis are given in Figure 4. It shows that the best focus of the pattern and the topography at the same position are correlated linearly with a very high coefficient of correlation: $R^2 = 0.81$.

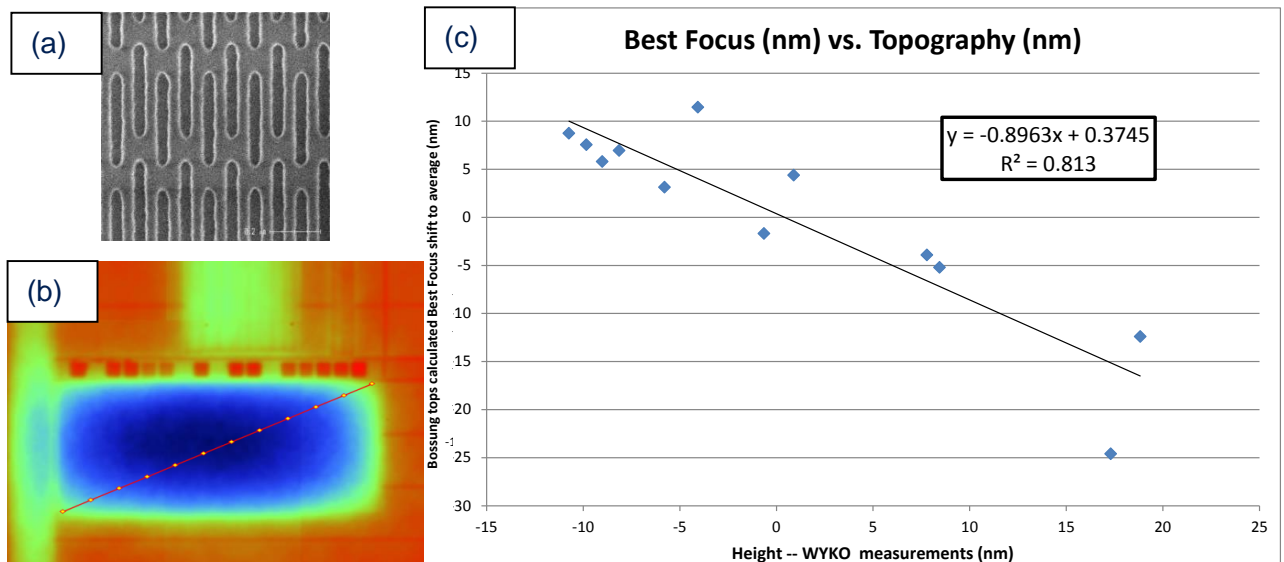


Figure 4: (a) Pattern measured – (b) Position of the measurement points on the area – (c) Best focus vs. topography correlation for care area 1.

The slope of the correlation curve is not 1 but this can be explained by the fact the reference topography measurements were done without any litho stack and that the tri-layer smooths the topography. Mask CD effects were not taken into account here and that may explain the shift of some point from the curve.

In order to test if this can also be seen on the product, some extra measurements were done within the logic, where the topography variations are much smaller. The same structure was measured in two parts of a logic block showing about 9nm height difference. The best focus shift between the different locations was about 11nm.

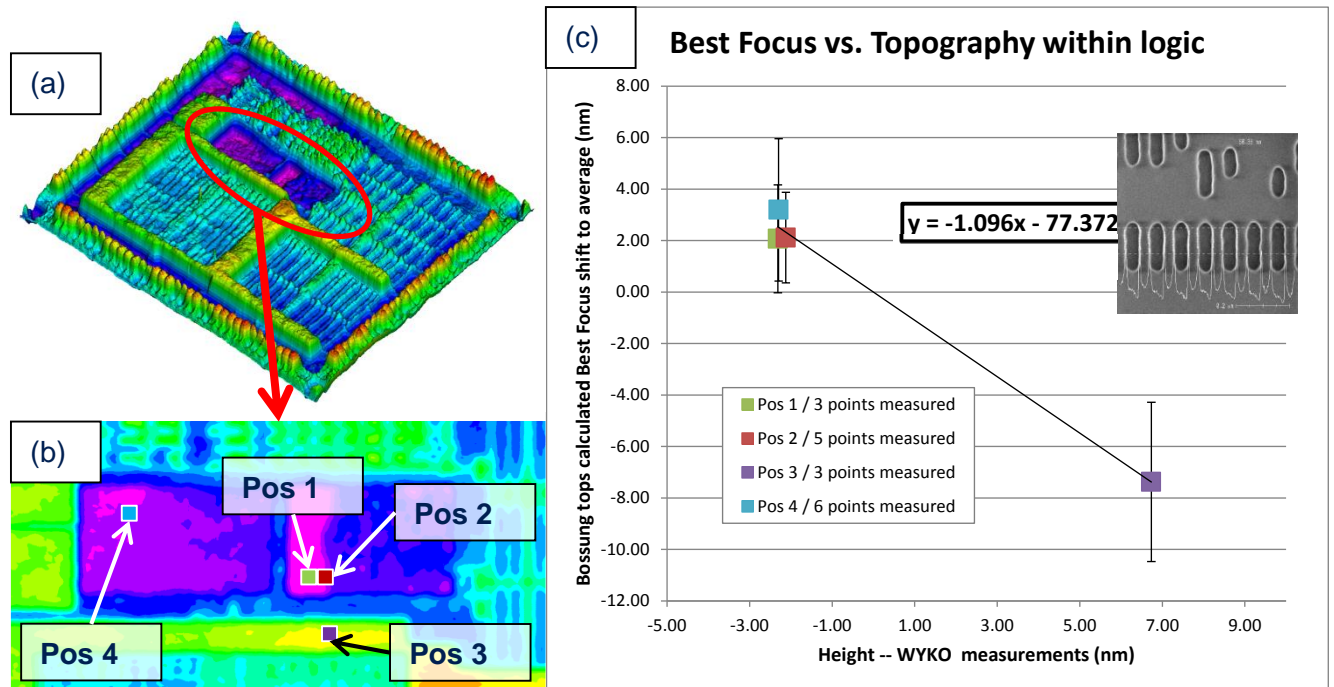


Figure 5: (a) Logic chip topography mapping – (b) Position of the measurement points on the area – (c) Best focus vs. topography correlation for logic area.

In both cases, a higher topography leads to a negative best focus shift which corresponds to how the focus is referenced in the scanner where a positive focus offset moves the imaging plane (wafer stage) away from the lens.

III – PRODUCT LAYOUT AWARE LEVELLING OPTIMIZATION

Dense topography information can be used to identify imaging critical locations that are most at risk of causing topography driven focus induced yield loss. Figure 6 shows an example of this process, where weights are specified for different locations in the field.

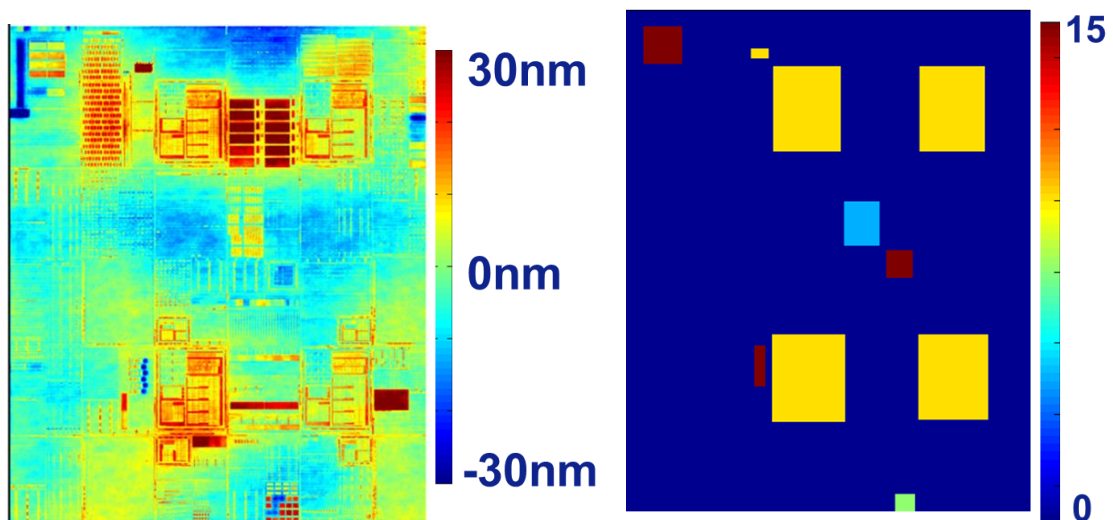


Figure 6: Intra-field topography measurements (left) and location of selected weight factors for optimization example (right)

Extra weight is given for the areas that contain most focus critical features. This is done by optimizing the slit z-offset and Ry rotation at each scan position. Figure 7 shows how the slit of the weighted fit is closer to the measured height of the features within the critical locations, with respect to standard levelling.

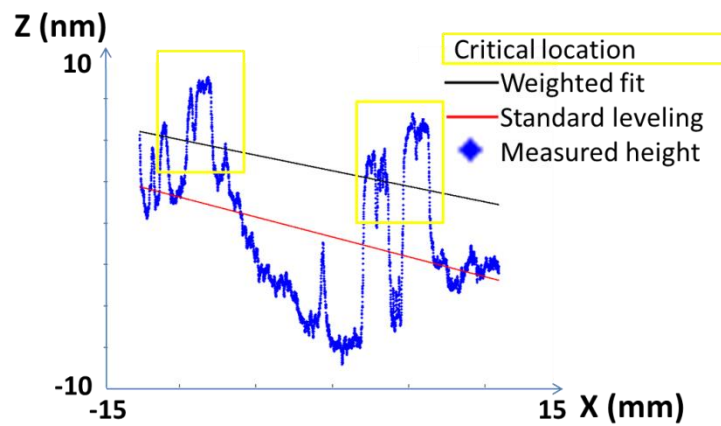


Figure 7: An example of the weighted fit versus standard levelling slit position at 1 scan location

The weighted levelling optimization results in visibly reduced levelling non-correctable moving average errors in the critical locations.

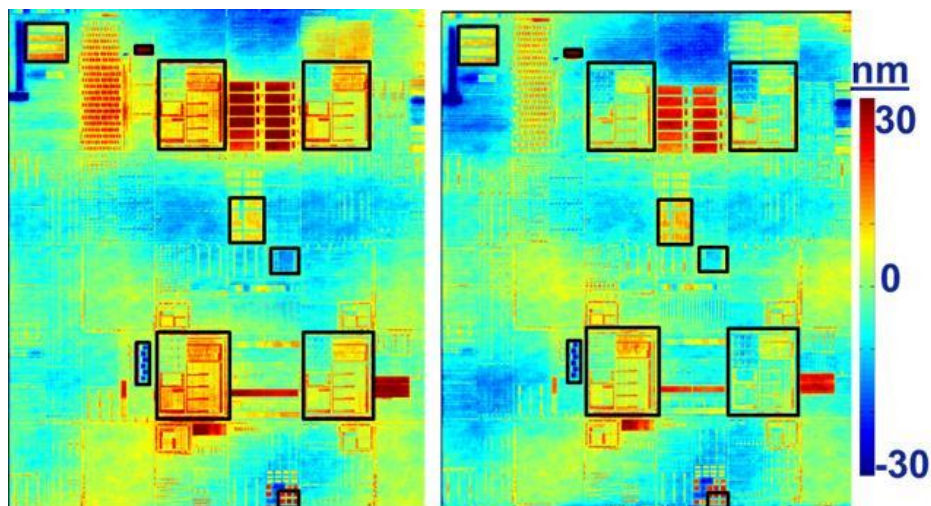


Figure 8: Levelling MA non-correctable error with regular levelling (left) and weighted levelling (right)

This method of optimizing levelling is a trade-off between improving performance in the critical areas and compromising on the performance outside the critical areas. This is visualized in figure 9. Within the critical area the amount of points with a levelling MA error <15nm increases from 90% of points to 95% of points. However, in the non-critical area the amount of points with a levelling MA error <15nm reduces from 95% to 93%. Using the topography information it is possible to visualize the impact applying weight factors before applying them on the scanner.

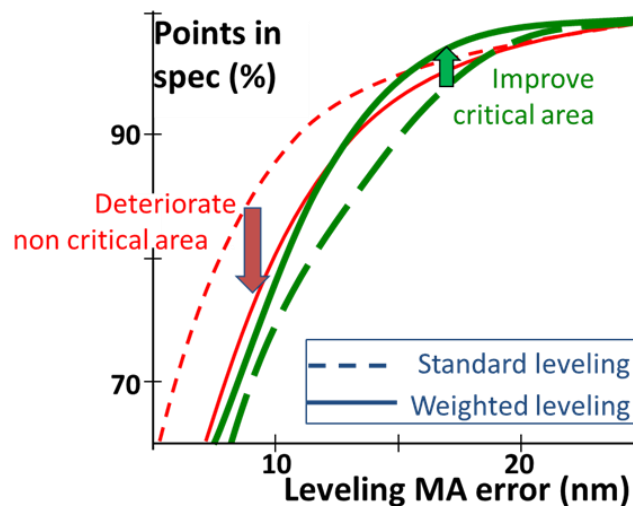


Figure 9: Levelling MA error points in spec comparison of standard levelling with respect to weighted levelling for critical and non-critical locations

CONCLUSION

Topography measurements have been correlated to on-product focus showing that a difference in height can cause focus excursions. Using full field topography measurements, it was possible to determine weight factors for different areas of the field. These weight factors were used to calculate an optimized levelling fit and ultimately correct what matters. This resulted in improved focus performance in critical areas. This method of optimization would be especially useful for development shuttles due to the fact that the process is still being optimized and that some non-critical test chips are present on the reticle.

REFERENCES:

- [1] Simiz, J-G., Hasan, T., Staals, F., Le-Gratiet, B., et al., "Predictability and impact of product layout induced topology on across-field focus control", Proc. of SPIE Vol. 9424, 2015
- [2] Seltmann, R., "28nm node process optimization: A Lithographic Centric View", EMLC 30, Proc. of SPIE 9231, 2014
- [3] Katakamsetty, U., Colin, H. et al., "Scanner correction capabilities aware CMP / Lithography hotspot analysis", Proc. of SPIE Vol. 9053, 2014
- [4] Colin, H., Bin, W. X., et al., "Hotspot Detection and Design Recommendation Using Silicon Calibrated CMP Model", Proc. of SPIE 7275, 2009
- [5] Jang, J. H., Park, T. et al., "Focus control budget analysis for critical layers of flash devices", Proc. of SPIE Vol. 9050, 2014
- [6] Hunsche, S., Jochemsen, M., et al., "A new paradigm for inline detection and control of patterning defects", Proc. Of SPIE Vol. 9424, 2015
- [7] Dettoni, F., et al., "High resolution nanotopography characterization at die scale of 28nm FD-SOI CMOS front-end CMP processes", Microelectronic Eng. 113, p105-108, 2014.

Verification and application of multi-source focus quantification

J-G. Simiz^{1,2}, T. Hasan⁵, F. Staals³, B. Le-Gratiet¹, W.T. Tel³, C. Prentice⁴, J-W. Gemmink³,
A. Tishchenko², Y. Jourlin²

¹STMicroelectronics, 850 rue Jean Monnet, F-38926 Crolles Cedex, France

²LaHC CNRS-UMR 5516, 18 Rue Professeur Benoît Lauras, F-42000 Saint-Étienne, France

³ASML, De Run 6501, 5504DR Veldhoven, the Netherlands

⁴ASML SARL, 459 Chemin des Fontaines, F-38190 Bernin, France

⁵ASML US, 399 W Trimble Rd, Jan Jose, CA 95131, United States

Abstract:

The concept of the multi-source focus correlation method was presented in 2015 [1, 2]. A more accurate understanding of real on-product focus can be obtained by gathering information from different sectors: design, scanner short loop monitoring, scanner leveling, on-product focus and topography.

This work will show that chip topography can be predicted from reticle density and perimeter density data, including experimental proof. Different pixel sizes are used to perform the correlation in-line with the minimum resolution, correlation length of CMP effects and the spot size of the scanner level sensor. Potential applications of the topography determination will be evaluated, including optimizing scanner leveling by ignoring non-critical parts of the field, and without the need for time-consuming offline topography measurements.

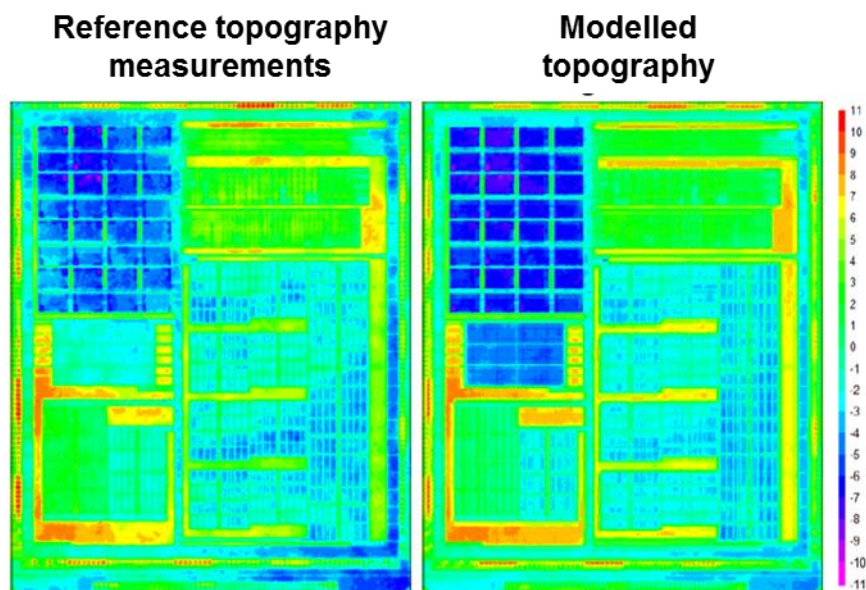


Fig 1: High resolution topography (Left: Measured; Right: Simulated) – Colour scale is in nm

KEYWORDS:

Depth of focus, intrafield, scanner leveling, topography, scanner, product design layout effect, PLS regression analysis

INTRODUCTION

The concept of the multi-source focus correlation method was presented in 2015 [1, 2]. A more accurate understanding of real on-product focus can be obtained by gathering information from different sectors: design, scanner short loop monitoring, scanner leveling, on-product focus and topography. In reference [1], the link between scanner monitoring and on-product focus was established as well as the correlation between design and the scanner level sensor measured intra field non-correctable errors after leveling. In [2], the on-product focus to topography correlation has been studied and the concept of smart leveling was proposed.

This work will investigate the design to topography and design to focus correlation. It has been shown that chip topography can be predicted from reticle density and perimeter density data, including experimental proof. To visualize the potential applications different pixel sizes will be used to perform the correlation in-line with (i) minimum resolution, (ii) correlation length of Chemical Mechanical Polishing (CMP) effects and (iii) scanner level sensor spot size. Potential applications of the topography determination will be evaluated, including optimizing scanner leveling by ignoring non-critical parts of the field.

I – INTRAFIELD FOCUS CONTROL

Best focus variation, coupled with reduced depth of focus, is a major contribution to tight process margins for the 28nm and 14nm logic nodes. For both intra wafer and intra field, process and tool fingerprints are creating local shifts from the best focus, causing printing issues leading to defectivity and yield losses.

Major contributors to interfield focus are scanner fingerprints [1], edge effects [7] and leveling non-correctable of the wafer topography [2, 3, 4, and 5].

Considering the intrafield, it is possible to distinguish two families of effects: the imaging effects and the spatial effects, summarized in Fig.2. From imaging, the best focus shift is caused by wave front deformations induced in a large part by the light passing through the mask so that the image focal plane of each pattern shift apart from each other making it more difficult to print [6]. Best focus is pattern shape dependant.

Spatial effects are characterized by a best focus change across field for the same pattern. The mask is not perfect and the same pattern may not have been printed exactly the same way for each of its occurrences. Imaging effects (i.e. image plane deviation measured by FOCAL) can also effect the BF of a single pattern differently across the field [15]. The wafer topography also has an intrafield component showing non-correctable systematics that are tightly linked to the on-product focus [2]. For a given pattern, best focus depends on the intrafield position of each of its occurrences across field. This part of the intrafield focus budget is the one studied in this paper.

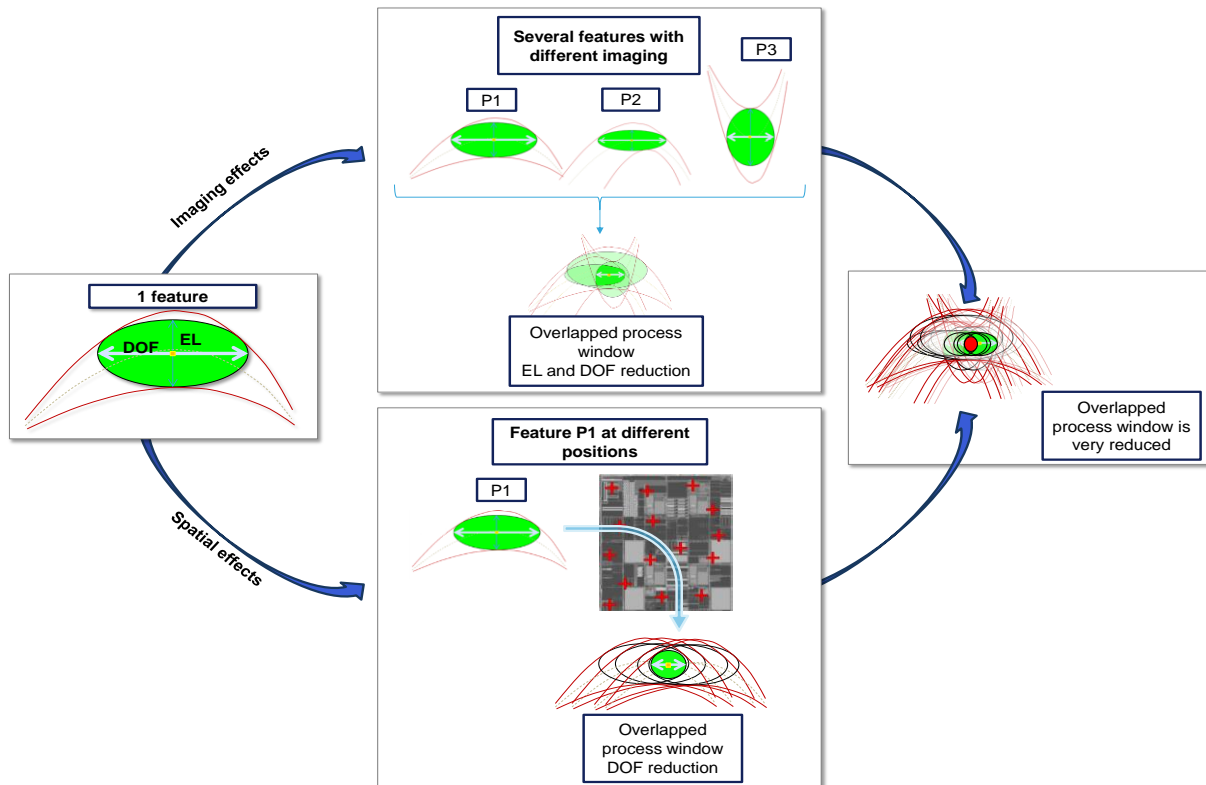


Figure 2: Brief overview of some of the sources of intrafield focus variability

II – THE TOPOGRAPHY INDUCED FOCUS NON-UNIFORMITY

As written above, focus is a function of the pattern and its position. It has been shown in reference [2] that focus distribution is mainly topography driven. Fig. 3 shows two things that can be derived from offline reference topography measurement done on a Veeco WYKO NT9300 tool in LETI without the photolithography process stack [8]. The height distribution in field (in green) which is mainly spread between 45 and 75nm but presents some excursions of a few tens of nanometres in atypical areas of the chip and the best focus to topography correlation (in blue) which shows a slope close to 1 with an excellent correlation factor.

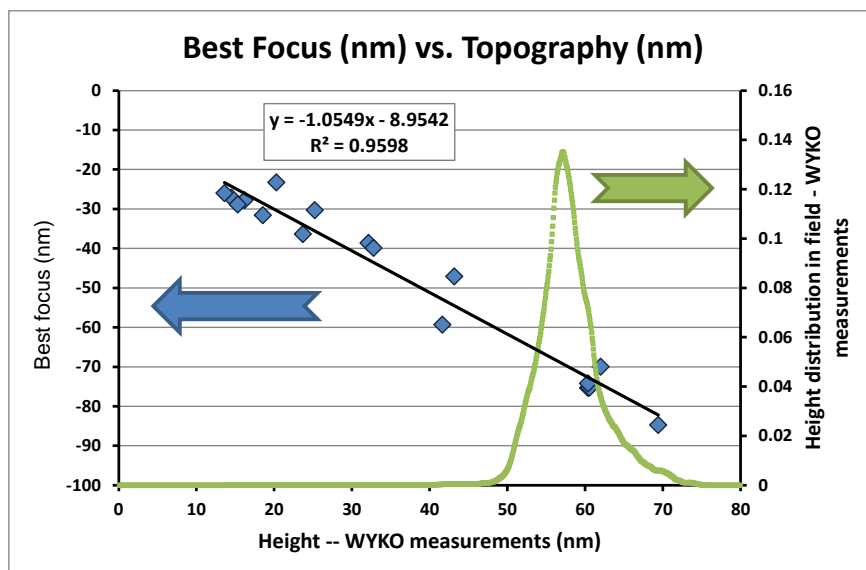


Figure 3: Correlation plot showing Bossung tops determined on-product focus vs. Local topography measured by a Veeco WYKO NT9300 tool and height distribution through chip.

A correlation was seen with the level sensor measured intrafield height map (with and without process dependency error corrected) and intra field leveling non-correctable errors. Process dependency is a measurement error due to optical stack effects on the older generation level sensor, and it is corrected by the AGILE Sensor. The PLS gives the Q^2 parameter that indicates how well a variable can be predicted. Level sensor measured intrafield height map and process dependency corrected map showed respectively $0.78Q^2$ and $0.50Q^2$ expected prediction capabilities. As one aspect of leveling is a measurement of the topography of the wafer, it is expected that the PLS coefficients w_j can be calculated for reference topography as well:

$$Topography_{Wyko\ Measured} = \sum_j(w_j LAYER_DENSITY_j) \quad (2)$$

Intrafield wafer topography is layout driven. Each part of the chip has a specific design linked to its electrical behaviour (logic cells, memory cells, electrical protection, analogic devices, antennas, etc.) which all have their specific design, dimensions and densities. This variety of different devices can create some local topography induced by patterning and polishing steps. Some tiling is done within the chip to homogenize the density and mitigate those effects but it is not always sufficient. As a consequence, the assembly of a chip within a mask field will lead to a specific topological map. Fig. 4 illustrates this fact by showing an example of 14FD-SOI (Fully depleted silicon on insulator 14nm technology node) test chip shuttle and the topography measured at the Contact layer.

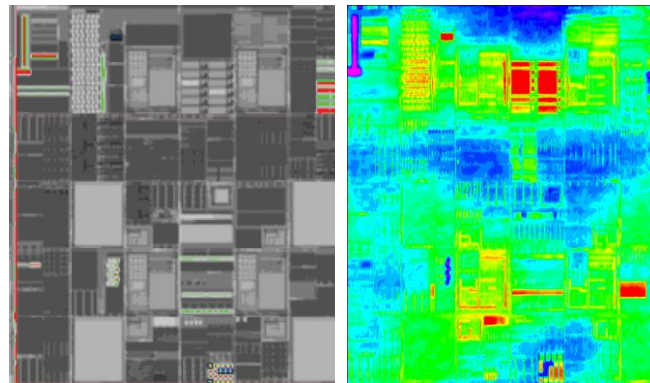


Figure 4: Product layout of the 14FD-SOI test chip shuttle and topography measured on a complete field before Contact exposure

Consequently, it is possible to create a predictive model of the intrafield topography using the GDS of the product as an input.

III – MODELLING TOPOGRAPHY WITH GDS DENSITIES

In this paper, the same methodology as the one described above (in part II) and in the SPIE 2015 paper [1] was applied in order to link the GDS to offline topography measurements. These were done on a Veeco WYKO NT9300 tool in LETI without the photolithography process stack [8]. A topography model was constructed using a partial least square linear prediction method by combining GDS densities of the underlying layers. Different models were investigated with different pixel sizes to be sensitive to several effects at multiple ranges.

- Short range effects

These effects are the ones that better match the capabilities of the Wyko in terms of spatial resolution. They are related to the direct impact of the design on a localized area.

- Long range effects

These effects are mainly process induced (deposition uniformities, CMP dishing, loading effects on etch). Some layout effects such as assembly can be responsible for these by creating areas with different designs and densities.

The PLS coefficients for each effect were calculated on a small part of the chip for 14FD-SOI Contact layer and then tested on a larger area of the field containing about 800 times more data points for validation.

This first graph (see Fig. 5) shows the comparison between the expected performance of the model regarding prediction capabilities Q^2 and the actual performances by the model R^2_{test} , for each of the pixel sizes chosen in this study.

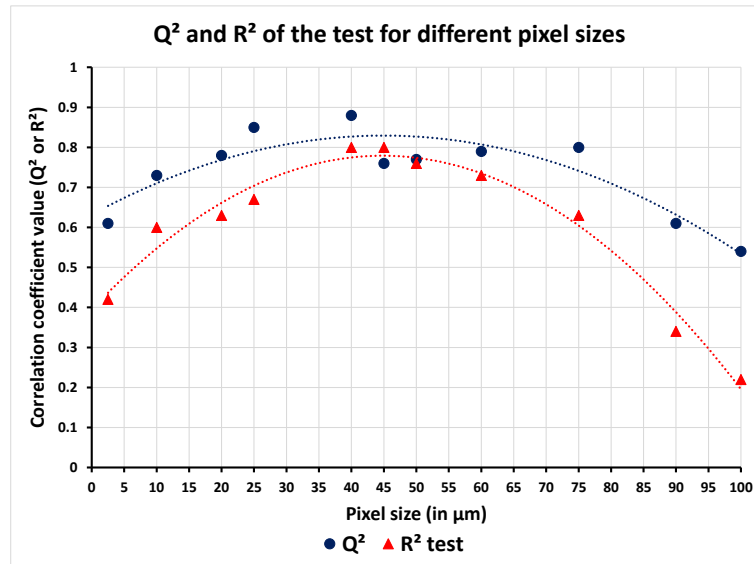


Figure 5: Graph comparing Q^2 and R^2_{test} values for several pixel sizes

The optimum model comes with pixel sizes of 40 to 50 μm . These showed the best description and expected prediction capabilities of the topography and the best results after testing on a larger area. In order to elect the optimal pixel size, it is necessary to look at other data given by the PLS model such as the predicted range of the values. A model can have a high correlation coefficient and still gives values that are not representative of measured topography. The closer the slope between measurement data and predicted data is to 1, the better the model is representative of the wafer actual topography.

The following figure, Fig. 6, gives the detailed results for the some of the best test cases. It compares model build-up, expected performances vs. actual performance (correlation coefficient, slope) and mappings of the data.

The slope of the correlation plot between measurement and model data is never exactly equal to 1. This can be explained by the fact that the models used in this work are empirical and based on the extraction of the density data at a given pixel size and also because the measurement data is also averaged to the same resolution to perform the validation of the results. This averaging, especially with large pixels, will induce a reduction of the total range of heights since highest peaks and lowest valleys are smoothed by the data averaging.

This second dataset shows best performing models to be the 40 and 45 μm -pixel ones. The model parameters were calculated without full use of the PLS analysis capabilities. This statistical method gives not only model coefficients and prediction capabilities forecast but also a ranking of the importance of each input by evaluating the VIP (Variable Importance in the Projection) of every component [12, 13]. This VIP gives the components that are actually discriminating in the regression – in this case, the layers that will have a substantial influence in the topography built-up.

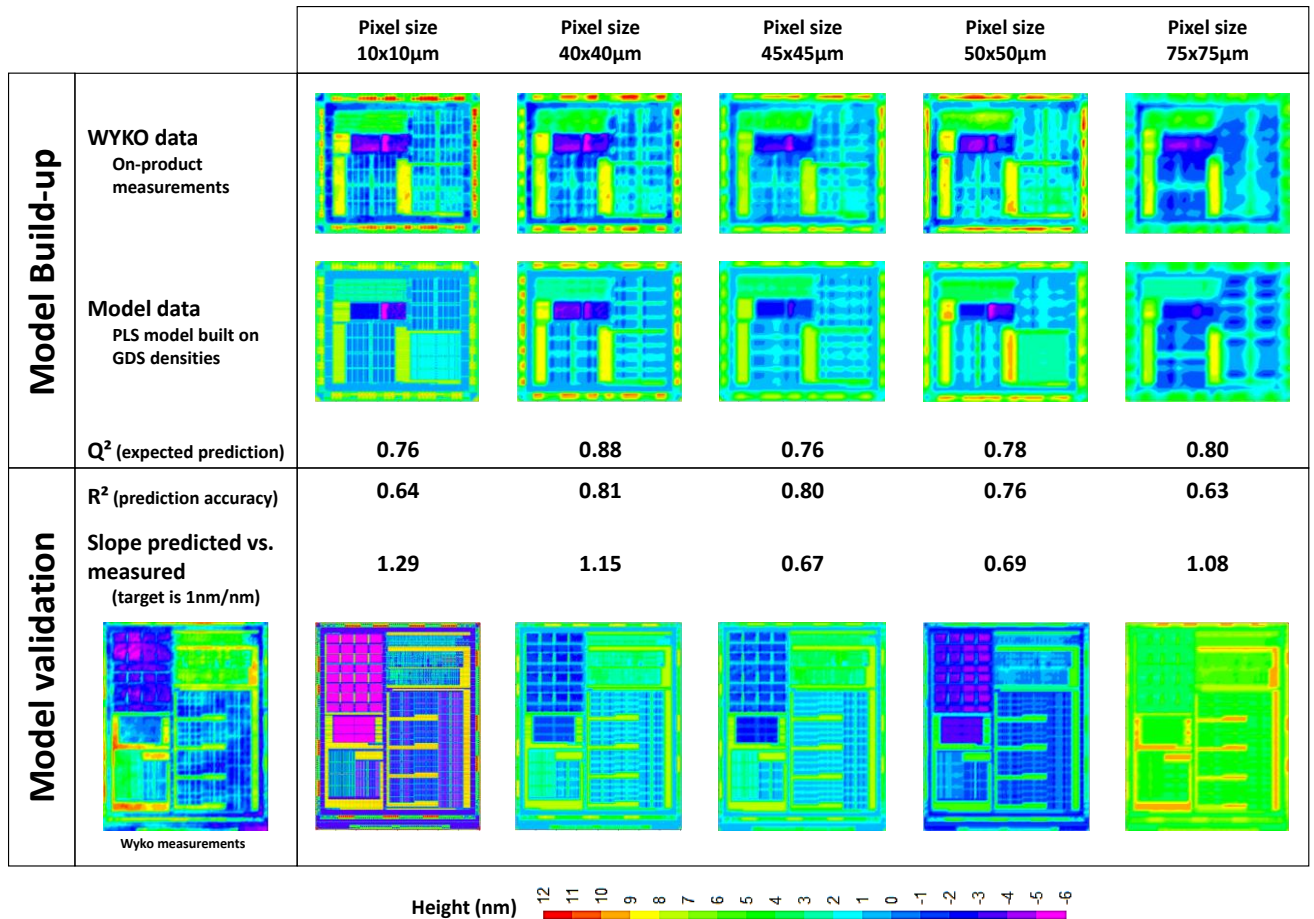


Figure 6: Performances of a few models for topography prediction

- Model build-up (~3mm²): R² shows the data description capability and Q² the data expected prediction capability of the model
- Model application on a larger area (20mm²) is compared to high frequency measurement data

It has been show above that the topography could be predicted using a linear combination of layer densities and perimeters determined by PLS analysis. As the focus is related to the local height by a linear relation since most of these high spatially frequent topography are scanner leveling non-correctable, local defocus can be predicted through this modelling. Combining relation (2) and the trend line given by Fig. 2 gives:

$$\begin{cases} Topography = \sum_j(w_j LAYER_{DENSITY_j}) \\ Best\ Focus = (-1) * Topography + b \end{cases} \Leftrightarrow Best\ Focus = (-1) * \sum_j(w_j LAYER_{DENSITY_j}) + b \quad (3)$$

This prediction enables the definition of care areas i.e. areas in which critical patterns in terms of imaging, determined by full chip LMC [9, 10], and high local topography variation might be found, causing high probability of patterning failure. The capability of defining smart care areas is key for process improvement efficiency. Using the VIP information as an input to the design may allow an improvement of the dummification process and chip layout. Simulated topography can also serve as an input for Process Window Optimizer (PWO) but for control plan optimisation by giving care areas to be monitored. It can also be used as an input for a smart scanner leveling capable of correcting preferentially the topography where it matters. This last solution was investigated in this work and is discussed in the next section of the paper.

IV – FOCUS CONTROL THROUGH SCANNER LEVELING OPTIMIZATION

High-resolution intrafield topography modelled from the design layout density can be a useful source to tailor scanner leveling for optimizing the focus control within the care area. Scanner leveling can be referred to as the process of correctly positioning the fields of the wafer in (best) focus during exposure. Leveling is limited by, among others, the physical shape of the exposure slit of the scanner. In the following paragraphs, we proposed a scheme for care and ignored area driven leveling.

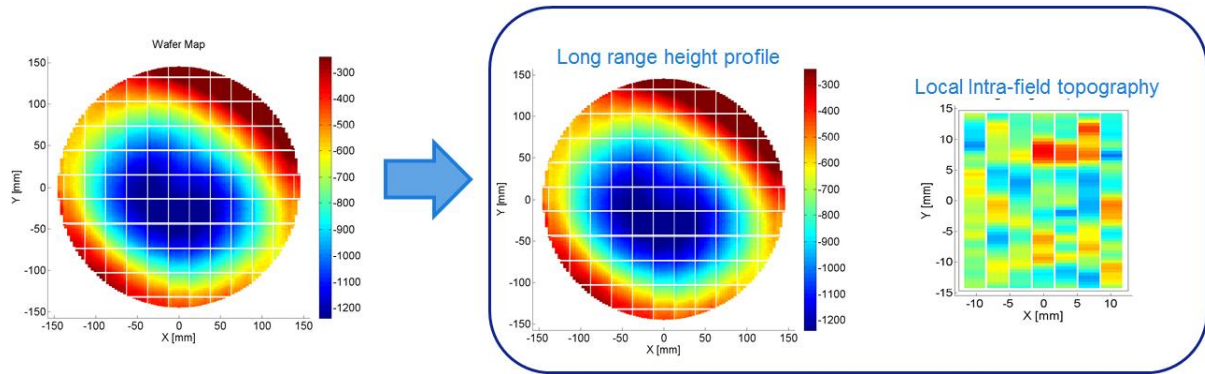


Figure 7: Wafer height map decomposed into long range height profile and local intrafield topography.

Leveling uses scanner level sensor measured surface height information of the chucked wafer, known by ‘wafer map’, as input. Figure 7 shows a wafer map which contains multiple areas corresponding to dies or fields. A full wafer map can be decomposed into two parts; a long range height profile (inter field) that represents the global shape of the wafer and a local (intrafield) height map that is sufficiently repetitive among fields. These two parts can be processed separately by the leveling algorithm, and then algebraically summed to generate the wafer stage positioning setpoints. Additionally, in our previous communication [1], and sections above, we showed that layout design density modelled topography map reasonably correlate with the level sensor measured local intrafield topography or even high-resolution topography measurement with external tools (e.g. Wyko). Consequently, we can substitute this local intra-field height map with the intra-field modeled topography map. Thanks to its high resolution, it will allow us to define accurately the care and ignore area and thus, to enable their preferential treatment in leveling for better focus control.

In figure 8 (left), some areas within the field are highlighted as care (green), ignore (red), and non-specific (orange) area. These areas are identified based on the process and design knowledge. However, they can also be identified using other computational lithography products. These defined areas are then mapped into the dense intra-field topography in figure 8 (right). Although we can, in principle, use the modelled intrafield topography, we used the external tool measured topography in this example since modelled topography was not available for the full field.

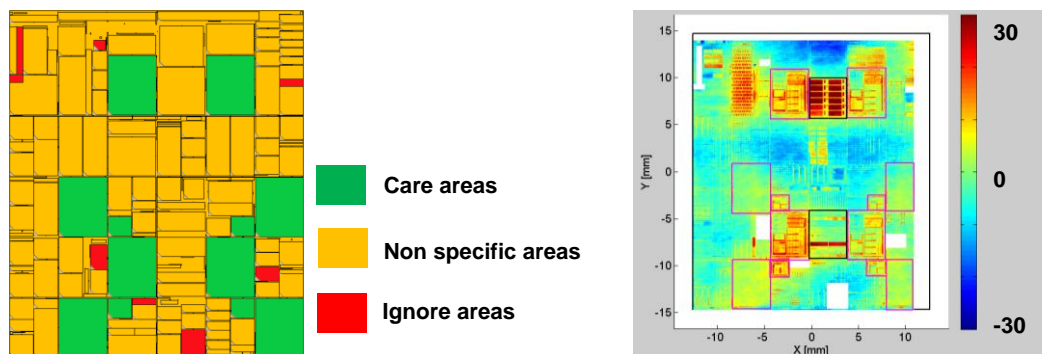


Figure 8: Care (green), ignore (red), non-specific (orange) area are highlighted on the design layout (left). Care and ignore area are then mapped into the intrafield topography (right). In right figure, height values of the ignore areas are removed (white) and care area are given higher emphasis and marked with rectangles.

A modified (weighted) scanner leveling algorithm was applied offline to this care-and-ignore area mapped intrafield topography. The non-correctable errors at care area after such leveling are reduced leading to improved focus control. Figure 9 compares (delta) the non-correctable error after the dense topography-assisted, care-and-ignore area driven leveling and the standard (all area are equally weighted) leveling. Positive (red) values show improvement. Figure 9 shows care area improved (up to 9nm in some area) with a tradeoff deterioration at the non-specific or ignore area. Figure 9 (right) shows the cumulative distribution of the non-correctable error samples (points) within the field. Within the care area, 3% more samples are now in the spec (15nm) without pushing any locations within the non-specific or ignore area beyond their spec (25nm). Spec for the non-specific/ignore area is larger than that of the care area. Note, although the choice of the spec values in this example is arbitrary, they are representative for this process. Further optimization can lead to larger improvement in the care area.

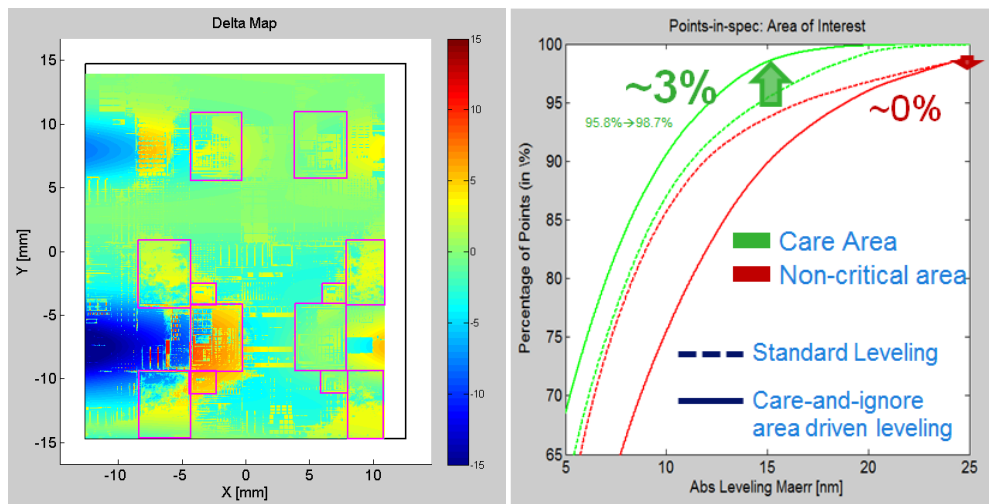


Figure 9: (left) Delta map of the non-correctable errors from care-and-ignore area driven leveling and the standard leveling. A positive value means improvement. The right figure shows cumulative distribution of the leveling non-correctable error.

Note, this proposed technique is not currently available in the scanner. Therefore, we also discussed a possibility to achieve similar performance in the scanner by a feed-forward correction mechanism exploiting the sub-recipe interface of the scanner. The delta between the care-and-ignore area optimized leveling and the standard one can be applied to the scanner through the feed-forward interface after converting them into a correction format suitable for that interface. This dense topography assisted care-and-ignore driven leveling can be used to find optimal configurations of scanner leveling improving edge die focus [14].

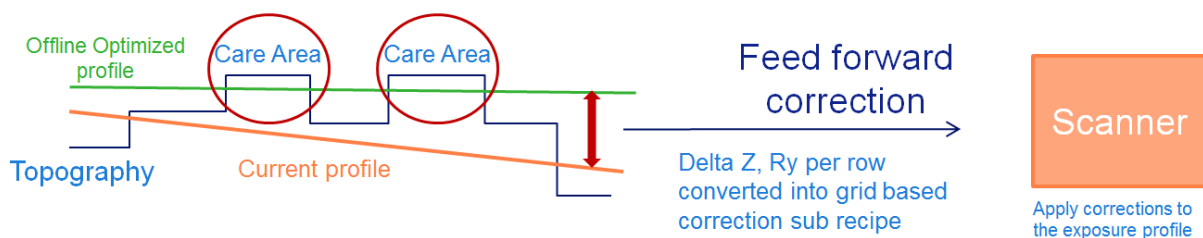


Figure 10: A schematic representation showing the possibility to enable care-and-ignore area based optimization through scanner's subrecipe interface feed forward subrecipe correction interface.

CONCLUSION

Topography measurements have been modelled and predicted with the Partial Least Square methodology with a high predictability capability up to 0.8Q². The topography being a cause of focus intrafield excursions, this method allows the definition of care areas where focus margin will be reduced and defectivity may occur at a higher rate.

The use of this data as an input for PWO [9, 10] will allow an improved defect prediction. The care areas were also used as an input for a smart leveling methodology where the scanner will correct preferentially the critical areas of the chip. Some other applications of the topography data, modelled or measured, would be tiling optimization and securing the reticle assembly.

REFERENCES:

- [1] Simiz, J-G., et al., "Predictability and impact of product layout induced topology on across-field focus control," Proceedings of SPIE 9424, (2015)
- [2] Simiz, J-G., et al., "Product layout induced topography effects on intrafield leveling," EMLC 31, Proceedings of SPIE 9661, (2015)
- [3] Lee, H., et al., "Improvement of depth of focus control using wafer geometry," Proceedings of SPIE 9424, (2015)
- [4] Seltmann, R., "28nm node process optimization: A Lithographic Centric View," EMLC 30, Proceedings of SPIE 9231, (2014)
- [5] Katakamsetty, U., Colin, H. et al., "Scanner correction capabilities aware CMP / Lithography hotspot analysis," Proceedings of SPIE 9053, (2014)
- [6] Zine El Abidine, N., et al., "Mask Blank Optimization through rigorous EMF approach from IDM perspective, for 28 nm node and beyond," Photomask Japan 22, (2015)
- [7] Le-Gratiet, B., et al., "An evaluation of edge roll off on 28nm FD-SOI (fully depleted silicon on insulator) product," Proceedings of SPIE 9778, (2016)
- [8] Dettoni, F., et al., "High resolution nanotopography characterization at die scale of 28nm FD-SOI CMOS front-end CMP processes," Microelectronic Eng. 113, p105-108 (2014)
- [9] Hunsche, S., Jochemsen, M., et al., "A new paradigm for inline detection and control of patterning defects," Proceedings Of SPIE 9424, (2015)
- [10] Fanton, P., et al., "Process Window Optimizer for pattern based defect prediction on 28nm Metal Layer," Proceedings of SPIE 9778, (2016)
- [11] Gatefait, M., et al., "AGILE integration into APC for high mix logic fab," EMLC 31, Proceedings of SPIE 9661, (2015)
- [12] Eriksson, L., Johansson, E., [Multi and Mega-Variate Data Analysis], Umetrics Academy, UMETRICS AB, pp.85 & 390 (2006)
- [13] Wold, S., et al., "PLS," In: Kubinyi, H., (ed.), [3D-QSAR in Drug design, Theory, Methods, and Applications], ESCOM Science, Ledien, 523-550 (1993)
- [14] Hasan, T., et al., "Holistic, model-based optimization of edge leveling as an enabler for lithographic focus control –Application to a memory use case," Proceedings of SPIE 9778, (2016)
- [15] T. A. Brunner, "Impact of lens aberrations on optical lithography," IBM Journal of Research and Development 41, 57 (1997)

Process Window Optimizer for pattern based defect prediction on 28nm Metal Layer

P. Fanton¹, R. La Greca⁴, V. Jain³, C. Prentice⁴, J-G. Simiz^{1,2}, S. Hunsche³, B. Le-Gratiet¹, L. Depre⁴

¹STMicroelectronics, 850 rue Jean Monnet, F-38926 Crolles Cedex, France

²LaHC CNRS-UMR 5516, 18 Rue Professeur Benoît Lauras, F-42000 Saint-Étienne, France

³ASML US, 399 W Trimble Rd, Jan Jose, CA 95131, United States

⁴ASML SARL, 459 chemin des Fontaines, F-38190 Bernin, France

ABSTRACT

At the 28nm technology node and below, hot spot prediction and process window control across production wafers have become increasingly critical. We establish proof of concept for ASML's holistic lithography hot spot detection and defect monitoring flow, process window optimizer (PWO), for a 28nm metal layer process. We demonstrate prediction and verification of defect occurrence on wafer that arise from focus variations exceeding process window margins of device hotspots. We also estimate the improvement potential if design aware scanner control was applied.

INTRODUCTION

Hot spot prediction and process window control across production wafers have become increasingly critical to prevent hotspots from becoming yield limiting defects. Traditional computational lithography applications are effective in eliminating major layout related issues before a mask is made, but are not directed towards predicting on-wafer performance, largely due to the unavailability of detailed data on actual process variations before tape-out. Process of record solutions to find these patterning defect locations are not sensitive enough or have low throughput and need improvements in order to be used as systematic defect detection tools.

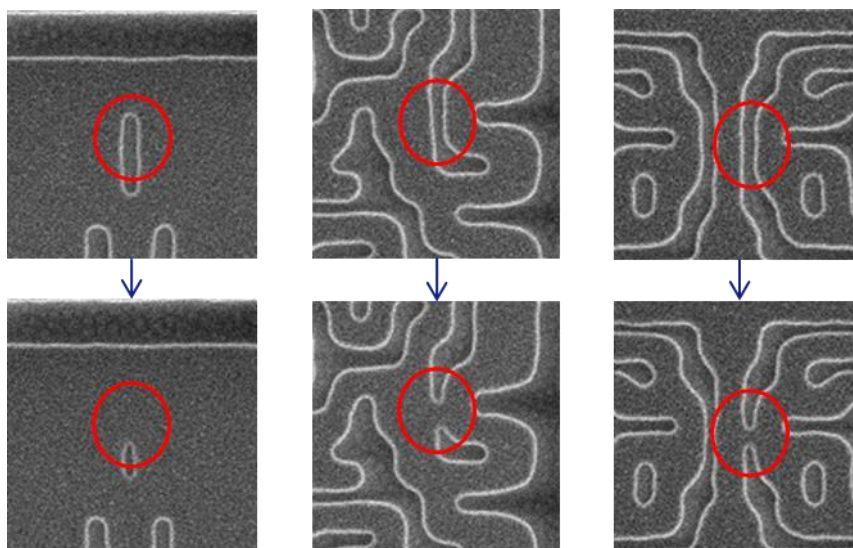


Figure 1: Even on product wafers focus variations can cause patterns to turn into real defects

Process Window Optimizer (PWO) extends the holistic lithography framework towards prediction, detection and verification of on-product patterning defects. The location of these process window limiting “hotspots” are

determined using computational lithography simulations combined with on-product focus measurements and in-line scanner metrology. In this paper we demonstrate prediction and verification of defect occurrences on wafer that arise from focus variations approaching or exceeding process window margins of device hotspots.

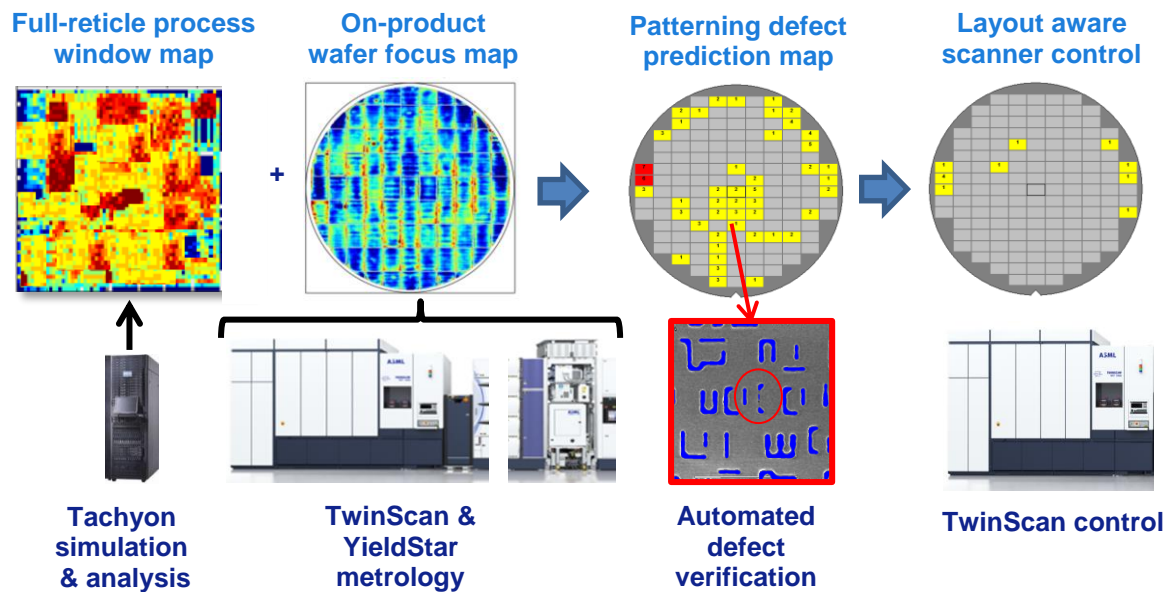


Figure 2: Holistic lithography concept for patterning defect prediction and layout aware scanner control [1]

Proof of concept has been demonstrated on a 28nm metal layer with a full-field product layout at STMicroelectronics. A three dimensional resist model (R3D) [2] was used to perform, resist-profile aware, simulations with the PWO software. The outcome of these simulations was used to predict the location of defects after etch. CD-SEM verification using Hitachi's contour extraction [3] for complex logic patterns is then shown. Focus variations are characterized by determining a systematic fingerprint that is combined with per-wafer topography data into dense focus maps to determine if and where across a production wafer focus-sensitive hotspots may turn into device defects.

Detecting the hot spots locations, with their CD estimations through focus can also help, to better control scanner focus correction in order to make sure that critical hotspots exposure are in their focus range. An estimation of this gain has been performed during this PWO evaluation.

TEST PLAN

A 28nm metal-1 product was selected to investigate the effectiveness of PWO at predicting defect locations. Each field was 26nm x 32.9 mm in size and consists of 7 different product blocks.

The exposures were performed on a NXT:1950i immersion scanner using short-flow wafers, with contact and metal stack only. This means that short-range topography effects from the front end of line will not influence the on-product focus measurements. These effects were studied in a separate paper [4].

Additionally, after etch, hotspot verification measurements were performed with a Hitachi CG 5000 and defect contours were extracted using the Hitachi contour extraction software. Extracted contours are analysed with PWO software to identify necking-type defects and quantify local minimum CDs. There was a significant delay between the exposure and etching of these wafers which meant that resist shrinkage occurred. To compensate for this effect a CD offset was applied to the measurement data.

Two PWO use cases were evaluated during this study. The first was computational defect prediction and control involving simulating the full chip process window and combining this with the effective focus map. The second

was using pre-selected hotspots, combined with the effective focus map to predict the location of defects on the wafer, then verifying these results using CD-SEM measurements.

HYBRID DENSE FOCUS MAP GENERATION

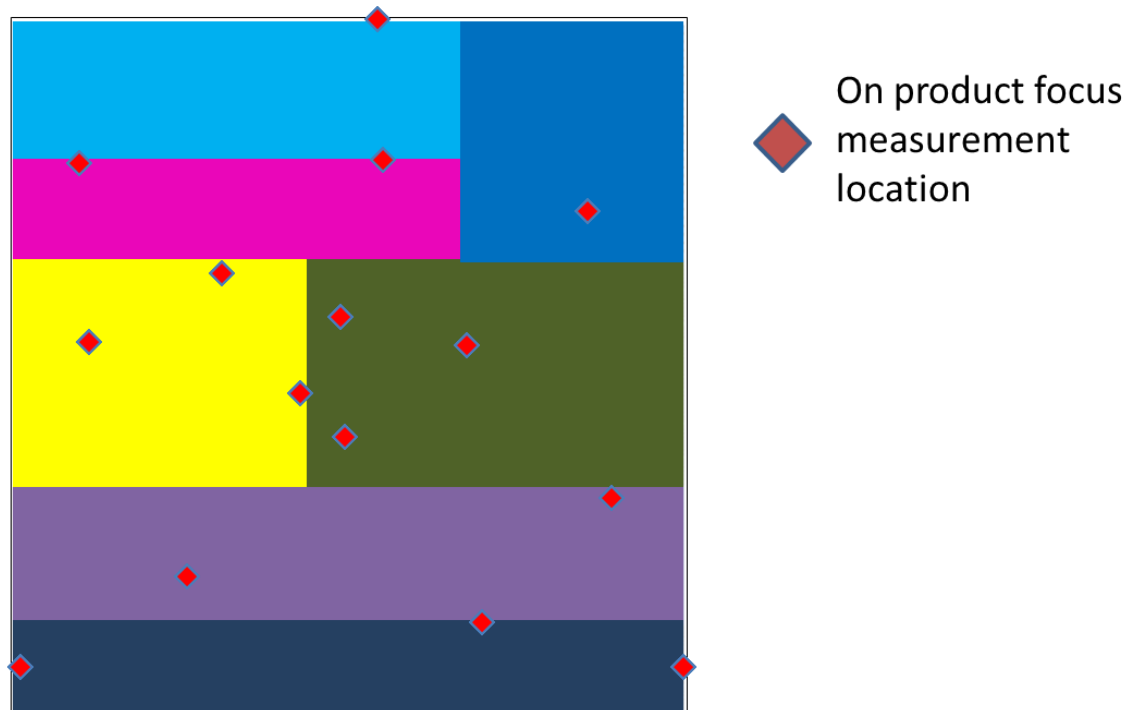


Figure 3: Field layout and on-product focus measurement locations of metal 1 product mask

An on-product focus map was generated using the multi wafer FEM methodology [5]. This step involves measuring focus sensitive isolated structures on the multiple product wafers exposed with a nominal focus delta between each wafer. Then re-constructing the Bossung curves for each measurement location and taking best focus as the Bossung top. For this, 15 locations per field were measured, after development (ADI), for every field. This method of determining best focus is only used for initial characterization, whereas diffraction based focus measurements would be used for inline focus measurements [6].

The hybrid dense focus map (HDFM) is generated using on-product focus measurements combined with scanner measurements on a higher spatial frequency. The scanner measurements are captured in-line and include the levelling non-correctable error (NCE) for every wafer and the lens image plane deviation (IPD). The levelling NCE is the residual which cannot be corrected by scanner levelling during wafer exposure and the lens IPD is a measurement of the focus offset induced by the lens across slit.

COMPUTATIONAL FULL CHIP DEFECT PREDICTION FLOW

The first step in the computational defect prediction flow is to find the patterns across the full chip that will limit the process window the most. Executing full field simulations across all focus conditions would take too much computation time. A method has been established in order to detect several locations that become simulated defects at worst focus conditions and, then, fine tune these locations in order to derive the most limiting hotspots. Once hot spots patterns are identified, pattern search and match across the field are performed in order to find all the pattern occurrences. Finally, CD through FEM for each hotspot was estimated by simulation.

From the hotspot list and data simulated, we generate a hotspot list together with two maps that characterize the product reticle in terms of Best Focus (BF) and Depth Of Focus (DOF). In order to build these BF and DOF maps, the reticle layout is split into 0.5x0.5mm pixels. The process limiting hotspots within each pixel are determined and the corresponding depth of focus and best focus is extracted. The combination of depth of focus and best focus tells us how much of a focus excursion is required to induce a defect at each location.

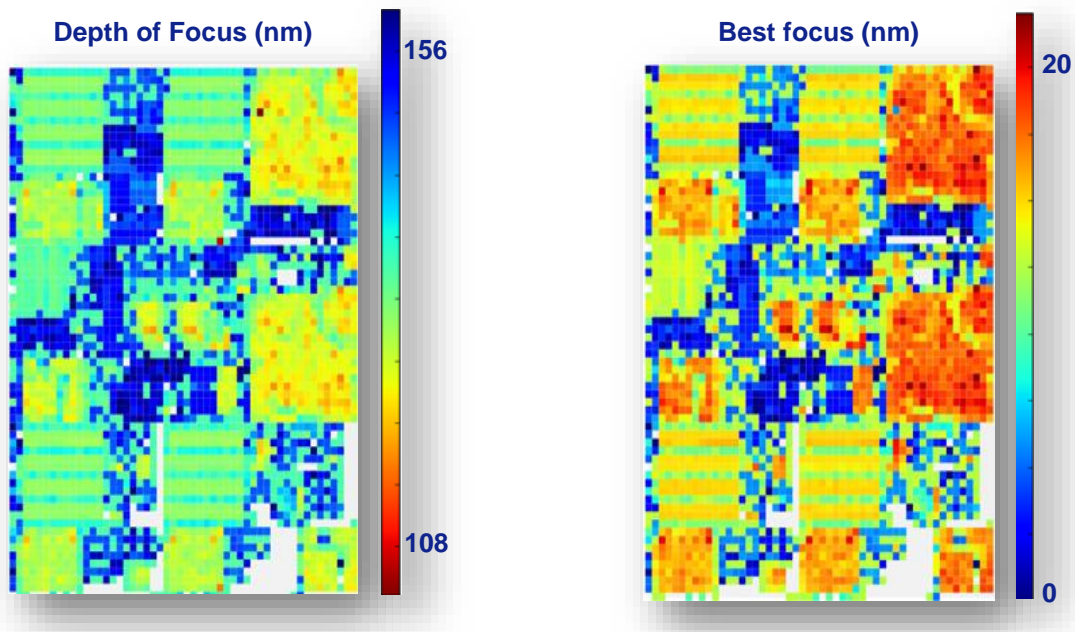


Figure 4: Full chip simulated depth of focus and best focus per location (white indicates a non-critical region)

The depth of focus of this product is most critical at the top right and center right of the field. These locations also have a higher best focus than the rest of the field. Optimizing scanner focus and levelling control to have a positive defocus at these locations may be a way to reduce defectivity.

The next step in the flow is to combine the simulated CD through FEM information with the effective focus values at hotspots locations in the hybrid dense focus map (HDFM). Wherever the defect CD for the corresponding wafer location focus fall below a given threshold (same as LMC threshold used for currently qualified verification in production), the hotspot will be flagged as a defect. A defect map was determined for each of the 9 wafers exposed through focus to visualize the increasing defectivity and induced defect locations at positive and negative focus.

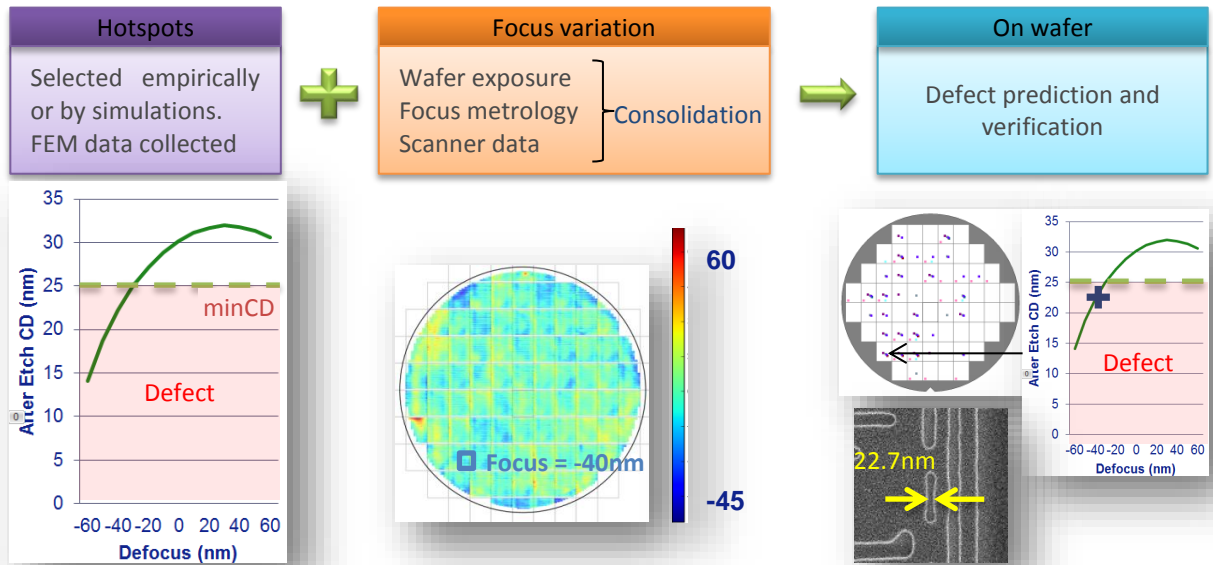


Figure 5: Simplified defect prediction flow

Predicted defect count is seen to increase as we move out of focus. Defects increase at a more rapid rate in negative defocus than positive defocus which can be considered for future process centering.

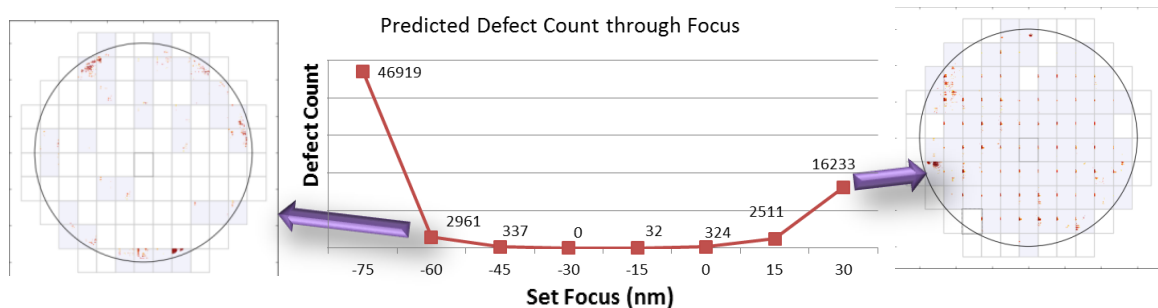


Figure 6: Predicted defect count and locations based on full chip simulation data

Using the overlapping process window information and the HDFM it was possible to generate a design aware scanner control output. This input aims to maximize the overlapping process window by applying focus corrections per exposure (z offset, Rx and Ry). The simulations show that a potential usable DOF improvement up to 17nm might be possible if correction was applied.

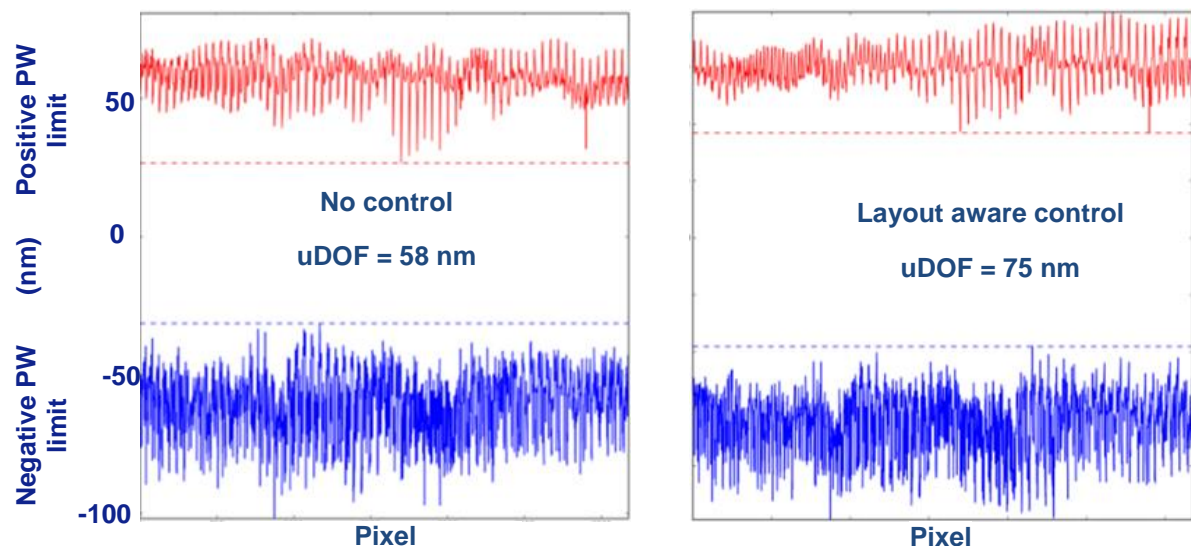


Figure 7: Simulated window comparison with and without PWO generated control recipe

Due to timing constraints it was not possible to verify the simulated defects on wafer in time for SPIE 2016. The complete computational full chip defect prediction flow is underway on a new test vehicle and will be presented in an upcoming conference.. The defect prediction and verification step will be presented in the next chapter using the user provided hotspot flow.

USER PROVIDED HOTSPOT FLOW

The user provided hotspot flow starts with the user selecting some known hotspots. These can be from simulation (standard LMC), measurement (process window qualification with bright field inspection tool) or empirically. For each of the 39 studied hotspots, we measured FEM data using CD-SEM image collection, combined with Hitachi contour extraction and applied ASML CD measurements to collect data through FEM. In order to select the most focus limiting hotspots, we ranked them and sorted out 7 hotspots kept for defect prediction.

Standard LMC is performed at worst process window conditions as opposed to PWO that will determine which hotspots are limiting the process window. LMC is performed at different focus and dose conditions with different resist heights (using Brion R3D resist profile calibrated models) defined as a standard and qualified production verification flow.

The focus offsets from the hybrid dense focus map at each position in the wafer are converted to predicted CD values, per hotspot, using the FEM curve. A hotspot is considered a defect if it falls below the CD defect threshold.

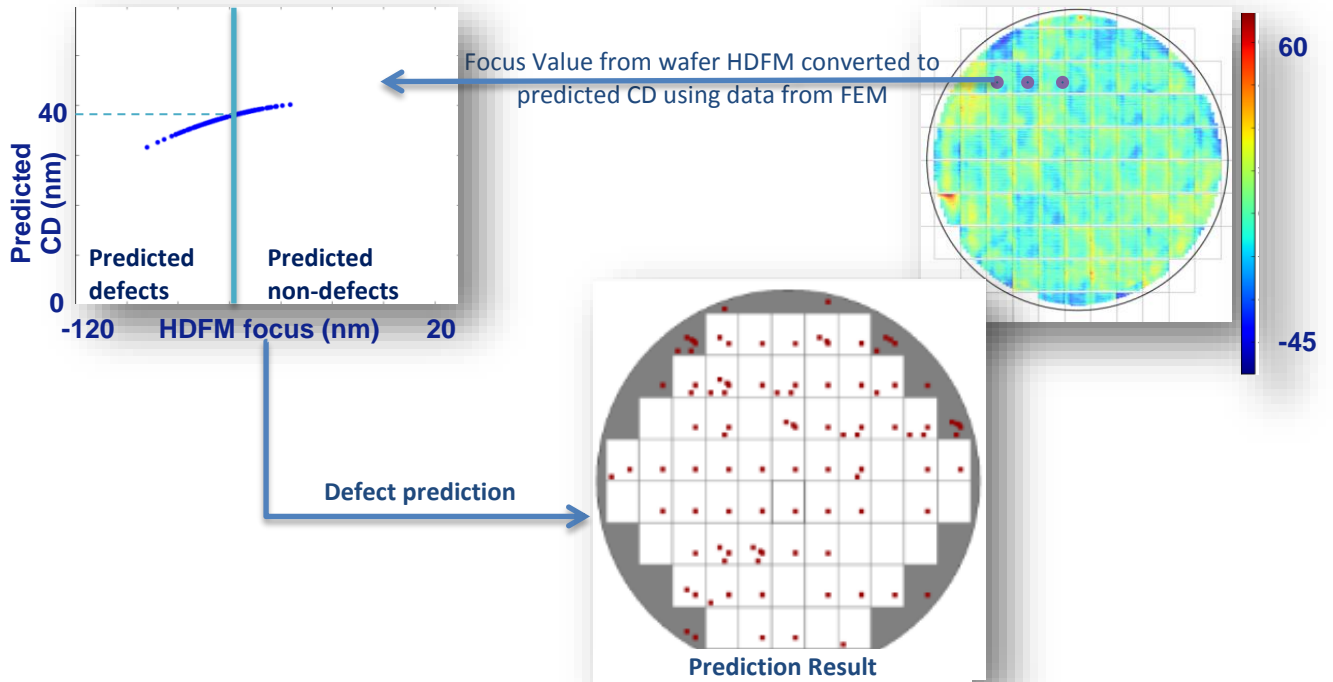


Figure 8: Defining defect and non-defect

In order to generate a significant occurrence of defects for evaluation purposes this study used an off-focus wafer. At -60nm set focus, defects are present sporadically across the wafer and at -90nm there are repeating defects in every field.

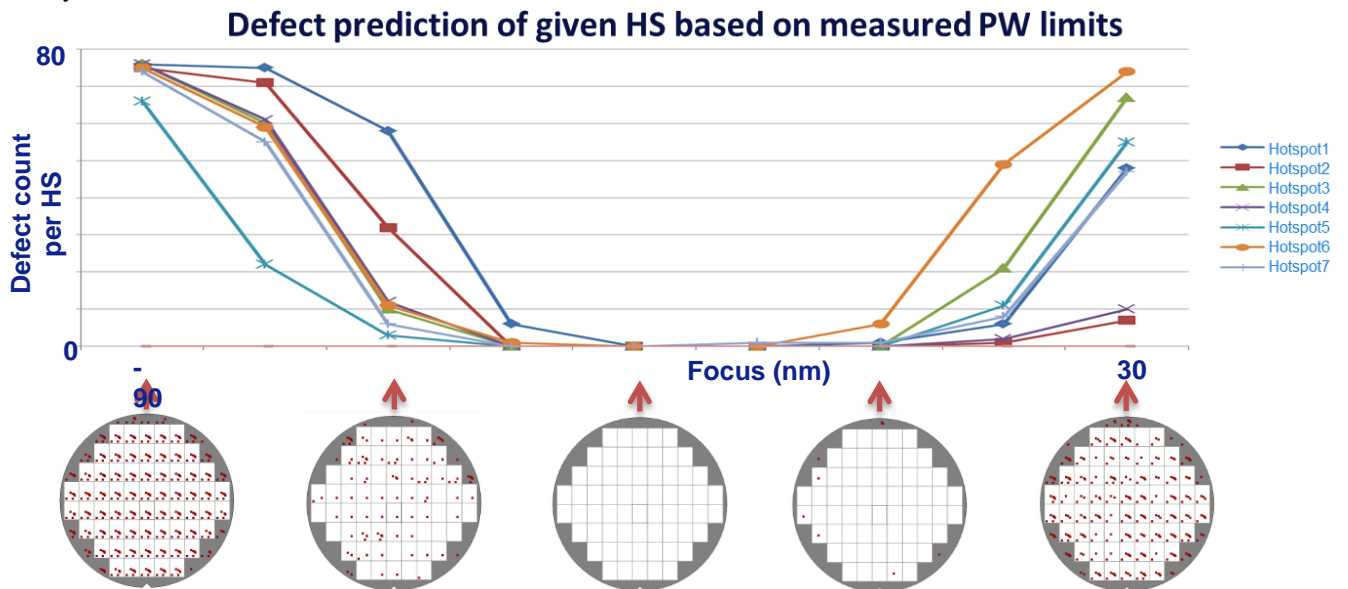


Fig 9: Defect prediction of given hotspots based on measured process window limiters

For this evaluation each hotspot location was measured. A check was done to see if the predicted defect locations the SEM revealed a defect or whether no defect was present. At predicted non-defect locations a check was also done to verify whether a non-defect was present or whether the simulation missed a defect location. The -75nm and -60nm set focus wafers were selected to check prediction accuracy.

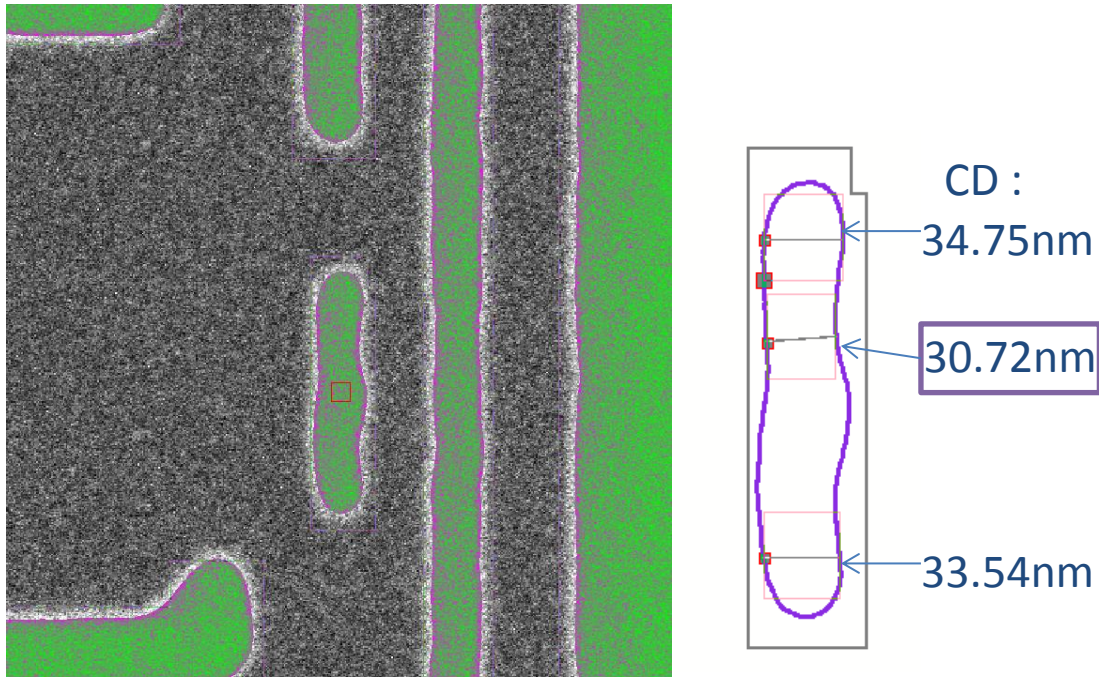


Figure 10: Example of neck location in contour extraction for selected hotspot

The defect measurements were performed, after etch using SEM. A contour was extracted from the measured images. Then LMC detectors were applied to check for defects. By applying a CD threshold to the LMC detector it is then possible to generate a binary yes/no defect state.

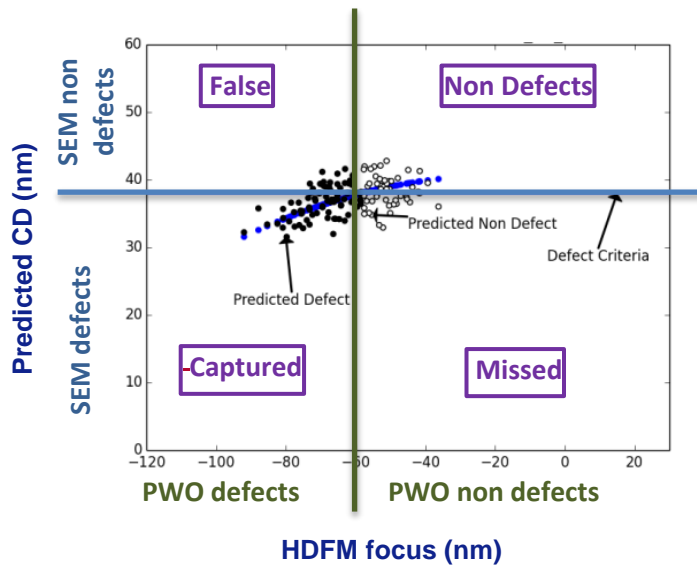


Figure 11: Visualization of PWO capture rate

Once measured CD data on CDU wafers are available, we were able to compare them to the predicted defects. The prediction accuracy is determined as a ratio of the number of accurate predictions with respect to the total number of hotspots. A matching location is also a requirement in order to have an accurate prediction.

$$\text{Capture rate} = \frac{\# \text{ Accurate predictions (captured or non defect)}}{\text{Total number of hotspots}}$$

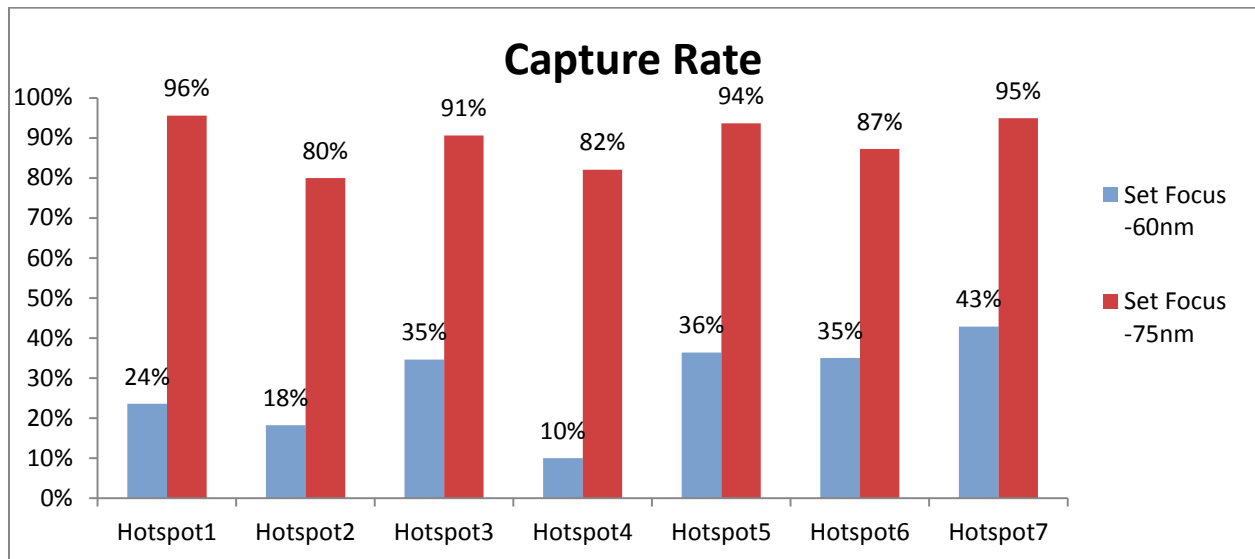


Fig 12: Capture rate of 8 user selected hotspots

The measured defect to predicted defect comparison gives an accurate capture rate. On the -75nm and -60nm wafers the average capture rate is 89% and 27%, respectively. An improvement of this metric would be expected using inline focus measurements to limit delay between lithography and etch. Including high frequent topography, etch fingerprint, mask and dose effects in the defect prediction is also expected to improve capture rate.

CONCLUSIONS

This evaluation has demonstrated the capability of using PWO from full chip layout layout to on-product defect prediction. A computational full chip simulation was combined with on product focus measurements and inline scanner metrology data to predict defect locations. Additionally, user selected hotspots were combined with the hybrid dense focus map and the prediction efficiency of PWO was investigated.

Resist simulation associated with inline data like dose and focus map is a powerful way to predict patterning issues. PWO demonstrates this and opens a new area for patterning control. PWO will help us to improve our inline defect detection. It will accurately predict defect locations. These locations will be used in our defect detection tool to improve the signal to noise ratio. It will also enable a new way to adjust process within wafer and within field for a given layout. PWO can predict the defect count for several process adjustments and thus help us to find the best within wafer/within field process condition dose/focus to minimize patterning defects.

KEYWORDS:

Defect prediction, hotspots, focus, topography, simulation, process window, holistic lithography

REFERENCES:

- [1] S. Hunsche *et al.*, "A new paradigm for in-line detection and control of patterning defects", Proc. of SPIE Vol. 9424, 2015
- [2] A. Szucs *et al.*, "Advanced OPC Mask-3D and Resist-3D modeling", Proc. of SPIE Vol. 9052, 2014
- [3] D. Fuchimoto *et al.*, "Measurement Technology to Quantify 2D Pattern Shape in sub-2x nm Advanced Lithography", Proc. SPIE 8681, 2013
- [4] J-G. Simiz *et al.*, "Product layout induced topography effects on intrafield levelling", Proc. of SPIE Vol. 9661, 2015
- [5] J-G. Simiz *et al.*, "Predictability and impact of product layout induced topology on across-field focus control", Proc. of SPIE Vol. 9424, 2015
- [6] K. D Park *et al.*, "Improvement of inter-field CDU by using on-product focus control", Proc. of SPIE Vol. 9050, 2014

RESUME

La complexification des intégrations sur les puces électroniques (co-intégration, diversification des matériaux utilisés, ...) et la course à la miniaturisation sont les deux moteurs actuels de la recherche en microélectronique. Les limites optiques de la lithographie sont déjà atteintes depuis longtemps et la double exposition est devenue une méthode usuelle pour continuer à diminuer les dimensions des motifs et augmenter la densité du circuit. Avec la multiplication de ces besoins, la fabrication doit aussi être contrôlée de plus en plus étroitement afin d'éviter des variabilités qui nuiraient au bon fonctionnement du produit. Le nombre d'effets croisés entre les différents éléments augmente donc (en nombre comme en proportion) alors que les besoins de contrôle se font plus stricts que jamais.

Cette thèse présente une approche holistique du contrôle d'un des paramètres les plus importants de la photolithographie : le focus. Celui-ci est directement lié à la qualité de l'image transférée dans la résine photosensible pendant l'exposition. Son contrôle est donc primordial. Les sources de variabilités du focus sur le wafer sont multiples et diverses : laser, masque, colonne optique, contrôle des moteurs dans le scanner, planéité de la plaque, intégration, design, réflectivité du substrat, qualité des matériaux, etc. Tous ces effets vont s'ajouter, pouvant provoquer la formation de défauts qui peuvent être catastrophiques (courts-circuits par exemple).

L'objectif de ce travail est tout d'abord de présenter les challenges actuels que nous proposent les nouvelles technologies développées chez STMicroelectronics, particulièrement en termes de photolithographie et de contrôle du focus. Un état des lieux de deux procédés critiques en BEOL 28nm FD-SOI et Contact 14nm FD-SOI est ensuite établi dans lequel tous les effets seront mis en évidence. Les effets du design à l'échelle macroscopique ont été évalués comme ayant une influence de l'ordre de 20% du budget total et presque 50% des effets intra-puce. La topographie représente la plus grande partie de ces effets et des mesures ont montré jusqu'à 32nm de variabilité 3σ au sein d'un même champ, ce qui risque de créer des défocus du même ordre. La profondeur de champ disponible étant de l'ordre de 60 à 70nm pour les niveaux étudiés, il apparaît évident que le contrôle du focus est ici très critique.

L'approche holistique de cet effet en particulier a conduit à l'utilisation d'outils de « data mining » telle la régression par la méthode des moindres carrés partiels (Partial least Square ou PLS) qui a permis de pointer les principales causes de cette topographie, de créer un modèle prédictif de la topologie mais aussi d'évaluer des solutions d'améliorations. On distinguera les solutions « palliatives » et les solutions « curatives ». Dans la première catégorie, on comptera l'amélioration des corrections qu'effectue le scanner permettant un meilleur contrôle généralisé de toutes les technologies sans toutefois changer l'intégration et le design ou encore la mise en place d'une méthode qui permet d'évaluer les erreurs de focus sur le wafer sans pour autant avoir recours à des mesures intensives sur silicium. Les solutions « curatives » s'attèleront à corriger les facteurs de risques à la source en modifiant le design afin de limiter la formation de la topologie de surface.

ABSTRACT

The increasing complexity in chip integration (co-integration, increasing diversity of materials...) and the race to dimension shrinkage are the two main drivers of research in microelectronics today. The optical limitations of lithography have been reached some years ago so that double patterning is now a typical process flow in production and helps reducing pattern size and increasing design density. Because of these, the manufacturing itself needs to be more tightly controlled in order to avoid marginalities. Which will affect the chip operation. The cross-effects between these elements are more numerous and their ratio in the total budget is larger whereas the needs for tighter process control are rising.

This thesis presents a holistic approach of the control of one of the main parameters for photolithography: focus. It is directly linked to the quality of the image transferred into the photoresist during exposure. Its control is then essential. Variability sources for focus are manifold and diverse: laser, mask, optical column, servo-controllers, wafer flatness, integration, design, substrate reflectivity, material quality etc. All these are added to each other, leading to the creation of defects which can be catastrophic such as shorts.

The first objective of this work was to show current challenges raised by STMicroelectronics new technologies, specifically photolithography-wise and focus-wise. A budget breakdown of two critical processes (Metal line patterning in 28nm FD-SOI and Contact patterning for 14nm FD-SOI) has been established which gives the impact of every effect. The product layout effects were evaluated to represent up to 20% of the complete budget and 50% of its intra-chip component. Topography contributes to a large part of these effects and offline measurements showed up to 32nm 3 σ of height variation in a single field. This may lead to local defocuses of the same order of magnitude. The usable depth of field being about 60 to 70nm for the studied layers, it is clear that focus control is really tight here.

The holistic approach of topology led to the use of data mining tooling as PLS regression (Partial least Square). It allowed the highlighting of main causes of topography, the creation of a predictive model of topology and the evaluation of several improvement solutions. One may distinguish "palliative" and "curative" solutions. In the first category, one may put scanner levelling improvements which might be effective for every technology without any modification to make on integration and design. The emulated wafer map methodology providing on-product focus non-uniformities without any measurements is also a solution for investigation. "Curative" solutions may concern the mitigation of risk factors by modifying the design topography built-up main factors.