Stealth Dicing Characterization, Optimization, Integration, and Operations Management for Ultra-Thin Stacked Memory Dies
by
Weng Hong Teh

B.Eng. in Electrical Engineering, University of Malaya, Kuala Lumpur, 1999
S.M. in Advanced Materials, National University of Singapore, Singapore, 2000
Ph.D. in Physics, University of Cambridge, England, 2004

Submitted to the MIT Sloan School of Management and the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degrees of
Master of Business Administration
and
Master of Science in Electrical Engineering and Computer Science
in conjunction with the Leaders for Global Operations Program at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2015
© 2015 Weng Hong Teh. All rights reserved.
The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.
Stealth Dicing Characterization, Optimization, Integration, and Operations Management for Ultra-Thin Stacked Memory Dies

by

Weng Hong Teh

Submitted to the MIT Sloan School of Management and the Department of Electrical Engineering and Computer Science on May 8, 2015, in partial fulfillment of the requirements for the degrees of Master of Business Administration and Master of Science in Electrical Engineering and Computer Science

Abstract

This dissertation presents original work in the development of multi-strata subsurface infrared (1.342 μm) nanosecond pulsed laser die singulation (stealth dicing) to enable defect-free ultra-thin stacked memory dies. The over-arching contribution is the first comprehensive and systematic experimental study of stealth dicing, encompassing process physics and simulation, characterization, optimization, and integration, as well as operations management, including statistical process control, sensitivities, interactions, and risk analysis. This work exploits the multi-strata interactions between generated thermal shockwaves and the preceding dislocation layers formed to initiate controlled crack fractures that separate the individual dies from within the interior of the wafer as a method for significant singulation-related defect reduction and die strength enhancement. A new partial-stealth dicing before grinding (p-SDBG) integration based upon the tandem use of three-strata stealth dicing followed by static loading from backgrinding to complete full kerf separation has successfully demonstrated defect-free eight die stacks of 25 and 46 μm thick 2D NAND memory dies on high backside reflectance wafers for the first time. This work resulted in a 3.5% mean increase in memory/system test yield and has been used to realize production-worthy 64 GB retail memory products after passing reliability tests. Based on unit loadings at SanDisk Shanghai for 2014, this translates to annual cost savings averaging $12.0M when extending this technology to all systems-in-package (SIP) products consisting of 4-, 8-, and 16-die stacks.

Thesis Supervisor: Roy Welsch
Title: Professor of Statistics and Management Science, MIT Sloan School of Management

Thesis Supervisor: Duane Boning
Title: Professor of Electrical Engineering and Computer Science, MIT Department of EECS
The author wishes to acknowledge the Leaders for Global Operations Program for its support of this work.
Acknowledgments

This dissertation would not have been possible without the selfless support of many. I swear that this dissertation (fourth!) will be my last ... academically. Ahem.

First and foremost, I am very grateful to my parents and my friends, old and new, for their unwavering support. You are empirical evidence that quality trumps quantity in relationships and those who matter do not mind and those who mind do not matter.

I would like to thank SanDisk Semiconductor Shanghai (SDSS) for the maiden international internship opportunity to stack memories, figuratively and literally, at the Paris of the East that is Shanghai, China. I have heard so much about China’s economic boom; I had to see it for myself. I am grateful for the cultural and professional experimentation offered.

Manish Bhatia (LGO ’99), SVP of Worldwide Ops for SanDisk, has continued to place his confidence in MIT and LGO to become new partners. Manish, together with the senior leadership of SanDisk, Gursharan Singh and K.L. Bock, have provided immense hospitality, flexibility, and empowerment to my work at SDSS. Thank you.

Duane Boning’s dedication to semiconductor research, particularly statistical metrology and modeling, and MIT’s international education partnerships, is inspiring. Specifically, I thank Duane for helping me navigate the complex web of academia-industry intellectual property convolution. Roy Welsch, a Jedi Knight in all things big data and statistics, has provided immense support to my internship, never failing to drop by, Far East or the West Coast, no matter how tight his schedule. I am grateful to Roy for giving me the flexibility to steer the management part of the dissertation toward more quants in the form of statistical control systems and DOEs, which I simply adore. I also thank him for opening unexpected doors for me, thus increasing my serendipity surface area.

K.L. Bock and C.U. Tan from SDSS have been extremely generous with providing me bureaucracy-busting access to all levels of personnel, from blue collar workers to senior leadership and everyone in between. K.L.’s suggestion to initiate, develop, and deliver case-based leadership classes to his directors has been pivotal to enable meaningful discussions on various leadership models, both East and West. C.U. is extremely inspiring with his leadership style of leading with humility. I thank him for his selflessness and for helping me mobilize
actions across different functions and perfecting the art of Guanxi. I am eternally grateful for his resolute support of my data-driven convictions. I hope my work has made him proud.

I owe my thanks to many SDSS colleagues in making this internship memorable and providing me with such a welcoming voice. They have been an important source of ideas to my work and cultural experience. From process engineering, special thanks to Tony for integration support, Lu Peng for getting me off the ground, Henry for his insightful sharing of Chinese leadership, Wu Jie for fulfilling my endless Taobao requests, Wang Xu for being my teaching assistant in the MBA L-series, and Wang Li for managerial support. From quality assurance, I thank Santosh for FA support, Cindy for helping me deploy SPC, and Yew Wee for leadership support. I also wish to thank Brant, Hope, Eric, Anna, and Elley from R&D for their help in answering many of my questions. Azure, Guan Jin, Yu Zheng, and Dandy deserve special thanks for agreeing to join my SPC taskforce and delivering the data I needed. Yuki and Joyce from HR/Admin have been very helpful with the logistics of the MBA L-Series. I also acknowledge Shrikar and Suresh from SanDisk Milpitas for discussions on a variety of technical topics. And to Kris and Shai, you guys rock!

From DISCO, I gratefully acknowledge Raphael for the immense operations support, always willing to go above and beyond to satisfy my endless experiments. Also, this work would not have been possible without the resources provided by DISCO Japan and the impeccable coordination by Heidi, my liaison to the Japanese partners.

I must thank Roy, Duane, Dave Hoyme, Scott Bromley, Kevin Lin, “J-Dub”, and Alex Whigham for proof-reading my thesis draft to ensure that my scrappy writing deserves to see the light of the day. Tom Sanderson has been utmost helpful in terms of LaTeX programming.

I am particularly indebted to my previous advisors, Gene from MIT, Charles from Cambridge University, and Quat Vu from Intel for agreeing to be my referees for LGO admission. Your kindness in dispensing good advice is greatly appreciated.

Last but not least, my heartfelt thanks for the generosity of the Noyce Foundation toward supporting my studies at MIT’s LGO program by awarding me the Robert N. Noyce scholarship. I am consistently reminded of how the freedom of daring to dream, daring to do is such an admirable hallmark of the American entrepreneurial heritage. And how I am one of the many billions of fortunate beneficiaries of the United States of America.
## Contents

### List of Figures

- 15

### List of Tables

- 29

### 1 Introduction

- 31
  1.1 System-On-Chip (SOC) and advanced packaging
  - 31
    - 31
  1.1.1 Multi-Chip Module (MCM)
  - 33
  1.1.2 System-In-Package (SIP)
  - 33
  1.1.3 System-On-Package (SOP)
  - 33
  1.2 Motivation for thin die: Scaling, performance, and applications
  - 34
    - 34
  1.2.1 Thin dies for More Moore miniaturization
  - 35
  1.2.2 Thin die for More than Moore functional diversification
  - 36
  1.2.3 Ultra-thin stacked memory die assembly: Goals
  - 37
  1.3 Dissertation overview
  - 38

### 2 Literature review: Die singulation technologies

- 41
  2.1 Introduction
  - 41
  2.2 Die singulation: Challenges
  - 42
    - 42
  2.2.1 Cu-low-κ interlayer dielectric interconnect structures
  - 42
  2.2.2 Ultra-thin dies fabrication
  - 43
  2.2.3 Dicing street reduction and complex metal structures
  - 43
2.2.4 Emerging applications: Different die sizes and non-rectangular shapes, 3D wafer-scale integrated devices, and MEMS ........................................ 44

2.3 Singulation quality characterization and responses .............................. 44
2.3.1 Topside (TSC) and Backside Chipping (BSC) ............................... 44
2.3.2 Kerf geometry ........................................................................... 45
2.3.3 Die sidewall damage and integrity .................................................. 46
2.3.4 Die surface contamination ........................................................... 46
2.3.5 Electrostatic discharge .................................................................. 47
2.3.6 Metal/Interlayer dielectric delamination and peeling ....................... 47
2.3.7 Die strength and dispersion .......................................................... 47

2.4 Die singulation technologies and thin-die integration schemes ............... 48
2.4.1 Mechanical blade dicing .............................................................. 49
2.4.2 Ablative laser dicing ..................................................................... 50
2.4.3 Laser/mechanical blade hybrid dicing ............................................. 51
2.4.4 Plasma dicing .............................................................................. 52
2.4.5 Post-dicing die strength enhancement ............................................ 52
2.4.6 Dicing After Grinding (DAG) and Dicing Before Grinding (DBG) ... 53

2.5 Pulsed laser induced subsurface separation: Stealth dicing .................... 54
2.5.1 Stealth dicing as a potential defect-free subsurface separation technology 55
2.5.2 Main principles and mechanisms of stealth dicing laser processing .... 55
2.5.3 Limitations and challenges to stealth dicing ................................... 57
2.5.4 Stealth dicing process integration schemes ..................................... 57
2.5.5 Stealth dicing applications .......................................................... 60

3 Experimental methods and analysis ...................................................... 61

3.1 Preparation of NAND flash memory Si substrates ............................... 61
3.1.1 2-D NAND flash memory monitor wafers .................................... 61
3.1.2 2-D NAND flash memory functional wafers ................................. 62

3.2 Experimental methods: Key processing modules ............................... 62
3.2.1 Stealth Dicing module experimental set-up .................................... 64
3.2.2 Wafer backgrinding module experimental set-up .......... 65
3.2.3 Die separation (DDS) module experimental set-up .......... 66

3.3 Experimental analysis: Measurement methods and concepts .......... 68
3.3.1 Absolute reflectance measurements .......... 68
3.3.2 Atomic Force Microscopy (AFM) .......... 68
3.3.3 X-Ray Powder Diffractometer (XRD) .......... 68
3.3.4 Three-point bend test .......... 69
3.3.5 Shadow Moiré measurements .......... 69

4 Stealth Dicing: Process physics and simulations .......... 71
4.1 Introduction and literature review .......... 71
4.2 Process physics and simulations: Temperature rise of silicon due to absorption of permeable nanosecond pulsed laser .......... 73
4.2.1 Analysis model and methods .......... 73
4.2.2 Effects of laser pulse energies and focal distance on the spatial and temporal heat affected zones within Si .......... 76
4.2.3 Effects of laser pulse energies and focal distance on the spatial and temporal heat affected zones at frontside Cu layer .......... 80
4.2.4 Radial distribution of heat affected zones at frontside Cu layer .......... 84
4.2.5 Effects of laser focal distance from wafer backside on heat affected zones .......... 86
4.3 Conclusions and perspectives .......... 87

5 Stealth Dicing: Process characterization, optimization, and integration .......... 89
5.1 Introduction and literature review .......... 89
5.2 Ultra-thin stacked memory die architecture .......... 90
5.3 Results and discussion: Stealth dicing process characterization .......... 92
5.3.1 Influence of wafer technologies .......... 93
5.3.2 Single-pass and multi-pass stealth dicing layer(s): Key responses .......... 97
5.3.3 Effects of laser beam optical parameters (BP) .......... 97
5.3.4 Effects of laser scanning speed .......... 102
5.3.5 Effects of laser average power .......... 109
5.3.6 Effects of multi-strata stealth dicing ........................................ 112

5.4 Results and discussion: Stealth dicing process optimization ............ 117
5.4.1 Development of production-worthy stealth dicing ...................... 117
5.4.2 Process-of-record (POR) stealth dicing to realize 25 μm ultra-thin die 124

5.5 Results and discussion: Stealth dicing process integration development ........................................ 126
5.5.1 Stealth Dicing Before Grinding (SDBG) integration: Challenges ...... 126
5.5.2 A new integration approach: Partial stealth dicing before grinding (p-SDBG) ........................................ 128
5.5.3 Development of p-SDBG integration: Stealth dicing related integrated defects ........................................ 128
5.5.4 Optimized p-SDBG quality demonstration ................................ 134
5.5.5 Conclusions and perspectives ........................................ 145

6 Stealth Dicing: Operations Management – Process control, sensitivities, interactions, and risks analysis ........................................ 147
6.1 Introduction: Lean Six Sigma operations ...................................... 147
6.1.1 Define, Measure, Analyze, Improve, Control (DMAIC) .................. 148
6.1.2 Identify, Design, Optimize, Validate (IDOV) ............................ 149
6.1.3 Stealth Dicing: Using the IDOV + C framework ......................... 150
6.2 Introduction: Statistical process control and robust optimization ........ 152
6.3 Results and discussion .................................................... 153
6.3.1 Stealth dicing process time-series baselining and statistical process control development ........................................ 153
6.3.2 Optimized solution space: Sensitivity and two-factor interaction analysis161
6.3.3 Stealth dicing quality fishbone diagram .................................. 173
6.3.4 Stealth dicing process Failure Modes Effects Analysis (FMEA) ....... 174
6.4 Conclusions and perspectives ........................................ 177

7 Conclusion and Future Work ........................................ 179
7.1 Review .................................................... 179
7.2 Perspectives ........................................ 182
7.2.1 Design 2.0 and More than Moore: Innovator’s dilemma for semiconductor packaging technologies .......................... 182
7.2.2 Applying software development methodologies to packaging and assembly technologies ..................................... 186
7.3 What next? .............................................................. 187
7.3.1 Short-term projects ............................................... 188
7.3.2 Medium to long-term projects ............................... 189

Bibliography .......................... 191
A Glossary .................. 199
B Publications ............... 203
# List of Figures

1-1  (a) System-On-Chip (SOC) based on a complete system on one chip, (b) Multi-Chip Module (MCM) based on interconnected components, (c) System-In-Package (SIP) based on a stacked chip/package for reduced form factors, and (d) System-On-Package (SOP) which optimizes the functions between ICs and the package while providing overall miniaturization .......................... 32

1-2  SOP concept of system integration of components and ICs .................. 34

2-1  Schematic illustrating top side chipping failing the specified defect free zone (DFZ) requirement ................................................................. 45

2-2  Schematics of problematic kerf (a) geometries (side view), and (b) straightness (top view) ................................................................. 46

2-3  Principal assembly process flow from completed wafer to final package. . . . . 48

2-4  Schematic illustration of the interaction between the mechanical blade and Si wafer ................................................................. 49

2-5  Schematic illustration of the interaction between laser beam and Si wafer . . . 50

2-6  Schematic illustration of laser and mechanical blade hybrid dicing: (a)-(c) show the different options for the laser scribe step. (d) is the mechanical blade dicing step post laser scribing ............................................ 51

2-7  Schematics representing the (a) Dicing After Grinding (DAG) and (b) Dicing Before Grinding (DBG) die preparation process integration flows. ............. 53
2-8 (a) 3-D schematic showing SD laser processing on a patterned Si wafer, (b) 2-D schematic showing the formation of the internal modified SD layer with its inset, (c) an SEM micrograph representation of a typical SD layer microstructure.

2-9 Process schematics representing the stealth dicing after grinding (SDAG) integration flow.

2-10 Process schematics representing the stealth dicing before grinding (SDBG) integration flow.

3-1 Process schematics representing the stealth dicing before grinding (SDBG) integration flow with identified intercept points for key metrology and characterization steps.

3-2 A schematic of the DISCO DFL7360 Stealth dicing tool.

3-3 A schematic of the linked DISCO DGP8761 Grinder/Polisher and DFM2800 Wafer Mounter integrated tool.

3-4 A schematic of the DISCO DDS2300 integrated die separation tool. The numbers on the 3-D schematic (right) shows the typical sequence of processing steps.

4-1 Schematic illustrations of the SD laser processing operations (left) and the formation of the SD-modified internal layer within Si as the SD laser scans along the dicing street (zoom-in).

4-2 A schematic illustrating the analysis model assumed in the simulations.

4-3 Time evolution of SD laser-induced (Pulse Laser Energy (PLE)/Laser Average Power of 15.4 $\mu$J/1.4 W) heat affected zones at a $z$ focal distance from the wafer frontside at (a) 20 $\mu$m, (b) 35 $\mu$m, (c) 50 $\mu$m, and (d) 70 $\mu$m. The zoom-in images show the temperature contours corresponding to 75 ns, 100 ns, and 200 ns pulse timing for $z = 35$ $\mu$m.

4-4 Similar simulation conditions to Fig. 4-3 but with Pulse Laser Energy (PLE)/Laser Average Power of 18.7 $\mu$J/1.7 W.
4-5 Similar simulation conditions to Fig. 4-3 but with Pulse Laser Energy (PLE)/Laser Average Power of 22.2 \( \mu J/2.0 \) W. 

78

4-6 Simulated peak power of the incident SD laser beam as observed at \( \sim 120 \mu m \) from the wafer frontside (before the beam converges into the focal point) and right above the Cu plane (as the beam diverges out of the focal point) at a \( z \) focal height of 35 \( \mu m \). Results of the simulations are similar for all the \( z \) focal heights considered. 

79

4-7 Time evolution of SD laser-induced (Pulse Laser Energy (PLE)/Laser Average Power of 15.4 \( \mu J /1.4 \) W) heat affected zones at the frontside Cu layer with SD laser focal height \( z \) of (a) 20 \( \mu m \), (b) 35 \( \mu m \), (c) 50 \( \mu m \), and (d) 70 \( \mu m \). The arrows indicate the corresponding pulse timing in which maximum temperature at the Cu layer is achieved. 

80

4-8 Time evolution of the SD laser-induced temperature at different radii positions, ranging from 0 \( \mu m \) (center) to 14 \( \mu m \) from the projected focal center of the SD laser beam on the frontside Cu layer plane with SD laser focal height \( z \) of (a) 20 \( \mu m \), (b) 35 \( \mu m \), (c) 50 \( \mu m \), and (d) 70 \( \mu m \). The Pulse Laser Energy (PLE)/Laser Average Power used here is 15.4 \( \mu J /1.4 \) W. 

81

4-9 Similar simulation conditions to Fig. 4-7 but with Pulse Laser Energy (PLE)/Laser Average Power of 18.7 \( \mu J/1.7 \) W. 

82

4-10 Similar simulation conditions to Fig. 4-8 but with Pulse Laser Energy (PLE)/Laser Average Power of 18.7 \( \mu J/1.7 \) W. 

83

4-11 Similar simulation conditions to Fig. 4-7 but with Pulse Laser Energy (PLE)/Laser Average Power of 22.2 \( \mu J/2.0 \) W. 

84

4-12 Similar simulation conditions to Fig. 4-8 but with Pulse Laser Energy (PLE)/Laser Average Power of 22.2 \( \mu J/2.0 \) W. 

85

4-13 Maximum temperature reached on the frontside Cu plane at different distances from the projected focal center of the SD laser at various \( z \) focal heights. The Pulse Laser Energy (PLE)/Laser Average Power used here is 15.4 \( \mu J /1.4 \) W. 

85

4-14 Similar simulation conditions to Fig. 4-13 but with Pulse Laser Energy (PLE)/Laser Average Power of 18.7 \( \mu J/1.7 \) W. 

85
4-15 Similar simulation conditions to Fig. 4-13 but with Pulse Laser Energy (PLE)/Laser Average Power of 22.2 μJ/2.0 W. 85

4-16 Time evolution of SD laser-induced (Pulse Laser Energy (PLE)/Laser Average Power of 22.2 μJ/2.0 W) heat affected zones at a z focal distance from the wafer frontside at (a) 740 μm (or 40 μm from the wafer backside) and (b) 755 μm (or 25 μm from the wafer backside). 87

5-1 A schematic of an eight layer memory die stack with single-sided bonding pads progressively staggered (staircase structure) in the form of two four-die blocks. This architecture is used as a mechanical p-SDBG test vehicle for 25 μm thick memory dies. 92

5-2 A schematic of an eight layer memory die stack with single-sided bonding pads staggered one-way progressively (staircase structure). This architecture is used as an electrically functional p-SDBG test vehicle for 46 μm thick memory dies to realize a 64 GB micro-SD memory card. 92

5-3 Measured spectral properties of the specularly reflecting wafer backside due to different thin film stacks for different wafer technologies: 19 nm, 1Y, and 1Z. 94

5-4 Backside surface root mean square (RMS) surface roughness $R_a$ for different wafer technologies 19 nm, 1Y, and 1Z. Inset: AFM image of a representative 1Y wafer backside surface. 95

5-5 XRD patterns of the 1Y wafer die sidewalls along the length and the width. Inset: XRD patterns of the 1Y wafer backside. Two 1Y dies are stacked on top of each other, measuring 1550 μm in total thickness, to facilitate the XRD measurements. 96

5-6 Cleavage planes $\{110\}$ of (100)-plane Si wafers. Note that the scribe lines are aligned with the $<110>$ directions. 96
5-7 Schematic of a representative multi-pass SD processing showing the formation of three SD layers (strata) with key quantifiable parameters SD layer focal z-height $Z_{SDi}$ and SD layer height $T_{SDi}$ identified. Both single pass and multi-pass SD processing are performed as part of the characterization and optimization efforts reported.

5-8 Focal plane $z$-height of SD layer, $Z_{SD1}$ as a function of SD laser defocus position (DF) for different beam optical parameter (BP) settings. Single-pass SD processing was performed at a scanning speed of 500 mm/s with an average power of 2.0 W (PLE $= 22.2 \mu$J), incident on the 2D NAND 1Y monitor wafer backside. Spatial dimensions were measured using cross-sectional optical microscopy.

5-9 (a) Mean SD layer height $T_{SD1}$ as a function of mean focal plane $z$-height $Z_{SD1}$ for different beam optical parameter (BP) settings is plotted with defects identified. (b)-(d) Optical micrographs (left: top view from wafer FS; right: cross-section view on die sidewalls) demonstrating clean SD processing ($Z_{SD1} = 72 \mu$m), FS ablation ($Z_{SD1} = 48 \mu$m), and FS ablation and interference ($Z_{SD1} = 23 \mu$m), respectively, for BP $= 15$. Single-pass SD processing was performed at 500 mm/s, 2.0 W (PLE $= 22.2 \mu$J), incident on the 1Y monitor wafer backside.

5-10 SEM micrograph of a zoom-in view on the interference-induced defects (stria-tions) between the bottom of the SD layer and the frontside of the wafer.

5-11 Example of identified key SD layer microstructural transitions as observed by magnified optical microscopy on the sidewalls of a full thickness 1Y die. *Ceteris paribus*, the magnitude of the dislocation density and polycrystallization of the SD-modified layer changes from (left to right): high (dense) to medium (optimal) to low (“fishbone”) as laser scan speed increases. Beam optical parameter (BP) is set at 15 for all cases given the $z$-range involved.
Experimental evidence of SD layer microstructural and dimensional transitions observed by optical microscopy on full thickness die sidewalls as a function of laser scanning speed (50 to 900 mm/s) for (a) 1Y, (b) 1Z, and (c) 19 nm monitor wafers. "*" indicates the optimal scan speed, positioned before a microstructural transition (represented by the dashed line where applicable) toward a low-density “fishbone” structure for a set of SD processing conditions and wafer technologies. Beam optical parameter (BP) is set at 15 for all cases given the z-range involved.

Mean SD layer height $T_{SD1}$ as a function of laser feed/scanning speed (50 to 900 mm/s) for different monitor wafer technologies (1Y, 1Z, and 19 nm). Inset: Microstructural and dimensional transitions observed by optical microscopy on full thickness die sidewalls for selected combinations of wafer technology and feed speed. Beam optical parameter (BP) is set at 15 for all cases given the z-range involved.

Schematic illustration of the separation of individual irradiation pulses as scanning speed increases for a given pulse laser energy.

Examples of (a, b) dense and (c) “fishbone” SD layer microstructures within 1Y wafers. Scanning electron microscope (SEM) micrographs of SD layer microstructure produced using scan speeds of (a) 200 mm/s, (b) 50 mm/s, and (c) 900 mm/s. The SD layers are generated using a laser average power of 1.7 W (PLE = 18.8 μJ) at a $Z_{SD1}$ of ~ 94 μm (DF = -182 μm). Inset of (a) shows a magnified SEM image of the circular voids measuring ~ 2 μm, produced at the lasers focal spot.
5-16 Experimental evidence of SD layer microstructural and dimensional transitions observed by optical microscopy on full thickness die sidewalls as a function of laser average power (1.0 W to 2.2 W, i.e., PLE from 11.1 μJ to 24.4 μJ) for (a) 1Y, (b) 1Z, and (c) 19 nm monitor wafers. Here “*” indicates the optimal laser average power, positioned before a microstructural transition toward a low-density “fishbone” structure for a set of SD processing conditions and wafer technologies. Beam optical parameter (BP) is set at 15 for all cases given the z-range involved.

5-17 Mean SD layer height (first SD layer), $T_{SD1}$, as a function of laser average power (1.0 W to 2.2 W, i.e., PLE from 11.1 μJ to 24.4 μJ) for different monitor wafer technologies (1Y, 1Z, and 19 nm). Inset: Microstructural and dimensional transitions observed by optical microscopy on full thickness die sidewalls for selected combinations of wafer technology and laser average power. For the 1Y wafer, two passes of SD processing were made within the wafer in order to facilitate manual separation. Beam optical parameter (BP) is set at 15 for all cases given the z-range involved.

5-18 Plot of mean focal z-height (first SD layer), $Z_{SD1}$, versus stealth laser average power for different for monitor wafer technologies (1Y, 1Z, and 19 nm). Inset: Representative cross-sectional optical microscope images for 19 nm and 1Y wafers at a laser average power of 1.7 W. Beam optical parameter (BP) is set at 15 for all cases given the z-range involved.

5-19 Experimental evidence of a dual-pass 2.0 W, 500 mm/s SD process on 1Y monitor wafers to show the effects of multi-strata SD processing in generating undesired cleavage plane $\{111\}$ defects when inter-strata distances are not optimized. Plots of measured (a) mean focal plane z-heights, $Z$, for SD1 and SD2 (left axis) and (b) Inter-strata distance, $D_{SD1-SD2}$ (right axis) as a function of SD2 defocus position, $DF_{SD2}$. Inset images show representative optical micrographs with and without inter-strata cleavage plane $\{111\}$ defect formation depending on the inter-strata distance for the given SD processing conditions.
5-20 Cross-sectional optical micrographs of a dual-pass SD process on 1Y monitor wafers for inter-strata distances of (a) 32 μm, (b) 40 μm, and (c) 61 μm. The SD layers are generated using a laser average power of 2.0 W (PLE = 22.2 μJ) at a scanning speed of 500 mm/s, focusing at Z_{SD1} ~ 70 μm and varying Z_{SD2} from 32 μm to 61 μm. (d) Cleavage planes \{111\} of (100)-plane Si wafer with scribe lines aligned with the <110> directions; cleavage plane \{111\} intersects surface plane (100) along the <110> direction at an angle of 54.74°.

5-21 (a) Angled and (b) cross-sectional SEM micrographs of a representative dual-pass SD process that generated a cleavage plane \{111\} defect in between the SD layers as a result of an inter-strata distance of ~ 50 μm.

5-22 A schematic illustrating the mechanism of cleavage plane \{111\} defect formation when the inter-strata distance between the SD layers (for multi-pass SD processing) exceeds a certain threshold resulting in the nominal cleavage fracture on the \{110\} plane deflecting into the \{111\} planes.

5-23 Evidence of sub-optimal inter-strata spacing definition. Cross-sectional optical micrograph of a full thickness 1Y die for a representative four-pass multi-strata SD process at 2.0 W, 500 mm/s. Inset illustrates the generation of three cleavage plane \{111\} \textquotedbl{}V\textquotedbl{} cracks between each pair of SD layers as a result of an inter-strata distance of ~ 55 μm.

5-24 Evidence of a severely sub-optimal SD process resulting in major SD defect modes.

5-25 Optical micrographs of SD-unique defects as a result of sub-optimal Z_{SD1} positioning. These include (a) frontside ablation defects, (b) Si interference defects, and (c) microcracks due to increased sensitivity to wafer handling. Insets show respective magnified images of the boxed regions.

5-26 Top view optical micrographs of frontside surface of adjacent dies on a full 300 mm 1Y wafer with SD kerfs initiated along the dicing streets. (a)-(c) show various SD kerf geometric defects (straightness issues, particles, sensitivity to TEG structures, and etc.) while (d) shows an optimized SD process with a generated straight kerf line that is robust even to complex TEG structures.
Comparison of frontside within wafer (WIW) SD kerf coverage against different SD multi-pass and scanning speed combinations before and after backgrinding down to 25 μm in thickness. All SD processes represented here are performed at 2.0 W at $Z_{SD1} \sim 70$ μm, $Z_{SD2} \sim 115$ μm, $Z_{SD3} \sim 158$ μm, $Z_{SD4} \sim 201$ μm, depending on the number of SD layers with $BP = 15$.  

Schematics illustrating frontside SD kerf formation with approximately (a) 20%, (b) 50%, (c) 70% and (d) 100% within wafer (WIW) coverage. SD laser is incident on the backside. 

Comparison of frontside SD kerf straightness and integrity against different SD multi-pass and scanning speed combinations. All SD processes represented here are performed at 2.0 W at $Z_{SD1} \sim 70$ μm, $Z_{SD2} \sim 115$ μm, $Z_{SD3} \sim 158$ μm, $Z_{SD4} \sim 201$ μm, where applicable with $BP = 15$. 

Cross-sectional optical microscope images of the developed three-pass POR base process for stealth dicing of 1Y memory monitor wafers before backgrinding down to 25 μm thick. The inset image shows a magnified optical image of the boxed region demonstrating well controlled definition of the three SD layers, SD1-SD3, with no undesired defects. 

Top view optical micrographs of frontside surface of adjacent dies on a full 300 mm 1Y wafer with well defined, high quality SD kerfs (identified by the arrows) initiated along the dicing streets regardless of the presence of complex TEG structures. There are no signs of kerf geometric defects such as kerf width, kerf loss, kerf perpendicularity, and kerf straightness issues. The POR three-pass SD process is used here. 

Process schematics representing the stealth dicing before grinding (SDBG) integration flow with an additional wafer backside pre-grinding step inserted right before SD processing.
5-33 Top view optical micrographs of frontside surface of adjacent dies at (a) center, (b) mid-radii, and (c) edge of a 25 μm thick 300 mm SD-processed 1Y wafer post backgrinding (BG) using the Lintec BG tape with adhesive layers measuring 125 μm thick. (d) Approximate locations for the optical inspections. Defined SD kerfs before BG are identified by the arrows. Heavy BG stains and Si residue can be observed post BG at the center of the thinned wafer. ............................................ 131

5-34 Schematics illustrating the different wheel rotations during wafer thinning: from edge to center (left) and center to edge (right). ......................... 131

5-35 Top view optical micrographs of frontside surface of adjacent dies of a 25 μm thick 300 mm SD-processed 1Y wafer post backgrinding (BG) using the Lintec backgrinding tape with the adhesive layers measuring 125 μm thick. (a), (c), and (e) show the respective center, mid-radii, and edge located images before wafer cleans, while (b), (d), and (f) show the corresponding images after wafer cleans using the integrated die separation module. ..................... 132

5-36 (a) Die attach pick up issues as a result of a marginal die separation process for the p-SDBG flow causing dies to remain on the edge of the wafer mounting frame. (b) A close-up view from the DAF/dicing tape side of the mounted wafer showing the nominal die-to-DAF edge separation distance required for successful die pick up (mainly located away from the edge of the wafer mounting frame). (c) Evidence of a sub-optimal DAF cut in the die separation module for adjacent dies located near the edges of the wafer mounting frame. ............................................ 133

5-37 Top view optical micrographs of defect-free frontside surface of adjacent dies at (a) center, (b) mid-radii, and (c) edge of a 25 μm thick 300 mm SD-processed 1Y wafer post backgrinding (BG) using the Lintec backgrinding tape with adhesive layers measuring 165 μm thick. (d) Approximate locations for the optical inspections. Defined SD kerfs before backgrinding are identified by the arrows. There is no evidence of heavy BG stains and Si residue post BG tape material change. ............................................ 135
5-38 Representative top view optical micrographs of defect-ridden frontside surface of adjacent dies singulated and thinned based on the DBG integration flow. (a) Evidence of top side chipping (TSC) due to DBG half-cut using mechanical dicing saw. (b)-(f) Various representative images post wafer backgrinding down to 25 μm thick where both TSC and micro-cracks can be seen.

5-39 Top view optical micrographs of defect-free frontside surface of adjacent dies at (a) center, (b) mid-radii, and (c) edge, of a 25 μm thick 300 mm SD-processed 1Y wafer post optimized die separation (DDS). (d) View from the wafer backside with a magnified view of the boxed region showing the die-to-DAF edge separation distance, which is critical for successful die pick-up. Post-DDS, the kerf width enlarges to 49 μm and 62 μm for channel 1 and 2, respectively.

5-40 Representative top view optical micrographs of defect-ridden frontside surfaces of adjacent dies singulated and thinned based on the DBG integration flow post DAF laser grooving. (a) and (b) Evidence of the protruding DAF layer with respect to the die sidewall. (c) and (d) Evidence of DAF residue and DAF strand defects, respectively as a result of this non-flush DAF-die sidewall profile.

5-41 Schematic illustrating typical DBG integration defect modes incoming to die attach process that are eliminated by the p-SDBG integration flow.

5-42 Top view SEM micrographs showing a defect-free eight layer memory die stack with single-sided bonding pads progressively staggered (staircase structure) in the form of two four-die blocks before (left) and after (right) wirebonding. This architecture is used as a mechanical p-SDBG test vehicle for 25 μm thick memory dies and the integration flow stops at this point.

5-43 Additional angled side-view SEM micrographs for the mechanical p-SDBG test vehicle previously seen in Fig. 5-42. The inset shows a magnified SEM image of the boxed region on the die side walls to better illustrate the integrity of the sidewalls and the flush profile across the 25 μm thick die to the 10 μm thick DAF.
Angled top view SEM micrographs showing a defect-free eight layer memory die stack with single-sided bonding pads staggered one-way progressively (staircase structure). This architecture is used as an electrically functional p-SDBG test vehicle for 46 μm thick memory dies to realize a 64 GB Micro-SD memory card. The inset shows a magnified SEM image of the boxed region on the die side walls to better illustrate the integrity of the sidewalls and the flush profile across the 46 μm thick die to the 10 μm thick DAF.

Three-point bend test (a) set-up configuration and (b) loading mechanism, bending moment profile, and related parameters. The samples have dimensions of 6.3 mm width, b, 46 μm thickness, h and 10.9 mm length, l with a 10 μm DAF attached to the Si side (backside).

Comparison of die strength between DBG and p-SDBG produced 46 μm thick memory dies to evaluate both circuit side and Si backside. If the orientation of the die is such that the circuit side is on the top side and the silicon side is on the bottom side of loading, then the die strength we measure is termed as Si-side die strength. The vice-versa of Si-side die strength is the circuit-side die strength, where the circuit side in on the bottom side of the loading. This experimental procedure was performed on ~ 10 sample dies per wafer for up to three different wafers.

Comparison of thinned die warpage measurements at room temperature between p-SDBG (left) and DBG (right) produced 46 μm thick memory dies.

Comparison of thinned die warpage measurements at elevated temperatures (up to reflow temperatures of 260 °C) between p-SDBG and DBG produced 46 μm thick memory dies.

X-Bar and Sigma control charts for respective SD layer height, T-SD_i (where i = 1, 2, or 3), with a sub-group size of nine per run.

X-Bar and Sigma control charts for respective SD layer focal z-height, Z-SD_i (where i = 1, 2, or 3), with a sub-group size of nine per run.
6-3 Attributes control charts for FS SD-kerf post-SD % coverage across wafer and different types of defects in terms of affected die counts. .......................... 160

6-4 SD1 layer focal plane Z-height, Z-SD1: (a) Actual by Predicted Plot, (b) Pareto Plot of Estimates, and (c) Two-factor Interaction Profiles. Laser defocus settings refer to that of SD1-layer, while SD2 and SD3 layer positions follow in lockstep mode, i.e., inter-strata DF positions are separated by a constant 10 μm, which translates to ~ 45 μm in actual “within Si” distance between each SD layers focal spot. ........................................ 166

6-5 Similar plots and two-factor interaction profiles as shown in Fig. 6-4 but using SD2 layer focal plane Z-height, Z-SD2, as the observed response. Laser defocus settings refer to that of SD1-layer. ................................. 167

6-6 Similar plots and two-factor interaction profiles as shown in Fig. 6-4 but using SD3 layer focal plane Z-height, Z-SD3, as the observed response. Laser defocus settings refer to that of SD1-layer. ................................. 168

6-7 SD1 layer height, T-SD1: (a) Actual by Predicted Plot, (b) Pareto Plot of Estimates, and (c) Two-factor Interaction Profiles. Laser defocus settings refer to that of SD1-layer while SD2 and SD3 layers positions follow in lockstep mode i.e. inter-strata DF positions separated by a constant 10 μm, which translates to ~ 45 μm in actual “within Si” distance between each SD layers focal spot. ........................................ 169

6-8 Similar plots and two-factor interaction profiles as shown in Fig. 6-7 but using SD2 layer height, T-SD2, as the observed response. Laser defocus settings refer to that of SD1-layer. ................................. 170

6-9 Similar plots and two-factor interaction profiles as shown in Fig. 6-7 but using SD3 layer height, T-SD3, as the observed response. Laser defocus settings refer to that of SD1-layer. ................................. 171
6-10 Prediction Profiler Plots for Z-SD$_i$ and T-SD$_i$ for $i = 1$, 2, and 3. Laser defocus settings refer to that of SD1-layer while SD2 and SD3 layers positions follow in lockstep mode i.e. inter-strata DF positions separated by a constant 10 m, which translates to $\sim$ 45 $\mu$m in actual "within Si" distance between each SD layers focal spot.

6-11 Stealth dicing quality fishbone diagram for singulating a 300 mm Si memory wafer mounted on a backgrinding adhesive tape. Factors affecting stealth dicing quality pre-die attach are categorized as wafer, wafer mounting, SD machine, SD process, and integration.
# List of Tables


1.2 Properties of silicon in comparison to selected metals .................. 36

4.1 Key irradiation and material conditions used in the simulations. ....... 75

5.1 Recommended optimal SD layer depth $Z_{SD1}$ and DF range for different BP settings. ........................................... 101

5.2 Experimentally observed minimum $Z_{SD1}$ to prevent FS damage and its corresponding maximum SD layer depth and calibrated DF values as a function of different BP settings before frontside damage occurs for single-pass SD processing at 500 mm/s and 2.0 W ($\text{PLE} = 22.2 \, \mu\text{J}$), incident on a 1Y monitor wafer backside. .......................................................... 102

5.3 Experimentally observed physical differences between thin dies fabricated from a p-SDBG process versus a nominal DBG process. .............. 130

6.1 Summary of calculated center line and three-sigma control limits for variable responses SD layer height, $T_{SDi}$ (where $i = 1, 2, \text{or } 3$). ............... 158

6.2 Summary of calculated center line and three-sigma control limits for variable response SD layer focal plane Z-height, $Z_{SDi}$ (where $i = 1, 2, \text{or } 3$). ....... 159

6.3 $2^3 + 3$ (center points) + 1 (replicate) = 12 factor level combinations with actual factor levels (laser scan/feed speed, laser average power, laser defocus position for SD1 layer). ............... 162
6.4 SD Process/p-SDBG integration Failure Modes and Effects Analysis (FMEA) as captured on WW51, 2014. Customer here refers to recorded Memory Test and System Test (MTST) unit yield. . . . . . . . . . . . . 175

6.5 FMEA rating descriptions for severity (SEV), occurrence (OCC), and detection (DET) scales to help assign one severity rank for each effect. . . . . . . . . . . . . . . . . . . . 176
Chapter 1

Introduction

Back in 1965, Gordon E. Moore predicted the future growth of the semiconductor industry via the exponential growth of the number of components that could be economically incorporated per integrated circuit [1]. Moore’s Law, as the exponential conjecture was subsequently known, became a defining notion and motivating factor for the industry in establishing a steady pace of miniaturization, doubling device density every 18 to 24 months. Now, 50 years later, fuelled by the rising number of convergent applications, the industry is now faced with a new development, More than Moore, where added value to devices is provided by incorporating functionalities that do not necessarily scale according to Moore’s Law [2]. The pervasion of More than Moore will influence the development of new integration platforms, both in technology and manufacturing. This includes test and reliability as well as design and modeling tools capable of handling heterogeneous subsystems. To this end, the synergy between integrated circuits (ICs) and packages required to enable this system convergence becomes increasingly important as new System-In-Package (SIP) and System-On-Package (SOP) concepts start to disrupt traditional packaging approaches [3, 4].

1.1 System-On-Chip (SOC) and advanced packaging

A microsystem or nanosystem integrates sensors, actuators, and/or electronic components on a small footprint with critical sizes from micrometers to a few nanometers, collects and interprets data, makes decisions, and enforces actions upon its environment [5]. In the past,
microsystems packaging played the main roles of providing I/O connections to and from devices and also interconnected both active and passive components at the package or board level [3]. More recently, in the interest of integrating various system functions on the same chip, heterogeneous integration of IC blocks toward full systems becomes of great interest, where computing, communication, storage, sensing, actuation, and other consumer functions (such as microprocessors, microelectromechanical systems (MEMS), memory, wireless, and graphics) are integrated with other required components (such as antennas, filters, switches, waveguides, and resonators) on the same chip, if possible. This SOC approach offers the promise for the highest performance and integration. However, in the long run, 2D integration within the same chip presents technological challenges that impose computing and integration limits for wireless communications as well as non-incremental costs to both [4].

Figure 1-1: (a) System-On-Chip (SOC) based on a complete system on one chip, (b) Multi-Chip Module (MCM) based on interconnected components, (c) System-In-Package (SIP) based on a stacked chip/package for reduced form factors, and (d) System-On-Package (SOP) which optimizes the functions between ICs and the package while providing overall miniaturization [3].

SOC challenges include long design time due to integration complexities and processing constraints for global optimization despite having individual-level optimization based on traditional scaling, i.e., *More Moore*. Thus, alternative ways to achieve systems integration are also being pursued by SIP and SOP technologies leveraging building blocks of bare and packaged ICs integrated through 3D stacking. In these alternatives, less per-
formance and time-to-market compromises have to be made when it comes to integrating disparate technologies. Fig. 1-1 shows different system integration approaches to enable ultraminiaturized, high-performance, multifunctional products: (a) System-On-Chip (SOC), (b) Multi-Chip Module (MCM), (c) System-In-Package (SIP), and (d) System-On-Package (SOP) technologies.

1.1.1 Multi-Chip Module (MCM)

Invented in the 1980s at IBM, Fujitsu, NEC, and Hitachi, unlike SOC which is essentially partial or full systems on a single IC with more than two disparate functions, MCM is a package-enabled integration of two or more chips interconnected horizontally [4]. Original MCMs started with high-temperature cofired ceramics (HTCCs) but were subsequently replaced with higher-performance low-temperature cofired ceramics (LTCCs) made of lower dielectric constant ceramics and better metallization. Third generation MCMs improved further with multi-layer organic dielectrics and high performing conductors.

1.1.2 System-In-Package (SIP)

In contrast to the horizontal nature of MCM, SIP is package-enabled IC integration with two or more similar or dissimilar chips stacking in the vertical direction. SIP includes the stacking of bare or packaged structures. This helps in improving latency especially if the stacked chips and/or packages are small and thin. Other major benefits include simpler design and design verification, smaller form factor, minimal time-to-market, and the ease of integrating disparate chip processing technologies. However, because SIP only involves IC integration and hence addresses a limited part of the overall system, the end-product system is limited by what it can achieve with only CMOS processing at or below nanoscale.

1.1.3 System-On-Package (SOP)

In contrast to purely IC integration by SOC on a chip-level and/or SIP on a package-level, SOP aims to achieve true system integration by additionally integrating functions and components such as power sources, boards, thermal structures, and passives. This helps
to significantly reduce the size of the non-IC part of the packaged system. The synergy also overcomes fundamental and integration CMOS-limited shortcomings of SOC and SIP. For example, while transistor density scaling has substantial performance and cost benefits in traditional digital components (in a Moore’s Law or More Moore approach), it is sub-optimal for RF components which may be best fabricated in the package with microns-thick dimensions rather than on silicon (e.g., producing higher Q-factors). SOP is also useful for optical system integration (i.e., chip-to-chip optical transmission for high I/O and high speed communications rather than moving onto SOC to replace copper interconnects). Fig. 1-2 illustrates the SOP concept of system integration of heterogeneous components and ICs into one small form-factor, low-cost, high performance packaged system without compromising local optimization of individual performance and yield.

1.2 Motivation for thin die: Scaling, performance, and applications

More Moore allowed the miniaturization at the IC level (i.e., through SOC integration), with single chip package scaling taking place at a slower rate until chip-scale packages (CSP) and 2D MCMs were introduced in the 1990s [3, 4]. A decade later, the harnessing of the vertical dimension helped to create a variety of 3D solutions such as SIPS and SOPs. However, to fully take advantage of these 3D solutions in support of continued miniaturization vis-à-vis More
Moore and added functionality vis-à-vis More than Moore, thin dies are needed more than ever. Thin dies not only help to reduce overall form factor of the final package, they can also improve package functionality and parametrics while enabling flexible electronics and display development due to improved mechanical properties of silicon. In addition, thin dies improve heat dissipation [6]. All of these advantages alongside system level package integration can be applied to mobile, client, Internet of Things (IoT), and wearable technologies applications as we move from packaging applications in the past with a computing focus to current packaging with wireless mobility, big data, and digital convergence focus [3]. The importance of having thin dies to allow for 3D system integration is underscored by its mention on the International Roadmap for Semiconductors (ITRS) as early as 2001 [7].

### 1.2.1 Thin dies for More Moore miniaturization

While the mention of the need for thin and very thin dies was made in the ITRS as early as 2001 and 2003, respectively, it was until 2005 that a defined thickness target was made (50 μm), with an increased focus on wafer thinning/handling and small/thin die assembly and packaging of stacked dies. The 2007 ITRS edition placed a stronger emphasis on the formation of through-silicon-vias (TSVs) and how they leverage on thinner dies in support of 3D IC solutions [7]. Table 1.1 shows the minimum wafer thickness projections in ITRS 2005, 2007, and 2008 updates with indications of manufacturing challenges [7].

<table>
<thead>
<tr>
<th>Year</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITRS-2005</td>
<td>50</td>
<td>25</td>
<td>20</td>
<td>20</td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>ITRS-2007</td>
<td>50</td>
<td>20</td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>ITRS-2008</td>
<td>20</td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 1.1: Minimum wafer thickness projections on ITRS 2005, 2007, and 2008 update [7].

3D system integration based on vertically connected active dies through TSVs creates
interconnect schemes with shorter lengths of intermediate and global wires compared to conventional planar IC. This is necessary to overcome the increase of effective interconnect length, which can cause signal delay and thus hinder circuit performance. Having thinner dies allows for a larger TSV processing solution space because of the processing limits represented by the TSV depth/width ratio. Minimizing die thickness will increase the planar TSV density by reducing TSV pitch and via parasitics [6]. In other words, the driving force behind having thin dies in the context of More Moore is to further enhance computational and communication performance.

1.2.2 Thin die for More than Moore functional diversification

Current packaging is increasingly driven by portable, wireless, and multi-functional applications. The 2005 edition of ITRS first projected the need for added functionality (More than Moore) instead of traditional scaling (More Moore). As the convergence of integrated applications starts to take place, there is less reliance on the electronic properties of silicon but more on its excellent mechanical properties (such as flexibility, stiffness, ultimate strength, and weight), thermal properties (such as thermal conductivity), and biocompatibility. Applications that harness these thin silicon film properties include wearable technologies, flexible electronics/displays, biochips, robot sensors/actuators, mobile devices, and IoT devices. Table 1.2 shows the properties of silicon as compared to selected metals.

<table>
<thead>
<tr>
<th></th>
<th>Young’s modulus (GPa)</th>
<th>Ultimate strength (MPa)</th>
<th>Thermal conductivity (W/mK)</th>
<th>Density (g/cm³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silicon</td>
<td>185</td>
<td>7000</td>
<td>150</td>
<td>2.33</td>
</tr>
<tr>
<td>Stainless steel</td>
<td>200</td>
<td>860</td>
<td>43</td>
<td>8.19</td>
</tr>
<tr>
<td>Cast iron</td>
<td>210</td>
<td>200</td>
<td>80</td>
<td>7.87</td>
</tr>
<tr>
<td>Aluminum</td>
<td>70</td>
<td>480</td>
<td>235</td>
<td>2.70</td>
</tr>
</tbody>
</table>

Table 1.2: Properties of silicon in comparison to selected metals [7].

With the integration of numerous system functions on a common silicon platform, thin dies help to enable the new More than Moore paradigm that calls for the efficient synergy between the IC-package-system using SOP.
1.2.3 Ultra-thin stacked memory die assembly: Goals

Recent memory packaging developments are trending toward smaller form factor, increased heterogeneity, higher performance, and improved heat dissipation. One method to address these challenges is to assemble and stack ultra-thin, high performance memory die with considerable yield. However, ultra-thin die typically generate higher sensitivity assembly interactions, which result in increased defect modes (e.g., chipping, die sidewall damage, microcracks, debris, delamination) and decreased quality characteristics (e.g., die strength, kerf geometry, reliability).

Advanced mechanical dicing or laser ablation based wafer dicing or combinations thereof (hybrid or sequential) based on dicing after grinding (DAG) or dicing before grinding (DBG) integration schemes have been developed to help address the challenges involving ultrathin wafers. Although these approaches help to improve yield loss from direct mechanical dicing, fundamentally, they still result in a highly uncontrollable surface edge perturbation (frontside and/or backside of wafer depending on the integration scheme) due to mechanical interactions or direct laser ablation with Si. In addition, die sidewall damage and surface contamination typically result from a wet, mechanical dicing approach and the risk rises to cause assembly, test, and reliability yield loss as the wafers get thinner.

The primary goal of the work presented in this dissertation is to enable defect-free, ultra-thin (less than 30 μm) stacked memory die assembly by developing a new permeable nanosecond pulsed laser-based wafer dicing technology, commonly known as stealth dicing (SD) and process integration scheme. This is in order to obtain high costs savings via improved assembly yield, which is critical to demonstrate a route toward production-worthy fabrication of highly defect-sensitive ultra-thin dies.

This goal is achieved through five specific objectives: (1) developing and performing process simulations with targeted characterizations to validate and advance the understanding of stealth dicing physics, (2) systematically characterizing and optimizing the stealth dicing process module, (3) developing novel stealth dicing process integration schemes that include solving complex integrated defect issues involving downstream modules, (4) conducting stealth dicing process baselining and multi-factor interaction studies as well as developing key
metrology responses to establish process control systems, and (5) performing failure modes effects analysis (FMEA) and developing quality fishbone diagrams to analyze stealth dicing manufacturability risks. This work demonstrates key steps towards bridging high volume manufacturing (HVM) readiness gaps for a newly developed partial-Stealth Dicing Before Grinding (p-SDBG) integration flow for high backside reflectance Si wafers. With robust equipment and process hardening, this work helps pave the way to accelerate the adoption of stealth dicing technology.

1.3 Dissertation overview

Chapter 2 presents a literature review on past, present, and emerging die singulation technologies and integration schemes with an emphasis on major challenges to wafer dicing, characterization of dicing quality, and post-dicing die strength enhancement practice. In addition, a sub-chapter provides a literature review of stealth dicing as a potential defect-free subsurface separation technology with discussion of the established and developing physics of laser processing and heat affected zones. A review of present stealth dicing integration schemes and applications is given.

Chapter 3 describes the experimental methods and analysis techniques used in this work. These include the preparation of experimental wafers, both functional and non-functional (monitor), experimental descriptions and set-ups of key processing modules, and a selection of key measurement methods and their basic concepts.

Chapter 4 reports the development and the results of stealth dicing process simulations in order to validate and advance the understanding of its physics within the solution space (less than 30 microns final die thickness) and integration flow of interest (SDBG). The analysis model and methods used to understand the temperature rise of silicon due to absorption of permeable nanosecond pulsed laser are discussed. The focal distance and laser energy effects of the 1.342 μm nanosecond pulsed laser on the temporal and 2-D spatial heat affected zones are presented.

Chapter 5 is the first of two chapters that present key experimental steps towards developing production-worthy stealth dicing technology. This chapter reports on characterizing
and optimizing the stealth dicing process module with an understanding of its dicing qual-
ity (defects and parametric performance) and solution space. Combined with solutions to
complex integrated defects involving downstream and upstream modules, this work led to
the development of a novel partial Stealth Dicing Before Grinding (p-SDBG) process inte-
gration scheme to envelope high backside reflectance Si wafers and improve cycle time. The
demonstration of a zero-defect eight die (8D) stack of 25 \( \mu m \) thick non-functional memory
dies and 46 \( \mu m \) thick functional memory dies is shown for the first time with clear evidence
of a significant increase in assembly unit yield.

**Chapter 6** reports the development and the results of stealth dicing technology opera-
tions management lean six sigma operations frameworks, statistical process control, sensi-
tivities, interactions, and risks analysis. Time-series process baselining to establish within
wafer (WIW) and wafer-to-wafer (WtW) statistical process control (SPC) limits, process
sensitivity and two-factor interactions analysis centering around the developed SD process
of-record (POR), parameters failure modes effects analysis (FMEA), and quality fishbone di-
agram definition are described quantitatively and qualitatively to help analyze stealth dicing
manufacturability risks.

**Chapter 7** concludes the dissertation by providing an overall review of the work and
suggests future directions for its continuation. Perspectives on the emerging innovator’s
dilemma for semiconductor packaging technologies and the application of software devel-
opment methodologies to avoid packaging and assembly technologies disruption are also
offered.
Chapter 2

Literature review: Die singulation technologies

2.1 Introduction

A key process step before die stacking is the singulation of wafers, either before or after backside grinding depending on the application. Die singulation, also known as wafer dicing, is part of the die preparation flow within the assembly process that separates the individual dies on a finished wafer for further packaging. Today, die singulation has been used for the packaging and assembly of integrated semiconductor devices (e.g., memory and microprocessors), discrete semiconductor devices (e.g., small-signal transistors and diodes, resistors, and light emitting diodes), solar cells, and microelectromechanical systems (MEMS) [6].

It is important to control yield loss factors, particularly when it comes to ultra-thin die fabrication. Die singulation is one of the most critical elements of the IC assembly process, where the reduction of defects and improvements in quality can make a significant contribution to final yield and per unit costs. Contact-based singulation technologies such as mechanical blade dicing and non-contact, global material removal techniques such as ablative laser dicing are facing growing challenges due to the adoption of copper/low-\(\kappa\) dielectric interconnect structures, ultra-thin wafers, die attach films (DAF), narrow dicing streets, different die sizes on the same wafer, polygonal-shaped work pieces, and complex stacked structures along the dicing streets. Several important dicing technologies emerged in
the early 2000s to address the tough defect challenges associated with incoming or outgoing thin wafers, MEMS wafers, or their derivatives/combinations. These include hybrid laser scribing/blade sawing [8], dicing before grinding (DBG) [9], stealth dicing, a laser induced subsurface separation technology [10–15], and plasma etching for wafer dicing [16]. At the same time, post dicing die strength enhancement is complementing most dicing technologies to achieve dies with high fracture strength [6].

2.2 Die singulation: Challenges

Since the last decade, the traditional mechanical dicing method has been facing many challenges from developments in advanced wafer fabrication and packaging technologies as well as increasingly stringent yield-driven quality control. For example, semiconductor wafer singulation is facing challenges with the adoption of copper/low-κ dielectric interconnect structures, ultra-thin wafers, narrow dicing streets with complex metal stacked structures, and emerging applications involving different die sizes and shapes, 3D wafer-scale integrated devices, and MEMS.

2.2.1 Cu-low-κ interlayer dielectric interconnect structures

Advanced ICs are trending towards higher interconnect density and increased performance in support of the exponential growth in device density and speed [6]. Since 1997, Cu has displaced Al as the interconnect material of choice for high-performance microprocessors (lower RC delay as a result of resistivity reduction from ~ 2.7 μΩ-cm for Al to 1.6 μΩ-cm for Cu, improved electromigration performance, reduced power loss, and increased heat dissipation) and at the 90 nm node, lower-κ (κ ≤ 3) dielectric materials were implemented to replace SiO₂ (κ ~ 4) to form the interlayer dielectric [17–19]. However, these low-κ materials have weak thermomechanical properties (elastic modulus below 10 GPa versus 70 GPa for SiO₂, and a coefficient of thermal expansion (CTE) of ~ 10-17 ppm/K versus 0.6 ppm/K for SiO₂. These properties, coupled with their lower fracture toughness and poor adhesion to adjacent films, cause major defect concerns in the form of delamination and cracking during die singulation. At the same time, Cu is more ductile than Al and as such, will enhance the
premature clogging of the dicing blade, causing excessive chipping [6].

2.2.2 Ultra-thin dies fabrication

Thin dies are needed to enable SIP and SOP solutions for continued miniaturization and improved package functionality. While dies for enterprise and client microprocessors are typically 300 \( \mu \text{m} \) or thicker, devices for mobile products (e.g., memory, logic, microprocessors) have been moving rapidly from 150-200 \( \mu \text{m} \) thickness range to 50-75 \( \mu \text{m} \) or smaller [6]. State-of-the-art memory dies have thicknesses as low as 15-25 \( \mu \text{m} \). Reducing die thickness has several advantages: improved integration, reduced form-factor, ease of TSV processing, increased mechanical flexibility, and improved heat dissipation. However, the handling of thin wafers and thin dies becomes challenging for assembly processes such as die singulation, backgrinding, die attach, and wirebonding. The larger variation in the handling load during die singulation can result in die cracking, die chipping, die peel-off, and die warping for these thin wafers. Additionally, these thin wafers can be more sensitive to internally-generated residual stresses due to the singulation process. DAFs (thermoplastic die attach adhesives in dry film form), when diced through together with the thin wafer, can also cause defect modes such as DAF debris (whiskers), die side wall cracking, and die chipping.

2.2.3 Dicing street reduction and complex metal structures

Shrinking die size and dicing street widths translates into more dies for a given wafer diameter and can lead to effective cost reduction. However, reducing the dicing street width demands tighter control on kerf width, kerf loss, and defect size in order to meet the specified defect free zone (DFZ), the distance from the edge of the street toward the kerf, within which chipping, cracking, or other defects are not tolerated [6]. Die singulation becomes more complex when the aspect ratio of wafer thickness to dicing street widths gets larger.

Test element groups (TEGs) or process control monitoring (PCM) structures, non-uniformly distributed along the dicing streets on the front side of the wafers, are metrology patterns that include representative devices and structures to evaluate processes, circuit characteristics and reliability without the need to directly test the actual device circuits.
They are destroyed when the wafer is diced. The interaction between these complex metal stacked structures that build TEGs and the die singulation process can cause increased immediate or delayed, integrated defects.

2.2.4 Emerging applications: Different die sizes and non-rectangular shapes, 3D wafer-scale integrated devices, and MEMS

Die singulation also faces other challenges in emerging applications: (1) Free-shape dicing for dies with non-rectangular shapes, such as polygonal and circular dies used in LED applications [20], (2) dicing of multi-stacked ultrathin wafers which are vertically bonded together to form 3D integrated devices thus introducing a super lattice of different materials to dice through [21], and (3) dicing of various MEMS devices which can be very sensitive to mechanical pressure, contamination, and electrostatic discharges introduced by the singulation process [22, 23].

2.3 Singulation quality characterization and responses

Die singulation has a significant impact to die packaging and assembly yield. It is one of the most critical processes to meet cost and overall yield targets because the process resides at a stage where the wafers are at their highest value. At the same time, package reliability tests have repeatedly shown that singulation related defects have been the main causes of die-level interconnection failures and/or structural integrity failures [6].

2.3.1 Topside (TSC) and Backside Chipping (BSC)

Chipping is defined as mechanical damage from singulation in which small fragments of the die are removed along the die edges. They are usually caused by cyclic stress induced by backgrinding, blade saw, and/or laser dicing whereby microcracks are initiated and subsequently coalesce after propagation along dislocation planes. Chipping can be classified as topside (TSC) and backside chipping (BSC) depending on the surface of the die where the edge chipping occurs. If TSC propagates into active circuitry area of the die, device failure
will occur. TSC outside the seal ring of the die may still lead to reliability failure of the die despite no immediate electrical failures [6,24]. Fig. 2-1 shows a schematic of TSC failing the specifications of an example DFZ requirement. Typically, in mechanical blade dicing, the maximum allowable TSC is half the dicing street width minus half the blade thickness [6].

While the specifications for BSC can be more forgiving than TSC, oversized BSC can increase the sensitivity of the devices to thermal cycling and cause imperfect molding. This can ultimately lead to reliability issues. BSC is harder to detect after die attach as compared to TSC because it will be covered by epoxy compound [6].

### 2.3.2 Kerf geometry

Kerf geometry covers all geometrical aspects of kerf width, kerf loss, kerf perpendicularity, kerf straightness, and surface finish [25]. For production worthy packaging and assembly, it is important to minimize the variations of kerf geometry. A consistent and reduced kerf width allows the reduction of dicing street width, facilitating the addition of dies on a wafer. Kerf perpendicularity is a depth-wise measure of the tilt of a cut and the kerf width consistency; this is important for products whose functionality depends on the die side wall. Kerf straightness measures the eccentricity of the kerf from the dicing direction; if not properly controlled, over-size die issues can cause problems with package fitting and wire bonding [6]. The same is true for uncontrolled kerf loss. Finally, surface finish, which depends on characteristics of the material being diced, along with blade and process parameters, can
affect applications related to optical filters [25]. Fig. 2-2 illustrates examples of (a) wrong kerf geometries (side view), and (b) kerf straightness issues (top view).

### 2.3.3 Die sidewall damage and integrity

Die sidewall damage can be both visible and non-visible. Visible damage include defects such as scratches, microcracks, microdeformations and microstructural modifications, as a result of mechanical stresses or thermal damage [6] such as those associated with mechanical blade dicing and laser ablation dicing, respectively. In the case of stealth dicing, due to the localized nature of the sub-surface heat affected zone (HAZ) damage, one can choose to remove the induced damage using backgrinding post dicing.

### 2.3.4 Die surface contamination

There are different types of die surface contamination such as dust residue, burrs, and debris formed on the front side, back side, and side walls of a singulated die. When terminal pads on the die get contaminated, the reliability of wire bonds on these pads can be affected. In the case of dicing wafers laminated to DAFs and dicing tapes, unwanted polymeric residues from the DAF and dicing tape material [26] left on the die are called whiskers. Another source of contaminants that can result in device failures are carbon-rich residues that form as a result of laser induced pyrolysis and carbonization of polymers [27] when wafers laminated to DAF is singulated via laser ablation. Additionally, certain MEMS devices can fail as a
result of stiction of moving parts when they are singulated using dicing mechanisms that are water-based, such as mechanical blade dicing or water-jet guided laser dicing.

2.3.5 Electrostatic discharge

Electrostatic discharge (ESD) can be harmful to devices by causing localized scorching, which then can lead to malfunction of sensitive electronic devices. For MEMS devices involving free-standing parts, ESD can result in not only physical damage but also unwanted actuation and/or a change of the device actuation characteristics. In mechanical blade dicing, ESD can be developed in high pressure cleaning stations due to high resistivity deionized (DI) water sprayed onto the wafer during high speed sawing [28].

2.3.6 Metal/Interlayer dielectric delamination and peeling

Metal/Interlayer dielectric (ILD) delamination and peeling are defect modes unique for low-$\kappa$ wafer singulation. This is because of the weak thermomechanical properties of low-$\kappa$ ILD and its poor adhesion with adjacent materials. One method to mitigate the propagation of these defects into the active device area is to use guard rings which are built surrounding each die [29]. These Cu filled guard rings will also help to prevent moisture penetration into the devices.

2.3.7 Die strength and dispersion

Die strength can be degraded by surface and edge defects resulting from die singulation such as chipping, side wall damage, and delamination [6]. Too low a die strength will make the die more susceptible to die fracture failure during assembly processes such as die pickup, die attach, and/or wirebonding of die stacks, particularly those involving the overhang portions of a stacked die. Reliability tests can also exacerbate die fracture failure. Brittle fracture, which is applicable to Si, is of a statistical nature and therefore, both the mean value and its dispersion are needed to better represent the die strength. Certain non-visible damage can also compromise the integrity of the die as a result of excessive residual stresses generated by the die singulation process.
2.4 Die singulation technologies and thin-die integration schemes

Fig. 2-3 shows the principal assembly process flow from completed wafer to final package. In general, the three main segments of assembly flow (nominally the dicing after backgrinding scheme) between wafer sort (in order to identify known good dies) and final card assembly/shipping are known as assembly front-of-line (FOL), end-of-line (EOL), and test. FOL is further divided into two integration flows, namely die preparation, and die attach (DA) wirebonding (WB) mold. Die preparation is the integration flow that prepares the dies from the incoming wafers for subsequent assembly with the substrates. Surface mount technology (not shown in fig. 2-3) is a parallel integration flow that prepares the substrates with passive components. The FOL DA-WB-mold integration flow integrates the dies with the substrates and ends with a hermetic encapsulation to protect the device. Finally, the EOL integration flow transforms the hermetically sealed devices into final, individual packages before testing, card assembly (optional), packing, and shipping to customers. The die singulation process falls under the die preparation integration flow.

![Figure 2-3: Principal assembly process flow from completed wafer to final package.](image-url)
2.4.1 Mechanical blade dicing

Before the use of mechanical blade dicing in the 1970s, the scribe and break process was used for mainstream semiconductor device production. The scribe (typically using diamond scribing) and break approach is a two-step process, where a scribe line is created first with a certain depth and sharpness in the wafer surface along a dicing street, to establish a stress concentration factor. The wafer is then cleaved along the scribed line under applied static or impact loading via non-contact (exposed air bridges) or contact mechanisms [6]. Due to an increasing need for a more consistent die geometrical accuracy, mechanical blade dicing, a grinding process in which material is sheared away from a wafer by diamond grit abrasive embedded on the edge of a rotary annular blade, has evolved into the predominant wafer dicing technology. This mechanical method has been used for both full cutting (e.g., dicing after grinding to thin the wafer) and half cutting (e.g., dicing before grinding (DBG), or laser/mechanical blade hybrid dicing) of wafers. Because Si is a highly brittle material, this method typically causes chipping in the front and back surfaces of the wafer (TSC and BSC). While the magnitude and frequency of the chipping can be reduced via the optimization of wafer circuit design (e.g., interconnect structure and TEG patterns), blade (e.g., abrasive grit material, grit size, and blade thickness), saw machine (spindle power, machine vibration, etc.), wafer mounting (DAF material, mounting tape/adhesive material, etc.), and process parameters (dicing mode, blade speed, etc.) including integration schemes (DBG and others), it can hardly be eliminated due to the mechanical contact. Fig. 2-4 shows a schematic illustration of the interaction between the mechanical blade and Si wafer resulting in TSC.
Ablative laser dicing is a die singulation process using laser ablation, in which material is removed via evaporation and melt ejection under irradiation of laser pulses [6]. The ablation rate depends on the optical and thermal properties of the material, and laser beam characteristics such as laser wavelength, pulse width, optical spot size, pulse energy, pulse frequency, and number of pulses overlapped (nominally controlled by laser scanning speed).

There are three commercially available ablative nanosecond laser dicing technologies: “zero-overlap” dicing (galvanometer-based positioning system to spatially distribute laser pulses as contiguous or spaced apart in a single pass), multiple laser beam (splitting of output laser into a plurality of laser beams to have better processing quality), and water-jet guided laser (coupling a pulse laser beam into a low pressure water jet to perform precision machining) [6, 30]. The thinner the wafer, the more the advantage shifts to laser technologies. Fig. 2-5 shows a schematic illustration of the interaction between laser beam and Si wafer [30].

![Figure 2-5: Schematic illustration of the interaction between laser beam and Si wafer](image)

In the laser ablative process, the laser works in the optimum absorption range of Si (i.e., Neodymium-doped Yttrium Aluminium Garnet (Nd:YAG) laser at a wavelength of 355 nm). As the UV laser hits the wafer (nominally coated with polyvinylalcohol for protection purposes), photons are absorbed, leading to increased energy density in the vicinity of the laser focal spot. This will then generate the ignition of plasma causing ablation and melting [see Fig. 2-5]. Particles and dust released from the wafer surface by the laser impact
are removed by an integrated vacuum exhaust system [30]. As shown in fig. 2-5, the heat impact zone has a characteristic length that defines the distance over which laser induced microstructural modifications occur which may impact device functionality. This length is usually $\sim 3-5 \mu m$ [30]. Typical defects generated by this die singulation process include: redeposited excess material on the die front side and sidewall that may impact device functionality, inconsistencies of kerfs due to the alteration of laser absorptions as a result of stacked passivation layers along the dicing street, and reduction of die strength due to laser induced stresses and hidden defect initiation within the heat impact zone [6,30]. Literature suggests that laser dicing leads to a significant decrease in die strength [6,30–34] in comparison to conventional blade dicing, most likely attributed to increased local tensile stress due to laser-induced melting and recrystallization effects.

2.4.3 Laser/mechanical blade hybrid dicing

The laser/mechanical blade hybrid dicing approach is usually adopted for thick wafers with fragile passivation layers and metal stacks. This is because a purely mechanical dicing process will easily create cracking, delamination, peeling, and smearing defects due to the fragility of these upper layers, while a purely laser ablation dicing approach cannot provide competitive

![Figure 2-6: Schematic illustration of laser and mechanical blade hybrid dicing: (a)-(c) show the different options for the laser scribe step. (d) is the mechanical blade dicing step post laser scribing [6].](image)
The hybrid approach can eliminate such defects and improve yield, at the cost of additional process complexity. Fig. 2-6 shows the schematic illustration of laser/mechanical blade hybrid dicing.

Figs. 2-6(a)-(c) show the different options for the laser scribe step, while fig. 2-6(d) is the mechanical blade dicing step post laser scribing [6]. The hybrid dicing approach involves two steps: the first step is laser scribing to remove fragile materials, cutting into bulk Si with much less defects than blade dicing does; the second step is to cut through the remaining wafer thickness using a mechanical blade.

2.4.4 Plasma dicing

Plasma dicing may afford an exceptionally high die strength post dicing because it is damage and stress free. Additionally, plasma dicing is a parallel process so its productivity is insensitive to the die size. It also enables free-shape die formation and separation. There are several integration flows to enable plasma dicing: (1) dicing of a wafer by applying plasma etch to the backside of the wafer [35], (2) applying laser scribing to first remove TEGs and thereafter, using plasma etching on the frontside of the wafer to dice through the wafer and DAF [36], and (3) using the dicing before grinding approach where a precut is generated with plasma etching [37,38]. The key challenge to plasma dicing is making it a cost-effective solution.

2.4.5 Post-dicing die strength enhancement

The need for die strength enhancement comes from the two-step process of backside grinding, i.e., coarse grinding at a removal rate of $\sim 5 \, \mu\text{m/s}$ followed by fine grinding at a slower removal rate of $\sim 1 \, \mu\text{m/s}$. It was demonstrated [39] that die strength increased markedly when its backside surface roughness, saw mark, and saw mark depth decreases. Thus, a “mirror finishing” or “stress relief” process is introduced after backside grinding to recover the die strength by removing microcracks and dislocations introduced by the mechanical grinding. These mirror finishing approaches include wet etching, dry polishing, plasma etching and chemical mechanical polishing (CMP) [31,39-42]. Post-dicing die strength enhancement is
particularly critical for thin dies because thicker dies with the same die strength are more likely to withstand the mechanical forces during the die pick up process, die attach, wire-bonding, and package reliability tests.

2.4.6 Dicing After Grinding (DAG) and Dicing Before Grinding (DBG)

Figs. 2-7(a) and (b) show schematics representing the DAG and DBG die preparation process integration flows, respectively. Typically, for a 300 mm Si wafer with a wafer thickness greater than 100 μm, the conventional DAG approach is adopted [see Fig.2-7(a)]. Here, the final fabricated Si wafer is firstly thinned to its target thickness via the backgrinding process, followed by DAF/dicing tape (DT) lamination on the wafer backside, and subsequently backgrinding tape peel-off. Finally, die singulation takes place to form the separated, individual dies. Coarse and fine backgrinding, dry polishing, DAF/dicing tape mounting, and backgrinding tape peeling are typically performed within the same module “link” to increase throughput and minimize defects generated via excessive handling movements.

To avoid incoming wafers to die singulation from being too thin, and thus easily warped with increased sensitivity to breakage due to handling and a host of other cracking and
chipping (frontside and backside) defects related to blade dicing, the dicing before grinding (DBG) approach [see Fig. 2-7(b)] has been explored [6,37-39,43-47]. This is despite the fact that post grinding “mirror finishing” processes helped to provide stress-relief in the DAG approach. While the advantage of using dicing tape is its ease of transportation, dicing across a Si wafer/dicing tape interface at its backside induces severe backside chipping, mainly due to blade clogging by the dicing tape adhesive, and the fact that the elastic modulus of the dicing tape is not able to sustain the dicing stress [39]. The DBG process can avoid backside chipping (but not frontside chipping), which has a large influence on die strength by avoiding the mechanical blade contact to the backside Si/dicing tape. In the principal DBG approach, a wafer is partially diced from its frontside to form trenches that are a little deeper than the desired final die thickness. This is accomplished by diamond blade sawing, laser scribing, wet etching, plasma etching or any combination thereof. The wafer is then laminated with backgrinding tape on its frontside, followed by backside grinding to physically separate the individual dies when the frontier of the backgrinded interface reaches the “half-cut” diced trenches. Laser DAF grooving is then performed to fully separate out the individual die along with the DAF layer, which is essential for the downstream die attach process. Finally, the individually separated dies with their respective DAF layers are spatially separated further using tape expansion. The die separation (DDS) module, which uses a cooling expansion process, ensures that the distance between individual die is large enough to allow for die detection and pick up, and also to increase DAF separation quality.

2.5 Pulsed laser induced subsurface separation: Stealth dicing

When a pulsed laser radiation is tightly focused inside a transparent material, high optical intensity can be achieved in the vicinity of focus without damaging material surface leading to significant increase in laser absorption. This causes internal material modification, thermal stress, and microcracks, which can then be exploited for laser subsurface machining of materials.
2.5.1 Stealth dicing as a potential defect-free subsurface separation technology

Stealth dicing (SD) was developed as a permeable pulsed laser die singulation technology by Hamamatsu Photonics K.K. [10–15]. In its optimal condition and depending on wafer type and process integration, high speed processing with no chipping and cracks on both the frontside and backside of the wafer are possible due to its non-contact processing approach. There is no damage at the surface layer and its edges because a laser wavelength permeable to Si is used. Additionally, there is virtually no cutting or kerf loss (removal of parts of the Si dicing street material) post die singulation and as a result, there is no debris contamination induced by SD directly. The completely dry process of SD can also eliminate some of the defects caused by wet processing-based die singulation technologies. Depending on the applications and the sensitivities of the final device, the subsurface internal modified-layer formed by SD processing can be removed by post-dicing backgrinding and/or different “mirror polishing” or stress relief techniques. SD can provide excellent cut quality with flexibility in dicing workpieces with different die sizes and polygonal shapes, supporting a wide range of applications. This is critically important to enable ultra-thin dies.

2.5.2 Main principles and mechanisms of stealth dicing laser processing

The principal SD method involves permeable laser-induced “perforation” in the form of an SD modified layer within the bulk Si, combined with the application of fracture mechanics in order to physically “cleave out” the individual dies. The SD laser beam consists of nanosecond short pulses, permeable to Si (near infrared), oscillating at a high repetition rate and can be highly condensed up to the diffraction threshold level. This highly localized beam is formed at an extremely high peak power density, both time and spatially compressed in the vicinity of the laser focal point, without damaging the material surface. When the SD laser transmitting through the Si wafer exceeds a peak power density (typically more than 100 MW/cm² during the condensing process, a nonlinear absorption effect causes a phenomenon in which extremely high absorption occurs at the focal point [10,12]. A localized temperature
field of larger than 1000 K in a confined volume of $10 \mu m^3$ within the vicinity of the focal spot is established within nanoseconds. As a result, at the focal point vicinity, a void ~ 1-3 $\mu m$ in size is formed due to the melting and vaporization of Si. There are also reports of partial recrystallization of Si and microcracking occurring near the focal spot during the cooling phase [6,10]. Thereafter, a controlled high dislocation density is generated due to the thermal shock wave produced upwards from the laser focal point vicinity. The heat affected zones (HAZs) grew only towards the beam incident surface because the absorption coefficient increases non-linearly with the increasing temperature [48-51]. As the SD laser scans in the horizontal direction, the interaction between the subsequent thermal shockwave generated and the previously formed high dislocation density layer will initiate a crack fracture that separates out the individual dies. SD laser dicing avoids issues associated with the use of conventional laser dicing methods operating at wavelengths that are highly absorbed by the materials to be diced (i.e., ablation) that unavoidably produce heat and debris at the surfaces. Optimization of SD processing is greatly dependent on wafer specifications (e.g., thickness, impurity elements and their concentration, backside film thickness and materials) and the final device applications (e.g., final thickness, defect sensitivity of device, and die size/shape), balancing system-level throughput via different SD process integration flows.

Figs. 2-8(a) and (b) show the 3-D and 2-D schematics of SD laser processing on a patterned Si wafer, respectively [11], while fig. 2-8(c) shows an SEM micrograph representation.
of a typical SD layer microstructure [10]. Post-SD singulation, the development of the perpendicular cracks facing the front and backside surfaces of the die from the top and bottom ends of the SD layer is essential to the die separation mechanism. The understanding and control of the formation of these SD-initiated cracks subsequently led to the development of a new integration approach produced in this work called partial stealth dicing before grinding (p-SDBG), in order to improve SD cycle time for high backside reflectance wafers and to minimize wafer macro-cracks.

2.5.3 Limitations and challenges to stealth dicing

There are several observed limitations to SD technology. First, highly doped substrates can cause too much absorption in the vertical direction (along the laser incidence) before the SD laser radiation reaches the focal plane, leading to insufficient internal layer modification at the focal position and SD layer. This will lead to lower separation ability. It was reported that for easy separation of dies using SD, the resistivity of bulk Si needs to be greater than 0.01 Ωcm [52]. Second, the presence of opaque materials in the dicing street such as metals blocks the SD laser radiation [53,54]. Suggestions are made to have the SD laser incident from the wafer backside, but sometimes this is not viable due to the presence of backside metallization. Efforts have thus been reported to design the metal stack patterns in a way that favors SD [55,56]. Other difficulties reported include the presence of multiple layers of different materials that can affect the propagation direction of the crack fracture initiated by SD [12], and the need for an exclusion zone having a width of 40% or more of the wafer final thickness to account for the high numerical aperture used by SD technology [12,57].

2.5.4 Stealth dicing process integration schemes

From a stealth dicing process point of view, the SD laser beam can be irradiated from the device surface (frontside irradiation) or from the back surface (backside irradiation) on full thickness or backgrinded Si wafers. In general, the limitations posed by frontside irradiation are more constraining. For example, it is difficult to position the SD laser away from the randomly distributed light-blocking TEGs and/or metallic thin films located along
the dicing street. At the same time, the frontside of the wafer tends to have complex stacks of passivation layers that can affect SD processing characteristics (although the backside of the wafer can also have these layers if diffusion-based deposition processes were used in the wafer fabrication). Also, the frontside of the wafer usually has surface topography, which can cause SD laser guidance issues due to incident angle and as a result, frontside irradiation will require non-light shielded areas to guide the SD laser beam. When it comes to silicon-on-insulator (SOI) wafers, the buried oxide (BOX) layer can significantly attenuate the SD laser beam depending on the intended location of the laser focal point and the side from which it is incident. There is also reported work [58] on using SD processing of thin TSV wafers from the micro-bump side or the C4 bump side.

From the literature, in order to realize singulated thin dies, SD processing can be performed using two main types of SD integration flows: Stealth dicing after grinding (SDAG) [10–13] or stealth dicing before grinding (SDBG) [59]. Figs. 2-9 and 2-10 show process schematics representing stealth dicing after grinding (SDAG) and stealth dicing before grinding (SDBG) integration flow, respectively. The SDAG process as shown in fig. 2-9 is very
similar to the DAG process pictured in fig. 2-7(a), except that die singulation is performed using stealth laser before DAF/DT lamination and backside grinding tape peel-off in the SDAG case. There is also an additional die separation step involving cooling expansion after tape peel-off in SDAG. Because the DAF layers are adjoining the SD-separated dies as a consequence of DAF/DT lamination post-SD die singulation, the cool expansion method is used to achieve stable DAF cleaving. Using characteristics of specifically engineered DAF such that it becomes brittle at low temperatures, tape expansion is performed at a low temperature to realize high precision DAF separation. Depending on the magnitude of Si cleaving initiated by SD in the first place, complete die separation in addition to DAF cleaving under radial tensile stresses can be achieved via tape expansion. Similar to the DBG method [see Fig. 2-7(b)], the DDS module also helps to ensure that the distance between individual die/DAF is large enough to allow successful die pick up for the subsequent die attachment process. Sagging that occurs around the periphery of the dicing tape after expansion can be eliminated by heat shrinking to enable the framed diced wafer workpiece to be transported to the following die attach process without having to replace the tape. It is worth noting that DDS here helps to cleave the DAF layers using a separate step, instead of using mechanical blade sawing that cuts through Si, DAF and parts of the DT in one step (which has a disadvantage of generating backside chipping, and DAF peels and folds defects). Because the adhesion strengths between the DAF and wafer, and between the DAF and the dicing tape, are extremely high, and also because the die clearance equivalent to the kerf width is near zero before tape expansion, a high quality DAF separation can be made in the same shape as the chip during the tape expansion. This is especially useful for ultra-small chips.

The SDBG process shown in Fig. 2-10, on the other hand, is similar to the DBG process portrayed in fig. 2-7(b), except that there is no mechanical blade “half-cut” process and DAF laser grooving steps in SDBG. Here, die singulation is performed using stealth laser before wafer backgrinding as opposed to after in the case of SDAG integration. SDBG is usually more applicable for thinner final device thicknesses, i.e., 50 μm or less, with the SD-modified layer removed during the backgrinding step. This ensures that no dicing marks or defects are left behind on the sidewalls of the die. SDBG also has the benefit of ease of wafer handling during the SD processing step. Having thicker wafers incoming to the SD process step helps
to prevent wafer warping issues associated with ultra-thin wafers during SD processing that can increase the variability of the process and make process control more difficult. However, SDAG provides a larger solution space for SD processing, because having thinned wafers coming to SD processing makes perpendicular cracking during SD laser scanning easier. However, it is important to make sure that the final device characteristics are not affected by the residual internally-modified SD-layer [11,15], especially in reliability tests. There is also usually an edge trimming step [60] before frontside taping for the SDAG process, to avoid wafer edge chipping when backgrinding to realize ultra-thin wafers.

2.5.5 Stealth dicing applications

Stealth dicing can be applied to a variety of applications other than enabling ultra-thin dies. In this dissertation, the key focus is using stealth dicing to enable ultra-thin (less than 30 μm thick) stacked memory die assembly. Other applications include: low-κ logic devices, MEMS devices, bonded TSV-based 3D ICs, sapphires for high brightness LED devices, glass, and III-V devices [10–15,58,59,61]. In applying to low-κ devices, the permeable SD laser offers a non-contact singulation technology that helps to eliminate debris and delamination of low-κ materials. Delamination typically results from the global heat generated from laser ablation using pulsed UV laser or from mechanical cutting using blade dicing. MEMS devices benefit from the dry processing and the near zero impact of the SD singulation process too.

By applying multi-passes with or without controlled load-impact breaking, SD processing is also able to singulate hard-to-cut materials such as AlN, GaN, alumina ceramics, and SiC. SD offers process flexibility by being able to configure Hasen Cut recipes [59], where the laser can be turned on and off at any point to process work pieces with different die sizes and polygonal-shaped work pieces, supporting a wide range of applications such as multi-project-wafer (MPW) processing and 3D MEMS devices. One important advantage of SD is the increased chip yield obtained from one wafer. Because of the zero kerf loss, the dicing street width can be drastically reduced from the present ~ 60-80 μm wide to as low as 5 μm. This can substantially increase the number of chips per wafer and lower the costs per unit die. Finally, SD has the potential to increase power savings by eliminating power-intensive pure water production used in blade dicing units [59].
Chapter 3

Experimental methods and analysis

3.1 Preparation of NAND flash memory Si substrates

There are basically two types of Si substrates used for the work reported in this dissertation: patterned 2-D NAND flash memory monitor wafers and functional wafers. Both types of wafers come in different technologies, namely 19 nm, 1Y, and 1Z, depending on the advanced 2-D NAND memory process node used, which ranges from 24 nm down to 15 nm patterning technology. The memory wafers provided are supplied by SanDisk and Toshiba Corporations joint venture facility in Japan, and assembled, packaged, and tested in SanDisk Semiconductor Shanghai in China. These wafers are used for different stages of stealth dicing process characterization, optimization, and integration to enable ultra-thin stacked memory die assembly.

3.1.1 2-D NAND flash memory monitor wafers

Boron doped, Czochralski grown, non-epitaxial grade, 12” diameter silicon (100) surface orientation wafers were used to build the monitor wafers. These P-type silicon wafers are 775±25 μm in thickness and have a resistivity in the range of 20-30 Ωcm. These wafers undergo 2-D NAND flash memory full-process fabrication proprietary steps but have low or zero known-good-die (KGD) yield at die sort and test and are therefore used mainly for mechanical test vehicle purposes. The monitor NAND wafers have a die size of 11.982 x 7.850
mm and can come in three different technology flavors, i.e., 19 nm, 1Y, and 1Z. The bulk of
the work presented in this dissertation focuses on using 1Y wafer technology because it has
the highest backside reflectance to the stealth dicing (SD) laser’s operating wavelength and
thus is able to be used as a representative test vehicle to envelope the remaining two wafer
technologies. These fully processed 2-D NAND monitor wafers have a patterned polyimide
frontside exposing the appropriate bond pads on the die for subsequent wire bonding with
a topography measuring $\sim$ 7 - 8 $\mu$m. The street width for the patterned wafers is $\sim$ 70 $\mu$m
wide. Depending on the wafer technologies, the wafer backside consists of multi-stack layers
of thin films such as SiO$_2$, polysilicon, and Si$_3$N$_4$, ranging from 350 to 1100 nm in total stack
thickness as a consequence of the wafer fabrication process. Prior to SD laser processing,
the wafers are laminated on the wafer frontside with a Lintec backgrinding tape with the
adhesive layers measuring either 125 $\mu$m or 165 $\mu$m in nominal thickness. The wafer notch,
cut during crystal grinding, has a standard $<110>$ orientation.

### 3.1.2 2-D NAND flash memory functional wafers

As for the 2-D NAND flash memory functional wafers, they are similar to the monitor
wafers except that die sort test yielded baseline levels of KGDs, thus facilitate subsequent
memory and card assembly electrical tests, and reliability tests to observe device circuit
characteristics. However, it is important to note that the NAND functional wafers used have
a smaller die size of 6.276 x 10.922 mm than on the monitor wafers, in order to facilitate
potential product intercept as a result of this work. Similarly, they can come in three different
technology flavors, i.e., 19 nm, 1Y, and 1Z. Prior to SD laser processing, the functional wafers
are laminated on the wafer frontside with a Lintec backgrinding tape with the adhesive layers
measuring 165 $\mu$m in nominal thickness.

### 3.2 Experimental methods: Key processing modules

Fig. 3-1 shows the stealth dicing before grinding (SDBG) process integration flowchart from
frontside taping of the wafer to the final framed diced, ultra-thin 300 mm diameter wafer
staged inside a specialty carrier, with identified intercept points for key metrology and char-

62
acterization steps. This SDBG process, where the SD laser is incident on the wafer backside, forms the baseline flow from which this work is derived, including the development of the partial-stealth dicing before grinding (p-SDBG) process for cycle-time optimization. The inputs to the wafer frontside taping step (Lintec backgrinding tape either 125 μm or 165 μm thick) are the fabricated and sort tested monitor wafers or functional wafers, while the output of the SDBG process shown in Fig. 3-1 will be for the die attach process, where the individual dies with DAF will be picked up and subsequently attached on the substrates before oven curing [see Fig. 2-3]. The three key processing modules involved in the SDBG integration flow are the SD module, the integrated wafer backgrinding module, and the die separation (DDS) module. Key characterization methods used include visual inspection, backside reflectance measurements, atomic force microscopy (AFM), optical microscopy, scanning electron microscopy (SEM), X-Ray diffraction (XRD), three-point bend tests, and Shadow Moiré measurements.

Figure 3-1: Process schematics representing the stealth dicing before grinding (SDBG) integration flow with identified intercept points for key metrology and characterization steps.
3.2.1 Stealth Dicing module experimental set-up

The SD processing experiments were conducted in an integrated 300 mm DISCO DFL7360 stealth dicing tool with the “SD Engine” developed by Hamamatsu Photonics K.K. Fig. 3-2 shows a schematic of the 300 mm DISCO DFL7360 stealth dicing tool from the top view as well as a 3D view (inset). This particular tool configuration is capable of supporting the process development of wafer-level SD processing with a route towards high volume manufacturing (HVM). The integrated tool consists of two load ports capable of supporting front opening unified pods (FOUPs) and custom metal carriers configurations, a notch pre-alignment table, a waiting table to allow for cascade processing, two wafer handling arms, one robot pick-up arm, a chuck table that moves the relative positions of the SD laser and wafer in order to scan the wafer at high speeds per the desired dicing pattern, the SD laser processing head that consists of both the measurement laser and the SD laser, and an IR camera microscope (detection range is ~ 800-1100 nm) with a halogen lamp light for wafer frontside pattern alignment.

The measurement laser operates at a near infrared wavelength of 830±20 nm and is primarily used to detect the backside surface of the wafer in order to account for undesired wafer warpage effects on the z-height focal point positioning of SD scanning. This is performed \textit{in-situ} to reduce cycle time by avoiding the need to conduct a wafer pre-scan for each dicing line. A combination of precise z-direction spatial offset between the measure-
ment laser’s focal spot and the SD laser’s focal spot, coupled with a calibrated displacement versus measured photodiode voltage curve (generated due to the reflection of the measurement laser beam from the wafer backside), allows not only wafer warpage compensation to the SD scanning line but also facilitates the “deep trace” capability of this integrated tool, i.e., the ability to form SD-layers deep within the wafer, beyond 500 μm, by ensuring that the operating regime for the warpage compensation stays linear. The “deep trace” function is particularly useful for thick wafers incoming to SD processing.

As for the SD laser light source, a 90 kHz, 1.342 μm near-infrared wavelength pulsed laser varying from 1.0 to 2.2 W in average power is used. Its proprietary operating pulse widths and energies are in the order of tens to hundreds of ns/pulse and μJ/pulse. Multi-strata SD layers within the wafer are defined by translating the chuck table relative to the position of the SD laser with scanning speeds ranging from 50 to 900 mm/s and z-height focal points positioned from ~ 25 μm to 200 μm, as measured from the wafer frontside surface. Wafers used for the SD processing experiments were not backgrinded beforehand, and thus come in their original full thicknesses with dicing tape laminated on their frontside. As a result, the work reported in this dissertation is subjected to the challenges associated with high backside reflectance from the wafers, varying from ~ 13% to as high as 83% depending on the wafer technologies (19 nm, 1Y, or 1Z) used.

3.2.2 Wafer backgrinding module experimental set-up

Wafer backgrinding, DAF/DT lamination, and backgrinding tape peel-off were performed in a “linked” wafer grinder/polisher and wafer mounter tool. Fig. 3-3 shows a schematic of the 300 mm linked DISCO DGP8761 grinder/polisher and DFM2800 wafer mounter integrated tool. This linking approach helps to optimize the handling layout to shorten the overall cycle time and minimize defects due to excessive wafer handling.

On the right hand side of Fig. 3-3, where the wafer grinder/polisher is housed, the backgrinder integrates backside grinding and stress relief processing and performs stable thin grinding to thicknesses less than 25 μm. There are three main grinding wheels, namely Z1, Z1, and Z3, and they are used for rough grinding, fine grinding, and dry polishing, respectively. The Z3 spindle enables dry polishing without chemicals or water to realize high
wafer strength while maintaining the grinding gettering effects. The baseline backgrinding process used in the experiments leverage the use of a three-step grinding process. For a final Si monitor wafer thickness of \( \sim 25 \ \mu m \), the Z1 spindle is used for coarse grinding to remove \( \sim 700 \ \mu m \) of bulk Si while the Z2 spindle is used for fine grinding to remove \( \sim 50 \ \mu m \) of bulk Si. Typically, spindles Z1 and Z2 operate at \( \sim 3200-3400 \ \text{rpm} \) with the chuck table rotating at \( \sim 300 \ \text{rpm} \). The feed speed for Z1 is higher at 1-6 \( \mu m/s \) as compared to Z2 at 0.2 \( \mu m/s \). For the final dry polishing step, Z3 removes \( \sim 2 \ \mu m \) of Si using ultra-high-mesh wheel grinding in order to remove all the backside saw marks of grinding as part of stress relief processing to improve wafer strength. As for a final Si functional wafer thickness of \( \sim 46 \ \mu m \), Z1, Z2, and Z3 removed \( \sim 677 \ \mu m, \ 60 \ \mu m, \) and \( 2 \ \mu m \) of bulk Si, respectively. The thinned wafers are then transferred to the wafer mounter, which is housed as shown on the left hand side of Fig. 3-3. It serves to mount the backside of the thinned wafers onto the integrated DAF (10 \( \mu m \) thick)/DT films (100 \( \mu m \) thick), followed by backgrinding tape peel-off from the frontside of the wafer. Before DAF/DT tape mounting, UV irradiation is performed to reduce the adhesion of the wafer to the existing backgrinding tape. After tape peeling, the frame mounted DAF/DT laminated thinned wafers are stored in a cassette.

3.2.3 Die separation (DDS) module experimental set-up

There are two main functions of the DDS module. A cool expansion process is used to cleave/dice the DAF film (which is attached on the backside of the wafer after wafer back-g
grinding) and to increase the distance between singulated, thinned dies for ease of die pick up during die attach. Using the characteristics of DAF in which it becomes brittle at low temperatures, expansion is performed in a low temperature environment to realize high precision DAF separation. In addition, for SD processes that allow the dies to be still integrated with the wafer (likely a SDAG approach), the DDS module can help to finish off separating the Si dies. In the SDBG approach reported in this work, DDS eliminates the need for DAF laser grooving as used by DBG. For a process that singulates thin wafers with DAF, there are issues such as DAF burrs forming on the cut surface when full-cut dicing is performed. Using DDS helps to improve the DAF cut quality.

The push-pull arm first pulls the SD-processed and thinned wafer out from the cassette and moves it to the frame centering stage for alignment before moving to the intermediate stage. The thinned wafer then moves to the cooler expansion stage. Here, the table holding the wafer cools down to 0 °C and stays there for 120 s before ascending upwards by 12 mm at a speed of 200 mm/s. It remains at the peak of the ascension for 10 s before returning to its home position. The framed wafer will then move to the heater shrink stage. Here, the heater, 20 mm separated from the table, is set at 200 °C. The table held at 60 °C ascends by 11 mm at a speed of 10 mm/s and stays for 10 s at the peak of the expansion. After this, it returns to the home position but this time, the table rotates at a rate of 3°/s so that the hot air is applied to the outer ring of the DT to shrink it [see Fig. 3-4]. This helps to reduce the DT sag due to the previous expansions. For this work, UV irradiation was not used, and thus the framed wafer is returned to its cassette after DDS processing.
3.3 Experimental analysis: Measurement methods and concepts

This section describes the primary measurement methods, which are used for the work reported in this dissertation.

3.3.1 Absolute reflectance measurements

Spectral properties of the specularly reflecting wafer backside surface consisting of multi-stack layers of thin films such as SiO$_2$, polysilicon, and Si$_3$N$_4$ were measured for the different wafer technologies (19 nm, 1Y, and 1Z) using the JASCO V-670/ARMN-735 UV-Visible-NIR spectrophotometer. The multi-layer thin film stacks on the backside surface of wafer technologies 19 nm, 1Y, and 1Z measure ~340 nm, 1100 nm, and 990 nm in total thickness, respectively. Photosensors were configured in the measurement system to detect the absolute reflectance as the light from the spectrophotometer is swept from 800 nm to 1400 nm in wavelength, in order to represent both the measurement laser and the SD laser which operate at wavelengths of 830 nm and 1342 nm, respectively. It is important to quantify the reflectance of the wafer backside because it can attenuate the incident SD laser.

3.3.2 Atomic Force Microscopy (AFM)

Surface roughness and morphology of the top-most deposited Si$_3$N$_4$ film at the wafer backside surface were characterized using a Digital Instruments Atomic Force Microscope at a scan rate of 1.001 Hz in tapping mode for the different wafer technologies 19 nm, 1Y, and 1Z. The AFM study utilizes a scan size of 10 $\mu$m.

3.3.3 X-Ray Powder Diffractometer (XRD)

A computer-controlled D8 ADVANCE Powder Diffractometer by Bruker using Cu K$\alpha$ radiation (1.54058 Å), operated at 40 kV and 40 mA, was used for phase and compound formation identification. The diffractometer was set to a lower secondary beam path with a continuous scan mode from a start angle of 10° to an end angle of 90° based on a 2Theta scan axis.
Using the diffractometer, microstructural analysis of the die sidewall after SD processing was carried out to understand its phase impact.

### 3.3.4 Three-point bend test

The three-point bending method (Shimadzu EZ Graph) was applied to get the SD-singulated thinned die (46 μm thick) breakage strength (for both the circuit-side/frontside and the backside of the die) and compare it to the existing baseline DBG process. The dies were laminated with a 10 μm thick DAF. The three-point bend test is a type of uniaxial bending test whereby the top surface (in contact with loading rollers) is in a compressive stress state and the bottom surface (in contact with supporting rollers) is in a tensile stress state, with the maximum bending stress at the center of the sample. The region seeing the highest stress is primarily the die backside surface, which also includes a portion of the die edge (on the other hand, the four-point bend test gives rise to a maximum bending moment that is constant in the region between its top two loading points; this region not only includes the surface of the die, but also a larger part of the die edge). Thus, the die strength measured from this method is more sensitive to the largest defects that are closest to the centerline, namely the damage caused by the wafer backgrinding process [63] and to a certain extent, its interaction with the SD-layers near the top edges. The ultimate failure stress is used as the measure of the die strength, while the actual stress is calculated from beam bending theory. The settings of the three-point bend test are such that the spacing between the supporting rollers is 2 mm with a displacement velocity of 4 mm/min (66.7 μm/s). It is important to note that due to the thin dies, the large deflection caused by the high flexibility of the samples as well as the plate structure of the dies (large width as compared to the thickness) can violate certain assumptions for the standard bending stress calculations.

### 3.3.5 Shadow Moiré measurements

Shadow Moiré measurements were carried out using the Akrometrix TherMoir AXP, a modular metrology solution that utilizes the shadow Moiré measurement technique, combined with automated phase-stepping, to characterize out-of-plane displacement for samples up to
400 mm x 400 mm. With time-temperature profiling capability, the TherMoir AXP captures a complete history of a sample’s behavior during a user-defined thermal profile. In this work, SD-singulated thinned die (46 μm thick) samples with 10 μm thick DAF were measured from room temperature to reflow temperatures of ~ 250 °C and back, in order to capture hysteretic (if any) warpage profile behavior as a function of temperature. Warpage profiles of thinned die incoming to die attach are important to be characterized to facilitate the optimization of the die attach process (both die pick up and attach) and hardware design, in addition to identifying potential die cracking and reliability risks. These measurements are also useful to understand potential risks to ultra-thin die stacking and the various temperature treatments they may be subjected to before the molding process.
4.1 Introduction and literature review

For the purpose of ultra-thin die fabrication via a combination of wafer-level singulation before or after backgrinding (or similar derivatives), mechanical blade dicing creates undesirable defects such as chipping, cracks, residual stress, and contamination in the wafer and final die due to its contact-based mechanical nature. This results in low yield, in particular for thinner and smaller dies. As a result, singulation methods using laser’s have been developed [6,30–34,64–66]. These laser’s are focused on the surface of the wafer with optical energy absorbed efficiently to produce ablation or to crystallographically deform the wafer surface. While they do not involve mechanical vibration the same way blade dicing does, the influence of the global heat and debris contaminants during the laser ablation process is of major concern. As a consequence, stealth dicing (SD) has been developed [10–15] in which the laser processing only occurs within the wafer due to the use of a permeable nanosecond pulsed laser which is condensed locally inside the wafer. When scanned in the horizontal direction along the dicing street, a belt-shaped polycrystalline layer with high dislocations density is formed at a controlled depth inside the wafer, which will eventually help to separate individual dies when cracks that spread from this SD-layer progress to the surface of the wafer. Using SD processing thus has several advantages: (1) enabling high-speed dicing
for thin wafers with no defects such as cracks, debris, and chipping typically caused by blade dicing or laser ablation, (2) being a completely dry process and as such, free of contaminants including ionic and water-mark residues, and (3) no kerf loss and highly localized heat affected zones (HAZ). In the literature, there is reported work on SD processing in the areas of thermal simulation to understand the growth of laser-induced temperature distribution with time \cite{10-12, 14, 15}, impact to device reliability \cite{11}, and thermo-elastic-plastic analysis using finite element methods \cite{10, 13}. Chen et al. also reported processing through-silicon-via (TSV) wafers with SD technology \cite{58}.

Figure 4-1: Schematic illustrations of the SD laser processing operations (left) and the formation of the SD-modified internal layer within Si as the SD laser scans along the dicing street (zoom-in).

As discussed in Chapter 2, the SD method uses a permeable laser-induced “perforation” in the form of an SD modified layer on a subsurface level, combined with the application of fracture mechanics to “cleave out” the individual dies. Fig. 4-1 shows schematic illustrations of the SD laser processing operations (left) and the formation of the SD-modified internal layer within Si as the SD laser scans along the dicing street (zoom-in). The SD laser propagates to the neighborhood of the laser’s focal point in the wafer using a wavelength permeable to Si so that there is no damage to the laser incident surface of the wafer. The purpose of this chapter is to present heat conduction simulations via an absorption model of condensed Gaussian beam in which the non-linear temperature dependence of absorption coefficient is considered \cite{48-51}. These simulations help explain the mechanics surrounding the formation of the SD-modified layer, with validation to first order through subsequent experiments.
4.2 Process physics and simulations: Temperature rise of silicon due to absorption of permeable nanosecond pulsed laser

This section reports on the development of the stealth dicing heat analysis model followed by process simulations in order to validate and advance the understanding of its processing physics and manufacturing constraints.

4.2.1 Analysis model and methods

The primary heat analysis uses an absorption model for a condensed Gaussian beam with a non-linear temperature dependence of the absorption coefficient in Si. The simulation model considers heat generation due to absorption, and heat diffusion into the surrounding Si and Cu layers. A Cu layer with a thickness of 0.5 \( \mu \text{m} \) represents the frontside patterning elements of the wafer, providing a Si to Cu interface for thermal diffusivity and light absorption purposes. The Cu layer here assumes 100% absorption of the SD laser beam should it reach that layer. Similarly, a Si to air interface is used to represent the operations of the SD laser’s entry point via the wafer backside. Different from the assumptions made in Ref. [11], the backside surface of the Si assumes 82.3% reflectance at the 1.342 \( \mu \text{m} \) operating wavelength of the SD laser to represent the actual reflectance measurements (to be reported in Chapter 5) conducted on 1Y technology wafers, the primary test vehicles used in this work. In addition, we also assume that the SD laser beam has a spectral transmission of 90% (10% absorption is assumed corresponding to \( \sim 1.6 \text{ cm}^{-1} \) in absorption coefficient) as it propagates through the bulk Si from an initial thickness of 780 \( \mu \text{m} \) to a z-height of 120 \( \mu \text{m} \) as measured from the wafer frontside. Because the band gap of Si is 1.1 eV, the SD laser beam at a wavelength of 1.342 \( \mu \text{m} \) is able to pass through with this high transmission. The 120 \( \mu \text{m} \) value is chosen because the simulation region of Si is a disk with a diameter of 60 \( \mu \text{m} \) and the thickness is 120 \( \mu \text{m} \). Fig. 4-2 illustrates the analysis model assumed in the simulations.

SD pulse width is assumed to be 100 ns (the actual SD laser pulse width is proprietary), comparable to the simulation and experimental space reported elsewhere where pulse width
is 150 ns [10–15]. Time of pulse center is assumed at $t = 0$. At the same time, the SD laser's repetition frequency is set at 90 kHz. The simulations (Mathematica, Wolfram Research, Inc.) were conducted based on a Gaussian beam SD laser light source with $\lambda = 1.342 \, \mu m$ (i.e., the operating wavelength of the SD laser) going through an objective lens without undergoing Fraunhofer diffraction, producing a 2 $\mu m$ final spot size at focus. This spot size is approximated by the average dimensions of the voids generated by the SD laser at its focal point as observed by cross-section SEM microscopy, which will be discussed further in Chapter 5. The simulations aim to capture the spatial (radial and z-direction) and temporal distribution of HAZ effects resulting from single pulse irradiation, and varying SD laser focal distance from the wafer frontside and backside. Table 4.1 lists the key irradiation and material conditions used in the simulations.

The principal heat conduction equation which should be solved is given by:

$$\rho C_p \frac{\partial T}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left( r K \frac{\partial T}{\partial r} \right) + \frac{\partial}{\partial z} \left( K \frac{\partial T}{\partial z} \right) + \omega$$

(4.1)

where $T$ is temperature, $\rho$ is density, $C_p$ is isopiestic specific heat, $K$ is thermal conductivity, and $\omega$ is internal heat generation per unit time and unit volume. The temperature dependence of isopiestic specific heat [67] and thermal conductivity [68] is considered.

In these simulations, similar to those reported in Refs. [10–15], we consider the SD laser beam to be axisymmetric and therefore introduce the cylindrical coordinate system $O-rz$ whereby the $z$-axis corresponds to the optical axis of the laser beam and the $r$-axis is situated on the backside surface of Si. We define the absorption coefficient in a lattice $(i,j)$
<table>
<thead>
<tr>
<th>Simulation parameters</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si thickness x diameter</td>
<td>120 μm x 60 μm</td>
</tr>
<tr>
<td>Pulse width $\tau_p$</td>
<td>100 ns</td>
</tr>
<tr>
<td>Repetition frequency $f$</td>
<td>90 kHz</td>
</tr>
<tr>
<td>Laser wavelength $\lambda$</td>
<td>1.342 μm</td>
</tr>
<tr>
<td>Wafer backside reflectance $R$</td>
<td>82.3 %</td>
</tr>
<tr>
<td>Spot size $x$</td>
<td>2 μm</td>
</tr>
<tr>
<td>Average laser power $P$</td>
<td>1.4 W, 1.7 W, 2.0 W</td>
</tr>
<tr>
<td>Pulse laser energy $E_p$</td>
<td>15.4 μJ, 18.7 μJ, 22.2 μJ</td>
</tr>
<tr>
<td>SD-layer focal spot distance from wafer frontside $Z_{SD}$</td>
<td>20 μm, 35 μm, 50 μm, 70 μm, 80 μm, 95 μm, 740 μm, 755 μm</td>
</tr>
</tbody>
</table>

Table 4.1: Key irradiation and material conditions used in the simulations.

whose temperature is $T_{i,j}$ using $\alpha_{i,j}(T_{i,j})$. When Lamberts law is applied to a small depth element $\Delta z$ from depth $z = z_{j-1}$ to $z = z_j$, the laser intensity $I_{i,j}$ at the depth $z = z_j$ is expressed by:

$$I_{i,j}^* = I_{i,j}e^{-\alpha_{i,j}\Delta z} \quad (4.2)$$

where $I_{i,j}$ is the laser intensity at the depth $z = z_{j-1}$.

The $1/e^2$ radius of the laser beam at the depth $z$ which is condensed with a lens is expressed by $r_c(z)$. When the laser light beam propagates from the depth $z = z_{j-1}$ to $z = z_j$, the condensing or divergence of the beam can be evaluated by the below ratio:

$$\gamma_j = r_c(z_j)/r_c(z_{j-1}) \quad (4.3)$$

where $\gamma_j$ is known as the ratio of condensation [11].

A beam is condensed when $\gamma_j$ is less than 1, and is diverged when larger than 1. Laser intensity $I_{i,j}$ at the depth $z = z_{j-1}$ for a finite difference grid $(i,j)$ can be further expressed using the law of energy conservation as follows:
Condensed: For $\gamma_{j-1} < 1$, $I_{i,j} = \frac{(\gamma_{j-1}^2 r_i^2 - r_{i-1}^2) I_{i,j-1}^* + (1 - \gamma_{j-1}^2) r_i^2 I_{i+1,j-1}^*}{\gamma_{j-1}^2 (r_i^2 - r_{i-1}^2)}$ (4.4)

Diverged: For $\gamma_{j-1} > 1$, $I_{i,j} = \frac{(\gamma_{j-1}^2 r_i^2 - r_{i-1}^2) I_{i-1,j-1}^* + (r_i^2 - \gamma_{j-1}^2 r_i^2) I_{i,j-1}^*}{\gamma_{j-1}^2 (r_i^2 - r_{i-1}^2)}$ (4.5)

Considering Eq. 4.2, internal heat generation per unit time and unit volume in the finite difference grid $(i, j)$ is thus:

$$w_{i,j} = (1 - e^{-\alpha_{i,j} \Delta z}) I_{i,j} / \Delta z$$ (4.6)

The analysis region of Si for the simulations is a disk with a diameter of 60 $\mu$m and thickness 120 $\mu$m. The temperature dependence of the absorption coefficient $\alpha_{i,j}(T_{i,j})$ is taken into account using Refs. [48–51].

### 4.2.2 Effects of laser pulse energies and focal distance on the spatial and temporal heat affected zones within Si

Figs. 4-3 to 4-5 show simulated time evolution (65, 70, 75, 80, 100, 150, 200, and 300 ns) of SD laser-induced heat affected zones (HAZs) at different $z$ focal distances, ranging from 20 to 70 $\mu$m as measured from the wafer frontside. These simulations are performed for laser average power of 1.4 W, 1.7 W, and 2.0 W, which correspond to pulse laser energies (PLEs) of 15.4 $\mu$J, 18.7 $\mu$J, and 22.2 $\mu$J, respectively (PLE = average power/frequency; ignoring wave deformation as it propagates through the Si wafer). The incident SD laser pulse shape in the form of a single shot irradiation is assumed to be Gaussian (in time) as it begins to rise in its peak power. The initial temperature is set to be room temperature of 293 K.

In all the cases for Figs. 4-3 to 4-5, full absorption of the laser’s energy occurs at $\sim$ 75 ns, as evident from the drastic drop in the peak power of the “through” beam as observed right above the Cu plane (i.e., below the laser’s respective focal points). It is important to note that at 75 ns, the SD laser beam has not reached its peak power. Fig. 4-6 shows the simulated peak power of the incident SD laser beam as observed at $\sim$ 120 $\mu$m from the wafer.
Figure 4-3: Time evolution of SD laser-induced (Pulse Laser Energy (PLE)/Laser Average Power of 15.4 µJ/1.4 W) heat affected zones at a z focal distance from the wafer frontside at (a) 20 µm, (b) 35 µm, (c) 50 µm, and (d) 70 µm. The zoom-in images show the temperature contours corresponding to 75 ns, 100 ns, and 200 ns pulse timing for $z = 35$ µm.

frontside (before the beam converges into the focal point) and right above the Cu plane (as the beam diverges out of the focal point) at a $z$ focal height of 35 µm. The results of the simulations are similar for all $z$ focal heights of 20 µm, 35 µm, 50 µm, and 70 µm.

At $\sim 75$ ns, due to the non-linear temperature dependence of the absorption coefficient and the high peak power intensity in the focal volume, complete laser absorption occurs abruptly in the focal point vicinity, and as a result, a void is generated. The presence of this void measuring $\sim 2$ µm in diameter has been validated in experiments conducted in this work, and are reported in Chapter 5. This void is caused by the instantaneous evaporation of Si as the temperature rises to about 18000 °C locally. It is noteworthy that if the absorption coefficient is independent of temperature and is the value at room temperature, the maximum temperature rise is only about 360 K, which is much smaller than the melting point of 1690 K ($\sim 1416$ °C) under atmospheric pressure [12]. This demonstrates the importance of the absorption coefficient having positive temperature dependence in order to modify the Si at the focal point. It can be seen from Figs. 4-3 to 4-5 that the heated regions expanded rapidly
Figure 4-4: Similar simulation conditions to Fig. 4-3 but with Pulse Laser Energy (PLE)/Laser Average Power of 18.7 μJ/1.7 W.

Figure 4-5: Similar simulation conditions to Fig. 4-3 but with Pulse Laser Energy (PLE)/Laser Average Power of 22.2 μJ/2.0 W.
and grew only towards the SD laser beam incident surface and did not grow towards the Si frontside. This is consistent with the simulations reported in Refs. [10-12]. It is believed that since the absorption coefficient increases non-linearly with increasing temperature, it becomes higher for the heated regions than at room temperature and thus, the incident laser beam is absorbed mainly around the tip of the leading edge of the high temperature region, yielding a thermal shockwave. This causes the heated regions to grow only towards the beam incident surface. Immediately after the void is generated, a thermal shockwave [see Fig. 4-1 for a schematic of mechanism] propagates toward the upper surface, and the high dislocation density layer, i.e., the SD-layer, is formed. When the thermal shockwave pushes up against the SD-layer formed from the previous pulse irradiation (as it scans), a crack, whose initiation is a dislocation, is produced. This mechanism for SD processing to enable die singulation is able to explain the simulation and experimental results [see Chapter 5] very well. The velocity of the thermal shockwave can be approximated from Figs. 4-3 to 4-5 to be about 300 m/s, which is much higher than the velocity of thermal diffusion. This explains the fact that as the thermal shock wave propagation is near complete, the contour of the high temperature area becomes gradually vague and less “sharp”.

Figure 4-6: Simulated peak power of the incident SD laser beam as observed at ~ 120 µm from the wafer frontside (before the beam converges into the focal point) and right above the Cu plane (as the beam diverges out of the focal point) at a z focal height of 35 µm. Results of the simulations are similar for all the z focal heights considered.
4.2.3 Effects of laser pulse energies and focal distance on the spatial and temporal heat affected zones at frontside Cu layer

Figure 4-7: Time evolution of SD laser-induced (Pulse Laser Energy (PLE)/Laser Average Power of 15.4 μJ/1.4 W) heat affected zones at the frontside Cu layer with SD laser focal height \( z \) of (a) 20 μm, (b) 35 μm, (c) 50 μm, and (d) 70 μm. The arrows indicate the corresponding pulse timing in which maximum temperature at the Cu layer is achieved.

From Figs. 4-3 and 4-4, it appears that possible laser ablation risks on the wafer frontside can be present, based on the observation of a secondary HAZ near the Cu layer when the \( z \) focal distance is set at 35 μm or less for an average laser power of 1.4 W and 1.7 W. When the average laser power is increased to 2.0 W, the laser ablation risks exist even up to a \( z \) focal distance of 50 μm, as evident from Fig. 4-5. Also, the distribution of the HAZs is fairly similar in Figs. 4-3 to 4-5 despite the difference in \( z \) focal distance, due to the transparency of Si to the operating wavelength of the SD laser.

To further understand the possible risks of ablation-related damage to the frontside of the Si wafer using SD processing, Figs. 4-7 to 4-12 aim to simulate the time evolution (65, 70, 75, 80, 100, 150, 200, and 300 ns) of SD laser-induced HAZs at the frontside Cu layer (as opposed to bulk Si) with SD laser focal height \( z \) of (a) 20 μm, (b) 35 μm, (c) 50 μm, and (d)
70 μm. Similar to Figs. 4-3 to 4-5, these simulations are performed for laser average power of 1.4 W, 1.7 W, and 2.0 W, which correspond to PLEs of 15.4 μJ, 18.7 μJ, and 22.2 μJ, respectively. In Figs. 4-7, 4-9, and 4-11, the arrows indicate the corresponding laser pulse timing in which the maximum temperature at the Cu layer is achieved during the entire course of a single pulse irradiation. Meanwhile, Figs. 4-8, 4-10, and 4-12 show the time evolution of the SD laser-induced temperature at different radii positions, ranging from 0 μm (center) to 14 μm from the projected focal center of the SD laser beam on the frontside Cu layer plane, for laser average power of 1.4 W, 1.7 W, and 2.0 W, respectively.

Figure 4-8: Time evolution of the SD laser-induced temperature at different radii positions, ranging from 0 μm (center) to 14 μm from the projected focal center of the SD laser beam on the frontside Cu layer plane with SD laser focal height z of (a) 20 μm, (b) 35 μm, (c) 50 μm, and (d) 70 μm. The Pulse Laser Energy (PLE)/Laser Average Power used here is 15.4 μJ/1.4 W.

Figs. 4-7 and 4-8 show that the maximum temperatures achieved at the frontside Cu plane at an PLE/average power of 15.4 μJ/1.4 W are 2872 °C, 1949 °C, 968 °C, and 508 °C for an SD laser focal height z of 20 μm, 35 μm, 50 μm, and 70 μm, respectively. From Figs. 4-9 and 4-10, the maximum temperatures achieved at the frontside Cu plane, for a somewhat larger PLE/average power of 18.6 μJ/1.7 W, are 3026 °C, 2266 °C, 1150 °C, and 609 °C for an SD laser focal height z of 20 μm, 35 μm, 50 μm, and 70 μm, respectively. Finally, Figs. 4-11 and 4-12 demonstrate that the maximum temperatures achieved at the frontside Cu plane, now at the largest PLE/average power of 22.2 μJ/2.0 W, are 3135 °C, 2785 °C, 1670 °C, and 955 °C for an SD laser focal height z of 20 μm, 35 μm, 50 μm, and 70 μm, respectively.
2536 °C, 1368 °C, and 721 °C for an SD laser focal height $z$ of 20 μm, 35 μm, 50 μm, and 70 μm, respectively. Not surprisingly, the maximum temperature is achieved at the center of the projected focal spot of the laser on the Cu plane in all three cases of different PLEs, and this temperature reduces drastically as one moves further away from the center point. From Figs. 4-7 to 4-12, it can also be deduced that the maximum temperature is reached at laser pulse timings around 75-80 ns for the different PLEs simulated.

![Diagrams showing temperature at the frontside Cu layer for different PLEs and pulse timings](image)

Figure 4-9: Similar simulation conditions to Fig. 4-7 but with Pulse Laser Energy (PLE)/Laser Average Power of 18.7 μJ/1.7 W.

The results from Figs. 4-7 to 4-12 also show that the maximum temperature reached at the Cu plane increases the most when $z$ is at 35 μm, i.e., an increase of 317 °C and 587 °C when incident power increases to 1.7 W and 2.0 W, respectively (from 1.4 W). The increment is lower when $z$ is larger or smaller than 35 μm. Given that the melting and boiling point temperatures for Cu are ~ 1085 °C and 2571 °C correspondingly, the simulation results predict possible damage risks due to laser ablation on the wafer frontside where the Cu layers are (in the form of metallization layers or TEG structures) when the SD laser beam is focused at or below a $z$ height of 35 μm, 50 μm, and 50 μm for laser average power of 1.4 W, 1.7 W, and 2.0 W, respectively. This was experimentally proven in Chapter 5 in the
Figure 4-10: Similar simulation conditions to Fig. 4-8 but with Pulse Laser Energy (PLE)/Laser Average Power of 18.7 μJ/1.7 W.

Figure 4-11: Similar simulation conditions to Fig. 4-7 but with Pulse Laser Energy (PLE)/Laser Average Power of 22.2 μJ/2.0 W.
case of laser power equal to 2.0 W. It is worthy to note that the specific heat capacity for Cu (0.39 kJ/kg K) is much lower than that of Si (0.71 kJ/kg K).

4.2.4 Radial distribution of heat affected zones at frontside Cu layer

Figs. 4-13 to 4-15 show the simulated maximum temperature reached on the frontside Cu plane at different distances from the projected focal center of the SD laser beam, at various $z$ focal heights (20 $\mu$m to 70 $\mu$m) for PLEs/laser average power of 15.4 $\mu$J /1.4 W, 18.7 $\mu$J /1.7W, and 22.2 $\mu$J /2.0 W, respectively. It can be seen that the maximum temperature reached at the Cu plane depends on the $z$ focal height, due to spot size spreading from the focal spot. In general, the higher the $z$ value is, the larger the spread of the maximum temperature simulated at the Cu plane.

In the case of the PLE/laser average power of 15.4 $\mu$J /1.4 W, it can observed from Fig. 4-13 that the maximum temperature exceeds 400 °C (the typical glass transition temperature limits for organic based passivation layers such as Polyimide found on the wafer frontside) only for an area within 8.5 $\mu$m in diameter from the projected focal point on the Cu plane. For a PLE/laser average power of 18.7 $\mu$J /1.7 W [see Fig. 4-14], this area increases to $\sim$ 9.3
Figure 4-13: Maximum temperature reached on the frontside Cu plane at different distances from the projected focal center of the SD laser at various z focal heights. The Pulse Laser Energy (PLE)/Laser Average Power used here is 15.4 μJ/1.4 W.

Figure 4-14: Similar simulation conditions to Fig. 4-13 but with Pulse Laser Energy (PLE)/Laser Average Power of 18.7 μJ/1.7 W.

Figure 4-15: Similar simulation conditions to Fig. 4-13 but with Pulse Laser Energy (PLE)/Laser Average Power of 22.2 μJ/2.0 W.
μm in diameter. When the PLE/laser average power is increased to 22.2 μJ /2.0 W, the area further amplifies to ~10 μm in diameter. These results show that the reduction of dicing street width for the benefit of reducing per unit die costs is likely limited by the spread of HAZ on the Cu plane at the wafer frontside, given that the kerf loss due to SD processing is much smaller (~2 μm on average). While one can reduce z to lower the beam spread, this actually risks more severe laser-induced ablation damage in the form of cracks, debris, and even chipping to the wafer frontside because of the very high temperatures induced. Also, the specific value of z selected also depends strongly on the exact thermal budget one seeks to avoid on the wafer frontside. Nonetheless, when compared to mechanical blade dicing, SD processing still has a significantly smaller kerf exclusion zone within the dicing street, demonstrating its substantial potential to reduce per unit die costs.

4.2.5 Effects of laser focal distance from wafer backside on heat affected zones

Fig. 4-16 shows the time evolution of SD laser-induced (for Pulse Laser Energy (PLE)/Laser Average Power of 22.2 μJ/2.0 W) heat affected zones at a z focal distance from the wafer frontside at (a) 740 μm (or 40 μm from the wafer backside) and (b) 755 μm (or 25 μm from the wafer backside). When the SD laser beam is focused at 25 μm from the wafer backside [see Fig. 4-16(b)], in addition to the sudden and localized laser energy absorption at its focal spot, the thermal shock wave propagates in the surface direction until about 200 ns when the laser absorption suddenly begins at the surface. Despite the fact that the laser power has already passed its peak and is gradually decreasing, the surface temperature can actually rise beyond 14000 °C, which is comparable to the maximum temperature reached at the focal spot within Si. This is mainly because of the fact that the heat near the surface can only be diffused in the inside of the lower half and as a result, the surface temperature becomes very high and is maintained for a comparatively longer time (diffusion velocity is much slower than the thermal shock wave velocity). This in turn will mean that severe ablation will occur at the wafer backside surface due to the high temperature generated. Experimental evidence of such a phenomenon has been reported in Ref. [12].
4.3 Conclusions and perspectives

In summary, this chapter has presented the development and the results of stealth dicing process simulations to validate and advance the understanding of its physics and mechanism within the solution space (less than 30 microns final die thickness) and integration flow of interest (SDBG). The simulation results show that the formation of the interior modified layer (SD layer), which consists of discrete voids and a belt layer of high density dislocations, are primarily caused by the non-linear temperature dependence of the absorption coefficient and the propagation of a high velocity thermal shock wave toward the laser incident surface. The thermal simulations also demonstrate the effects of laser pulse energies and focal distance on the spatial and temporal HAZs within Si, on the frontside Cu plane, and on the Si wafer backside (laser incident surface). These lead to the conclusion that a suitable focal plane depth and laser irradiation condition is necessary for an optimal SD process, in order to avoid laser-induced ablation damage on the wafer frontside and backside. The highly localized confinement of the HAZs within the Si, as well as the relatively tight radial distribution
of HAZs at the frontside Cu layer show that an optimal SD process is indeed a strong die singulation candidate to enable ultra-thin die stacked die assembly with zero defects (using the SDBG integration process) coupled with the potential of reduced dicing street width in future designs.
Chapter 5

Stealth Dicing: Process characterization, optimization, and integration

5.1 Introduction and literature review

A review of literature pertaining to stealth dicing research and technology indicates a general focus on theoretical frameworks and simulations, with limited scope and depth in systemic empirical work [10–15]. The few who have conducted a more systematic, albeit limited, experimental work on stealth dicing have assessed SD processing quality based on microstructure and photodiode device characteristics [11, 12, 15], demonstrated the importance of focal plane depth in preventing the thermal shockwave from reaching the backside surface [12], and reported SD-related defects [15], die strength [15, 69], and stress distribution analysis [15]. Most of the SD-related experiments have been performed to realize singulated 50 pm thick Si die using the Stealth Dicing After Backgrinding (SDAG) approach [11, 12, 15]. Attempts have also been made to apply SD processing on TSV wafer 100 μm thick [58], laser subsurface machining of transparent materials [70], low-κ dielectrics [59], and to enable MEMS devices [53, 54, 57, 59]. It is worth noting that when using the SDAG approach, for very thin wafers, i.e., less than 30 μm thick, precise control of the focused SD laser beam
within the wafer can be very challenging [69]. Also, the sensitivity of SD processing to the crystallographic anisotropy of Si [71, 72] is a concern that needs to be addressed [73].

The purpose of this chapter, the first of two experimental chapters in this dissertation, is to disseminate a comprehensive and systematic experimental study of stealth dicing (SD) technology in terms of process characterization, optimization, and integration development. Specifically, this chapter focuses on the characterization and optimization of the SD process in order to realize zero-defect, singulated ultra-thin die (as thin as, but not limited to 25 m) using the Stealth Dicing Before Grinding (SDBG) approach. In this chapter, the influence of wafer technologies, SD-modified microstructures, laser scanning speed, laser average power, laser focal position, test element groups (TEGs), and multi-strata SD processing is investigated to assess impact to dicing quality, both quantitatively and qualitatively. Additionally, design of experiments (DOEs) are formalized to include detailed sensitivity analysis, while paying attention to multi-factor interactions and their impact on the determined responses using the optimized SD process conditions as the center point. This is necessary to determine a production-worthy solution space. In combination with solutions to complex integrated defects involving downstream modules such as the backgrinding and the die separation (DDS) processes, this work has led to the development of a novel partial Stealth Dicing Before Grinding (p-SDBG) integration scheme to encompass high backside reflectance Si wafers and to improve cycle time. Finally, demonstrations of a zero-defect eight die (8D) stack of 25 μm thick non-functional memory dies, and of 46 μm thick functional memory dies (64GB), are shown for the first time, with a product intercept ultimately realized.

### 5.2 Ultra-thin stacked memory die architecture

Mobility-driven consumer products like smart phones, digital cameras, phablets, tablets, ultra-thin laptops, wearable technologies, and other emerging wireless devices in the realm of Internet of Things (IoT) require maximum functional integration in the smallest footprint, lowest profile, high yield, and low cost package. Chip scale packages (CSPs) have minimized the footprint to achieve a chip/package area ratio of about 80%. 3-D packaging has increased that ratio to an impressive level of larger than 200% without increasing the thickness or the
footprint of the package [74]. High density packaging of semiconductor devices, in particular
flash memory, on a substrate is accomplished by stacking bare semiconductor die or stacking
packages in the z-direction and interconnecting them with wire bonding.

In 3-D packaging, there are two main challenges. First, the stacking of dies in a thinner
package stretches the performance envelope of all assembly processes, materials, and equip-
ment. High yield and reliable assembly of thin packages are critical. Second, the testing
of the final module requires that package design allow access to all of the chips and this
increases the test complexity [74]. In stacked die assembly, low cost of ownership of the die
preparation module requires that all stacked die are high yielding, otherwise good die will be
wasted if one die of the stack is bad since rework is not an option [74]. And to make matters
worse, ultra-thin die typically generate higher sensitivity assembly interactions, which result
in increased defect modes and decreased quality characteristics. Therefore, it is very impor-
tant to first produce zero-defects in singulated die in order to realize a high density stack of
ultra-thin memory dies packaged with considerable yield. This is where SD processing as a
viable zero defect singulation technology comes in. In addition to the need for developing
a production-worthy zero-defects singulation solution using SD processing, it is critically
necessary to understand how final stacked die assemblies respond to SD singulated dies in
terms of integrated defects and parametric effects.

For this purpose, Fig. 5-1 shows a schematic of a mechanical p-SDBG test vehicle for
25 µm thick memory dies, consisting of an eight layer memory die stack with single-sided
bonding pads progressively staggered (staircase structure) in the form of two four-die blocks.
This architecture is used to identify SD-specific integrated defects and the lowering of pro-
cess margins generated with downstream process modules including die attach, wire bonding,
molding, etc. Here, it is worth pointing out that in the actual assembly integration experi-
ments involving the mechanical p-SDBG test vehicle, the thick 25 µm wire embedded film
(WEF) as seen in Fig. 5-1 is replaced with a thin 10 µm die attach film (DAF), similar to
the rest of the remaining dies in order to simplify the integration logistics. Also, due to the
use of DAF instead of a thicker WEF, there is no wire bonding performed between die D4 to
D3. This minor variation should not deter the test vehicle from achieving its original intent.

Meanwhile, Fig. 5-2 shows the schematic of an eight layer memory die stack with single-
sided bonding pads staggered one-way progressively (staircase structure). This architecture is in turn used as an electrically functional p-SDBG test vehicle for 46 \( \mu \text{m} \) thick memory dies to realize a 64 GB Micro-SD memory card. The main focus of this functional test vehicle is to understand the parametric and reliability impact due to SD processing, if any.

Figure 5-1: A schematic of an eight layer memory die stack with single-sided bonding pads progressively staggered (staircase structure) in the form of two four-die blocks. This architecture is used as a mechanical p-SDBG test vehicle for 25 \( \mu \text{m} \) thick memory dies.

Figure 5-2: A schematic of an eight layer memory die stack with single-sided bonding pads staggered one-way progressively (staircase structure). This architecture is used as an electrically functional p-SDBG test vehicle for 46 \( \mu \text{m} \) thick memory dies to realize a 64 GB micro-SD memory card.

5.3 Results and discussion: Stealth dicing process characterization

This section reports a thorough quantitative and qualitative process characterization of stealth dicing in order to increase the understanding of its processing quality (defects and parametric performance) and solution space.
5.3.1 Influence of wafer technologies

As described in Chapter 3, both the monitor and functional 2-D NAND flash memory wafers can come in three different technology flavors, i.e., 19 nm, 1Y, and 1Z. In the SDBG process, backside reflectance and surface roughness play an important role in SD processing because they influence the amount of laser energy and its propagation properties when entering the subsurface of bulk Si. In addition, these surface characteristics can also influence the proper identification of the backside surface profile by the SD measurement laser, which can affect warpage correction and SD-layer z-height variability. Knowing the crystal orientation on the side wall of the dies patterned on the wafers is also important to understand cleaving performance and defects.

A. Backside reflectance measurements

Fig. 5-3 shows the measured spectral properties of the specularly reflecting wafer backside due to different thin film stacks for different wafer technologies, specifically 19 nm, 1Y, and 1Z. The backside surface of wafer technologies 19 nm, 1Y, and 1Z appear visibly purple, green, and light brown, respectively [see inset images of Fig. 5-3] as a result of the different multi-stack layers of thin films such as SiO₂, polysilicon, and Si₃N₄. The multi-layer thin film stacks on the backside surface of wafer technologies 19 nm, 1Y, and 1Z measure ~ 340 nm, 1100 nm, and 990 nm in total thickness, respectively. It can be seen from Fig. 5-3 that the absolute backside reflectances for 19 nm, 1Y, and 1Z wafers at the SD laser operating wavelength of 1342 nm are 13.4%, 82.3%, and 65%, respectively. This means that 1Y wafers will attenuate the SD laser energy the most as laser light passes through the air/Si backside interface in the SDBG integration approach. In other words, ceteris paribus, applying SD processing to 1Y wafers is highly inefficient from an energy coupling point of view and risks not creating enough SD-modified controlled damage to allow for the subsequent singulation of the wafer. This can potentially make the SD process non-manufacturable. It also means that by characterizing and optimizing SD process on 1Y wafers, the solution space defined can be used to encompass or envelope the remaining two wafer technologies. The successful optimization of SD processing to 1Y wafers provides flexibility to the ambiguity of product mix and loading using the different wafer technologies. From the results of Fig. 5-3, it
can also be seen that the absolute backside reflectances for 19 nm, 1Y, and 1Z wafers at the SD measurement laser operating wavelength of 830 nm are 53.3%, 93.9%, and 26.6%, respectively. Because all three reflectance values are greater than 10%, the SD measurement laser has no issues performing the backside surface profile mapping in order to compensate for inherent wafer warpage effects, due to high sensitivity of its detection capability.

![Graph showing reflectance values for 19 nm, 1Y, and 1Z wafers.](image)

Figure 5-3: Measured spectral properties of the specularly reflecting wafer backside due to different thin film stacks for different wafer technologies: 19 nm, 1Y, and 1Z.

**B. Backside surface morphologies**

Fig. 5-4 plots the backside surface root mean square (RMS) surface roughness, $R_a$, for different wafer technologies 19 nm, 1Y, and 1Z, with the inset showing an atomic force microscope (AFM) image of a representative 1Y wafer backside surface. As can be seen from the inset of Fig. 5-4, there are certain backside surface artifacts as a result of adhesive residue from DAF removed manually from the wafer backside. The AFM imaging of the film surface on all three types of wafer technologies reveals that the films have a fairly matched surface roughness of ~ 0.4 to 0.8 nm with no porosity or voiding observed. The films appear amorphous and homogeneous. This is not surprising given the fact that the uppermost films on the backside of all three wafer technologies are silicon nitride films ranging from 15-23 nm in thickness. The results from Figs. 5-3 and 5-4 therefore show that the collective spectral
property of the entire stack of thin films, including interfacial interactions, is the dominant factor influencing the coupling of the incident SD laser beam via the wafer backside into the wafer. This has been empirically measured with results shown in Fig. 5-3.

C. Sidewall wafer crystal orientation

Fig. 5-5 shows the XRD patterns of the 1Y wafer die sidewalls along the length and the width, with the inset showing the XRD patterns of its backside. In order for the die sidewalls to fully encapsulate the large spot size (∼ 1 mm) of the XRD beam (with no auto alignment capability), two 1Y dies are stacked on top of each other, measuring 775 µm x 2 = 1550 µm in total thickness. These full thickness dies are manually separated using a simple scribe and break technique using a diamond pen along the patterned dicing streets. It can be concluded from Fig. 5-5 that the cleavage plane for the 1Y wafer (also for 19 nm and 1Z wafers since they are patterned in the same way with respect to the wafer notch) is \{110\} and the frontside/backside plane is \{100\}. This is consistent with the diamond structure of \{100\}-plane Si wafers where \{111\} and \{110\} are the two principal cleavage planes [see Fig. 5-6]. Fig. 5-6 illustrates the cleavage planes \{110\} of \{100\}-plane Si wafers, where the scribe lines are aligned with the \langle 110 \rangle directions [6]. The \langle 110 \rangle propagation directions (scribe line directions), as the easiest slip direction family, are preferred on both cleavage planes [6]. The crack propagation on the \{110\} plane is very anisotropic, whereas the cleavage

Figure 5-4: Backside surface root mean square (RMS) surface roughness $R_a$ for different wafer technologies 19 nm, 1Y, and 1Z. Inset: AFM image of a representative 1Y wafer backside surface.

95
anisotropy on the \{111\} plane is very minor [6,71,72]. These XRD measurements are useful to understand the generation of inter-strata cleave \{111\} plane defects in multi-pass SD processing under certain conditions. This will reported in later sections of this chapter.

Figure 5-5: XRD patterns of the 1Y wafer die sidewalls along the length and the width. Inset: XRD patterns of the 1Y wafer backside. Two 1Y dies are stacked on top of each other, measuring 1550 \(\mu\)m in total thickness, to facilitate the XRD measurements.

Figure 5-6: Cleavage planes \{110\} of (100)-plane Si wafers. Note that the scribe lines are aligned with the \(<110>\) directions [6].
5.3.2 Single-pass and multi-pass stealth dicing layer(s): Key responses

In terms of the experimental responses, both qualitative (e.g., microstructural morphologies, defect modes, and crystal orientation) and quantitative (e.g., SD-layer dimensional changes, kerf dimensional changes, die warpage, DAF separation performance, die strength, electrical tests, and reliability) observations are made. Key SD-layer dimensional responses are identified in Fig. 5-7, which shows a schematic of a representative multi-pass SD processing scan that generates three SD layers (strata) with quantifiable parameters $Z_{SDi}$ and $T_{SDi}$ (where $i = 1-3$). Both single pass and multi-pass SD processing are performed as part of the characterization and optimization efforts reported.

Figure 5-7: Schematic of a representative multi-pass SD processing showing the formation of three SD layers (strata) with key quantifiable parameters SD layer focal $z$-height $Z_{SDi}$ and SD layer height $T_{SDi}$ identified. Both single pass and multi-pass SD processing are performed as part of the characterization and optimization efforts reported.

5.3.3 Effects of laser beam optical parameters (BP)

Fig. 5-8 shows the results of the calibration between focal plane $z$-height of SD layer $Z_{SD1}$ as a function of SD laser defocus position (DF) for different beam optical parameter (BP)
settings. In this experiment, single-pass SD processing was performed at a scanning speed of 500 mm/s with an average power of 2.0 W (PLE = 22.2 μJ), incident on the monitor wafer backside. After SD processing using different BP parameter settings, i.e. 14, 15, and 16, on the same wafer (to avoid wafer-to-wafer variation) along alternating dicing streets, the individual dies on the wafer are separated by manual breaking using the SD1 layer as a “perforated” guide for separation. Z_{SD1} is then measured from the voids located at the bottom of the SD layer to the wafer frontside based on cross-sectional optical microscopy inspection (at 50x magnification) as shown in Fig. 5-7. The different BP settings are used to tune the optical lens in order to reduce its spherical aberration so that the optical parameters are optimal to the intended process at a specific depth within the Si wafer. By choosing different BP settings, one can control the amount of “leaked” SD laser energy that reaches the bottom (frontside) of the wafer by calculating and adjusting the parameters that minimize spherical aberration of the condensed beam. The general trend recommended is that one should move to higher BP settings when focusing the SD laser beam deeper within the Si wafer to avoid undesired damages near the wafer frontside [61].

It can be seen from Fig. 5-8 that for all three BP settings, for a DF value ranging from -201 to -190 μm, Z_{SD1} correlates linearly with DF but with slightly different Z_{SD1} to DF ratios. The DF parameter is a recipe-driven setting that facilitates the ability to form SD layers deep within the wafer using the “deep trace” function. The linear correlation observed validates the “deep trace” linear operating regime, where a certain DF z-direction offset is applied to shift the distance between the measurement lasers focal spot on the surface and the SD lasers focal spot in order to maintain the linearity achieved between a previously calibrated displacement versus measured photodiode voltage curve (generated due to the reflection of the measurement laser beam from the wafer backside) deep within the wafer. Depending on the material (and on the BP settings) to which the SD processing is applied, different Z_{SD1} to DF position ratios are obtained, mainly driven by the refractive index of the material (in this case, Si) and the focused spot properties.

The results in Fig. 5-8 show that for BP settings of 14 and 15, mean depth (measured from backside of wafer starting from a full thickness of 775 μm) to DF ratio is ~ 3.72-3.73, whereas for a BP setting of 16, this is lowered to 3.67 (the refractive index of Si is ~ 3.5
Figure 5-8: Focal plane z-height of SD layer, $Z_{SD1}$ as a function of SD laser defocus position (DF) for different beam optical parameter (BP) settings. Single-pass SD processing was performed at a scanning speed of 500 mm/s with an average power of 2.0 W (PLE = 22.2 μJ), incident on the 2D NAND 1Y monitor wafer backside. Spatial dimensions were measured using cross-sectional optical microscopy.

at a wavelength of $\sim 1.3 \, \mu m$). In other words, the change in the optical focal properties of the condensed beam (by reducing the spherical aberration of the lens) via increasing the BP setting to 16 results in a slightly shallower depth for the SD layer given the same DF positions. This provides a tighter focal point for the SD laser beam across all entry points on the lens and is best used when one wishes to scan at a depth larger than 750 μm with minimal energy leakage to the wafer frontside [61]. One can also deduce from Fig. 5-8 that the z-direction scanning range is about 40 - 41 μm based on the calculated ratios.

Fig. 5-9 shows experimental evidence of frontside (FS) Si ablation damage and interference induced defects as a result of non-optimal selection of BP settings and focal plane z-height. Fig. 5-9(a) plots the mean SD layer height $T_{SD1}$ as a function of mean focal plane z-height $Z_{SD1}$ for different BP settings, with SD-induced defects identified, i.e., FS Cu ablation and/or FS Si ablation. Figs. 5-9(b)-(d) show optical micrographs (left: top view from wafer frontside; right: cross-section view on singulated die sidewalls) demonstrating clean SD processing ($Z_{SD1} = 72 \, \mu m$), FS ablation defects ($Z_{SD1} = 48 \, \mu m$), and FS ablation
and interference defects ($Z_{SD1} = 23 \, \mu m$), respectively, for BP = 15. Here, single-pass SD processing was performed at a scanning speed of 500 mm/s with an average power of 2.0 W (PLE = 22.2 pJ), incident on the 2D NAND 1Y monitor wafer backside.

It is evident from the trend shown in Fig. 5-9 that ceteris paribus, depending on the BP settings, there is a minimum $Z_{SD1}$ below which damage to the frontside of the Si wafer occurs. As can be seen from Figs. 5-9(b)-(d), this damage is in the form of Cu and/or Si ablation on the FS surface, including possible interference defects. This FS damage is primarily driven by laser energy “leakage” toward the FS due to spherical aberration induced by a non-optimal lens setting for a given $Z_{SD1}$. The higher the BP setting, the lower the $Z_{SD1}$ damage threshold. This is consistent with the supplier recommended optimal $Z_{SD1}$ range for
a given BP setting; the range shifts downwards, deeper into the Si from the incident backside surface as BP increases as summarized in Table 5.1 [61].

<table>
<thead>
<tr>
<th>OP settings (a.u.)</th>
<th>Recommended SD layer depth (µm)</th>
<th>Corresponding DF (µm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>~600 to 675</td>
<td>-158 to -176</td>
</tr>
<tr>
<td>14</td>
<td>~650 to 725</td>
<td>-170 to -188</td>
</tr>
<tr>
<td>15</td>
<td>~700 to 775</td>
<td>-182 to -200</td>
</tr>
<tr>
<td>16</td>
<td>~750 to 825</td>
<td>-194 to -212</td>
</tr>
</tbody>
</table>

Table 5.1: Recommended optimal SD layer depth $Z_{SD1}$ and DF range for different BP settings [61].

In addition, it is noteworthy to point out from Figs. 5-9(c) and (d) that Cu FS ablation is noticed on the Cu-based test element groups (TEGs) non-uniformly placed on the patterned dicing streets, while Si-only FS ablation can be detected on dicing streets devoid of these TEGs. Below the accepted $Z_{SD1}$ FS damage threshold, the ablation usually occurs first on the Cu layer (if such a layer exists on the dicing street) followed by Si ablation and/or bulk Si interference defects. One possible reason for this is the fact that Cu has a lower melting point ($\sim 1085 ^\circ$C) as compared to Si ($\sim 1414 ^\circ$C) so they tend to be ablated first. Also, Cu has a lower specific heat capacity. The interference defects in the form of physical striations located at distinct distances from the wafer FS below the SD layer as seen in Fig. 5-9(d) could be a result of SD laser reflection from the FS surface (and interfering as a result of superpositioning of leaked and reflected waves) due to the leaked energy. Fig. 5-10 shows a scanning electron microscope (SEM) micrograph of a zoom-in view on the interference-induced defects between the bottom of the SD layer and the frontside of the wafer.

Table 5.2 summarizes the experimentally observed minimum $Z_{SD1}$ before FS damage occurs, and its corresponding maximum SD layer depth and calibrated DF values as a function of different BP settings for single-pass SD processing at 500 mm/s and 2.0 W (PLE = 22.2 µJ), incident on a 1Y monitor wafer backside. For the purpose of the work reported in this dissertation, a BP setting of 15 is preferred for the range of final die thickness (25 µm to 46 µm) to be realized using the p-SDBG approach, where the lowest SD-layer is optimized at a $Z_{SD1}$ of $\sim 70$ µm (DF value of -190 µm). Results shown in Figs. 5-3 and 5-9 as well as Tables 5.1 and 5.2 demonstrate the importance of having a robust design for manufacturing
(DfM) methodology when it comes to TEG designs and placement along the dicing street, as well as the types of wafer technologies (and their backside films) in order to realize the full benefit of SD processing in producing defect free ultra-thin dies with maximum flexibility and energy coupling efficiency.

![SEM micrograph of a zoom-in view on the interference-induced defects (stripes) between the bottom of the SD layer and the frontside of the wafer.](image)

<table>
<thead>
<tr>
<th>OP settings (a.u.)</th>
<th>Minimum $Z_{SD1}$ to prevent FS Cu and/or Si damage ($\mu$m)</th>
<th>Corresponding maximum SD layer depth ($\mu$m)</th>
<th>Corresponding DF ($\mu$m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>$\sim 52$</td>
<td>$\sim 723$</td>
<td>-193.8</td>
</tr>
<tr>
<td>15</td>
<td>$\sim 48$</td>
<td>$\sim 727$</td>
<td>-195.4</td>
</tr>
<tr>
<td>16</td>
<td>$\leq 17$</td>
<td>$\geq 758$</td>
<td>-206.5</td>
</tr>
</tbody>
</table>

Table 5.2: Experimentally observed minimum $Z_{SD1}$ to prevent FS damage and its corresponding maximum SD layer depth and calibrated DF values as a function of different BP settings before frontside damage occurs for single-pass SD processing at 500 mm/s and 2.0 W (PLE = 22.2 $\mu$J), incident on a 1Y monitor wafer backside.

5.3.4 Effects of laser scanning speed

Fig. 5-11 shows representative images of identified key SD layer microstructural transitions as observed by magnified optical microscopy on the sidewalls of a full thickness 1Y die. The observations suggest that there are two key microstructural transitions: from a dense to an optimal SD-layer structure, and from an optimal to a “fishbone” structure, as the pulse laser energy (PLE) decreases by way of increasing laser scanning speed. *Ceteris paribus*, the magnitude of the dislocation density and polycrystallization of the SD-modified layer changes...
from (left to right of Fig. 5-11) high (dense) to medium (optimal) to low ("fishbone") as laser scan speed increases. Because the SD laser is operating via rapid single shot irradiations in the form of nanosecond (in the hundreds) pulsing with a repetition frequency of 90 kHz, the lowering of scanning speed in effect increases the magnitude of the SD "belt-layer" exposure dose due to pulse overlapping. In Fig. 5-11, the representative dense, optimal, and suboptimal "fishbone" SD layers are produced on 1Y monitor wafers with a laser average power of 2.0 W (PLE = 22.2 μJ) at a ZSD1 ~ 70 μm with scanning speeds of 50 mm/s, 500 mm/s, and 900 mm/s, respectively.

We define the optimal SD-layer structure by examining and identifying the type of SD layer microstructure, both qualitatively and quantitatively, given a set of processing conditions (including the type of wafer technology) such that the density of the dislocations and amount of polycrystallization induced by the thermal shock wave is "just enough" to generate a controlled crack fracture toward the frontside of the wafer (instead of fracturing across the entire full thickness of the wafer as reported in Refs. [10–15]. This technique of controlled SD led to the development of a new integration scheme called p-SDBG, introduced in this dissertation and discussed later in Section 5.5.2 together with the stealth dicing integration development. Fig. 5-11 also illustrates various important characteristics, divided into different zones, of the SD layer. Zone A consists of uniformly spaced voids measuring ~ 2 μm in size along the z-focal plane of the SD laser. These voids form, as previously described in Chapter 2, as a result of the heating phase induced by the extremely high absorption phenomenon that occurs when a certain peak power density is exceeded during the beam condensing process. Zone B is the larger microstructural region where the internally modified SD "belt" layer with high dislocation densities is generated as a consequence of the thermal shock wave produced upwards from the laser focal point vicinity. The systematic characterization produced in this work shows that the density, positioning, and the dimensional characteristics of zone B (along with inter-strata distance for multi-pass SD) play a dominant role in controlling the crack fracture mechanics for high quality and low cycle-time SD processing. Zone C represents the region at the tail-end of zone B where the thermal shock wave propagation is near complete and thermal diffusion starts to take over. It is important to point out that vertical micro-cracks are formed above and below zone B. The
microcracks generated upwards have a higher crack length and frequency as compared to the ones propagating downwards. This is due to the directional nature of the high-velocity thermal shockwave going upwards. It is believed that the micro-cracks propagating downwards are actually an artifact from the manual breaking of the samples required to conduct the optical microscopy. Therefore, they pose minimal risk to the integrity of the quality of the die sidewall after backgrinding down to the final desired thickness. As for the sub-optimal SD microstructure, despite the lower effective dose of the PLE at high scanning speeds, the upper micro-cracks remain consistently large thus producing a “fishbone” like structure.

Figure 5-11: Example of identified key SD layer microstructural transitions as observed by magnified optical microscopy on the sidewalls of a full thickness 1Y die. *Ceteris paribus*, the magnitude of the dislocation density and polycrystallization of the SD-modified layer changes from (left to right): high (dense) to medium (optimal) to low (“fishbone”) as laser scan speed increases. Beam optical parameter (BP) is set at 15 for all cases given the z-range involved.

Fig. 5-12 shows experimental evidence of SD layer microstructural and dimensional transitions observed by optical microscopy on full thickness die sidewalls as a function of laser scanning speed (50 to 900 mm/s) for (a) 1Y, (b) 1Z, and (c) 19 nm monitor wafers. The
symbol “*” indicates the optimal scan speed to produce the optimal cycle-time SD layer microstructure, positioned before a microstructural transition toward a low-density “fishbone” structure for a set of SD processing conditions and wafer technologies. Here, slightly different single-pass SD processing conditions were used for different wafer technologies in order to ensure easy separation of full thickness die (for optical inspection) by manual breaking. For example, higher laser power and a lower ZSD1 are needed for higher backside reflectance wafers such as 1Y wafers. Meanwhile, Fig. 5-13 plots the mean SD layer height TSD1 as a function of laser feed/scanning speed (50 to 900 mm/s) for different monitor wafer technologies (1Y, 1Z, and 19 nm) with the inset illustrating microstructural and dimensional transitions observed by optical microscopy for selected combinations of wafer technology and feed speed.

From Figs. 5-12 and 5-13, several trends, both qualitative and quantitative, are evident. First, as the scanning speed increases from 50 mm/s to 900 mm/s, the SD layer height TSD1 decreases non-linearly [see Fig. 5-13] across all wafer technologies. Qualitatively, it is found that at significantly lower effective energy doses (achieved with higher scanning speeds and higher backside reflectance wafer technologies), the SD layer becomes less dense with dislocation damage and vertical microcracks becoming more prominent. Backside reflectance, R, of the SD laser at its operating wavelength of 1.342 μm is highest for 1Y wafers (R1Y = 82.3%) followed by 1Z wafers (R1Z = 65%) and 19 nm wafers (R19NM = 13.43%). For example, at a laser average power of 2.0 W (PLE = 22.2 μJ), “fishbone” SD layer microstructure arises at laser scan speed v = 900 mm/s. When comparing across wafer technologies, it is obvious that 1Y wafers have a generally lower TSD1 than 1Z and 19 nm wafer technologies because of its significantly higher backside reflectance that limits the effective dose entering the Si wafer from the backside surface. For the 19 nm and 1Z wafer technologies, despite scanning at 900 mm/s with a lower laser average power of 1.7 W (PLE = 18.8 μJ), a clear transition to the “fishbone” microstructure is not immediately obvious. As a result, for lower backside reflecting surfaces, the optimal scanning speed v can be set at higher values, i.e., 700 mm/s for 19 nm and 1Z wafers instead of 500 mm/s for 1Y, and thereby improving the throughput time for SD processing. These optimal scanning speeds for a set of given conditions are extracted not only qualitatively from microstructural observations [see Fig. 5-12], but also from the
Figure 5-12: Experimental evidence of SD layer microstructural and dimensional transitions observed by optical microscopy on full thickness die sidewalls as a function of laser scanning speed (50 to 900 mm/s) for (a) 1Y, (b) 1Z, and (c) 19 nm monitor wafers. "*" indicates the optimal scan speed, positioned before a microstructural transition (represented by the dashed line where applicable) toward a low-density “fishbone” structure for a set of SD processing conditions and wafer technologies. Beam optical parameter (BP) is set at 15 for all cases given the z-range involved.
non-linear dependency plotted in Fig. 5-13 where $T_{SD1}$ starts to plateau off beyond a certain scan speed. The plateauing of $T_{SD1}$ can be explained by the fact that as irradiation pulses separate further and further from one another as scan speed increases, it reaches a point where no overlap of individual irradiation pulses begin to occur. When no overlapping starts to happen, $T_{SD1}$ remains similar because the effective dose becomes a constant thereafter [see Fig. 5-14]. One can expect the “fishbone” structure to emerge when the vertical microcracks remain similar in size while $T_{SD1}$ starts to decrease and plateau off. Results from Figs. 5-12 and 5-13 show that $T_{SD1}$ can indeed be well-controlled by scanning at different speeds. This dependency enables good control of the SD “damage” necessary to realize a reproducible fracture mechanism to begin singulating individual dies from the wafer.

To obtain a better understanding on the SD layer microstructure, Fig. 5-15 shows scanning electron microscope (SEM) micrographs of the SD layers produced using scan speeds of (a) 200 mm/s, (b) 50 mm/s, and (c) 900 mm/s. The inset of Fig. 5-15(a) further shows a magnified SEM image of the circular voids measuring ~ 2 $\mu$m, produced at the lasers focal spot. The SD layers are generated using a laser average power of 1.7 W (PLE = 18.8
Individual laser pulse irradiation

Figure 5-14: Schematic illustration of the separation of individual irradiation pulses as scanning speed increases for a given pulse laser energy.

\[
\text{Increasing laser scan speed, } v
\]

Scan direction

Figure 5-15: Examples of (a, b) dense and (c) “fishbone” SD layer microstructures within 1Y wafers. Scanning electron microscope (SEM) micrographs of SD layer microstructure produced using scan speeds of (a) 200 mm/s, (b) 50 mm/s, and (c) 900 mm/s. The SD layers are generated using a laser average power of 1.7 W (PLE = 18.8 \( \mu \)J) at a ZSD1 of \( \sim 94 \mu m \) (DF = -182 \( \mu m \)). Inset of (a) shows a magnified SEM image of the circular voids measuring \( \sim 2 \mu m \), produced at the laser’s focal spot.

\[
\mu J
\]

at a ZSD1 of \( \sim 94 \mu m \) (DF = -182 \( \mu m \)) from the wafer frontside. Figs. 5-15(a) and (b) demonstrate a dense SD layer while Fig. 5-15(c) shows a “fishbone” SD layer, both generated within 1Y wafers. As described previously, the SD laser beam penetrates through the wafer backside surface and is then absorbed locally at the vicinity of the focal point, producing a void and an SD layer. The voids, which are generated at the early stage of the laser absorption, are evident in Fig. 5-15 in both circular [see Figs. 5-15(a) and (c)] and elongated [see Fig. 5-15(b)] shapes. It is believed that the elongated voids are due to the severely overlapping SD laser focal spots as they scan at a significantly reduced speed of 50 mm/s. This produces high aspect ratio voids elongated to the direction of the optical axis.
5.3.5 Effects of laser average power

Fig. 5-16 shows qualitative experimental evidence of SD layer microstructural and dimensional transitions observed by optical microscopy on full thickness die sidewalls as a function of laser average power (1.0 W to 2.2 W, i.e., PLE from 11.1 µJ to 24.4 µJ) for (a) 1Y, (b) 1Z, and (c) 19 nm monitor wafers. As opposed to Fig. 5-12, the symbol “*” in Fig. 5-16 refers to the laser average power (a representation of PLE given a fixed repetition frequency) to produce an optimal cycle-time SD layer microstructure, positioned before a microstructural transition toward a low-density “fishbone” structure for a set of SD processing conditions and wafer technologies. From Fig. 5-16, it can be seen that the identified optimal average laser power for wafer technologies 1Y, 1Z, and 19 nm are 2.0 W (22.2 µJ), 1.7 W (18.8 µJ), and 1.7 W (18.8 µJ), respectively, for a scan speed of 500 mm for 1Y wafers and 700 mm/s for both 1Z and 19 nm wafers. Again, slightly different single-pass SD processing conditions were used for different wafer technologies in order to ensure easy separation of full thickness die (for optical inspection) by manual breaking.

Fig. 5-17 plots the mean SD layer height $T_{SD1}$ as a function of laser average power (1.0 W to 2.2 W, i.e., PLE from 11.1 µJ to 24.4 µJ) for different monitor wafer technologies (1Y, 1Z, and 19 nm) with its inset showing evidence of microstructural and dimensional transitions observed by optical microscopy on full thickness die sidewalls for selected combinations of wafer technology and laser average power. For the 1Y wafer, two passes of SD processing were made within the wafer in order to facilitate manual separation. The results of the experiments shown in Figs. 5-16 and 5-17 show qualitative and quantitative evidence that as the laser average power increases from 1.0 W (PLE = 11.1 µJ) to 2.2 W (PLE = 24.4 µJ), the SD layer height, $T_{SD1}$, increases non-linearly, in a sigmoidal fashion [see Fig. 5-17]. Qualitatively, analogous to the results shown in Fig. 5-12, it is found that at significantly lower effective energy doses (achieved with lower laser average power and higher backside reflectance wafer technologies), the SD layer becomes less dense with dislocation damage and vertical microcracks becoming more prominent. For example, at a laser average power of 1.0 W (PLE = 11.1 µJ), “fishbone” SD layer microstructure arises at laser scan speed $v$ of 500 mm/s.
Figure 5-16: Experimental evidence of SD layer microstructural and dimensional transitions observed by optical microscopy on full thickness die sidewalls as a function of laser average power (1.0 W to 2.2 W, i.e., PLE from 11.1 μJ to 24.4 μJ) for (a) 1Y, (b) 1Z, and (c) 19 nm monitor wafers. Here “*” indicates the optimal laser average power, positioned before a microstructural transition toward a low-density “fishbone” structure for a set of SD processing conditions and wafer technologies. Beam optical parameter (BP) is set at 15 for all cases given the z-range involved.

Similar to the results in Fig. 5-12, when comparing across wafer technologies, it is obvious that 1Y wafers have a generally lower T_{SD1} than 1Z and 19 nm wafer technologies because of the significantly higher backside reflectance in 1Y wafers that limits the effective dose entering the Si wafer from the backside surface. For the 19 nm and 1Z wafer technologies, despite a low laser average power of 1.0 W (PLE = 11.1 μJ) with a higher scan speed at 700 mm/s, there is no obvious transition to the “fishbone” microstructure. Therefore, for lower backside reflecting surfaces, the optimal laser average power can be set at a lower value, i.e., 1.7 W for 19 nm and 1Z wafers instead of 2.0 W for 1Y, and thereby improving laser lifetime (cost of ownership) for SD processing. In addition to qualitative observations, the optimal laser average power for a given SD condition can also be validated from the non-linear plot shown in Fig. 5-17. From Fig. 5-17, it can be seen that T_{SD1} starts to increase as laser average power increases but begins to plateau beyond a certain point, thus resembling a sigmoidal curve (this is more apparent for 19 nm and 1Z wafer technologies given the
sweeping range for the average power). The decreasing sensitivity of $T_{SD1}$ to laser average power as the latter increases reinforces the need to fully comprehend the minimal “safety” $T_{SD1}$ (or $T_{SDi}$ if using multi-passes) to initiate crack fracture without unnecessarily high laser power usage, in order to reduce the costs of ownership for the SD process. It is clear from Figs. 5-16 and 5-17 that $T_{SD1}$ can be well-controlled by using different PLEs on top of different scanning speeds. This dependency enables another knob to control the magnitude of SD “damage” necessary to realize a production-worthy die singulation technology. Optimal laser average power is selected to complement high scanning speeds in order to reduce SD processing cycle time, balancing the need for power stability and laser head lifetime. For this reason, the optimal condition is usually close to the vicinity of the rising edges on the plot shown in Fig. 5-17.

Fig. 5-18 shows a plot of mean focal $z$-height (first SD layer), $Z_{SD1}$, versus stealth laser average power for different for monitor wafer technologies (1Y, 1Z, and 19 nm). The inset of
Fig. 5-18 shows representative cross-sectional optical microscope images for 19 nm and 1Y wafers at a laser average power of 1.7 W. For the 1Y wafer, two passes of SD processing at a rate of 500 mm/s were made at DFs of -190 μm (ZSD1 = 72.4 μm) and -180 μm (ZSD2 = 105 μm) within the wafer in order to facilitate ease of manual separation (for cross-sectional optical microscopy), while single pass SD processing was used for the 19 nm and 1Z wafers at a rate of 700 mm/s at a DF of -183 μm (ZSD1 = 94.2 μm). It can be observed from Fig. 5-18 that for the given range of laser average power from 1.0 W to 2.2 W, ZSD1 remains fairly stable (± 2 μm) indicating a minimal change in the size of the void (which will affect the measurement of ZSD1) created at the lasers focal spot.

5.3.6 Effects of multi-strata stealth dicing

Multi-strata SD layers are created by scanning the SD laser beams focal spot at different z-heights within the Si wafer along a common dicing street. This technique is usually performed in cases where there is inadequate effective laser exposure dose when using single pass and/or a larger SD “damage” threshold is needed to initiate a well-controlled fracture crack. The need for having more than one SD layer is exacerbated by wafers with high backside reflectance to the SD lasers operating wavelength such as the 1Y wafers. The trade-off in having multi-strata SD processing is an increase in cycle time, as multiple passes are needed.
for each dicing street. Optimal multi-strata positioning, dimensions, and microstructure definition enables a minimum number of SD passes (and hence, minimum cycle time) required to initiate a high quality, defect-free die singulation.

To understand the effects of multi-strata SD, we experimented with controlled dual-pass SD processes on 1Y monitor wafers. Fig. 5-19 shows experimental evidence of multi-strata SD processing at 2.0 W and 500 mm/s in generating undesired cleavage plane \{111\} defects when inter-strata distances are not optimized and go beyond a certain threshold, \(D_{th,SD_{i-1}}\).

Figs. 5-19(a) and (b) show the measured mean focal plane z-heights \(Z_{SD1}\) and \(Z_{SD2}\) (left axis) and inter-strata distance \(D_{SD1-SD2}\) (right axis) as a function of SD2 defocus position, \(DF_{SD2}\). Inset images in Fig. 5-19 show representative optical micrographs with (\(D_{SD1-SD2} = 40 \mu m\)) and without (\(D_{SD1-SD2} = 27 \mu m\)) inter-strata cleavage plane \{111\} defects. Here, it can be seen that, for a given SD processing condition, as the inter-strata distance increases, there is a higher likelihood for the cleavage plane \{111\} defects to arise. In the example shown in Fig.
5-19 where $Z_{SD1} \sim 70 \mu m$ while $Z_{SD2}$ varies from 121 $\mu m$ to 155 $\mu m$, we establish $D_{th,SD1-SD2}$ to be $\sim 40 \mu m$, beyond which the cleavage plane $\{111\}$ defect is generated between the SD layers as a horizontal fracture line.

![Cross-sectional optical micrographs of a dual-pass SD process on 1Y monitor wafers for inter-strata distances of (a) 32 $\mu m$, (b) 40 $\mu m$, and (c) 61 $\mu m$. The SD layers are generated using a laser average power of 2.0 W (PLE = 22.2 $\mu J$) at a scanning speed of 500 mm/s, focusing at $Z_{SD1} \sim 70 \mu m$ and varying $Z_{SD2}$ from 32 $\mu m$ to 61 $\mu m$. (d) Cleavage planes $\{110\}$ of (100)-plane Si wafer with scribe lines aligned with the $<110>$ directions; cleavage plane $\{111\}$ intersects surface plane (100) along the $<110>$ direction at an angle of 54.74°.](image)

Figs. 5-20 to 5-22 provide results from various attempts to understand the mechanism behind the formation of the cleavage plane $\{111\}$ defect. Figs. 5-20(a)-(c) show magnified cross-sectional optical micrographs of a dual-pass SD process on 1Y monitor wafers for inter-strata distances of 32 $\mu m$, 40 $\mu m$, and 61 $\mu m$, respectively. Here, the SD layers are generated using a laser average power of 2.0 W (PLE = 22.2 $\mu J$) at a scanning speed of 500 mm/s, focusing at $Z_{SD1} \sim 70 \mu m$ and varying $Z_{SD2}$ from 32 $\mu m$ to 61 $\mu m$ in order to increase the $D_{SD1-SD2}$. It can be seen from the results in Fig. 5-20 that the cleavage plane $\{111\}$ defects are formed when $D_{SD1-SD2} \geq 40 \mu m$. Meanwhile, Fig. 5-20(d) shows cleavage planes $\{110\}$ of a (100)-plane Si wafer with scribe lines aligned with the $<110>$ directions. It can be seen that the cleavage plane $\{111\}$ intersects surface plane (100) along the $<110>$ direction at an angle of 54.74°.
To fully appreciate its microstructure, Figs. 5-21(a) and (b) show angled and cross-sectional SEM micrographs of the generated cleavage plane \{111\} defect in a representative dual-pass SD process that resulted in a $D_{SD1,SD2}$ of $\sim 50 \ \mu m$. The cleavage plane \{111\} defect can be characterized by a horizontal "V"-shape fracture where two $<111>$ fracture directions meet, as shown in Fig. 5-21(b).

To understand the formation of this cleavage plane \{111\} defect, Fig. 5-22 shows a schematic illustrating its formation mechanism. When the inter-strata distance between the SD layers (for multi-pass SD processing) exceeds a certain threshold, the in-built inter-strata tensile stress will result in the nominal cleavage fracture on the \{110\} plane deflecting into the \{111\} planes. The far field stress generated as the distance between the SD layers

---

**Figure 5-21:** (a) Angled and (b) cross-sectional SEM micrographs of a representative dual-pass SD process that generated a cleavage plane \{111\} defect in between the SD layers as a result of an inter-strata distance of $\sim 50 \ \mu m$.

**Figure 5-22:** A schematic illustrating the mechanism of cleavage plane \{111\} defect formation when the inter-strata distance between the SD layers (for multi-pass SD processing) exceeds a certain threshold resulting in the nominal cleavage fracture on the \{110\} plane deflecting into the \{111\} planes.
becomes larger increases the likelihood for plane “slip” or deflection to arise. The “V”-shape formation is not surprising, given that propagation along the $<111>$ direction is difficult to achieve and never gives a flat fracture surface [71, 72]. While these defects (along with the SD layers) can be subsequently removed using backgrinding in the SDBG process (whereas for the SDAG process, these defects will be left behind and thus may cause serious yield and reliability issues), the cleavage plane $\{111\}$ defects are undesired because, once generated, they usually affect the integrity and control of the crack fracture initiated by the cumulative “damage” of the SD layers. This will, in turn, increase the variability of the SD process which is undesirable for low $Z_{SD1}$ values that are typically more effective for ultra-thin die fabrication using the SDBG process.

Fig. 5-23 shows further repeatability and evidence of sub-optimal inter-strata spacing definition through cross-sectional optical micrographs of a full thickness 1Y die for a representative four-pass multi-strata SD process at 2.0 W and scanned at 500 mm/s. The inset of Fig. 5-23 illustrates a magnified view of the boxed region and shows the generation of three cleavage plane $\{111\}$ “V” cracks between each pair of SD layers as a result of an inter-strata distance of $\sim 55 \, \mu m$. 

Figure 5-23: Evidence of sub-optimal inter-strata spacing definition. Cross-sectional optical micrograph of a full thickness 1Y die for a representative four-pass multi-strata SD process at 2.0 W, 500 mm/s. Inset illustrates the generation of three cleavage plane $\{111\}$ “V” cracks between each pair of SD layers as a result of an inter-strata distance of $\sim 55 \, \mu m$. 

116
5.4 Results and discussion: Stealth dicing process optimization

Wafers have the highest value at the die singulation step, and while singulation does not add value to the finished dies, it has a profound impact on die packaging and assembly yield and reliability. In many applications requiring fabrication of ultra-thin dies for 3D die stacking, wafer singulation has become the most challenging process in meeting both yield and costs targets, as it is one of the most important steps to maximize the number of packaged units per wafer. Based on the previous characterization results enabling us to understand the empirical process trends, the inherent non-linear nature and various first-order constraints (e.g., equipment limitation on optical performance, power and speed, defects generation, backside reflectance, and process cycle time) of the SD process and associated wafer technologies have been recognized.

By knowing the various effects of the major modulators concerning internal SD processing, we next proceed to evaluate and optimize die singulation quality from an “external” appearance point of view on a wafer-level. This is important in order to produce a production-worthy SD process. For example, Fig. 5-24 shows that a sub-optimal SD process not only results in an unsuccessful wafer-level singulation despite using up to four SD passes (thus resulting in a high cycle time per wafer), but also, it creates a large number of SD-unique defects internal to the wafer. An optimized SD process on a module-level will subsequently be followed by the development and optimization of the entire process integration consisting of backgrinding and die separation, before moving on to die attach. Along the integration flow, integrated defects and parametric responses such as electrical characteristics, die strength, warpage, and reliability are captured.

5.4.1 Development of production-worthy stealth dicing

The focus of the work reported in this section is to realize an optimized stealth dicing process module on 1Y wafer technology for eventual die stacking, assembly, and packaging into mobile data storage products.
A. Cycle time and SD laser power limitations

It is important to ensure that for every new process developed for a particular objective, be it quality or output, or both, a solid understanding of its cycle time impact is made. While learning curves and economies of scale can eventually reduce the effective cycle time, it is useful (where possible) to ensure that the new cycle time falls somewhere within a reasonable distribution of the existing cycle time of the process it is trying to disrupt, while maintaining the goodness it serves to offer. This ensures a relatively smooth transition into the new process technology while offering less disruption to the existing capacity of processes upstream and downstream from the new process insertion.

On this note, based on earlier results, we know that laser power influences microstructural transition with an indirect impact to cycle time. With a larger SD layer “damage” induced by higher SD power, a similar crack fracture initiation can be made with higher laser scanning speeds. This offers a lower cycle time per wafer. Therefore, it is in our favor to maximize PLEs (subjected to no known quality or precision impacts) based on the results shown in Figs. 5-16 to 5-18. Given the existing tool capabilities and laser lifetime impact, the laser average power is ultimately set at 2.0 W. The current SD laser has a maximum output power of $\sim 2.2 - 2.3$ W and a maximum laser scanning speed of $\sim 900$ mm/s based on empirical studies. The correct balance and buffering between laser average power, scanning speed, and number of SD passes determine the cycle time of the SD process and its quality.

B. Surface ablation, interference defects, and beam optical parameters

Figure 5-24: Evidence of a severely suboptimal SD process resulting in major SD defect modes.
Based on the characterization studies performed, it can be shown that the SD1 layer not only sets the starting point for the crack fractures propagation to singulate the wafer but also serves as the last “damaged” layer (first in, last out) to be removed by backgrinding. Therefore it is important to position the SD1 layer optimally so that efficient fracturing occurs and no damaged layer is left behind. At the same time, correct $Z_{SD1}$ positioning (i.e., larger than $\sim 50 \, \mu m$ for a given laser effective dose of 2.0W at 500 mm/s) also helps to prevent FS surface ablation and Si interference defects [see Fig. 5-25] This positioning will ultimately determine the beams optical parameter settings in order to minimize the spherical aberration of the condensed beam. From the results of Figs. 5-8 to 5-10 and Tables 5.1 and 5.2, we set $Z_{SD1} \sim 70 \, \mu m$ at a BP = 15 to create a larger solution space for the subsequent optimization of scanning speed to obtain 100% wafer-level SD kerf formation. Eventual baseline studies of the process-of-record (POR) recipes developed show three-sigma variations of $\sim 10 \, \mu m$ for run-to-run $Z_{SD}$ variations of mean and $\sim 4 \, \mu m$ for run-to-run $T_{SD}$ variations of mean. The 20 $\mu m$ buffer given by $Z_{SD1} \sim 70 \, \mu m$ will produce a sustainable process that is robust to natural process and 1Y wafer variations.

C. Cleavage plane \{111\} defects

Figure 5-25: Optical micrographs of SD-unique defects as a result of sub-optimal $Z_{SD1}$ positioning. These include (a) frontside ablation defects, (b) Si interference defects, and (c) microcracks due to increased sensitivity to wafer handling. Insets show respective magnified images of the boxed regions.
Results from Figs. 5-19 to 5-24 help to establish a maximum inter-strata distance of ~32 $\mu$m - 40 $\mu$m, beyond which cleavage plane $\{111\}$ defects will arise. For this reason, inter-strata distances were set at ~25 $\mu$m for subsequent wafer-level SD uniformity optimization. This distance corresponds to a laser defocus (DF) stepping of 10 $\mu$m in bulk Si for each SD layer focal positioning. Experiments show that there is minimal benefit to SD kerf quality and the uniformity of wafer-level SD singulation by reducing this inter-strata distance further. While its increase can potentially result in a larger effective tensile stress to initiate a frontal crack fracture with minimal number of SD passes (and thus reduce the overall cycle time), this direction is unfortunately limited by a stress-relieving mechanism via generated cleavage plane $\{111\}$ cracks.

D. Kerf width, loss, straightness, and effects of test element structures (TEG)

Kerf geometry includes kerf width, loss, perpendicularity, and straightness. To enable high yielding assembly and packaging, it is important to control the variations of kerf geometry [see Chapter 2]. In terms of kerf perpendicularity, in the case of SD processing, this is less concerning because of the proximity of the SD layer (which is usually in the tens of microns) to the frontside of the wafer, which assures a near-zero depth-wise tilt of the “cut”. Because SD singulation is essentially performed via the generation of a crack fracture to the frontside of the wafer, the kerf loss is practically zero, while the kerf width is measured to be ~1-3 $\mu$m. The small kerf loss allows further reduction of the dicing street width to reduce costs per unit die.

Figure 5-26: Top view optical micrographs of frontside surface of adjacent dies on a full 300 mm 1Y wafer with SD kerfs initiated along the dicing streets. (a)-(c) show various SD kerf geometric defects (straightness issues, particles, sensitivity to TEG structures, and etc.) while (d) shows an optimized SD process with a generated straight kerf line that is robust even to complex TEG structures.
Fig. 5-26 shows top view optical micrographs of the frontside surface of adjacent memory dies on a full 300 mm 1Y wafer, with SD kerfs initiated along the dicing streets with a sub-optimal SD process. Specifically, Figs. 5-26(a)-(c) show various SD kerf geometric defects (straightness issues, particles, sensitivity to TEG structures, etc.), while Fig. 5-26(d) shows an optimized SD process (using the developed POR process, which involves three SD layers operating at 2.0 W at 500 mm/s; this will be described in Section 5.4.2) with a generated straight kerf line that is robust even to complex TEG structures. For the most part, kerf geometry integrity is dictated by optimizing the SD layers dimensions, positions, and microstructure type against the characteristics of the multiple layered stacks found in the TEG structures.

E. Coverage of high-quality SD kerf production across wafer

This section of stealth dicing process optimization emphasizes die singulation quality from a wafer-level “external” appearance point of view, with a route towards creating a production-worthy SD process. A good understanding of the process uniformity, variability, and reproducibility is critical to avoid potential process robustness issues during high volume ramping. Fig. 5-27 shows a comparison of frontside within wafer (WIW) SD kerf coverage (%) against different SD multi-pass and scanning speed combinations before and after backgrinding down to 25 μm in thickness. In generating these results, the SD processes operate at a laser average power of 2.0 W and at focal Z-heights of $Z_{SD1} \sim 70 \, \mu m$, $Z_{SD2} \sim 115 \, \mu m$, $Z_{SD3} \sim 158 \, \mu m$, and $Z_{SD4} \sim 201 \, \mu m$, where applicable. These focal positions ensure no FS ablation defects, interference defects, and inter-strata cleave plane \{111\} defects with an adequate PLE without creating any adverse effects. In addition, the beam optical parameter (BP) is set at 15 for reasons described earlier. The estimation of FS WIW kerf coverage is based on optical microscope inspection of the wafers both before and after the backgrinding process (wafer is previously mounted on a Lintec backgrinding tape with the adhesive layer thicknesses measuring \sim 125 \, \mu m). For example, Figs. 5-28(a)-(d) show schematics illustrating frontside SD kerf formation with approximately 20%, 50%, 70%, and 100% WIW coverage.

Some generalizations can be made from the results of Fig. 5-27. The findings suggest the need to increase the number of SD passes in order to progressively increase the SD kerf.
coverage across the wafer up to 100% coverage before the backgrinding process. This can be explained by the fact that with more SD layers stacking on top of each other, the effective tensile stress is higher and will assist the crack fracture propagation toward the frontside of the wafer previously generated by the interaction between the thermal shockwave and the preceding “damaged” SD layer. This will result in a more uniform and consistent coverage of the SD across the entire wafer before backgrinding. Failing to achieve 100% coverage, it can be seen from Fig. 5-27 that backgrinding helps to provide a static load to “complete” the 100% kerf coverage on the wafer frontside. While one may suggest that SD cycle time can be improved if only the response for full SD kerf coverage across the wafer post backgrinding matters, the results from Fig. 5-29 suggests otherwise.

In the case of 1Y memory monitor wafers where there are non-uniformly distributed TEG structures along the dicing street, the failure to realize a 100% SD kerf coverage across the

*Figure 5-27:* Comparison of frontside within wafer (WIW) SD kerf coverage against different SD multi-pass and scanning speed combinations before and after backgrinding down to 25 μm in thickness. All SD processes represented here are performed at 2.0 W at \( Z_{SD1} \sim 70 \, \mu m \), \( Z_{SD2} \sim 115 \, \mu m \), \( Z_{SD3} \sim 158 \, \mu m \), \( Z_{SD4} \sim 201 \, \mu m \), depending on the number of SD layers with \( BP = 15 \).

*Figure 5-28:* Schematics illustrating frontside SD kerf formation with approximately (a) 20%, (b) 50%, (c) 70% and (d) 100% within wafer (WIW) coverage. SD laser is incident on the backside.
wafer before backgrinding will result in kerf straightness and integrity [see also Fig. 5-26] issues post backgrinding. This is a result of inadequate force to thrust the cracks propagation normal to the TEG metallization layers resulting in crack deviation. At the same time, the benefits of reduced cycle time (running at 700 mm/s versus 500 mm/s for a three pass SD process) are small when compared to the need for a robust SD process to envelope the performance variability of interacting tools, such as backgrinding and wafer mounting, for high volume manufacturing. Fig. 5-29 shows the comparison of frontside SD kerf straightness and integrity against different SD multi-pass and scanning speed combinations similar to the one shown in Fig. 5-27. These results clearly demonstrate that kerf geometric quality is compromised whenever SD kerf formation coverage before backgrinding is incomplete. Also, it is much harder to obtain good kerf geometric quality on TEG structures as compared to bare Si, due to the multiple different materials involved to form the TEG structures. From the results shown in Figs. 5-27 and 5-29, the use of three SD layers at a scan speed of 500 mm/s at 2.0 W with the positioning given above forms the core process-of-record (POR) stealth dicing recipe to realize 25 μm ultra-thin die for subsequent stacking, having struck the right balance between achieving a low SD cycle time, i.e., ~ 3.75 mins per wafer, including alignment time and high die kerf quality. This SD cycle time is considerably smaller than mechanical blade dicing through full thickness wafers.

Figure 5-29: Comparison of frontside SD kerf straightness and integrity against different SD multi-pass and scanning speed combinations. All SD processes represented here are performed at 2.0 W at $Z_{SD1} \sim 70 \mu m$, $Z_{SD2} \sim 115 \mu m$, $Z_{SD3} \sim 158 \mu m$, $Z_{SD4} \sim 201 \mu m$, where applicable with BP = 15.
F. Topside, backside, and edge chipping and microcracks

As expected from SD process technology, the permeable (to Si) SD laser offers a practical, non-contact way that helps to eliminate topside, backside, and edge chipping, provided that no surface defects are generated from a sub-optimal SD process. Evidence of a zero chipping process can be seen from the results of Fig. 5-30, to be described in the next section. As for microcracks, the best test vehicle for this is to exacerbate these defects by downstream processing such as backgrinding, die separation, die pick-up, and die attach, followed by visual inspection, electrical characterization, and reliability tests. Meanwhile, as far as die chipping goes, the solution space using SD processing is fairly large as compared to other quality indicators such as kerf geometric integrity and internal SD damage variation unable to be removed by backgrinding.

5.4.2 Process-of-record (POR) stealth dicing to realize 25 μm ultra-thin die

Starting from the appreciation of the SD process physics coupled with simulation work reported in Chapter 4 and reinforced by the subsequent process characterization and optimization work described earlier, a POR SD process can be developed to meet the needs of a 300 mm wafer-level production worthy die singulation process to enable ultra-thin stacked memory die assembly. The empirical demonstrations of good WIW kerf generation, zero defects, and minimal kerf loss with a process cycle time that is significantly lower than that of a mechanical blade dicing process in the dicing-after-grinding (DAG) or dicing-before-grinding (DBG) schemes will complement further exploration of the manufacturability of system-in-package (SIP) and system-on-package (SOP). Fig. 5-30 shows cross-sectional optical microscope images of the developed three-pass POR base process for stealth dicing of 1Y memory monitor wafers before backgrinding down to 25 μm thick. The inset of Fig. 5-30 shows a magnified optical image of the boxed region demonstrating well controlled definition of the three SD layers, SD1-SD3, with no undesired defects.

At the same time, Fig. 5-31 shows top view optical micrographs of the frontside surface of adjacent dies on a full 300 mm 1Y wafer with well defined, high quality SD kerfs (identified
Figure 5-30: Cross-sectional optical microscope images of the developed three-pass POR base process for stealth dicing of 1Y memory monitor wafers before backgrinding down to 25 μm thick. The inset image shows a magnified optical image of the boxed region demonstrating well controlled definition of the three SD layers, SD1-SD3, with no undesired defects.

Figure 5-31: Top view optical micrographs of frontside surface of adjacent dies on a full 300 mm 1Y wafer with well defined, high quality SD kerfs (identified by the arrows) initiated along the dicing streets regardless of the presence of complex TEG structures. There are no signs of kerf geometric defects such as kerf width, kerf loss, kerf perpendicularity, and kerf straightness issues. The POR three-pass SD process is used here.

by the arrows) initiated along the dicing streets regardless of the presence of complex TEG structures. As can be seen from all the images shown in Fig. 5-31, there are no signs of kerf geometric defects such as kerf width, kerf loss, kerf perpendicularity, and kerf straightness issues when using the developed POR three-pass SD process. Kerf width measures about 2 μm wide on average with near zero kerf loss observed as expected.
5.5 Results and discussion: Stealth dicing process integration development

Building off the developed POR stealth dicing process, this section reports on the development of a novel p-SDBG process integration scheme to encompass high backside reflectance Si wafers and improve cycle time. The developed process integration scheme includes solutions to complex integrated defects involving downstream and upstream modules.

5.5.1 Stealth Dicing Before Grinding (SDBG) integration: Challenges

With the developed POR SD process, it is critically important to define, characterize, and optimize its process integration to enable successful high volume manufacturing (HVM). For stealth dicing before grinding (SDBG) process integration, the critical integration loop was previously shown in Fig. 3-1 and discussed in Chapter 3; it starts from wafer frontside tapping (with backgrinding tape) and ends after die separation, producing zero-defects thinned, singulated dies on a wafer frame, ready for the die attach process. Beyond this short integration loop, the remaining assembly and packaging process ought to be straight-forward, except for some minor optimization tweaks if certain new materials or process conditions are introduced within the SDBG integration loop that interact with downstream modules.

In the case of 1Y memory wafer technologies, which have a very high backside reflectance, i.e., larger than 82%, developing an SD process that is able to produce a crack fracture across the entire thickness of the wafer before backgrinding, if at all possible, significantly increases the SD process cycle time. Previous results such as the ones seen in Fig. 5-23 and those from others [75] suggest that despite using up to five SD passes per dicing street on 1Y wafers, a full SD kerf across the wafer thickness is not achieved. One has no choice but to move to an SDAG integration flow, similar to those reported in Refs. [11, 12, 15]. However, SDAG integration has severe limitations in fabricating ultra-thin dies because the SD layers consisting of high density dislocations are left behind without getting backgrinded off; this would compromise reliability tests especially for high performing memory products.
A circumvent to this issue is to have a pre-grinding step preceding the SDBG flow [see Fig. 5-32], which emphasizes back-grinding out the 350 to 1100 nm thick thin film stacks on the wafer backside that are mainly responsible for the SD laser back reflection. But such a pre-grinding step has severe limitations in terms of material removal control and process variations, even if one aims to backgrind up to 50-120 μm off in the z-direction, because of the need for good control in blade dressing stability and also for feedback current control on the backgrinding tool. In addition, if one were to backgrind off an amount larger than 120-150 μm to circumvent the issues described, this will unfortunately compromise the “deep trace” function of the SD tool, decapitating its ability to perform in-situ mapping of the wafer backside surface to compensate for wafer warpage while performing the SD scans. This will significantly increase the SD process cycle time. And finally, adding an extra processing step, i.e., pre-grinding, only serves to increase integration complexity, costs of ownership, higher likelihood for defects entry points, and increased process variations. There are no significant process time savings for the downstream bulk backgrinding process when using an earlier pre-grinding step because the dominant backgrinding step relies on the Z1 blade for coarse backgrinding, which will still require the subsequent Z2 and Z3 spindle steps for fine backgrinding and dry polishing (as described in Chapter 3).

Figure 5-32: Process schematics representing the stealth dicing before grinding (SDBG) integration flow with an additional wafer backside pre-grinding step inserted right before SD processing.
Therefore, combined with solutions to integrated defects involving downstream modules, the work reported in this dissertation contributes the development of a novel p-SDBG integration scheme to encompass high backside reflectance Si wafers and improve cycle time. The p-SDBG integration flow enables the demonstration of a zero-defect eight die (8D) stack of 25 μm thick non-functional memory dies and 46 μm thick functional memory dies (64GB) for new product intercepts, as described in Section 5.5.4.

5.5.2 A new integration approach: Partial stealth dicing before grinding (p-SDBG)

This new p-SDBG integration scheme, similar to the schematic flow shown previously in Fig. 3-1, is based upon the use of stealth dicing to initiate controlled crack fracture toward the frontside of the wafer only, and not both the backside and frontside. It then relies on the subsequent static loading from backgrinding to “finish the job” of full kerf separation of individual dies – improving the cycle time and solution space of SD processing. This is because the magnitude of “damage” required for die separation from SD processing (as a result of the frontal crack fracture initiation) becomes less, and thus saves processing time. This is especially critical for high backside reflectance wafers, such as 1Y wafers, where the amount of SD energy coupling into the wafer is limited (unless an extra pre-grinding step to remove the reflecting thin films is used). As a consequence of the non-linearity of SD-induced damage in response to effective PLE dosing [see Figs. 5-12 and 5-13], the p-SDBG is a meaningful step toward developing a production-worthy integration scheme for high backside reflectance wafers that would otherwise take a non-realistic number of SD passes to create (if it creates) SD kerfs across the entire thickness of the wafer. The developed POR SD process to enable p-SDBG was reported in the earlier sections.

5.5.3 Development of p-SDBG integration: Stealth dicing related integrated defects

First, it is important to understand that there are certain physical output differences between the dies singulated using the SDBG process versus the DBG process. Besides the significant
defects and cycle time reduction offered by stealth dicing via the p-SDBG process, there are key physical differences at a singulated die-level between both types of integration flows, given that the DBG process involves mechanical blade dicing from the frontside to create the “half-cut” followed by wafer thinning and DAF laser grooving to separate out the defined dies from the DAF attached to the thinned wafers. While these differences are fertile areas for DfM opportunities, they can also result in integrated issues (defects and/or parametric) as a result of subsequent process interactions. Table 5.3 lists the main physical differences between thin dies fabricated from a p-SDBG process versus a nominal DBG process. The main factors driving the differences stem from the near-zero kerf loss and the use of DAF mechanical separation via the die separation process in the p-SDBG integration flow. The former effect, coming from the SD process, will result in a nominally larger die size (~ 15-20 μm) than those generated by the DBG process, making it important to revisit design rules such as die edge to package edge accepted tolerances. As for the latter effect, the use of the die separation method to perform the DAF cut on the SD singulated dies will result in a near-flush die-DAF sidewall profile, as compared to the DBG method where the DAF actually protrudes ~ 5-10 μm beyond the die sidewall. The protrusion is due to the differences in the amount of DAF material removed using the DAF laser grooving method versus the amount of Si removed using mechanical blade dicing. Additionally, the major difference in kerf width, i.e., ~ 1-3 μm for p-SDBG versus 40-45 μm for DBG, will introduce certain integrated defects (for a description, see sub-section A below) due to the interaction between backgrinding and the SD processed wafers, and will require a material solution involving an upstream process. The extremely thin kerf generated from the SD process in the p-SDBG flow will also render new DAF-related defects (for a description, see sub-section B below) when using the default process die separation recipe nominally used in the DBG flow, as shown in Fig. 2-7(b).

The following is a discussion of different integrated defects unique to the p-SDBG integration flow that are generated and/or exacerbated as a result of subsequent processing steps, interacting with any form or part of the materials or processing associated with stealth dicing. These integrated defects tend to show up a process step or more away from the actual stealth dicing process, and thus involves the need for a better understanding of the overall
Physical parameter | p-SDBG | DBG
--- | --- | ---
Kerf loss | Near zero | ~ 35 – 40 μm
Kerf width | ~ 1 – 3 μm | ~ 40 - 45 μm
Die-DAF side profile | Near flush Die-DAF sidewalls | DAF protrudes ~ 5-10 μm beyond die sidewall
Die size | p-SDBG dies are nominally larger than DBG dies by ~ 15 – 20 μm each side

Table 5.3: Experimentally observed physical differences between thin dies fabricated from a p-SDBG process versus a nominal DBG process.

system-level integration. With a system-level understanding of the integration flow, one can then troubleshoot it more effectively by modulating different process knobs across modules.

A. Silicon residue and particulates post wafer backgrinding

Figs. 5-33(a)-(c) show top view optical micrographs of the frontside surface of adjacent dies at the center, mid-radii, and edge of a 25 μm thick 300 mm SD-processed 1Y wafer post backgrinding (BG), respectively. The defined SD kerfs before backgrinding are identified by the arrows. Fig. 5-33(d) shows the approximate locations for the optical inspections performed.

As can be seen from Fig. 5-33, there is evidence of heavy BG stains and Si residue at the center of the thinned wafer primarily on the dicing streets of the wafer post backgrinding. The magnitude and spread of these defects appears to be lower at the edge of the wafer and almost non-existent at the mid-radii of the wafer. The location dependency of these defects is closely related to the backgrinding recipe, specifically the rotation direction of the backgrinding wheel. Limited experiments have shown that for a wheel rotation moving from the edge to the center (with reference made to the backgrinding table), the residues seen in Fig. 5-33 are center heavy, while center to edge rotation makes the edge worst. This may be explained by the fact that edge to center rotation direction results in more Si dust generated in combination with grinding water towards the center, while the other direction is true for center to edge rotation. Fig. 5-34 shows schematics illustrating the different wheel rotations during wafer thinning. More importantly, these integrated defects are a direct consequence
of the extremely thin SD kerf ($\sim 1-3 \mu m$) for p-SDBG when compared to the much larger 40-45 $\mu m$ for DBG integration. This thin kerf, when combined with the 5-7 $\mu m$ step height between the surface of the polyimide passivation layer on the wafer frontside and the actual surface of the dicing street, will render a higher likelihood for Si dust entrapment. These Si dust particles are created during wafer thinning and they enter through the small kerf into the “undercut” resulting from the step height of the wafer frontside. The water cleaning, which is part of the backgrinding process module, is unable to dislodge and remove these Si dust particles effectively through such a narrow aperture of the thin kerf when compared to the apertures produced by the mechanical dicing blade in the DBG process.

Despite quickly adding another wafer cleaning step (queue time control of less than three
hours after backgrinding) using the die separation module right before the die separation process, the residues can only be partially removed, as can be seen from the results shown in Fig. 5-35. Fig. 5-35 shows top view optical micrographs of the frontside surface of adjacent dies of a 25 μm thick 300 mm SD-processed 1Y wafer post BG, where Figs. 5-35(a), (c), and (e) show the respective center, mid-radii, and edge located images before wafer cleans, while (b), (d), and (f) show the corresponding images after wafer cleans using the integrated die separation module. In the die separation module, the cleaning step has been optimized to fully take advantage of a single-wafer cleaning mechanism, typical in state-of-the-art Si wafer cleans processing, involving a high pressure nozzle and controlled radial sweeps across the wafer while having the wafer rotating at high speeds. Notwithstanding many attempts to optimize this cleaning process which compromise the overall cycle time of the die separation module significantly, the residues remain at large and can gravely cause die cracking during die stacking. Fortunately, one solution to this particular integrated defect lies with the adhesive thickness choice for the backgrinding tape and will be discussed in the next section.

### B. Die Attach Film (DAF)-related defects

The extremely thin kerfs generated from the SD process in the p-SDBG flow, compared to the larger kerf widths from mechanical blade dicing in the DBG flow, will render the
need for a larger tensile force generated during die separation in order to create an adequate space (larger than ~0.2 mm) between the edge of the die and the edge of the cut DAF. This die-to-DAF edge separation also needs to be uniform across the framed wafer for successful die pick up during the die attach process.

When using the die separation recipe nominally used in the DBG flow (where in addition to the larger kerf widths to begin with, there is no need for DAF cut using the die separation technique since it was already performed using DAF laser grooving, as seen in Fig. 2-7(b)), the magnitude for table ascension in the cooling expansion stage can be much less. As a consequence, when the same recipe is applied to SD-processed wafers, the recipe becomes marginal, in particular at the wafer edges. Fig. 5-36(a) shows die attach pick up issues as a result of the marginal die separation process for the p-SDBG flow, causing dies to remain on the edge of the wafer mounting frame. In Fig. 5-36(b), the view from the DAF/dicing tape side of the mounted wafer shows the nominal die-to-DAF edge separation distance required for successful die pick up, which in this sub-optimal case applies only to adjacent dies located near the edges of the wafer mounting frame.
away from the edges of the mounted wafer. It is clear from Fig. 5-36(c) that at the edges of the mounted wafer, there is evidence of unsuccessful DAF cut due to inadequate tensile force generated during the ascension of the wafer stage. To resolve this integrated defect, the die separation process has been optimized to employ the process conditions previously described in Chapter 3.

5.5.4 Optimized p-SDBG quality demonstration

Armed with the results from SD process characterization, optimization, and integration based on the p-SDBG flow, significant progress has been made to advance the understanding of SD-assembly interaction and its potential yield and reliability impacts. As a result, an optimal p-SDBG integration flow has been developed to enable HVM-readiness using a suite of newly created POR recipes for SD, backgrinding, and die separation. The POR recipe for SD processing encompasses the fabrication of both 25 µm and 46 µm thick memory dies and has been described in Section 5.4.2 with the SD characteristics shown in Figs. 5-30 and 5-31. As for the POR recipes surrounding wafer backgrinding down to 25 µm and 46 µm final wafer thickness, these have been presented in Section 3.2.2. Finally, the die separation recipe with higher induced tensile forces at the wafer edges, which is needed to circumvent the DAF-related integrated defects, has been presented in Section 3.2.3. A Lintec backgrinding tape with adhesive layers measuring 165 µm thick is used to mount the wafer prior to SD processing to eliminate the residue defects generated by backgrinding. All these process and integration developments lead to the following p-SDBG demonstration of significant defect elimination when compared to the DBG integration scheme. This POR p-SDBG integration flow has been applied to both an eight layer die stack mechanical test vehicle for 25 µm thick memory dies [see Fig. 5-1] as well as an eight layer die stack electrically functional test vehicle for 46 µm thick 2-D NAND memory dies [see Fig. 5-2]. The latter is used to realize a 64 GB Micro-SD memory card for retail production intercept at the time of writing. The results shown below serve to segment out the different stages of quality inspections in order to better contrast (where applicable) the benefits of the developed p-SDBG integration flow as compared to the existing DBG integration flow to enable ultra-thin stacked NAND memory die assembly.
A. Integration post backgrinding

Figs. 5-37(a)-(c) show the respective top view optical micrographs of the frontside surface of adjacent dies at the center, mid-radii, and edge locations of a 25 µm thick 300 mm SD-processed 1Y wafer right after backgrinding (BG). The approximate locations for the optical inspections are shown in Fig. 5-37(d). These thinned wafers use the Lintec backgrinding tape with the adhesive layers measuring 165 µm thick. It can be seen from the results shown in Fig. 5-37 that the SD kerfs (identified by the arrows) are well-defined and there is no clear evidence of defects (direct or integrated) post backgrinding. It is evident that the heavy BG stains and Si residue seen previously [see Figs. 5-33 and 5-35] have been eliminated via a BG tape material change. This change is much more effective than inserting a wafer cleaning step right before die separation [see Fig. 5-35].

![Figure 5-37: Top view optical micrographs of defect-free frontside surface of adjacent dies at (a) center, (b) mid-radii, and (c) edge of a 25 µm thick 300 mm SD-processed 1Y wafer post backgrinding (BG) using the Lintec backgrinding tape with adhesive layers measuring 165 µm thick. (d) Approximate locations for the optical inspections. Defined SD kerfs before backgrinding are identified by the arrows. There is no evidence of heavy BG stains and Si residue post BG tape material change.](image)

When contrasting this with the results shown in Fig. 5-38, it can be seen that the current DBG flow, up to the point of post-backgrinding, despite being optimized, has an abundance of physical defects such as top side chipping (TSC) and microcracks (direct or indirectly induced). This is not surprising given the fact that TSC is a direct result of the DBG
Figure 5-38: Representative top view optical micrographs of defect-ridden frontside surface of adjacent dies singulated and thinned based on the DBG integration flow. (a) Evidence of top side chipping (TSC) due to DBG half-cut using mechanical dicing saw. (b)-(f) Various representative images post wafer backgrinding down to 25 μm thick where both TSC and microcracks can be seen.

half-cut produced using mechanical dicing. Additionally, microcracks can be seen from Figs. 5-38(e) and (f), which are generated from weak spots surrounding die chipping sites due to the static loading caused by backgrinding and potential handling. Ultra-thin wafers are more susceptible to such defects and could cause serious assembly yield issues.

B. Integration post die separation

Figs. 5-39(a)-(c) show the respective top view optical micrographs of defect-free frontside surfaces of adjacent dies at center, mid-radii, and edge of a 25 μm thick 300 mm SD-processed 1Y wafer post optimized die separation (DDS). Meanwhile, Fig. 5-39(d) shows the view from the wafer backside, with a magnified view of the boxed region showing the die-to-DAF edge separation distance, which is critical for successful die pick-up. It can be seen that in addition to the defect-free and high SD kerf quality results from Fig. 5-39, post-DDS, the kerf width enlarges to 49 μm and 62 μm for channel 1 and 2, respectively, from an initial recorded kerf width of ~ 1-3 μm right after SD processing. This is to be expected as a result of the die separation process, which induces tensile force to widen the die separation distance via a cold expansion process, as previously described in Section 3.2.3.
Figure 5-39: Top view optical micrographs of defect-free frontside surface of adjacent dies at (a) center, (b) mid-radii, and (c) edge, of a 25 μm thick 300 mm SD-processed 1Y wafer post optimized die separation (DDS). (d) View from the wafer backside with a magnified view of the boxed region showing the die-to-DAF edge separation distance, which is critical for successful die pick-up. Post-DDS, the kerf width enlarges to 49 μm and 62 μm for channel 1 and 2, respectively.

Also, using the developed POR recipe for the die separation process, the die-to-DAF edge separation distance ranges from 1.0 to 3.0 mm across the wafer, comfortably exceeding the 0.2 mm requirement to allow a successful die pick up process, i.e., peeling off the thinned, singulated die with DAF attached on its backside from the dicing tape. An increase in the separation distance between the separated die edge and the DAF edge leads to a reduction in the DAF-to-dicing tape effective contact area, and thus reduces the adhesion at the DAF-to-dicing tape interface. This is important to prevent an excessive die pick up force impinging on the thinned die in order to “peel” it off for subsequent die attach. On the other hand, having too large a die-to-DAF edge separation distance will also lead to the possibility of dies “flying” off if excessive handling or rotational based processes are used. For smaller die size, the solution space for an optimal die separation process is much smaller because of the reduced contact surface area between the DAF and the dicing tape, which is essential to provide better tensile force coupling.

Meanwhile, Fig. 5-40 shows representative top view optical micrographs of defect-ridden frontside surfaces of adjacent dies singulated and thinned based on the DBG integration flow post DAF laser grooving. Specifically, Figs. 5-40(a) and (b) show evidence of the protruding
DAF layer with respect to the die sidewall, while (c) and (d) show evidence of DAF residue and DAF strand defects, respectively, as a result of this non-flush DAF-die sidewall profile, unique to the DBG integration flow. Fig. 5-41 shows a schematic illustrating typical DBG integration defect modes incoming to the die attach process that are eliminated by the p-SDBG integration flow. The primary drivers for defect reduction are the non-contact mode, the flush die-to-DAF profile, and the elimination of DAF laser grooving process step as part of the p-SDBG integration flow.

C. Integration post multi-die stacking

Figs. 5-42 and 5-43 show the respective top view and angled top view SEM micrographs of a defect-free eight layer memory die stack with single-sided bonding pads progressively
staggered (staircase structure) in the form of two four-die blocks. Fig. 5-42 shows the assembled ultra-thin memory die stack before (left) and after (right) wirebonding. This stack architecture is used as a mechanical p-SDBG test vehicle for 25 μm thick memory dies, and the integration flow stops at this point. The inset of Fig. 5-43, which shows a magnified SEM image of the boxed region on the die side walls, illustrates the integrity of the die sidewalls and the flush profile across the 25 μm thick die to the 10 μm thick DAF, both of which are characteristics enabled by an optimal SD process and p-SDBG integration flow.

As for the electrical test vehicle, Fig. 5-44 shows angled top view SEM micrographs of a defect-free eight layer memory die stack with single-sided bonding pads staggered one-way progressively (staircase structure). This architecture is used as an electrically functional p-SDBG test vehicle for 46 μm thick memory dies to ultimately realize a 64 GB Micro-SD memory card. The inset image of Fig. 5-44 shows a magnified SEM image of the boxed region on the die side walls to better illustrate the integrity of the sidewalls and the flush profile across the 46 μm thick die to the 10 μm thick DAF, again enabled by an optimal SD process and p-SDBG integration flow. This electrical test vehicle will continue to proceed downstream along all the assembly and packaging processes, including assembly electrical tests and reliability tests, to ascertain the manufacturability of the developed p-SDBG integration for actual retail data storage products in the form of SD-micro cards.

Figure 5-42: Top view SEM micrographs showing a defect-free eight layer memory die stack with single-sided bonding pads progressively staggered (staircase structure) in the form of two four-die blocks before (left) and after (right) wirebonding. This architecture is used as a mechanical p-SDBG test vehicle for 25 μm thick memory dies and the integration flow stops at this point.
Figure 5-43: Additional angled side-view SEM micrographs for the mechanical p-SDBG test vehicle previously seen in Fig. 5-42. The inset shows a magnified SEM image of the boxed region on the die side walls to better illustrate the integrity of the sidewalls and the flush profile across the 25 μm thick die to the 10 μm thick DAF.

Figure 5-44: Angled top view SEM micrographs showing a defect-free eight layer memory die stack with single-sided bonding pads staggered one-way progressively (staircase structure). This architecture is used as an electrically functional p-SDBG test vehicle for 46 μm thick memory dies to realize a 64 GB Micro-SD memory card. The inset shows a magnified SEM image of the boxed region on the die side walls to better illustrate the integrity of the sidewalls and the flush profile across the 46 μm thick die to the 10 μm thick DAF.
D. Die strength characterization: Three-point bend test

As described in Chapter 3, the three-point bending method is applied to measure the SD-singulated thinned die (46 μm thick) breakage strength (for both the circuit-side, i.e., frontside, and the backside of the die) and flexibility, and compare them to the existing baseline DBG process [see Fig. 5-46]. Fig. 5-45(a) illustrates the three-point bend test set-up configuration, while Fig. 5-45(b) shows the loading mechanism, bending moment profile, and related parameters. The samples measured have dimensions of 6.3 mm width, \( b \), 46 μm thickness, \( h \), and 10.9 mm length, \( l \), with a 10 μm DAF attached to the Si backside. The bending moment profile shown in Fig. 5-45(b) reflects this type of uniaxial bending test whereby the top surface (loading rollers) is in a compressive stress state and the bottom surface (supporting rollers) is in a tensile stress state, with the maximum bending stress at the center of the samples backside. Using this method, the ultimate failure stress is used as the measure of die strength, while the actual stress from the test is calculated from the beam bending equation 5.1 below. As can be seen from Eq. 5.1, the die strength is a function of the load recorded at the time of fracture, \( W \), the thickness of the die, \( h \), the width of the die, \( b \), and the distance of the lower span, \( l \).

\[
\sigma_{3\text{-point}} = \frac{3W}{2bh^2} \tag{5.1}
\]

The Si-side and circuit-side die strengths for DBG and p-SDBG produced 46 μm thick memory dies are given in Fig. 5-46. From the results shown (the experimental procedure was performed on \( \sim 10 \) sample dies per wafer for up to three different wafers), it is clear that the overall die strengths for both Si-side and circuit-side for the p-SDBG integration flow are higher than their equivalent measurements for the dies produced by DBG. This is more obvious for the circuit-side (p-SDBG has die strengths \( \sim 26\% \) higher on average for the circuit-side versus \( 17\% \) for the Si-side when compared against DBG), and is due to the fact that the DBG flow causes TSC on the circuit side (frontside) of the die [see Fig. 5-38], which will reduce its circuit-side die strength. Also, specifically for the DBG integration flow, it can be shown from the results in Fig. 5-46 that the overall mean of the Si-side die strength is higher than the circuit-side by \( \sim 23\% \) for similar reasons.
As for the Si backside, both DBG and p-SDBG integration flows are free from backside chipping due to a similar final dry polishing step, although there are slight differences in terms of static loading interactions. In the DBG process, the backgrinding impinges on a wafer with relatively large half-cut “kerfs” measuring 40-45 μm on its frontside (from mechanical dicing), and this may result in micro-damage not immediately visible depending on the alignment of the stages and/or the parameter settings, whereas in the p-SDBG process, the SD kerfs are extremely small, i.e., 1-3 μm wide. This difference can potentially explain the slightly lower Si-side die strength for DBG when compared with p-SDBG.

It is important to note that due to the thin dies, the large deflection caused by the high flexibility of the samples as well as the plate structure of the dies (large width as compared to the thickness) can violate certain assumptions for the standard bending stress calculations. For this reason, and also accounting for added variations due to the DAF attachment to the back of the thin dies, relative comparisons to the current DBG baseline are more meaningful than absolute values. The influence of DAF, which has different mechanical properties than Si, most likely cause the larger scatter in the die strength results for the Si-side. Nonetheless, the positive results seen for p-SDBG continued to pave the route towards manufacturability and product qualification.
E. Die warpage characterization: Shadow Moiré measurements

Shadow Moiré measurements were carried out, as described in Chapter 3. Fig. 5-47 shows the comparison of thinned die warpage at room temperature between p-SDBG and DBG produced 46 µm thick memory dies.

From the results in Fig. 5-47, it can be seen that the measured warpage for the p-SDBG process is ~79 µm, which is lower than the DBG process at ~95 µm. This means that there is minimal risk to the subsequent die pick up process because the pick-up head is set at room temperature. The substrate temperature is nominally ~125 °C.

In addition, Fig. 5-48 extends the comparison of thinned die warpage measurements at elevated temperatures (up to reflow temperatures of 260 °C) to further understand if p-SDBG produced thinned dies behave differently as a result of internal thermal damage caused by the SD laser. As evident from the results shown in Fig. 5-48, both warpage curves are relatively matched indicating no major edge damage internal to the Si die that may render a change in mechanical properties.
Figure 5-47: Comparison of thinned die warpage measurements at room temperature between p-SDBG (left) and DBG (right) produced 46 μm thick memory dies.

F. Ultra-thin stacked die assembly yield and reliability

Finally, we investigated the impact of the developed POR p-SDBG integration flow on an electrically functional eight layer 2-D NAND memory die stack [see Fig. 5-2] packages, which are used to realize retail 64 GB micro-SD memory cards using responses obtained from assembly memory test and system test (MTST) yield and POR reliability tests. A total of up to 5K unit packages in six separate lots were fabricated using the p-SDBG integration flow, and when compared to the current baseline MTST yield for similar retail memory products, we obtain an increase of up to a mean of 3.5% MTST yield (post test yield recovery). Based on approximated unit loadings for 2014, by extending the developed POR SD processing
and the p-SDBG integration flow to all micro-SD memory card product stock keeping units (SKUs) consisting of 4-die, 8-die, and 16-die stacks with final die thicknesses lower than 100 m (a configuration deemed low risk in terms of manufacturability using p-SDBG), costs savings between $1.5M to $6.0M can be made per year. When one further extends this technology to all SIP product SKUs, again consisting of 4-die, 8-die, and 16-die stacks with final die thicknesses lower than 100 µm, annual costs savings between $3.0M to $12.5M can be made. If future roadmaps use p-SDBG for all products with final die thicknesses lower than 100 µm, a large annual costs savings of between $12.0M to $99.0M can be realized. MTST yield is a better representation of the benefits realized because it is 100% sampled when compared with assembly yield.

Additionally, all units fabricated using the p-SDBG integration flow above passed 100% POR reliability tests, which consist of various tests such as a minimum of 500 cycles of temperature cycle component level test (TMCL) from -65 °C to 150 °C, 96 hours of highly accelerated stress test (HAST) at 130 °C, 85% RH, and 500 hours of high temperature stress test (HTST) at 150 °C. At the time of writing, the developed POR SD process and the p-SDBG integration flow have been approved to intercept high volume production of retail 64 GB micro-SD memory cards.

5.5.5 Conclusions and perspectives

In summary, this work has contributed the most comprehensive and systematic known experimental study of SD technology in terms of process characterization, optimization, and integration development. It has been conclusively shown that the newly developed POR SD process and the p-SDBG integration flow are indeed able to significantly increase the MTST yield for the eight layer (46 µm thick die) 2-D NAND memory die stack packages, which are used to realize retail 64 GB micro-SD memory cards. These unit packages passed all POR reliability tests. When extending this technology for all SIP packages loaded in 2014 consisting of 4-die, 8-die, and 16-die stacks with final die thicknesses lower than 100 µm, substantial costs savings can be realized based on the representative MTST yield increase. This capability and set of characterization results will benefit the semiconductor assembly and packaging community, enabling them to produce high yielding ultra-thin dies.
for stacked memory die assembly. However, better energy coupling into the Si for high backside reflectance wafers either through the backside or frontside with well thought out DfM practices have to be solved before one can proceed to fabricate defect-free thick dies for other applications (e.g., automotive) because they would require a higher crack fracture frontal force when one increases the $Z_{SD1}$ height using the p-SDBG flow.
6.1 Introduction: Lean Six Sigma operations

Broadly speaking, Lean Six Sigma is a business improvement methodology that spans across the entire business value chain and maximizes stakeholder value by achieving the highest rate of improvement in customer (internal and external) satisfaction, cost, quality, process speed, and invested capital. Examples of typical value segments across the Si technology innovation pipeline include research, pathfinding, development, ramp, technology transfer, and high volume manufacturing. A process can be defined as a repetitive and systematic series of steps or activities where inputs are modified to achieve a value-added output. Lean Six Sigma initiatives eliminate causes of customer (internal and external) critical-to-quality issues, lead-time delays, and sources of variation in any process. Eliminating those causes provides the greatest opportunity for improvement in cost, quality, capital, and lead-time.

In Lean Six Sigma, set differently from its individual constituents is the recognition that
one cannot deliver “just quality” or “just speed,” one needs a balanced process that can help an organization focus on improving quality, as defined by the customer (internal and external), within a set time limit. The focus on process is critical because it helps us to (1) understand how and why work gets done, (2) characterize input and output relationships, (3) manage for maximum customer satisfaction while utilizing minimum resources, (4) appreciate the process from beginning to end as it is currently performed, and (5) establish and validate models based on objective data. The Lean component centers around efficiency and the elimination of waste allowing us to produce what is needed, when it is needed, and with the minimum amount of materials, equipment, labor, and space. As for the Six Sigma component, the focus is on effectiveness and variation reduction identifying, eliminating, and controlling factors that contribute to the most underperforming process and making sure these processes meet and exceed customer expectations every time. Sigma, which is a measure of variation, is determined by identifying the optimum quality level beyond which the costs of quality improvement exceed the expected costs savings from a reduced number of defects.

There are two categories of improvements: improving the existing process or creating a new process altogether. When we have an existing process and we want to improve the process, we usually deploy the DMAIC methodology. When creating a new process or completely revamping the existing process, the IDOV methodology is usually deployed. For stealth dicing process and integration flow development, the IDOV framework is more relevant.

### 6.1.1 Define, Measure, Analyze, Improve, Control (DMAIC)

In the Lean Six Sigma literature, a typical systematic variable reduction framework for existing processes is known as DMAIC, which stands for define, measure, analyze, improve, and control. At its core, the essence of DMAIC is to identify key inputs to a process and through iterative sequences, reduce them to a selected critical few and then relentlessly improve and control them.

Under “Define”, customer needs are stated and the process to be improved is identified. Here, the problem or opportunity, goal, process, and customer requirements are described
clearly. In “Measure”, the process performances are comprehended with the input and output variables of the processes defined, and with measurement systems validated. At this stage, gathering data about the problem and the processes is very important. During the “Analyze” stage, the data collected is used to establish key process inputs that affect the process outputs, and thus allowing the root cause(s) to be identified. The iterative use of hypothesis and validation occurs at this stage to find the root cause(s), clearly documenting and analyzing supporting and refuting evidence. When the “Improve” stage is reached, the improvements have been identified to optimize the outputs and eliminate or reduce the defects and variation. Here, the iterative sequence of development, implementation, and refinement is executed. Finally, when it comes to “Control”, the documentation, monitoring, and the assignment of accountability to sustain the gains made by the process improvement(s) are implemented based on a plan for stability and measurement.

### 6.1.2 Identify, Design, Optimize, Validate (IDOV)

To create new processes or to completely revamp certain processes involving higher levels of ambiguity, the IDOV six sigma framework, which stands for identify, develop, optimize, and validate, is used. The IDOV framework can be rather different than the DMAIC framework.

The identification phase involves identifying the voice of the customer (VOC), developing a team and team charter, performing competitive analysis, and identifying critical-to-quality (CTQ) factors. Some of the main Six Sigma tools used in this phase include quality function deployment (QFD), failure modes effects analysis (FMEA), supplier, input, product, output, customer product map (SIPOC), integrated product delivery system (IPDS), target costing, and benchmarking [76].

Based on the CTQs identified, the design phase formulates the concept design, identifies the risk elements, identifies the design parameters, and defines the experimental and manufacturing plans. Some of the tools used in this phase include smart and simple design, risk assessment, FMEA, engineering analysis, materials selection, simulation, and design of experiments (DOEs) [76].

The optimize phase utilizes the CTQs to assess the tolerance level of the process and find ways to maximize performance with minimum resources. This phase may involve assessment
of process capabilities, optimization of design parameters, development of design for robust performance and reliability, error proofing, and establishment of tolerance measurement objectives. Some common tools used in this phase include manufacturing database and flow back tools, DfM, process capability models, robust design, and Monte Carlo methods [76].

The last phase of the IDOV framework, i.e., validation, involves testing the selected design and validating its outcome. Some of the main steps in this phase include prototype testing and validation, assessment of performance, failure modes, reliability and risks, design iteration, and final phase review.

### 6.1.3 Stealth Dicing: Using the IDOV + C framework

We need ultra-thin dies incoming to die attach for multi-die stacking to be free from all defects in order to obtain a high memory test and system test (MTST) unit yield. The wafer singulation process within the die preparation flow is one of the most critical processes causing losses to MTST yield mainly driven by various die chipping and microcracks, directly or indirectly in the highly sensitive dicing before grinding (DBG) integration flow is used for 50 μm or below die thicknesses.

For the development of the stealth dicing process and the p-SDBG integration flow, the main framework used is IDOV. However, in order to bring the process to a production-worthy and sustainable state for effective process transfer, it is important to establish solid process control with the identification of key CTQs. As a result, a modified IDOV + C framework is used instead. Some of the key applications and learnings gained when using this hybrid framework for stealth dicing technology development are described below; detailed results from application of the “control” aspect of this framework to stealth dicing are discussed in Section 6.2.

Early stealth dicing process screening and process physics understanding fell under the “identify” phase. Through the use of SIPOC, the quality needs, existing issues, and critical input/output parameters had to be fully understood and incorporated into the pathfinding process improvement metrics as part of the IDOV framework. In the “identify” phase, direct and integrated defect dependencies were documented across the entire segment of the die preparation flow from wafer mounting until right before die attach. Physical models explain-
ing defects and process cause and effects were hypothesized. New metrology and responses had to be researched due to the immaturity of the stealth dicing process in the memory die assembly industry. At the same time, parts of the process pathfinding and development needed to be managed through external suppliers while waiting for internal experimental tool time. These interactions had to be well understood and defined to minimize and eliminate miscommunication errors. An early FMEA was created based on a thorough identification of all failure modes and effects, with some centered on “gross reality check” quick experimentation. FMEA is a time tested tool used to identify and reduce the unexpected and to take counter measures.

As for the “design” phase, we conducted systematic stealth dicing process characterization and integration flow design. Because of limited resources and tool time in a manufacturing environment, experiments needed to be designed to have the highest return on investment (ROI), i.e., to recognize the point of diminishing marginal returns associated with additional data collection and then balance the desire for detailed problem understanding versus risk taking from implementing a quicker solution based on less perfect data. The work on stealth dicing was done in addition to other engineering experiments for a variety of other projects, and thus the data was expected to sometimes be confounded. Because of this, we had to be extra diligent in terms of data interpretation (correlation vs. causation), data integrity (attention to details, commercial bias, etc.), and also data communication (language, etc.) from our foreign partners. In addition, to help with the design of experiments, process simulations were performed on key identified parameters and responses to predict process constraints. Despite thorough planning, unexpected issues will surface. These should be embraced as opportunities to improve and to understand the process.

In the “optimize” and “validate” phases, the iterative use of process optimization and validation for both the stealth dicing process and the p-SDBG integration flow was emphasized. Here, because the number of samples available for experiments can be low in a research and development environment, it was important to perform time-paired experiments whenever possible. A strong understanding of metrology and the interpretation of the raw data (and not just the summary data) was critical to understand the output of the optimization. It was also important to be hands-on in order to build better experimental intuition. Additionally,
it is always useful to be resourceful in finding experimental opportunities to tally the results performed in-house versus the ones performed at supplier sites or reported in the literature, especially for critical recipe decisions. The new stealth dicing process and p-SDBG integration flow was then validated by repeat builds and process baselining to ensure that clear, measurable process improvements and defect elimination were achieved.

Finally, under the “control” phase, the statistical process control, sensitivities, interactions, and risks analysis were assessed before potential product intercept. This is to ensure process robustness for high volume manufacturing. Based on all the results obtained as well as the gain of system-level engineering intuition, a stealth dicing quality fishbone diagram and final process FMEA were created.

6.2 Introduction: Statistical process control and robust optimization

Statistical process control (SPC) for production processes in industry is mainly used to monitor variation and to distinguish the usual random variation, i.e., common causes, from abnormal changes, i.e., special causes [77, 78]. The development of these monitor charts are useful to ensure that key characteristics of interest remain as close to specific target values are possible. These statistics are usually from a time series, and the patterns they exhibit over time provide indications to problems, process capability, and/or opportunities for process, equipment, and integration improvements. These charts may or may not have a specifications limit, depending on the responses they monitor, but they always have variation based control limits to ensure that a process is in a state of statistical control. The basic premise in the development of SPC assumes that a process remains in control until acted upon by some outside force, called an assignable cause. Hence, attempts to repeatedly change the process settings to drive the mean closer to the process target, which some people call tinkering, will actually introduce more variability into the process and thus do more harm than good. The use of SPC techniques became popular in the 1980s as industry began to better understand the issues of quality, after the pioneering effort of Japanese industry and
under the leadership of W. Edwards Deming and Joseph Juran. There are two types of SPC charts, namely variable charts and attribute charts. Variable control charts are used when the quality characteristic to be monitored is measured on a continuous scale, whereas attribute control charts are used for qualitative characteristics (ordinal or categorical). Both types of charts are developed for the introduction of stealth dicing technology.

The philosophy of experimentation calls for experiments that proceed sequentially where, ideally, we apply what we learn from previous experiments to plan each subsequent one [77]. Depending on the resources available and the costs-benefits analysis for further optimization, factorial experiments centered at the settings that have given us the best performance so far are conducted. This is usually done after using a series of simple, quick turn-around screening experiments to construct a model in the important factors, i.e., constructing a path of steepest ascent or descent, depending on whether we wish to maximize or minimize the response [77]. Additionally, one needs to realize that an optimized process (at a certain stage) does not always mean to maximize or minimize one characteristic. In fact, we call a process that reaches a target condition for a characteristic of interest with minimal variability a robust process, because it can achieve the target over a wide range of operating conditions. For the latter reason, custom DOEs have been constructed using the developed POR stealth dicing conditions to fully explore the solution space, by performing sensitivity and two-factor interaction analysis. This also helps to build further engineering intuition on the process.

6.3 Results and discussion

This section reports the development and the results of stealth dicing technology operations management statistical process control, sensitivities, interactions, and risks analysis.

6.3.1 Stealth dicing process time-series baselining and statistical process control development

After the initial qualification of the developed SD process and p-SDBG integration flow, we were ready to run a small, controlled pilot in the assembly line in order to start the new
POR process baselining (time-series) and collect random samples to build SPC charts. Over the time period of 11 days (9/1/14-9/11/14), we ran two sets of runs per day, separated by at least four hours. One set of runs consisted of two random 1Y monitor wafers, one representing stealth dicing (SD) module monitors and the other representing p-SDBG integration monitors, with different responses recorded. The former monitors focus on creating variables control charts, where characteristics of the SD layers are quantified by destructive cross-sectional optical microscope imaging and measurement. The latter monitors (in combination with the former monitors where applicable, e.g., observing SD kerf defects post-SD) focus on creating attributes control charts, primarily for defects in terms of number of affected die counts. A total of 22 runs are processed over 11 days bringing the total number of wafers used to 44 pieces, splitting across SD and p-SDBG sets.

A. Variable control charts

The SD process for each run has the characteristics described earlier from Figs. 5-7 (schematic with measurement definitions) and 5-30 (actual cross-sectional images) of Chapter 5. It is essentially a developed three-pass POR base process for stealth dicing of 1Y memory monitor wafers before backgrinding down to 25 μm thickness. The results of the process baselining can be seen from Figs. 6-1 and 6-2 below.

Figs. 6-1 and 6-2 show the X-Bar and Sigma, i.e., STDEV, control charts for SD layer height $T_{SD_i}$ (where $i = 1, 2, \text{or } 3$) and SD layer focal z-height $Z_{SD_i}$ (where $i = 1, 2, \text{or } 3$), respectively. Given the fact that the sub-group size is nine per run, we use the X bar / Sigma charts, since the range statistic is a poor estimator of process sigma for large subgroups. In fact, the within-subgroup sigma is always a better estimate of subgroup variation than subgroup range. In the X-bar charts, we assume that the variance is constant, i.e., that the Sigma chart is in control, which is validated by the right-hand-side charts shown in Figs. 6-1 and 6-2. We also utilize the central limit theorem to assume normality, which is a reasonable assumption given that the raw data is not highly skewed and sample size is fairly large ($N = 21$). Additionally, the process is in control for the base period considered for the most part.
Figure 6-1: X-Bar and Sigma control charts for respective SD layer height, $T-SD_i$ (where $i = 1, 2,$ or $3$), with a sub-group size of nine per run.
Figure 6-2: X-Bar and Sigma control charts for respective SD layer focal z-height, $Z_{SD_i}$ (where $i = 1, 2, 3$), with a sub-group size of nine per run.
From Fig. 6-1, the time-series pattern appears relatively stable, i.e., no systematic effects such as sawtooth or strong trending behavior. The variation is also consistent with engineering intuition and SD equipment performance capabilities. A similar conclusion can be made for Fig. 6-2, although the mean for SD layer focal z-height appears to start slightly lower before trending back up to the center line. This is believed to be caused by a minor conditioning effect of the tool, which impacts the positioning of the focal spot more so than the thermal shock wave affected area. From these results, we recommend the need to run SD conditioning recipes on dummy wafers at least once every 48 hours to ensure a more stable laser behavior.

Table 6.1 summarizes the calculated center line and three-sigma control limits for variable responses SD layer height, T-SDi, and SD layer focal z-height, Z-SDi (where i = 1, 2, or 3). From the results shown in Table 6.1, it can be seen that all three SD layer heights have a well-controlled grand mean of \( \sim 19-20 \) \( \mu \text{m} \) with a run-to-run (RtR) mean variability (one-sigma) of \( \sim 1.3-1.4 \) \( \mu \text{m} \). As for the within wafer (WIW or within subgroup) variation, all three SD layer heights have a grand mean of \( \sim 1.4-2.3 \) \( \mu \text{m} \) with a variability (one-sigma) of \( \sim 0.6-0.8 \) \( \mu \text{m} \).

As for the data shown in Table 6.2, it can be seen that SD layer focal plane Z-height for SD1, SD1, and SD3 layers have respective well-controlled grand means of 69 \( \mu \text{m} \), 115 \( \mu \text{m} \), and 158 \( \mu \text{m} \). Its run-to-run (RtR) mean variability (one-sigma) ranges between 3.2-4.0 \( \mu \text{m} \). As for the within wafer (WIW i.e. within subgroup) variation, all three SD layer heights have a grand mean of \( \sim 1.9-3.4 \) \( \mu \text{m} \) with a variability (one-sigma) of \( \sim 0.6-1.3 \) \( \mu \text{m} \).

The values shown in Tables 6.1 and 6.2 demonstrate the potential for SD technology to enable controlled fabrication of thinned, singulated die measuring 25 \( \mu \text{m} \) and below in thickness in combination with backgrinding technology, because the size and the positioning of the SD “damaged” layers within Si has a very low variation; much lower than that of backgrinding.
<table>
<thead>
<tr>
<th>Response (Units)</th>
<th>Process flow</th>
<th>Experiment type</th>
<th>No. of runs</th>
<th>Samples per run</th>
<th>No. of days</th>
<th>Success criteria and/or control systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height of SD1 layer, TSD1</td>
<td>No pre-grind SDBG flow</td>
<td>SEQ</td>
<td>21</td>
<td>9</td>
<td>11</td>
<td>19.62 +/- 1.35</td>
</tr>
<tr>
<td>Mean (µm)</td>
<td></td>
<td></td>
<td></td>
<td>9/1/14 - 9/11/14</td>
<td></td>
<td>N = 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CL = 19.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UCL = 23.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LCL = 15.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.44 +/- 0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UCL = 3.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height of SD2 layer, TSD2</td>
<td>No pre-grind SDBG flow</td>
<td>SEQ</td>
<td>21</td>
<td>9</td>
<td>11</td>
<td>20.02 +/- 1.30</td>
</tr>
<tr>
<td>Mean (µm)</td>
<td></td>
<td></td>
<td></td>
<td>9/1/14 - 9/11/14</td>
<td></td>
<td>N = 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CL = 20.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UCL = 23.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LCL = 16.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.27 +/- 0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UCL = 4.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height of SD3 layer, TSD3</td>
<td>No pre-grind SDBG flow</td>
<td>SEQ</td>
<td>21</td>
<td>9</td>
<td>11</td>
<td>19.51 +/- 1.43</td>
</tr>
<tr>
<td>Mean (µm)</td>
<td></td>
<td></td>
<td></td>
<td>9/1/14 - 9/11/14</td>
<td></td>
<td>N = 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CL = 19.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UCL = 23.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LCL = 15.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.32 +/- 0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UCL = 4.82</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1: Summary of calculated center line and three-sigma control limits for variable responses SD layer height, T-SD_i (where i = 1, 2, or 3).
<table>
<thead>
<tr>
<th>Response (Units)</th>
<th>Process flow</th>
<th>Experiment type</th>
<th>No. of runs</th>
<th>Samples per run</th>
<th>No. of days</th>
<th>Success criteria and/or control systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal plane Z</td>
<td>No pre-grind</td>
<td>SEQ</td>
<td>21</td>
<td>9</td>
<td>11</td>
<td>69.04 +/- 3.20</td>
</tr>
<tr>
<td>height of SD1</td>
<td>SDBG flow</td>
<td></td>
<td></td>
<td></td>
<td>9/1/14 -</td>
<td>N = 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9/11/14</td>
<td>CL = 69.04</td>
</tr>
<tr>
<td>Mean (µm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LCL = 59.45</td>
</tr>
<tr>
<td>WIW STDEV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.87 +/- 0.64</td>
</tr>
<tr>
<td>(µm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UCL = 3.80</td>
</tr>
<tr>
<td>Focal plane Z</td>
<td>No pre-grind</td>
<td>SEQ</td>
<td>21</td>
<td>9</td>
<td>11</td>
<td>115.33 +/- 3.98</td>
</tr>
<tr>
<td>height of SD2</td>
<td>SDBG flow</td>
<td></td>
<td></td>
<td></td>
<td>9/1/14 -</td>
<td>N = 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9/11/14</td>
<td>CL = 115.33</td>
</tr>
<tr>
<td>Mean (µm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UCL = 127.28</td>
</tr>
<tr>
<td>WIW STDEV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LCL = 103.38</td>
</tr>
<tr>
<td>(µm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.56 +/- 1.18</td>
</tr>
<tr>
<td>Focal plane Z</td>
<td>No pre-grind</td>
<td>SEQ</td>
<td>21</td>
<td>9</td>
<td>11</td>
<td>158.20 +/- 3.73</td>
</tr>
<tr>
<td>height of SD3</td>
<td>SDBG flow</td>
<td></td>
<td></td>
<td></td>
<td>9/1/14 -</td>
<td>N = 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9/11/14</td>
<td>CL = 158.20</td>
</tr>
<tr>
<td>Mean (µm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UCL = 169.39</td>
</tr>
<tr>
<td>WIW STDEV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LCL = 147.01</td>
</tr>
<tr>
<td>(µm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.43 +/- 1.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UCL = 7.42</td>
</tr>
</tbody>
</table>

Table 6.2: Summary of calculated center line and three-sigma control limits for variable response SD layer focal plane Z-height, Z-SD_i (where i = 1, 2, or 3).
B. Attribute control charts

In Chapter 5, in addition to “internal” SD layer characteristics to ascertain SD processing performance, we also evaluated quality responses from an “external” appearance point of view on a wafer level.

Figure 6-3: Attributes control charts for FS SD-kerf post-SD % coverage across wafer and different types of defects in terms of affected die counts.

Fig. 6-3 shows the attribute control charts recording these qualitative (or ordinal) responses to ensure that the control of SD performance translates to increased quality characteristics for the singulated dies. These attributes include FS SD-kerf post-SD % coverage across wafer, FS SD-kerf geometric defects, topside and backside chipping, edge damage,
and residue/stain defects. The responses for defects (direct and integrated) are measured in terms of total affected die counts, sampled at five different locations on each wafer. It is clear from the results in Fig. 6-3 that the developed SD process and the p-SDBG integration flow have shown zero defects recorded so far, significantly better than the current DBG integration flow (especially when compared against frontside defects).

6.3.2 Optimized solution space: Sensitivity and two-factor interaction analysis

The following response surface methodology (RSM) experiments were conducted around the POR SD recipe as a center point, to understand the continuous response sensitivities and two-factor interactions (in addition to the main effects) surrounding the developed SD process. These are performed using a set of treatment combinations with center runs and replicates added to the base DOE design. Using center runs allow a “reality check” while having replicates serve to let us estimate an error term.

A randomized, full $2^3 + 3 + 1$ factorial experiment was performed with three repeats of the current POR conditions for the SD process and one replication of a random treatment combination. This assessment aimed to study the relationship among six selected responses (focal plane Z-height and SD height for SD1, SD2, and SD3 layers) and three selected controllable factors (laser scan/feed speed, laser average power, and laser defocus positions for all SD layers “locked in step” trailing SD1 layers defocus position). Each factor was studied at two levels (coded + and -), and tests were run in random order giving a total of $8 + 3 + 1 = 12$ factor level combinations, as summarized in Table 6.3. The conditions of the experiments follow that of Table 6.3 with other conditions being similar to the POR SD recipe described in Section 5.4.2. The cross-products of the different factors were used to estimate the corresponding two-factor interaction effects. This characterization study, which involved incremental changes to the current POR recipe parameters, allowed processing space and interaction profiles for selected factors to predict process margins and future process improvements for different integration schemes as they emerge.

Figs. 6-4 to 6-6 show the observed responses of the factorial experiments in terms of the
Table 6.3: \(2^3 + 3\) (center points) + 1 (replicate) = 12 factor level combinations with actual factor levels (laser scan/feed speed, laser average power, laser defocus position for SD1 layer).

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Laser scan/feed speed (mm/s)</th>
<th>Laser average power (W)</th>
<th>Laser Defocus Position for SD1 (µm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>+++</td>
<td>600</td>
<td>2.2</td>
<td>-195</td>
</tr>
<tr>
<td>000</td>
<td>500</td>
<td>2</td>
<td>-190</td>
</tr>
<tr>
<td>-+-</td>
<td>400</td>
<td>2.2</td>
<td>-185</td>
</tr>
<tr>
<td>+--</td>
<td>400</td>
<td>1.8</td>
<td>-195</td>
</tr>
<tr>
<td>000</td>
<td>500</td>
<td>2</td>
<td>-190</td>
</tr>
<tr>
<td>++-</td>
<td>400</td>
<td>2.2</td>
<td>-185</td>
</tr>
<tr>
<td>--+</td>
<td>400</td>
<td>1.8</td>
<td>-195</td>
</tr>
<tr>
<td>000</td>
<td>500</td>
<td>2</td>
<td>-190</td>
</tr>
<tr>
<td>++-</td>
<td>400</td>
<td>2.2</td>
<td>-185</td>
</tr>
<tr>
<td>+--</td>
<td>400</td>
<td>1.8</td>
<td>-195</td>
</tr>
</tbody>
</table>

focal plane Z-height for SD1, SD2, and SD3 layers, respectively. Actual by predicted plots, pareto plots of estimates, and two-factor interaction profiles are graphical methods chosen to illustrate the relationship between the selected responses and the controllable factors. It should be pointed out that in Figs. 6-4 to 6-6, the laser DF settings refer to that of SD1-layer, while SD2 and SD3 layer positions follow in lockstep mode, i.e., consistently having inter-strata DF positions separated by a constant 10 µm (in terms of DF settings), which translates to \(\sim 45\) µm in actual “within Si” distance (predominantly due to the index of refraction of Si as described in the earlier Chapters) between each SD layers focal spot.

Fig. 6-4 shows the observed response of the average SD1 layers focal z-height, Z-SD1. Fig. 6-4(a) indicates that the observed response values are highly correlated with the predicted values (RSq of 1.00). The seemingly perfect fit is a result of overfitting due to the use of the three repeats of the POR condition. However, this does not detract us from the general trends observed. Figs. 6-4(b) and (c) show that, as expected, the laser DF position has the greatest effect on Z-SD1. Also, all three two-factor interactions appear to be negligible. The laser DF position, in absolute terms, has a large negative effect on Z-SD1. Figs. 6-5 and 6-6 show comparable behaviors to the results shown in Fig. 6-4, only this time the observed responses are Z-SD2 and Z-SD3, respectively. However, it deserves to be pointed out that
based on the results seen in Figs. 6-5 and 6-6, two-factor interactions involving laser feed speed and laser average power, and between laser feed speed and laser DF position, have a slightly higher effect on both Z-SD2 and Z-SD3. This may be a result of laser power “leakages” and interactions with the SD layers below them. Nonetheless, the effects seen are within 1-2 sigma variation as quantified and summarized in Table 6.2. Because of the relatively higher error margins during the measurements of the SD layer positioning using cross-sectional optical microscopy, strong engineering judgment and experience, as well as understanding of calibration sensitivity, are needed when leveraging these results.

Figs. 6-7 to 6-9 show the observed responses of the factorial experiments in terms of the SD layer height for SD1, SD2, and SD3 layers, respectively. Actual by predicted plots, pareto plots of estimates, and two-factor interaction profiles are again used to depict the relationship between the selected responses and the controllable factors. Similarly, in Figs. 6-7 to 6-9, the laser DF settings refer to that of SD1-layer, while SD2 and SD3 layer positions follow in lockstep mode, i.e., as before, always having inter-strata DF positions separated by a constant 10 μm (in terms of DF settings), which translates to ~ 45 μm in actual “within Si” distance between each SD layers focal spot. Different from the results in Figs. 6-4 to 6-6, Figs. 6-7 to 6-9 show significant two-factor interaction effects on the specified responses.

Fig. 6-7 shows the observed response of SD1 layer height, T-SD1. Fig. 6-7(a) indicates that the observed response values are well correlated with the predicted values (RSq of 0.98), demonstrating that the fitted model is an adequate approximation. Fig. 6-7(b) clearly shows strong two-factor interactions for two out of three two-factor combinations. From the results in Fig. 6-7(b), laser average power appears to have the largest positive effect on T-SD1, while laser feed speed comes in at a close second, albeit as a negative effect. Fig. 6-7(b) also shows strong interaction effects of laser feed speed*laser defocus position and laser average power*laser defocus position. However, the effects from the remaining laser feed speed*laser average power interaction and laser DF position appear to be negligible. These observations, with the given experimental conditions, generally agree with previous development work, in that to increase SD layer “damage” affected areas, higher power and/or lower laser feed speeds are needed, and with a lower absolute DF value, i.e., z-height focal positions closer to the laser source, a higher sensitivity of T-SD1 to feed speed is obtained. But interestingly,
we see that at a lower DF absolute value, a lower sensitivity of T-SD1 to laser average power is observed. It seems that above a certain average power value (≈ 2 W), one actually gets a higher T-SD1, i.e., more damage is recorded for a higher DF absolute value than for a lower one. We suspect that this threshold may be attributed to the beginning of “SD damage layer”-negating interaction effects between “leakage laser energy” and its back reflection from the wafer frontside as power increases.

Fig. 6-8 shows the observed response of the SD layer height for the SD2 layer, T-SD2. Fig. 6-8(a) indicates that the observed response values are poorly correlated with the predicted values (RSq of 0.45). From Fig. 6-8(b), the laser feed speed*laser average power interaction is much stronger than the other two interactions. Also, the results show that the laser feed speed*laser average power interaction has the greatest positive effect, while surprisingly, as compared to the results in Fig. 6-7, laser average power has a significant negative effect. Equally counter-intuitive is the fact that laser feed speed has a far weaker effect and exhibits non-linear characteristics, as seen in Fig. 6-8(c). Other than that, the general trends remain similar, albeit with differing values of T-SD2 sensitivities to laser average power and laser DF position, depending mainly on the values of laser feed speed. We think that the non-linearity seen and the counter-intuitive observations are likely a result of the combination of interactions between the laser with the SD1 layer and/or the wafer frontside, compounded with higher error margins incurred during the measurements. Dimensional measurements are much more difficult for higher stratas, i.e., SD2 layers and above, due to the presence of convoluting vertical microcracks all over. This likely explains the poor correlation shown in Fig. 6-8(a).

Fig. 6-9 shows the observed response of the SD layer height for the SD3 layer, T-SD3. From Fig. 6-9(a), it can be seen that the observed response values and the predicted values have a slightly weaker correlation (RSq = 0.73). From Fig. 6-9(b), the laser average power has the greatest positive effect, similar to the results seen in Fig. 6-7(b). This is followed by the laser DF position, which has a significant negative effect and can be explained by the fact that at a much higher SD layer (SD3 in this case), the exact positioning of the laser focal spot can determine the magnitude of the SD layer damage by virtue of the higher incoming effective laser dose. Fig. 6-9(b) suggests non-negligible two-factor interactions for
laser average power*laser defocus position, while the remaining factors have minimal effects. Similar to the results seen in Fig. 6-8(c), Fig. 6-9(c) shows that the laser feed speed has a far weaker effect and exhibits non-linear characteristics. Other than that, the general trends remain similar, albeit with differing values of T-SD3 sensitivities to laser average power and laser DF position, also depending mainly on the values of laser average power. Again, we think that the non-linearity seen and the counter-intuitive observations are likely a result of the combination of interactions between the laser with the SD2 layer, SD1 layer, and/or the wafer frontside, compounded with higher error margins incurred during the measurements.

Finally, Fig. 6-10 shows the prediction profiler plots for Z-SD_i and T-SD_i for i = 1, 2, and 3. The results shown in Fig. 6-10 clearly demonstrate the negative effects of laser feed speed, mainly on the first SD layer. T-SD2 and T-SD3 values are less sensitive to laser feed speed for reasons explained above. As for the laser average power, it has a significant positive effect on T-SD1 and T-SD3 and a strong negative effect on T-SD2. This observation is not well understood, but is believed to be a combination of laser interactions with SD layers located below, and compounded with the difficulties in measurements due to the omnipresence of microcracks across stratas. As expected, there are no significant effects of laser feed speed and laser average power on Z-SD1, Z-SD2, and Z-SD3, in general. When it comes to the laser DF position, there is a higher positive effect on T-SD3 as compared to T-SD1 and T-SD2, in general. This can be explained by the fact that more “damage” sensitivity to laser DF position is recorded for the layer closest to the incident beam. As for Z-SD1, Z-SD2, and Z-SD3, their significant negative effects from the laser DF positions are well understood as being due to the modulation of their respective laser focal positions.
Figure 6-4: SD1 layer focal plane Z-height, Z-SD1: (a) Actual by Predicted Plot, (b) Pareto Plot of Estimates, and (c) Two-factor Interaction Profiles. Laser defocus settings refer to that of SD1-layer, while SD2 and SD3 layer positions follow in lockstep mode, i.e., inter-strata DF positions are separated by a constant 10 μm, which translates to ~45 μm in actual “within Si” distance between each SD layers focal spot.
Figure 6-5: Similar plots and two-factor interaction profiles as shown in Fig. 6-4 but using SD2 layer focal plane Z-height, Z-SD2, as the observed response. Laser defocus settings refer to that of SD1-layer.
Figure 6-6: Similar plots and two-factor interaction profiles as shown in Fig. 6-4 but using SD3 layer focal plane Z-height, Z-SD3, as the observed response. Laser defocus settings refer to that of SD1-layer.
Figure 6-7: SD1 layer height, T-SD1: (a) Actual by Predicted Plot, (b) Pareto Plot of Estimates, and (c) Two-factor Interaction Profiles. Laser defocus settings refer to that of SD1-layer while SD2 and SD3 layers positions follow in lockstep mode i.e. inter-strata DF positions separated by a constant 10 μm, which translates to ~ 45 μm in actual “within Si” distance between each SD layers focal spot.
Figure 6-8: Similar plots and two-factor interaction profiles as shown in Fig. 6-7 but using SD2 layer height, T-SD2, as the observed response. Laser defocus settings refer to that of SD1-layer.
Figure 6-9: Similar plots and two-factor interaction profiles as shown in Fig. 6-7 but using SD3 layer height, T-SD3, as the observed response. Laser defocus settings refer to that of SD1-layer.
Figure 6-10: Prediction Profiler Plots for Z-SD$_i$ and T-SD$_i$ for $i = 1$, 2, and 3. Laser defocus settings refer to that of SD1-layer while SD2 and SD3 layers positions follow in lockstep mode i.e. inter-strata DF positions separated by a constant 10 m, which translates to $\sim 45 \mu$m in actual “within Si” distance between each SD layers focal spot.
6.3.3 Stealth dicing quality fishbone diagram

Various factors impact the SD quality pre-die attach process. Fig. 6-11 shows a stealth dicing quality fishbone diagram for singulating a 300 mm Si memory wafer mounted on a backgrinding adhesive tape. The factors affecting stealth dicing quality pre-die attach are categorized as wafer, wafer mounting, SD machine, SD process, and integration. The brief summary shown in Fig. 6-11 illustrates the complexity in the direct and indirect dependencies involved in the SD process.

Figure 6-11: Stealth dicing quality fishbone diagram for singulating a 300 mm Si memory wafer mounted on a backgrinding adhesive tape. Factors affecting stealth dicing quality pre-die attach are categorized as wafer, wafer mounting, SD machine, SD process, and integration.
6.3.4 Stealth dicing process Failure Modes Effects Analysis (FMEA)

FMEA is a risk assessment procedure and tool to (1) help identify every possible failure mode of a process (or product), (2) determine the severity of the effect of the failure mode, (3) determine the frequency of the cause of the failure mode, (4) assess the controls (or lack thereof) around the containment of failures, and (5) identify and prioritize actions to eliminate or reduce potential failures [76]. FMEA was originally developed by NASA in the 1960s in the aerospace industry during the Apollo missions and has since spread to many different industries, including the semiconductor industry. One of the key benefits of FMEA is to proactively identify possible risks of a newly designed process or for a new product development in the design phase, so that one can narrow down key process input variables to a critical few for further improvement or further analysis.

Table 6.4 summarizes the FMEA as captured on WW51, 2014 with Memory Test and System Test (MTST) unit yield as the “customer”. The FMEA is performed on the “as-is” developed POR SD process and p-SDBG integration flow and is a result of engineering experience, process mapping, brainstorming with suppliers, specifications, quality fishbone diagrams, etc.

Table 6.5 shows the FMEA rating descriptions for severity (SEV), occurrence (OCC), and detection (DET) scales to help assign one severity rank for each effect. The Risk Priority Number (RPN) is the overall risk of each effect, a product of the SEV, OCC, and DET rankings. Table 10 shows that in terms of priority, in order to facilitate excursion-free p-SDBG integration for manufacturing, proper floor production expectations and error-proof recipe controls need to be implemented as soon as possible. At the same time, a taskforce needs to be set up to eliminate the bubble effect seen along the dicing street using the current 165 μm thick BG tape.
<table>
<thead>
<tr>
<th>Process/Step/Input</th>
<th>Potential Failure Mode</th>
<th>Potential Failure Effects</th>
<th>S/N</th>
<th>Potential Causes</th>
<th>O/C/C</th>
<th>Current Controls</th>
<th>D/E/P</th>
<th>R/P</th>
<th>N</th>
<th>Actions Recommended</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backgrinding receives SD-processed wafer</td>
<td>incoming wafer integrity issues i.e. macro-crack</td>
<td>BG tool errors and/or non-POR BG related defects</td>
<td>8</td>
<td>Wafer handling and transportation</td>
<td>6</td>
<td>Wafer are engineering run</td>
<td>1</td>
<td>48</td>
<td>Implement formal error-proof recipe control systems for the production floor</td>
<td></td>
</tr>
<tr>
<td>SD Process receives wafer</td>
<td>Non-POR incoming wafer</td>
<td>Non-POR SD kerf formation</td>
<td>7</td>
<td>Particles issues on wafer</td>
<td>6</td>
<td>Assembly cleanliness environment: no clear gloving guidelines when handling wafers</td>
<td>1</td>
<td>42</td>
<td>Start with mandatory nitrile gloves donning when handling wafers for SDBG flow</td>
<td></td>
</tr>
<tr>
<td>SD Process receives wafer</td>
<td>Non-POR BG tape mounting</td>
<td>Residual/distortion defects along die-cut line</td>
<td>2</td>
<td>Sub-optimal mounting recipe</td>
<td>9</td>
<td>None at this stage</td>
<td>2</td>
<td>36</td>
<td>Develop optimal mounting recipe to eliminate defect</td>
<td></td>
</tr>
<tr>
<td>Backgrinding receives SD-processed wafer</td>
<td>Non-POR SD processing</td>
<td>Below 100% total kerf formation after thinning</td>
<td>8</td>
<td>SD process excursion</td>
<td>2</td>
<td>Wafer are engineering run</td>
<td>2</td>
<td>32</td>
<td>Implement formal error-proof recipe control systems for the production floor</td>
<td></td>
</tr>
<tr>
<td>SD Process receives wafer</td>
<td>Non-POR BG tape mounting</td>
<td>Residual/distortion defects along die-cut line</td>
<td>2</td>
<td>Sub-optimal mounting recipe</td>
<td>9</td>
<td>None at this stage</td>
<td>2</td>
<td>36</td>
<td>Develop optimal mounting recipe to eliminate defect</td>
<td></td>
</tr>
<tr>
<td>Backgrinding receives SD-processed wafer</td>
<td>Non-POR SD processing</td>
<td>Below 100% total kerf formation after thinning</td>
<td>8</td>
<td>SD process excursion</td>
<td>2</td>
<td>Wafer are engineering run</td>
<td>2</td>
<td>32</td>
<td>Implement formal error-proof recipe control systems for the production floor</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4: SD Process/p-SDBG integration Failure Modes and Effects Analysis (FMEA) as captured on WW51, 2014. Customer here refers to recorded Memory Test and System Test (MTST) unit yield.
<table>
<thead>
<tr>
<th>Severity Scale</th>
<th>Description</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>The effect is not noticeable by the customer</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Very slight effect, noticeable by the customer; however, does not prompt customer to seek service.</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Slight effect that causes customer annoyance; however, does not prompt customer to seek service.</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Slight effect that causes customer annoyance, prompting customer to seek service.</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Minor effect that causes customer inconvenience; however, does not prompt customer to seek service.</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Minor effect that causes customer inconvenience, prompting customer to seek service.</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Moderate effect that causes degradation in the design performance leading to a hard failure or a failure that will eventually cease design functions.</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Significant effect resulting in hard failure; however, does not put the customer's safety at risk and does not result in a significant failure cost.</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Critical customer dissatisfaction effect, cease design functions, significant failure cost and slight safety risk (non-life threatening or permanently disabling) for the customers.</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Hazardous, life threatening, or permanently disabling, or other significant failure cost that places the organization's ability to continue to operate at risk.</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Occurrence Scale</th>
<th>Description</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely remote, highly unlikely</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Remote, unlikely</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Slight chance of occurrence</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Small number of occurrences</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Occasional number of failures expected</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Moderate occurrence</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Frequent occurrence</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>High occurrence</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Very high occurrence</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Certain occurrence</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Detection Scale</th>
<th>Description</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almost certain to detect: Very effective system, mature quality and reliability programs, state of the art process controls. Prevents excursion from happening</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Very high probability of detection: Effective design, implementation of quality and reliability programs, highly automated process controls. Problem detected mid-lot. Lot completed before problem detected.</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>High probability of detection: Full quality program implemented, partial reliability programs, automated process for the majority of operations. Lot completed before problem detected.</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Moderate chance of detection: Quality program in place, awareness of reliability (no formal program), a mix of automated and human intervention process controls. Few lots processed before problem detected. A few lots are produced before problem detected.</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Medium detection: Quality programs developed, but not fully implemented, some automated process controls, dependent on adherence to standard operating procedures. Several lots processed before problem detected.</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Low detection: Initial stages of the quality program implemented, little automated process controls, partial implementation of operating procedures. Problems not detected until lots reach test or FQA.</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Slight detection: Partial implementation of quality methods, sampling inspection plans and random audits, 100 percent inspection. Problems not detected until lots reach test or FQA.</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Very slight detection: Completely dependent on operator self-inspection with periodic quality control inspection, no implementation of quality methods, no formal procedures.</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Remotely possible that detection of a problem will occur: Completely reactive to problems identified during manufacturing, no formal programs, some awareness of product quality.</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Absolute certainty of non-detection of excursion. No detection system implemented. Based on individual discretion.</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.5: FMEA rating descriptions for severity (SEV), occurrence (OCC), and detection (DET) scales to help assign one severity rank for each effect.
6.4 Conclusions and perspectives

In summary, this chapter has presented thorough framework development and experimental studies unique to SD technology operations management. Using the IDOV + C framework, which we think is ideal for the introduction of a previously alpha proven (but not beta proven) technology, the work has been designed accordingly. The work here predominantly reports on the final control stage, which encompasses all the critical aspects important for production-worthy intercept of the developed stealth dicing technology, i.e., process control, sensitivities, interactions, and risks analysis. Based on the results obtained, as well through system-level engineering hands-on experience, a stealth dicing quality fishbone diagram and a final process FMEA have been created.

Through the development of variables and attributes control charts using time-series process baselining and statistical process control methodologies, it has been conclusively shown that the developed POR SD process and the p-SDBG integration flow enable controlled fabrication of thinned, singulated die measuring 25 μm and below in thickness in combination with backgrinding technology. The size and the positioning of the SD “damaged” layers within Si have a very low variation; much lower than that of backgrinding. This facilitates a technology that can last through several memory assembly technology generations and is worthy of calculated investment in building its capacity. Various sensitivity and two-factor interaction analyses have also been performed, with interesting differing characteristics between the first SD layer and subsequent higher-strata SD layers uncovered. Having the stealth dicing quality fishbone diagram and the SD process/p-SDBG integration FMEA, we are able to plan and put into action the priorities needed to facilitate excursion-free p SDBG integration for manufacturing, as well as identify areas for continuous improvement.
Chapter 7

Conclusion and Future Work

7.1 Review

Overall, this dissertation describes original work in the development of subsurface infrared (1.342 μm) nanosecond pulsed laser die singulation technology (stealth dicing) for next-generation 2-D NAND memory products. Specifically, this work reports one of the most comprehensive and systematic experimental studies of stealth dicing technology in terms of process physics, characterization, optimization, integration, and operations management. The primary goal for this work is to enable defect-free, ultra-thin (25 μm and 46 μm) stacked memory die assembly by developing a new process-of-record (POR) stealth dicing process and partial-stealth dicing before grinding (p-SDBG) process integration scheme. This is in order to obtain high cost savings via improved assembly yield, which is critical to bridge high volume manufacturing (HVM) readiness gaps for future memory technologies involving highly defect-sensitive ultra-thin dies. Not only has this work successfully demonstrated defect-free eight die (8D) stacks of 25 μm thick non-functional memory dies and 46 μm thick functional memory dies (64GB) for the first time, but also, this work has been transferred and tested to realize production-worthy retail 64 GB micro-SD memory cards (8D stacks of 46 μm 2-D NAND memory dies staggered one-way progressively) based on the POR product qualification responses obtained from assembly memory test and system test (MTST) yield and reliability tests. We obtain a mean increase of up to 3.5% MTST yield (post test yield recovery). Based on approximated unit loadings for 2014, this translates to annual
cost savings averaging $\sim $12.0M when extending this technology to all internal systems-in-package (SIP) products consisting of 4-die, 8-die, and 16-die stacks with final die thicknesses lower than 100 $\mu$m. If future internal roadmaps use p-SDBG for all products with final die thicknesses lower than 100 $\mu$m, an annual cost savings of up to $99.0M is projected.

Building off the work from others [10–12,14,15], stealth dicing process simulations are developed and performed in order to validate and advance the understanding of its processing physics and manufacturing constraints, at the solution space (less than 30 microns final die thickness) and integration flow of interest. The analysis model and methods are described and refined to provide insights into understanding the temperature rise of silicon due to absorption of permeable nanosecond pulsed laser, and how that can lead to a promising, defect-free singulation technique. This is despite the use of a high backside reflectance (82.3%) value, which is representative of the 1Y wafers used for the bulk of the experiments.

The thermal simulations have demonstrated the effects of laser pulse energies (15.4-22.2 $\mu$J) and focal distance (20-750 $\mu$m from the wafer frontside) on the spatial and temporal heat affected zones (HAZs) within Si, on the frontside Cu plane (0.5 $\mu$m thick), and on the Si wafer backside (laser incident surface). These lead to the conclusion that a suitable focal plane depth and laser irradiation condition is necessary for an optimal SD process, in order to avoid laser-induced ablation damage on the wafer frontside and backside. For example, the observation of a secondary HAZ near the Cu layer, when the $z$ focal distance is set at 35 $\mu$m or less for an average laser power of 15.4 $\mu$J and 18.7 $\mu$J, is observed. Also, if the SD laser beam is focused at 25 $\mu$m (at 22.2 $\mu$J) from the wafer backside, the close proximity will result in the thermal shock wave reaching the backside surface, and as a result cause severe backside ablation. Finally, simulation results show a highly localized confinement of the HAZs within the Si, as well as a radial distribution of HAZs at the frontside Cu layer. This provides confidence in the potential of the SD process to be a strong defect-free die singulation candidate to enable ultra-thin die stacked die assembly, with the potential of
reduced dicing street width in future designs.

As part of the work to develop a production-worthy stealth dicing technology, thorough quantitative and qualitative process characterization and optimization are performed to enhance the understanding of its dicing quality (defects and parametric performance) and solution space. Different aspects of wafer technologies (e.g., backside reflectance ranging from 13.4 to 82.3%, backside surface morphology with surface roughness ranging from 0.4 to 0.8 nm, and \{111\} sidewall crystal orientation) are evaluated against the effectiveness of SD processing. Also, other key responses (e.g., SD-layer dimensional changes, kerf dimensional changes, die warpage, die attach film (DAF) separation performance, die strength, electrical tests, and reliability) are characterized against the effects of laser beam optical parameters (BP), laser scan/feed speed, laser average power or pulse laser energies (PLEs), and multi-strata stealth dicing. One major finding of the characterization work helps to establish a maximum inter-strata distance of $\sim 32 - 40 \mu m$, beyond which cleavage plane \{111\} defects will arise, explained by a proposed stress-relieving mechanism. With the information gathered and analyzed, we then evaluate and optimize die singulation quality from an “external” appearance point of view on a wafer-level. These include the prevention of surface ablation, interference defects, cleavage plane \{111\} defects, and topside, backside, and edge chipping/microcracks. At the same time, we make sure that the SD produced kerf width, kerf loss, and kerf straightness are robust against the effects of test element structures (TEG), and we show a 100% consistent coverage of high-quality SD kerf production across the wafer. Factoring the need for a reduction in cost of ownership (COO) combined with developed solutions to complex integrated defects (e.g., residue/strains post backgrinding and DAF-related defects post die separation) involving downstream and upstream modules, this work contributes the development of a novel p-SDBG process integration scheme to encompass high backside reflectance Si wafers and improve cycle time. Finally, demonstrations of zero-defect 8D stacks of 25 \mu m thick non-functional memory dies and 46 \mu m thick functional memory dies (64GB) are shown for the first time, with clear evidence of up to 3.5% increase in MTST unit yield.

This dissertation also reports on operations management and manufacturability aspects of the developed SD technology statistical process control (within wafer and wafer-to-wafer)
and process robustness, process sensitivities and interactions, and system-level quality/risks analysis. We present an IDOV + C framework and accompanying development and experimental studies unique to SD technology operations management. Through the development of variable and attribute control charts using time-series process baselining and statistical process control methodologies, we have conclusively shown that the developed POR SD process and the p-SDBG integration flow enable controlled fabrication of thinned, singulated die measuring 25 μm and below in thickness in combination with backgrinding technology. The size and the positioning of the SD “damaged” layers within Si have very low variation, which will facilitate a technology that can last through several memory assembly technology generations. The sensitivity and two-factor interaction analyses have also uncovered differing characteristics between the first SD layer and subsequent higher-strata SD layers, with explanations provided. With a new stealth dicing quality fishbone diagram and SD process/p-SDBG integration FMEA, we are also able to plan and put into action the priorities needed to facilitate excursion-free p-SDBG integration for manufacturing, as well as identify new areas for continuous improvement.

7.2 Perspectives

In this section, perspectives on the emerging innovator’s dilemma for semiconductor packaging technologies and the application of software development methodologies to avoid packaging and assembly technologies disruption are given.

7.2.1 Design 2.0 and More than Moore: Innovator’s dilemma for semiconductor packaging technologies

The semiconductor industry, born around 1960, is a collection of companies that design and manufacture semiconductor devices. It is a cyclical sector that thrives during periods of economic prosperity as a result of increased corporate and consumer spending. By the same token, it suffers in economic downturns as consumers shy away from the latest gadgets, and IT budgets shrink in the corporate world. Nonetheless, one never ceases to be amazed
at the exponential conjecture called Moore’s Law (named after Intel co-founder Gordon E. Moore) [1], which states that the number of transistors on integrated circuits doubles approximately every two years. This exponential improvement provides massive economies of scale to the industry resulting in the rapid decline of computing costs to the consumers. The end result is the ubiquitous impact of digital electronics in all segments of the world economy, producing a force of technological and social change.

The semiconductor industry is widely recognized as a dominant driver of economic growth due to its network effects. The electronics and semiconductor industries, key building blocks of the global economy, combine macro- and micro economics in a variety of ways. From a worldwide base semiconductor market of $213 billion in 2004, the industry enables the generation of nearly $1,200 billion in electronic systems business and $5,000 billion in services, representing close to 10% of the worlds GDP [79]. Because of its staggering rate of price-performance improvement, changes in the market and innovation occur extremely rapidly. One major consequence of this rapid change is that established market strongholds can be displaced very quickly.

Now, five decades later, with increasing evidence of the bubbling up and sustaining of Design 2.0, refreshingly new hardware engineering (along with the influence from the Web. 2.0 software community) creativity is being unleashed [80]. Design 2.0 can be loosely defined as a type of “maker movement” framework where instead of hardware design being inhibited by a complex, bureaucracy-ridden set of corporate design methodologies and standardized industry practices, rapid low-cost experimentation to reduce friction in innovation is encouraged. This movement is assisted by a communal pool of tools (e.g., open source hardware specifications and customizations, hardware hackathons, accelerators, incubators, and crowdfunding sources) and private partnerships (e.g., joint research and development between nimble hardware and solution providers, and the occasional vertically integrated manufacturer), generally benefiting from the proliferation and level-setting of global knowledge and networking through the Internet [80].

When one reinforces the Design 2.0 framework with an analogous new development in the semiconductor technology space known as More than Moore, we begin to see the rapid uprising of potentially disruptive innovations and supply chain models where the boundaries
of ownership and value extraction are blurred. The pervasion of *More than Moore*, where added value to devices is provided by incorporating functionalities that do not necessarily scale according to *Moore’s Law* [2], will influence the development of new integration platforms, both in technology and manufacturing, including test and reliability as well as design and modeling tools capable of handling heterogeneous subsystems. This is especially prominent in the semiconductor assembly and packaging space given its relatively lower costs structure, less complicated technological needs, and much more relaxed U.S. technology export law restrictions, as compared to Si wafer fabrication technology. These factors will naturally lead to more global inclusion, especially from Asia, which is increasingly known as the *mecca* for electronics assembly manufacturing and rapid hardware “hacking” activities.

Disruptive innovation, a term coined by Clayton Christensen [81], describes a process by which a product or service takes root initially in simple applications at the bottom of a market and then relentlessly moves up market, eventually displacing established competitors. Christensen suggests that successful companies often place too much emphasis on customers’ current needs, and fail to adopt new technology or business models that will meet their customers’ unstated or future needs. He argues that such companies will eventually fall behind and be “disrupted”. The idea of the “innovators dilemma” stems from the fact that businesses (especially large, vertically integrated companies) will reject innovations, based on the assumption that customers cannot currently use them or that the profit margin is too low to make a significant dent to their bottom line, thus allowing these ideas with great potential to go to waste until it becomes too late.

Increasingly, with the advantages gained from system level package integration to enable minimal supply chain *friction*, extraction of parts of adjacent businesses, and significant reduction in packaging form-factor, semiconductor packaging technology has evolved from applications in the past with a brute-force, computing focus to current packaging with wireless mobility, big data, and digital convergence focus [3]. While *More Than Moore* packaging technology pursuits often involve new System-In-Package (SIP) and System-On-Package (SOP) concepts that disrupt traditional packaging approaches [see Fig. 1-2] the *Design 2.0* movement additionally gives rise to various rapid experiments to disrupt traditional in-house or out-source assembly and test (OSAT) packaging houses. For example, in order to pack
more into their ultra-thin MacBook Airs, Apple started to solder/pack microchips, memory, and solid-state drives (SSDs) directly into their motherboards without their traditional packaging architecture/housing from the respective suppliers. The same holds true for manufacturers of mobile devices such as smartphones, tablets, phablets, etc. One can only expect large e-commerce companies like Amazon, Alibaba, and Xiaomi, with large data storage needs for their integrated mobile product offerings and their data server farms, to do the same in the near future, if they have not already done so. This approach also makes sense from a strategic intellectual property (IP) point of view, since by experimenting on ways to extract innovation from adjacent supplier technology, they can secure strategic differentiation, whether that is in terms of user experience or their own system-level performance optimization.

The development of Design 2.0 and More Than Moore also help to create fertile ground for Internet of Things (IoT), machine-to-machine (M2M), and wearable technology consumer spaces, where many companies from many industries can imagine, develop, and find innovative growth opportunities. The applications are usually industry agnostic. Because of the “innovator’s dilemma” in that these applications may not fit perfectly well anywhere within the company, exacerbated by relatively lower per unit profit margins, large companies are especially susceptible to their own success trap. At worst, these companies may be completely disrupted in terms of their existing offerings, and at best, a significant portion of their value add may be extracted away by small, nimble competitors, emerging from behind. In the IoT and M2M consumer spaces, disruptive innovations may not attempt to bring better packaging solutions, but rather solutions that are not as good as, but which are simpler, more convenient, and less expensive than existing items.

There is no sure and fast way to prevent the “innovator’s dilemma”. All companies should always keep a pulse on upcoming technologies and applications in their primary and adjacent industries, to maintain or increase competitive advantage. One proven way to help a large company understand and respond better to rapid changes is through the effective deployment of corporate entrepreneurship, by architecting a nimble strategic design while paying attention and taking advantage of cultural and political forces within the organization. Smart, large companies will rely on a bottoms-up approach along the lines of corporate
entrepreneurship to champion disruptive ideas and business opportunities which, at that
time, may not seem to fit perfectly well anywhere within the company. This will enable them
to find their next breakthroughs and not miss out on key inflection points. Organizations
capable of disruptive growth need to be architectured with relentless support from the top
ranks so that the correct deployment of resources, incentive alignment, values, the right
organizational home, the creation of an expert team of movers and shapers, the training of the
troops, and the art of starting something before you need to are all important building blocks
[82]. The ability to recognize the potential of disruptive innovations rather than focusing
merely on how large the “economic pie” is at the present moment is key to integrating the
dots necessary to avoid being disrupted.

7.2.2 Applying software development methodologies to packaging
and assembly technologies

The systems development life cycle (SDLC) is typically used in information systems and
software engineering to describe a process for planning, creating, testing, and deploying an
information system. While different SDLC methodologies are adopted to manage different
levels of complexities, agile software development and rapid prototyping are two models that
can complement the technology development cycle of packaging and assembly technologies.

Agile software development, guided by the Agile Manifesto [83], is a group of software
development methods in which requirements and solutions evolve through collaboration be-
tween self-organizing, cross-functional teams. This helps to promote adaptive planning,
evolutionary development, early delivery, continuous improvement, and encourages rapid
and flexible response to change. By architecting a similar methodology with the assignment
of key technical leaders and product managers interweaving with existing organizational de-
sign, the use of agile development can complement existing development cycles in traditional
packaging and assembly technology groups, so that the next breakthrough in technology or
business model will not be missed. Lightweight agile methods provide a creative escape path
away from the heavily regulated, regimented, micromanaged, and over-incremental method-
ology in traditional semiconductor technology development by avoiding the need to commit
to misleading perceptions of “divisional lines and ownerships”. In addition, it is imperative to form strong, global alliances within the ecosystem of suppliers and customers so that one can continually innovate with them and be involved with potential disruptions early on, in order to course-correct quickly when needed.

Rapid prototyping typically simulates only a few aspects of, and may be completely different from, the final product [84]. In the software industry, prototyping has several benefits. For example, the software designer and implementer can get valuable feedback from the users early in the project. Additionally, the client and the contractor can compare at an early stage if the software produced matches the software specification. It also allows the software engineer some insight into the accuracy of initial project estimates and whether the deadlines and milestones proposed can be successfully met [84]. Rapid prototyping is in fact one of the tools involved in the agile development methodology.

There are two major types of prototyping: throwaway prototyping and evolutionary prototyping. Specifically for assembly and packaging technologies, which involve a small number of processing steps, mainly centering on mechanical engineering, manufacturing, and architectural design on a small scale, throwaway prototyping is better suited to facilitate agile development cycles. Throwaway or rapid prototyping refers to the creation of a working model of various parts of the system at a very early stage, after a relatively short evaluation. The method used in building it is rather informal, with velocity and turn-around time being the most important factors. The model then becomes the starting point from which users can re-examine their expectations and clarify their requirements. When this has been achieved, the prototype model is discarded and the system is formally developed based on the identified requirements [84]. This prototyping model can be applied to joint packaging and assembly development projects forged between customers and suppliers, to allow meaningful technological iterations to flourish in a cost-effective manner.

7.3 What next?

Although extensive characterization, optimization, integration, and operations management of stealth dicing technology have been presented, most of the results are geared toward en-
abling assembled 2D NAND memory die stacks with individual final die thicknesses between 25-46 μm thickness for POR 1Y wafer technologies. Some of the most promising directions where the work described in this dissertation can be improved or addressed to increase its impact are given below, segmented by different estimated timelines based on experience.

7.3.1 Short-term projects

A set of short-term opportunities focusing on the output of the FMEA analysis, examination of factory equipment, capacity, and product intercept anticipated for p-SDBG, development of new DAF and wire-embedding film (WEF) materials, and SD equipment (software and hardware) improvements are recommended.

- Take immediate action based on the developed FMEA to further reduce production risks. These include, but are not limited to, establishing the proper floor production expectations and error-proof recipe controls for critical processes in the p-SDBG integration flow. Also, a taskforce needs to be set up to eliminate the bubble effect seen along the dicing street using the current 165 μm thick backgrinding (BG) tape.

- Analyze factory equipment and process capacity for dicing-before-grinding (DBG) integration flow versus different wafer loading scenarios forecasted for the future, and contrast it with the p-SDBG integration flow in order to plan future capital equipment spending and COO savings. One needs to also ascertain the different flavors of product intercept anticipated for p-SDBG. Proper sensitivity analysis to loading changes below and above forecast, as well as the need for flexibility, should be performed with a strategic decision made by looking at key commonalities and trends.

- Work with materials suppliers to develop new DAF and wire-embedding film (WEF) materials that are optimized for die separation. This is important to extend p-SDBG integration to stacked die systems using a variety of DAF/WEF combinations.

- Work with the SD equipment supplier to improve software and hardware systems. For example, given the importance of laser pulse energies as a major modulator in optimizing the SD process, the flexibility of existing laser power attenuation and the range
of offered optical systems need to be improved. Software and digital signal processing improvements can also be made to better manage wafer backside surface focusing and frontside pattern recognition requirements for a variety of wafer technologies and products.

7.3.2 Medium to long-term projects

A set of medium and long-term efforts are recommended, in order to maximize the benefits of stealth dicing and expand its versatility in various product mix, including future NAND technologies and other unique device applications.

- Work on developing a POR SD process and integration flow for future 3D NAND memory and MEMS technologies, including those using low-\( \kappa \) dielectrics and/or different crystal orientations.

- Work on developing a POR SD process and integration flow to enable the fabrication of defect free, singulated dies with thicknesses larger than 100 \( \mu m \).

- Work on developing a POR SD process and integration flow based on SD laser incident on the wafer frontside as opposed to the backside. This will generate a variety of future product intercept possibilities.

- Develop a robust design-for-manufacturing (DfM) set of design rules to maximize the benefits of SD technology, especially towards the pursuit of dicing street width reduction in combination with new TEG designs, in order to significantly reduce per unit costs. This will be extremely beneficial given the zero kerf loss of SD processing.

- Co-develop an integrated and modular SD tool platform, with the possibilities of an inline backgrinding tool and DAF laser grooving module incorporated to significantly enhance the flexibility of the process.
Bibliography


, display-, and bio-industry,” in Lasers and Applications in Science and Engineering,

[65] B. Richerzhagen, D. Perrottet, and Y. Kozuki, “Dicing of wafers by patented water-
jet-guided laser: the total damage-free cut,” in Proc. of the 65th The Laser Materials

[66] P. Chall, “ALSI’s low power multiple beam technology for high throughput and low
damage wafer dicing,” in Proc. of the 65th The Laser Materials Processing Conference,

[67] J. S. for Mechanical Engineers, JSME Data Book: Heat Transfer. Tokyo, Japan:

[68] Y. S. Touloukian, R. W. Powell, C. Y. Ho, and P. G. Klemens, “Thermophysical prop-
ties of matter-the TPRC data series. volume 1. thermal conductivity-metallic elements

and dicing technology for thin wafer,” in 3D System Integration, 2009. 3DIC 2009.


[73] A. L. Hartzell, M. G. Da Silva, and H. R. Shea, MEMS reliability. No. EPFL-BOOK-

[74] M. Karnezos, “3D packaging: Where all technologies come together,” in Electronics


to Statistics and Data Analysis Using JMP, Fifth Edition. Cary, NC: SAS Institute,


## Appendix A

### Glossary

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFM</td>
<td>Atomic Force Microscopy</td>
</tr>
<tr>
<td>BOX</td>
<td>Buried Oxide</td>
</tr>
<tr>
<td>BP</td>
<td>Beam (Optical) Parameter</td>
</tr>
<tr>
<td>BS</td>
<td>Backside</td>
</tr>
<tr>
<td>BSC</td>
<td>Backside Chipping</td>
</tr>
<tr>
<td>CMOS</td>
<td>Complementary Metal Oxide Semiconductor, a technology for making low power integrated circuits</td>
</tr>
<tr>
<td>CMP</td>
<td>Chemical Mechanical Polishing</td>
</tr>
<tr>
<td>COO</td>
<td>Cost Of Ownership</td>
</tr>
<tr>
<td>CSPs</td>
<td>Chip-Scale Packages</td>
</tr>
<tr>
<td>CTQ</td>
<td>Critical-To-Quality</td>
</tr>
<tr>
<td>DAF</td>
<td>Die Attach Film</td>
</tr>
<tr>
<td>DAG</td>
<td>Dicing after grinding</td>
</tr>
<tr>
<td>DBG</td>
<td>Dicing before grinding</td>
</tr>
<tr>
<td>DDS</td>
<td>Die Separation</td>
</tr>
<tr>
<td>DET</td>
<td>Detection</td>
</tr>
<tr>
<td>DF</td>
<td>Defocus parameter</td>
</tr>
<tr>
<td>DfM</td>
<td>Design for Manufacturing</td>
</tr>
<tr>
<td>DFZ</td>
<td>Defect Free Zone</td>
</tr>
</tbody>
</table>
DI water Deionized water
DMAIC Define, Measure, Analyze, Improve, Control
DOE Design Of Experiment
DT Dicing Tape
ESD Electrostatic Discharge
FMEA Failure Modes Effects Analysis
FOUP Front Opening Unified Pod
FS Frontside
HAZ Heat Affected Zone
HTCCs High-Temperature Cofired Ceramics
HVM High Volume Manufacturing
I/O Input/Output
ICs Integrated Circuits
IDOV Identify, Design, Optimize, Validate
IDOV+C Identify, Design, Optimize, Validate plus Control
ILD Interlayer Dielectric
IoTs Internet of Things
IP Intellectual Property
IPDS Integrated Product Delivery System
ITRS International Roadmap for Semiconductors
KGD Known-Good-Die
LTCCs Low-Temperature Cofired Ceramics
M2M Machine to Machine
MCM Multi-Chip Module
MEMS Microelectromechanical Systems
More Moore A semiconductor roadmapping trend where a traditional scaling “technology push” approach is used, assuming the validity of Moore’s Law
More Than Moore A semiconductor roadmapping trend where added value to devices is provided by incorporating functionalities that do not necessarily scale according to Moore’s Law

200
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition/Explaination</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPW</td>
<td>Multi-Project-Wafer</td>
</tr>
<tr>
<td>MTST</td>
<td>Memory Test and System Test</td>
</tr>
<tr>
<td>Nd:YAG</td>
<td>Neodymium-doped Yttrium Aluminium Garnet</td>
</tr>
<tr>
<td>OCC</td>
<td>Occurrence</td>
</tr>
<tr>
<td>OSAT</td>
<td>Out-Source Assembly and Test</td>
</tr>
<tr>
<td>PCM</td>
<td>Process Control Monitoring</td>
</tr>
<tr>
<td>PLE</td>
<td>Pulse Laser Energy</td>
</tr>
<tr>
<td>POR</td>
<td>Process-Of-Record</td>
</tr>
<tr>
<td>p-SDBG</td>
<td>partial-Stealth Dicing Before Grinding</td>
</tr>
<tr>
<td>QFD</td>
<td>Quality Function Deployment</td>
</tr>
<tr>
<td>RMS</td>
<td>Root Mean Square</td>
</tr>
<tr>
<td>RPN</td>
<td>Risk Priority Number</td>
</tr>
<tr>
<td>RtR</td>
<td>Run-to-Run</td>
</tr>
<tr>
<td>SD</td>
<td>Stealth Dicing</td>
</tr>
<tr>
<td>SDAG</td>
<td>Stealth Dicing After Backgrinding</td>
</tr>
<tr>
<td>SDBG</td>
<td>Stealth Dicing Before Grinding</td>
</tr>
<tr>
<td>SDLC</td>
<td>Systems Development Life Cycle</td>
</tr>
<tr>
<td>SDSS</td>
<td>SanDisk Semiconductor Shanghai</td>
</tr>
<tr>
<td>SEM</td>
<td>Scanning Electron Microscopy</td>
</tr>
<tr>
<td>SEV</td>
<td>Severity</td>
</tr>
<tr>
<td>SIP</td>
<td>System-In-Package</td>
</tr>
<tr>
<td>SIPOC</td>
<td>Customer Product Map: Suppliers, Inputs, Process, Outputs, Customers</td>
</tr>
<tr>
<td>SKUs</td>
<td>Stock Keeping Units</td>
</tr>
<tr>
<td>SOI</td>
<td>Silicon-On-Insulator</td>
</tr>
<tr>
<td>SOP</td>
<td>System-On-Package</td>
</tr>
<tr>
<td>SPC</td>
<td>Statistical Process Control</td>
</tr>
<tr>
<td>SSD</td>
<td>Solid-State Drive</td>
</tr>
<tr>
<td>TEGs</td>
<td>Test Element Groups</td>
</tr>
<tr>
<td>TSC</td>
<td>Topside Chipping</td>
</tr>
<tr>
<td></td>
<td>201</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>TSVs</td>
<td>Through-silicon-vias</td>
</tr>
<tr>
<td>VOC</td>
<td>Voice of the Customer</td>
</tr>
<tr>
<td>WEF</td>
<td>Wire Embedded Film</td>
</tr>
<tr>
<td>WIW</td>
<td>Within Wafer</td>
</tr>
<tr>
<td>WtW</td>
<td>Wafer-to-Wafer</td>
</tr>
<tr>
<td>XRD</td>
<td>X-Ray Diffraction</td>
</tr>
</tbody>
</table>
Appendix B

Publications


